Name: Pratik Rakesh Bhutada
Student_id: 24840332

1. Data Understanding and Exploration

1.1 Meaning and Type of Features:

The dataset 'Adverts.csv' contains information about cars sold through AutoTrader, a major UK automotive marketplace, up to the year 2021. This dataset, has 402,005 entries and 12 features, including: 'public_reference', 'mileage', 'reg_code', 'standard_colour', 'standard_make', 'standard_model', 'vehicle_condition', 'year_of_registration', 'price', 'body_type', 'crossover_car_and_van', and 'fuel_type'. These attributes can be further classified into qualitative and quantitative categories.

The "vehicle_condition" feature is a categorical variable with two values: "NEW" and "USED." This binary classification is crucial for understanding vehicle value, as "NEW" vehicles generally command higher prices. The pie chart reveals a significant imbalance in the distribution of vehicle condition, with "Used" vehicles constituting the vast majority at 92.2%. In contrast, "New" vehicles account for only 7.8% of the dataset. This distribution likely reflects the dynamics of the used car market, where the demand and availability of used vehicles typically exceed those of new vehicles in the United Kingdom. (Refer to Fig 1.1)

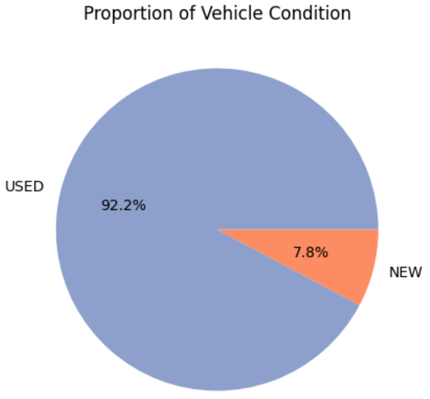


Fig 1.1

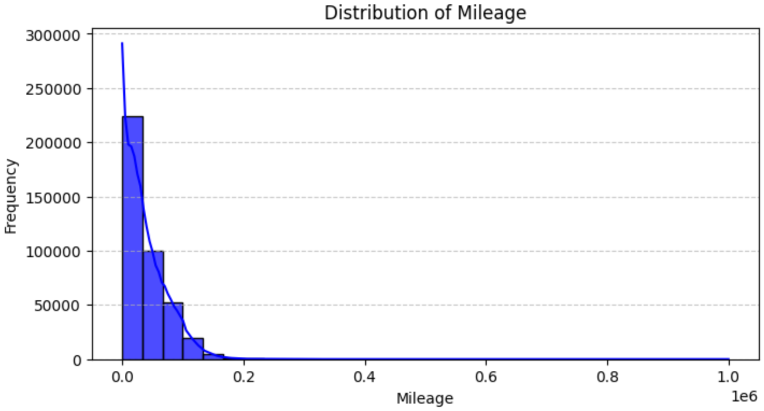


Fig 1.2

"Mileage," a continuous numerical variable, represents the distance a vehicle has traveled. This feature is a significant indicator of vehicle wear and tear and plays a crucial role in determining its value. Higher mileage typically corresponds to lower prices due to increased wear and tear. A moderate negative correlation between mileage and price supports this observation. The distribution of the Mileage feature shows a clear right-skewed pattern. This means that the majority of vehicles in the dataset have relatively low mileage, while smaller proportions have high mileage. This distribution likely reflects real-world trends, as vehicles tend to depreciate over time, and many people prefer to purchase vehicles with lower mileage due to their higher quality and longer lifespan. (Refer to Fig 1.2)

"Year_of_registration," a numerical variable, represents the year in which the vehicle was registered. This feature is highly influential in determining vehicle value, as newer vehicles generally depreciate less rapidly and are equipped with more advanced features. A moderate positive correlation between year of registration and price reflects this trend. However, attention must be given to potential outliers, such as very old or unusually recent registration years. The year of registration distribution is multimodal with a right-skewed tail, showing several peaks at specific years and more vehicles registered in recent years. This suggests a trend of increasing newer vehicles in the dataset. The peaks may reflect fluctuations in vehicle sales or registration trends, possibly due to economic factors, new model releases, or government incentives. Additionally, the distribution may be influenced by data collection biases, with more recent data showing higher numbers of newer vehicles. (Refer to Fig 1.3)

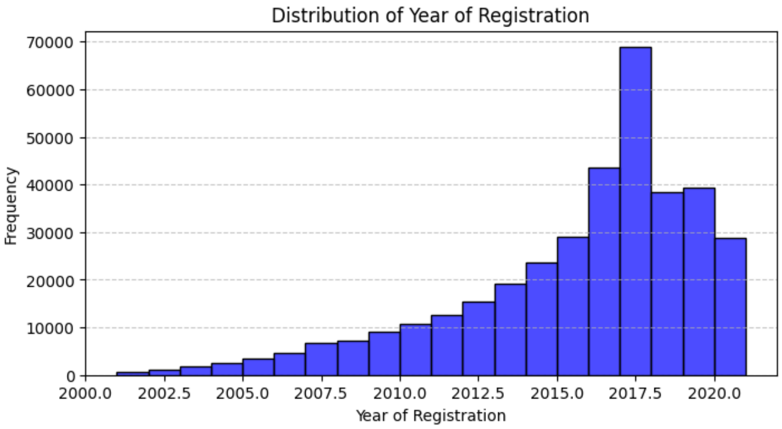


Fig 1.3

1.2. Analysis of Distributions:

Price:

The price distribution is right-skewed, meaning most vehicles are priced low, while fewer are priced high. This matches real-world trends where vehicles lose value over time, and buyers often prefer affordable options. There's a sharp rise in the number of low-priced vehicles, showing they dominate the dataset. The long tail of higher prices indicates some expensive vehicles are still present, but they are much fewer in number. (Refer to Fig 1.4)

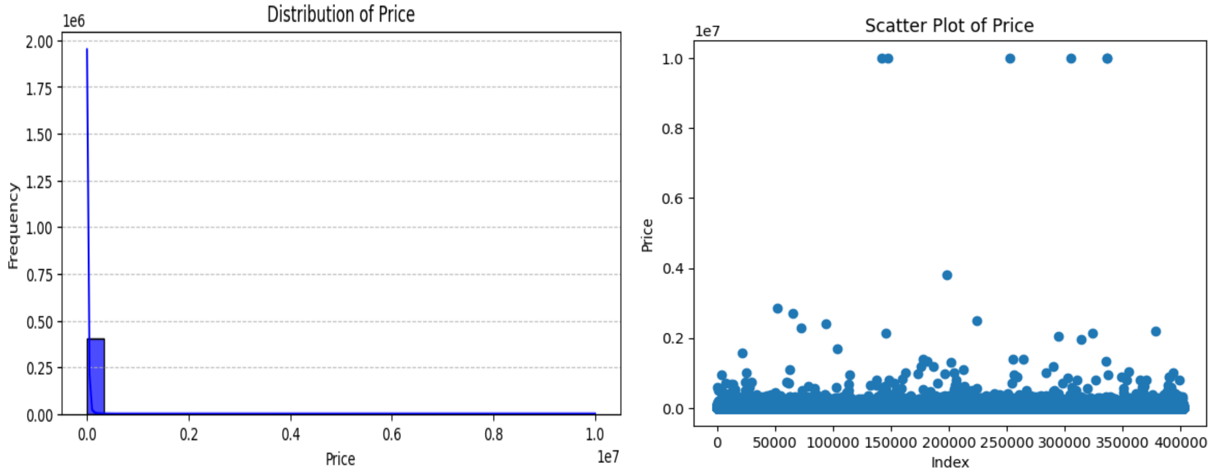


Fig 1.4

Body type:

The body type distribution is imbalanced, with Hatchbacks as the most common, followed by SUVs and Saloons. Other body types, like Estates, Coupes, and MPVs, are much less frequent. This likely reflects real-world trends, as Hatchbacks and SUVs are popular for their practicality, and Saloons are favored for comfort and style. The less common body types likely represent niche markets with smaller demand. (Refer Fig 1.5)

Standard colour:

The car color distribution is dominated by "Black," which is the most common color by far. "White" and "Grey" are also popular, though less so than "Black." Other colors like "Blue," "Silver," and "Red" are much less frequent. This pattern likely mirrors real-world preferences, where black, white, and grey are favored for their appeal, practicality, and higher resale value. (Refer Fig 1.5)

Standard make:

The dataset shows that BMW is the most common car make, followed by Audi and Volkswagen, with other makes like Vauxhall, Mercedes-Benz, and Toyota being less frequent. BMW’s dominance indicates it is the most popular make in the dataset, though other makes like Audi and Volkswagen also appear fairly often. This pattern likely reflects market popularity, driven by factors like brand reputation and consumer choice. Additionally, the data collection process might have influenced the distribution, capturing trends or preferences from specific regions or times. (Refer Fig 1.5)

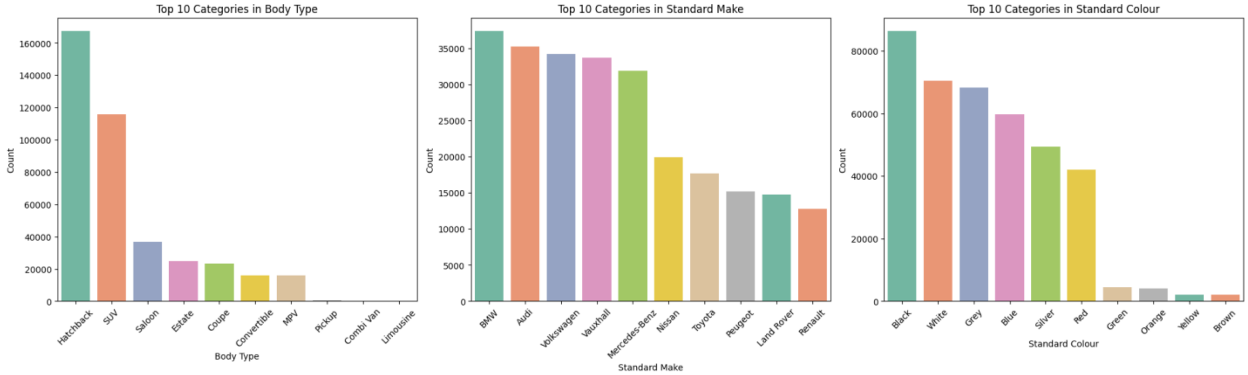


Fig 1.5

Standard model:
The distribution of car models reveals that "Golf" is the most popular, with around 12,000 vehicles in the dataset. "Corsa" and "C Class" come next, each with about 10,500 and 8,500 vehicles, respectively. Other models like "3 Series," "Polo," "Qashqai," "1 Series," "Astra," "Hatch," and "A Class" are less common, with vehicle counts ranging from 6,500 to 7,500. This pattern likely reflects the popularity and market presence of these models, with more well-known and widely sold models appearing more frequently in the data. (Refer to Fig 1.6)

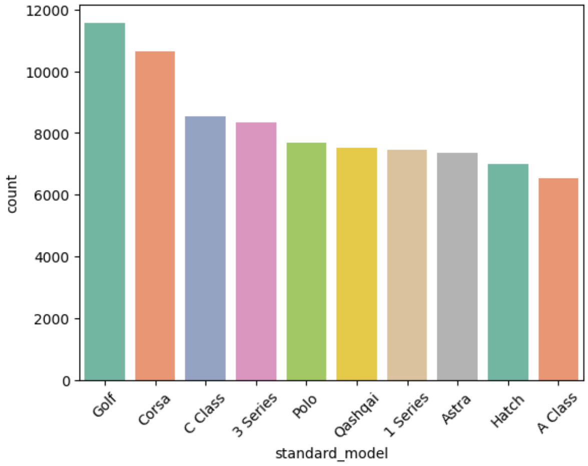


Fig 1.6

2. Data Pre-Processing

2.1 Data Cleaning

The Data Cleaning process was to ensure the dataset's quality and usability for subsequent analysis and modeling. Missing values, if left unaddressed, could skew results, or lead to invalid conclusions. For example, imputing year_of_registration and reg_code for "NEW" vehicles ensures that such entries remain complete and consistent with their expected characteristics. Similarly, imputing mileage based on average usage patterns allows for more realistic representations of "USED" vehicles, preventing biases from incomplete data. Removing outliers like implausible year_of_registration values avoids distortions in trends and prevents models from being misled by erroneous data. These steps were guided by logical assumptions and data-driven techniques, ensuring that the cleaned dataset accurately reflects the underlying phenomena it represents. By addressing these issues, the dataset becomes more reliable, interpretable, and relevant, ultimately improving the outcomes of any subsequent analysis or predictive modeling.

2.2 Feature Engineering

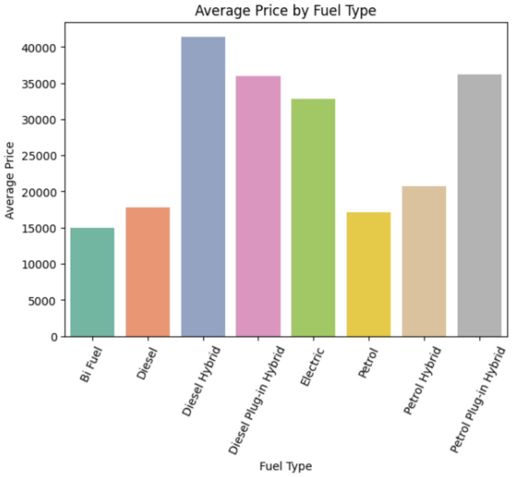
model_name (Combining standard_make and standard_model):
By combining the vehicle's make and model into one feature, we create a unique label for each car, making it much easier to group and analyze the data. This helps us spot trends specific to certain vehicle types, like the most common colors or how mileage varies across different models. It ensures that when we look at the data, we're focusing on the unique combination of make and model, which is key to understanding patterns and relationships.

Imputation Using Most Frequent Color for standard_colour:
When the color of a vehicle is missing, we want to fill that gap with something that makes sense. Using the most common color for a particular model is a smart way to do this, as it keeps the dataset consistent with the typical characteristics of that model. This ensures that the missing values don't stand out as anomalies and helps maintain the overall quality of the data. If there's no clear "most frequent" color, we use "Others" to cover any rare cases, making sure there are no holes in the dataset.

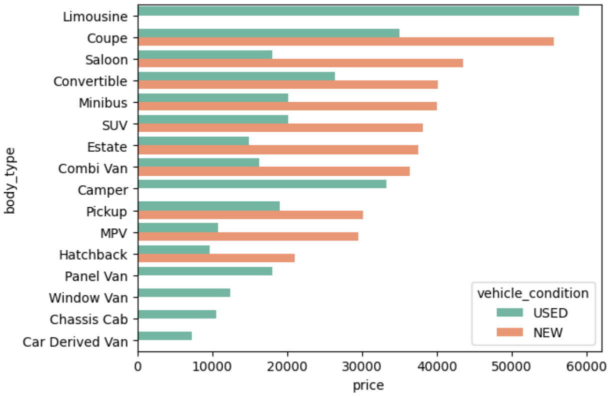
average_mileage_per_year (Calculated as mileage/Age):
This feature helps us better understand how much a car is actually being driven by taking the total mileage and dividing it by the vehicle's age. This gives us a clearer picture of whether a car is getting more or less use than we would expect for its age. It’s a useful metric because it helps us spot trends in how cars are being used, which could impact things like wear and tear, maintenance needs, or even their resale value down the road.

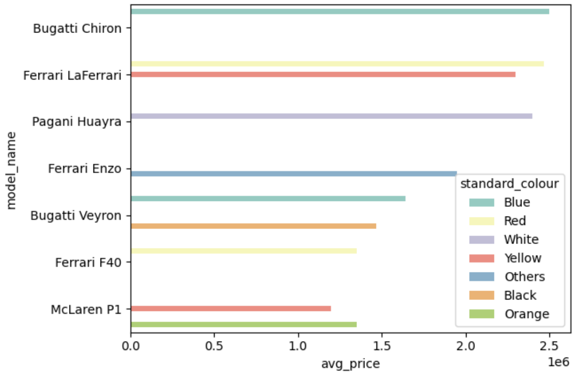
2.3 Subsetting

Insight 1:
This analysis reveals significant price variations across different fuel types. Diesel Hybrid and Petrol Plug-in Hybrid vehicles command the highest average prices at £41,400 and £36,251 respectively. In contrast, Bi Fuel vehicles have the lowest average price at £14,929. To capitalize on this, a strategic focus on promoting Electric Vehicles (EVs) is warranted. Leveraging government incentives, emphasizing long-term cost savings, and implementing targeted advertising campaigns highlighting the advanced features and environmental benefits of EVs can attract a wider audience. Further analysis should consider factors like vehicle age, mileage, and brand to gain a deeper understanding of price variations and inform more refined marketing strategies. (Refer to graph below)



Insight 2:
Analysis of the bar plot reveals significant price variations across body types and vehicle conditions. Limousines, even with high mileage, command premium prices, indicating that prestige and brand value can outweigh mileage concerns for certain body types. In contrast, Car Derived Vans and Window Vans, despite being used, have lower average prices, suggesting lower demand for these types of vehicles. For NEW vehicles, Coupes, Saloons, and Convertibles have the highest average prices, reflecting their luxury and advanced features. To capitalize on these insights, a multi-pronged approach is recommended. For high-end NEW vehicles, focus on marketing to premium customers by emphasizing luxury features, advanced technology, and exclusive services. For used vehicles with higher mileage, highlight their affordability and reliability, while offering options like certified pre-owned programs to increase buyer confidence. Additionally, segmenting the target audience based on factors like age, income, and lifestyle will enable more effective marketing and sales strategies. (Refer to graph below)

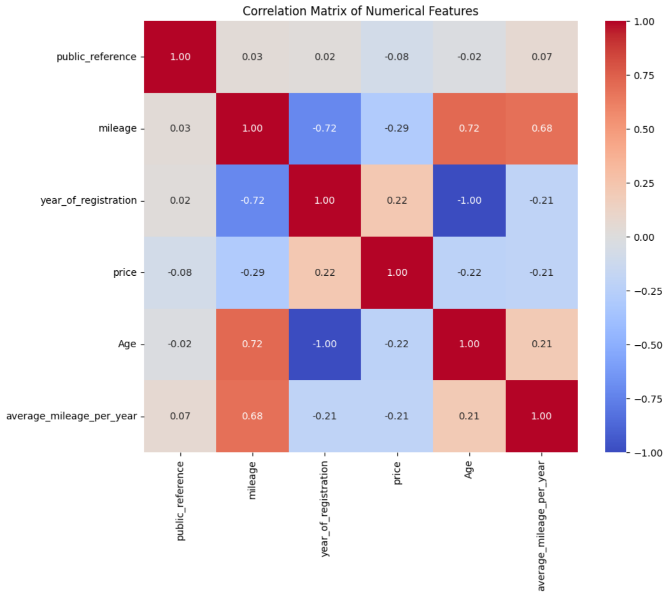




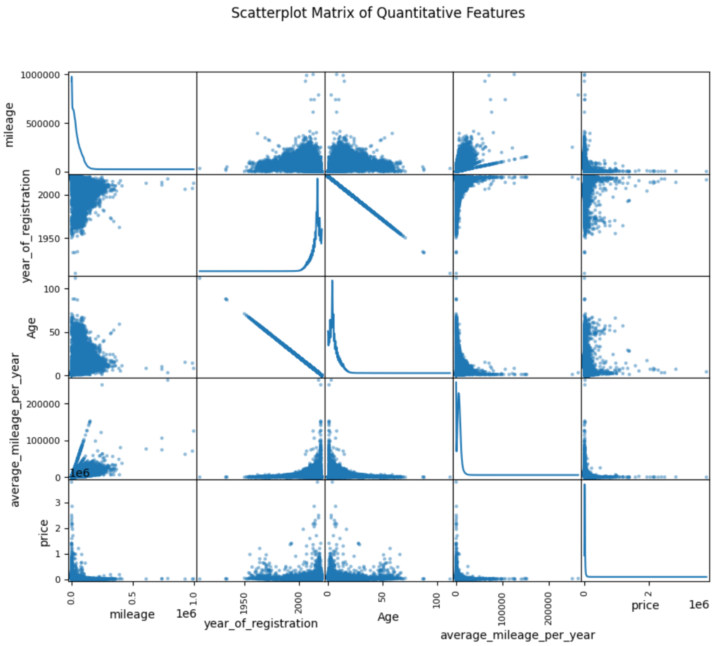
Insight 3:
The bar plot reveals two distinct segments in the automobile market: premium and budget-friendly vehicles. Premium models, characterized by low mileage, vibrant colors like yellow and blue, and exceptionally high prices, are primarily targeted towards luxury enthusiasts and collectors. These vehicles, often with limited production runs, command premium prices due to their exclusivity and investment potential. In contrast, budget-friendly vehicles, with high mileage and more affordable price points, cater to cost-conscious buyers who prioritize functionality and value. To capitalize on these insights, a differentiated approach is crucial. For premium models, a strategy focused on exclusivity, prestige, and brand building is essential. Collaborating with luxury brands, hosting exclusive events, and emphasizing the investment potential of these vehicles through targeted marketing campaigns can attract high-net-worth individuals and collectors. For budget-friendly models, a focus on affordability, reliability, and value is key. Offering tailored financing options, maintenance packages, and highlighting unique features like specific color options or utility features can attract a wider range of cost-conscious buyers. Analyzing the competitive landscape within each segment and understanding the impact of mileage on pricing will further refine these strategies and enable businesses to effectively target their respective markets. (Refer to 2nd plot above)

3. Analysis of Associations and Group Differences

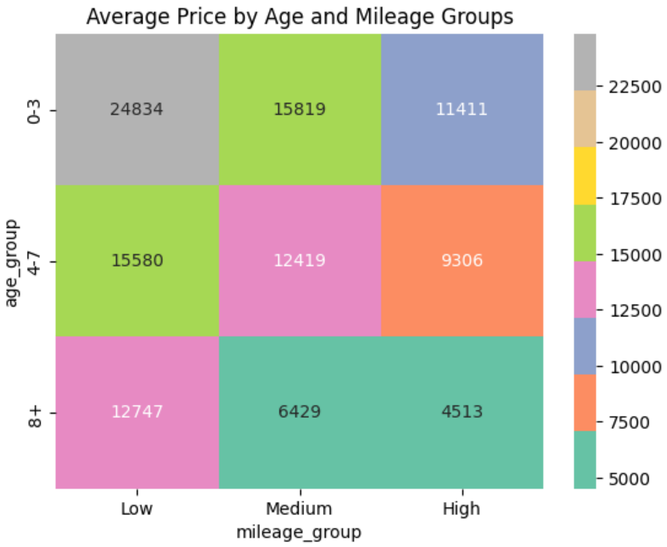
3.1 Quantitative – Quantitative



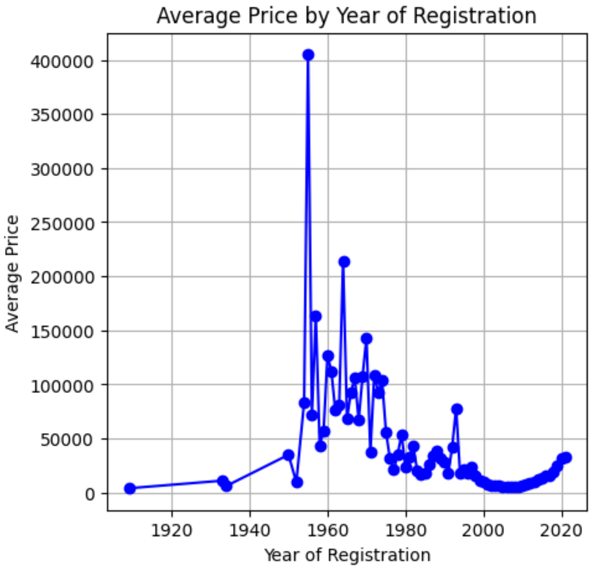
The correlation matrix highlights key relationships between numerical features. It shows a strong negative correlation between 'mileage' and 'year_of_registration', as newer cars typically have lower mileage. There is also a moderate negative correlation between 'mileage' and 'price', with higher mileage cars generally being cheaper. 'Age' and 'mileage' exhibit a strong positive correlation, as older cars tend to have higher mileage. A moderate negative correlation exists between 'age' and 'price', as older cars are usually priced lower. Other feature pairs show weak or no correlation, indicating that factors like 'public_reference' have little influence on others. (Refer to graph above)



The analysis highlights some interesting patterns in how car features relate to price. Cars with higher mileage tend to cost less, which makes sense since more mileage usually means more wear and tear. Similarly, older cars are generally cheaper due to depreciation over time. Also, older cars also tend to have higher mileage. Cars that rack up more miles each year are usually less expensive, likely because they experience faster wear. When it comes to registration year, newer cars typically cost more, but there’s an exception for some older models, like vintage cars, which can hold significant value. These findings give us a clearer picture of how different factors affect car prices. (Refer to graph above)

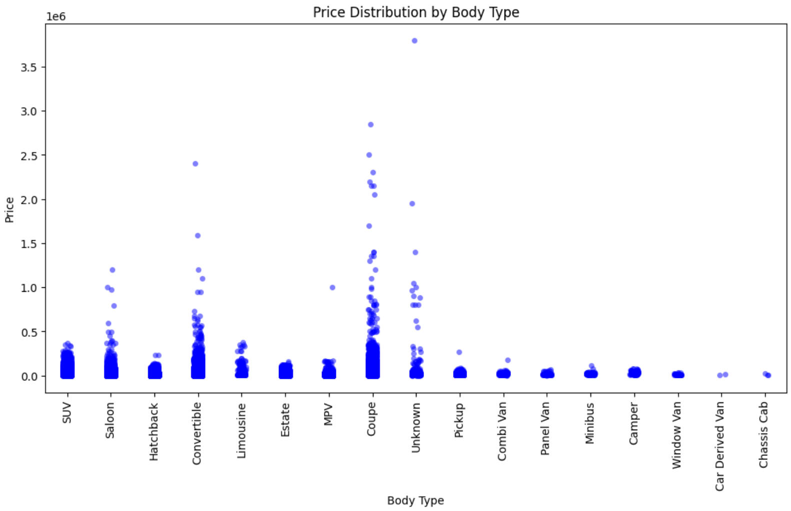


The heatmap highlights how car age and mileage influence average prices. Prices decrease as cars age and mileage increases, with older cars showing a stronger price drop between "Low" and "High" mileage groups. This suggests that mileage impacts the condition of older cars more significantly. Depreciation, wear and tear, and market demand likely explain these trends. (Refer to graph above)

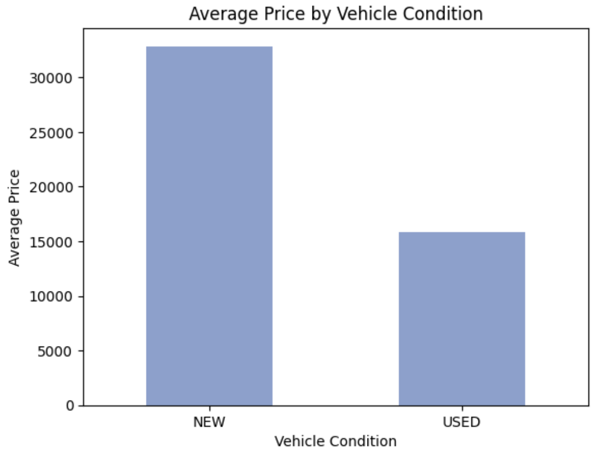


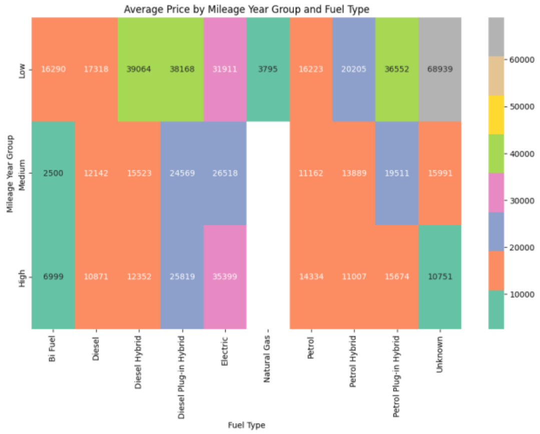
The graph shows a complex relationship between car prices and year of registration. Vintage cars from the 1920s and 1930s command high prices due to rarity and collectability. In contrast, cars from the mid-20th century show lower prices, reflecting depreciation. From the late 1990s onward, prices rise with newer registration years, driven by advanced features and market demand. While this trend highlights key influences like vintage value and technology, distribution and other factors like make and model warrant further exploration for a comprehensive understanding. (Refer to graph above)

3.2 Quantitative – Categorical



The scatter plot shows significant price variation within each body type, indicating that factors beyond body type, such as brand, model, age, mileage, and condition, influence vehicle prices. Outliers are present, particularly for Limousines and Convertibles, which have higher prices. While prices vary across body types, luxury types like Limousines and Convertibles generally command higher prices, while utilitarian body types like Panel Vans and Chassis Cabs tend to be priced lower, reflecting their functional nature. (Refer to graph above)

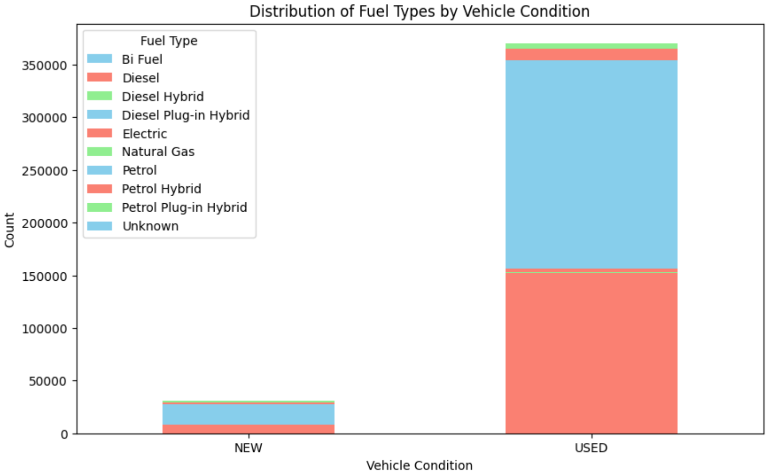




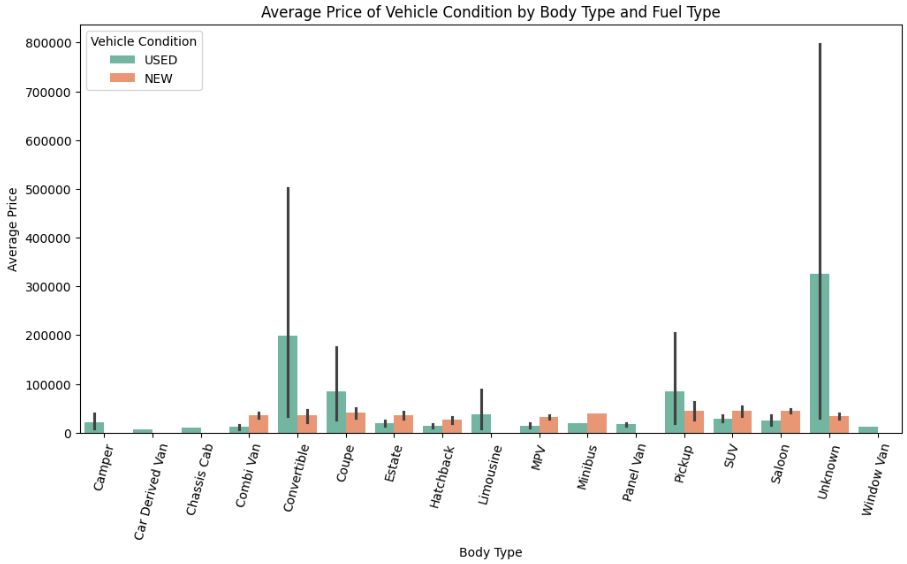
The graph highlights a clear distinction in average vehicle prices based on condition, with "NEW" vehicles priced significantly higher than "USED" ones. This stark difference reflects the impact of market value depreciation, as vehicles lose value once used. It also underscores the price sensitivity to condition, with buyers willing to pay a premium for new vehicles due to factors like minimal wear and tear, lower risk of mechanical issues, and higher perceived reliability. This visualization emphasizes the importance of condition as a key determinant of vehicle pricing. (Refer to graph above)

The plot paints a clear picture of how a few key factors—mileage, fuel type, and other features—work together to influence vehicle prices. Cars in the "Low" Mileage Year Group, which are likely newer and less driven, tend to have higher prices, while those with more miles ("High" or "Medium") are generally more budget-friendly. Fuel type also plays a big role, with Electric and Natural Gas vehicles standing out as more expensive, thanks to their efficiency and growing popularity. The colors in the graph bring it all together, showing how these factors combine to reflect market trends and buyer preferences. (Refer to 2nd graph above)

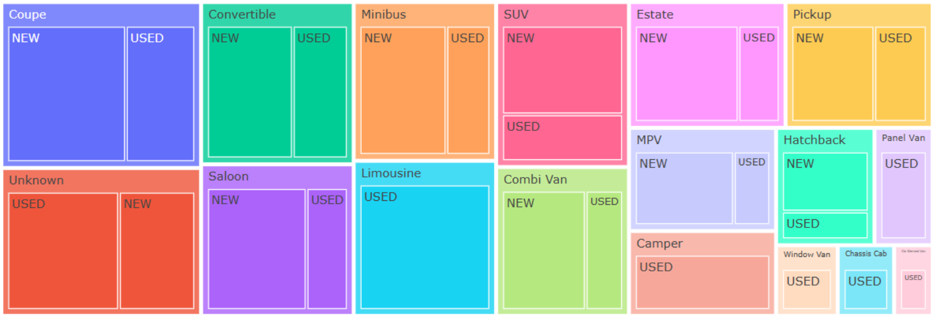
3.3 Categorical – Categorical



The data shows that petrol vehicles dominate the market, both for new and used cars, with diesel taking second place. Electric, plug-in hybrids, and natural gas vehicles are less common, and bi-fuel vehicles are the least frequent across both categories. New cars are mostly petrol, while used cars tend to be diesel. From a business perspective, the dominance of petrol and diesel highlights a stable market for traditional vehicles, but the growing presence of electric and hybrid cars signals a shift towards eco-friendly options. This trend could present new opportunities for businesses to invest in green technologies, expand electric vehicle offerings, and meet changing consumer preferences. Understanding the shift in fuel type preferences could help businesses adjust marketing strategies and inventory to align with future demand, particularly as consumers increasingly prioritize fuel efficiency and sustainability. (Refer to graph above)



The graph reveals how body type, vehicle condition, and fuel type impact vehicle prices. Luxury body types like Limousine and Convertible tend to be pricier, while new vehicles always cost more than used ones. Fuel type also plays a role, with eco-friendly options likely influencing higher prices. For businesses, understanding these factors helps in pricing strategy. Luxury models and newer cars can be marketed at a premium, while expanding electric and hybrid offerings can meet growing demand for sustainable vehicles. (Refer to graph above)



Treemap of Body Type and Vehicle Condition with Average Price

The treemap visually highlights how body type and vehicle condition influence average vehicle prices. Larger rectangles represent body types like Limousine and Convertible, which tend to have higher prices. New vehicles are consistently more expensive than used ones within each body type. The color coding may indicate fuel types or other variables. If you click on different body_type, you will be able to get average price from the treemap. For businesses, this visualization emphasizes the importance of focusing on luxury body types and new vehicles for premium pricing strategies, while considering the impact of fuel type on consumer preferences. (Refer to graph above)

References:

https://pandas.pydata.org/ (Accessed: 25 December 2024).

https://www.autotrader.co.uk/ (Accessed: 25 December 2024).

https://matplotlib.org/ (Accessed: 25 December 2024).

'Vehicle registration plates of the United Kingdom' (2021). Available at: https://en.wikipedia.org/wiki/Vehicle_registration_plates_of_the_United_Kingdom(Accessed: 20 December 2024).