MATKADATA-ANALYYSI

Matkalaskujärjestelmät ja hiilijalanjälki

HAMK
HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

Ammattikorkeakoulututkinnon opinnäytetyö

Tietojenkäsittelyn koulutus, Hämeenlinnan korkeakoulukeskus

syksy, 2021

Leena Pasanen

HAMK
HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

Tietojenkäsittelyn koulutus — Tiivistelmä

Hämeenlinnan korkeakoulukeskus

Tekijä	Leena Pasanen	Vuosi 2021
Työn nimi	Matkadata-analyysi Matkalaskujärjestelmät ja hiilijalanjälki	
Ohjaaja	Miranda Kosova-Alija	

Opinnäytetyön tarkoituksena oli selvittää Power BI avulla lento- ja automatkustamisesta aiheutuneen hiilijalanjäljen muutokset ja kokonaisjäljen kehittyminen Hämeen ammattikorkeakoulussa. Datan tutkinta on rajattu vuosiin 2010–2020. Erityisenä kiinnostuksen kohteena on jaottelu toimipisteiden välisiin matkoihin ja työmatkoihin. Tavoitteena oli löytää loppukäyttäjälle ymmärrettävä visualisointimenetelmä. Opinnäytetyön toimeksiantaja oli Hämeen ammattikorkeakoulu.

Opinnäytetyön teoreettisessa osuudessa tarkasteltiin hiilijalanjälkilaskentaa, data-analytiikan malleista tutkimusdata analyysia ja CRISP-DM analyysin etenemismalleja. Teoria osuudessa käytiin läpi myös käytetyimpiä data-analytiikan alustoja sekä seikkoja, joita datan visualisoinnissa tulee ottaa huomioon. Opinnäytetyö on toiminnallinen ja sen tuloksena toteutettiin data-analyysi Power BI -alustalla, jossa visualisointi oli merkittävässä roolissa.

Johtopäätöksenä voidaan todeta, että järjestelmien tulee täyttää yritysten tarve vaadittavista analyyseista ja raporteista. Järjestelmiä tulee kehittää vastaamaan korkeakoulun tarpeita sekä vastaamaan muuttuvia lainsäädäntöjä. Hiilijalanjälkilaskenta on helposti otettavissa osaksi matkalaskujärjestelmää, kunhan asiakkaat sitä osaisivat vaatia. HAMK oli tyytyväinen kehittämistyön tuloksiin ja valmis ideoimaan työtä eteenpäin.

Avainsanat: Data-analytiikka, CRISP-DM, Power BI, Tiedon visualisointi, Hiilijalanjälki

Sivut 43 sivua ja liitteitä 5 sivua

HAMK
HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

Degree Programme in Business Information Technology
Hämeenlinna University Centre

Abstract

Author Leena Pasanen
Year 2021

Subject Travel data analysis
Travel invoicing systems and the carbon footprint

Supervisor Miranda Kosova-Alija

The purpose of the thesis was to use Power BI to find out the changes in the carbon footprint caused by air and car travel and the development of the overall footprint at Häme University of Applied Sciences. The examination of the data is limited to the years 2010–2020. A special object of interest is division into inter-office and business trips. The goal was to find a visualization method that was understandable to the end user. The thesis was commissioned by Häme University of Applied Sciences.

The theoretical part of the thesis examined carbon footprint calculation, research data analysis from data analytics models, and CRISP-DM analysis progress models. The theory section also covered the most used data analytics platforms as well as aspects that should be taken into account in data visualization. The thesis is functional and as a result data analysis was performed on the Power BI platform, where visualization played a significant role.

In conclusion, the systems should meet the need of companies for the required analyzes and reports. The systems must be developed to meet the needs of the university and to meet changing legislation. Carbon footprint accounting can be easily integrated into a travel invoice system as long as the customers understand require it. HAMK was satisfied with the results of development work and ready to come up with ideas for the work ahead.

Keywords: Data Analytics, CRISP-DM, Power BI, Data Visualization, Carbon footprint

Pages 43 pages and appendices 5 pages

KÄYTETYT LYHENTEET JA SANASTO

Algoritmit	ovat yksityiskohtainen kuvaus siitä, miten tehtävä tai prosessi suoritetaan.
Avoin lähdekoodi	(open source) on ohjelmiston tai kirjaston lähdekoodi, joka on julkisesti saatavilla ja sitä voi kuka tahansa hyödyntää haluamiinsa tarkoituksiin.
Big Data	on nopeasti liikkuvia suuria tietomassoja, joiden rakenne voi muuttua.
CO2ekv	on hiilidioksidiekvivalentti, joka kuvaa eri kasvihuonekaasupäästöjen ilmastoa lämmittäviä vaikutuksia muuttamalla ne vastaamaan hiilidioksidipäästöjä.
CRISP-DM	tarkoittaa toimialan välistä tiedon louhintaprosessia. CRSP-DM menetelmä tarjoaa jäsennellyn lähestymistavan tiedonlouhintaprojektin suunnitteluun.
Data Lake	on keskitetty arkisto, jonka avulla voi tallentaa kaikki jäsennelty ja strukturoimaton tieto missä tahansa mittakaavassa.
Gartner	on erittäin tunnettu kansainvälinen ICT-alan tutkimus- ja konsultointiyritys, jonka pääkonttori sijaitsee Connecticutussa Yhdysvalloissa. Gartner liikeideana on myydä maa- tai maanosakohtaisia markkinatietoja sekä konsultointipalveluita. Gartner julkaisemia lehtiuutisia ja lukuja seurataan ICT – alan lehdistössä tarkkaan.

GHG Protocol	Greenhouse Gas Protocl on kattojärjestö, joka tarjoaa standardeja, ohjeita, työkaluja ja koulutusta yrityksille ja hallituksille mittaamaan ja hallitsemaan ilmaston lämpenemisestä aiheutuvia päästöjä.
GRI raporttijärjestelmä	on maailmanlaajuinen aloite yhtenäisten yhteiskuntavastuun raporttien käytäntöjen kehittämiseksi.
Data Warehouse	eli tietovarastot ovat eräänlainen tiedonhallintajärjestelmä, joka on suunniteltu mahdollistamaan ja tukemaan liiketoimintatiedon (BI) toimintaa, erityisesti analytiikkaa. Tietovarastot ovat tarkoitettu kyselyjen ja analyysien suorittamiseen, ja ne sisältävät usein suuria määriä historiallisia tietoja.
Data Mart	ovat tietovaraston osajoukko, joka keskittyy tiettyyn liiketoiminta-alueeseen, osastoon tai aihealueeseen. Data Martit asettavat tietyn tiedon tietyn käyttäjäryhmän saataville, jonka ansiosta käyttäjät voivat nopeasti käyttää tärkeitä oivalluksia tuhlaamatta aikaa koko tietovaraston etsimiseen.
Data pipeline	(Dataputki) on hallittu toimintokokonaisuus datan jalostukseen ja liiketoiminta-arvoa tuottavien datatuotteiden luontiin. Datatuote voi olla esimerkiksi raportti tai koneoppimisalgoritmin tuottama ennuste, jota käytetään rajapinnan kautta.
ETL/ELT	ETL tarkoittaa prosessia, joka poimii, muuntaa ja lataa dataa useista lähteistä tietovarastoon tai muuhun yhdistettyyn tietovarastoon. Suurin ero ETL:n ja ELT:n välillä on toimintajärjestyksen ero. ELT kopioi tai vie tiedot lähdepaikoista, mutta sen sijaan, että ne ladattaisiin pysähdysalueelle muutosta

	varten, se lataa raakatiedot suoraan kohdetiedostoon muunnettavaksi tarpeen mukaan.
Tietomallit	ovat tiedon rakenteen määritelmä.
Tiedonlouhinta	on menetelmä, jolla suuresta datamassasta karsitaan oleellinen.
Koneoppiminen	on tekoälyn alaryhmä, jonka tarkoituksena on saada ohjelmisto toimimaan paremmin pohjatiedon ja käyttäjän toiminnan perusteella. Koneoppimisessa kone kykenee parempiin tuloksiin itsenäisesti ilman erillistä toimintaohjetta.

Sisällys

Kuvat

Ohjelmakoodit

Taulukot

Liitteet

1 Johdanto

Yhdellä kuvalla voi kertoa enemmän kuin pitkällä tarinalla. Visuaalisessa muodossa on helpompi hahmottaa vuosien takaista dataa ja saada välitettyä nopeasti paljon informaatiota. Hyvän visualisoinnin takana on kuitenkin pitkä analyysiprosessi, jossa suurista datamääristä muokataan esille haluttu tieto.

Ilmastonmuutos on yksi suurimmista globaaleista puheenaiheista ja sitä voidaan pitää yhtenä suurimmista haasteista, joka koskettaa kaikkia organisaatioita ja kaikkia yksilöitä. Organisaatioiden omien ilmastopäästöjen aito tietämys antaa motivaatiota vähentää hiilijalanjälkeä ja talkoisiin tarvitaan kaikkien organisaatioiden eri yksiköiden panosta. Kun mennyt tunnetaan, on helpompi asettaa tavoitteita tulevaan. Tämän opinnäytetyön tavoitteena on visuaalisessa muodossa, Microsoft Power BI:ta apuna käyttäen, hahmottaa HAMK henkilöstön lento- ja automatkustamisen hiilijalanjälkeä ja sen muutoksia pitkällä aikavälillä.

HAMKilla on matkustuksen dataa kertynyt kahdesta eri matkalaskujärjestelmästä usean vuoden takaa, ja tavoitteena on analysoida hiilijalanjäljen muutokset ja kokonaisjäljen kehittyminen Power BI:n avulla. Erityisenä kiinnostuksen kohteena on HAMKin toimipisteiden välisen ns. toimipistematkojen hiilijalanjälki. Toimipisteiden väliset matkat ovat HAMKin 7:n eri kampuksen välisiä matkoja. Muut matkat käsittävät esim. asiakaskäynnit ja edustusmatkat. Datan tutkinta on rajattu vuosiin 2010–2020. Tavoitteena oli löytää selkeä visualisointimenettely, joka oli loppukäyttäjän ymmärrettävissä.

Tutkimuskysymykset:

- Miten matkadatasta saa yhtenäisen ja vertailukelpoisen?
- Millaisia laskentakäytänteitä hiilijalanjäljen laskentaan on käytössä?
- Miten matkadatan saa esitettyä visuaalisessa muodossa ymmärrettävästi ja reaaliajassa?
- Miten matkalaskujärjestelmät saadaan tukemaan hiilijalanjälkilaskentaa?

2 Ilmastomuutos

Ilmastonmuutosta pidetään historian suurimpana maailmanlaajuisena ympäristökriisinä. Kasvihuonekaasujen, joihin lukeutuvat mm. hiilidioksidi ja metaani, kasvaneet määrät ilmakehässä ovat aiheuttaneet ilmaston lämpenemistä, kuivuutta, rankkasateita sekä tulvien yleistymistä. Ihmisen toiminnan seurauksena kasvihuonekaasujen määrää ilmakehässä on lisääntynyt. (Ilmasto.Org, n.d.)

Maapallon lämpeneminen on jo lähtenyt käyntiin. Mallit ennustavat, että maapallon keskilämpötila tulee nousemaan lähivuosikymmeninä ja kymmenessä vuodessa kasvu olisi 0,2 astetta. Lämpötilan nousu ei tule olemaan tasaista. 1 asteen lämpeneminen on saavutettu jo vuonna 2017. 1,5 asteen nousu saavutetaan vuosien 2030–2052 välillä, jos ei ryhdytä voimakkaisiin vähennystoimenpiteisiin. (Ilmasto-Opas.Fi, n.d.)

Pariisin ilmastosopimuksessa vuonna 2015 on kirjattuna 2. artiklaan tavoite, että maapallon keskilämpötilan nousu on pidettävä selvästi alle kahden asteen. Sopimuksen osapuolten on ryhdyttävä välittömiin tehtäviin, että lämpeneminen saadaan alle 1,5 asteen. Sopimuksen osapuolet ovat sitoutuneet päästöjen vähentämiseen sekä nopeisiin ja tehokkaisiin toimiin ilmastonmuutokseen sopeutumiseksi. (Pariisin sopimus, 2. artikla, 2016)

2.1 Hiilijalanjälki

Hiilijalanjälki kuvaa ilmastovaikutusta numeerisin arvoin. Hiilijalanjälki kuvaa yksittäisen henkilön, yrityksen, tuotteen, toiminnan tai palvelun aiheuttamaa ilmastokuormaa. Hiilijalanjäljellä on tarkoitus kuvata toiminnan, palvelun ja tuotteen koko elinkaaren aikana aiheutuneita päästöjä. Kasvihuonekaasupäästöillä mitataan vaikutusta, joita syntyy prosessin eri kohdissa joko välillisesti tai suoraan. Näissä laskelmissa on mukana käytöstä aiheutuvien päästöjen lisäksi tuotantoprosessin päästöt. (Seppälä et al., n.d.)

Keskivertokansalaisella hiilijalanjälki on noin 10 300 kg CO2ekv/vuosi. Keskiverto suomalaisen hiilijalanjälki jakautuu suunnilleen seuraavalla tavalla: matkustus 29%, ruoka 18 %, asuminen 20 %, ja muu kulutus 33 %. (Sitra, n.d.)

Hiilidioksidiekvivalentti tai hiilidioksidipäästöjen määrä ovat arvoja, joilla hiilijalanjälki ilmoitetaan. Hiilidioksidiekvivalentti sisältää eri kasvihuonekaasujen yhteenlasketut vaikutukset ja vertaa niitä hiilidioksidin vaikutuksiin. Päästöjen määrä ilmoitetaan suuruuden mukaan joko grammoina, kilogrammoina tai tonneina. Hiilijalanjälki ilmoitetaan usein per vuosi. Käytämme tällä hetkellä yli 1,5 kertaa planeettamme resurssit. (OpenCO2.Net, n.d)

EU-direktiivi NFR (Non-financial reporting) vaatii, että yli 500 henkilöä työllistävien yritysten ja organisaatioiden tulee laatia vastuullisuusraportti, josta selviää, kuinka heillä huomioidaan ympäristö, talous ja sosiaaliset tekijät. Raportista tulee käydä ilmi, 1. ja 2. luokan päästöt, mutta 3. luokan päästöjen raportointi ei ole pakollista. 3. luokan päästöt ovat kokoluokaltaan usein suuria ja muodostavat yleensä suuren osan koko yrityksen päästökertymistä. (Green Carbon, n.d.)

Raportointia varten tarvitaan standardeja eli sääntöjä, jotta raportointi on yhdenmukaista ja vertailukelpoisia toisiin yrityksiin ja organisaatioihin nähden. Säännöt ohjaavat sitä, kuinka jokin asia tulisi tehdä. Hiilijalanjäljen laskentaa ohjaavia ympäristöstandardeja on lukuisia, jotka ovat kansainvälisesti hyväksyttyjä. (Green Carbon, n.d.)

Green Gas Protocol lyhennettynä GHG-Protocol eli kasvihuonekaasuprotokolla on yksi suosituimmista käytössä olevista ympäristövaikutusten laskemiseen kehitetyistä standardeista. GRI- raportointijärjestelmä on osa laajempaan raportointijärjestelmään, jonka painopisto on ympäristön ohella myös yritysten yhteiskunnallinen ja taloudellinen vaikuttaminen. (Green Carbon, n.d.)

Kasvihuonekaasuprotokollassa päästöt on jaoteltu kolmeen alueeseen:

- 1. luokan alueeseen sisältyvät sellaiset päästöt, joita yritys voi itse kontrolloida ja vaikuttaa niihin suoraan. Päästöt ovat seurausta yrityksen omasta

toiminnasta, ja ne ovat syntyneen yleensä yrityksen omissa toimitiloissa. Esimerkiksi omien moottoriajoneuvojen polttoainepäästöt.

- 2. luokan alueeseen sisältyvät sellaiset päästöt, joita yritys kuluttaa epäsuorasti ja maksaa niistä hinnan. Esimerkiksi sähkön ja lämmön tuotannosta aiheutuneet päästöt.
- 3. luokan alueeseen kuuluvat kaikki epäsuorat päästöt, joita on syntynyt myytyjen tuotteiden loppukäytöstä ja tavaroiden ja palveluiden hankinnasta. Esimerkiksi jätehuollon, vesihuollon, logistiikan, materiaalien hankinnan päästöt. (Green Carbon, n.d.)

2.2 Hiilijalanjäljen laskentaperiaatteet

Hiilijalanjäljen laskenta pohjaa ohjeistuksiin ja kansainvälisiin standardeihin sekä näiden lisäksi myös toimialojen uusimpiin suosituksiin. Yleisimmät standardit, jotka ovat käytössä kasvihuonekaasuprotokollan laskennassa ovat GHG Protocol mukaiset standardit. Laskennoissa on yleensä mukana myös muista erilaisia laskentastandardeja ja -ohjeita, joita hyödynnetään soveltuvin osin kuten ISO 14040, 14064, 14044, 14067, 14069 ja PAS2050. (Ranganathan et al., n.d.)

Standardi ISO 14067 sisältää mm. hiilijalanjäljen laskentaa sekä raportointiohjeistuksia. ISO-14064-1 määrittelee 5 pääperiaatetta (taulukko 1), joiden tehtävänä on varmistaa, että kasvihuonekaasulaskelmista saadut tulokset ovat todenmukaisia.

Taulukko 1 5 pääperiaatetta kasvihuonekaasulaskelmista todenmukaisuudelle. (Ranganathan et al., n.d.).

Periaate	ISO 14067 Standardi
Merkitys	Selvitykseen valitaan sellaiset kasvihuonekaasujen päästölähteet, tiedot ja käytettävät menetelmät, jotka parhaiten palvelevat selvityksen käyttötarkoitusta.
Täydellisyys	Selvitykseen sisällytetään kaikki merkittävät kasvihuonekaasupäästöt ja -poistumat. Mukaan tulisi sisällyttää kaikki prosessit ja virrat, joiden vaikutus hiilijalanjälkeen on merkittävä. Merkittävän päästölähteen rajaaminen ulkopuolelle vääristää hiilijalanjälkiselvityksestä saatuja tuloksia merkittävästi, jolloin se ei yllä standardin vaatimalle tasolle.
Johdonmukaisuus	Menetelmiä ja informaatiota käytetään siten, että saatuja tuloksia pystytään vertailemaan keskenään (ISO 14064. Greenhouse gases). Kaikki tulokset tulisi tällöin ilmoittaa samoissa mittayksiköissä määriteltyjen tavoitteiden ja soveltamisalan mukaisesti.
Tarkkuus	Hiilijalanjälki lasketaan tarkasti ja huolehditaan siitä, että siinä mahdollisesti ilmentyvät epävarmuudet ovat mahdollisimman vähäisiä (ISO 14064. Greenhouse gases). Tämä on luonnollisesti tärkeää saatujen tulosten luotettavuuden kannalta.
Läpinäkyvyys	Kasvihuonekaasuihin liittyvä informaatio raportoidaan avoimesti ja asianmukaisesti. Tarkoituksena hiilijalanjälkiselvityksessä on esittää kaikki käytetyt menetelmät ja tietolähteet totuudenmukaisesti ja ymmärrettävästi. Tällä pyritään siihen, että tehdyn selvityksen voi luottaa olevan sellainen, kun se raportissa tuodaan ilmi. (ISO 14067. Kasvihuonekaasut)

2.3 Hiilifiksu järjestön laskuri

Helsingin yliopiston Metsätieteiden osaston toteuttama ja Sitran rahoittamassa Hiilifiksu järjestö – hankkeessa tuetaan järjestöjä kohti ilmastovastuullisempaa toimintatapaa. Syksyllä 2018 kehitettiin järjestöille suunnattu hiilijalanjälkilaskuri osaksi järjestöjen raportointia. Laskurin tavoitteena on konkreettisesti tuoda esille havainnot siitä, miten järjestöjen hiilijalanjälki muodostuu ja toimia työkaluna hiilijalanjäljen pienentämisessä. Laskurin ajanjakso on vuosi ja lopputulos ilmoitetaan hiilidioksidiekvivalentteina. Laskurin tuottamat luvut ovat vain suuntaa antavia arvoja eri toiminnoista syntyneistä päästöistä. Laskurissa on eri taulukot hankinnoille, matkustamiselle, jätteille, hankintojen, energian, palveluiden ja tapahtumien hiilijalanjäljen laskemiseen. Toiseksi suurin tekijä on usein

matkustus, joka aiheutuu lentomatkustuksesta sekä työmatkaliikenteestä. Erityisesti lentomatkustus aiheuttaa suuret päästöt. (University of Helsinki, 2018)

2.4 Hiilineutraali Hämeen ammattikorkeakoulu

Hämeen ammattikorkeakoululla on laadittu kestävän kehityksen ohjelma ja tahtotilana on olla hiilineutraali vuoteen 2030 mennessä. Kestävän kehityksen toimintakokonaisuutta seurataan systemaattisesti esim. ammattikorkeakoulujen yhteisen hiilijalanjälkimittarin avulla. (Hämeen ammattikorkeakoulu, n.d.)

Vuodesta 2020 eteenpäin korkeakoulujen tulee ilmoittaa laskelmat, toiminnasta aiheutuvista hiilidioksidipäästöistä opetus- ja kulttuuriministeriöön. Vuonna 2020 Arene (ammattikorkeakoulujen rehtorineuvosto ry) valmisteli laskentamallia korkeakoulujen päästöjen laskentaa varten, jolla pyydetyt tiedot toimitettiin ministeriöön. Laskentamallina käytetään Hiilifiksu järjestön laskuria. (Arene, 2020)

3 Data-analytiikka

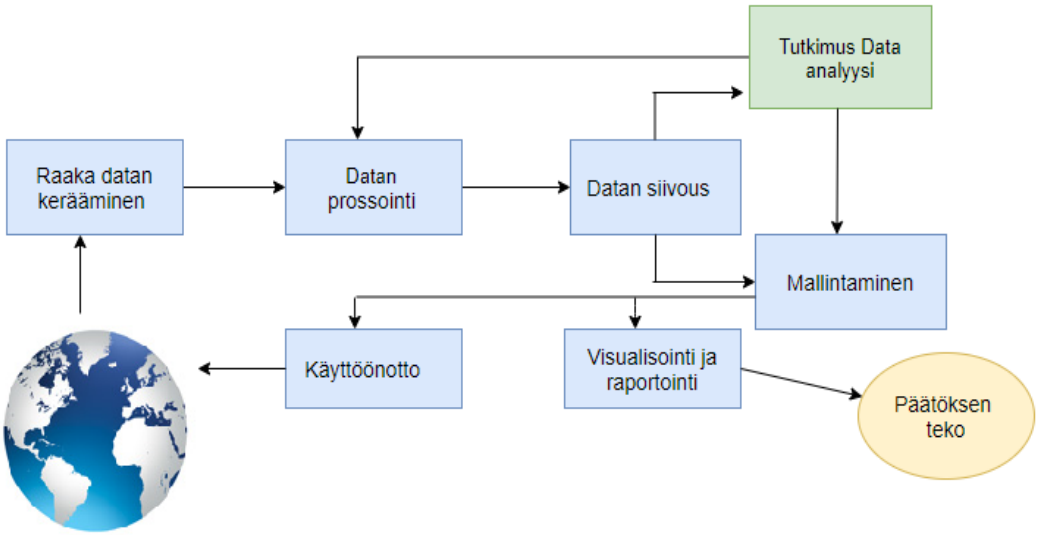
Data-analytiikkaa pidetään prosessina, jonka tavoitteena on muodostaa kerätystä datasta halutun näköistä, käyttäen hyväksi kehitettyjä analytiikkaohjelmistoja ja eri matemaattisia kaavoja. Data-analytiikka tarkoittaa menetelmiä, joilla raakadatasta saadaan muodostettua informaatiota päätöksenteon tueksi. Analytiikalla on tarkoitus kuvata sekä nykyistä aikaa, että ennustaa tulevaa. Data-analytiikan avulla voidaan esimerkiksi ennustaa Covid-19 taudin leviämistä ja rokotusstrategian pitkäaikaista vaikutusta. (Kaearney, 2021)

Data-analytiikan avulla on mahdollista seurata tavoitteiden, toimintatapojen tai säädösten toteutumista. Analytiikka hyödyntää laskennallisia malleja, jotka perustuvat eri käyttötarkoitukseen ja lähtökohtiin sekä rikastettuun tietopohjaan. (Markkula & Syväniemi, 2015, s.61)

3.1 Tutkimusdatan analyysimalli eli Exploratory Data Analysis (EDA)

Piirtäminen ei yleensä ole vaikeaa, vaan se miten aloittaa piirtämisen tyhjälle valkoiselle paperille. Samoin datatieteessä on vaikea päästä alkuun tietojoukon vastaanottamisen jälkeen, kun tietoa on niin paljon. Tutkimusdata analyysimalli eli Explodatory Data Analysis (EDA) on lähestymistapa tietojoukkojen analysointiin pääpiirteiden tiivistämiseksi, usein visuaalisilla menetelmillä. Kuvan 1 tietojen keräämisen jälkeen tieto prosessoidaan ja puhdistetaan käytettävään muotoon, jonka jälkeen voidaan siirtyä tutkimusdata analyysin puolelle. Tutkimusdata analyysin jälkeen voidaan palata tarvittaessa myös takaisin datan prosessointiin. Analyysillä pyritään ymmärtämään tietoa ja löytämään vihjeitä tiedosta, muotoilla oletuksia ja hypoteeseja mallinnuksen avulla. (Weng, 2019)

Kuva 1 Exploratory data Analysis mukaillen. (Weng, 2019)



Tutkimusdata- analyysia (EDA) on erinomainen menetelmä yrityksien käyttöön, sillä

- Pääpaino on visuaalisen analyysin suorittamisessa, kuten tiivistelmien tekeminen kerätystä datasta.
- Käytetään liiketoiminnan lisäämisen ymmärtämiseen.
- Antaa ymmärrettävää tietoa tehdä liiketoimintapäätöksiä.

(Weng, 2019)

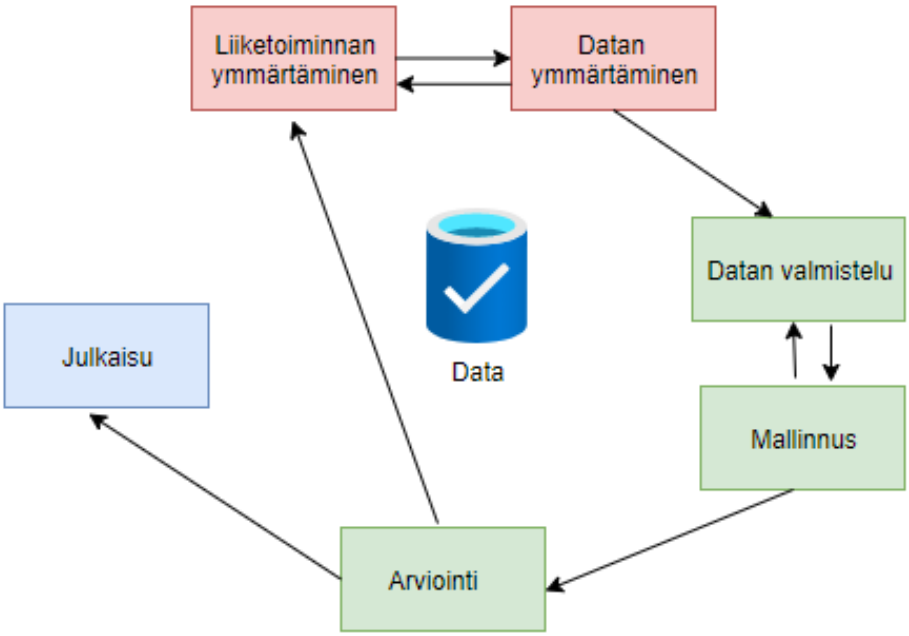
3.2 CRISP-DM analytiikan etenemismalli

Analytiikan etenemisen standardimallina pidetään CRISP-DM (Cross Industry Standards Process for Data Mining) mallia. Malli on helppotajuinen lähestymistapa analytiikan viemiseen liiketoimintalähtöisesti ja järjestelmällisesti tuotantoon. Mallin tarkoituksena on kuvata analytiikan tuotantoprosesseja, erilaisia tiedonlouhinnan työvaiheita sekä kytkeä analytiikka liiketoimintalähtöisesti ja järjestelmällisesti tuotantoon. Prosessin vaiheet eivät välttämättä seuraa tiiviisti toisiaan, vaan ovat usein omia iteroitavia kokonaisuuksia. (Markkula & Syväniemi, 2015, ss. 95–97)

CRISP-prosessin työryhmän perustajajäseniin kuuluvan Colin Shearerin mukaan CRISP- malli luotiin tiedonlouhinnan hyödyntämisen edistämiseksi. ”Tavoitteena on luoda selkeä ja

helposti ymmärrettävä prosessimalli, joka kuvaa vaadittavat tiedonlouhintaprojektin työvaiheet. Tiedonlouhinta ei ole vain tekninen tehtävä, vaan se linkittyy suoraan liiketoiminnan tavoitteisiin ja prosesseihin” Shearer painottaa. (Shearer, 2020)

Kuva 2 The CRISP-DM analytiikan etenemismalli mukaillen. (Shearer, 2020)



3.2.1 Liiketoiminnan ymmärrysvaihe

Liiketoiminnan ymmärrysvaiheessa määritellään muutamia kysymyksiä (käyttötapauksia), johon prosessin avulla pyritään löytämään ratkaisu. Datan analysoinnille on aina jokin syy tai tavoite. Tavoitteiden määrittely on lopputulosten hyödyntämisen kannalta tärkeää vaihe. Ymmärrysvaiheessa tulisi myös saada yleiskuva käytettävissä olevista ja vaadituista resursseista sekä suunta seuraaville vaiheille. Tekemisen laajuus ja analysoinnin kustannukset tulisi myös arvioida ja mitoittaa oikein ymmärrysvaiheessa. (Elsevier B.V., 2021)

3.2.2 Datan ymmärrysvaihe

Tietojen kerääminen tietolähteistä, niiden tutkiminen ja kuvaaminen sekä tietojen laadun tarkistaminen ovat olennaisia tehtäviä datan ymmärrysvaiheessa. Datan ymmärrysvaiheessa validoidaan käyttötapaukset ja selvitetään vastaako käytössä oleva data siihen mitä analysoinnilta lähdetään hakemaan. Tärkeätä on pohtia onko käyttötapaukset mahdollista toteuttaa olemassa olevan datan avulla. Dataymmärrystä voidaan hankkia kyselyillä, esimerkki tietorivien käyttämisellä, asiantuntijoiden haastatteluilla, raporteilla sekä datan visualisoinnilla. Ymmmärrysvaihe edellyttää riittäviä datan luku- ja tarkasteluoikeuksia. Vaiheen keskeisiä tavoitteita ovat datamuotojen ja datan sisällön ymmärtäminen sekä datan laadun selvitys. (Markkula & Syväniemi, 2015, ss. 95–96, Elsevier B.V., 2021)

3.2.3 Datan valmisteluvaihe

Valmisteluvaihe on tulosten oikeellisuuden kannalta keskeinen ja usein kaikista eniten aikaa vievin työvaihe. Usein valmisteluvaihe vie 50-70% koko projektin ajasta ja resursseista. Datan valmisteluvaiheessa työskennellään analytiikan työkaluilla datan saattamiseksi mallinnukselle vastaavaan muotoon. Datan valmisteluvaiheen työvaiheisiin liittyy lukuisia toimenpiteitä: siivous, yhdistäminen, aineiston karsinta, ominaisuuksien luonti ja valinta sekä normalisointi. Esimerkiksi siivouksessa virheellistä tietoa voidaan korjata ennen tiedon hyödyntämistä. Yhdistämisessä aineistoa yhdistetään eri lähteistä saman muotoiseksi kokonaisuudeksi ja ominaisuuksien valinnassa karsitaan turhat tiedot analyysin kannalta. (Markkula & Syväniemi, 2015, ss. 95–96)

Data Cleasing eli datan puhdistuksella tarkoitetaan datan tarkistamista ja muuttamista laadukkaaseen muotoon. Tämä tarkoittaa epätäydellisen tai duplikaattidatan korjaamista ja poisto sekä datan saamista koneluettavaan muotoon. Suurin osa data-analytiikkaprojektin ajasta kuluu datan puhdistukseen, sillä data tulee erilaisissa muodoissa ja monessa tapauksessa virheellistä. (Hovi, 2018)

3.2.4 Mallinnusvaihe

Datamallinnusvaihe koostuu mallintamistekniikan valitsemisesta, testitapauksien ja mallien rakentamisesta. Kaikkia tiedonlouhintatekniikoita voidaan käyttää mallinnusvaiheessa. Yleensä valinta riippuu liiketoimintaongelmasta ja tiedosta. Tiedonlouhinta on yksi vaihe tietämyksen muodostamiseen. Tiedonlouhinnalla tarkoitetaan erilaisia tilastollisia menetelmiä, joilla pyritään löytämään tarvittava ja oleellinen tieto suurista tietomääristä. Tiedonlouhinnalla tuotettuja tuloksia ja yhteenvetoja sanotaan usein malleiksi tai hahmoiksi. Tavoitteena on, että tietämys datasta kasvaa ja dataa pystytään hyödyntämään erilaisilla tavoilla. Tietämyksen muodostamisessa on monta vaihetta. Se alkaa datasta ja ongelman määrittämisestä päättyen tulosten hyödyntämiseen ja tietämyksen syntymiseen. (Joutsijoki, n.d., Elsevier B.V., 2021)

Tilastollisten ja matemaattisten menetelmien käyttöä hyödynnetään mallinnusvaiheessa liiketoimintaongelman ratkaisemiseksi. Mallinnukset tehdään usein iteratiivisesti. Tyypillisesti ajetaan oletusarvoja käyttäen ja säätäen useita malleja ja jopa palataan datan valmisteluvaiheeseen muokkaamaan data sopivaksi valittuun menetelmään. (Markkula & Syväniemi, 2015, ss. 95–96)

3.2.5 Arviointivaihe

Arviointivaiheessa tulokset tarkistetaan suhteessa määriteltyihin liiketoiminnan tavoitteisiin. Suurin osa projektin työstä on jo takana. Arviointivaiheen tärkein tehtävä on, että tulokset antavat ratkaisun määriteltyyn ongelmaan ja täyttävät työn alussa linjatut tavoitteet ja tavoitellun tarkkuustason. Arviointivaiheessa on hyvä mitata ainakin determinaatiokertoimen luottamusväli ja ennustevirhe. Tuloksia on tutkittava ja jatkotoimet on määriteltävä, että työn tuomat opit voidaan ottaa suoraan käytäntöön. (Markkula & Syväniemi, 2015, ss. 96–97, Elsevier B.V., 2021)

3.2.6 Käyttöönottovaihe

Käyttöönottovaiheessa tehdyn prosessin toimivaksi todetut tulokset otetaan käyttöön. Käyttöönottovaihe kuvataan usein käyttöoppaaksi, ja se voi olla esim. loppuraportti tai ohjelmistokomponentti. Käyttöoppaassa kuvataan, että käyttöönottovaihe koostuu käyttöönoton suunnittelusta, seurannasta ja ylläpidosta. Tuloksia voivat olla ostotodennäköisyysennusteet, markkinoinnin kohdennuksen pohjana toimiva asiakassegmentointi tai vaikkapa huoltoprosessin optimoinnissa käytettävä prosessiin kuuluvan osan elinkaariennuste. (Elsevier B.V., 2021, Markkula & Syväniemi, 2015, s.96)

Analytiikkaa voidaan pitää automatisoinnin onnistumisen jakoavaimena. Analytiikan tekeminen on arvotonta siihen asti, kunnes sen tulosten avulla pystytään saamaan parempien päätösten kautta lisätuottoja tai kustannusvähennyksiä. Analytiikan tulosten hyödyntämisessä on usein lukuisia haasteita. Onnistuessaan analytiikka on hyödyntäjälle positiivinen tukipilari, joka havainnollistaa tekemistä ja auttaa työskentelemään tehokkaammin ja tuloksekkaammin. Tulosten hyödyntäminen saattaa olla vaikeaa, jos analyysin tuotokset ovat epämääräisiä tai liian vaikeasti ymmärrettävissä. Pahimmillaan analytiikka taas on epäluottamusta herättävää, josta palataan nopeasti tuttujen Exceleiden pariin. (Varila, 2019)

Analytiikan tulokset tulee jalkauttaa hyödyntäjille ymmärrettävästi ja varmistua siitä, että niitä todella käytetään. Analytiikka on jatkuvaa tekemistä laajalla rintamalla, ei yksittäinen taikatemppu. Paras lopputulos saadaan aikaiseksi yleensä systemaattisella analytiikan hyödyntämisellä pitkällä aikavälillä ja laajalla rintamalla. (Varila, 2019)

3.3 Datan laatu ja oikeellisuus

Datan oikeellisuus tulee varmistaa ennen se käyttöä. Datan määrä vaikuttaa osaltaan työn määrään. Suurista data määristä virheiden havaitseminen on vaikeampaa kuin pienistä data määristä. Automatisoitu tarkistusprosessi voi parantaa virheiden havaitsemisessa sekä niiden vähentämisessä. Perinteisesti yrityksissä on voitu tarkistaa listoja manuaalisesti

tietojen oikeellisuuden varmistamiseksi. Näin tehdyt tarkistukset ovat alttiita virheille ja vie paljon aikaa ja on erittäin kallista. (Markkula & Syväniemi 2015, ss. 56–57)

Datan taltioinnin, käsittelyn ja tietosisällön osaamista tarvitaan, että voidaan varmistua datan laadusta. Organisaatiossa tulee olla yhtenäinen näkemys osaamisen tärkeydestä, normaali-ilmiöiden sekä karkeiden virherajojen ymmärtämisestä. Kun voidaan todeta, että osaaminen ja ymmärrys ovat riittävällä tasolla, voidaan päättää oikeanlaiset ja parhaat välineet tarkastusprosessiin. Mitä tarkemmalla tasolla dataa tallennetaan, sen tarkemmin on mahdollista seurata yksittäisten tapahtumien tasoa sekä seurata historiaa ja tunnistaa virhetilanteita. Algoritmeja hyödyntämällä on mahdollista ihmistä paremmin huomata asioita. Organisaatio hyötyy automatiikasta, sillä vapautuvat resurssit voidaan hyödyntää kehittämällä asioita eteenpäin. (Markkula & Syväniemi 2015, ss. 56–57)

Virheiden korjaus on ehdottoman välttämätöntä, että voidaan varmistua tiedon oikeellisuudesta. Järjestelmästä tai datasta riippuen korjaustoimenpide tulee tehdä oikein, jotta virheen vaikutus prosesseihin, analytiikkaan tai raportointiin olisi huomaamaton. Väärä tieto voi vaikuttaa tunnuslukujen vääristymiseen ja aiheuttaa päätöksentekijöissä epäluottamusta analytiikkaa kohtaan. Analytiikka tehostaa työaikaa, säästää resursseja ja mahdollistaa sen, että hyvästä datasta saadaan tehtyä useammin raportteja. (Markkula & Syväniemi 2015, ss. 59–60)

Tietosuoja on tärkeää datan keräämisessä, käyttämisessä ja tallentamisessa. Lainsäädäntö asettaa vaatimuksia ja EU:n tietosuoja-asetus vaikuttaa erityisesti henkilötietojen käyttöön ja keräämiseen. Henkilötietojen käsittely ja kerääminen tulee olla tarkkaan harkittua käyttötarkoitukseen perustuen. Lainsäännön mukaan yrityksellä tulee olla dokumentoitu toimenpiteet tietosuojan ja tietoturvan varmistamiseksi. Tietojen käyttöön liittyvät toimenpiteet, prosessit sekä tietomurtoihin varautuminen tulee dokumentoida auditointia varten. Tietosuoja tulee raportoida yrityksen johdolle säännöllisesti. Laiminlyönneistä seuraa tuntuvia sakkoja. (Markkula & Syväniemi 2015, ss. 63–64)

3.4 Datan säilytys ja turvaaminen

Datan keräämiseen, käyttämiseen, tallentamiseen, siirtämiseen ja hallintaa on yrityksillä ja eri organisaatiolla oikeus. Yritykset ja eri organisaatiot voivat hallita kaikkea muuta, mutta eivät henkilötietoja. Heillä on myös vapaus valita pilvipalveluista mistä tahansa EU- maasta sekä käyttää haluamiaan datakeskuksia. Yrityksissä ja organisaatioissa laaditaan usein datan tallennusjärjestelmäarkkitehtuuri, jossa on suunnitelma datan tallennukseen, säilytykseen ja avaamiseen. Datan arkaluonteisuus, jatkuvat datavirrat ja suuret datamäärät tuovat omat erityisvaatimuksensa datanhallintaa. Datan säilytys ja turvaaminen aiheuttaa myös kustannuksia, joten turhaa dataa ei kannata säilyttää kalliilla levypinnalla. Datamassat kasvavat koko ajan, joten yksinkertaistaminen ja kustannustehokkuus ovat tärkeitä. (Puuronen, 2020, Your Europe, 2020)

Suurten tietojen ja analytiikan hallitsemilla markkinoilla Data Martit ovat yksi avain tietojen tehokkaaseen muuttamiseen oivalluksiksi. Tietovarastot (data warehouses) käsittelevät yleensä suuria tietojoukkoja, mutta tietojen analysointi vaatii helposti löydettävää ja helposti saatavilla olevaa tietoa. (talend, n.d.)

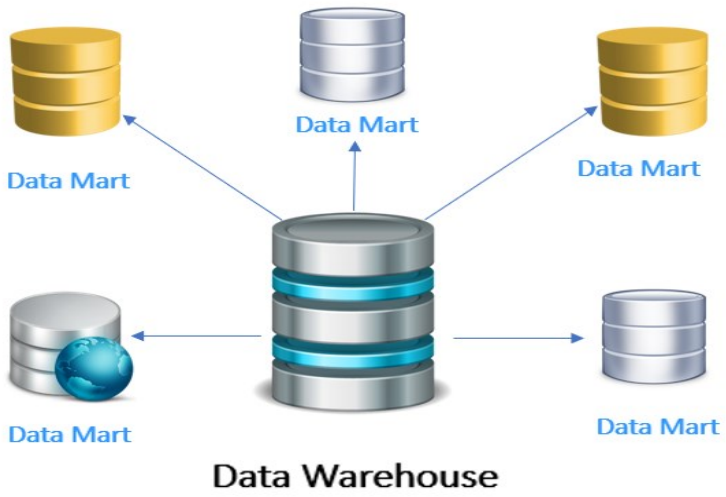
Data Martit ovat aihekeskeisiä tietovarastoja, joka ovat jaettuja segmenttejä yrityksen tietovarastosta. Data Martissa oleva tieto vastaa tyypillisesti tiettyä liiketoimintayksikköä, kuten myyntiä, rahoitusta tai markkinointia. Data Martit nopeuttavat liiketoimintaprosesseja sallimalla pääsyn asiaankuuluviin tietoihin tietovarastossa tai operatiivisessa tietovarastossa muutamassa päivässä, toisin kuin aikaisemmin. Myös käyttöoikeuksia on helpompi hallita Data Martien avulla. Data Mart sisältää vain tietylle liiketoiminta-alueelle sovellettavat tiedot, ja se on kustannustehokas tapa saada nopeasti käyttökelpoista tietoa. (talend, n.d.)

3.4.1 Data Mart vd Data Warehouse

Data Mart ja Data Warehouse ovat molemmat erittäin jäsenneltyjä arkistoja, joissa tietoja säilytetään ja hallitaan, kunnes niitä tarvitaan. Kuva 3 selventää eroavaisuuksia eli ne eroavat kuitenkin tallennettujen tietojen laajuudesta. Data Warehouse on rakennettu toimimaan

koko yrityksen keskusvarastoina, kun taas Data Mart täyttää tietyn divisioonan tai liiketoimintatoiminnon pyynnön. Data Warehouse sisältää koko yrityksen tietoja, joten käyttöoikeuksien hallinta on rajattava erittäin tarkasti. Data Warehouse:ssa tarvittavien tietojen kyseleminen on yleensä uskomattoman vaikea tehtävä yritykselle. Data Martin ensisijainen tarkoitus on eristää tai jakaa osiin pienempi datajoukko kokonaisuudesta, jotta loppukäyttäjät voivat saada helpommin tietoja. (talend, n.d.)

Kuva 3 Data Mart vs Data Warehouse mukaillen. (talend, n.d.)



3.4.2 Data Martin edut

Suurten tietojen hallinta ja arvokkaan liiketoimintatiedon saaminen on kaikkien yritysten kohtaama haaste, joihin pyritään usein löytämään ratkaisu Data Marteista.
Data Mart on aikaa säästävä ratkaisu tietyn tietosarjan käyttämiseen yritystietoja varten. Data Martit voivat olla edullinen vaihtoehto yrityksen tietovaraston kehittämiselle, jossa vaaditut tietojoukot ovat pienempiä. Riippumaton Data Mart voi olla toiminnassa viikon tai alle. Riippuvat ja hybridi Data Martit voivat parantaa Data Warehousin suorituskykyä ottamalla käsittely taakan vastaamaan analyytikon tarpeisiin. Kun riippuvaiset Data Martit sijoitetaan erilliseen käsittely yksikköön vähentää se merkittävästi myös analytiikan käsittelykustannuksia. (talend, n.d.)

Data Martin muita etuja:

- Tietojen ylläpito. Eri osastot voivat omistaa ja hallita tietojaan.
- Yksinkertainen asennus. Yksinkertainen muotoilu vaatii vähemmän teknisiä taitoja.
- Analytiikka. Keskeisiä kustannusmittareita (KPI) voidaan helposti seurata.
- Helppo syöttö. Data Mart:it voivat olla tulevan yrityksen tietovarastohankkeen rakennuspalikoita.

  (talend, n.d.)

3.4.3 Data Martsien tulevaisuus on pilvessä

Vaikka Data Marttien tarjoama joustavuus ja tehokkuus paranevat, big datasta- ja suuryrityksistä on yhä tulossa liian suuria monille paikallisille ratkaisuille. Kun Data Warehousit ja DataLakes siirtyvät pilveen, niin tapahtuu myös Data Marteille. (talend, n.d.)

Jaetun pilvipohjaisen alustan avulla voidaan luoda ja tallettaa dataa, jolloin pääsystä tietoon kiinni sekä analytiikasta tulee paljon tehokkaampaa. Lyhytaikaista analyysiä varten voidaan luoda ohimeneviä tietoklustereita ja pitkäaikaista analyysiä varten pitkäikäisiä klustereita. (talend, n.d.)

Pilvipohjaisten riippuvaisten ja hybridi Data Martsien etuja ovat:

- Joustava arkkitehtuuri pilvipohjaisilla sovelluksilla.
- Yksi säilytyspaikka, joka sisältää kaikki tiedot.
- Resurssit kulutetaan pyynnöstä.
- Välitön reaaliaikainen pääsy tietoihin.
- Lisää tehokkuutta.
- Resurssien yhdistäminen, joka alentaa kustannuksia.
- Reaaliaikainen, interaktiivinen analytiikka.

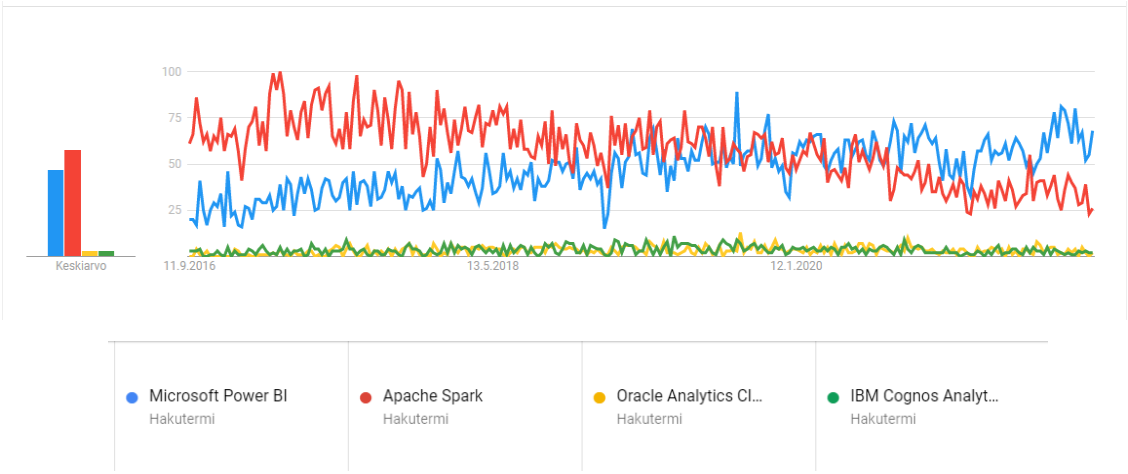
  (talend, n.d.)

3.5 Data-analytiikka alustat

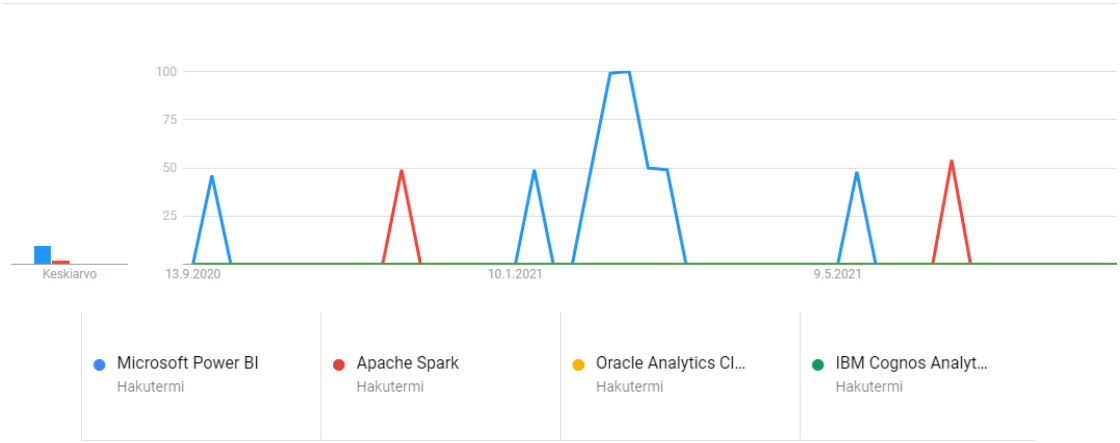
Dataa voidaan lähteä visualisoimaan monella alustalla ja ohjelmistolla. Kuuluisa teknologian alan vaikuttaja ja liiketoiminnan ajatusjohtaja Bernarn Marr, kirjoittaa artikkelissa, The 10 Best Data Analytics And BI Platforms And Tools In 2020, hyvän yleiskatsauksen parhaista ja tällä hetkellä suosituimmista analytiikka- ja liiketoiminta tiedotusaloista listan kymmenen kärjen vuonna 2020. Tähän listalla mahtuu Microsoft Power BI, Oracle Analytics Cloud, IBM Cognos Analytics, ThoughtSpot, Qlik, Apache Spark, Sisense, Talend, Salesforce ja SAS Viya. Alustat vaihtelevat avoimen lähdekoodin ratkaisuista täysin huollettuihin kaupallisiin paketteihin. Useammat ohjelmistoyritykset tarjoavat ilmaisia kokeiluita, joilla voi testata ohjelmiston soveltuvuutta omiin tarkoituksiin. (Bernard, 2020)

Tässä työssä tehdään pienimuotoinen vertailu neljästä alustasta, jotka ovat valittu täysin the 10 best – listalta sattumanvaraisesti. Valituksi ovat
tulleet Microsoft Power Bi, Apache Spark, Oracle Analytics ja IBM Cognos Analytycs. Google Trends on nopea ja helppo työkalu antamaan yleiskuvan tämänhetkisestä googlehakujen määrästä kyseisistä alustoista. Google Trendsin avulla voi vertailla esimerkiksi näitä neljää data-analytiikan ohjelmaa ja katsoa minkälaisessa asetelmassa ne ovat toisiaan vasten. Voi esimerkiksi vertailla, kumpi on suositumpi hakutulosten perusteella, joko maakohtaisesti tai maailmanlaajuisesti. Tarkasteltavan ajan pituutta voidaan myös muuttaa, jos halutaan tarkempaa tietoa lyhyemmältä ajalta. Tätä kappaletta varten tehtiin nopea Google Trends vertailu neljälle data-analytiikka ohjelmiston. Power BI on merkattu tässä vertailussa sinisellä, Apache Spark punaisella, Oracle Analytics Cloud keltaisella ja IBM Cognos Analytics vihreällä värillä. Kuva 4 kuvaa Google hakuja maailmanlaajuisesti viimeisen viiden vuoden aikana ja kuva 5 Suomen googlehakuja viimeisen vuoden aikana. Microsoft Power Bi on sekä maailmanlaajuisesti, että Suomessa selkeästi vahvassa asemassa.

Kuva 4 Vertailtavien ohjelmistojen Googlehaut viimeiset 5 vuotta. Koko maailma.



Kuva 5 Vertailtavien ohjelmistojen Googlehakujen määrä viimeiset 5 vuotta. Suomi.



3.5.1 Microsoft Power BI

Vuoden 2021 Gartner julkaisussa Microsoft Power BI valittiin analytiikka-alustojen johtajan paikalle. Se täyttää nykypäivän työkalun käyttövaatimukset nopeuden, helppouden ja hyödyntämismahdollisuuksien suhteen. Ehdoton etu on siinä, että se on tutun tuntuinen suurelle osalle alan ihmisistä, koska se kuuluu Office 365 tuoteperheeseen. Power BI:n avulla voidaan muuttaa olevassa oleva data helposti luettavampaan muotoon, kuten taulukoiksi ja graafeiksi. Tämä helpottaa ja nopeuttaa informaation lukemista. Power BI ohjelmistoa voidaan käyttää suoraan esimerkiksi Google Analytics- tai esimerkiksi Excel -tiedostojen käsittelemiseen. Power BI toimii myös muiden Office -pakettiin kuuluvien sovellusten kanssa, kuten Teamsin tai SharePointin. Tämän yhteyden avulla dataa voidaan jakaa

organisaation muiden jäsenten kanssa, joka tekee työntekijöiden välisestä kommunikaatiosta ja yhteistyöstä helpompaa. (Microsoft Power Bi. n.d.)

Power BI:tä voidaan käyttää paikan päällä tai pilvessä. Power BI:n avulla käyttäjä voi valita haluamansa datan lähteen, kuten esimerkiksi Excelin tai Azuren. Power BI:n avulla voidaan analysoida hallussa olevaa dataa ja luoda tämän informaation avulla ennustuksia tulevaisuudesta. Tämän avulla voidaan suunnitella paras mahdollinen tapa saavuttaa halutut tavoitteet. (Microsoft Power Bi. n.d.)

Power BI voi hyödyntää myös tekoälyä datan analysoimisessa. Ohjelmalla voidaan analysoida tekoälyn avulla esimerkiksi numeerisen datan lisäksi myös tekstiä ja jopa kuvia. Tämä tekee esimerkiksi asiakaspalautteiden ja sosiaalisen median analysoinnista huomattavasti helpompaa. Power BI:ssä on myös ominaisuus, jossa voidaan esittää kysymyksiä aineistolta ja tämän jälkeen Power BI etsii datasta haettua asiaa ja esittää nämä hakutulokset. (Microsoft Power Bi. n.d.)

3.5.2 Apache Spark

Apache Spark kehitettiin alun perin UC Berkeleyssä vuonna 2009. Apache Spark on avoimen lähdekoodin alusta, jota käytetään datan prosessointiin. Apache Sparkia käytetään pääasiassa erilaisissa data-analyysia vaativissa tehtävissä ja koneoppimisessa. Data voi olla levossa tai reaaliaikaisiin datavirtoihin. Apache Sparkia voidaan käyttää tietyllä koneella paikallisesti tai resursseja voidaan jakaa ryppäässä olevien koneiden kanssa. Apache Spark on kirjoitettu käyttäen Scala-ohjelmointikieltä, mutta sillä on mahdollista lukea sovelluksia, jotka ovat kirjoitettu SQL-, Scala-, Python-, Java- ja R-ohjelmointikielillä. Apache Spark tähtää nopeaan tiedon laskemiseen sekä helppokäyttöisyyteen käyttäen apunaan RDD:tä. RDD on Sparkin muuttumaton hajautettu tietorakenne. (Apache Spark, n.d.)

Apache Sparkin etuina pidetään nopeutta, helppokäyttöisyyttä ja yhtenäistä moottoria. Nopeus perustuu laajamittaisessa tietojenkäsittelyssä mahdollisuuteen hyödyntää muistilaskentaa ja muita optimointeja. Spark on myös nopea, kun tiedot tallennetaan levylle ja sillä on tällä hetkellä maailmanennätys laajamittaisesta levyn lajittelusta. Sparkissa on

helppokäyttöiset sovellusliittymät, joita voidaan käyttää suurissa tietojoukoissa. Spark on varustettu korkeamman tason kirjastoilla, mukaan lukien tuki SQL- kyselyille, suoratoistotiedoille, koneoppimiseen ja kaavioiden käsittelyyn. Nämä vakiokirjastot lisäävät kehittäjien tuottavuutta, ja ne voidaan yhdistää saumattomasti monimutkaisten työnkulkujen luomiseksi. (Apache Spark, n.d.)

Internet -voimayhtiöt, kuten Netflix, Yahoo ja eBay, ovat ottaneet Sparkin käyttöön laajamittaisesti ja käsittelevät yhdessä useita petatavuja dataa yli 8000 solmun klustereista. Sparkista on nopeasti tullut suurin avoimen lähdekoodin yhteisö big datassa, ja sillä on yli 1000 avustajaa yli 250 organisaatiosta. (Apache Spark, n.d.)

3.5.3 Oracle Analytics

Oracle Analyticsiä pidetään tietokantojen osalta kukkulan kuninkaana, sillä Oracle on viime vuosina uudistanut ja käynnistänyt uudelleen tuote - ja palveluvalikoimansa vastaamaan pilvi- ja tekoälyn aikakautta. Sen kielitoiminnot ovat alan kehittyneimpiä, ja ne hyväksyvät kyselyt yli 28 kielellä, eli enemmän kuin mikään muu alusta. Oracle painottaa voimakkaasti itsenäisen tietokannan käsitettä. Tämä tarkoittaa koneoppimisalgoritmien käyttöä monien toimintojen suorittamiseen, jotka olisivat aiemmin vaatineet organisaatioita palkkaamaan kalliin ihmisen tietokannan järjestelmänvalvojaksi. Tämä sisältää tiedonhallinnan, tietoturvapäivitykset ja suorituskyvyn säätämisen. (Bernard, 2020)

3.5.4 IBM Cognos Analytycs

IBM Cognos Analytics on web-pohjainen liiketoimintatiedon hallinta- ohjelmisto. Cognos Analytics ohjelmistoon on sisällytetty raportointi, analysointi, tuloskortti sekä tapahtumien hallintaominaisuudet. Cognos Analytics sisältää yhtenäisen työalueen yrityksen liiketoimintatiedon hallinnoimiseen ja analysoimiseen tavoitteena vastata liiketoiminnan tärkeimpiin ydinkysymyksiin ja olla parempi kilpailijoihin nähden. (IBM, n.d.)

Cognos Analytics avulla voidaan kerätä, liittää yhteen ja jakaa dataa tehokkaasti yrityksen liiketoiminnan menestyksen tueksi. Cognos Analytics sisältää paljon valmiita reaaliaikaisia

raportteja, joita voi hyödyntää ja muokata tietyn tuotteen, ajanjakson ja liiketoimintayksikön suhteen. Cognos Analytics on huomioinut tuotteissaan erikokoiset organisaatiot. Cognos Analytics-raportti sisältää viimeisimmät tiedot operatiivisten järjestelmien tietokannasta. Raportin muodostuminen voidaan ajastaa ja raportti on mahdollista tallentaa, lähettää sähköpostilla tai tulostaa. Raportin voi tuottaa monessa muodossa: HTML, PDF, Microsoft Excel, CSV, XML. (IBM, n.d.)

4 Datan visualisointi

Datan visualisoinnin ensisijainen tarkoitus on esittää luvut ymmärrettävässä muodossa. Hyvä visualisointi ei jätä tulkinnan varaa, vaan esittää datan pääkohdat yksinkertaisesti ja selkeästi. Visualisointi on samaan aikaan datan jäsentämistä ja analysointia, sekä viestintää ja esittämistä. Kun tieto on visualisoitu järkevästi, on kuvaa helppo ymmärtää. (Pengon, 2020)

Visualisointitutkija Robert Kosaran esittää visualisoinnin määritelmän seuraavasti:

- Visualisointi perustuu (ei-visuaalisiin) dataan ja sen tarkoituksena on tiedon välittäminen.
- Visualisointi muuttuu näkymättömästä näkyväksi ja tuottaa kuvan.
- Lopputuloksen on oltava tulkittavissa ja tunnistettavissa.
- Visualisoinnin on tarjottava tapa oppia jotain datasta.
  (Kosara Rober, 2008)

Tiedon omaksuminen ja numeerisen datan ymmärtäminen on helpompaa, kun tieto on esitetty visuaalisessa muodossa. Visuaalinen esitystapa tarjoaa laaja-alaiset tiedot ymmärrettäväsi, yksityiskohtaisesti ja tehokkaasti sekä auttaa hahmottamaan poikkeamia. Kuvan avulla voi nähdä tietoa nopeasti ja pienessä koossa. (Pengon, 2020)

4.1 Data kuviksi

Datan visualialisointi aloitetaan määrittelemällä tarinan esitettävän tiedon asiayhteys. Esimerkiksi ketkä ovat yleisönä, mitä yleisön halutaan oppivan, mikä on yleisön etukäteisymmärrys esitettävästä aiheesta ja mikä on esittäjän suhde yleisöön. Tärkeätä on ennalta tuntea hyvin yleisö ja viestintäympäristöön. Tärkeänä on myös varmistua, että ymmärtää analysoitavan datan ja sen mitä sillä halutaan tuoda esille. Tärkeätä on myös miettiä minkälainen visuaalinen esitys juuri tähän esitykseen ja siihen asetettuihin tavoitteeseen tarvitaan. Kokonaisuus tulee pitää koko ajan selkeänä ja viedä yleisön ajatukset tärkeimpiin kohtiin ja kertoa tarinaa datalla. (Nussbaumer Knaflic, 2015, ss. 20–29)

Ennakoivien ominaisuuksien avulla kuten ikonien käytöllä, kirjasin koolla, värien käytöllä ja sivunasettelu voidaan ohjata yleisön huomio siihen kohtaan mihin halutaan. 5 hyvää käyttökohdetta datan visualisoinnille:

1. Poikkeamien korostaminen
2. Trendien kuvaaminen
3. Eri aikajaksojen vertaaminen
4. Suuren tietomäärän kiteyttäminen
5. Tärkeän tiedon alleviivaaminen

(Pengon, 2020)

Ymmärtämällä kuinka yleisö näkee ja prosessoi tietoa voidaan parantaa omaa asemaa tiedon välittäjänä. Ihmisen muisti voidaan jakaa pitkäkestoiseen, lyhytkestoiseen ja ikoniseen muisti alueeseen. Ikonista muistia pidetään nopeana ja lyhytkestoisena. Ikoninen muisti kiinnittää huomion visuaalisiin korostuskohtiin. Viesti siirtyy ikonisesta muistista lyhytkestoiseen muistiin helposti ja pysyvästi, kun tieto on visuaalisesti esitetty. Lyhytkestoiseen muistiin mahtuu rajallisesti tietoa, mutta ryhmittelyllä parannetaan muistamista. Ryhmittelyn logiikkaa on hyvä tapa käyttää hyödyksi visuaalisessa suunnittelussa. Kun asioita toistaa useita kertoja siirtyy tieto lyhytkestoisesta muistista pitkäkestoiseen muistiin. Pitkäkestoinen muisti on hyvin laaja muistivarasto. Pitkäkestoinen muisti on visuaalisen ja verbaalisen muistin yhdistelmä. Pitkäkestoisesta muistista mieleen palauttaminen on yleensä vaikeaa, mutta visuaalisessa muodossa oleva tieto palautuu helpommin. (Nussbaumer Knaflic, 2015, ss. 99–103)

Tekemällä tiedotettavasta asiasta kaikille riittävän yksinkertaista, helpotetaan yleisön muistikuormaa ja uuden asian mieleen painamista. Datan visualisoinnissa ei pidä unohtaa ulkoasun merkitystä viestinnän keinona. Miellyttävä ulkoasu parantaa yleisön kiinnostusta asiaan ja edistää viestinnän tavoitteisiin pääsyssä. (Nussbaumer Knaflic 2015, ss.138–139)

Yleisön on helppo seurata, kun esitetyssä tarinassa on selkeä punainen lanka, joka on koottu tiedon eri osa-alueista loogisiksi ja juonteelliseksi tarinaksi data-analyysin keinoin. Tarinassa tulee olla yksinkertainen alku, josta käy selville lähtökohdat ja asiayhteys. Tarinan ydin on juoni, jonka avulla selviää ne tekijät, jotka ovat vieneet kohti muutosta. Tarinan lopusta

nähdään johtopäätökset ja jatkotoimenpiteet, joihin tulee ryhtyä. (Nussbaumer Knaflic 2015, ss. 167–168)

Laadukkaasti tehty ja innostusta luova tarina saa aikaan positiivisia mielikuvia. Positiiviset mielikuvat jäävät yleisön muistiin usein helpommin. Mitä nopeammin ja helpommin yleisö pystyy esitetyn tiedon omaksumaan, on päätöksien tekeminen sitä nopeampaa ja varmempaa. (Nussbaumer Knaflic 2015, ss. 95–97)

4.2 Visualisoinnin laatukriteeristö

Laadukkaan datan visualisoinnin suunnittelu vaatii ymmärrystä useasta eri aihepiiristä ja niiden keskinäisistä suhteista. Tietoa ja näkemystä tarvitaan hahmottamisesta, käytettävyydestä, saavutettavuudesta ja visualisointitavoista. Tiedolla on monta eri ulottuvuutta ja niiden välinen liikkuminen aiheuttaa haasteita. Mitä monipuolisemmat tekniset mahdollisuudet ovat, sitä laadukkaampia visualisointeja on mahdollista tuottaa. Tiedostamalla miten ihminen havainnoi ja ajattelee ympärillä olevista asioista, voidaan kehittää visualisointeja, jotka hyödyntävät ihmisen havainnoinnin ja merkityksellistämisen vahvuuksia. (Paukkeri. J., 2014)

Taulukko 2 Hyvä visualisointi täyttää laatukriteerit. (Paukkeri, J. 2014)

Osa-alue	Kriteeri (heuristinen sääntö)
Havainnointi	1. Visuaalisen vihjeen tehokkuus. 2. Värien kontrasti suhteessa toiseen. 3. Kognitiivisen ylikuorman huomioiminen.
Ymmärtämisen tukeminen	4. Visualisoinnin ja asian yhteys. 5. Vastaanottajan itseohjautuvuuden kannustaminen. 6. Yhtenäisyys ja johdonmukaisuus niin sisäisesti kuin ulkoisesti.
Analyysiprosessin tukeminen	7. Vuorovaikutus tukemaan analyysin tavoitteita. 8. Todennettavuus. 9. Vuorovaikutuksen tehokkuus. 10. Käyttäjien erilaisuuksien huomioiminen. 11. Datan uudelleen käyttö.
Saavutettavuus	12. Teknologian käyttö tulisi olla saavutettava. 13. Tiedon esittäminen on esteetöntä esim. mobiililaitteet.

5 Tutkimusosio matkadatan analyysi

Datan analysoinnille on aina jokin syy tai tavoite. Matkadata-analyysin tutkimusosiosta löytyvät vastaukset kysymyksiin siitä mitä, miksi ja miten tämä matkadata-analyysi on viety eteenpäin. Tutkimusosiossa matkadatan analysointia on lähdetty viemään eteenpäin CRISP-DM analytiikan etenemismallin mukaan.

Käytettävissä olevana resurssina on tämä opinnäytetyö ja tavoitteena on antaa suunta seuraaville vaiheille. Tekemisen laajuus on suhteutettu vaadittaviin opintopistetavoitteisiin. Tässä työssä on analysointi työvälinenä käytetty Power BI, sillä sen käytön monipuolinen opettelu on opinnäytetyön tekijän yksi henkilökohtainen tavoite. Power BI:n käyttö on myös edullisin vaihtoehto ja helpoiten saatavilla Microsoftin tuoteperheestä.

5.1 Liiketoiminnan ymmärrysvaihe

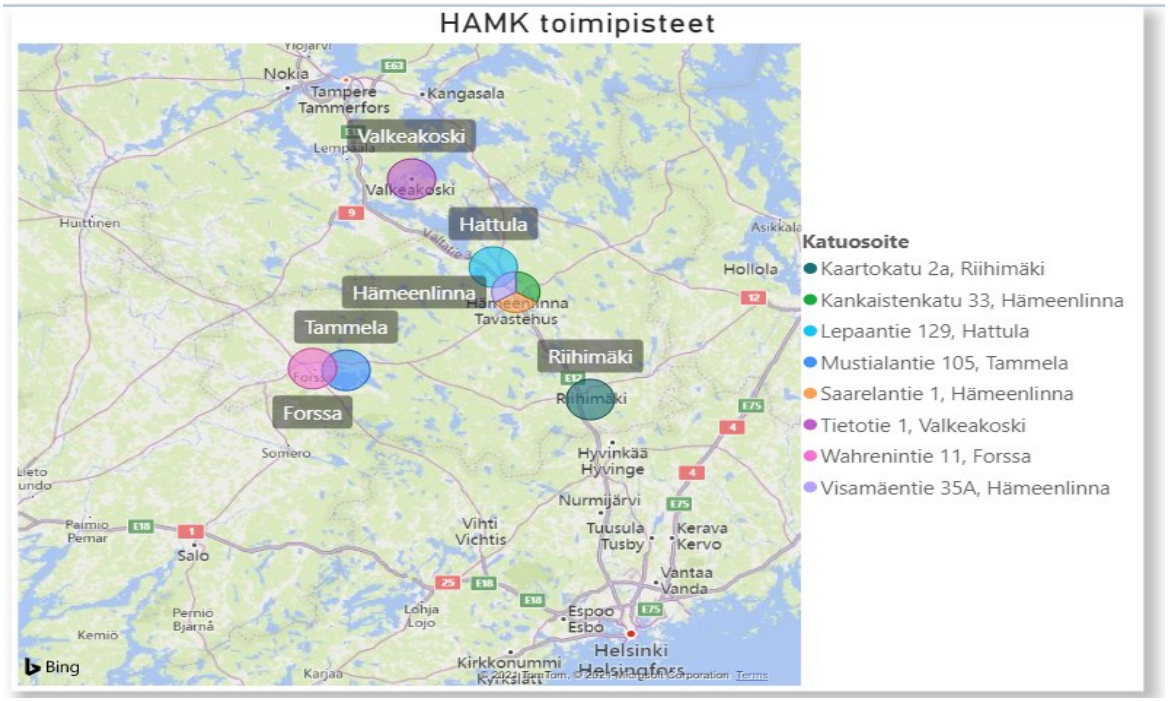
Liiketoiminnan ymmärrysvaiheessa pyritään konkreettisesti selvittämään liiketoiminta ympäristö ja käytettävissä oleva tieto. Määritettyinä käyttötapauksina toimii tämän opinnäytetyön tutkimuskysymykset, joihin tutkimusosion prosessin avulla pyritään löytämään ratkaisu.

Lyhyt katsaus teoriaosuudessa hiilijalanjäljen laskentaperiaatteisiin osoittaa, että hiilijalanjäljen laskenta on hyvin moniulotteinen kokonaisuus, joka vaatii syventävää erikoisosaamista. Hiilijalanjälki osaamista ja laskentaa tarjoaa lukuisat yritykset ja heidän tuotteitansa ostamalla yritykset voivat osoittaa sitoutumisensa päivittäisen toiminnan ympäristövaikutusten vähentämiseen. Tässä opinnäytetyössä on käytetty Hiilifiksu järjestö -laskuria, joka on järjestöille ja yhdistyksille suunnattu hiilijalanjälkilaskuri ja jota myös HAMK käyttää vuosittain ilmoittaessa laskelmat toiminnasta aiheutuvista hiilidioksidipäästöistä opetus- ja kulttuuriministeriöön.

5.1.1 HAMK henkilöstö ja toimipisteet

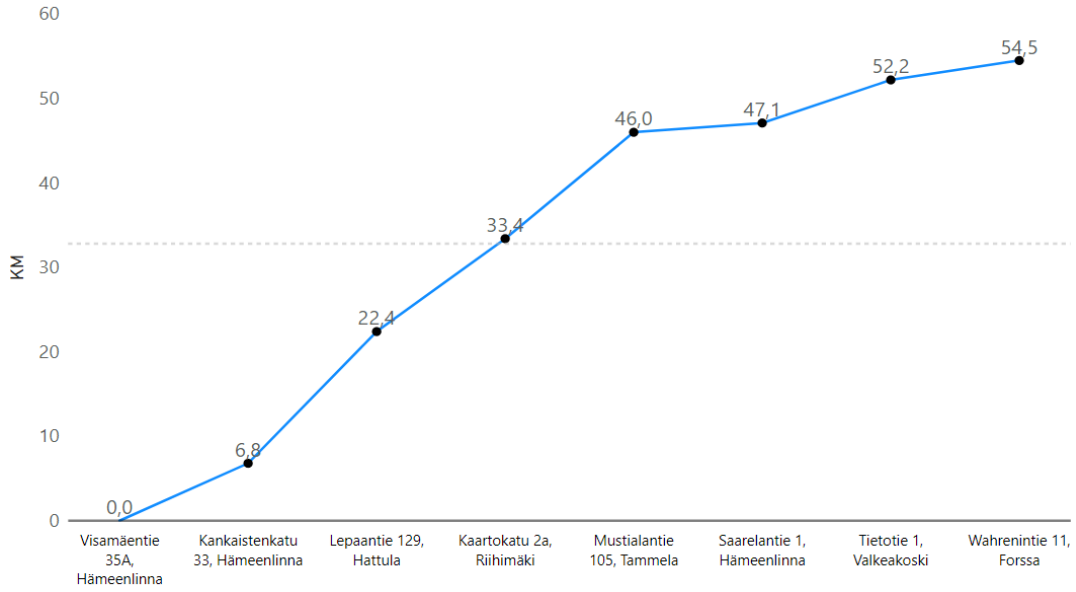
HAMK ammattikorkeakoululla on yhteensä kahdeksan toimipistettä, joista suurin on Korkeakoulukeskus, Visamäki Visamäentie 25A, Hämeenlinnassa. Visamäellä on kahdeksan eri käyntiosoitetta.

Kuva 6 HAMK toimipisteet kartalla.



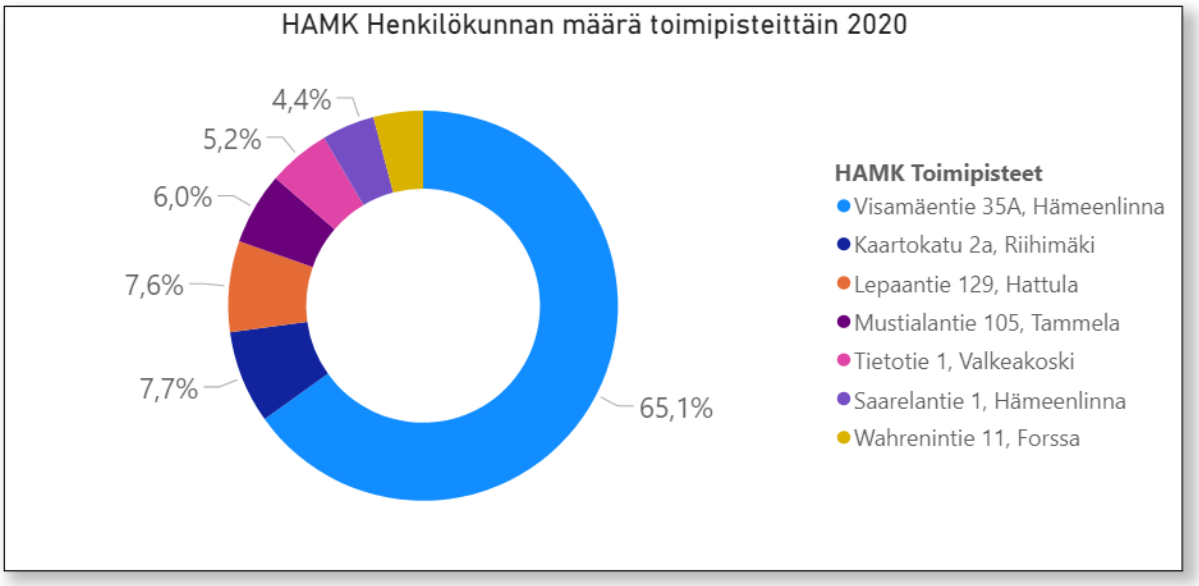
Toimipisteistä kaukaisin on Forssa, jonne on matkaa Hämeenlinnan Visamäentieltä yhteensä 54,5 kilometriä. Edestakainen matka tekee 109 kilometriä. Toimipisteiden keskimääräinen etäisyys on Visamäentieltä 32 kilometriä.

Kuva 7 Toimipisteiden etäisyys Visamäentiestä Hämeenlinnasta.



Hämeen ammattikorkeakoulussa oli vuoden 2020 tiedon mukaan yhteensä 754 työntekijää, joista HAMK 683, Hami 66 ja HAMK akatemia 5. (HAMK, 2020) Hämeenlinnan Visamäentiellä on eniten työntekijöitä 65,1%. Osa henkilökunnasta liikkuu toimipisteiden välillä työmatka tehtävissä ja näin ollen heille kertyy lukuisia kilometrejä työmatkatehtävissä. Osalla henkilökunnasta on myös edustus- ja muuta työmatkustusta niin Suomessa kuin myös maailmanlaajuisesti.

Kuva 8 HAMK Henkilökunnan määrä toimipisteittäin vuonna 2020.



5.1.2 Matkalaskujärjestelmä

Matkalaskujärjestelmiä HAMKissa on kymmenen vuoden aikana ollut kaksi, ja nykyisen matkalaskujärjestelmä kehittäminen on lopetettu. Matka- ja kululaskut saavat alkunsa HAMKissa tehdyistä työmatkoista ja toimipisteiden välisistä matkoista. Matkojen seurauksena työntekijälle korvataan matkakuluja matkustukseen liittyvistä lipuista, kilometrikorvaukset oman auton käytöstä sekä päivärahat voimassa olevien käytäntöjen mukaan. Matkalaskun sisällöstä saadaan selville matkustamistapa, kohde ja aikataulu.

HAMKissa työmatkan tekijä vastaa kulujen tiedottamisesta, ja matkalaskun laadinnassa kaikki kulut ovat jo selvillä. Ohjelma laskee valmiiksi kaikki korvaukset ja vaatii joitakin pakollisia tietoja täytettäväksi. Kilometrikorvausten osalta pakolliset täytettävät kentät ovat päivämäärä, ajettu reitti, ajetut kilometrit sekä onko kyseessä oma tai talonauto. Päivämäärä, ajetut kilometrit sekä auton omistajuus ovat vetovalikon takana. Ajettu reitti on avoin kenttä, johon käyttäjä itse täyttää reitin omalla tyylillä.

Lentojen ja hotellien matkakulut ovat luottokortilla tai suoraan matkatoimistosta ostettuja tapahtumia ja ne liitetään matkalaskulle. Lennettyjen kilometrien määrää kohteeseen ei käyttäjällä ole tiedossa. Lennettyjen kilometrien kokonaismäärä Arenen hiilijalanjälki laskuriin tulee osittain arviona matkatoimistosta.

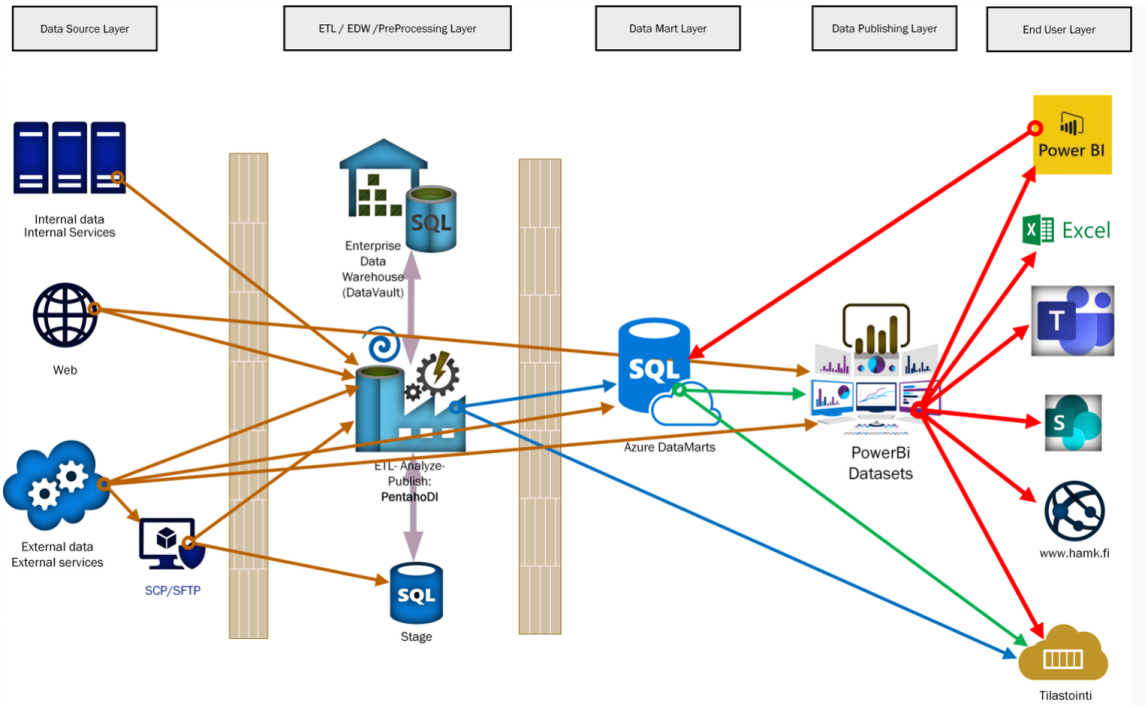
5.1.3 Datan hallintomalli HAMK

Datan hallintaa, strategiaa, arkkitehtuuria ja menetelmiä, joilla hallita, mallintaa, tallentaa, integroida, rikastaa ja valmistella dataa käyttöä varten on HAMKissa piirretty kuvan 9 mukaisen IT - infrastruktuuri malli. Infrastruktuuri mallilla viitataan tietohallinnon johtamaan palvelukokonaisuuteen, johon voidaan katsoa kuuluvaksi IT-infran loppukäyttäjäpalvelut, tietoliikennepalvelut, palvelin- ja kapasiteettipalvelut sekä käyttäjähallinta. Infrastruktuuri mallilla viitataan kaikkiin laitteistoihin, ohjelmistoihin ja verkon resursseihin, joita HAMK hyödyntää IT- ympäristön hallinnassa.

HAMKilla on käytössä Azuren DataMarts, joka on tietovarastoympäristö, jossa tieto on tallennettu tietokartoittain tai toisin sanoen aihekeskitettyihin siivuihin, joka palvelee kapeaa käyttäjäryhmää. Usein tarvitaan vain osajoukko tietojen tietovaraston täydellisistä taulukoista. Tätä opinnäytetyötä varten tallennettiin erilliseen Azuren DataMartsiin tietopaketti, jota lähdettiin työstämään edelleen.

Azuren ympäristö tarjoaa HAMKille rajattomasti mahdollisuuksia olemassa olevan tiedon hyödyntämiseen esim. tiedon ajastamiseen, dataputkien rakentamiseen ja ylipäätä suorittaa tarvittavat ETL/ELT- prosessin vaiheet.

Kuva 9 HAMK IT -infrastruktuuri malli.



5.2 Datan ymmärrysvaihe

Datan ymmärrysvaiheessa selvitetään mitä dataa on käytettävissä ja pohditaan onko käyttötapaukset mahdollista toteuttaa olemassa olevan datan avulla. Ymmmärrysvaihe edellyttää riittäviä datan luku- ja tarkasteluoikeuksia. Vaiheen keskeisiä tavoitteita ovat

datamuotojen ja datan sisällön ymmärtäminen sekä datan laadun selvitys. Myös puuttuvan datan selvitys ja sen luominen tai rakentaminen.

5.2.1 Tutkittava lentodata aineisto puutteineen

Tutkittavaa lentodata aineistoa oli kertynyt vuosilta 2010–2020. Vuodelta 2020 viimeinen merkintä on 11.8.2020. Data noudettiin riippumattomasta Azuren Data Marista, ja se oli Excel muodossa, jossa oli 3 587 riviä. Aineistosta oli tiedossa id, päivämäärä, maakoodi sekä kohdekaupunki taulukon 3 mukaisesti.

Taulukko 3 Lentodata.xlsx aineisto.

	A	B	C	
1	ID	TravelStartDate	CountryID	DestinationDescrition
2	0007cda84fafdcf42f96c4f4adb7f8ce	2019-06-23	SE	Tyresta kansallispuisto, Tyresö
3	001ab2fa029c064a45e41f8b2644a292	2019-04-14	GB	Aberdeen, Skotlanti
4	005532dd7e2a6deb1c2c9aed60434d56	2012-03-10	AT	Muu kohde
5	007065f4d37233c6716157eb32e6f915	2013-09-14	EE	Tallinna
6	00785f258c5fb19d5ef22c0e558806dc	2016-08-02	SE	Göteborg
7	00987d87e4dae5036fa5d8a00dc44e24	2015-10-18	EE	Tallinna
8	00a87ba9ee6920ffb37c266d40937346	2012-10-30	CN	Peking
9	00b97aff19b541dffe9bce7224e3047d	2015-02-15	DK	Muu kohde
10	00dba3250685bdef834e279a6adddf05	2016-05-08	KE	Nairobi

Id oli kryptattu eli käyttäjän henkilötiedot on salattu numerokoodin taakse, että henkilöä ei voida tunnistaa. ID on uniikki eli yksi id vastaa yhtä matkaa. 181 riviä sisälsi kohdekaupunkina Muu kohde, jolloin kohde kaupunki ei ollut tiedossa eli näitä rivejä ei voinut käyttää analysoinnissa. 5 % tutkittavasta aineistosta jää siis analyysin ulkopuolelle. Erilaisia maakoodeja oli yhteensä 83. Maakoodi on matkalaskujärjestelmän vetovalikossa, joka vähentää käyttäjän virheitä.

Kohdekaupunki oli matkalasku järjestelmässä avoin kenttä, jonka henkilön tuli itse kirjoittaa, joka on aiheuttanut sen, että kirjoitustyylejä ja tapoja oli useita. Sama kohde saattoi olla kirjoitettuna useammalla eri tavalla, mutta tarkoitti samaa kohdetta esimerkiksi Aarhus, Aarhus (Kalö) ja Aarhus C. Luotettavimmaksi analysointi kentäksi jää maakoodi, jota tässä tehtävässä on käytetty.

Puuttuvana tietona hiilijalanjäljen laskemiseen ovat lentokilometrit lennettävään kohteeseen, että tarvittava data on kasassa. Puuttuva tieto laskettiin koordinaatiopisteiden avulla erilliseksi maatauluksi. Maataulun esimerkkimalli on liitteessä 2.

Maatauluun kerättiin noin 170 maan koordinaatiopisteet ja niiden avulla laskettiin etäisyys Helsingistä kilometreinä. Apuna käytettiin google map- sovellusta sekä valmiita maaluetteloita leveys- ja pituusasteista. (DistanceLatLong. n.d.)

Ohjelmistokoodi 1 on laskettu esimerkki Helsingin ja Brysselin välin etäisyys kilometreissä, käyttäen Helsingin ja Brysselin koordinaatiopisteitä. Koordinaatiopisteen laskennassa on käytettyä ohjelmakoodi 1 mukaista kaavaa.

Ohjelmakoodi 1 Koordinaattiopisteiden avulla laskettu etäisyys. (richard512, 2021)

```
=ACOS(COS(RADIAANIT(90-B2)) *COS(RADIAANIT(90-B3)) +SIN(RADIAANIT(90-B2))
*SIN(RADIAANIT(90-B3)) *COS(RADIAANIT(C2-C3))) *6371
```

	A	B	C	D	E	F
1	Kaupunki	Lat	Lon			
2	Helsinki	60,1695	24,9354		Etäisyys:	1700,2
3	Brussels	50,4459991	3,939004			

5.2.2 Tutkittava autodata aineisto puutteineen

Tutkittava autodata aineistoa oli kertynyt vuosilta 2010–2020. Viimeisin merkintä on 31.12.2020. Data noudettiin riippumattomasta Azuren Data Martista, ja se oli Excel muodossa, jossa oli 35550 riviä. Aineistosta oli tiedossa id, rivi id, päivämäärä, kuvaus ja kilometrit.

Id-sarake oli kryptattu eli käyttäjän henkilötiedot oli salattu numerokoodin taakse, että käyttäjää ei voida tunnistaa tätä tehtävää varten. Kuvaus kohta oli avoin kenttä matkalaskujärjestelmässä, johon henkilö täyttää matkareitin. Erilaisia reitti kuvauksia oli yhteensä 11 527, jotka olivat eri tavalla kirjoitettu. 750 rivissä kuvaus oli tyhjä eli 2 %

tutkittavasta autodata aineistosta jää osittain tutkimuksen ulkopuolelle. Kuvaus sarake oli matkalaskujärjestelmässä avoin kenttä, jonka matkalaskun täyttäjä itse kirjoittaa.

Puuttuvana tietona oli se, onko ajettu matka ollut toimipistematka vai työmatka. Tätä tietoa ei matkalaskujärjestelmä tuota. Hiilijalanjälki laskentaa varten muita puuttuvia tietoja ei ole. Kuvauksen pilkkominen toimipistematkoihin ja työmatkoihin kuvataan seuraavassa kappaleessa.

Taulukko 4 Autodata.xlsx aineisto.

	A	B	C	D	E
1	ID	Rivi_ID	Paiva	Kuvaus	Km
2	00002b0913ad5462eb608f6af1275131	1	2015-08-17	Valkeakoski-Forssa-Valkeakoski	154
3	0000b707e9f12028f624642af21e5e09	1	2018-06-20	Hml Keskusta	10
4	000387181cc159bf29eed10d12cd3767	1	2010-10-18	Tietotie 1, Vlk - Laajamäentie 1, Hml	54
5	00075d330afde31a529cce55a357561f	1	2017-10-12	Visamäki-Sairio-Visamäki	16
6	00083b0499173bd4ef7aef6b7862418c	1	2013-04-16	Riihimäki-Hämeenlinna-Riihimäki	68
7	000aea08334884bd71b08f94cd7df6bf	1	2010-08-24	Riihimäki - Hämeenlinna - Riihimäki	68
8	000e0b7a77c4960f4a8fa314b1608df5	1	2014-11-24	Forssa-Valkeakoski-Forssa	151
9	000ff22de44f2dd8652930314da249b8	1	2018-09-04	Hml-Rmk-Hml	84
10	0011c5b4883fac3573f17043796f6896	1	2018-10-22	Riihimäki - Forssa - Riihimäki	136
11	00126b47d5502dfb7d01f750ad23d813	0	2019-12-04	Lepaa --- Evo	120
12	001285e5a3cc5e183133a148ffbe1760	2	2011-11-09	Rmk-Kt-Keskus-Kynttilätie Havi- Rmk Kt-Keskus	14
13	0013185c2492be8fc6a0368bb3cf086f	1	2017-01-09	Koti-Lahti-Koti (-20Km)	150

5.2.3 Autodatan kuvaus sarakkeen pilkkominen

Taulukon 4 mallin mukaan kuvaus sarakkeen kaikki solut käytiin läpi ja pyrittiin löytämään nimiä, mitä HAMK toimipisteistä on käytetty. Erilaisia nimiä löytyi noin 170 kpl ja niistä koottiin toimipistetaulukko, josta on esimerkki liitteessä 3. Taulukko 5 on esimerkki, kuinka Forssan toimipisteestä on käytetty 29 erilaista kirjoitustapaa 10 vuoden aikana. Erilaiset kirjoitustavat, joita tulee esimerkiksi lyhenteistä, välilyönneistä, isoista- ja pienistä alkukirjaimista, lempinimistä, pisteitä ja pilkuista aiheuttaa analysoinnin siivousvaiheessa haasteita.

Taulukko 5 Esimerkki Forssa -nimen kirjoitustavoista.

Frs	1
Fssa	1
Forssa	1
Foossa	1
Forrsa	1
Fossa	1
Forrssa	1
Fo	1
Forssa Hamk Wahreninkatu 11	1
Forssa Wahreninkatu	1
Forssa, Wahreninkatu	1
Forssa Hamk	1
HAMK/Forssan campus	1
Frs Hamk	1
Forssa/Hamk	1
Hamk/Forssa	1
HAMK/Forssan campus	1
Visit to Forssa campus	1
Forssa Kampus	1
Forssa, Hamk	1
Hamk, Foprssa	1
Hamk Forssa	1
Hamk, Forssa	1
Hamk.Forssa	1
Hamk, Frs	1
Hamk Frs	1
Hamk;Frs	1
Wahreninkatu 11 (Forssa)	1
Forssa, Jossa Sisäistä Ajoa	1

Kuvaus sarake jaettiin Power Querissä kolmeen sarakkeeseen, että matkakohteet pystyttäisiin tunnistamaan taukukko 6 mukaan. Toimipistetaulukon avulla pystytään tunnistamaan arvolla 1, onko kyseessä HAMK toimipiste. Jos kaikkien kuvaus 1–3 sarakkeiden arvot lasketaan yhteen ja tulos on 3, on tässä oletus, että kyseessä on HAMK toimipistematka. Tehtävässä on käytetty Power Query Merge toimintoa.

Taulukko 6 Kuvaus sarakkeen pilkkominen.

Paiva	Kuvaus.1	Kuvaus.2	Kuvaus.3
2.1.2010	Riihimäki	Hämeenlinna	Riihimäki
2.1.2010	Riihimäki	Hämeenlinna	Riihimäki

5.2.3 Hiilijalanjälki kertoimet

Lentodatan osalta hiilijalanjälki kohteeseen laskettiin taulukko 7. kertoimilla suoraan maataulukkoon, joka on liitteessä 2. Näin jokainen maakoodi sai taulukkoon valmiiksi arvon

hiilijalanjäljestä. Kohteena on käytetty maan pääkaupunkia. Maataulukosta on poistettu sellaiset maat, jotka ovat hyvin epärealistisia matkustuskohteita HAMK henkilöstölle.

Taulukko 7 Päästökertoimet lentodata. (University of Helsinki, 2018)

Lennot	CO_2ekv (g/hkm)
Lyhyet lennot < 463 km	571
Pitkät lennot, Eurooppa >463 km <3700km	328
Kaukolennot >3700km	297

Matkalaskujärjestelmässä ei ole kenttää, johon voisi täyttää minkälaisella autolla henkilö on työmatkansa tehnyt. Tässä tehtävässä on käytetty kertoimena arvoa 184 CO_2ekv (kg) eli kun käytettyä polttoainetta ei ole tiedossa**.** Arvo on keskiarvoa korkeampi. Hiilijalanjälki kerroin lisättiin erilliseksi sarakkeeksi autodata.xlsx, joka kerrottiin ajetuilla kilometreillä. Näin saatiin valmis sarake hiilijalanjäljestä.

Taulukko 8 Päästökertoimet henkilöautoliikenne. (University of Helsinki, 2018)

Henkilöautoliikenne	CO_2ekv (g/km)
Ajetut km:t (diesel)	171
Ajetut km:t (bensiini)	192
Ajetut km:t (kaasu)	85
Ajetut km:t (p.aine ei tiedossa)	184
Taksi	184
Ajetut km:t (sähköauto)	55
Ajetut kilometri hybridi	100

5.3 Mallinnusvaihe

Mallinnusvaiheessa on analysoitu valmista dataa tuloksien saamiseksi. Testattu kaavojen toimivuutta, rakennettu aputaulukoita ja muokattu dataa yhteen sopivaksi toisten taulukoiden kanssa. Turhia rivejä ja sarakkeita on poistettu. Oletusarvoja käyttäen on ajettu erilaisia malleja ja säädetty visuaalista ilmettä valitulle kuulijakunnalle. Mallinnusvaiheesta on myös palattu takaisin datan valmisteluvaiheeseen muokkaamaan dataa sopivaksi valitulle menetelmälle.

Tietämys tutkittavasta datasta on kasvanut ja ymmärrys, miten dataa pystytään hyödyntämään erilaisilla tavoilla. Asetettuja tutkimuskysymyksiä on myös mietitty uudestaan ja arvioitu vastaako tehty projekti tavoiteltuihin kysymyksiin.

5.4 Arviointivaihe

Liitteeseen 4 on kerätty lentodatasta tehtyjä kuva analyysejä. Kuvat ovat tehty tarkoituksella erilaisilla tekniikolla, jotta erilaiset ulkonäölliset seikat nousevat esille. Ensimmäisestä kuvasta, HAMK henkilökunnan lento matkakohteet kartalla vuonna 1.1.2020 – 11.8.2021, on helppo hahmottaa kuinka laaja-alaisesti HAMK henkilökunta on tehnyt työmatkoja viimeisen 10 vuoden aikana. Suurin osa matkoista suuntautuu kuitenkin Euroopan alueelle. HAMK henkilökunnan edestakaisten lentomatkojen kpl määrä vuosina 2010–2020 on helppo havaita, että vuosi 2018 on ollut matkustuksen hurjin vuosi. Korona pysäytti matkustuksen vuonna 2020 ja matkoja oli tuona vuonna vain 88kpl. Yhteen lasketut lentokilometrit noudattavat pääsääntöisesti lentomatkojen kpl määrää, mutta vuonna 2020 tehdyissä matkoissa on paljon kaukomatkoja, sillä lentokilometrien määrä on noussut kovin korkealle.

Keskimääräinen hiilijalanjälki maanosittain on laskettu maataulukon mukaan ja kuvan mukaan lentomatka Amerikkaan aiheuttaa suurimman hiilijalanjäljen. HAMK henkilökunnan lentomatkustuksesta aiheutuneesta hiilijalanjäljestä on grafiikka kuva.

Liitteeseen 5 on kerätty autodatasta tehtyjä kuva-analyysejä. HAMK automatkalaskujen kpl määrä oli hurjin vuonna 2010, jolloin kirjattiin yhteensä 4013kpl matkalaskua. Matkan tarkoitus; toimipistematka ja työmatka on saatu eroteltua seuraavassa kuvassa. Toimipisteden välinen matkustus on vähentynyt puoleen korona vuonna 2020 verraten vuotta 2019. Automatkustamisen yhteenlasketut kilometrit vuodessa vastaavat tehtyjä matkalaskuja eli vuonna 2010 on myös ajettu eniten kilometrejä, 512 819 km. Hiilijalanjälki noudattaa ajettujen kilometrien yhteismäärää. Korona vuosi 2020 romahdutti ajettujen kilometrien määrän, mikä näkyy kuvassa selkeäsi.

5.5 Käyttöönottovaihe

Lopputulos vastaa määriteltyjä tarpeita ja odotuksia. Tulokset ovat vain suuntaa antavia, mielenkiintoa ja kysymyksiä herättäviä seuraaville toimenpiteille hiilijalanjälkilaskennassa.

Tämän tehtävän tuloksena syntyneen maataulun ja toimipistetaulun avulla saadan helposti selville mille tasolle HAMK lento ja automatkustaminen on sijoittunut korona ajan jälkeen vuonna 2021. Koronapandemian takia vuosi 2020 on ollut poikkeuksellinen myös hiilijalanjäljessä matkustamisen jyrjän lasku takia. Ulkomaan matkoista on tehty vain pakolliset matkat, joka näkyy lentomatkojen suurena pudotuksena. Etätöiden lisääntymisen myötä kampuksilla vietetty aika on vähentynyt, joka on vähentänyt selvästi myös toimipisteiden välistä matkustusta. Myös muu työmatkustus oli selvästi vähentynyt vuonna 2020.

Tämän tehtävän avulla hiilijalanjälki saadaan helposti vietyä myös henkilötasolle, niin että jokainen matkalaskun tekijä on selvillä aiheuttamastaan hiilijalanjäljestä. Tavoitteena on herättää keskustelua, voisiko kokouksia enemmän järjestää jatkossa etäkokouksina tai olisiko vaihtoehtoisia tapoja matkustukselle.

6 Johtopäätökset ja pohdinta

Datan manuaalinen työstö osoittautui hitaaksi ja työlääksi. Analysoitava data oli rosoista ja työlästä siivottavaa. Riskejä virheille oli lukuisia, joka lisäsi riskiä lopputuloksen oikeellisuudesta. Esimerkkinä tästä mainittakoon, että alkuperäisen datan ID:t olivat lentodatan osalta 7 kertaisia ja autodatan osalta 8 kertaisia, vaikka ID on uniikki. Tiedon ajaminen tietokannasta ja useammasta eri järjestelmän tietokannasta samaan pakettiin sisältää riskin virheille. Datasta puuttui myös oleellista tietoa, jota matkalaskujärjestelmä ei tuottanut.

Lentodata analyysia varten puuttuva tieto oli kilometrit lennettävään kohteeseen ja välilaskujen määrä. Oletuksena pidettiin, että kaikki lennot ovat suoria lentoja. Pituus- ja leveyspiiriä hyväksi käyttäen tehty laskuri on tehty aina maan pääkaupunkiin. Kohteena ei ole aina ollut kohdemaan pääkaupunki vaan jokin muu kaupunki, joten luku vääristyy tässä kohdassa. Lentodatan analyysi tehtiin maakoodi kautta, joka todellisuudessa osoittautui liian laaja-alaiseksi analyysin perustaksi. Esimerkiksi Venäjän kohdalla saman arvon sai Karjalan alue, Kazan, Moskova ja Pietari. Kaikki nämä kohteet saivat saman arvon, vaikka etäisyys Helsingistä vaihtelee 200 km ja 800 km välillä.

Autodatan analyysissa puuttuvana tietona oli ajettavan auton tyyppi: diesel, bensiini, kaasu, sähkö tai käytetty polttoaine ei tiedossa. Eri ajettavan autotyypin hiilijalanjälki kerron vaihtelee 85–192 CO_2ekv (g/km) välillä, eli kyseessä on merkittävä ero oikeaan hiilijalanjälki kertoimessa. Tätä tietoa 10 vuoden takaa ei edes uskalla lähteä arvaamaan tai laskemaan millään kaavalla. Datasta puuttuvana tietona oli myös jaottelu toimipistematkoihin ja työmatkoihin, jota matkalaskujärjestelmässä ei kysytä. Jaottelun tekeminen avoimet kentästä osoittautui vain jonkinlaiseksi hyväksi arvaukseksi, että jonkinlainen jako matkojen kohteista saataisiin. Avoimen kentän pilkkominen kolmeksi kentäksi ja vertaamalla niitä tehtyyn toimipiste väline matka – taulukkoon osoittautui parhaaksi tavaksi, joskaan ei ihan luotettavaksi. Jos kenttä oli toimipistetaulukossa, sai se arvon 1, ja jos kaikki kolme kenttää sai arvon 1 oli kyseessä toimipistevälinen matka.
Sekä lentodatassa että autodatassa oli noin 2–5 % täyttämättömiä kenttiä, jolloin nämä rivit piti jättää analyysin ulkopuolelle.

Oma kokemukseni on osoittanut, että järjestelmien käyttötapa vaikuttaa datan käsittelyyn, analysointiin ja hyödyntämiseen. Vaikka ennalta on sovittu, että järjestelmään syötettävä data merkitään sovitulla mallilla, saa merkintä usein käyttäjän näköisen vivahteen. Tietojen syöttämisessä saattaa syntyä vaikkapa näppäily virheitä tai tiedot merkitä hyvin eri tavoin. Kiire on usein syynä poiketa sovitusta toimintatavasta, jolloin tärkeitä tietoja saattaa jäädä merkitsemättä. Kun tietoja lähdetään myöhemmin etsimään järjestelmistä tai järjestelmävaihdos tulee ajankohtaiseksi, eri tavalla kirjatut asiat saattavat jäädä löytymättä, siirtymättä tai esiintyä väärissä paikoissa. Käyttäjien erilaisuus näkyy järjestelmien osissa, joissa käyttäjät voivat syöttää itse tietoa. Tästä syystä käyttäjille tehdyt yhteiset ohjeet olisi hyvä päivittää tietyin väliajoin ja nostaa ne esille tietyin väliajoin.

Vapaat kentät järjestelmissä ovat aina hyvin haastavia analysoinnille, sillä tapoja kirjoittaa on niin monta kuin on käyttäjiä. Tiedon analysoinnissa selviää myös puutteet järjestelmien kentistä ja kysyttävistä tiedoista. Data tulisi olla helposti käytettävissä olemassa muodossa analysoinnille.

Koko tämän tehtävän ajan mielessäni kaiversi kysymys, että miksi matkalaskujärjestelmissä ei jo valmiiksi voida huomioida hiilijalanjälkeä, sillä sen koodaaminen olisi suhteellisen yksinkertaista, kun hiilijalanjälki kerroin on tiedossa. Hiilijalanjälki laskennan ympärillä leijuu vaikea, mystinen ilmapiiri, jota on vaikea saada rikottua. Totuuden mukaisia hiilijalanjälki arvoja oli vaikea löytää ja ne olivat kaupallistettu varsin tehokkaalla tavalla.

Tutkittuani lukuisia matkalaskujärjestelmiä netistä, jouduin toteamaan, että yksikään järjestelmä ei tuottanut suoraa hiilijalanjälki laskentaa. Myös haastattelu erääseen suuren eurooppalaisen matkalaskujärjestelmän markkinointipäällikköön vahvisti epäilykseni oikeaksi, että kysyntää hiilijalanjälki laskennalle ei vielä ole ollut, kuin Valtion yhtiöiden osalta, vaikka EU- direktiivi NFR (Non-financial reporting) velvoittaa yli 500 työntekijän yritykset laatimaan vastuullisuusraportin, eli kuinka yritys huomioi ympäristön, talouden ja sosiaaliset tekijät.

Järjestelmien tulee täyttää yritysten tarve vaadittavista analyyseista ja raporteista. Järjestelmiä tulee kehittää vastaamaan yritysten tarpeita sekä vastaamaan muuttuvia lainsäädäntöjä. Lukuisat yritykset tuskastelevat samojen kysymysten kanssa, eli mistä saadaan oikeanlainen data hiilijalanjälki laskentaan kustannustehokkaasti ja riittävän vaivattomasti.

HAMKissa ollaan harkitsemassa uutta matkalaskujärjestelmää, joten hiilijalanjälki laskenta tulisi vaatia kaupanpäällisenä yhdeksi perusominaisuudeksi matkalaskujärjestelmässä. Matkalaskujärjestelmään syötettävät kentät tulisi tukea paremmin data-analyysin tekoa tulevaisuudessa, että HAMK pystyy paremmin täyttämään heille asetetut vaatimukset. Esimerkkinä tästä matkalaskujärjestelmän tulisi pystyä luokittelemaan käytetty auto malli: diesel, bensiini, kaasu, sähköauto tai hybridi, jolloin hiilijalanjälkilaskelma datasta olisi lähempänä todellista ilmoitettua arvoa. Myös jaottelu toimipistematkoihin ja työmatkoihin olisi uuden järjestelmän vaatimus. Avoimia kenttiä uudessa järjestelmässä tulisi välttää, jopa poistaa kokonaan. Turhaa tietoa, ei tulevaisuudessa kannata kerätä. Uuden matkalaskujärjestelmän tulisi tuke myös henkilötasolle vietyä hiilijalanjälkilaskentaa, sillä HAMK henkilöstä tulee saada tietoisiksi jättämästään hiilijalanjäljestä. Järjestelmien hankinnoista vastaavat henkilöt tulisivat pohjata päätöksensä uudesta mahdollisesti hankittavasti matkalaskujärjestelmästä HAMK kestävän kehityksen arvoihin.

ICT ja kestävän kehityksen arvot pitää pystyä yhdistämään. Jokainen voi tehdä oman pienen osansa kestävässä kehityksessä.

7 Yhteenveto

Opinnäytetyön tavoitteena oli analysoida hiilijalanjäljen muutokset ja kokonaisjäljen kehittyminen Power BI:lla vuosina 2010–2020. Tavoitteena oli myös antaa tukea ja työkaluja jatkossa ilmoittaa matkustuksesta aiheutuneet hiilijalanjäljet visuaalisesti.

Tutkimuskysymyksiin löytyi vastaukset, ja matkadata analyysi osoittautui mahdolliseksi tehdä, mutta virhemarginaalia on vaikea arvioida. Datan pystyi muokkaamaan yhtenäiseksi suuren ja monimutkaisen työn avulla. Visuaalisten kuvien avulla on helppo hahmottaa, miten matkustus on HAMK muuttunut viimeisen 10 vuoden aikana. Toimipiste ja työmatkojen muutokset covid-19 myötä on myös helppo havaita.

Hiilijalanlaskenta on itsessään hyvin monimutkaisen laskentamekanismin takana, eikä ajantasaisia laskentakertoimia ollut saatavilla. Myös käytetty hiilifiksun laskentakaava vaatisi päivitystä. Hiilijalanjälki laskenta on myös kaupallistettu erittäin taidokkaasti, vaikka se olisi helposti otettavissa monen järjestelmän oletus näytöksi.

Kaiken kaikkiaan opinnäytetyö onnistui mielestäni hyvin ja tavoiteltu osaamisen kasvattaminen toteutui. Opinnäytetyön pohjalta minulla on selkeä ymmärrys mitä kannattaa lähteä analysoimaan ja miten työlästä datan saattaminen oikeanlaiseen muotoon on. Datan muokkausvaihe vie kaikista eniten resursseja ja se on tehtävä huolella, sillä muuten lopputuloksen oikeellisuus vaarantuu. Järjestelmien tulee tukea vaadittavia ja tarvittavia raportteja ja analyyseja. Uutta järjestelmää harkittaessa tarvittavat ominaisuudet tulee listata niin, että tarvittavat loppuraportit ja analyysien data tarve täyttyy. Minulla on huomattavasti vahvempi ymmärrys ja osaaminen data-analytiikasta, Power BI, Excelistä ja datan visualisoinnista kuin ennen opinnäytetyön tekemistä. Mielestäni valittu viitekehys tuki hyvin tutkimusosiota. Itse uskon siihen, että tulevaisuudessa hiilijalanjälki laskuri tulee olemaan monessa järjestelmässä oletuksena, kunhan järjestelmän käyttäjät uskaltavat alkaa vaatimaan laskentaa. Moni meistä seuraa jo nyt unen laatua, juostuja kilometrejä niin miksi ei hiilijalanjälkeä?

Lähteet

Apache Spark (n.d.). Apache Spark. Haettu 11.9.2021 osoitteesta https://databricks.com/spark/about

Arene. (2020). Arenen toimintakertomus 2020. Haettu 8.7.2021 osoitteesta https://www.arene.fi/ajankohtaista/arenen-toimintakertomus-2020/

Bernard Marr.2020. The 10 Best Data Analytics And BI Platforms And Tools In 2020. Haettu 7.9.2021 osoitteesta https://bernardmarr.com/the-10-best-data-analytics-and-bi-platforms-and-tools-in-2020/

Elsevier B.V. 2021. ScienceDirect. A Systematic Literature Review on Applying CRISP-DM Process Model. Haettu 8.7.2021 osoitteesta https://www.sciencedirect.com/science/article/pii/S1877050921002416?via%3Dihub

Green Carbon. (n.d.). Mikä ihmeen scope 1, 2, 3. Haettu 8.7.2021 osoitteesta https://greencarbon.fi/mika-ihmeen-scope-1-2-3/

IBM. (n.d.). Cognos Analytics – Features. Haettu 11.9.2021 osoitteesta https://www.ibm.com/products/cognos-analytics/features

DistanceLatLong. (n.d.) Country List with Latitude and Longitude. Haettu 8.7. osoitteesta https://www.distancelatlong.com/all/countries/

HAMK. 2020. HAMKin vuosikertomus 2020. Haettu 5.10.2021 osoitteesta https://www.hamk.fi/tietoa-hamkista/vuosikertomus/

Hovi Johannes. 2018 06.06. Data-alan termien selitykset ja kuvaukset. Haettu 8.7.2021 osoitteesta https://www.arihovi.com/3274-2/

Hämeen ammattikorkeakoulu. (n.d.). Kestävä HAMK. Haettu 8.7.2021 osoitteesta https://www.hamk.fi/tietoa-hamkista/kestava-hamk/

Ilmasto.org. (n.d.). Usein kysytyt kysymykset. Haettu 7.7.2021 osoitteesta http://www.ilmasto.org/ilmastonmuutos/usein-kysytyt-kysymykset.html

Ilmasto-opas.fi. (n.d.). Ilmastonmuutos ilmiönä. Haettu 7.7.2021 osoitteesta https://ilmasto-opas.fi/fi/ilmastonmuutos/ilmio/-/artikkeli/6c5a9908-7033-47a8-9855e745b4fa7604/maapallon-ilmasto-tulevaisuudessa.html

Joutsijoki Henry. (n.d.). Tiedonlouhinta ja sen mahdollisuudet. Haettu 8.7.2021 osoitteesta https://docplayer.fi/108135176-Tiedonlouhinta-ja-sen-mahdollisuudet.html

Kaearney Steve. 2021.15.1. How analytics can improve vaccine distribution and

administration. SAS Voices. Haettu 7.9.2021 osoitteesta https://blogs.sas.com/content/sascom/2021/01/15/3-ways-analytics-can-improve-vaccinedistribution-and-administration/

Kosara Rober. 2008 24.07. What is Visualization? A Definition. Haettu 8.7.2021 osoitteesta https://eagereyes.org/criticism/definition-of-visualization

Markkula Tuulikki & Syväniemi Antti. (2015). Analytiikkamatka - Datasta tietoon ja tiedolla johtamiseen (Suomela Susanna, Ed.; Suomen Liikekirjat). Suomen Liikekirjat.

Microsoft Power BI. (n.d.). Why Power BI - Features & Benefits. Haettu 7.9 osoitteesta https://powerbi.microsoft.com/en-us/why-power-bi/

Nussbaumer Knaflic Cole. 2015. Storytelling with Data - A data visualization guide for business professionals. John Wiley Sons Inc.

OpenCO2.net. (n.d.). Taustaa. Haettu 8.7.2021 osoitteesta https://www.openco2.net/fi/taustaa

Pariisin sopimus, 2 artikla/2016. Haettu 28.9.2021 osoitteesta https://www.finlex.fi/fi/sopimukset/sopsteksti/2016/20160076

Paukkeri, J. (2013). *Informaation visualisoinnin laadukkuuskriteerit.* [Diplomityö, Tampereen teknillinen yliopisto]. https://core.ac.uk/download/pdf/250162429.pdf

Pengon. 2020. Tiedon visualisointi -opas. Haettu 14.9.2021 osoitteesta https://pengon.fi

Puuronen, I. (21.9.2020). Kasvava määrä dataa voi olla automatisoidusti aina oikeassa paikassa- ratkaisulla, joka säästää myös rahaa. Tivi.

Ranganathan, J., Corbier, L., Schmitz, S., Oren, K., Dawson, B., Spannagle, M., Bp, M. M., oileau, P., Canada, E., Frederick, R., Vanderborght, B., Thomson, H. F., Kitamura, K., Woo, C. M., Naseem, &, Kpmg, P., Miner, R., Pricewaterhousecoopers, L. S., Koch, J., … Camobreco, V. (n.d.). The Greenhouse Gas Protocol. Haettu 9.7.2021 osoitteesta https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf

richard512.2021. Github. Haettu 9.7.2021. https://gist.github.com/richard512/dd6cd0c258e463b22ece9331435b4f12

Seppälä, J., Saikku, L., Soimakallio, S., Lounasheimo, J., Regina, K. & Ollikainen, M. (2019) HIILINEUTRAALIUS ILMASTOPOLITIIKASSA-VALTIOT, ALUEET JA KUNNAT. Suomen ilmastopaneeli Raportti 5a/2019. Haette 9.7.2021 osoitteesta

https://www.ilmastopaneeli.fi/wpcontent/uploads/2019/09/Hiilineutraalius_ilmastopaneeli_2019_FINAL.pdf

Shearer Colin. 2020. Journal of Data Warehousing. The CRISP-DM Model: The New Blueprint for Data Mining. Haettu 7.9.2021 osoitteesta https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-newblueprint-for-data-mining-shearer-colin.pdf

Sitra. (n.d.). Keskivertosuomalaisen hiilijalanjälki. Haettu 8.7.2021 osoitteesta https://www.sitra.fi/artikkelit/keskivertosuomalaisen-hiilijalanjalki/

Tiedonlouhinta | Ite wikin digitalisoinnin opas. (n.d.). Haettu 7.9.2021 osoitteesta https://www.itewiki.fi/opas/tiedonlouhinta/

talend. n.d. What is a Data Mart. Haettu 27.9.2021 osoitteesta https://www.talend.com/resources/what-is-data-mart/

University of Helsinki. 2018. Hiilifiksu järjestö – Askelia kohti hiilifiksumpaa järjestösektoria. Haettu 8.7.2021 osoitteesta https://blogs.helsinki.fi/hiilifiksu/

Varila Mikko. 2019.10.01. Oppeja analytiikasta – näin lähdet liikkeelle. Digia. Haettu 7.7.2021 osoitteesta https://blog.digia.com/oppeja-analytiikasta-nain-lahdet-liikkeelle

Weng Jiahao. 2019 24.09. Exploratory Data Analysis: A Practical Guide and Template for Structured Data. Towards Data Science. Haettu 7.9.2021 osoitteesta https://towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide- and-template-for-structured-data-abfbf3ee3bd9

Your Europe, 2020. Datan käyttäminen, tallentaminen ja siirtäminen. Haettu 29.9.2021 osoitteesta https://europa.eu/youreurope/business/running-business/developing-business/using-storing-transferring-data/index_fi.htm

Liite 1: Aineistonhallintasuunnitelma

Opinnäytetyön tutkimusosiota varten datapaketti haettiin kertaluontoisilla tunnuksilla Azuren Data Martista. Tunnukset olivat käytössä vain tietyn ajan. Luovutettu data-aineisto sisälsi matkalaskujärjestelmästä ajettua dataa lento- ja automatkustuksesta vuosilta 2010–2020. Ajot oli tehty kahdesta eri matkalaskujärjestelmästä. Lentodata sisälsi henkilöID, päivämäärän, maatunnuksen sekä kohdekaupungin. Autodata sisälsi henkilöID, päivämäärän, kuvauksen sekä kilometrit. Luovutetussa data-aineistossa henkilötiedot olivat salattuja, niin että henkilöitä ei pystytä yksilöimään. Aineisossa oli henkilön osoitetietoja, mutta niitä ei voida kohdistaa henkilöön.

Tutkimusosion aikana on pidetty myös päiväkirjaa, johon on kerätty tietoja tutkimusosion etenemisestä ja aikataulusta. Kaikkia opinnäytetyön tutkimusosiossa syntyneitä aineistoja ja opinnäytetyön versioita säilytetään tekijän tietokoneen C-asemalla sekä Power BI työstön projektit HAMK palvelimella tekijän oma P-asemalla. Työstön aikana on tehty säännöllisesti varmuuskopioita ulkoiselle kovalevylle. Tutkimusosion materiaaleja säilytetään ulkoisella kovalevyllä vähintään yhden vuoden opinnäytetyön valmistumisesta. Tämän jälkeen aineistot tuhotaan.

Opinnäytetyön tilaaja omistaa opinnäytetyön tutkimusosion aineistot ja tulokset. HAMK voi käyttää kehitysprojektin aineistoja ja tuloksia haluamallaan tavalla jatkossa.

Liite 2 Maataulukko esimerkki

Maataulukossa on koordinaatiopisteiden kautta laskettu kilometrit Helsingistä. Yksi suunta Helsingistä sekä meno-paluu Helsingistä. Yhden suunnan hiilijalanjälki on saatu koordinaatiopisteiden avulla lasketun kilometri * hiilijalanjälki kerroin.

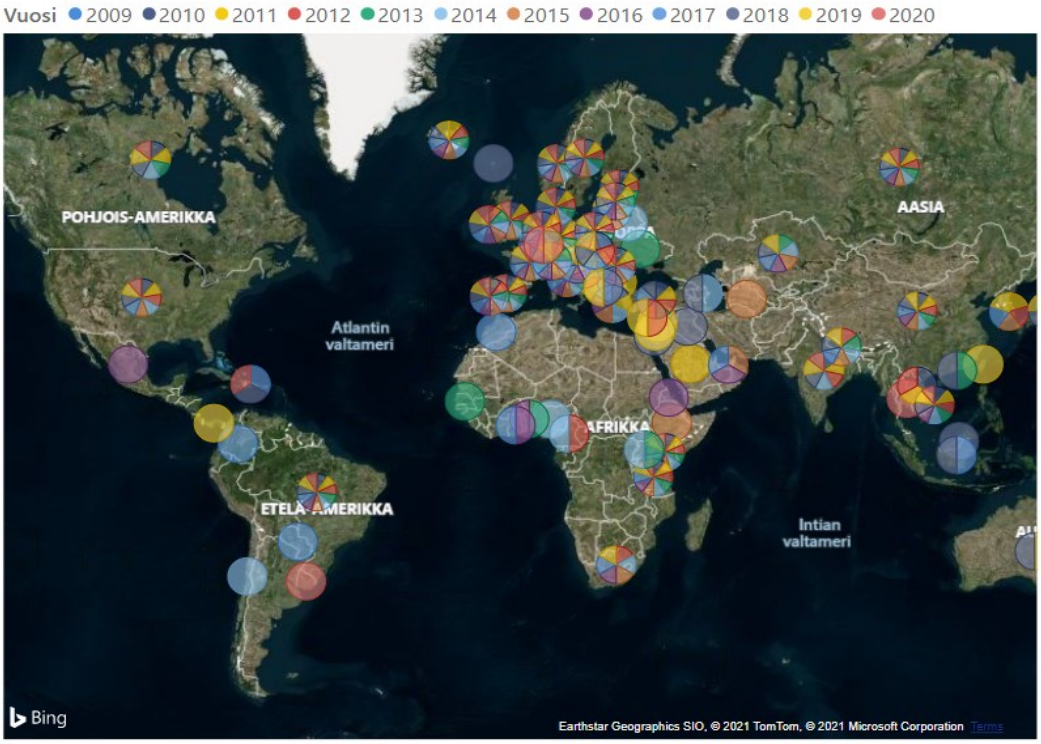
1	Country	Latitude	Longitude	Oneway from HEL	Return from HEL	Oneway Carbon footprint	Return Carbon footprint
3	AL	41,51899817	19,79700359	2103,456222	4206,912444	689,9336409	1379,87
7	AD	42,50000144	1,516485961	2522,908329	5045,816658	827,5139319	1655,03
13	AT	47,51669707	9,766701588	1714,438386	3428,876772	562,3357906	1124,67
19	BE	50,44599911	3,939003561	1700,212736	3400,425473	557,6697775	1115,34
23	BA	44,21997398	17,91998083	1834,370031	3668,740063	601,6733703	1203,35
32	BG	43,13799911	24,71900459	1893,872803	3787,745606	621,1902794	1242,38
35	BY	53,13684572	26,01344031	784,7392405	1569,478481	257,3944709	514,79
56	HR	43,7272222	15,9058333	1926,02997	3852,059939	631,73783	1263,48
58	CY	34,9170031	33,63599757	2877,250232	5754,500463	943,738076	1887,48
59	CZ	50,66299816	14,08100455	1256,148826	2512,297652	412,0168149	824,03
61	DK	55,70900103	9,534996498	1032,038665	2064,07733	338,5086821	677,02
69	EE	58,9430556	23,5413889	157,3661928	314,7323855	89,85609607	179,71
70	FO			7026,169131	14052,33826	2086,772232	4173,54
74	FI	60,99699611	24,47199954	95,42982541	190,8596508	54,49043031	108,98
75	AX			7026,169131	14052,33826	2086,772232	4173,54
76	FR	45,89997479	6,116670287	2010,038312	4020,076624	659,2925664	1318,59
85	DE	49,98247246	8,273219156	1543,252362	3086,504725	506,1867748	1012,37
87	GI	36,3243495	-5,37807483	3413,755993	6827,511985	1119,711966	2239,42
89	GR	38,89899915	22,43400358	2371,641535	4743,283071	777,8984236	1555,80
99	VA	41,90001223	12,44780839	2202,54716	4405,094319	722,4354684	1444,87
102	HU	47,09099714	17,91099957	1524,18271	3048,36542	499,9319289	999,86
103	IS	64,56950277	-21,86232219	2403,548206	4807,096412	788,3638116	1576,73
108	IE	53,6333333	-8,1833333	2111,464572	4222,929145	692,5603797	1385,12
110	IT	40,64200213	15,7989965	2260,86152	4521,723041	741,5625787	1483,13
124	LV	56,50002545	27,3165649	430,9810728	861,9621457	141,3617919	282,72
127	LI	47,13372377	9,516669473	1760,809016	3521,618032	577,5453572	1155,09
128	LT	55,74002016	24,37002641	493,6590343	987,3180687	161,9201633	323,84
129	LU	49,88330105	6,166701555	1644,161857	3288,323714	539,285089	1078,57
136	MT	35,89973248	14,51471065	2799,909538	5599,819077	918,3703286	1836,74
141	MC	43,73964569	7,406913173	2171,396363	4342,792725	712,2180069	1424,44
143	MD	47,2630556	29,1608333	1461,073229	2922,146458	479,2320191	958,46
144	ME	42,46597251	19,26630692	2005,802781	4011,605563	657,9033123	1315,81
152	NL	53,00000109	6,550002585	1372,024873	2744,049746	450,0241584	900,05
165	NO	58,46475606	8,766000553	934,3255081	1868,651016	306,4587666	612,92
178	PL	53,80003522	20,48003129	757,5290158	1515,058032	248,4695172	496,94

Liite 3 Toimipistetaulukko esimerkki

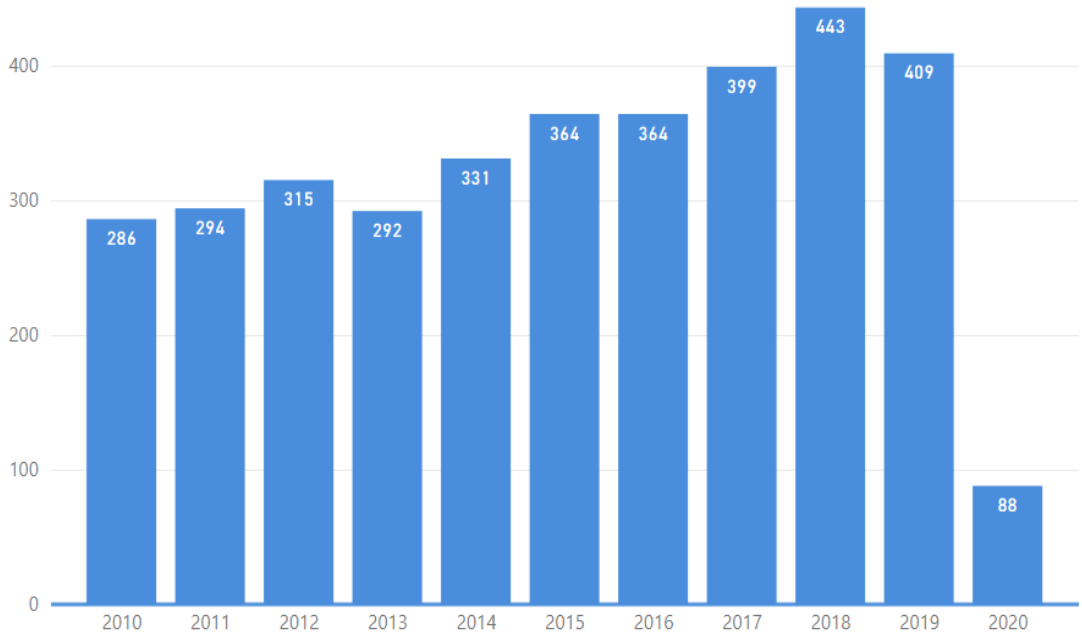
Kohde	ToimipisteM
Hamk Lepaa	1
Lepaa	1
Lepaa, Hattula	1
Lepaantie 129, Lepaa	1
Lepaantie 192, Lepaa	1
HAMK Lepaa	1
Hamk/Lepaa	1
Leppa	1
Hattuala	1
Hatula	1
Hat	1
Hattula Parola	1
Riihimäki	1
Hamk Riihimäki	1
HAMK Riihimäki	1
HAMK Riihimäen kampus	1
Riihimäki Kampus	1
HAMKin Riihimäen toimipiste	1
Hamk Rmk	1
Rmäkihamk	1
Riihimäki/Hamk	1
Riihimäkihamk	1
Rmk	1
Riihiumäki	1
Riihmäki	1
Riksu	1
Rki	1
Rm	1
Rmäki Hamk	1
Rmäki	1
Riihimäki / HAMK	1
Riihimäki HAMK	1
Hamk Valkeakoski	1
Hamk Vlk	1
Valkeakoski	1
Valkeakoski(Tietotie1)	1

Liite 4 Lentodata analyysi kuvina

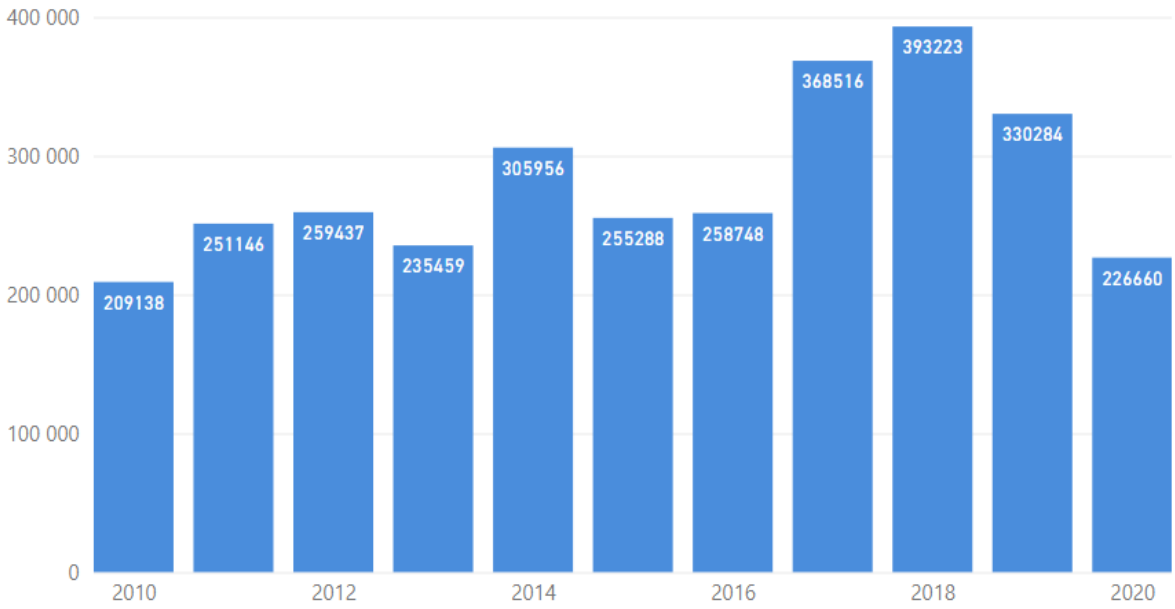
HAMK henkilökunnan lento matkakohteet kartalla vuonna 1.1.2020 – 11.8.2020



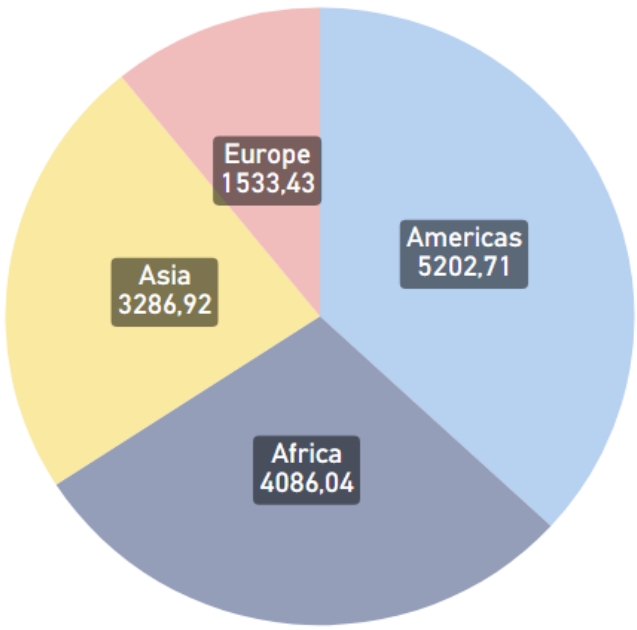
HAMK henkilökunnan edestakaisten lentomatkojen kpl määrä vuosina 1.1.2010-11.8.2020



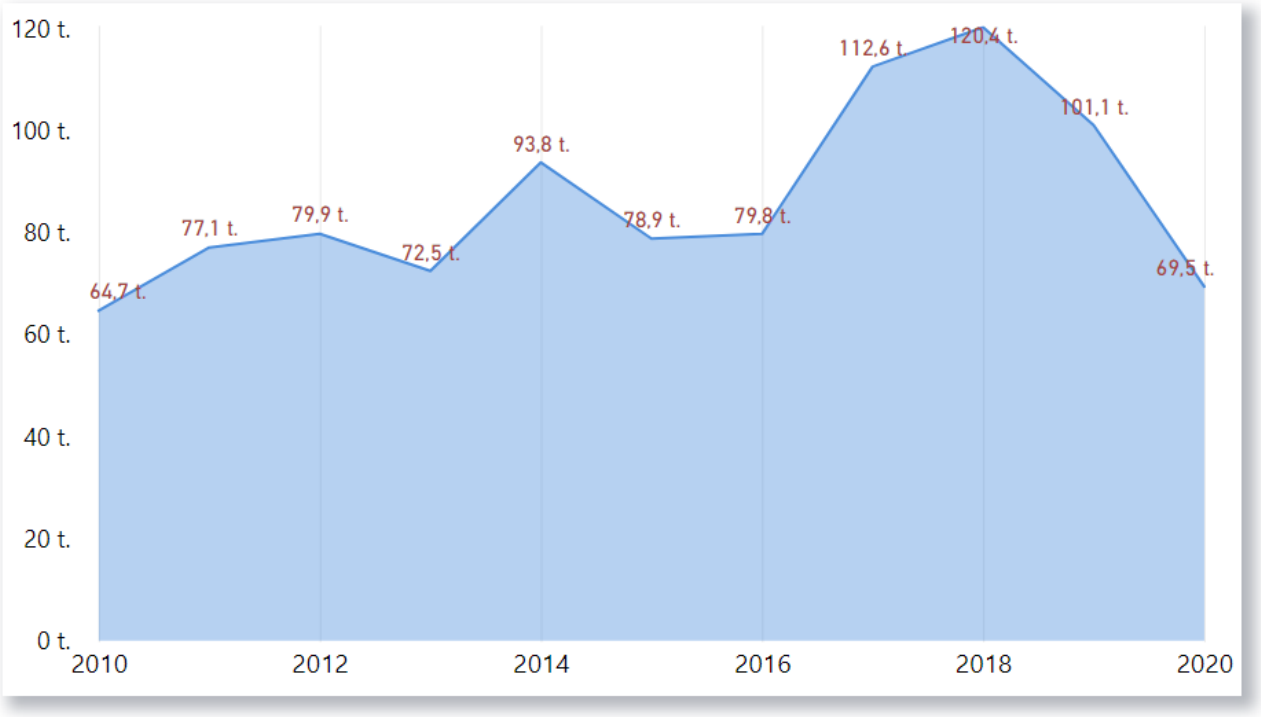
HAMK henkilökunnan lentokilometrit keskimäärin vuosittain 1.1.2010-11.8.2020



Keskimääräinen hiilijalanjälki (kg CO2ekv) per matka maanosittain maataulukon mukaan

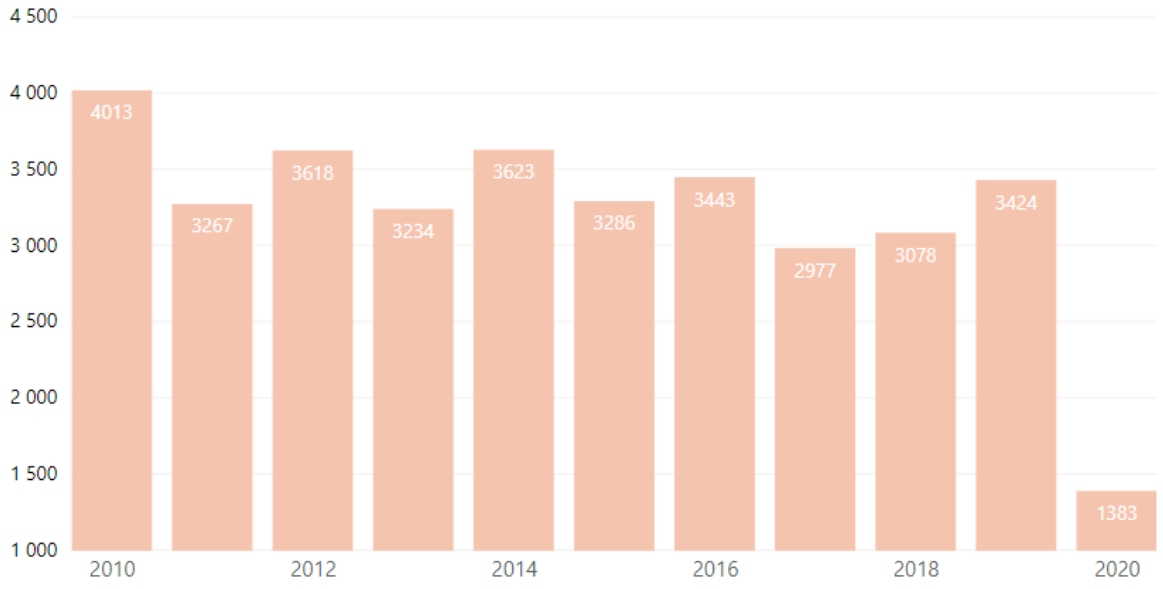


HAMK henkilökunnan lentomatkustuksen hiilijalanjälki (kg CO2ekv) 1.1.2010-11.8.2020

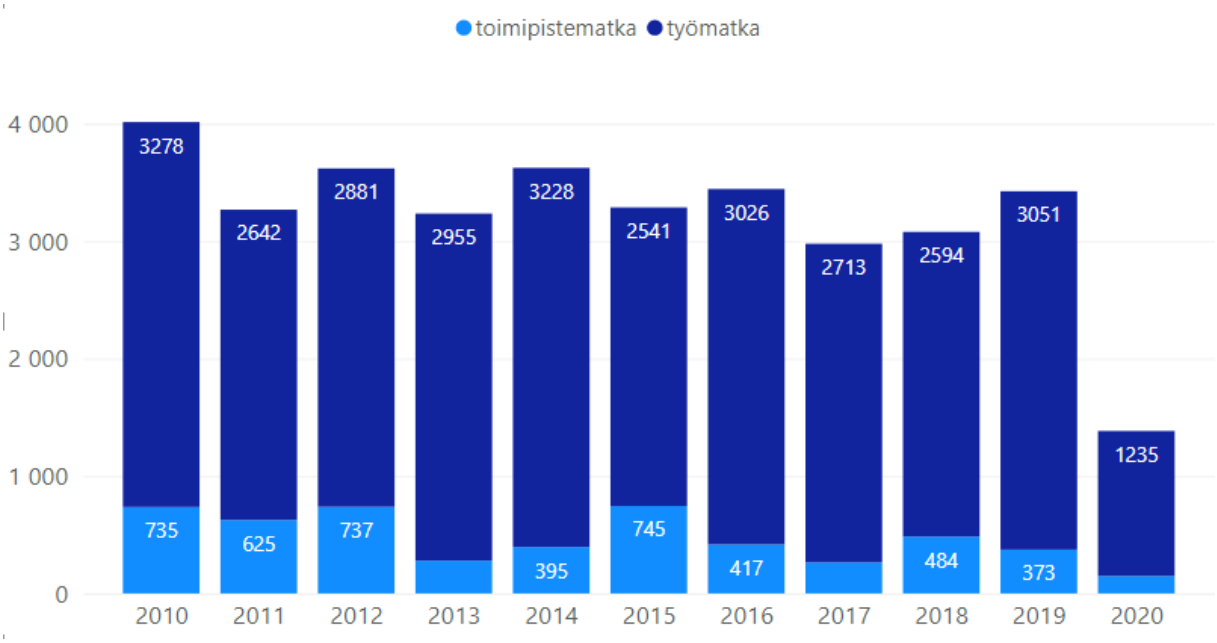


Liite 5 Autodata analyysi kuvina

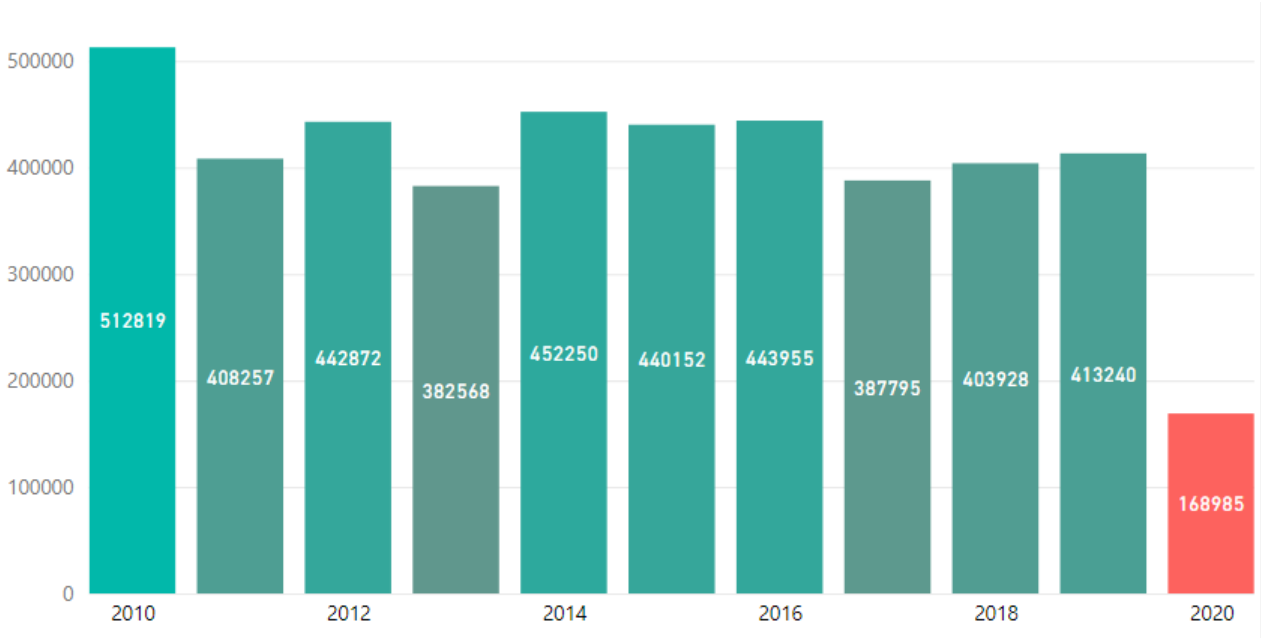
HAMK automatkalaskujen kpl määrä vuosina 2010–2020



HAMK henkilökunnan automatkustuksen ajojen jakauma toimipistematkoihin ja työmatkoihin



HAMK henkilökunnan automatkustaminen kilometreissä 2010–2020



HAMK henkilökunnan automatkustamisen hiilijalanjälki (kg CO2ekv) 2010–2020