# Introduction

An introduction to the methodology used within this project.

Skip to results

## Initial proposal

Not all cars are made the same and the price usually reflects that. I want to examine the key factors that drive the price of used vehicles, whether that be the car's make, model year or other factors. Then, I would like to further analyze which models retain their value longer or shorter than the average, as certain models are still in demand many years later. The current hypothesis is that popular vehicle makes in Slovakia (Skoda, etc.) will be in demand more due to the availability of spare parts.

## Objectives

find out **which factors influence the price the most**

find out **which makes retain their value the best**

build a **web application** to display the data

## Used data

All data have been scraped from the online portal with used cars - **autobazar.eu**.

## Acquiring data - part I.

The first step was collecting the URL addresses of all offers on the portal. The relevant script for this was `cars_spider.py`. The script scraped the autobazar.eu website for all car listings. However, the website was limited to 500 pages of results. To circumvent this, I had to split the scraping into different price intervals/categories to obtain all (most) results. Furthermore, the price was limited to the interval [500, ∞) to filter out results without a price, as well as rentals (more expensive rentals might still be included). The addresses were then saved to the file `car_links.txt`.

## Preprocessing data - part I.

Any duplicates within the collected addresses had to be subsequently removed. This was handled by the script `unique.py`. Results were saved to the file `car_links_unique.txt`. This has the same functionality as the Linux commands `sort | uniq` or the Windows commands `type | sort -unique`.

## Acquiring data - part II.

.The next step was scraping the individual offers for data. `car_details_spider.py` scraped all the unique addresses for data. It accessed relevant elements via their respective xpath values. It also automatically converted string numbers to integers and formatted strings. It set a standard for caterogical data and filtered out entries with nmissing values (except color). All this was then saved into the database.

## Preprocessing data - part II.

After all the aforementioned scrips had finished running, manual validation had to be done. SQL queries were run to verify there were no invalid data and no alternative spellings for the car makes (e.g. VW and Volkswagen). Due to their low amount, these were manually corrected.

# Results

Initial results of the data collection.

Skip to analysis

**Number of listings**

38821

cars for sale

**Number of manual transmissions**

15260

**Number of automatic transmissions**

23561

**Most frequent make**

Škoda

4911 offerings

**Most frequent year**

2024

5370 offerings

**Most frequent drivetrain**

FWD

22697 offerings

**Average mileage**

111405

kilometers

**Average cost**

27304

euros

**Average power**

132

kW

# Analysis

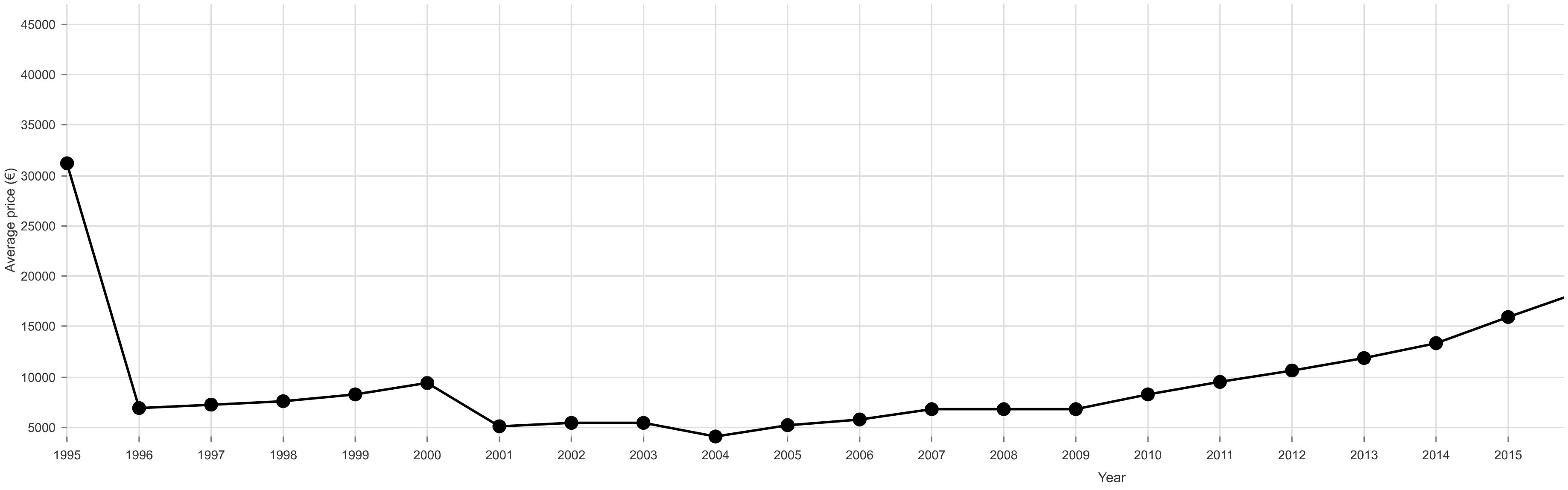
Analysis of the obtained data.

Back to beginning

# Obtained results

In the previous part we have confirmed our hypothesis that Skoda would be the most popular car make. Surprisingly, automatic transmissions are more abundant despite the prevalence of manual transmissions in Europe.

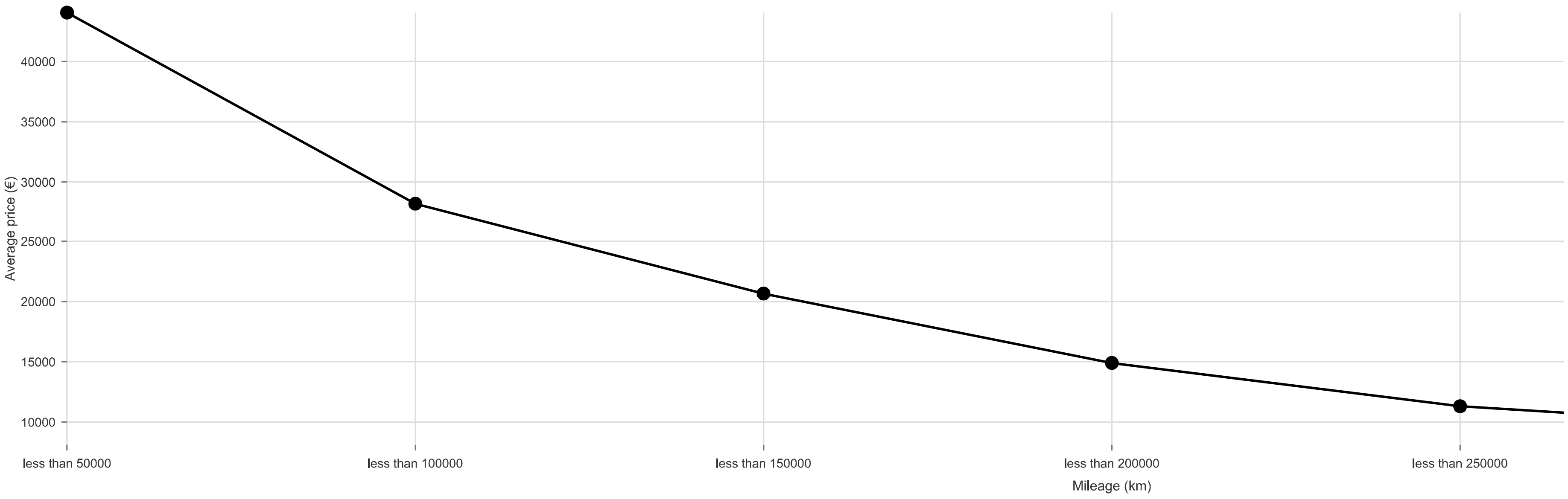
### Cost by model year (past 30 years)



# Analysis

Here, we see a steady exponential increase in price with increasing model year, as expected. This growth, however, gets slower as we approach the present year. This is most likely due to cars not depreciating in value as much within the first few years. Cars made before 2000, and even more so before 1996, sell for higher price - most likely due to their rarity, i.e. lower supply.

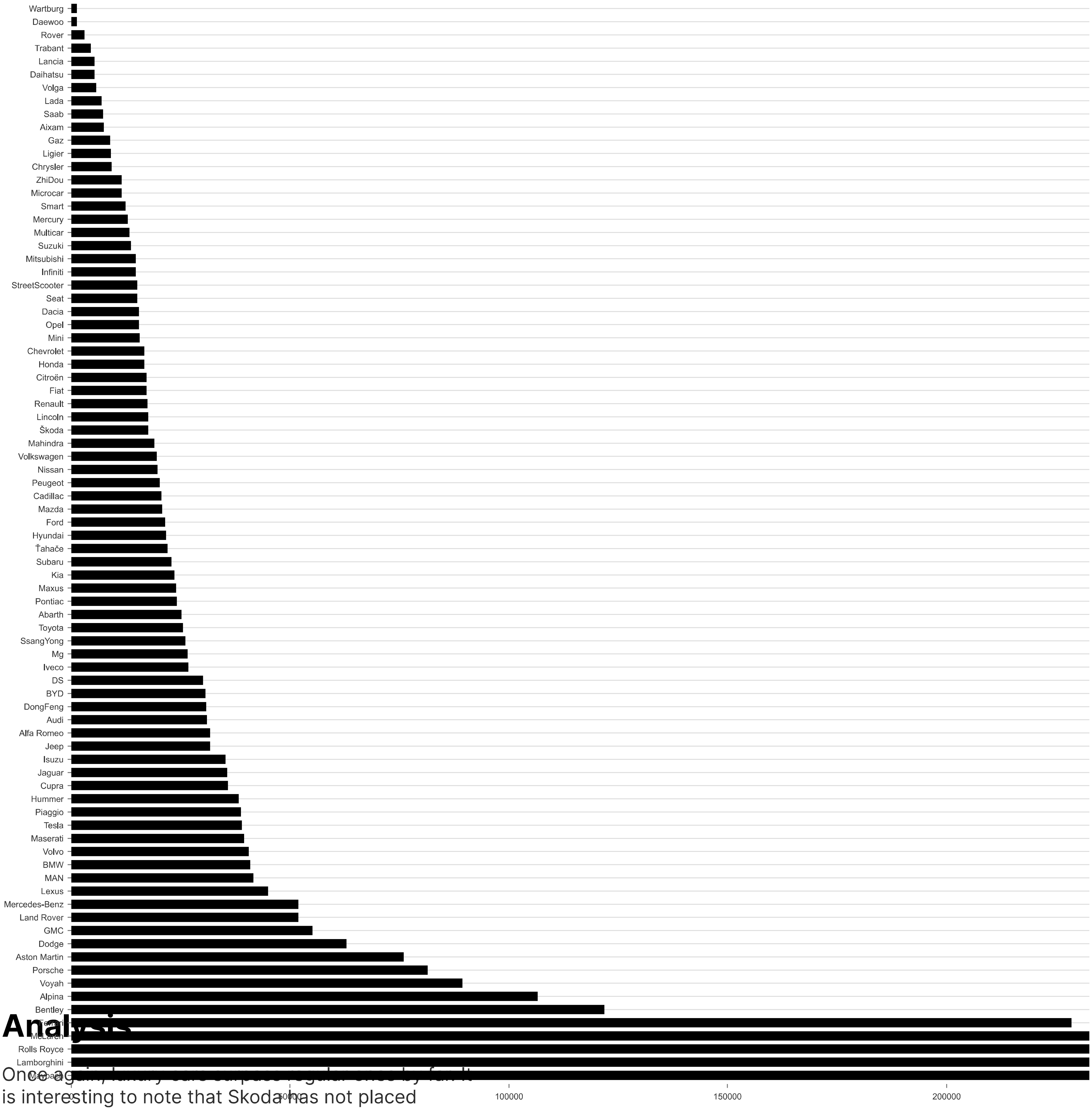
### Cost by mileage



# Analysis

Nothing unexpected here, either. The price goes down with increasing mileage, as the cars get less reliable and more expensive to fix.

### Cost by make



## Analysis

Once again, luxury cars surpass regular ones by far. It is interesting to note that Skoda has not placed within the cheapest makes. Toyota has also placed somewhat high, especially compared to other Japanese brands within the same category, e.g. Honda. Maybach has outdone Rolls Royce, despite Rolls Royce being generally the more expensive brand.

## Final thoughts

Most of my hypotheses were during the course of this project confirmed. Although, it was interesting to compare the general trends of the market with some commonly thought concepts.

As I have decided to create a web app, I was met with lots of difficulties, most of them related to the technologies that I have chosen. As I wanted to work mostly with front end, connecting to the database has been rather difficult as most libraries are back end only. Upon further research I have come upon the `sql.js` library which has solved my problems.

Another difficult part has been the visualization, as the library used (nivo) is not very well documented.

This part was mostly trial and error. The non-web related parts were rather easy.

As a result, I have learned many new technologies (scrapy, React, many JS libraries and frameworks) which I look forward to using in the future. I would not do anything differently as I think everything was a good learning experience.