The Fundamental Statistics and Data Visualization
Group Assignment

Chang Jia Qian ANQSCL
Nguyen Hong Gia Bao ELS9CY
Nguyen Thanh Trung I6P5GG
Zhanbolat Baiekenov EL8SVB

Dataset chose: Cars

Table of content

Background

Our group has chosen the Cars dataset for the Fundamentals of Statistics and Data Visualization group assignment. Our task is to analyze the dataset and make a report based on the observations. We can also Discuss some relevant social and/or economic questions that can be answered through statistical analysis.

Inspired by an Instagram reels showing a successful airplane dealer showing his clients the right private airplane for them based on his huge dataset, our group has decided to imagine ourselves as the biggest and best car dealer in the world to make things fun and creative. The analysis our group has done can help to understand our imaginary company better for internal management and also provide a suitable car for our imaginary clients based on their needs and wants. Our group will also use the data provided to address economic and environmental issues.

From this point onwards, members of this group assignment will name ourselves as “our company”, “the company”, “our group”, “we” interchangeably and we will call our readers “clients” or “professors” interchangeably. Since this is an imaginary company, names of the departments of the company can also be used throughout this report to describe the analysis. For example, the marketing department found out that most people would buy a car for a price of 500,000HUF (mode) and so we can market our cars to most of our clients that are priced around this number.

Research Questions

- How does the popularity of car brands relate to selling prices**?**
- What is the market concentration in each region based on selling price and the effect on our company's marketing and distribution of cars in different regions?
- What is the relationship between the mileage and selling price and how does it influence some of our economic conscious customers?
- What is the effect of the vehicle's age and its selling price?
- How does the popularity of car brands relate to selling prices**?**
- What is the relationship between car brands and their fuel type and how does it impact the environment?
- How can our company market our cars in different regions based on the car's transmission type and the reason behind it?
- How does the number of seats in a car relate to its selling price and how would it affect different kinds of customers?

Codebook/Metadata

Variable Name	Description	Type	Possible Values	Units	Scale of Measurement
brand	The brand or manufacturer of the car	Categorical	e.g., Toyota, BMW, Ford, Audi, etc.	-	Nominal
origin	The continent where the car brand originates	Categorical	Asia, Europe, North America, etc.	-	Nominal
Vehicle age	Age of the car, calculated based on the year of manufacture	Numeric	Positive integer (e.g., 1, 2, 5, 10, etc.)	Years	Ratio

	and current year				
Km driven	The total kilometers driven by the car	Numeric	Positive integer (e.g., 5000, 120000, etc.)	Kilometers (km)	Ratio
Seller type	Type of seller offering the car for sale	Categorical	Individual, Dealer	-	Nominal
Fuel type	The type of fuel the car uses	Categorical	Petrol, Diesel, Electric, Hybrid	-	Nominal
Transmission type	The type of transmission system in the car	Categorical	Manual, Automatic	-	Nominal
mileage	The car's fuel efficiency in terms of distance traveled per liter of fuel	Numeric	Positive number (e.g., 10, 15.5, 20, etc.)	Kilometers per liter (km/l)	Ratio
engine	The engine capacity of the car, usually in	Numeric	Positive integer (e.g., 1000, 1500, 3000)	Cubic centimeters	Ratio

	cubic centimeters (cm^3)				
seats	The number of seats in the car	Numeric	Positive integer (e.g., 2, 5, 7)	Seats	Ratio
Selling price	The price at which the car is being sold	Numeric	Positive integer (e.g., 500000, 1500000)	Hungarian Forint (HUF)	Ratio

Additional Notes:

- **Brand**: This is typically a categorical variable, representing the name of the car manufacturer. You can include a list of the possible brands in the dataset.
- **Origin**: This variable gives a sense of the geographical origin of the car brand. Each brand would have one associated continent of origin (e.g., Toyota is from Asia, BMW is from Europe).
- **Vehicle Age**: This is calculated by subtracting the car's year of manufacture from the current year. It helps analyze the relationship between age and selling price, mileage, and other factors.
- **Km Driven**: This represents how much the car has been driven. Higher kilometers often correspond to lower selling prices due to wear and tear. It’s important to consider how this variable correlates with the car's selling price and mileage.
- **Seller Type**: This indicates whether the car is being sold by an individual or a dealership. Seller type may influence the pricing, as dealers often offer warranties and other services.
- **Fuel Type**: This tells us about the energy source powering the car. Different fuel types might have varying impacts on the vehicle's efficiency, running costs, and environmental considerations.

- **Transmission Type**: Transmission type (manual or automatic) may affect the car's price and demand in certain markets.
- **Mileage**: Represents the fuel efficiency of the car. A higher mileage typically suggests a more fuel-efficient vehicle, which can influence both the car's price and potential buyer preferences.
- **Engine**: The engine size in cubic centimeters (cm^3) reflects the power of the engine, which can affect both the performance and price of the car.
- **Seats**: The number of seats is typically a factor that relates to the car's size and purpose. Cars with more seats might be considered more suitable for families, potentially influencing their pricing.
- **Selling Price**: This is the target variable in our dataset, representing the price at which the car is sold. It is often the result of multiple factors like age, brand, mileage, engine capacity, etc.

Introduction

Our company is selling cars from a total of 38 brands. These cars can be divided into 2 main categories, luxury cars and normal cars. The luxury brands include Land Rover, BMW, Jaguar, Volvo, Mercedes-Benz, Porsche, Bentley, Lexus, Audi, Maserati, Mercedes-AMG and Mini. The rest are labeled as normal.

In our original dataset, 8 extreme outliers were included. However, for this analysis, we have excluded it as it has highly affected our analysis. The 8 extreme outliers excluded are very rare and uncommon models our company sells. Hence, to better understand our company's business, we excluded it as selling those cars are not part of the company's usual operations.

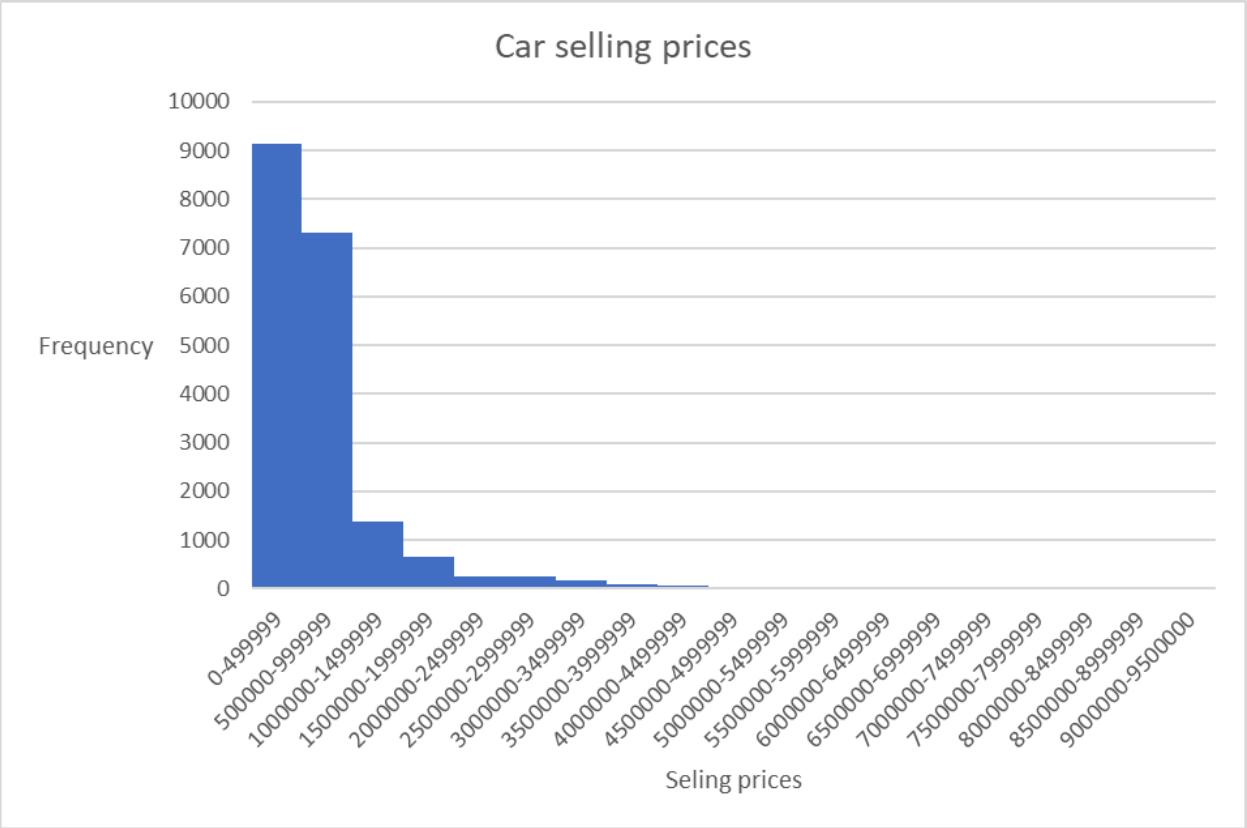
There are a total of 1,493 luxury cars and 18,039 economical or non-luxury cars in the filtered dataset. The highest price for the luxury and non-luxury cars in our company is priced at HUF 9,200,000. However the average price and lowest price for the luxury cars are HUF 2,19,462 and HUF 315,000 respectively which are higher than the average price and lowest price of the economic cars which are HUF 578,645 and HUF 30,000 respectively.

Central Tendencies and Distribution Characteristics (overall data)

Measurement of location	Selling Prices (HUF)
Mode	HUF 450,000.00
Median	HUF 525,000.00
Mean	HUF 734,549.69
Minimum	HUF 30,000.00
D1	HUF 230,000.00
Q1	HUF 349,000.00
Q3	HUF 785,000.00
D9	HUF 1,350,000.00
Max	HUF 9,200,000.00
Number of observation (N)	19532
Range	HUF 9,170,000.00
IQR(0.5)	HUF 436,000.00
IDR(0.8)	HUF 1,120,000.00
IDR - IQR	HUF 684,000.00
Variance	645646101641
Standard Deviation	803521
Relative standard deviation	1.09
Pearson skewness of measurement	0.78
F0.25	0.19
F0.1	0.47
α3	4.23
α4	24.43
K for kurtosis	0.195
Tukey's lower fence	-305000
Tukey's upper fence	2465000
Number of cars that are outliers	788
% of cars that are outliers	4.03%

Table 1.

The dataset reveals that the mode of car prices is HUF 450,000. However, in our large continuous dataset, this result is less helpful to understand our business. In our large dataset, the repeated value might not repeat enough to count as a useful point for the company to consider as there are so many cars priced at different values. Therefore, it is better to consider the bin of prices rather than a precise point in this case to bring out a more meaningful interpretation. Hence, based on Graph 4, the most frequent price range is from HUF 0 to HUF 499,999. This reflects a concentration of lower-priced cars that likely appeal to budget-conscious customers.



Copy of Graph 2. Distribution of Car selling prices

Num of observation	19532
Number of bins(K)	15
Min	30000
Max	9200000
Bin length	611333

Table 2. Calculation of number of bins and bin length with Sturge's rule

The median, at HUF 525,000, signifies the midpoint of the dataset, with half of the prices falling below and half above this value. Notably, the mean is significantly higher at HUF 734,549.69, highlighting the influence of high-value outliers. The disparity among the mean, median, and mode suggests a positively skewed distribution, where a small number of high-priced cars elevated the average.

The positive skewness is confirmed by a Pearson skewness coefficient of 0.78 and an α_3 value of 4.23, which both indicate a pronounced rightward tail. The value for Pearson skewness coefficient's value is small because it rarely takes a larger value than 1 in a positively skewed

distribution. This is because it does not handle highly skewed datasets well. However the value for α_3 is high because α_3 is sensitive to outliers. It takes all the data points into consideration. Hence being more precise compared to Pearson skewness of measurement. On a quick look, Pearson skewness coefficient helped us to understand that the prices are distributed in a right skewed manner and α_3 helped us to understand how highly the distribution is skewed to the right. The distribution can be seen on Graph 4 .

This distribution implies that while most cars are relatively affordable, the presence of high-priced models plays a significant role in shaping the overall price structure. A further examination of kurtosis reveals an α_4 value of 24.43, indicating a leptokurtic distribution. This sharp peak and heavy tails suggest that most prices are tightly clustered around the central tendency, with substantial outliers influencing the upper range.

Measures of Dispersion (overall data)

The range of HUF 9,170,000, calculated from HUF 30,000 to HUF 9,200,000 demonstrates the large variability in selling prices. This reflects a diverse inventory catering to different market segments, from economy models to luxury vehicles. The interquartile range (IQR), at HUF 436,000, narrows the focus to the middle 50% of prices, providing a clearer picture of central price variation. The standard deviation of HUF 803,521 indicates that, on average, car prices in the company's dataset vary by about HUF 803,000 from the mean. This reflects substantial variability in the prices, with some cars priced significantly higher than the average. The large standard deviation also suggests the presence of outliers that are driving this variation.

Analysis of quartiles and deciles provides deeper insights. The lower quartile (HUF 349,000) and upper quartile (HUF 785,000) indicate that the central 50% of prices fall within this range, with an interquartile range (IQR) of HUF 436,000. Meanwhile, the lower decile (HUF 230,000) and upper decile (HUF 1,350,000) show that 80% of prices span a broader range, with an interdecile range (IDR) of HUF 1,120,000. The IDR is HUF 684,000 wider than the IQR, indicating a significantly broader spread of prices when considering the middle 80% of the data, compared to the middle 50%. This suggests that while most prices are clustered around the central range of 50%, the broader distribution captures more extreme values, extending beyond

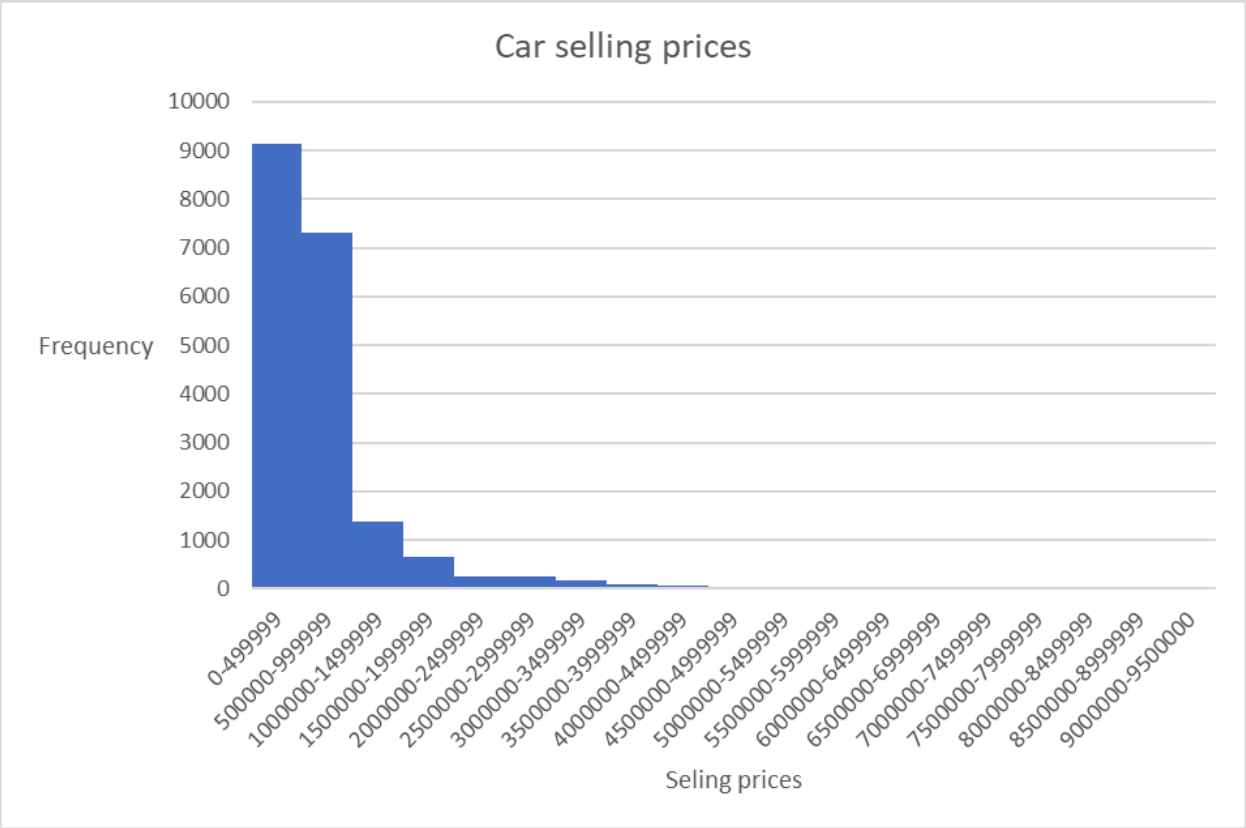
the central 50% of the dataset. These findings underscore the importance of understanding price dispersion to address varying customer needs effectively.

Box plot analysis

Graph 1. Car selling prices using box plot

Based on calculation, Tukey's lower fence is -305,000 HUF. However in real life, this is impossible as no car will be sold with a negative price, hence in reality the Tukey's lower fence is 0 HUF. The Tukey's upper fence is at 2,465,000 HUF. This means that all the 788 cars(4.03% of total cars) that are priced higher than the Tukey's upper fence are considered outliers. Therefore the prices are not common and it would increase the mean of the price of the cars.

Histogram analysis and market implications (overall)



Graph 2. Distribution of Car selling prices

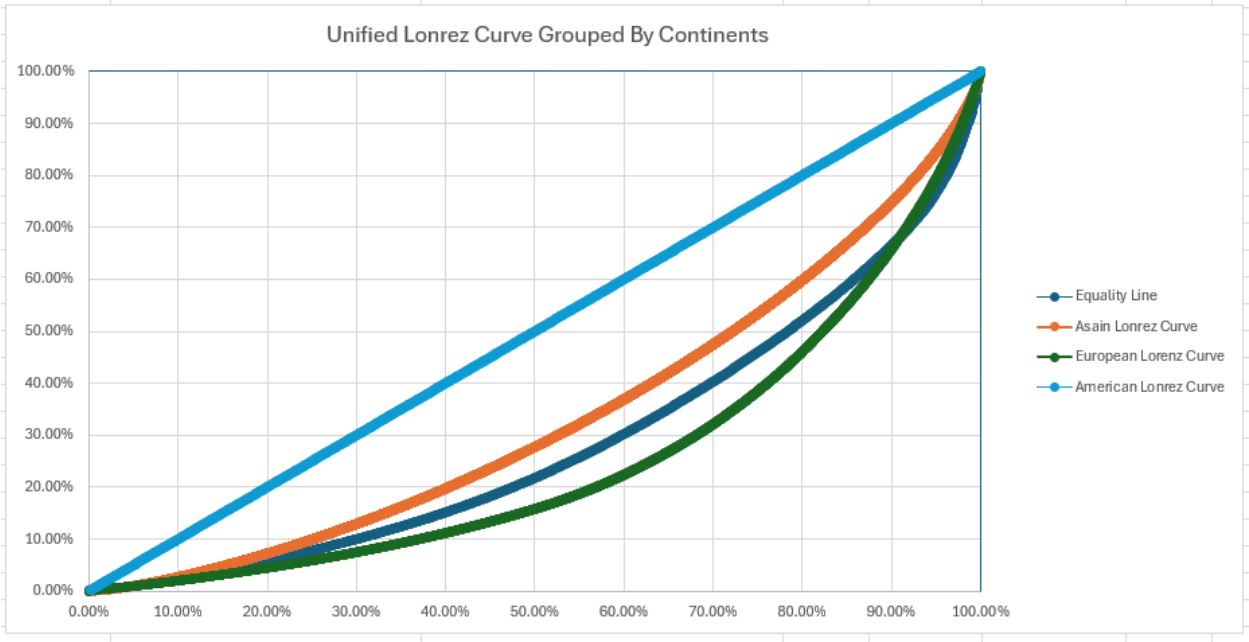
The dataset contains a total of 19,532 observations. To visualize the distribution of prices, we decided to use 15 bins based on Sturges' Rule. The prices range from a minimum value of 30,000 HUF to a maximum of 9,200,000 HUF. The bin length was calculated by dividing the price range by the number of bins, resulting in a bin length of approximately HUF 611,333. However in our histogram we use the bin length of HUF 500,000 to make the grouping more easier and even.

A positively skewed histogram would depict most prices clustered near the lower end, around the mode and median, with a long tail extending to higher price points. This visual representation aligns with the data's characteristics, highlighting the affordability of most cars while showcasing the impact of high-value models.

From the company's business perspective, the pricing structure reveals distinct opportunities. The clustering of prices near the lower end indicates a high demand for affordable models. To cater to this segment, the company should ensure consistent inventory levels and

competitive pricing. Conversely, the high-value outliers represent an opportunity to market premium vehicles more effectively, emphasizing their unique features and justifying their higher prices.

Market concentration



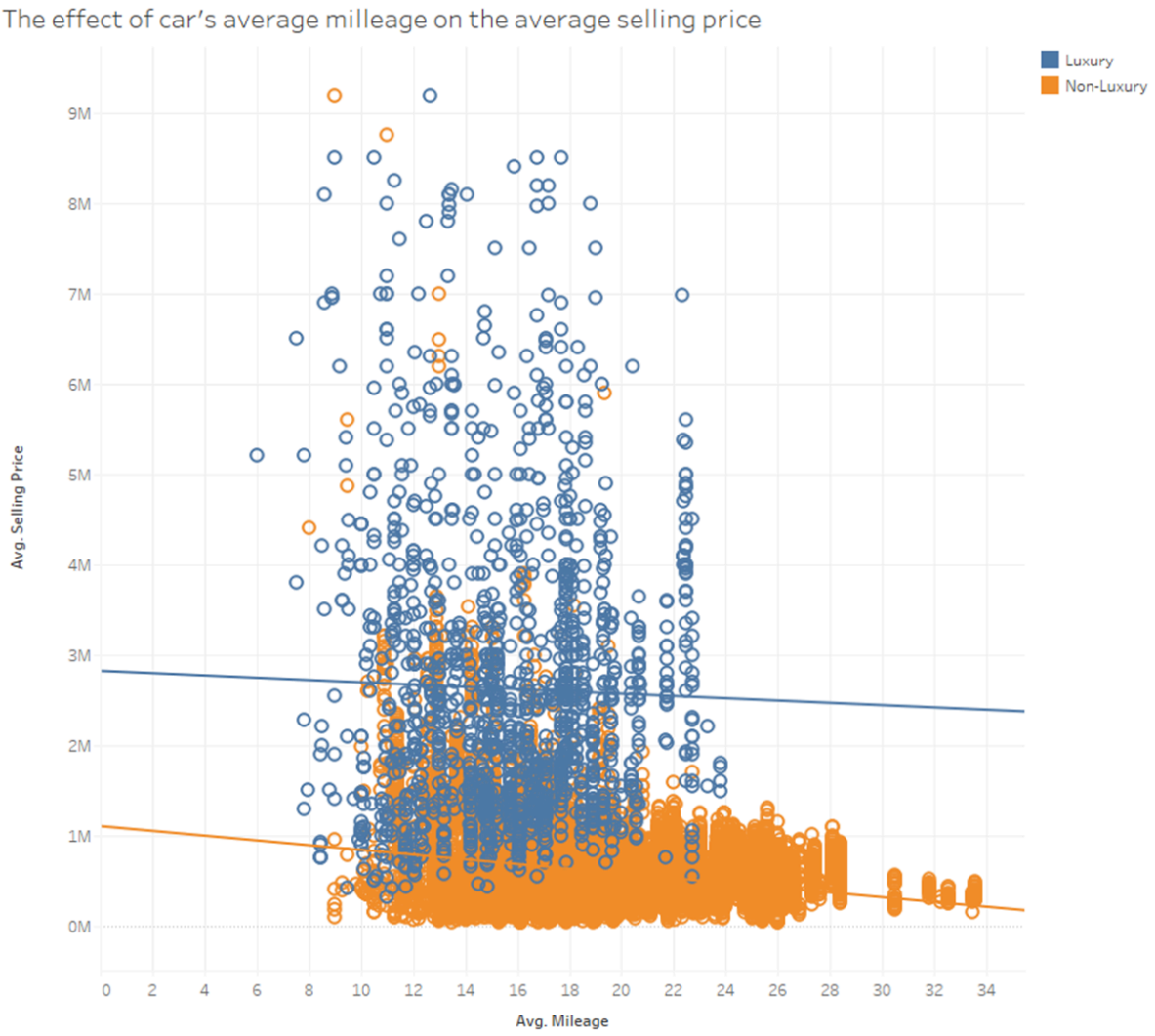
Graph 3. Lorenz Curve with different variables

The Gini coefficient is a measure of inequality, ranging from 0 (perfect equality) to 1 (maximum inequality). It can be interpreted based on its value range to understand the level of inequality in a distribution. A Gini coefficient between 0 and 0.2 indicates very low inequality, where the distribution is nearly equal. Between 0.2 and 0.4, the inequality is considered low to moderate, meaning the distribution is fairly even but with some disparities. A Gini coefficient from 0.4 to 0.6 represents moderate to high inequality, suggesting more significant disparities between groups. Values between 0.6 and 0.8 indicate high inequality, where there are large differences between the groups, and a coefficient from 0.8 to 1.0 reflects very high inequality, meaning the distribution is extremely unequal. This classification helps assess the extent of inequality within a given data set.

In the context of our company's car sales data grouped by continents based on selling price, the Gini coefficient reveals important insights into the distribution of prices in each market. For the American market, the Gini coefficient of 0.42 suggests a moderate level of inequality, meaning that while there is some disparity in car prices, it is not excessively skewed. The Asian market, with a Gini coefficient of 0.328, shows relatively low inequality, indicating a more balanced distribution of car prices where the differences between high and low prices are smaller. On the other hand, the European market has the highest Gini coefficient of 0.49, signifying a relatively high degree of inequality in car prices. This suggests that there are significant price differences, likely driven by the presence of luxury cars or high-end models that dominate the upper end of the price range.

Overall, these Gini coefficients indicate that the European market experiences the most pronounced inequality in car prices, while the Asian market shows the least. The American market falls somewhere in between, with moderate inequality. These findings provide useful insights into market dynamics, highlighting the varying levels of stratification in car prices across different continents. Based on this analysis, our company can adjust its strategy to better suit the geography of each market, identifying which types of cars (luxury or traditional) should be prioritized for sale depending on the region’s price distribution. This strategic approach will help tailor our product offerings to match market preferences and maximize sales potential.

Understanding the effect of the average car’s average mileage on the price



Graph 4. Relationship between the average mileage and average selling price for each brand type

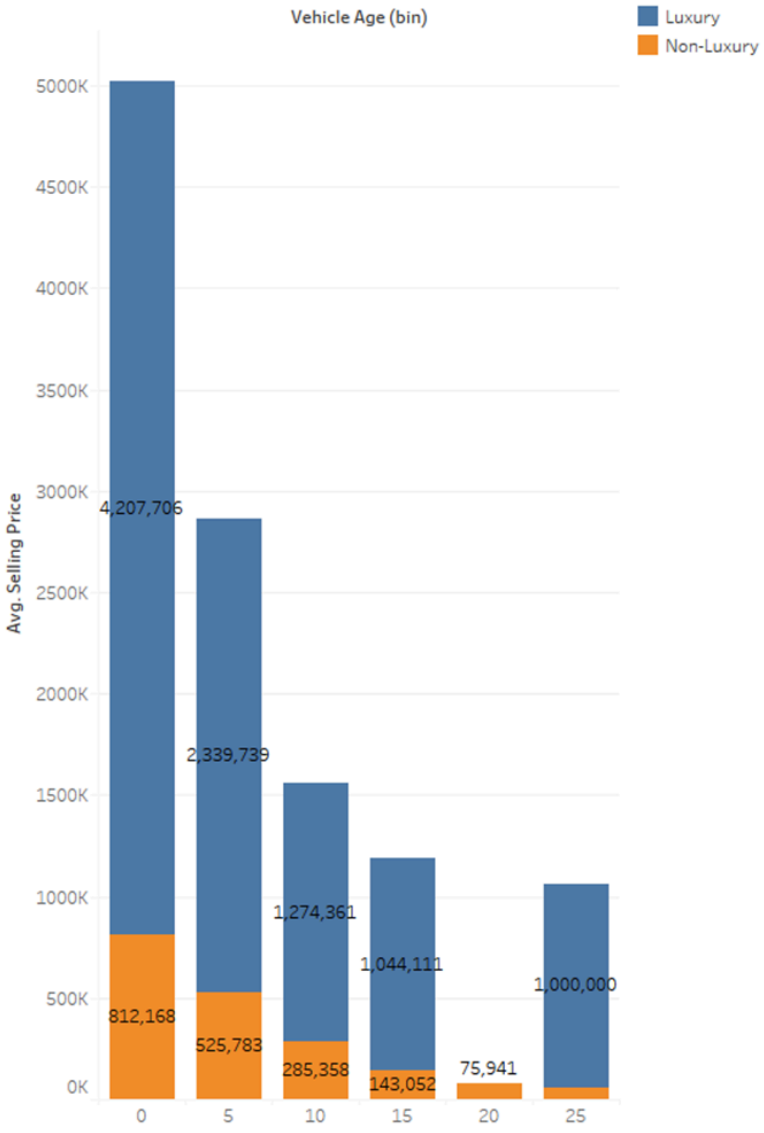
The relationship between a car's mileage and its price is an important factor in the car market. From the visualization, although the starting point of luxury brands is around HUF 2.8 million which is obviously higher than the value of HUF 1.1 million of the non-luxury ones, we can see both trend lines have a negative slope, which means the price of both luxury cars and non-luxury cars decreases as the mileage increases. The reason is higher mileage typically indicates more wear and tear, reducing the vehicle's overall condition and appeal to buyers.

However, the decline is more noticeable in non-luxury brands compared to luxury ones. This could be due to several factors. For non-luxury cars, higher mileage may significantly

impact their value, as they typically don't have the same level of durability or high-end features as luxury brands, so higher mileage tends to negatively affect their resale value more significantly. Additionally, non-luxury cars often have fewer features and lower build quality, meaning that their long-term value is more heavily impacted by high mileage, making them less appealing in the market. As a result, the price drop for non-luxury brands is steeper compared to luxury cars

On the other hand, the decline in luxury cars is less significant compared to non-luxury cars. Luxury cars are often built with superior materials, advanced technology, and high-quality engineering, which contribute to their durability and long lifespan. Even though fuel efficiency improves with higher mileage, the brand's reputation, the quality of the car, and its advanced features still help maintain its value. Buyers may still consider these factors valuable and are often willing to pay a premium for a well-maintained luxury vehicle, even if it has a higher mileage. Therefore, while the price does drop, it's less steep compared to non-luxury brands.

Understanding the effect of the number of vehicle ages on the prices



Graph 5. Relationship between vehicle age and average selling price based on brand type

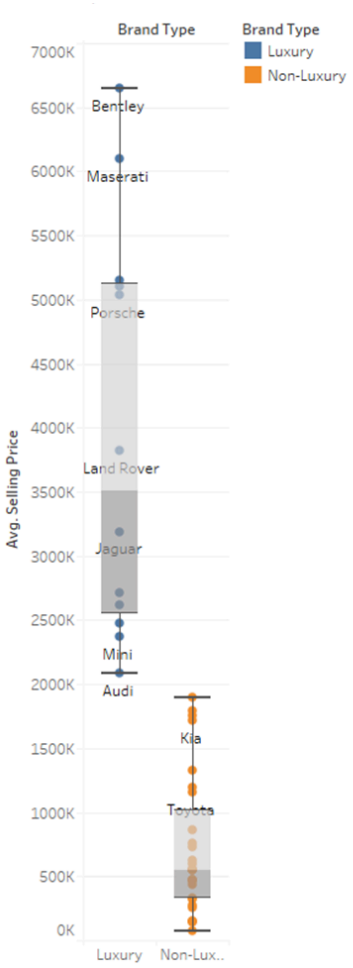
By using a side-by-side bar chart, we can see a significant trend in the depreciation of luxury brands. For luxury cars, there appears to be a critical age when the price experiences a dramatic drop, with an approximate decrease of 1,900,000 units after 5 years. This sharp decline suggests that luxury vehicles, although they maintain their value well during the initial years, face a significant drop in price once they pass this key age. This could be due to various factors

such as market perception, the high cost of maintenance, or the introduction of newer models with advanced features. All of which may contribute to a faster depreciation rate in the later years of a luxury vehicle's life cycle. Also as I mentioned above, they do not have a high level of durability or high-end features, therefore the age affects their value directly.

In contrast, the depreciation of non-luxury brands follows a much more consistent trend. When observing the data for non-luxury models, we can see that the price decreases steadily by about HUF 250,000 for the first 5 years; HUF 200,000 units for the next 5 years; and HUF 150,000 in the following 5 years, which shows no significant change in the depreciation rate as the vehicle ages. This consistent decline suggests that they have a negative linear relationship which means economy cars tend to lose value at a stable rate over time. Unlike luxury cars, non-luxury vehicles do not experience sharp drops in price after a certain age, indicating that they depreciate steadily rather than drastically.

In conclusion, while luxury cars may depreciate slower in their initial years, they experience a sharp depreciation after reaching a certain age, whereas non-luxury cars exhibit a more predictable and gradual depreciation pattern over time.

Understanding the effect of brand reputation on price

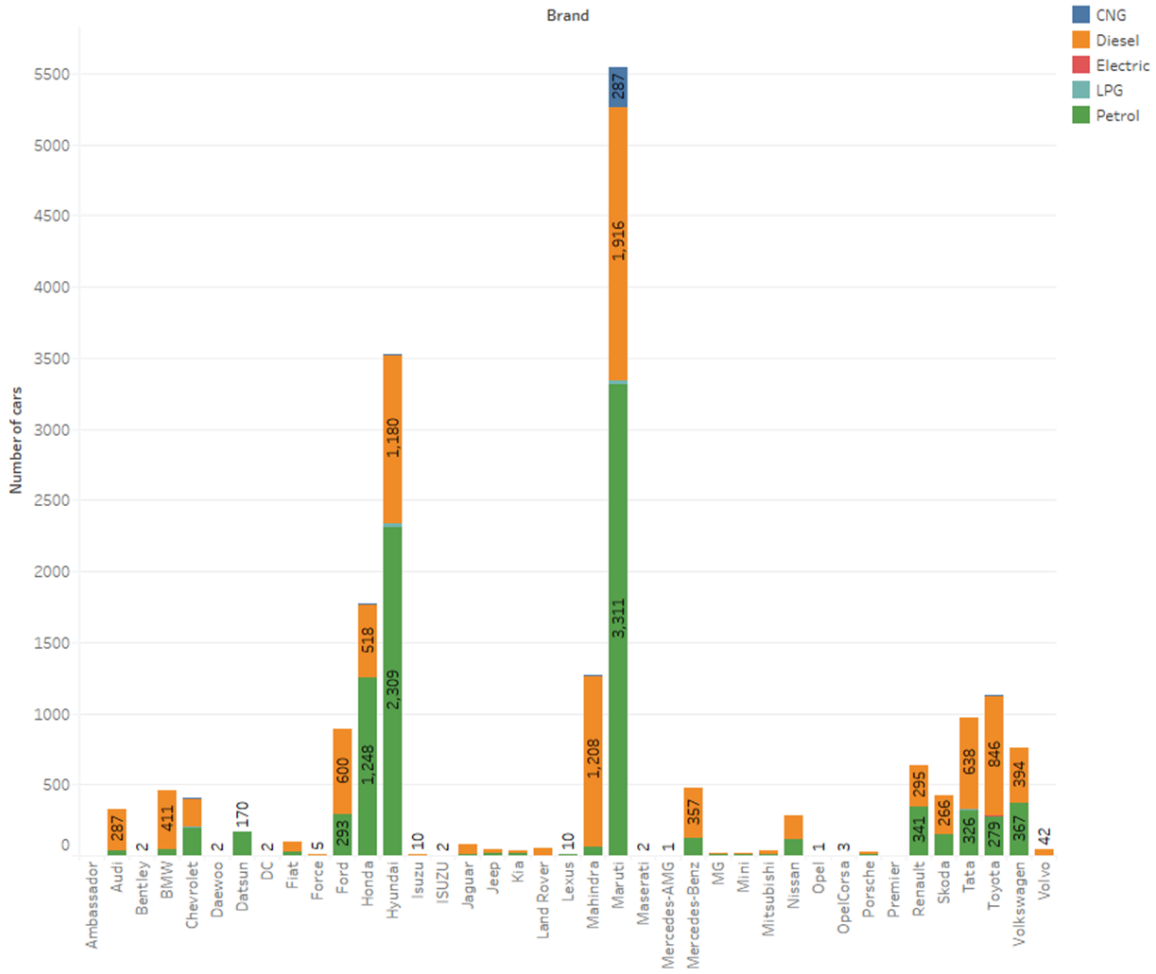


In the third visualization, we can see that the average selling price for luxury brands, such as Bentley, Maserati, and Porsche, is significantly higher compared to non-luxury brands.

The chart highlights the economic distinction between luxury and non-luxury car markets. Buyers of luxury cars care less about price and more about things like brand reputation value and performance. Non-luxury brands, on the other hand, target cost-conscious consumers, with affordability being a primary factor in purchasing decisions.

Graph 6. Relationship of the reputation of different brands and their average selling price

Understanding the relationship between brand and fuel type



Graph 7. Relationship between Brand and Fuel Type

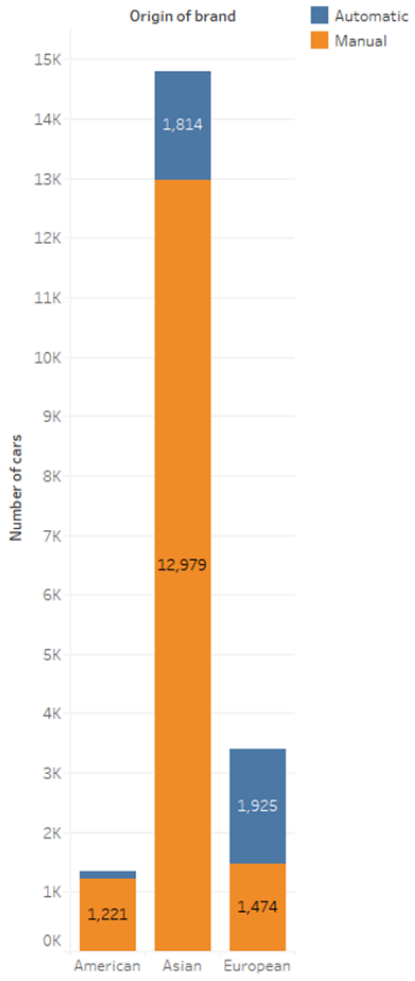
From the visualization, we can see that the most common fuel type of both luxury brand cars and non-luxury brand cars are petrol and diesel. But there is a significant difference between the behavior of each brand.

For non-luxury brands like Maruti, Hyundai, Honda, and Toyota, petrol is the dominant fuel type. The reason probably is because petrol engines are typically quieter and smoother than diesel engines, making them ideal for city driving and the maintenance and servicing of petrol engines are generally less expensive.

On the other hand, luxury brands tend to favor diesel cars, with most of their vehicles running on this fuel type. The reason is because diesel engines are more fuel-efficient, especially on highways and for long-distance driving. They also generally offer more power, meaning they may be a better choice if you need a larger vehicle or expect to regularly tow and/or carry a heavy load. (source: https://www.traffic.gov.scot/greener-travel/what-cars-best-for-environment)

This shift toward diesel in luxury brands can have a considerable impact on the environment. Diesel engines are typically more fuel-efficient than petrol engines, but they also produce higher levels of nitrogen oxide, which are harmful to both air quality and public health. Diesel emissions contribute to smog formation and respiratory issues, which are a growing concern in urban areas. While luxury brands may argue that diesel's higher fuel efficiency justifies their use, the long-term environmental consequences of increased diesel use cannot be ignored. Therefore, while luxury cars are more likely to use petrol for performance and efficiency, their preference for diesel fuel type still poses environmental challenges, especially in cities already grappling with air pollution.

Understanding the relationship between origin of brand and transmission type



Graph 8. Relationship between origin of brand and transmission type

The relationship between the origin of the brand and transmission type reveals interesting insights about regional preferences and trends. In Asia, manual transmission cars are significantly more popular, with 12,979 manual cars compared to just 1,814 automatic cars. This preference could be influenced by factors such as cost, fuel efficiency. Additionally, in many Asian countries, learning to drive a manual transmission is seen as a rite of passage. It lays the groundwork for driving automatic cars. Driving schools often teach manual cars, and manual licenses are more common.

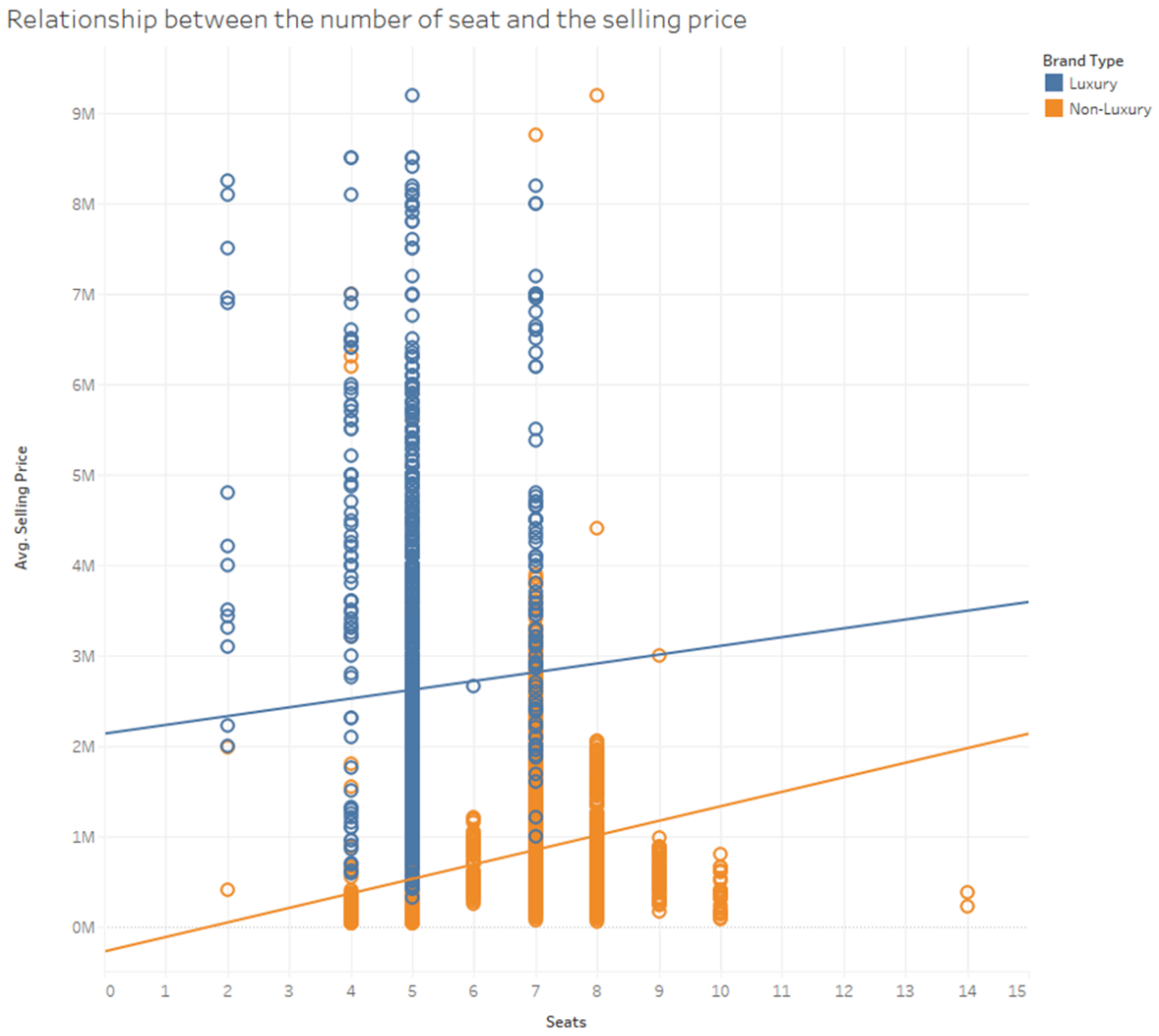
In the American market, the data shows that manual cars are more prevalent than automatics. In reality, this idea is actually quite unusual because America has been leading the

world in the car industry for many years . However, if we are considering specific segments of the market or certain trends, it may be due to their lower cost, better fuel efficiency, and preference among drivers for control in specific conditions. (source: theguardian.com)

Europe, on the other hand, shows a preference for automatic transmissions, but the difference is not as pronounced. This could be driven by the growing trend towards convenience and technology in European markets, where automatic cars offer smoother driving experiences in busy cities with heavy traffic.

Overall, the data shows that while manual transmissions remain popular in Asia and are still significant in America, Europe is leading the shift toward automatic transmissions. This shift reflects broader trends in regional driving habits, vehicle preferences, and technological advancements that prioritize ease of use and fuel efficiency.

Understanding how the number of seats affect the selling price



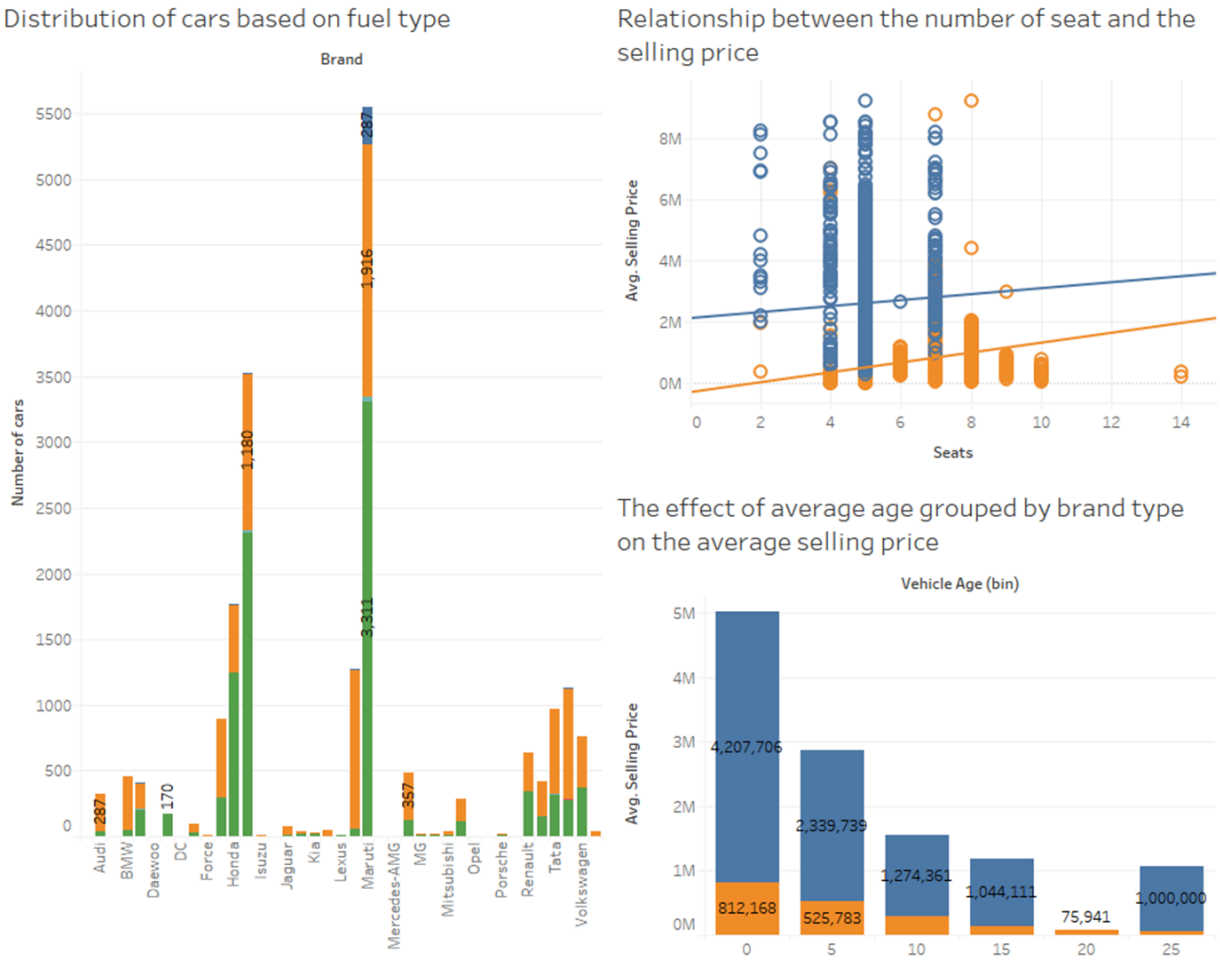
Graph 9. Relationship between number of seats and price

The graph reveals a relationship between the number of seats in a vehicle and its price, but this varies significantly between luxury and non-luxury vehicles. For non-luxury vehicles, there is a clear upward trend, with prices increasing as seat count rises. For instance, vehicles with 5 seats in the non-luxury category are priced around 1 million, while those with 8 seats rise to about 2 million or more. This demonstrates that larger vehicles, designed to accommodate more passengers, cost more due to higher production and design costs, making the number of seats a key driver of price.

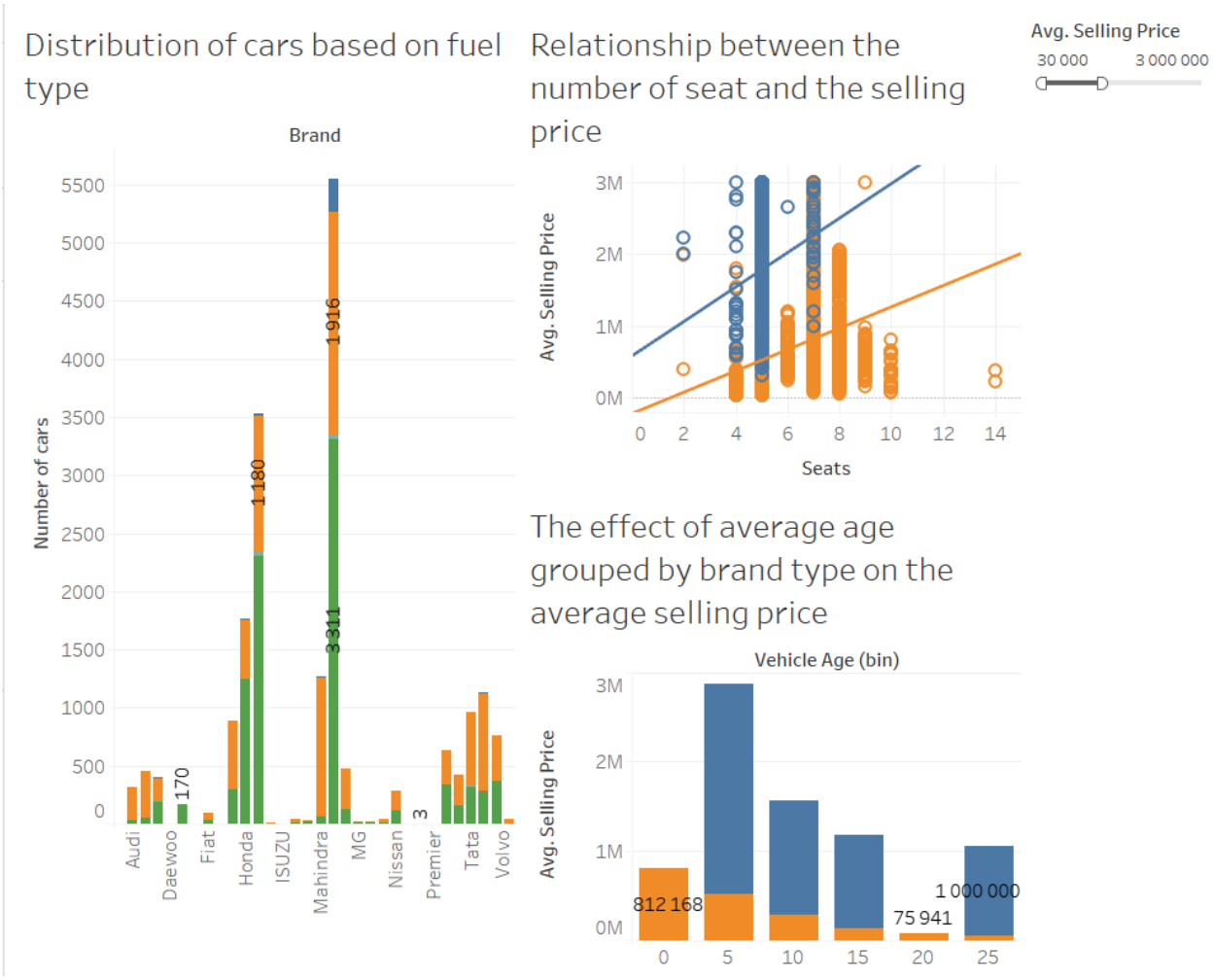
In contrast, luxury vehicles show a much weaker link between seats and price. Luxury vehicles with 4 seats are priced much higher, averaging between 3 million and 7 million, while those with 5 seats climb even further to over 8 million. However, the price differences here are less about the seat count and more about brand prestige, high-end materials, and advanced technology. For example, luxury vehicles with 4 seats are still priced far above non-luxury vehicles with 7 or more seats, indicating that factors beyond seat count dominate pricing for luxury brands.

In summary, while non-luxury vehicles clearly show a direct connection between seat count and price, with prices doubling as seats increase from 5 to 8, luxury vehicles rely more on brand and features. The higher prices for luxury vehicles across all seat ranges, often exceeding 7 million even for smaller cars, highlight the secondary role of seating in this category.

Dashboard: Conclusion



Graph 10. Dashboard



Graph 11. Dashboard with an interactive filter (price)

This dashboard helps customers choose cars within their budget by analyzing how price relates to fuel type, seating capacity, and vehicle age. The price filter is fully customizable, allowing users to adjust the range to suit their specific needs. The minimum price that can be set is 30,000 HUF and the maximum is 9,200,000 HUF. For example, in this dashboard, the average price has been filtered to display cars priced between 30,000 and 3,000,000, showcasing a range of affordable and mid-priced options.

The first graph shows the distribution of cars by fuel type and brand. Brands like Hyundai and Maruti have a high number of options across petrol and diesel categories, making them ideal for budget-conscious buyers looking for variety in fuel types within the selected price range.

The second graph highlights the relationship between seats and price. Non-luxury cars show an upward trend, with prices increasing from around HUF 500,000 for 4-seaters to over HUF 2 million for 8-seaters. Luxury cars, on the other hand, start at HUF 1 million for 4 seats and exceed HUF 3 million for larger models, emphasizing the premium pricing of luxury brands even within the filtered range.

The third graph examines how vehicle age affects price. Cars less than 5 years old are the most expensive, with luxury models averaging HUF 2.3 million and non-luxury models at HUF 812,000. As cars age, prices drop significantly, with non-luxury cars over 15 years old averaging HUF 75,941. This helps buyers consider the trade-off between affordability and the reliability of newer vehicles.

With the customizable price filter, buyers can focus on cars that match their budget while analyzing how fuel type, seating capacity, and age impact their choices. This flexibility ensures that the dashboard is a practical tool for finding cars tailored to individual preferences and financial constraints.