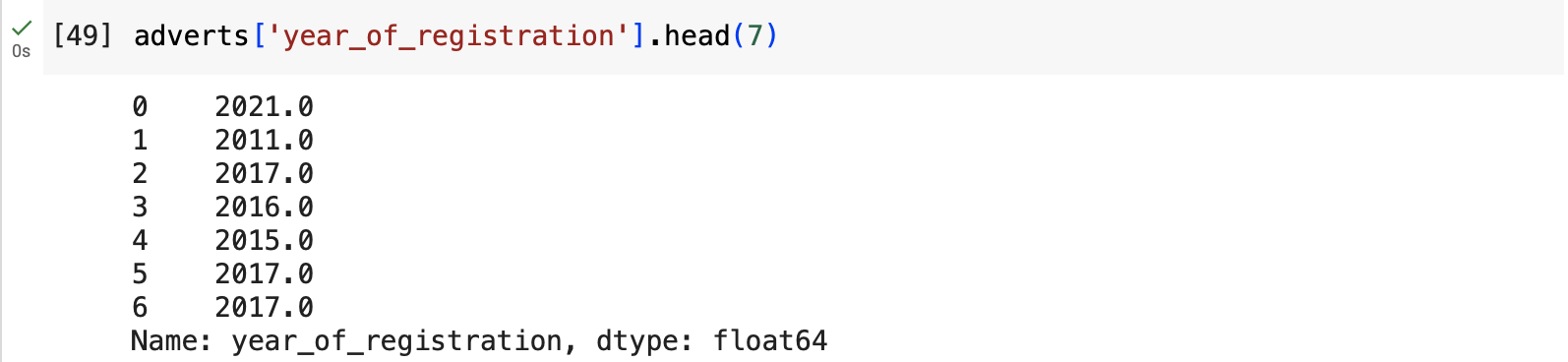
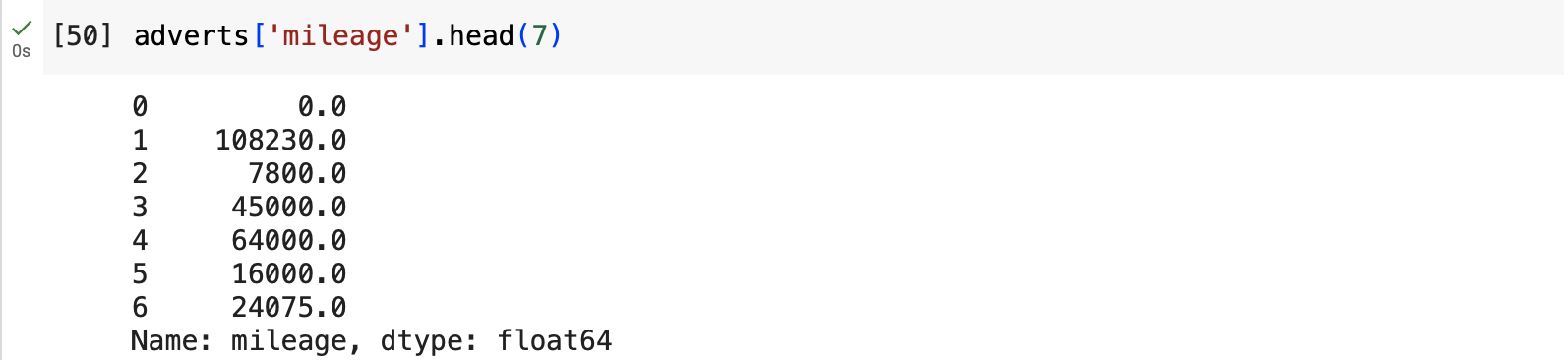
**1.1 Meaning and Type of Features**

1.Body Type: A car body type is the categorisation of vehicles based on their design, shape, size and the kind of space it has inside of it, this body type can include both streamlined body to foldable roof/body, body types depend on different factors like the year it was made, to number of doors, to style of roof and boot design/space, to its uses. They consist of Hatchback, SUV, Van, Saloon etc. The Body type consists of objects data type.

2.Year of Registration: The year of registration is the year a car was registered, it doesn’t necessarily mean the year it was manufactured, as it can be manufactured this year and registered on the next year. Year of registration is also an important factor in determining prices of car, a recent year would have a higher price compared to an old year. The year of registration has a float data type.

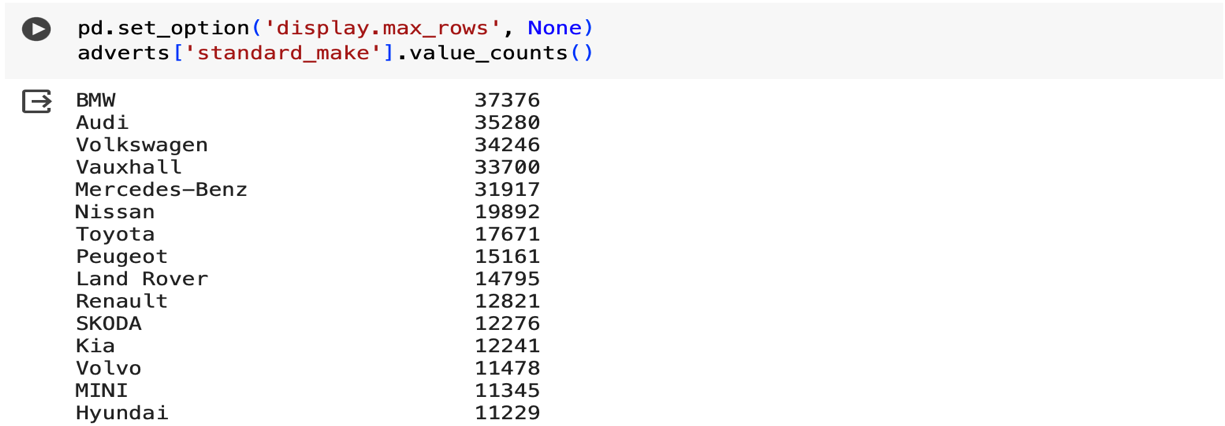


3.Mileage: A car mileage is the number of kilometers the car was driven for since it was manufactured or the total distance it has travelled since production, this number can also represent how often the car has been used and it automatically affects how it would be priced. A high mileage will automatically mean that the car has been used for a long time and driven long distance, zero mileage means car is brand new and low mileage means car has not been driven for a long time. This feature is a numerical feature because it is all numbers and is represented as a float.



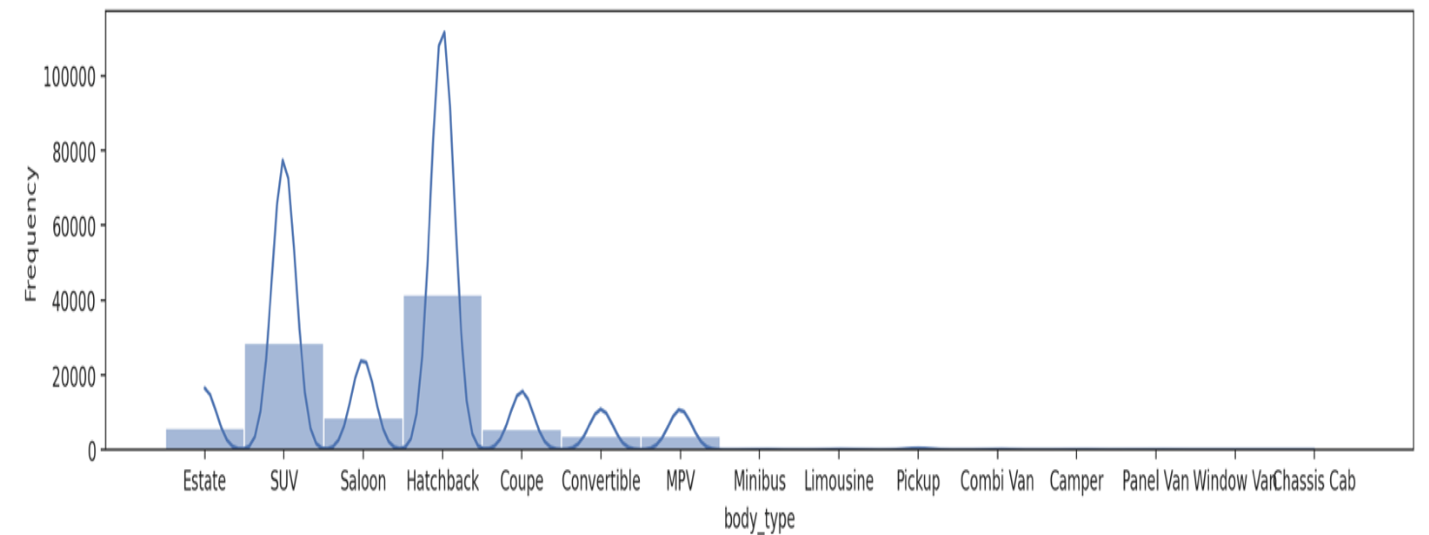
4. Standard Make:

The make of a car is also a great factor in influencing car price, known car brands would first get attention or get an offer first before other brands. For example, some car brands have more prestige than other common brands. From the picture below a BMW SUV would be of a higher price compared to a Hyundai or Ford SUV due to the prestige and reputation of the BMW brand.

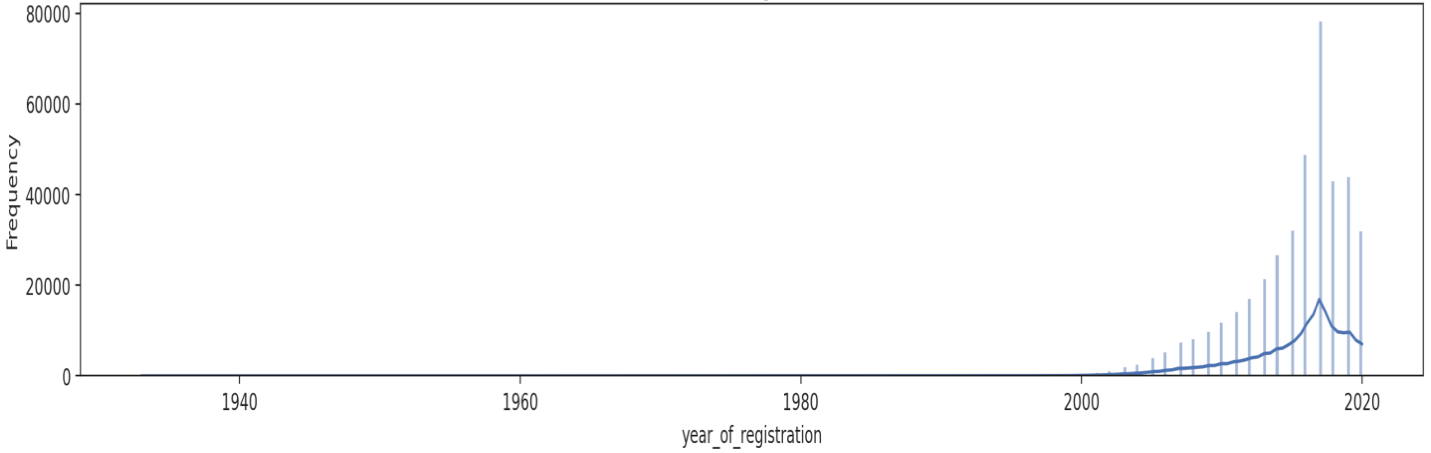


**1.2 Analysis of Distributions**

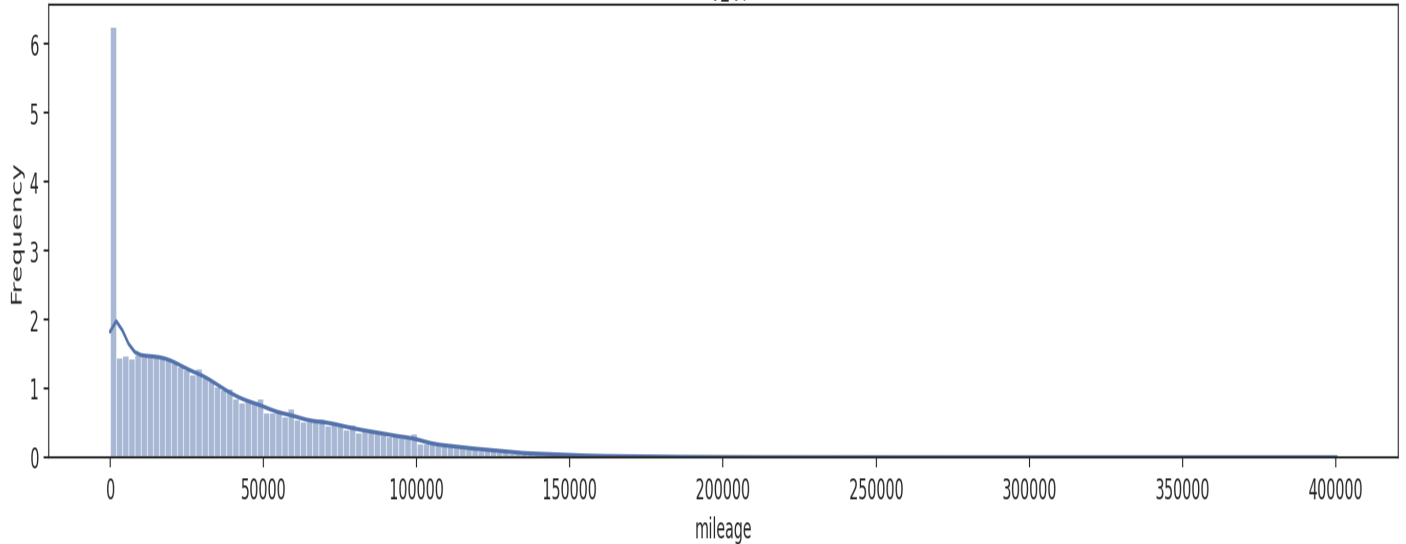
1. Body Type: The body type feature has a bimodal distribution because the hatchback and SUV are more frequent to occur in the dataset i.e., there are two highest main points or prominent peaks for the hatchback and SUV giving it a bimodal structure. This two car body types have the most frequency of occurrence.



2. Year of Registration: The year of registration has skewed distribution, the distribution shows more values are concentrated to the right side of the distribution, while the left is longer, this is negatively skewed distribution. This shows more frequency on recently registered cars compared to old ones.

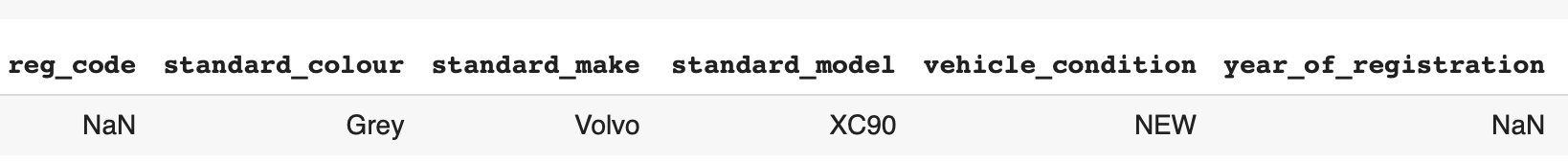


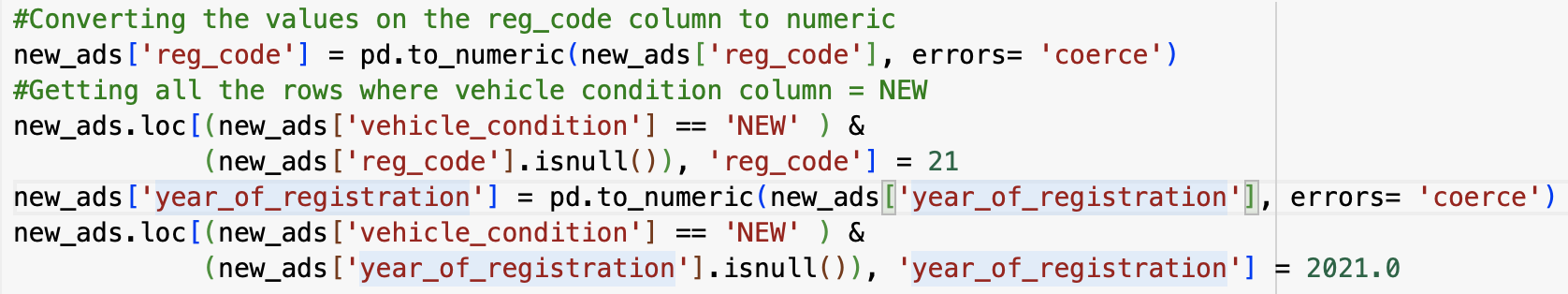
3. Mileage: the mileage has also a skewed distribution, the distribution shows more values are concentrated to the left side of the distribution, while the right is longer, this is a positively skewed distribution. This shows more frequency on brand new cars compared to used cars.

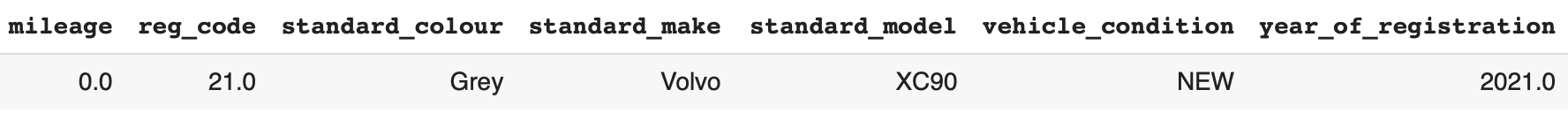


**2.1 Data Cleaning** (e.g., dealing with incorrect values, outliers)

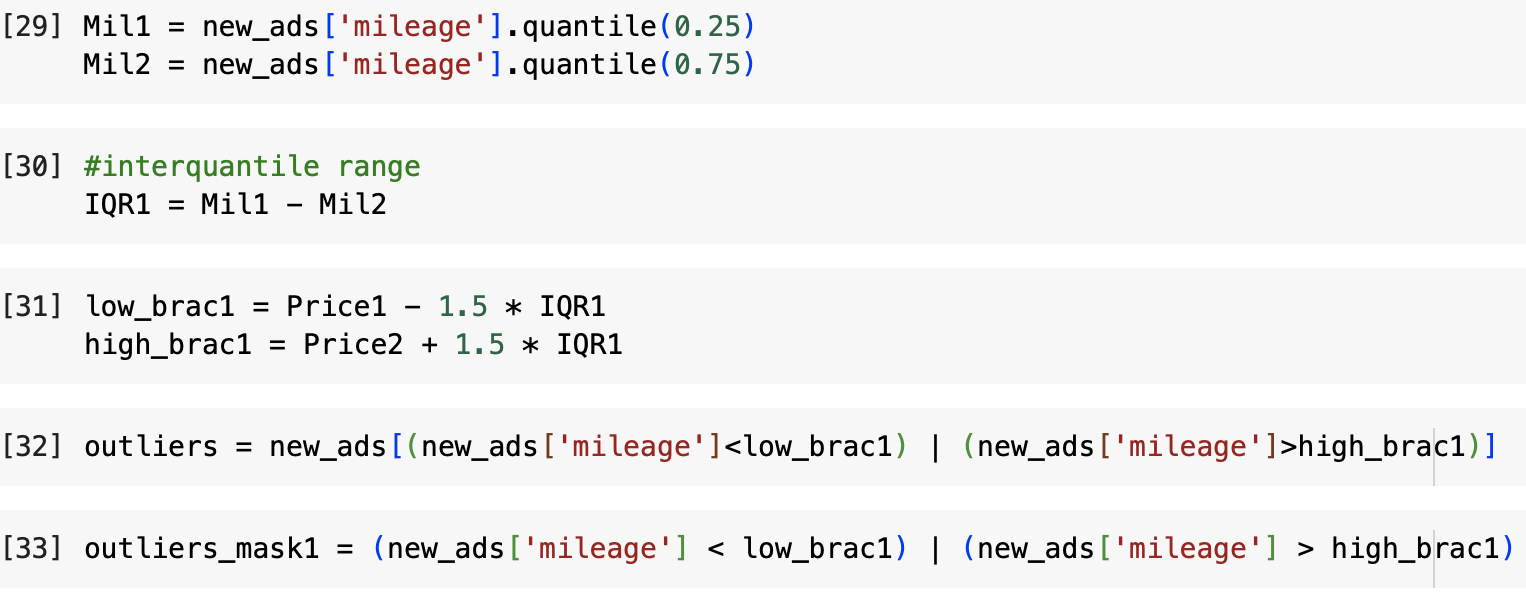
1. Regcode and Year of registration: I did fill up the null values for Regcode and Year of Registration with 2021. I queried the data set for the latest year, using adverts.[year\_of\_registrationmode].max it shows the modal year as 2020. so, I decided to fill the missing values with 2021 as the current year.



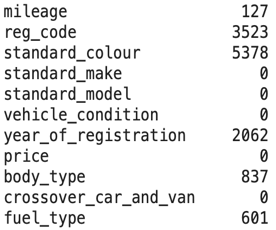
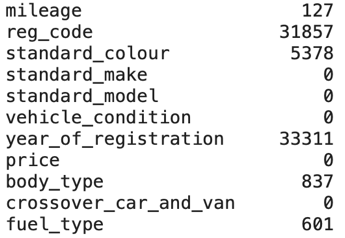
I converted the values on the reg\_code column to numeric, and for every row where vehicle\_condition equals new, I changed it to 21 and also for every row where reg\_code equals 21, I set year as 2021.



2.Handling Outliers for Mileage and Price: I calculated the first quartile (Mil1) and third quartile (Mil2) of the column ‘mileage’, then I got the Interquartile range (IQR). I used this method to detect and

potentially address outliers based on the distribution of data within the interquartile range.

3.Droping missing values: after reducing missing values for reg code and year\_of\_registration to the minimum by filling up, I dropped all missing values, using the df.dropna command to drop all missing values.

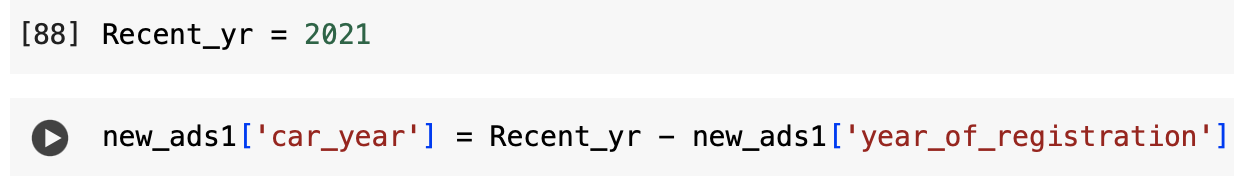


**2.2 Feature Engineering** (e.g., deriving informative features)

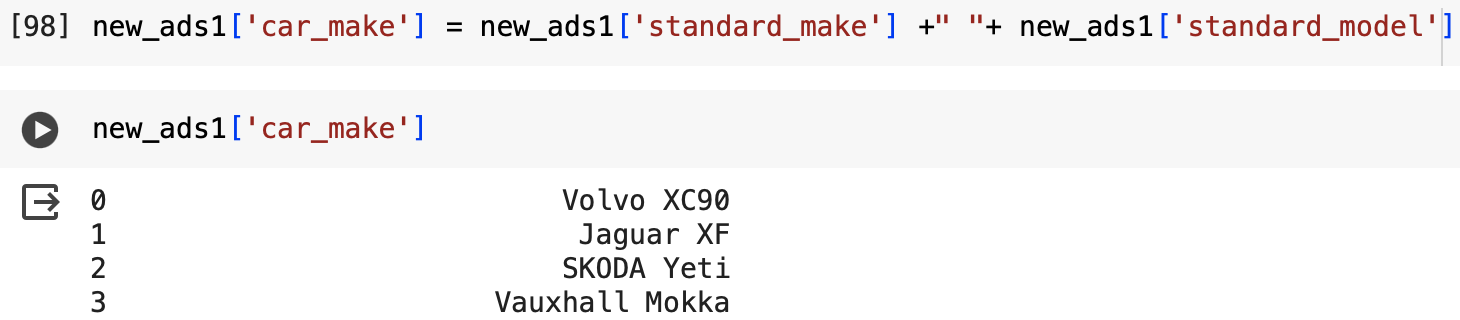
1. Dropping features: I dropped features like Reg\_code because it holds likely the same value as the year\_of\_registration and both reflect vehicle age, I will also be dropping the standard colour, public reference and crossover\_car\_and\_van, because they have no direct correlation with price.



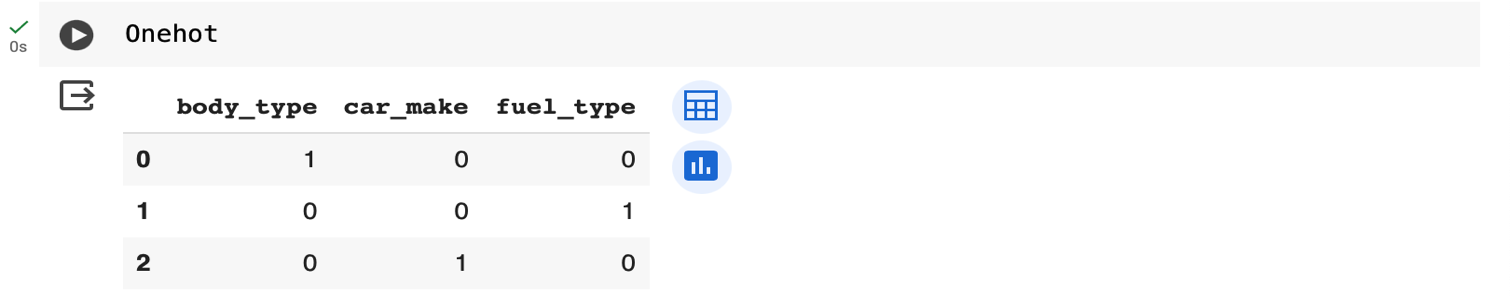
2. Creating new features: I created a new feature called the car\_year, using 2021 as recent year, I deducted car year from recent year to give me a new column called the car\_year.



I also created a new feature called the car\_make, I combined both the standard\_make and standard\_model as the car\_make.



3. Encoding categorical features: I used one-hot encoding to combine and encode "standard\_colour", "body\_type", "fuel\_type", and also "car\_make" also.

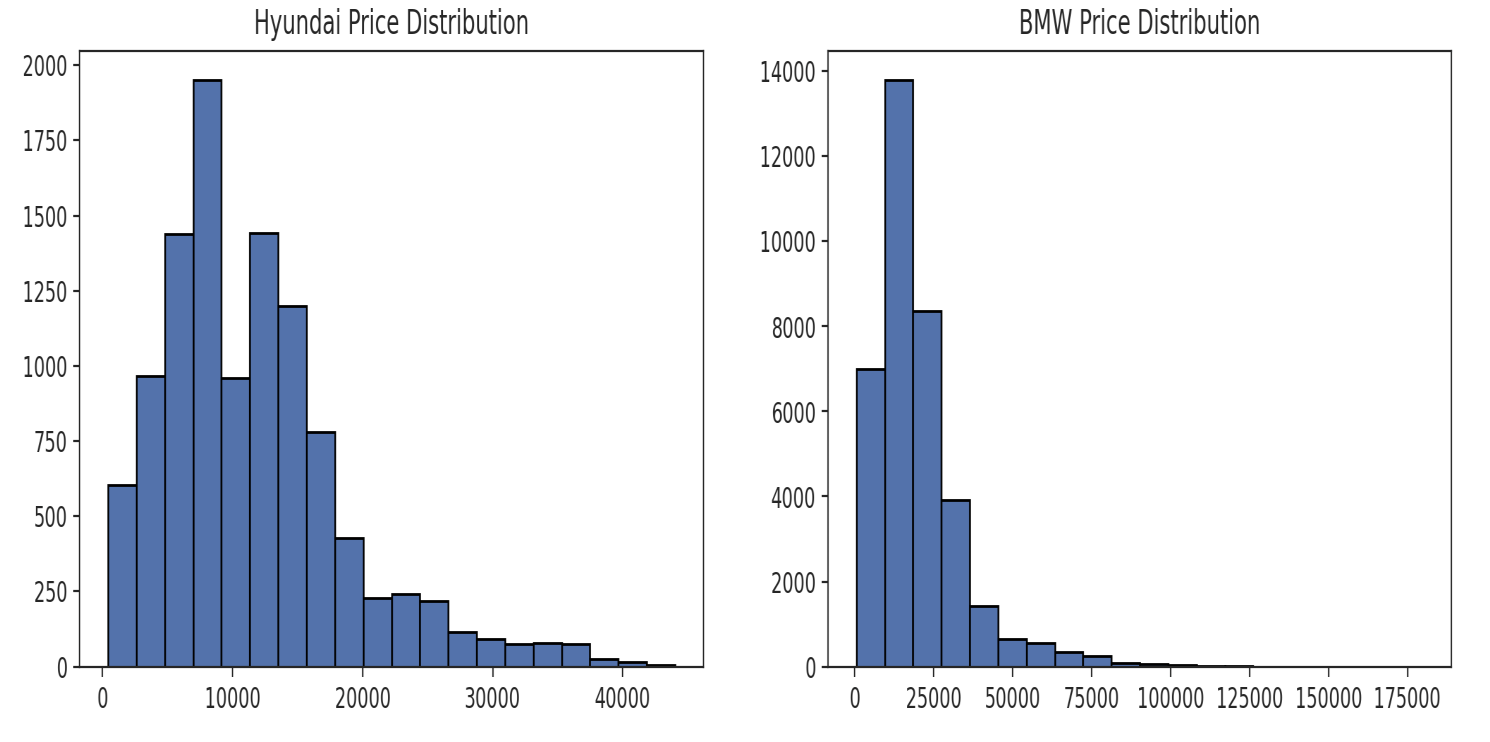


I also used ordinal encoding for vehicle\_condition, where it is “NEW” is used 1 and when it is “USED” I used 0,

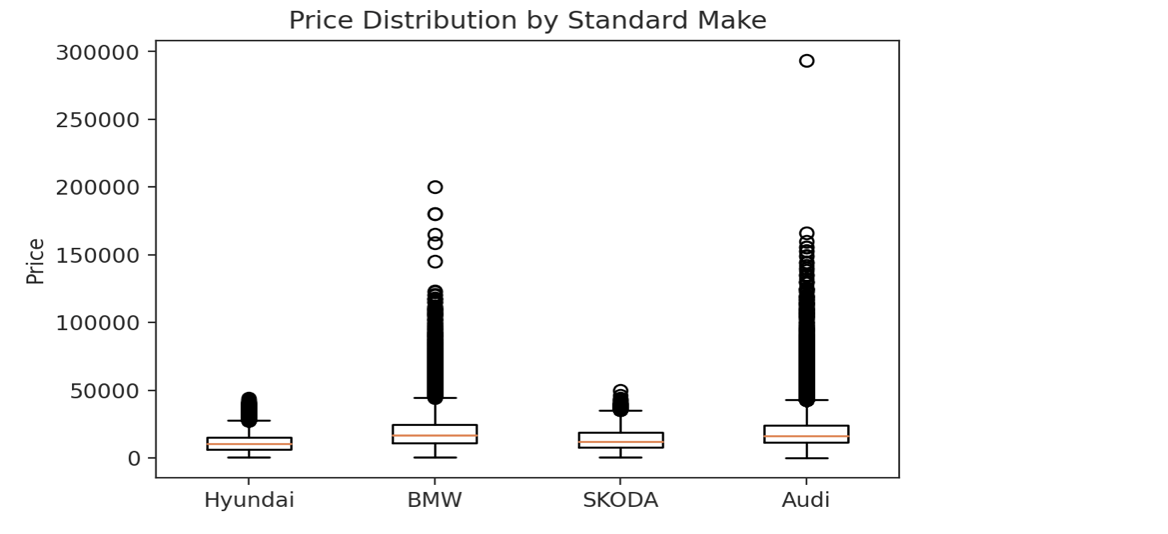


**2.3 Subsetting (e.g., feature selection and row sampling)**

1). I did a row sampling by taking the subset of standard make using rows like the Hyundai cars and BMW cars to check their price distribution.

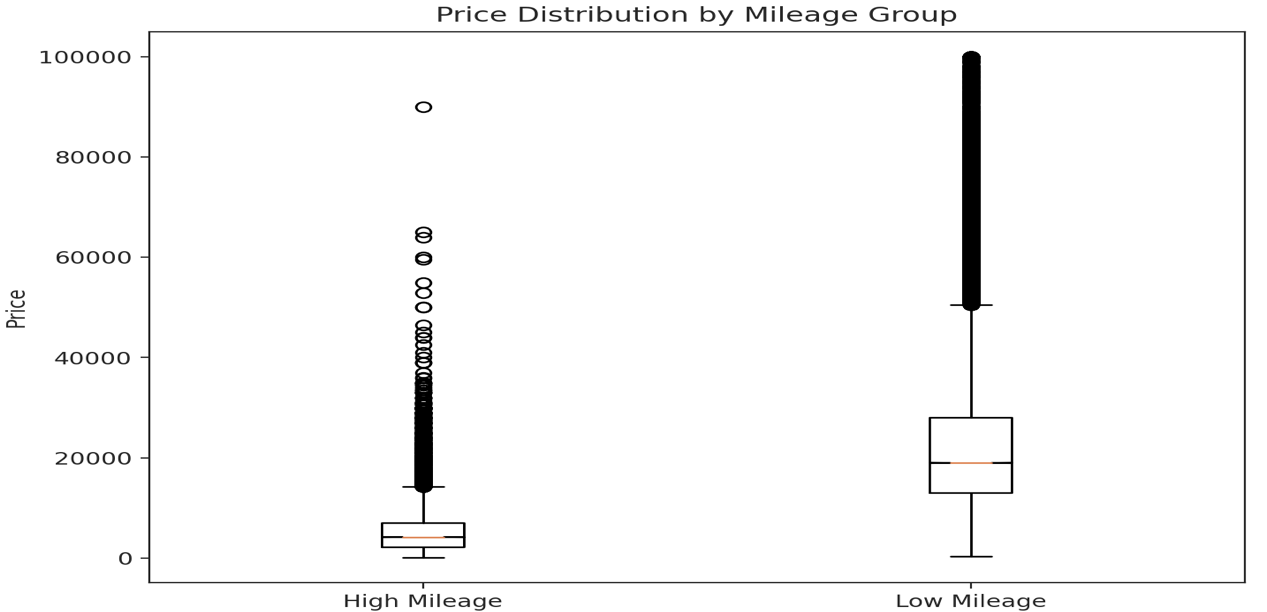


From the chart we can see that BMW cars have higher price distribution than the Hyundai, with the multitude of Hyundai cars being priced below 30,000 and BMW cars being priced above 70,000. This also shows the relationship of car make with their prices. Also, the price distribution of Hyundai, BMW, SKODA and Audi.



This shows the influence a car make can have with the price. A Hyundai SUV would be cheaper compared to a BMW or Audi SUV.

2. Also checking the price distribution by mileage group. We can see that cars with low mileage have good prices compared to cars with high mileage. This shows direct influence of mileage on car prices.



**Analysis of Associations and Group Differences**

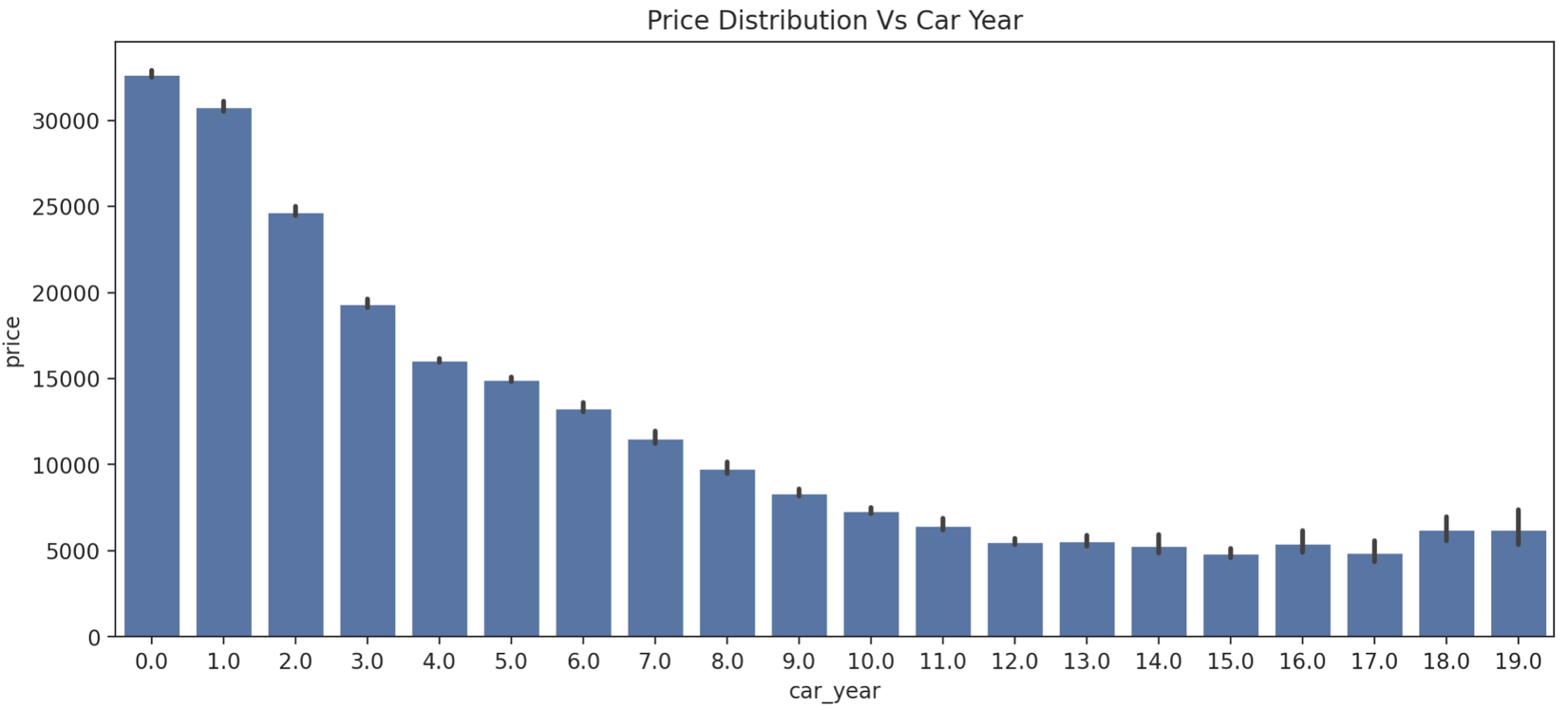
**3.1 Quantitative - Quantitative**

Pearson correlation: to assess the strength and direction of the linear relationship between two quantitative variables which are mileage and price.

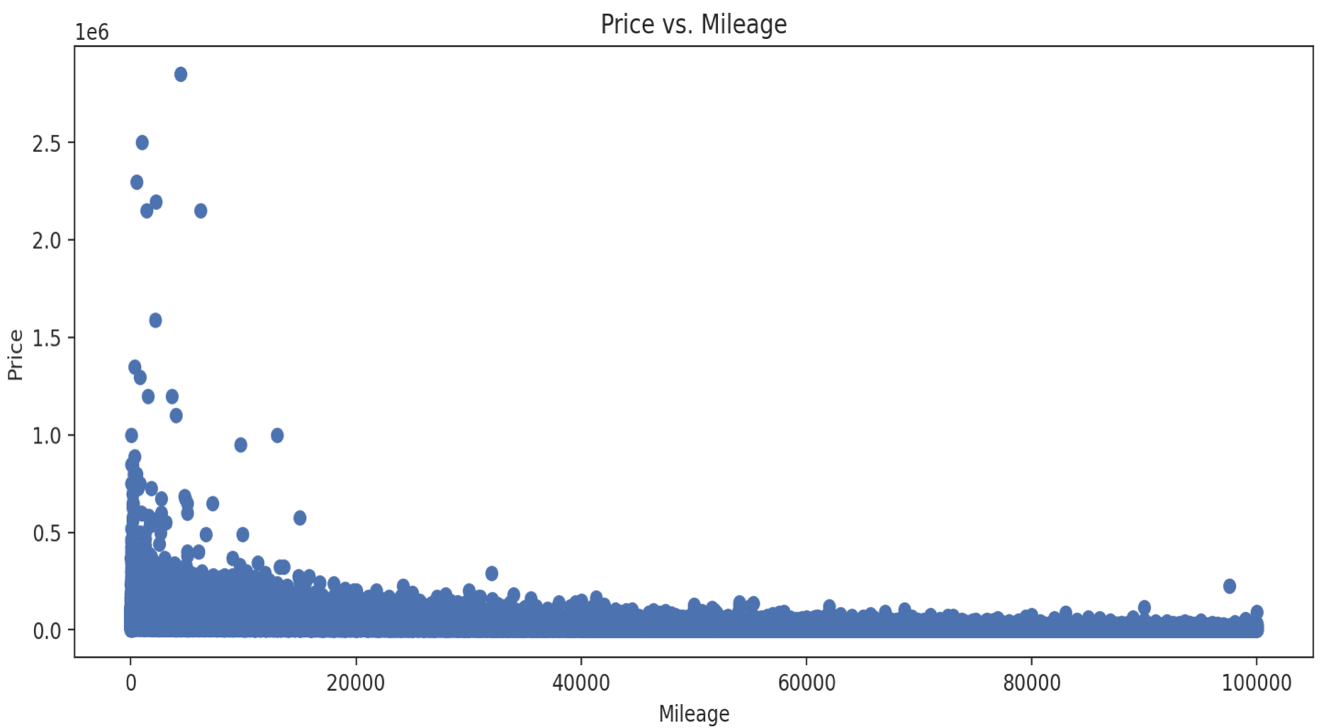


The value -0.330 shows a weak negative linear relationship between mileage and price. This shows that an increase in mileage will result to a corresponding decrease in price. The strength of this relationship is relatively weak and the value -0.330 measures the extent of this weak negative correlation.

Bar plot:



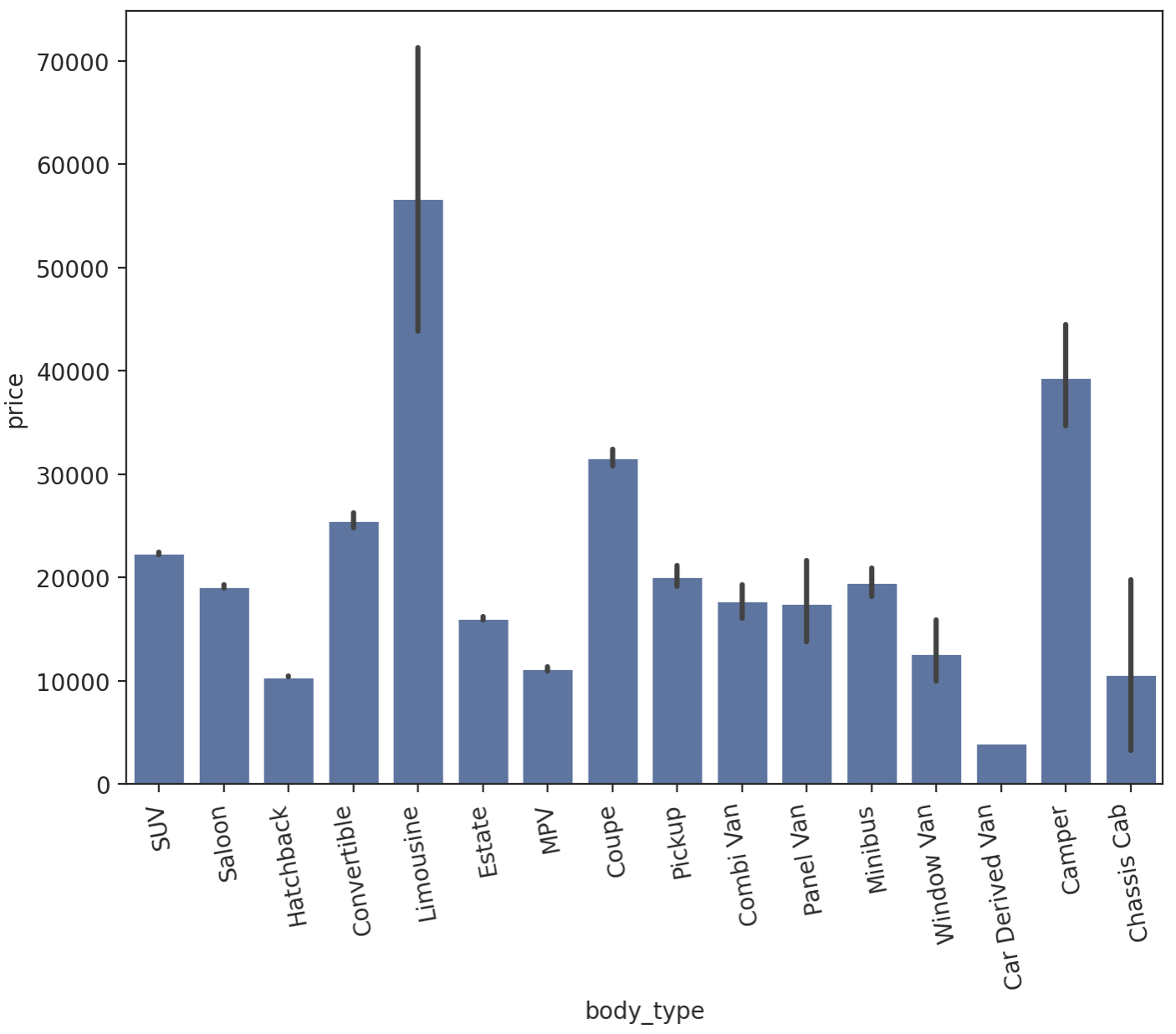
From the bar plot we can see an increase in car prices for cars produced in recent year while the is a continuous decrease for car prices for older years. Also, for the scatterplot, prices of new cars and cars with low mileage are higher than used cars or cars with high mileage.



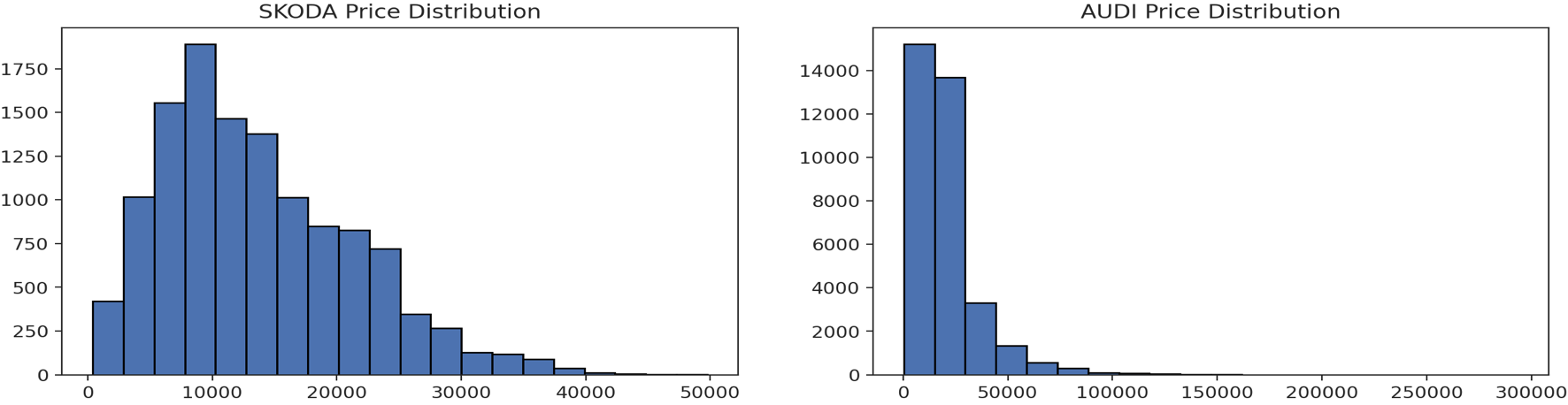
**3.2 QUANTITATIVE – CATEGORICAL**

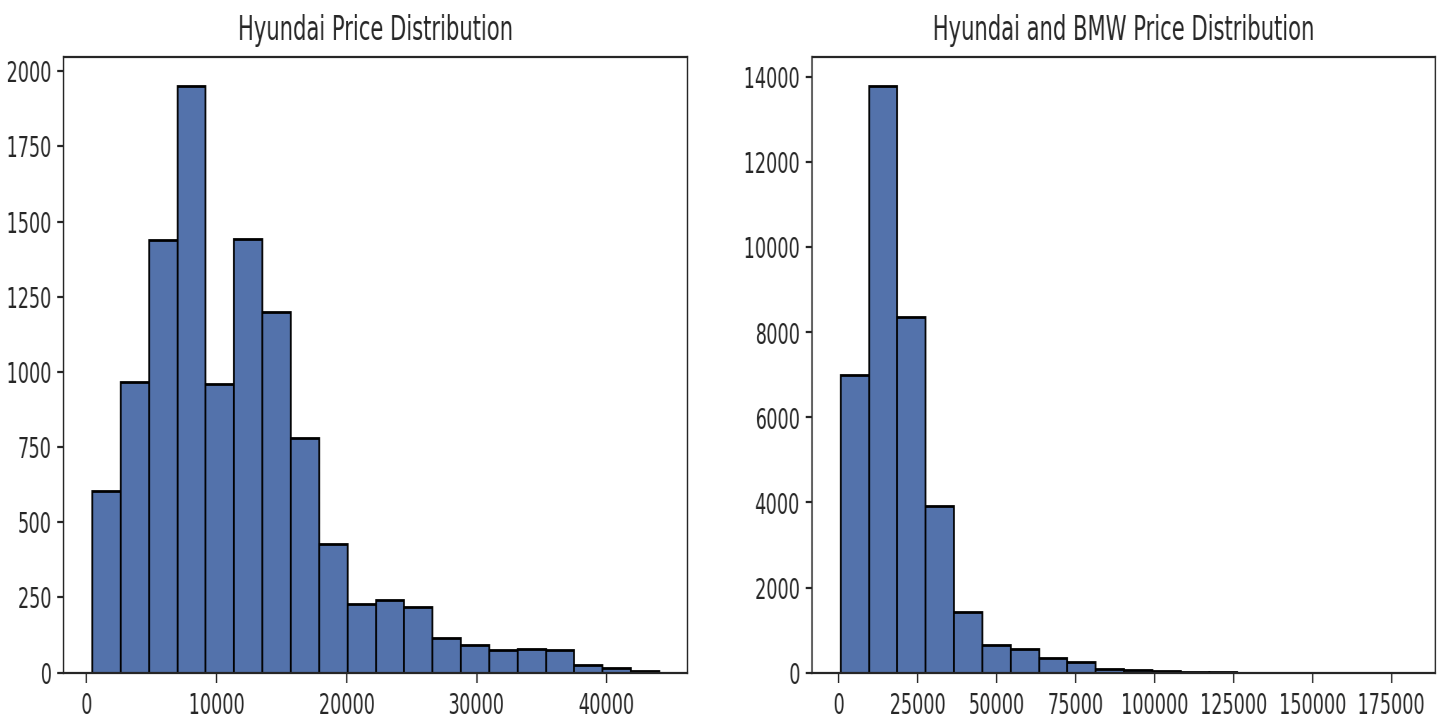
Relationship between the car body type and price:

We can see from the bar plot below that the price is also determined by the body type, the price of a Coupe or Limousine cannot be compared to that of a hatchback. This shows distribution of car body types and how prices vary amongst them.

****

Also, taking a subset for categorical feature like the standard make to check price distribution, we can see from the chart how prices differ for different car makes. Audi and BMW cars are more priced than SKODA.

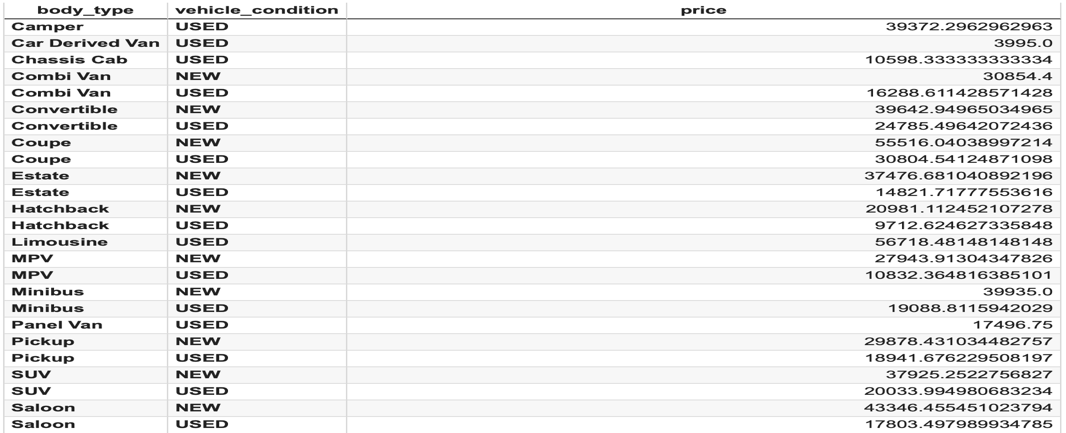
****

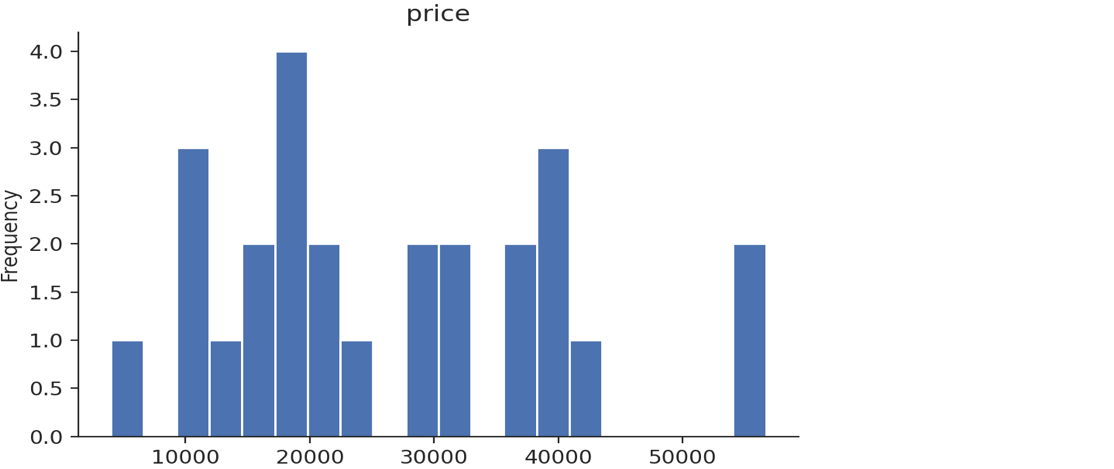
****

**3.3 Categorical – Categorical**

Pivot Table: Using Pivot tables to show the average price for features like body\_type, fuel\_type, vehicle\_condition and how they affect price.

Body\_type and Vehicle\_condition:





Body\_type and Fuel\_type:

