Satakunnan ammattikorkeakoulu
Satakunta University of Applied Sciences

VILMA HIETALA

Effective Visualization in EDA: Guidelines for Choosing and Designing Plots

DEGREE PROGRAMME IN DATA ENGINEERING
2025

ABSTRACT

Hietala, Vilma: Effective Visualization In EDA: Guidelines for Choosing and Designing Plots
Bachelor's thesis
Degree programme in Data Engineering
August 2025
Number of pages: 28

The objective of this thesis was to create a practical framework for designing and selecting effective plots that support decision making and data understanding. The target group was individuals without extensive experience in data analytics.

The study was conducted using a two-part approach. First, a literature review was carried out to identify common plot types, their use cases and special characteristics, the common data types, and basic principles of visualization design. In the second part, these findings were tested in a case study using a dataset compiled from the Finnish Public Libraries Statistics website.

It was found that a thorough understanding of both the dataset and the visualization topic is crucial, along with knowledge of the characteristics and use cases of common plot types. This thesis supports these conclusions by compiling key information on data types and common plot types into a single, accessible resource.

Keywords: Exploratory Data Analysis (EDA), data visualization, library statistics, plot selection, graphs, data types.

CONTENTS

1 INTRODUCTION

The term Exploratory Data Analysis (EDA) was introduced by statistician John W. Tukey in his 1977 publication. He described it as a form of detective work, where the analyst must have the appropriate tools to perform graphical, numerical, or counting analysis, but also adopt an investigative mindset to succeed. (Tukey, 1977, p. 1)

Although the core characteristics of the term EDA have remained unchanged since Tukey's publication in 1977, the massive amounts of data generated by modern organizations have fundamentally transformed the ways industries make data-driven decisions (McAfee & Brynjolfsson, 2012).

At the same time, there is increased access to powerful visualization tools, making it possible for almost anyone to create visualizations (Nussbaumer Knaflic, 2015, Introduction). Due to this, it's even more important to understand how to design visualizations effectively and avoid the common pitfalls that can lead to ineffective or misleading visualizations.

1.1 Goals and Approach

This thesis aims to offer a structured framework for selecting effective plots based on the characteristics of the data. It is designed as a practical resource for students and professionals, especially those who are new to data analytics, and need to create effective visualizations in their work.

In this thesis, the term *plot* refers to a graphical representation, which is often also referred to as a graph or a chart in the literature.

This research addresses the following question:

“What are the best practices for selecting and designing effective plots in Exploratory Data Analysis to optimize data understanding and decision-making?”

In order to answer the research question, this thesis follows a two-part approach. First, a research-based literature review, which inspects existing academic and professional sources to identify the importance of data visualization in EDA, common plot types, and their use cases, and the basic principles of designing the visualization. The second part covers a case study, that applies these findings to a real dataset.

This thesis does not cover the technical aspects of creating visualizations, such as preprocessing the data, or using programming languages like Python or R.

1.2 Use of Artificial Intelligence (AI) in This Thesis

All the ideas and thoughts presented without citation are the author's own or represent common knowledge. All the sources have been personally reviewed and follow academic standards.

The following tools have been used to support the writing process:

- ChatGPT: for text formatting and improving academic English, proofreading, planning the structure of this thesis, brainstorming, and assisting with code improvements related to graph creation. ChatGPT has not been used to generate information or to analyze research data.
- Grammarly: to ensure grammatical accuracy.
- DeepL: for translation purposes.

2 LITERATURE REVIEW

Visualizing data through plots is the essence of a successful exploratory data analysis. With plots, a person can explore the data at hand in detail. Anomalies, patterns in the data, and meaningful insights can be spotted from the right type of plot. Depending on the data type, a suitable plot style is selected. (Unwin, 2020, The What and Why of Data Visualization section)

Well-performed EDA plays a critical role in supporting data-driven decision-making. When plots are made taking into account the type of data and they are well designed, they can enhance the ability of decision-makers to interpret complex information and take informed action. (Burnay et al., 2019, pp. 853–855)

This literature review dives into the heart of designing effective plots. It will cover the most common plot types, look into the characteristics of different data types, and explore the principles of designing plots. With the information gained, we can answer the research question introduced in Chapter 1.

2.1 Types of Data

The first step in selecting a suitable plot type is to understand the type of data at hand. As shown in Figure 1, data can be classified into two categories: qualitative (also called categorical) and quantitative (also called numerical). These categories are further broken down into four main scales of measurement. In Figure 1, the four main scales of measurement, from left to right, go from the least detailed measurement to the most detailed level of measurement. (Sahay, 2016, pp. 13-15; Koponen & Hildén, 2019, p. 94)

Nominal data is the most basic level of measurement. It involves classifying observations into categories that do not have a logical order or scale between them. (Sahay, 2016, pp. 13-14; Myatt & Johnson, 2014, Section 2.3) For

example, students can be grouped by the school they attend or by the name of their tutor teacher.

As the name suggests, ordinal data can be ordered in a meaningful way. This could be, for example, the level of satisfaction asked in a survey or the ranking of some sort, such as 1st, 2nd, 3rd, and so on. There is a clear order, but the scale or distance between the classes is not defined or measurable. (Sahay, 2016, p. 14; Koponen & Hildén, 2019, p. 94)

A common feature of nominal and ordinal data is that mathematical calculations cannot be applied to them, except for determining the median in the case of ordinal data. Both represent qualitative scales of measurement. (Koponen & Hildén, 2019, pp. 94-95)

According to Koponen & Hildén (2019, p. 95), quantitative scale is divided into interval and ratio scales of measurement. An interval scale has both meaningful order and distance between values on the scale, such as temperature (Celsius or Fahrenheit) or years. A ratio scale, the highest level of measurement, differs by having a non-arbitrary absolute zero value. Examples include temperature in Kelvin, weight, or unemployment rate.

A discrete variable means that the number of possible values is finite. For example, a person might have 3 children, but not 3.2. Continuous variables, on the other hand, can take any value within a certain range, such as a height of 156.34 centimetres. (Myatt & Johnson, 2014, Section 2.3; Koponen & Hildén, 2019, p. 95)

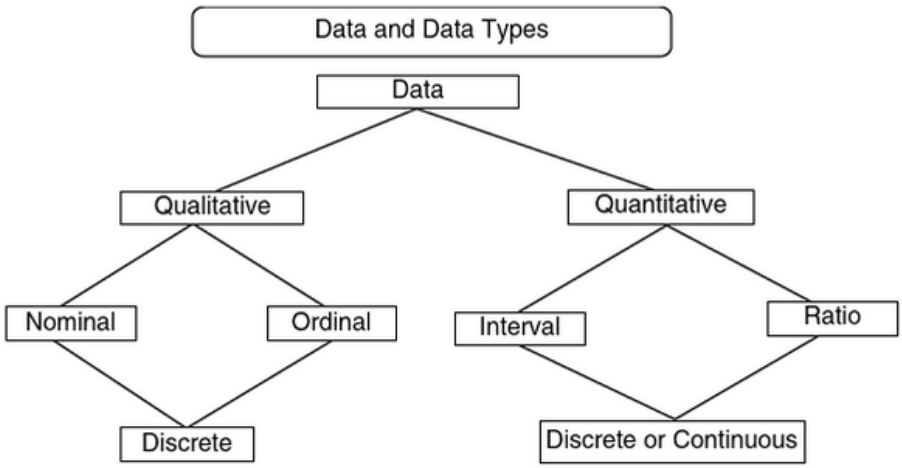

Figure 1 Classification of data (Sahay, 2016, p. 15)

2.2 Common Plot Types and Their Use Cases

2.2.1 Scatter Plot

A scatter plot (also known as XY chart) is used to display the relationship between two quantitative variables. The appropriate type of data for a scatter plot is measured on interval or ratio scales. (Myatt & Johnson, 2014, Section 4.2.1) For example, this could include a person's daily internet usage (in hours), number of years of education, house prices, or salaries.

Typically, the independent variable is plotted on the horizontal x-axis, and the dependent variable is plotted on the vertical y-axis. Each data point represents a single observation of these two variables. There are use cases for scatter plots both in the scientific and business field and according to Koponen & Hildén, it's one of the most commonly used plot types in scientific publications. (Nussbaumer Knaflic, 2015, Chapter 2; Koponen & Hildén, 2019, p. 190)

A relevant aspect to examine from the scatter plot is the strength and direction of the relationship, for example, if they increase or decrease together (Nica, 2013, p. 150). This relationship between variables is called correlation in

statistical terms. It's important to notice that correlation does not necessarily mean causation! The most commonly used statistical method to calculate correlation is the Pearson correlation coefficient. (Yau, 2013, Chapter 4; Koponen & Hildén, 2019, p. 192)

A scatter plot also reveals other important aspects of the data, for example clusters, outliers, and overall patterns. (Nica, 2013, p. 150; Koponen & Hildén, 2019, p. 190).

As seen in Figure 2, when there is a positive relationship between the data points, the cluster will form a shape that goes from left to right in an upward-facing pattern. This tells us that when values in the first variable increase, the values in the second variable increase too. (Myatt & Johnson, 2014, Section 4.2.1)

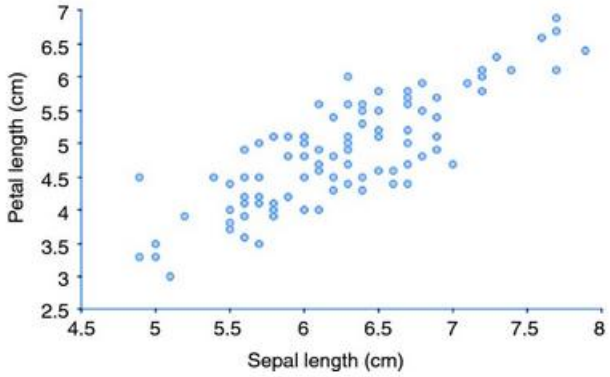

Figure 2 A scatterplot showing positive relationship (Myatt & Johnson, 2014, Section 4.2.1)

2.2.2 Box Plot

A box plot (also known as a box-and-whisker plot or chart) is used to visualize the variation of data. It is an informative visualization that presents key statistical measures such as the median, quartiles, and potential outliers. (Koponen & Hildén, 2019, p. 196; Firdose, 2024, Chapter 10)

As the use of statistical measures implies, box plots are suited for numerical data, where such measures are meaningful.

A demonstration of a box plot is presented in Figure 3. The box in the middle represents the interquartile range (IQR), where 50% of all values are located. The lower quartile is called Q1, and the lower fence is 1.5 × IQR below Q1. All values smaller than this are considered outliers. The upper quartile is called Q3, and the upper fence is 1.5 × IQR above Q3. All values greater than this are also considered outliers. The vertical line inside the box represents Q2, the median. (Firdose, 2024, Chapter 10)

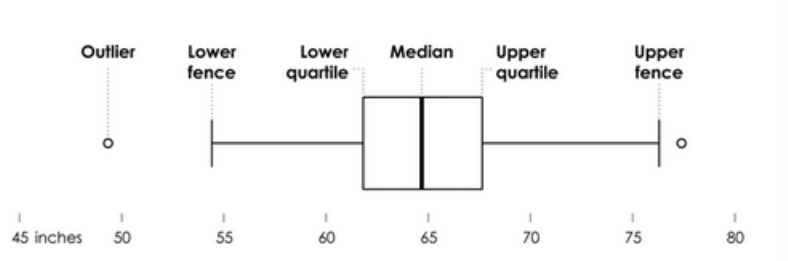

Figure 3 Box plot (Yau, 2013, Chapter 4)

The lines extending from the box are called whiskers, and they provide information about the skewness and variability of the data. If one whisker is considerably longer than the other, it suggests skewness in that direction. Similarly, longer whiskers and a wider box indicate greater variability, while shorter ones suggest the opposite. (Firdose, 2024, Chapter 10)

2.2.3 Line Chart

Line charts (also known as line plots or line graphs) are suited for continuous data, and are typically used for time series (Koponen & Hildén, 2019, p. 184). A line chart is a good option when there is a need to visualize trends or changes in variables over time (Firdose, 2024, Chapter 10). Line charts are typically not used with categorical data. In a line chart, data points are indicated

with markers, and a line is drawn between the data points. If there is a large number of data points, the data points should not be marked in the line chart to avoid clustering the plot. (Nussbaumer Knaflic, 2015, Chapter 2; Koponen & Hildén, 2019, p. 184)

Line charts have some advantages over bar charts, which are discussed in the next section. For example, the vertical axis in a line chart does not need to start at zero, and the chart still works with time series data that is unevenly spaced. (Koponen & Hildén, 2019, p. 184; Magnuson, 2016, Chapter 2)

Figure 4 shows an example of a single series line chart. Similar to a scatter plot, the independent variable in line charts is plotted on the horizontal x-axis, and the dependent variable is plotted on the vertical y-axis. In Figure 4, the independent variable is time series data, Year, and the dependent variable is Temperature (°C). (Firdose, 2024, Chapter 10)

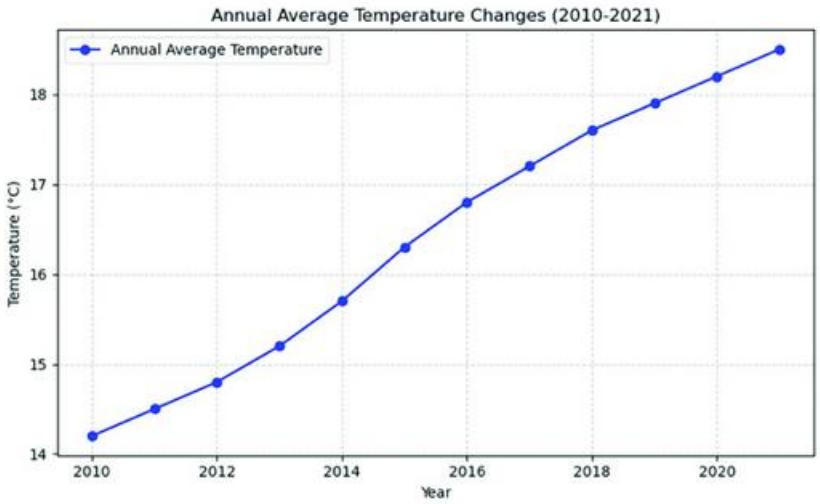

Figure 4 Line chart (Firdose, 2024, Chapter 10)

2.2.4 Bar Chart

A bar chart is used to visualize differences between categories or groups. One axis displays the categories, while the other shows the corresponding

numerical values as the height or length of a bar. It is simple to interpret and commonly undervalued because of its simple nature. It is important to ensure that the scale of the value axis starts at zero, so the chart does not mislead the reader. (Nussbaumer Knaflic, 2015, Chapter 2; Firdose, 2024, Chapter 10)

As seen in Figure 5, it's common to use a vertical bar chart (also called a column chart) for numerical scales and a horizontal bar chart for categorical scales, especially when category labels are long and descriptive. However, if there is a categorical variable with only ≤4 classes, a vertical bar chart may be used. (Koponen & Hildén, 2019, p. 180)

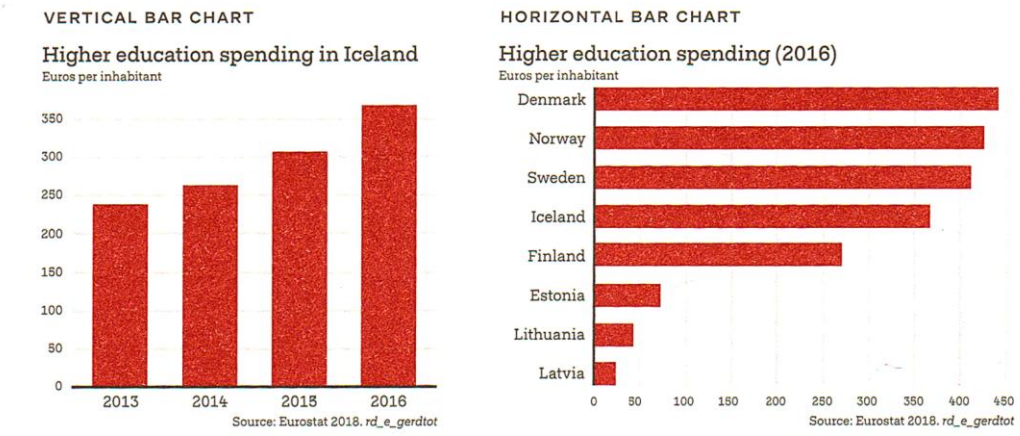

Figure 5 Vertical and horizontal bar chart (Koponen & Hildén, 2019, p. 180)

2.2.5 Histogram

A histogram is a type of bar chart used for visualizing continuous data. It is a powerful tool for illustrating the distribution, shape, and potential skewness of the data. The x-axis represents “bins” (also called “buckets”), which are equal-width intervals into which the data is grouped. If too few bins are used, important features such as the distribution shape or skewness may be obscured. The y-axis illustrates the corresponding number of data points that occur in every bin. (Firdose, 2024, Chapter 10; Koponen & Hildén, 2019, p. 182)

Figure 6 illustrates how the same data is shown with a different number of bins and how too few bins fail to accurately show the details of the distribution of the data. The histogram's peak can provide a rough indication of the data's mean or median, and this can also be seen in Figure 6 (Firdose, 2024, Chapter 10). Notice how the bars in histograms do not have a gap between them, unlike in bar charts.

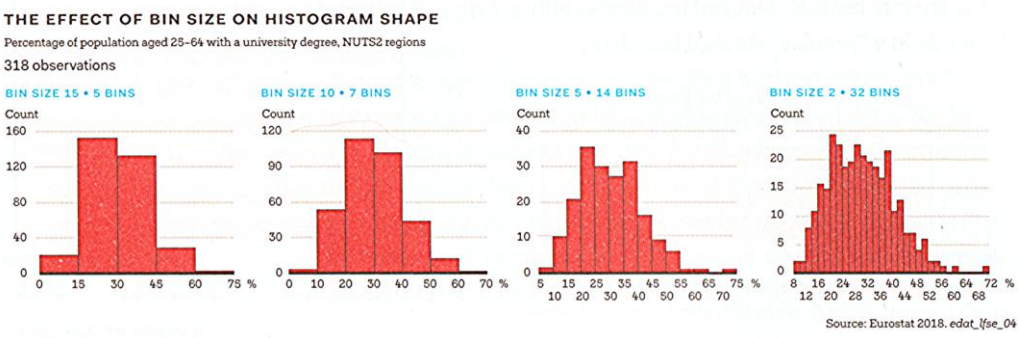

Figure 6 Histogram and the effect of bin sizes (Koponen & Hildén, 2019, p. 182)

2.2.6 Pie Chart

Pie charts are a controversial topic in data visualization. Some authors recommend avoiding them altogether, while others acknowledge that they can be useful when applied appropriately. Nussbaumer Knaflic expresses a particularly strong opinion on pie charts, by titling a section of her book as "Pie charts are evil". (Magnuson, 2016, Chapter 2; Nussbaumer Knaflic, 2015, Chapter 2)

The decision to use a pie chart may be justified when no other plot type serves the purpose more effectively. Pie charts are most suitable for visualizing percentage distributions of categorical variables when an approximate understanding is sufficient. If precise understanding is required, or the percentage differences between categories are small, an alternative plot type should be selected. Even though authors disagree on whether pie charts should be used at all, there is a consensus that if they are used, 3D versions should be strictly

avoided, and a flat 2D should be used instead. (Koponen & Hildén, 2019, pp. 188-189; Nussbaumer Knaflic, 2015, Chapter 2)

Another common mistake when using pie charts is including too many categories. According to Koponen & Hildén (2019, p. 188), a pie chart should not have more than 6-7 categories.

In Figure 7, a well-designed pie chart is presented. The chart includes only three categories, each labelled clearly, which enhances readability. Even though this is a good example of pie chart usage, the usual limitations still apply: it supports an approximate understanding of proportions, but it does not allow the reader to extract precise percentage values.

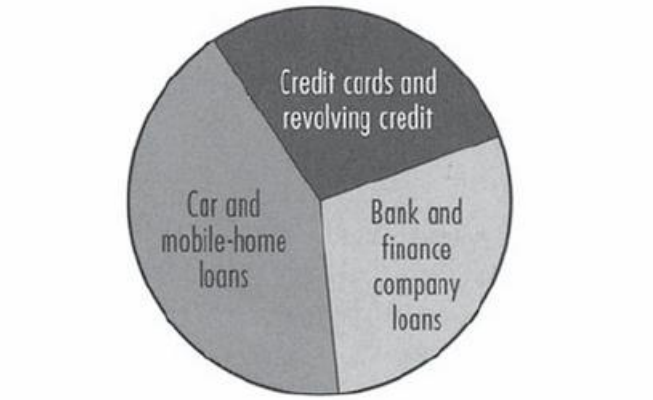

Figure 7 Pie chart (adapted from Kosslyn, 2006, p. 39)

2.2.7 Heatmap

Heatmaps and tables are similar to each other. They both have rows and columns, but the singular values are converted into a colour format. Heatmaps are less common in data visualization, but can be created using either categorical or numerical data. The advantage of heatmaps is the ability to visualize a large amount of data in a single graph. (Magnuson, 2016, Chapter 2) According to Koponen & Hildén (2019, p. 194), heatmaps are also often used with time series data.

According to Magnuson (2016, Chapter 2), one downside of heatmaps is that they are relatively uncommon. This may suggest that they are harder to interpret outside the field of data analytics and, therefore, not recommended for general reports. As illustrated in Figure 8, heatmaps can be generated without displaying the actual cell values. This reduces readability and supports the interpretation that they may not be ideal for general-purpose reporting.

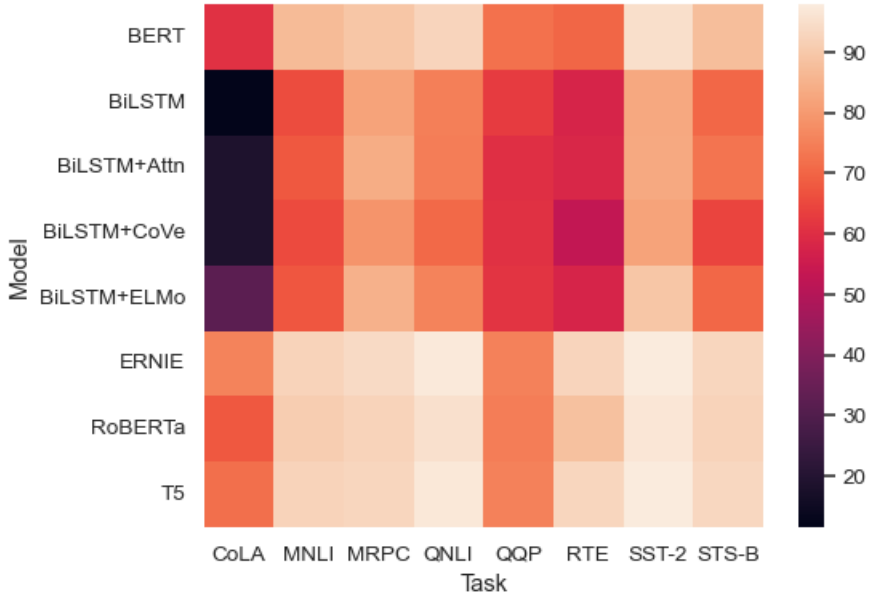

Figure 8 Heatmap without cell values (Seaborn, n.d.)

Figure 9 illustrates how a table can be converted into a heatmap, allowing for quicker visual identification of high and low values. The readability remains high when the cell values are displayed.

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

LOW-HIGH

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Figure 9 A table and a corresponding heatmap (Nussbaumer Knaflic, 2015, Chapter 2)

2.3 Principles of Effective Visualization Design

As discussed in previous sections, the type of data greatly affects the choice and design of a plot. When the goal is to create a clear and easy-to-read visualization, the number of design considerations increases. This section focuses on the readability of plots and how to avoid misleading the reader.

2.3.1 Designing a Colour Scheme

Around 8% of men and 0.5% of women have a colour vision deficiency (Nussbaumer Knaflic, 2015, Chapter 4). It's important to take this into consideration when choosing a colour scheme for a plot.

According to the National Eye Institute (2023), the most problematic colour pair for people with colour vision deficiencies is red and green. This combination should be avoided, and a safer alternative is to use blue and orange (Koponen & Hildén, 2019, p. 66).

Figure 10 demonstrates how red-green colour blindness affects the perception of colours. The top image shows the original colour scale, while the image below simulates how the same scale appears to someone with red-green colour blindness. Red and green become difficult to distinguish, whereas blue and orange remain easily distinguishable.

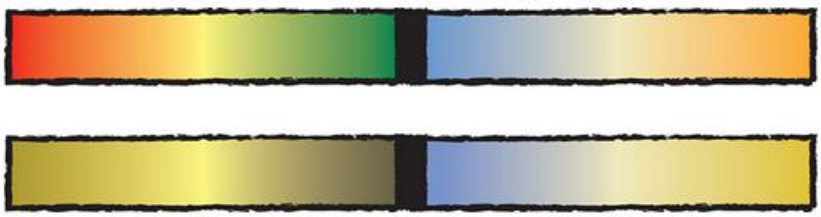

Figure 10 Colour scale appearance with and without red-green colour blindness (Brath & Jonker, 2015, Chapter 5).

Besides, colours are a powerful element when designing plots. It is advisable to avoid using too many colours in a single plot; instead, selecting one primary

colour and using its different shades can help draw attention more effectively. (Nussbaumer Knaflic, 2015, Chapter 4)

Figure 11 illustrates how colour choices influence readability and guide the viewer's attention. It's easier to spot the high and low values when using one primary colour than when using multiple strong colours.

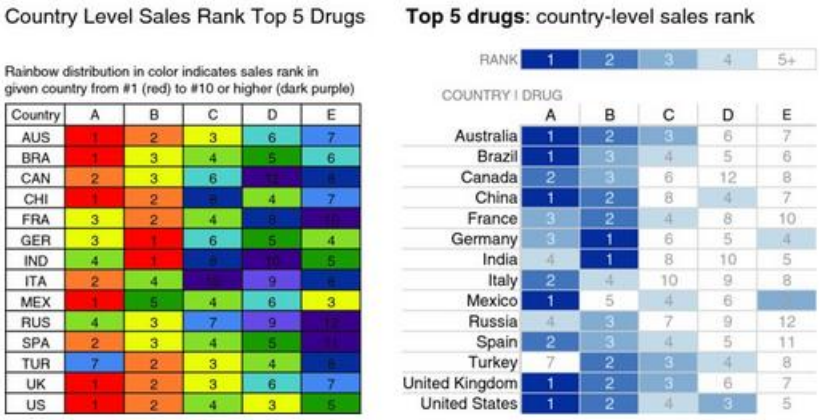

Figure 11 Use of colours (Nussbaumer Knaflic, 2015, Chapter 4)

As a final note on colour schemes: avoid changing the colours of your plots without a clear purpose. Colour changes should serve a function, such as directing the viewer's attention or signalling a topic shift. (Nussbaumer Knaflic, 2015, Chapter 4)

2.3.2 Direction of Text Elements

An effective way to increase the readability of a plot is to consider the direction of text elements. Text such as titles or variable names should be written horizontally, while vertical alignment should be avoided when possible. The effect of text direction can be seen below in Figure 12. According to Koponen & Hildén (2019, p. 257), a 45° angle may be used as a compromise in terms of readability, but horizontal text should be used whenever possible. As seen in some of the previous figures, the title of the vertical axis is still often placed vertically, even though this is not considered optimal in the terms of readability.

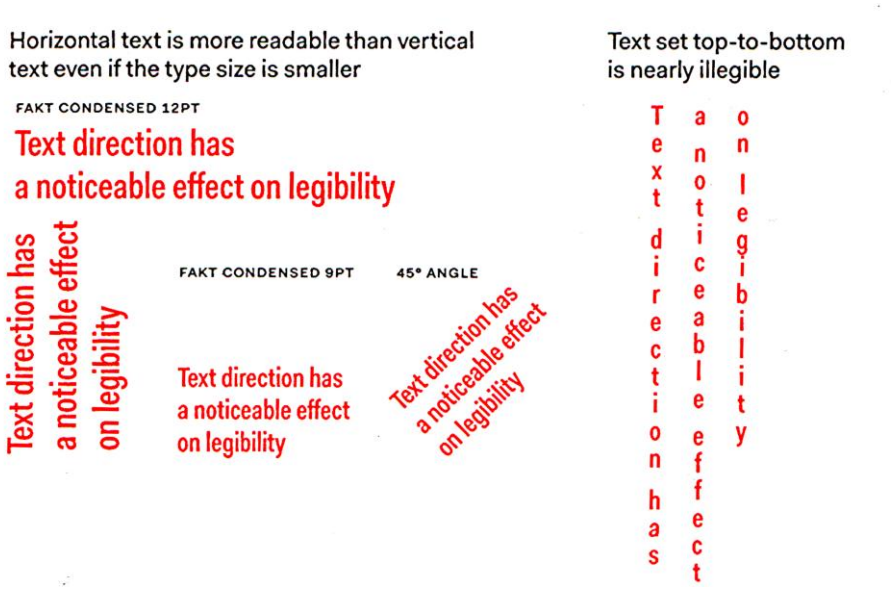

Figure 12 Effect of text direction (Koponen & Hildén, 2019, p. 257)

2.3.3 Misleading and Ineffective Visualizations

There are various ways in which visualizations can mislead the viewer or fail to communicate information effectively. This section presents only a selection of common pitfalls, but it is important to stay aware of the risk of misleading the viewer throughout the entire process of creating visualizations.

Some of the earlier chapters have already addressed factors that may mislead the reader, each within a specific context. For example, as discussed in the chapter on designing a colour scheme, colour changes should only be made with a clear purpose, a principle that applies to all aspects of visual design. In general, design elements should not be altered without a justified reason. As statistician Edward Tufte famously stated: “Show data variation, not design variation” (Tufte, 1983, as cited in Koponen & Hildén, 2019, p. 83).

When making comparisons using the same dataset, it is important to keep axis scales consistent across all plots. As shown in Figure 13, changing the scale can significantly mislead the viewer about the actual differences in the data. It creates the illusion that the two spending levels in the bar charts are similar.

However, when both are shown using the same scale, it becomes clear that government spending is only a fraction of that by the private sector. (Koponen & Hildén, 2019, p. 84)

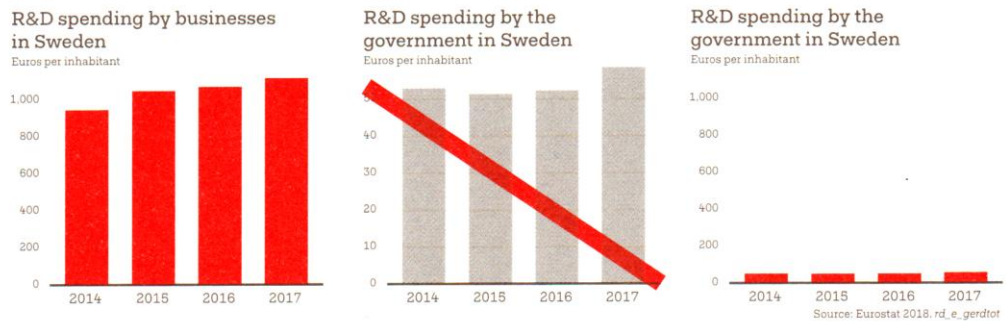

Figure 13 Misleading scales (Koponen & Hildén, 2019, p. 84)

As discussed in the chapter on pie charts, 3D visualizations should generally be avoided. They rarely add value to any type of graph and often make interpretation slower and more difficult. Even in bar charts, where values remain somewhat readable, the added dimension makes it slower to interpret. (Koponen & Hildén, 2019, p. 106)

Finally, it's best to avoid filling your graphs with elements that do not add any information. Some visualization tools add automatically borders, or other visual elements that are not necessary, and make the data stand out less. (Nussbaumer Knaflic, 2015, Chapter 3) Similarly, designers can add visual features that have no informative value, such as decorating a chart with illustrations of the topic (Koponen & Hildén, 2019, p. 104).

3 METHODOLOGY

3.1 Dataset Selection and Description

The dataset used for the case study is compiled from the Finnish Public Libraries Statistics website, which provides yearly reports for Finnish libraries. The

years selected for analysis are 2015 to 2024, and the cities included are Pori, Rauma, Ulvila, Vaasa, Helsinki, and Tampere. It is important to note that in the dataset, the municipality of Nakkila is grouped with Pori and cannot be separated. For clarity, the term *Pori* is used in this case study to refer to the combined data of both Nakkila and Pori. (Finnish Public Libraries Statistics, n.d.)

The choice of Finnish libraries as the topic was guided by the author's personal interest, the reliability of the data source, and the suitable characteristics of the available data.

City selection was also influenced by personal relevance: Pori was chosen as the home city of the author's university, Satakunta University of Applied Sciences. Ulvila and Rauma serve as relevant points of comparison, as they are neighbouring towns. Tampere, Helsinki, and Vaasa provide useful contrasts as larger cities in Finland.

The dataset contains 228 columns covering various topics, such as collections, loans, locations, and personnel costs in each city, making it a versatile source for visualizations.

3.2 Graph Selection

The selection of graphs began with preprocessing and exploration of the data and its characteristics. Based on this examination, relevant topics were identified for visualization, along with suitable graph types for each case.

The following topics were selected for visualization. In Chapter 4, there are visual examples with insights of each graph.

CASE 1: Book loans by language in Helsinki (2024)

The first case is to compare the differences between the book loans by languages in the city of Helsinki in 2024. As noted earlier in Chapter 2.1, the first

step in selecting a suitable graph is to identify the type of data. When examining the variables “Loans: Finnish books”, “Loans: Swedish books”, and “Loans: Books in other languages” and referring to Figure 1, they can be identified as quantitative, discrete variables measured on a ratio scale.

Since the goal is to compare the differences between loans based on language groups, a bar chart is a suitable choice. As discussed in Chapter 2.2.4, bar charts are often undervalued but serve well in simple situations. There is no need for complex graphics when the comparison to be made is simple. Since the labels of the groups are long, a horizontal bar chart was selected.

CASE 2: Trend in the number of events by city (2015–2024)

The second case illustrates the change in the number of events from 2015 to 2024 across the cities of Pori, Rauma, Tampere, and Vaasa. The data type is a discrete variable on a ratio scale, as it consists of countable values (number of events) and includes a non-arbitrary absolute zero value, which would be zero events. Because the goal is to illustrate changes in the number of events over time, a line chart was selected.

CASE 3: Relationship between the expenses and loan counts of e-materials

The third case examines the relationship between the number of loans and the expenses of e-materials in all six previously selected cities. Since both variables, e-material costs and e-material loans, are quantitative, and the goal is to explore their relationship, a scatter plot was chosen to visualize the data. The assumption is that e-material expenses affect the number of e-material loans.

4 CASE STUDY

4.1 Visualizations and Insights

The first case is illustrated in Figure 14, which presents a horizontal bar chart comparing book loans by language in the city of Helsinki in 2024.
Readability has been considered in the choice of colour, text orientation, and by numbering the highest value. As discussed in Chapter 2.3.1, blue is a safe choice in terms of accessibility, as it is generally distinguishable for people with colour vision deficiencies.

Horizontal text alignment improves clarity, and labelling the highest value provides essential context. Consider the graph below: without that number, would a reader be able to grasp the scale of the loans based solely on the bar length and the x-axis scale? The answer is likely no; the reader may not recognize the notation “1e6,” which denotes one million. Using a bold font further guides the viewer’s attention.

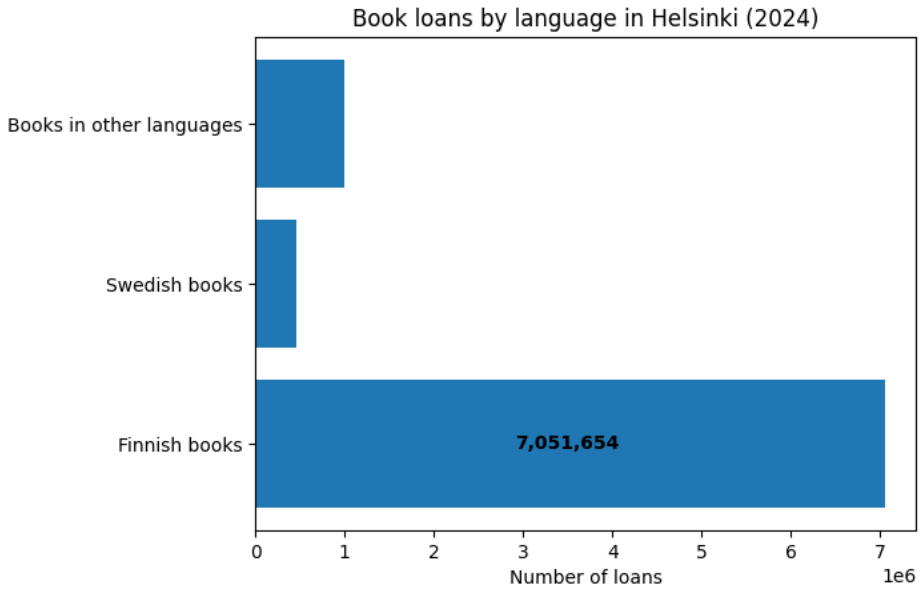

Figure 14 CASE 1: Book loans by language in Helsinki (2024)

The second case, which shows the change in the number of events from 2015 to 2024, is illustrated in Figure 15 using a line chart.

During the design process, a choice was made between using clearly distinct colours for each city (e.g., blue and orange) or relying on a single main colour with varying shades. While the choice is mostly a matter of preference in line charts, it is generally more effective to stick with one main colour, as noted in Chapter 2.3.1. This guideline was also followed in this case.

Besides the choice of colours, colour vision deficiencies were also considered when selecting markers. Each city was assigned a unique symbol, which supports both accessibility for colour-blind viewers and legibility if the chart is printed in grayscale.

The need for axis headings was considered, since the years can be interpreted without a label, and the main title already indicates that the y-axis shows the number of events. However, it was decided to include the headings and to follow Firdose's (2024, Chapter 10) approach to line plot design. The orientation of y-axis title follows Firdose's example and reflects standard practice, despite its slight impact on readability (see Chapter 2.3.2).

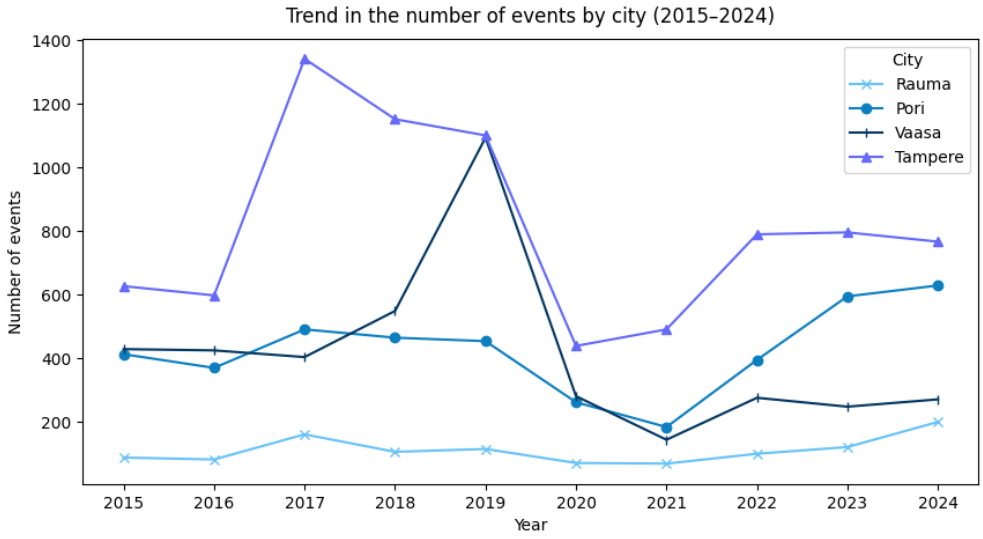

Figure 15 CASE 2: Number of events by city (2015–2024)

The third case is illustrated in Figure 16 using a scatter plot. As discussed in Chapter 2.2.1 regarding scatter plots, the independent variable is plotted on

the x-axis, and the dependent variable on the y-axis. Since the assumption is that e-material expenses affect the number of e-material loans, the variable "Expenses" was plotted on the horizontal (x) axis, and "Number of loans" on the vertical (y) axis.

In the process of creating this graph, it became clear that a good understanding of the dataset is important, not only for creating the graph itself but also for making it useful in supporting decision-making. It was therefore considered important to indicate the currency and specify that value-added tax (VAT) is excluded, as this information is relevant for interpreting the graph in a decision-making context.

Additionally, the content of Chapter 2.3.3 on ineffective visualizations was taken into account. This chapter includes a quote from statistician Edward Tufte about changes in visual design. The guidelines presented there influenced the colour choice for this plot. To maintain consistency and align with the principles discussed in Chapter 2.3.3, the same shade of blue that was used in Figure 15 was applied to the data points in Figure 16.

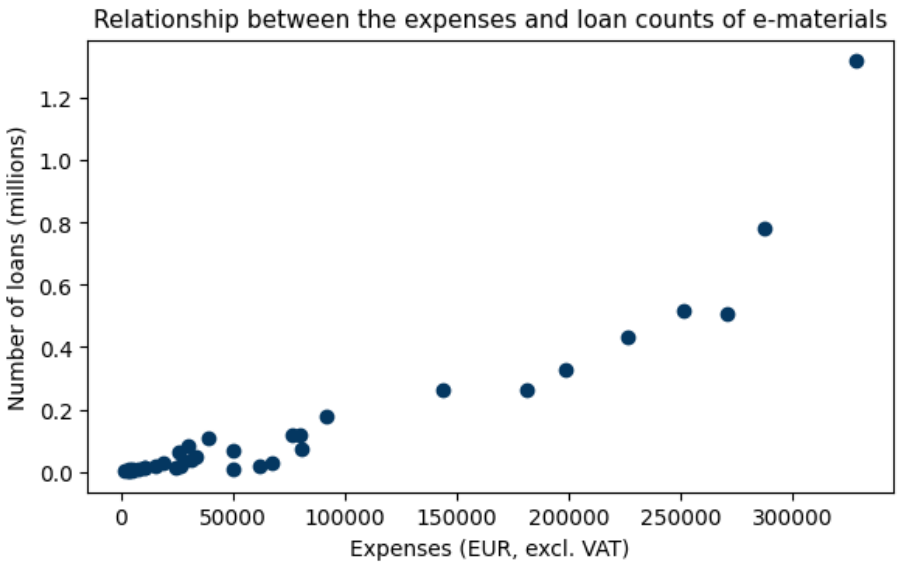

Figure 16 CASE 3: Relationship between the expenses and loan counts of e-materials

5 DISCUSSION

5.1 Interpretation of Results

The case study demonstrated that the foundation of effective data understanding lies in having a thorough understanding of the dataset and the visualization topic, identifying the data types, and selecting a suitable graph accordingly.

In addition, careful use of colours was found important due to their strong visual impact and the need to ensure accessibility for people with colour vision deficiencies.

The case study also reinforced the findings from the literature review regarding the importance of simplicity in graphics. Visuals that focus on a single topic and avoid unnecessary elements and design variations were found to support data understanding.

Finally, the case study revealed the importance of being familiar with different types of graphs and their characteristics, as this understanding supports the selection of graphs based on the type of data and the topic of the visualization.

5.2 Future Work

Future work could explore whether companies and organizations would benefit from a more comprehensive guide of this kind. Currently, many existing guides focus primarily on the technical aspects of creating graphs, while deeper knowledge about different graph types remains limited, and available information is scattered across various sources.

6 CONCLUSION

This thesis addressed the need for a framework to support the selection and design of effective plots, particularly for users without extensive experience in data analytics, in response to the growing volume of data generated by modern organizations and the increasing availability of visualization tools.

The research question introduced in Chapter 1 asked:
“What are the best practices for selecting and designing effective plots in Exploratory Data Analysis to optimize data understanding and decision-making?”

To address this question, a two-part approach was applied: first, a literature review was conducted to identify common plot types, their use cases, and basic principles of designing visualizations; second, these findings were tested through a case study using a real dataset. The results from both parts were aligned and supported the same conclusions.

The conclusions are as follows:
To choose an appropriate graph, the analyst must thoroughly understand both the dataset and the topic of the visualization. Identifying the data types and the purpose of the visualization is essential. In addition, familiarity with the characteristics of common plot types is crucial, as this knowledge enables the selection of a graph that fits both the nature of the data and the topic being explored.

This thesis supports these conclusions by compiling key information on data types and common plot types into a single, accessible source. The resulting framework can assist individuals new to data analytics in selecting and designing effective visualizations that support data understanding.

REFERENCES

Brath, R., Jonker, D. (2015). Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data. John Wiley & Sons. https://doi.org/10.1002/9781119183662

Burnay, C., Dargam, F., Zarate, P. (2019). Special issue: Data visualization for decision-making: an important issue. Operational Research. 19(4), 853–855. https://doi.org/10.1007/s12351-019-00530-z

Finnish Public Libraries Statistics. (n.d.). Yearly reports: 2015-2024 [Dataset]. Retrieved July 22, 2025, from: https://tilastot.kirjastot.fi/yearlyreports.php

Firdose, T. (2024). Ultimate Pandas for Data Manipulation and Visualization: Efficiently Process and Visualize Data with Python's Most Popular Data Manipulation Library (English Edition). Orange Education PVT Ltd.

Koponen, J., Hildén, J. (2019). Data visualization handbook. Aalto University.

Kosslyn, S. (2006). Graph Design for the Eye and Mind. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195311846.001.0001

Magnuson, L. (2016). Data Visualization: A Guide to Visual Storytelling for Libraries. Rowman & Littlefield Publishers.

McAfee, A., Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review. 90(10), 60-68. https://hbr.org/2012/10/big-data-the-management-revolution

Myatt, G., Johnson, W. (2014). Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining (second edition). John Wiley & Sons. https://doi.org/10.1002/9781118422007

National Eye Institute. (2023, August 7). Types of Color Vision Deficiency. U.S. Department of Health and Human Services. Retrieved July 1, 2025, from: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/color-blindness/types-color-vision-deficiency

Nica, M. (2013). Principles of business statistics. OpenStax CNX.

Nussbaumer Knaflic, C. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. John Wiley & Sons. https://doi.org/10.1002/9781119055259

Sahay, A. (2016). Data Visualization, Volume I : Recent Trends and Applications Using Conventional and Big Data. Business Expert Press.

Seaborn. (n.d.). API reference: seaborn.heatmap. Retrieved July 10, 2025, from: https://seaborn.pydata.org/generated/seaborn.heatmap.html

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

Unwin, A. (2020). Why Is Data Visualization Important? What Is Important in Data Visualization? Harvard Data Science Review. 2(1). https://doi.org/10.1162/99608f92.8ae4d525

Yau, N. (2013). Data Points: Visualization That Means Something. John Wiley & Sons.