1. Data Understanding and Exploration

1.1. Meaning and Type of Features (3-4)

In this section, reg_code, year_of_registration, mileage, and vehicle_condition features are going to be described before the data cleaning process. For each feature, a brief description, datatype, and their relationship with other given features is explained.

Two features of reg_code and year_of_registration are related to each other based on the information that is provided in the Age Identifiers table. By having one of them, the other one can be guessed and this process is possible by defining a dictionary and applying mapping to it.

1.1.1. reg_code

This numerical feature represents a specific letter or number relevant to a range of 6 months of a specific year in the UK car's plates. The Dtype attribute of this feature is Object, which indicates the values are string. It's noticed that numbers are also stored as string values in this column. There are 31,857 null values for this feature in the original data set. The valid values of this feature, for the given timeframes are:
a) 1963 – 1982: A, B, C, D, E, F, G, H, J, K, L, M, N, P, R, S, T, V, X, Y
b) 1983 - 2001: A, B, C, D, E, F, G, H, J, K, L, M, N, P, R, S, T, V, X, Y
c) 2001 – 2021: [2, 21] and [51,71]

1.1.2. year_of_registration

This numerical feature shouldn't be mistaken with the car's age because a car might be sold, re-register and eventually get an updated year_of_registration and reg_code. The Dtype attribute for this feature is float64, which indicates the values are entered as decimal numbers. There are 33,311 null values for this feature in the original data set. The values of this feature are in a range of 999. to 2020. There are two probabilities for observing odd values for this feature in the dataset. They might be outliers or based on the relevant reg_code in their record, these values were entered into the dataset incorrectly during the process of data entry. (e.g. for the record below, we have the registration year of 1007.0 instead of 2007, regarding its relevant registration code which is 07 (Output1.))

	public_reference	mileage	reg_code	standard_colour	standard_make	standard_model	vehicle_condition	year_of_registration	price	body_type	crossover_car_and_van	fuel_type
59010	202006270588110	14000.0	7	Blue	Toyota	Prius	USED	1007.0	7000	Hatchback	False	Petrol Hybrid

Output1. Example of Wrong Data Entry in year_of_registration

In this study, it's assumed that this 6 months' time frame doesn't have a huge impact on the target price, so for a given reg_code, the 12 months of a year is considered. (e.g. reg_code of 7, is implying to year_of_registration of 2007). Another issue is the 31,570 records that are observed with null values for both of these two features which will be handled in the next sections.

1.1.3. vehicle_condition

This categorical feature contains binary string values of USED and NEW with no null values. The NEW value for vehicle_condition is just observed when we have null values both for year_of_registration and reg_code. It shows that, in this dataset, the vehicle condition will be considered as NEW only when the car hasn't been registered yet (Code Snippet1.). Since, the last value for year_of_registration is entered 2020, later in the data cleaning process, these vehicles which are 31'249 of the records, will be filled with the year after which is 2021.

```
df.loc[df['vehicle_condition']=="NEW"].equals(df.loc[(df["vehicle_condition"]=="NEW") & (df["year_of_registration"].isnull()) & (df["reg_code"].isnull())])
```

```
True
```

Code Snippet1. Not Registered Vehicles

It can be understood the criteria in the data entry process for vehicle_condition is just the year_of_registration in this dataset and mileage doesn't have a role in defining a car as NEW or USED (Code Snippet2.).

```
df.loc[df["vehicle_condition"]== "NEW", "mileage"].agg(['min', 'max']).to_dict()
```

```
{'min': 0.0, 'max': 100.0}
```

```
df.loc[df["vehicle_condition"]== "USED", "mileage"].agg(['min', 'max']).to_dict()
```

```
{'min': 0.0, 'max': 999999.0}
```

Code Snippet2. Value of 0 for the mileage Feature is Observed in both NEW and USED Vehicles

1.1.4. mileage

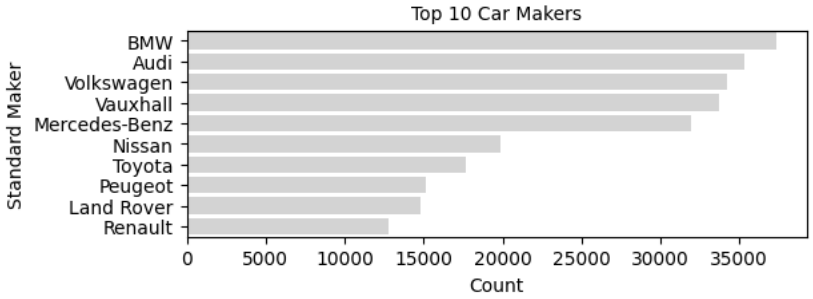
Mileage is a numerical feature which represents the distance the vehicle travelled in a year inferred from its negative correlation with price and year_of_registration. The values of this feature contain decimal numbers in the range of 0.0 to 999999.0 with 127 null values. There are 15,397 cars with NEW condition but not 0 mileage which is due to driving the distance from the firm to the customer. There are also 355 vehicles with USED condition that have 0 value for their mileage. It is due to the logic behind the data entry process which is considering a vehicle condition NEW only when it hasn't been registered or just because of a mistake in entering the data into the dataset.

1.2. Analysis of Distribution (3-4)

In this section, the distribution of a few features is described by visualising the relevant plots.

1.2.1. Distribution of Vehicles Produced by Top 10 Car Companies in the Dataset

This bar plot (Plot1.) illustrates the top 10 car companies that have the most instances in the dataset. BMW is the most frequently represented brand and is followed by Audi, Volkswagen, Vauxhall and Mercedes-Benz which is displayed by the right skewness in the plot. It suggests that these few brands dominate the dataset by having a significant share of the records. It probably reflects the higher production rate of these companies, or the market demand.



Plot1. Distribution of Top 10 Standard Makers

1.2.2. Distribution of Car Prices by Fuel Type

The box plot (Plot2.) illustrates the distribution of car prices across different fuel types. It seems that vehicles with Petrol Plug-in Hybrid and Diesel Hybrid fuel types have the widest price ranges according to the length of their boxes in the plot. Regarding the plot and also checking the median price of each group of fuel types in Code Snippet3. and Output2., we can confirm that the vehicle with Diesel Hybrid, Diesel Plug-in Hybrid and Petrol Plug-in Hybrid fuel type are more expensive than others.

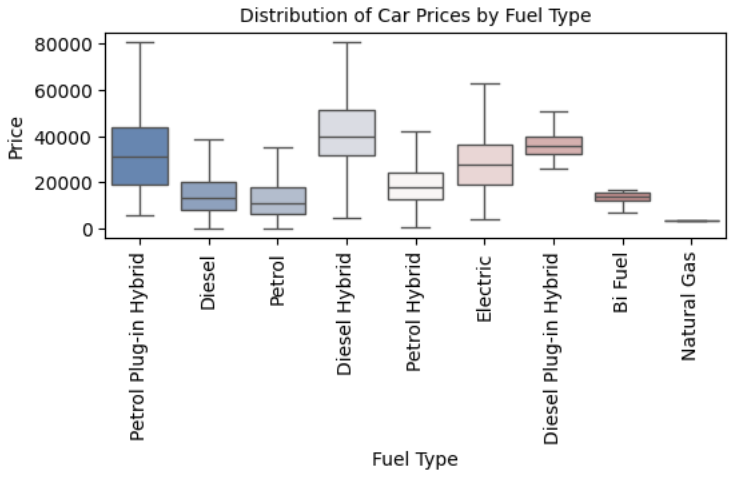
```
# Calculate the median price for all fuel types
median_prices = df.groupby('fuel_type')['price'].median().reset_index()

# Sort the median prices for all fuel types in descending order
sorted_median_prices = median_prices.sort_values(by='price', ascending=False).reset_index(drop=True)
sorted_median_prices
```

Code Snippet3. Group and Sorting the Median Price in Vehicles based on their Type of Fuel

	fuel_type	price
0	Diesel Hybrid	39990.0
1	Diesel Plug-in Hybrid	35991.0
2	Petrol Plug-in Hybrid	30995.0
3	Electric	27894.0
4	Petrol Hybrid	17814.0
5	Bi Fuel	14000.0
6	Diesel	13495.0
7	Petrol	11000.0
8	Natural Gas	3795.0

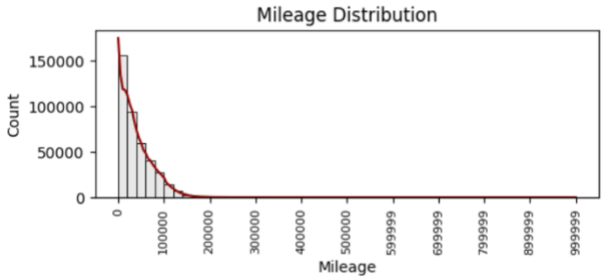
Output2. Median Price of Each Fuel Type Group



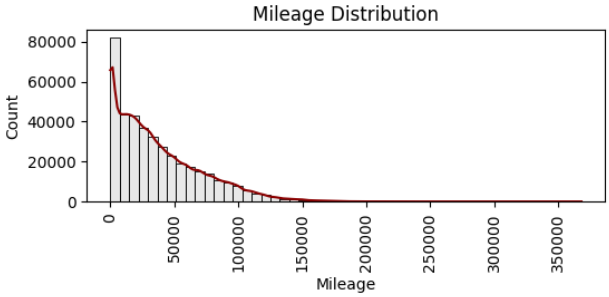
Plot2. Distribution of Car Prices Across Fuel Type

1.2.3. Distribution of Mileage

In this histogram (Plot3.) the distribution of the numerical feature, mileage is visualised. The distribution is right-skewed and the peak indicates the lower values of mileage, and the long tail indicates the higher values of mileage. It can be interpreted that the majority of vehicles in this dataset have lower amounts of mileage. So, most vehicles are lightly used and are in good condition. The long tail and the extremely high amounts of mileage values indicate the presence of outliers or probably old or heavily used vehicles which are less frequent. In the data preprocessing procedures, the missing values will be filled by median value of mileage in each group that made after applying grouping method and is based on standard_make and standard_model. Outliers will be managed by capping method, which replaces high outliers with the maximum value (upper bound) allowed (Plot4.).



Plot3. Distribution of Mileage Before Data Pre-Processing



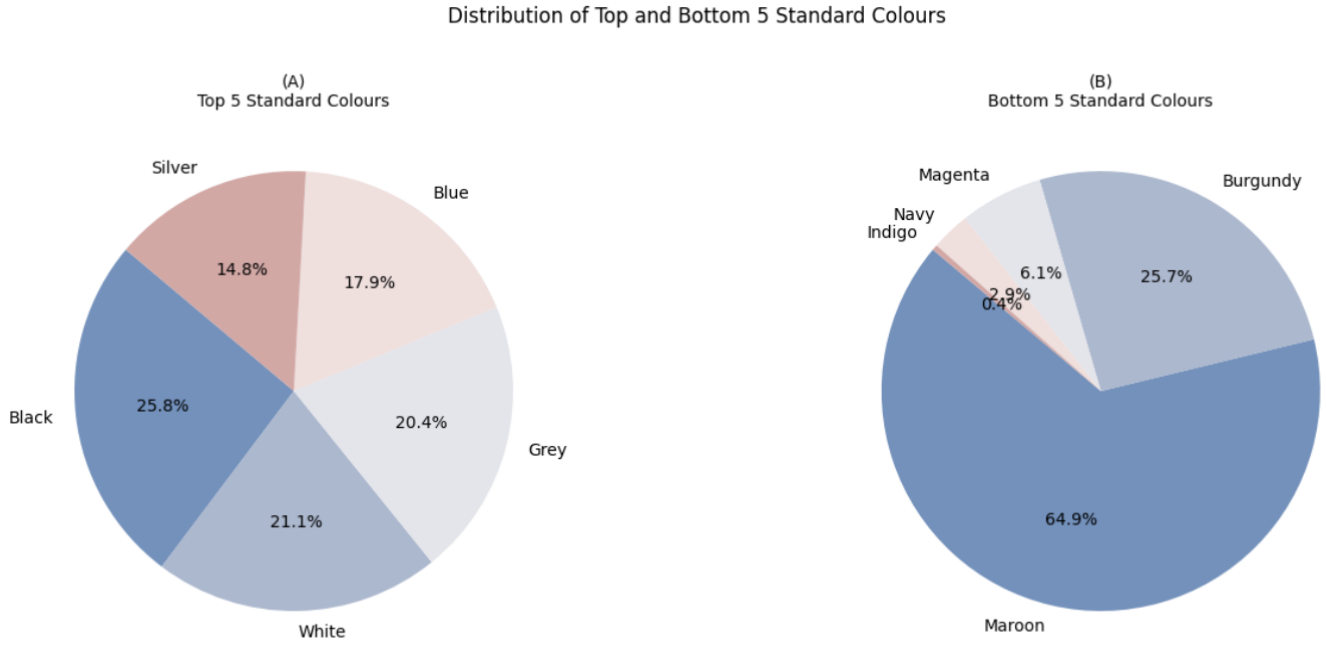
Plot4. Distribution of Mileage After Data Pre-Processing

1.2.4. Distribution of Colour

This pie chart (Plot5.) illustrates the distribution of the most and least vehicle colours within the dataset. Among the top 5 colours, black with the proportion of 25.8% is the most frequently observed instance and it is followed by white with the share of 21.1% and grey with 20.4%. The customer preference for these colours might be the result of availability and ease of resale for vehicles with these shades which is displayed in Plot5.A. below.
Plot5.B. may not necessarily be representative of the least preferred colours. With a closer look at the records, we can find Jaguar F-Type in Indigo colour, a high-priced luxury sports car which can be considered a valid outlier in this dataset (Output3.). Therefore, the reason for the low representation of these given colours might not be the lack of preference in the market but the rarity of these vehicles.

	public_reference	mileage	reg_code	standard_colour	standard_make	standard_model	vehicle_condition	year_of_registration	price	body_type	crossover_car_and_van	fuel_type
331032	202010064661534	54927.0	64	Indigo	Jaguar	F-Type	USED	2014.0	27000	Convertible	False	Petrol

Output3. An Example of a Rare Observed Colour in the Dataset



Plot5. Distribution of 5 Most(A) and Least(B) Observed Vehicle Colours

2. Data Preprocessing

2.1. Data Cleaning (2-3)

In this segment, three procedures in data pre-processing which are used to address issues such as missing values and outliers are described. While similar approaches are applied to other features, we focus on year_of_registration and reg_code to have a thorough comprehension of the changes made and maintain consistency.

2.1.1. Handling Missing Values and Dealing with Incorrect Values: Filling Values Based on Other Features

The reg_code and year_of_registration features are filled based on the defined dictionary (year_to_reg_codes) for years 1963-2001. Three decisions are made in this stage based on the condition:
a) If both reg_code and year_of_registration have null values, and “NEW” for the vehicle_condition, we assume the year 2021 which is the year after the last year is recorded in the dataset.
b) If the reg_code value is valid, year_of_registration will be filled based on the relevant reg_code for that record.
c) If the reg_code value is invalid, we look into the year_of_registration feature and by mapping, we assign the relevant value of year_of_registration to the reg_code.

2.1.2. Dealing with Outliers: Dropping the Records

One way to deal with the outliers is by removing them from the original dataset. Although some of these records are valid (Output4.) (Fig.1.), their removal simplifies the process of defining the mentioned dictionary and improves the accuracy of statistical measures by reducing noise and ensuring consistency across the dataset.
There are 44 observed records from years before 1963 which are decided to be removed from the dataset. Removing this amount of data (0.01%), won’t significantly affect the insights derived from the dataset.

	public_reference	mileage	reg_code	standard_colour	standard_make	standard_model	vehicle_condition	year_of_registration	price	body_type	crossover_car_and_van	fuel_type
156562	202009264242828	48000.0	FW	Black	Morris	10	USED	1934.0	5995	Saloon	False	Petrol

Output4. Example of a Valid Outlier

2.1.3. Handling Missing Values and Dealing with Incorrect Values: Grouping and Imputation Values

By using the dictionary and applying the mapping process, the issue in records that have incorrect or null values of year_of_registration and reg_code is solved. However, this approach falls short when one feature contains a null value while the other has an incorrect value and this is because the dictionary is defined based on only the valid values for both of these features. To resolve this, an imputation function (fill_missing_years) was introduced to fill the year_of_registration based on the standard_make and standard_model features. The logic behind the decision to use standard_make and standard_model features for grouping is their association with year_of_registration. The year_of_registration is linked to those given features because certain vehicle models are produced in specific periods of time. By grouping the dataset by standard_make and standard_model, and filling the year_of_registration with the observed mode in each group (Output.5.), and reapplying the dictionary, there will be no more records with null values in reg_code and year_of_registration.

```
grouped_modes = df.groupby(['standard_make', 'standard_model'])['year_of_registration'].agg(lambda x: x.mode()[0] if not x.mode().empty else None).reset_index()
grouped_modes.columns = ['standard_make', 'standard_model', 'year_of_registration_mode']
grouped_modes
```

Code Snippet4. Data Grouping and Mode Imputation

Fig1. Example of a Valid Outlier [1]

	standard_make	standard_model	year_of_registration_mode
0	AC	Cobra	2017.0
1	AK	Cobra	2011.0
2	Abarth	124 Spider	2018.0
3	Abarth	500	2009.0
4	Abarth	500C	2012.0
...	...	...	...
1195	Westfield	Se	2017.0
1196	Westfield	Sei	1992.0
1197	Westfield	Sport	1997.0
1198	Wolseley	6/110	1964.0
1199	Zenos	E10	2018.0

1200 rows × 3 columns

Output5. Groups of Data with Relevant Registration Year

The same grouping procedure based on standard_make and standard_model is applied to other features including mileage, standard_colour, body_type, and fuel_type due to the strong association of these features with the vehicle model. Unlike year_of_registration, these features still have null values after grouping. This is likely due to some records not fitting into the defined groups. To solve that issue, a global mode for categorical features such as standard_colour, body_type, and fuel_type and a global median for numerical features like mileage is defined to be applied in case of having null values after grouping procedures. This ensures that all missing values are handled appropriately after the grouping process.

2.2. Feature Engineering (2-3)

2.2.1. Applying Binning Method

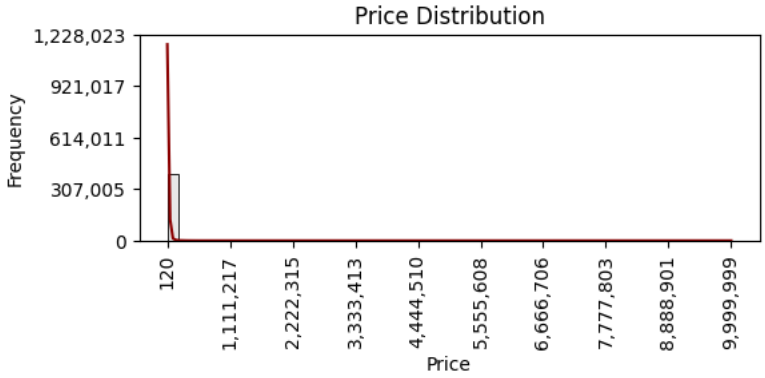
One of the approaches to analyse the association of vehicle prices with other features is through the binning method which divides the continuous numerical values of price into 3 distinct categories: Low, Medium, and High(premium) price. The price_category column is added to the original dataset through this implementation. These categories are defined based on the price distribution percentiles as follows:

1) Low Price Vehicles: Records below the 25th percentile
2) Medium Price Vehicles: Records between 25th and 75th percentiles
3) High Price Vehicles: Records above the 75th percentile

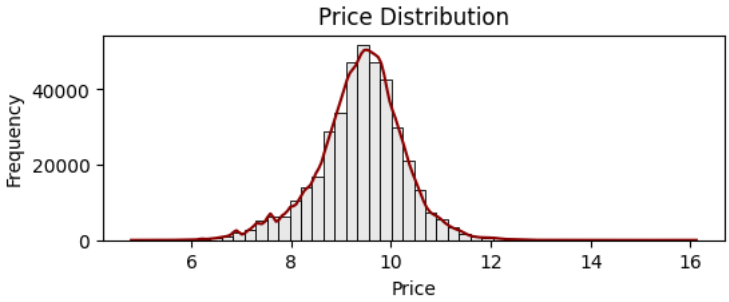
The binning method was applied using this logic because of the significant differences between high-priced vehicles and other records across other features. These kinds of records will cause issues, including bias in the model training process which affect the accuracy and reliability of the predictions in the future so, it is important to treat high-priced vehicles differently.

2.2.2. Applying Log Transform Method

Log transformation is applied on the price feature in this dataset and the transformed values are creating a new column named log_price. By applying this method, we can improve the price distribution normality and reshape it into a more symmetric bell-shaped curve. So, we can have a clearer visualization, valid accuracy and better model performance in later procedures. The plots below illustrate the price distribution before and after applying the log transformation. We can observe positive skewness distribution in Plot6. because of outliers' presence and in the Plot7. the range of outliers is reduced and made the curve more balanced.



Plot6. Price Distribution



Plot7. Log Transformation Effect on Price Distribution

2.3. Subsetting (2-3)

2.3.1. Feature Selection

For this part, the association of features in the dataset is evaluated with the target feature price to ensure which features are actually important for us to be considered in future procedures such as model training.

The association of categorical features are examined with feature price_category and for this purpose, the Chi-squared test is applied to compute the Chi-square statistic and p-value for the hypotheses test of independence of observed frequencies [2]. The assumption for H0, is that there's no association among the selected features. By having a look to the result of this function (Output6.) and the measured p-values, we can confirm that all of those features have significant impacts on our target price. Also, based on the Chi-squared scores in this output, we can find out that the standard_model has the highest association with the price.

The association of numerical features is examined with feature price and to do so, the correlation function by using Spearman method is applied. Unlike Pearson correlation, Spearman correlation can make us capable of capturing non-linear relationships among the features. This function result displays that both mileage and year_of_registration features have a strong negative correlation with each other (Output7.).
As a result, these given features are also should be kept for future procedures.

	Chi2 Statistic	p-value
standard_colour	15373.486452	0.000000e+00
standard_make	109029.141594	0.000000e+00
standard_model	272198.169999	0.000000e+00
vehicle_condition	43173.888094	0.000000e+00
body_type	67203.945993	0.000000e+00
crossover_car_and_van	585.865613	6.038075e-128
fuel_type	25474.799895	0.000000e+00

Output6. Association of Categorical Features with Target Feature of price_category

	mileage	year_of_registration	price
mileage	1.000000	-0.861406	-0.645278
year_of_registration	-0.861406	1.000000	0.704944
price	-0.645278	0.704944	1.000000

Output7. Association of Numerical Features with Target Feature of Price

2.3.2. Filtering

In several parts of this project, data filtering was used to narrow down the records that we wanted to focus on as a group of data with given conditions and have a thorough comprehension of the whole dataset. One of the examples of implementing this strategy is shown below (Code Snippet5.).
In this part, after grouping the average price of vehicles based on the body_type category, we got Limousines as the most expensive vehicles in the output. In the next stage, a specific condition such as Limousine for the body_type is set and the records are sorted based on their average prices. While it is expected that one of the high-priced coupes, convertibles or campers has the highest price, a Green Phantom Rolls-Royce limousine with price of £374,950 has this title(Fig2.).

```
df.loc[df['body_type']=='Limousine'].sort_values(by='price', ascending=False).head(5)
```

	mileage	standard_colour	standard_make	standard_model	vehicle_condition	year_of_registration	price	body_type	crossover_car_and_van	fuel_type	price_category	log_price
273840	103.0	Green	Rolls-Royce	Phantom	USED	2020.0	374950	Limousine	False	Petrol	High(premium)	12.834551
273821	1450.0	Purple	Rolls-Royce	Phantom	USED	2018.0	350000	Limousine	False	Petrol	High(premium)	12.765691
273820	19.0	Blue	Rolls-Royce	Phantom	USED	2020.0	345000	Limousine	False	Petrol	High(premium)	12.751303
273830	2942.0	Red	Rolls-Royce	Phantom	USED	2019.0	334950	Limousine	False	Petrol	High(premium)	12.721740
273793	63.0	NaN	Rolls-Royce	Phantom	USED	2019.0	329900	Limousine	False	Petrol	High(premium)	12.706548

Code Snippet5. An example of Data Filtering, Sorting the Average Price of Limousine Body Types

Figure2. Green Phantom Rolls-Royce limousine [3]

3. Analysis of Associations and Group Differences

3.1 Quantitative – Quantitative (2-3)

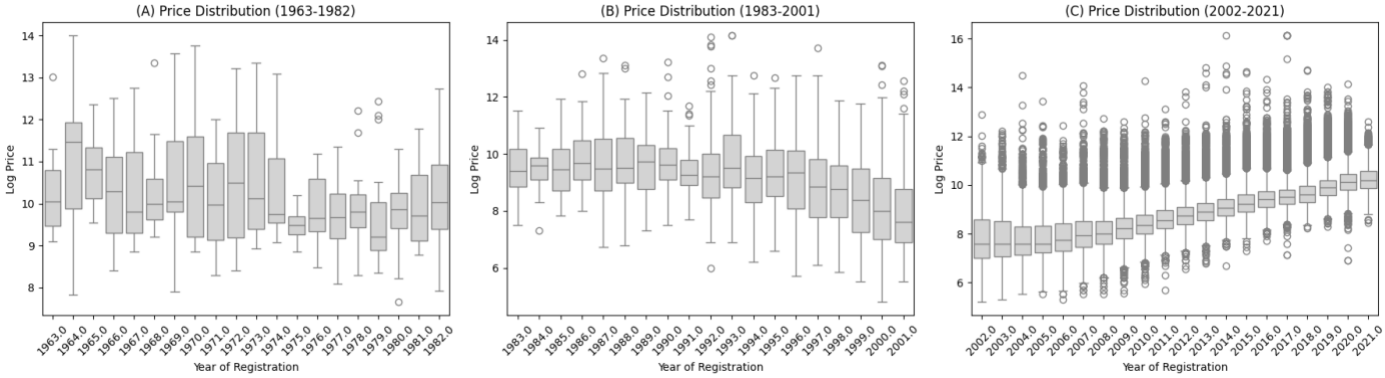
3.1.1. Distribution of Price Log Transformation over Registration Years

The boxplot (Plot8.) illustrates the log-transformed price across distinct time frames. Log transformed form of price is used because of the clarity of the plot. Three distinct time frames are as follows:

A) 1963-1982: The price pattern through the years indicates irregularity, high variability, and inconsistency in medians. The number of outliers is not very noticeable.

B) 1983-2001: The price pattern becomes more stable, with moderate variability and also more consistency in median values. Compared to the 1963-1982 time frame, more outliers are observed.

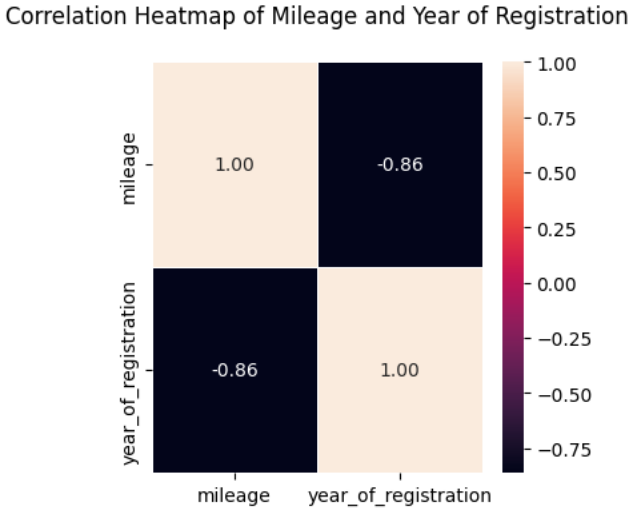
C) 2002-2021: The price pattern shows an upward in median prices and a higher number of outliers. The reason for the price growth occurrence might be technological advancement, increase in market demand, and the introducing vintage and premium vehicles.



Plot8. Price Distribution through Registration Years

3.1.2. Correlation of Registration Years and Mileage

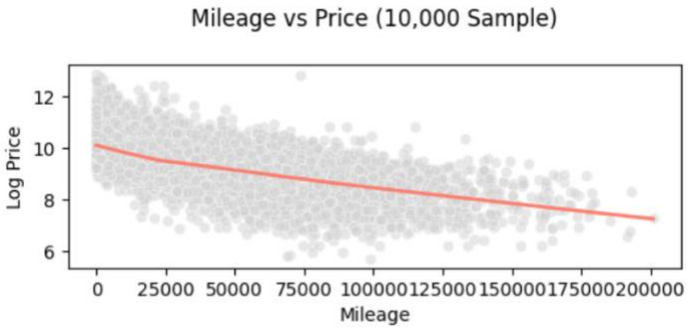
This heat map (Plot9.) visualises the correlation between year_of_registration and mileage features. These features have correlation amount of -0.86 which shows a strong negative correlation and we can interpret that by having higher amounts of mileage in vehicles, the amount for year_of_registration is going to be lower.
As mentioned previously in section 1.1.2, we know that year_of_registration is not necessarily representative of the vehicle's age, but we still can assume that vehicles registered in recent years (probably newer vehicles) generally, have lower mileage values, which might be attributed to technological advancements in the car production industry and also the improvements in fuel consumption.



Plot9. Mileage and Registration Year Correlation Heatmap

3.1.3. Distribution of Price across Mileage

This scatter plot (Plot10.) displays how target feature price which log transformation is applied on it, varies across different mileage values in a selected sample of 10,000 records in the dataset. The regression line highlights the negative relationship of these features with each other. As mileage increases, we can see the decline in the log-transformed price which means vehicles with high amounts of mileage tend to have lower prices. Also, price reduction is more noticeable at lower amounts of mileage due to the steeper slope that we can observe in the regression line. It means, in vehicles with very high amounts of mileage, price is still decreasing but with a smoother rate, indicating mileage has less effect on price at higher values.

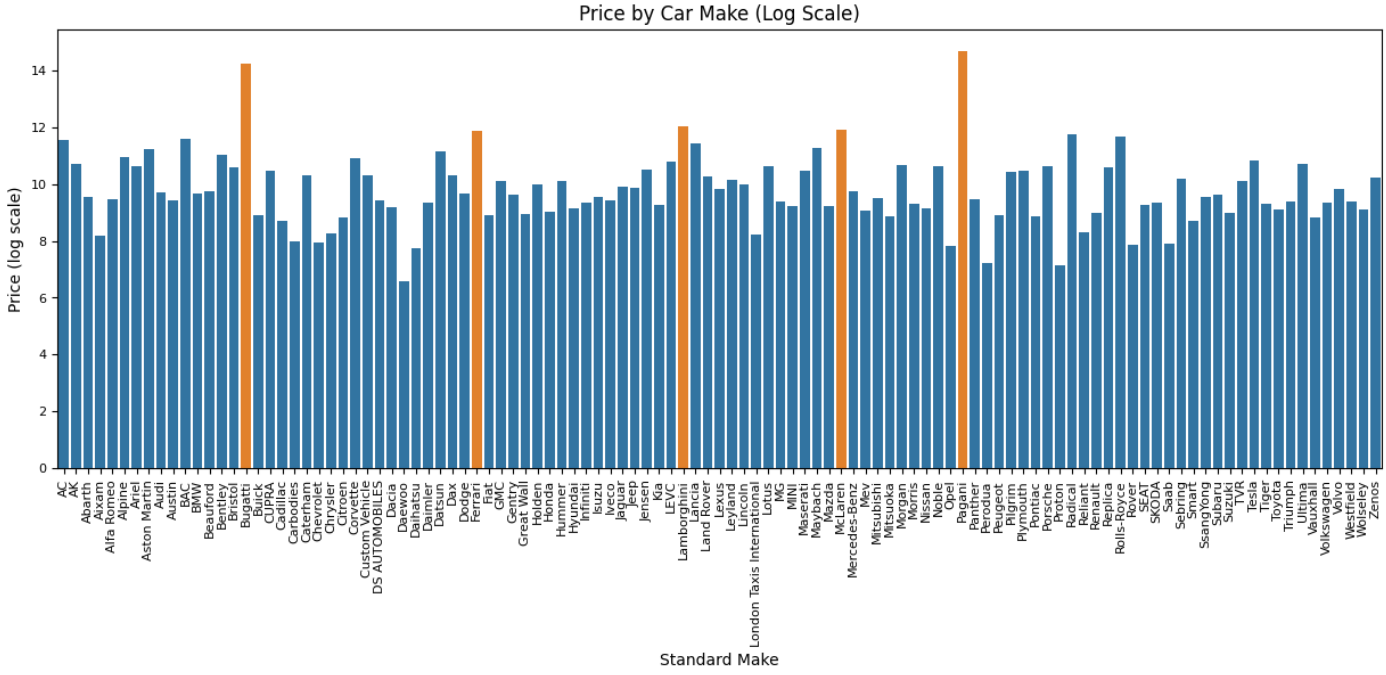


Plot10. Distribution Log Transformed Price over Mileage Using a Sample of 10,000 Records from the Dataset

3.2. Quantitative – Categorical (2-3)

3.2.1. Price Log Transformation over Standard Makers

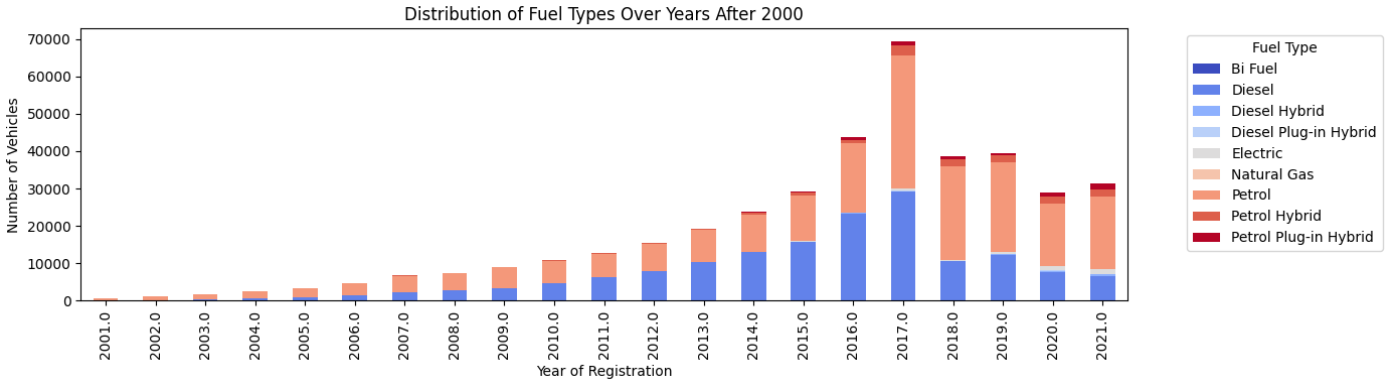
This bar plot (Plot11.) illustrates the price distribution across different car companies on a logarithmic scale. It also indicates the dominance of high-priced and luxury brands including Pagani, Bugatti, Lamborghini, Ferrari, and McLaren which are represented by orange color. Although the Pagani has the highest observed price among all manufacturers, this observation isn't entirely reliable because we only have one record of this company and this value may not represent the true average price of this company so, its price average is biased due to insufficient records in comparison other companies with a larger number of records in the dataset which provide a more reliable representation of their price distribution.



Plot11. Average Price of Manufacturers

3.2.2. Distribution of Fuel Type across Registration Years

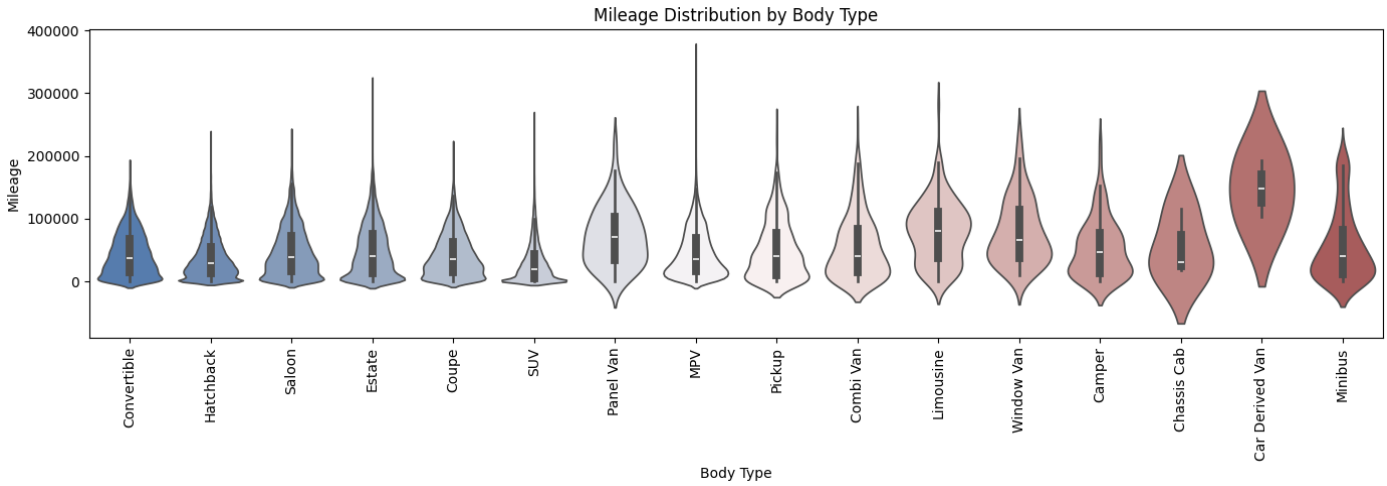
This stacked bar plot (Plot 12.) displays the distribution of fuel types in vehicles which are registered after year 2000. Petrol and Diesel are the most dominant fuel types through all these years. Petrol has a significant increase in 2017 and Diesel fall is started from that year. We can observe the emergence of Petrol Plug-in Hybrid from year 2014 and growth in number of Electric vehicles in years 2020 and 2021. Fuel types like Bi Fuel and Natural Gas rarely observed across these years. The peak in the 2017, followed by a decline, might indicate the influence of economic factors, shifts in market demand and vehicle registration as the result.



Plot12. Distribution of Fuel Types Over the Years After 2000

3.2.3. Distribution of Mileage across Body Type

This violin plot (Plot13.) displays the mileage distribution across vehicle body types. Regarding the median values of each category, we can observe that Car Derived Van, Limousine, and Panel Van show higher mileage values. This is aligned with the obtained result from the code which is grouping the records by their body type feature and calculates the median mileage for each group. Due to the services that these kinds of vehicles are used for, which involve transportation, deliveries, and logistics, they travel longer distances and that's the reason of having higher mileage amounts. On the other hand, SUVs and Hatchbacks have personal use cases, show wider base in their violin plots indicating large number of instances with lower mileage values. They also have narrower Interquartile range which means most of the records in these categories have similar use cases and less variation in the annual mileage.



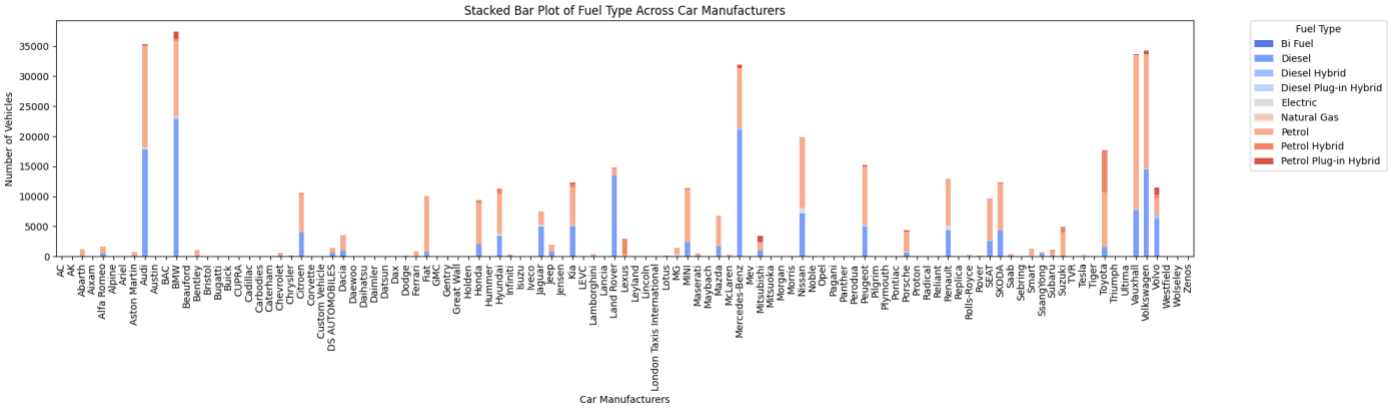
Plot13. Mileage over Vehicle Body Types

3.3. Categorical – Categorical (2-3)

3.3.1. Distribution of Fuel Type Over Standard Makers

This stacked bar plot (Plot14.) illustrates that Petrol and Diesel are the most common fuels across all manufacturers and this might be the result of market demands and manufacturer strategies. There are some interesting observations in this dataset:

- A large proportion of Petrol Hybrid vehicle production is limited to a few manufacturers such as Lexus and Toyota.
- Tesla's only focus is on producing Electric vehicles.
- Although manufacturers such as BMW, Mercedes-Benz, Toyota, and Volvo produce vehicles with a variety of fuel types, Petrol and Diesel remain the dominant preferences in their production lines.



Plot14. Distribution of Fuel Type in Manufacturers

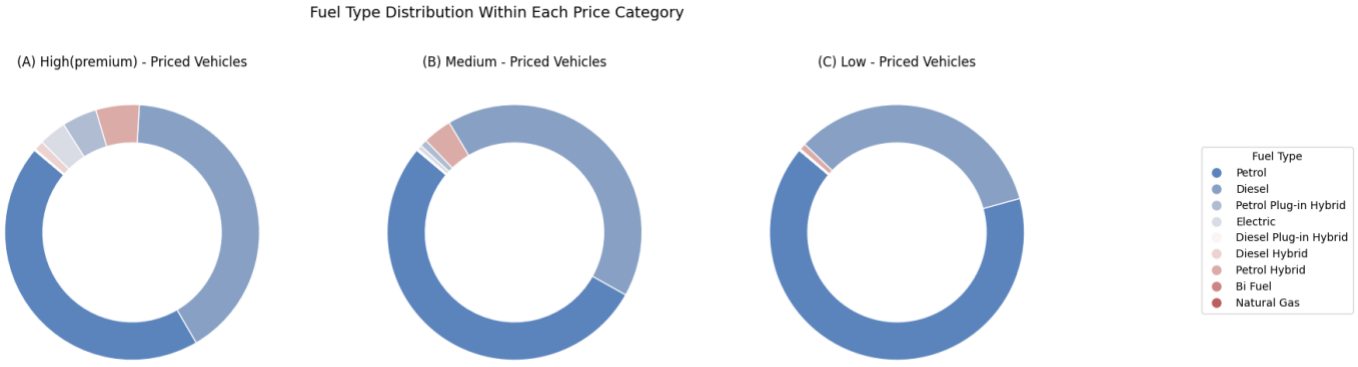
3.3.2. Price Category across Fuel Type

These donut charts (Plot15.) display the distribution of fuel types across three defined price categories:

A) High-priced vehicles: In this category, a wider variety of fuel types is observed beside the dominance of Petrol and Diesel types, in comparison to the other price categories. Also, there is a larger proportion of Electric and Petrol Plug-in Hybrid vehicles which can suggest that high-priced vehicles are often take advantage of having more advanced and modern technologies.

B) Medium-priced vehicles: this category is also shows the dominance of Petrol and Diesel fuel types. Petrol Hybrid, Petrol Plug-in Hybrid and Electric vehicles are also present in this price category but in smaller proportions meaning, these vehicles typically are not affordable within this price range.

C) Low-priced vehicles: this category, Petrol and Diesel have the largest shares, suggesting that low-priced vehicles still rely on traditional fuel types which are less costly for the manufacturer in the production process.



Plot15. Distribution of Fuel Type in Price Categories

References

[1] Wikipedia contributors, “Morris Ten,” Wikipedia, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Morris_Ten.
[2] SciPy Developers, “scipy.stats.chi2_contingency,” SciPy Reference Guide, 2023. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html.
[3] WrapStyle, “Rolls Royce Phantom Green Chrome Wrap,” WrapStyle, 2023. [Online]. Available: https://www.wrapstyle.com/gallery/58-rolls-royce-phantom-green-chrome-wrap.

Chatgpt and Grammarly tools are used in this text for text revising and code optimisation.