Activity 4 – data visualization and analytics – Group 2

Alexandre Baptista
Informatics Department
Faculdade de Ciências da Universidade de Lisboa
Lisbon Portugal
fc64506@alunos.fc.ul.pt

Vram Davtyan
Informatics Department
Faculdade de Ciências da Universidade de Lisboa
Lisbon Portugal
fc64691@alunos.fc.ul.pt

André Pires
Informatics Department
Faculdade de Ciências da Universidade de Lisboa
Lisbon Portugal
fc64347@alunos.fc.ul.pt

INTRODUCTION

This activity is a sequel of activity 2 GitHub, where the research question, “What is the type of car with the biggest sales between 2004 and 2024 in the USA?”, was defined. Activity 2 focused on analyzing trends and patterns in the automotive industry over two decades using the publicly available "Used Cars Dataset" created by Austin Reese on Kaggle.

The original dataset contains attributes that provide a robust foundation for analyzing the used car market, comprising 426,000 advertisements and 26 attributes that describe either car details or sales advertisements. For Activity 2, this dataset was transformed and reduced to implement an effective data cleaning strategy and address the research question. The script used for automation in the previous activity will be extended, and the dataset for this activity contains only information about the number of advertisements categorized by year and car type.

The goal of this activity is to explore various data visualizations to answer the research question and test the principles of graphical excellence. Four visualizations, each representing a unique category according to A. Abela (2006), are produced to showcase different perspectives: comparison, distribution, composition, and relationship. Additionally, two visualizations will be subjected to critical comparison, where one follows the best principles outlined in class, and the other intentionally breaks several of these principles.

The project seeks to bridge theoretical principles with practical application, leveraging Python and open-source tools, namely Pandas, which is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool[2], and matplotlib, to create the visualizations[3].

CONTRIBUTIONS

Each group member dedicated approximately 6 hours to the project, distributed across key tasks: 3 hours were spent creating visualizations in Python, 1 hour was dedicated to discussing and providing feedback on the visualizations, and 2 hours were allocated to collaboratively writing the report, documenting processes, insights, and conclusions.

QUESTION 1

The first question involved producing four different types of visualizations to represent the dataset from multiple perspectives: comparison, distribution, composition, and relationship.

Comparison visualizations are used to contrast values across different categories, often highlighting differences or ranking among them. A radar chart was chosen to depict the predominant selling car types. While this chart provides an easy overview of which car types dominated sales, it does not effectively convey temporal trends due to the overlapping lines and the limited scale of the chart area.

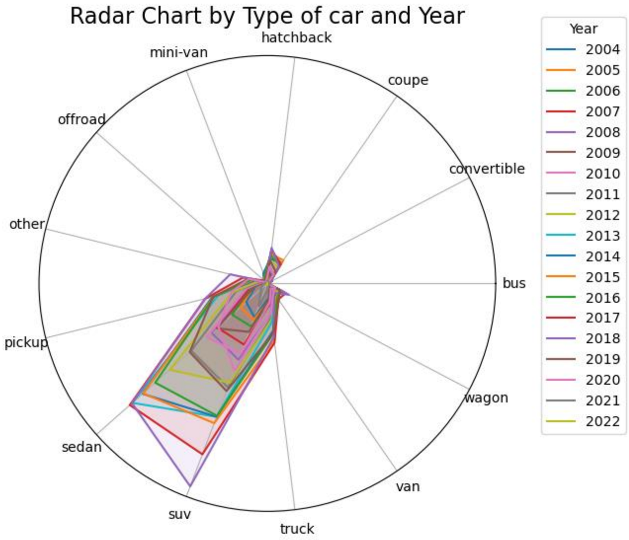

Figure 1: Radar chart

Distribution visualizations illustrate how values in a dataset are spread, enabling the identification of patterns, variations, and anomalies. A scatter plot was utilized in this project to display overall data distribution. While the scatter plot effectively highlights trends over the years, it falls short in clearly representing less frequently sold car types due to data clustering.

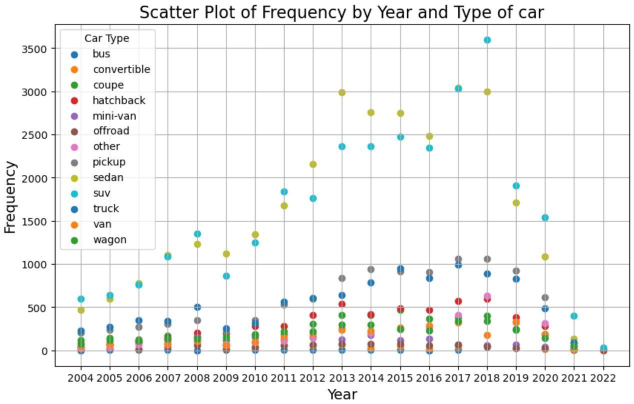

Figure 2: Scatter plot

Composition visualizations show how a whole is divided into parts, illustrating the proportions or contributions of different components. A 100% stacked column chart was used to compare the relative proportions of car types sold each year. This approach allows easy identification of how the composition of car sales evolved over the time period, considering each year as 100% of total sales.

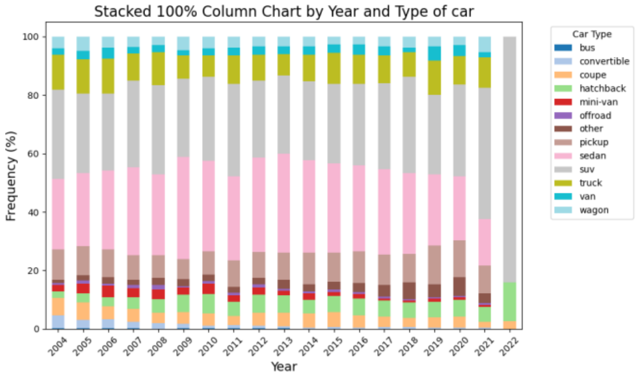

Figure 3: 100% stacked column chart

Relationship visualizations demonstrate how variables are interconnected, revealing patterns, correlations, or connections. A bubble chart was employed to enhance the multidimensional representation of car type frequencies. While this visualization captures relationships effectively, it is important to note that relationships do not imply causation. Interpretations should be made with caution.

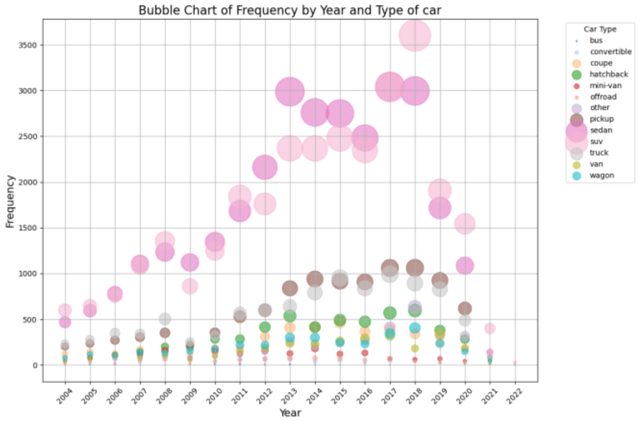

Figure 4: Bubble chart

QUESTION 2

The second task explores the application of Principles of Graphical Excellence in data visualization, as outlined by Edward Tufte. Edward Tufte emphasizes four key principles that lead to an effective data visualization: clarity, precision, efficiency, and the balance between maximizing ideas and minimizing ink.

Starting by visualization that follows the best principles, figure 5, reach its purpose in making the data accessible, accurate, and visually easy to communicate. The use of colors, detailed axis labels, and a balanced scale ensures that viewers can quickly grasp trends and relationships.

Visualization that breaks several of the best principles, figure 6, demonstrates the pitfalls of ignoring best practices. The starting data points of the Y-axis at 300 distorts the data, the lack of colors for the differentiation between car types, several years, in the X-axis, are missing, which leave viewers guessing about the data's evolution, and the title of the Y-axis is missing.

The principles are broken multiple times in the chart of figure 6. Regarding clarity, the absence of colors and reliance on black lines without a legend reduce readability, especially where overlapping lines create visual clutter. Regarding precision, the omission of intermediate years on the X-axis diminishes the ability to interpret temporal changes precisely, creating confusion. Regarding efficiency, without a legend or clear differentiation between car types, viewers spend more time trying to understand the chart, reducing efficiency. Regarding maximizing ideas and minimizing ink, the chart lacks visual distinction and missing labels reduces the information that is displayed but fails to maximize ideas. So, Figure 6 illustrates how poorly designed data visualizations can result in ineffective communication of information.

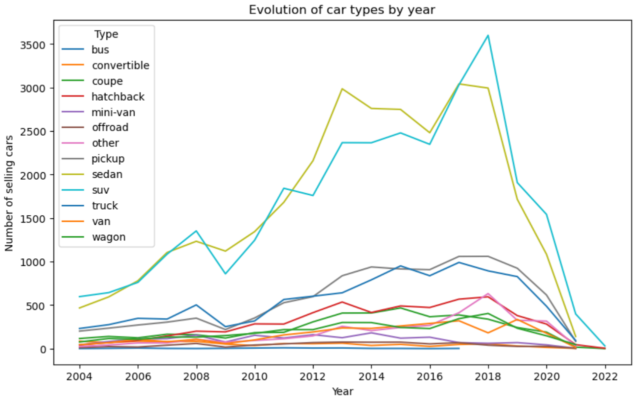

Figure 5: visualization that follows the best principles

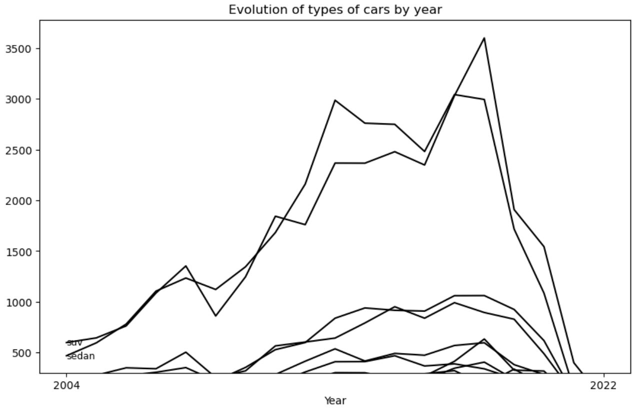

Figure 6: visualization that breaks several of the best principles

More Examples

- **Plot 1** uses a stacked bar chart, providing a clean, simple, and easy-to-interpret visualization with clear axis labels and minimal clutter.
- **Plot 2** uses a grouped bar chart but introduces unnecessary complexity with missing gridlines, 90 degree labels, making it visually cluttered and harder to read.

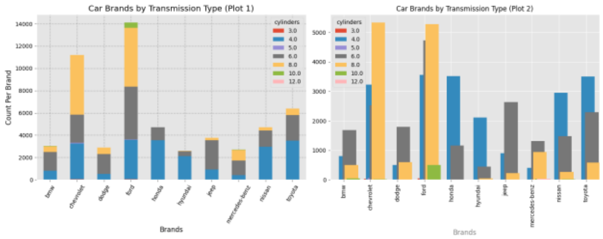

Figure 7: Comparison of Car Brands by Transmission Type: Stacked vs. Grouped Bar Charts

- **Left Plot**: A scatter plot that displays the top 20 car models by average price, where each point represents a car model and its average price. The points are clear and easy to interpret with minimal clutter.
- **Right Plot**: A horizontal bar chart displaying the same data, but as bars. The bars provide a more precise way to compare the prices across different models, though it takes up more space. Bars are not sorted, making it harder to interpret.

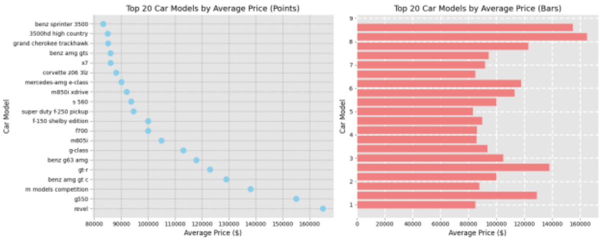

Figure 8: Comparison of Top 20 Car Models by Average Price: Scatter vs. Bar Plot

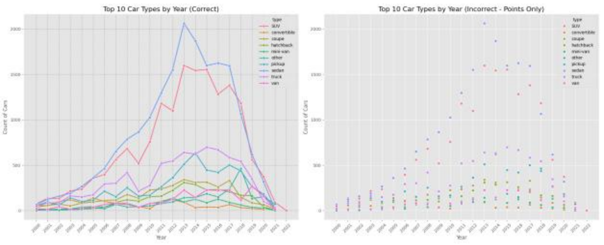

Figure 9: Slightly different approach to the previous figures (5 and 6)

Although everything in the grouped pie chart appears visually correct at first glance, in some cases, the labels may become squished.

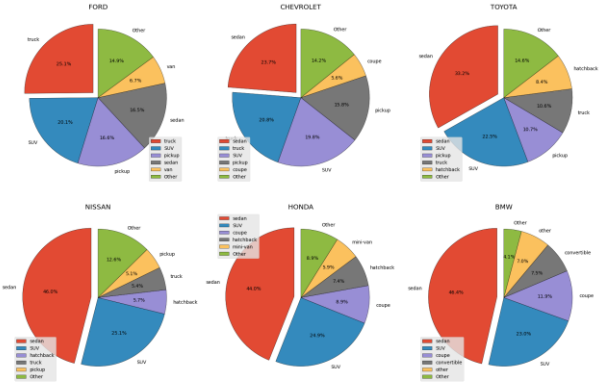

Figure 10: Pie Chart Visualization of Car Type Distribution

Here’s an alternative approach.

- **The first example** uses pie charts to display the distribution of car types for the top 6 brands, highlighting the largest segments with labels showing the percentage.
- **The second example** uses radial bar charts, offering a more modern and visually distinct design, where car types are represented by bars arranged in a circle. It sorts the data, adds percentage labels outside the chart with color-coded dots, and removes axis ticks for a cleaner presentation.

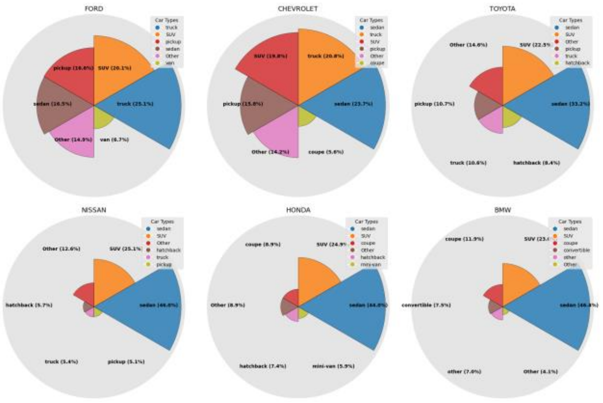

Figure 11: Radial Bar Chart Visualization of Car Type Distribution

REFERENCES

[1] https://pandas.pydata.org
[2] https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
[3] GitHub

ANNEX

Original data in table

	type	year	Frequency
0	bus	2004.0	2
1	bus	2005.0	4
2	bus	2006.0	7
3	bus	2007.0	3
4	bus	2008.0	2
...	...	...	...
228	wagon	2017.0	345
229	wagon	2018.0	405
230	wagon	2019.0	235
231	wagon	2020.0	148
232	wagon	2021.0	47