

**College: Engineering and Information Technology**

**Department: Information Technology**

**Program: Data Analytics**

**Data Visualization Course Project**

# **Ford UK Used Cars Analysis & Visualization**

**Prepared by:**

**202211479 Abdullah Kheshfa**

**202110251 Abdallah khaled,**

**Habeeb Mohammad Khaleel Habeeb**

**202011575,**

**Fathy Al Ghandoor 202211770,**

**Abdul rahman al husseini 202211325**

**Supervised by:**

**Dr. Salam fraihat**

**Academic Year 2024- 2025 – Spring**

The goal of your project is to use the maximum number of visualization charts (seen in course or new) to explore and analyse a list of datasets for useful insight (For example Anomalies Classification, Sales analysis…Etc).

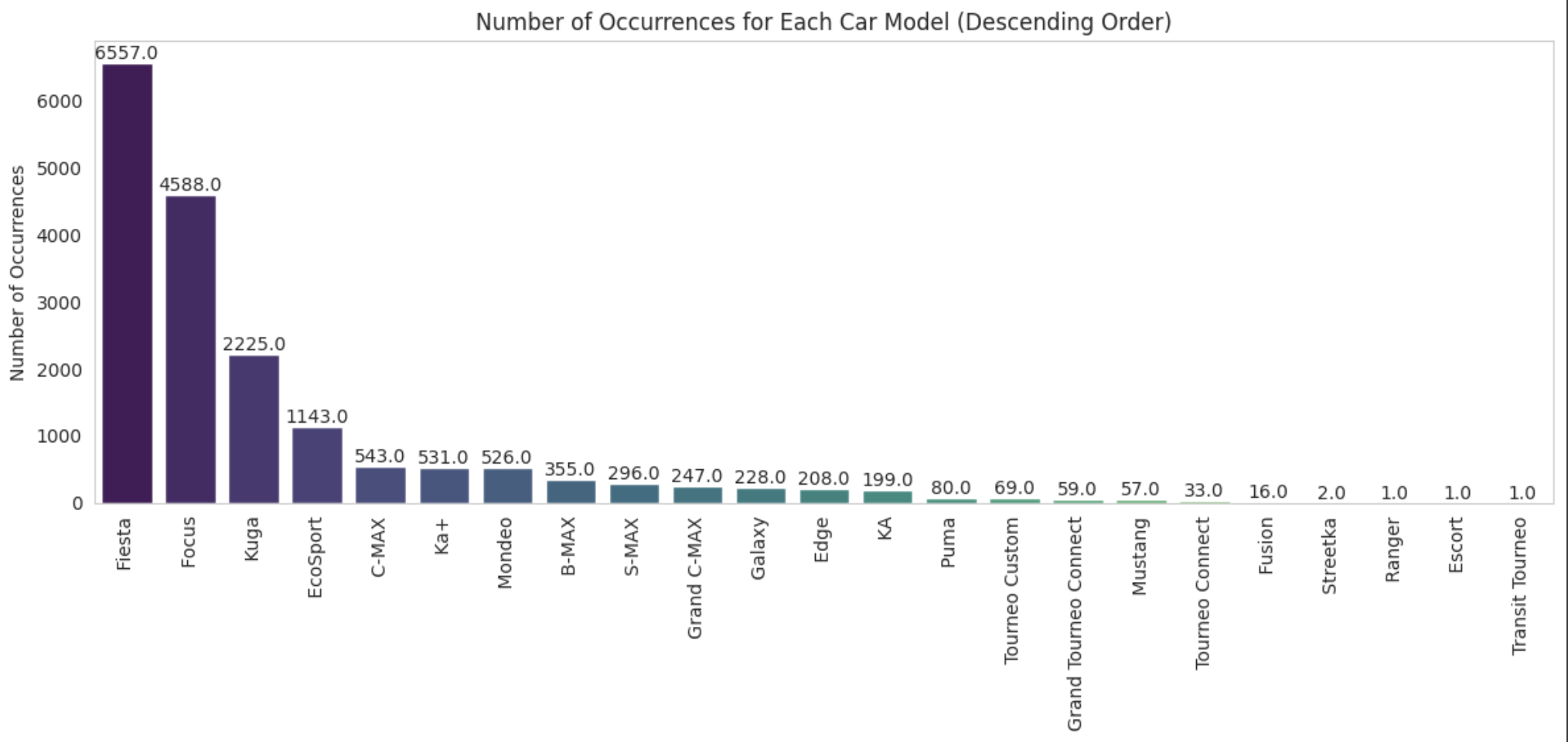
In your project you should use the appropriate charts to:

1. Data understanding (CLO A, B) – What data do we have / need? Is it clean?...
2. Data preparation (CLO C)– How do we organize the data for modeling?
3. Modeling – apply any machine learning model for classification/ Regression/ Clustering.
4. Evaluation (CLO D) – using charts, which model give best accuracy?

| **Criteria** | **Weight** | **Description** |
| --- | --- | --- |
| **Report Format, and Language** | **2** | Respected correct format (section numbering, page numbering, figures), Correct Language, grammar, code well commented. **(Writing Communication Skills)** |
| **Presentation** | **2** | speaking, tone of voice, body language, creativity, and delivery **(Oral Communication Skills)** |
| **Teamwork skills** | **2** | Works well with others |
| **Problem** | **1** | Correct case study Problem |
| **Solution Design** | **2** | Correct and complete solution design **(Problem solving skills)** |
| **Solution Implementation** | **4** | Correct implementation of the design **(Technical skills)** |
| **Results interpretation** | **2** | Correct interpretation and discussion of the results **(Analytical skills)** |
| **Total** | **15** |  |

**Data Understanding(Abdullah Khalid And Fathy):**

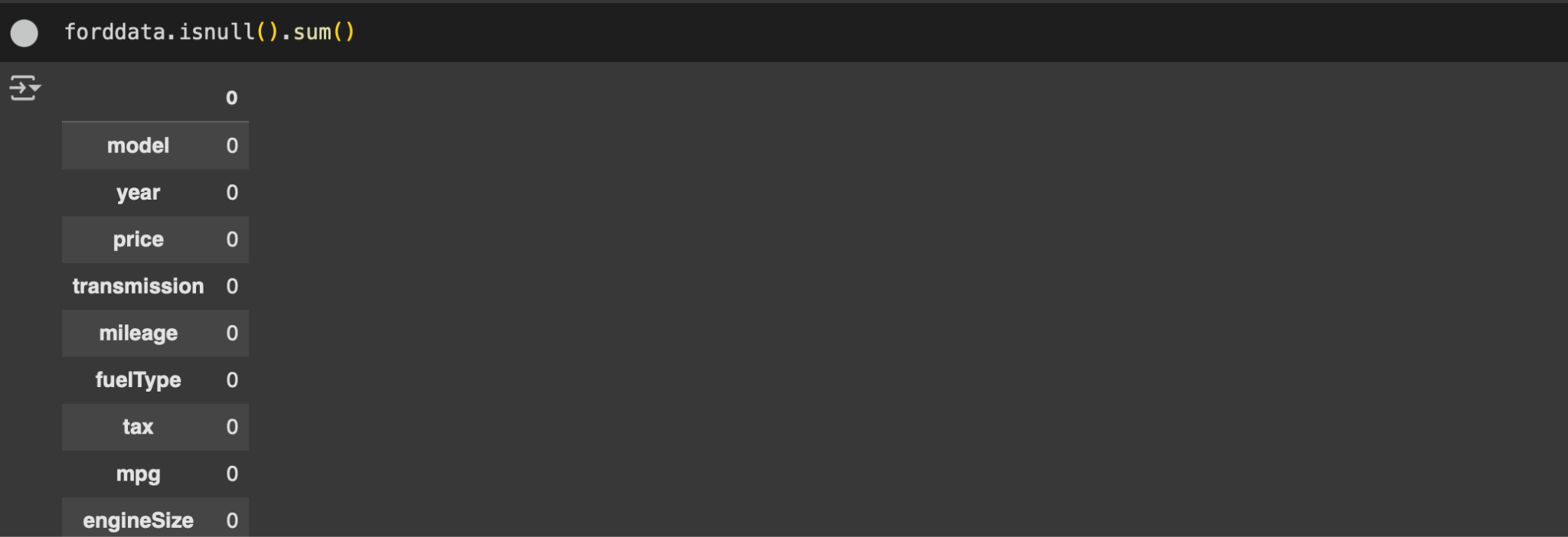
Data understanding is a very important part of data analysis,in this project the dataset we are working with consists of various attributes related to Ford car listings, including mileage, price, year, engine size, fuel type, transmission, and model. These variables allow us to explore market trends of used cars in the United Kingdom, evaluate pricing strategies, and assess vehicle performance characteristics. We’ve begun assessing data quality by identifying missing values, detecting outliers, and evaluating the consistency of data types. We started by gathering information about the distribution in the dataset as shown below.

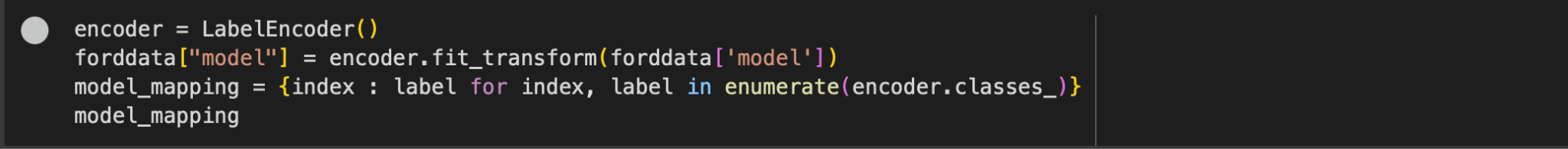
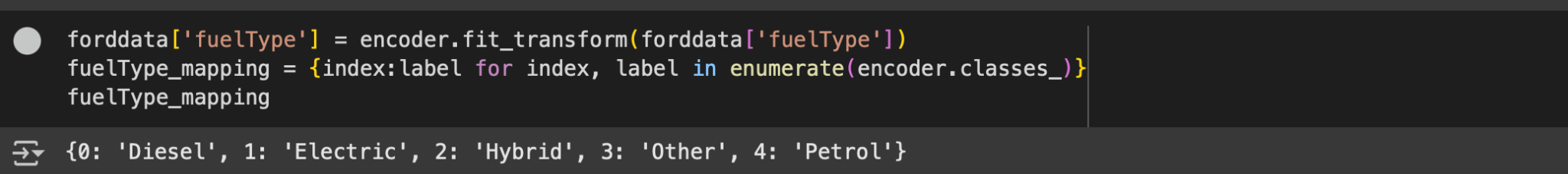


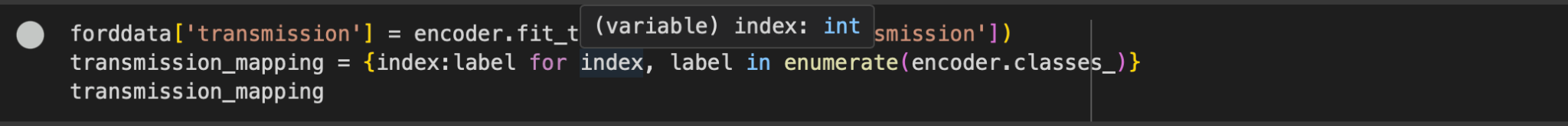
During the data understanding phase, we noticed that some features like mileage and price contained outliers. We used the Interquartile Range (IQR) method to remove outliers from the mileage column due to its skewed distribution, while we applied Z-score filtering on the price column, which was more normally distributed. Furthermore, we explored missing values and null entries to ensure no critical information gaps would affect our analysis.



We also made that the data is clean of null values in all features to prevent a low performing model.

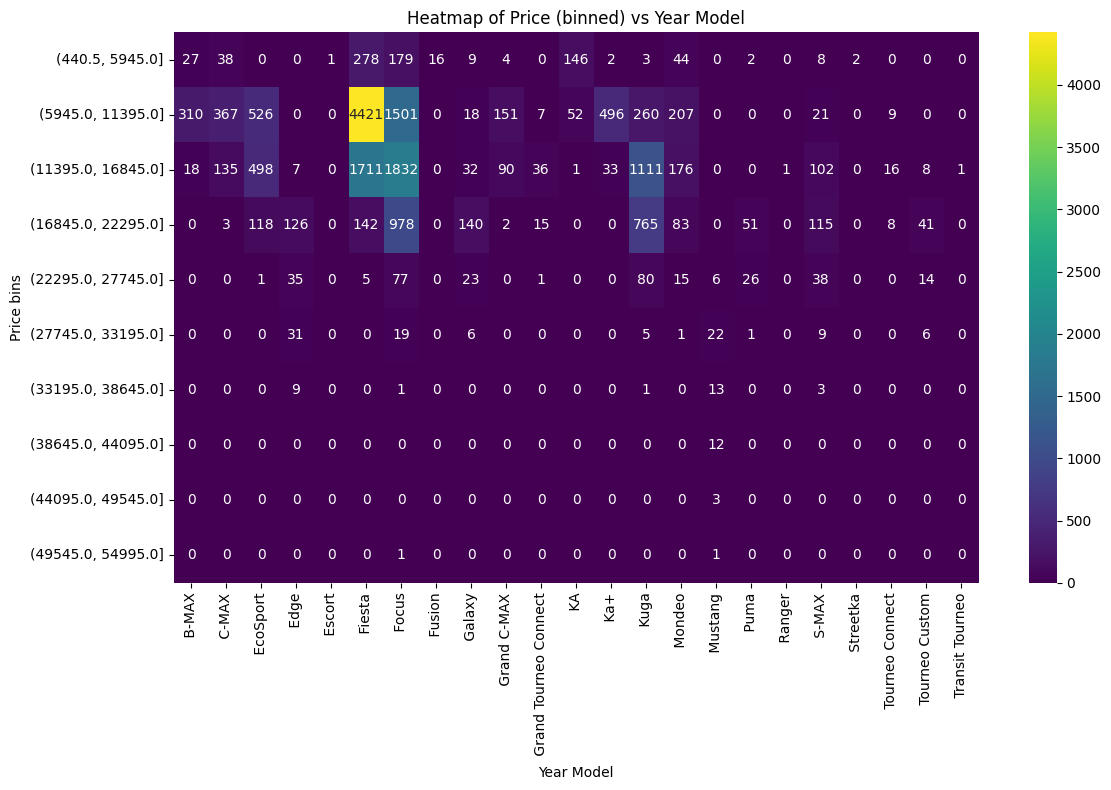


For categorical data, such as Transmission and Fuel type, we performed label encoding to make the data suitable for machine learning models. Additionally, we verified data types and ensured consistency across features. After these transformations, we conducted further statistical and visual analysis to confirm the integrity and usability of the cleaned dataset.

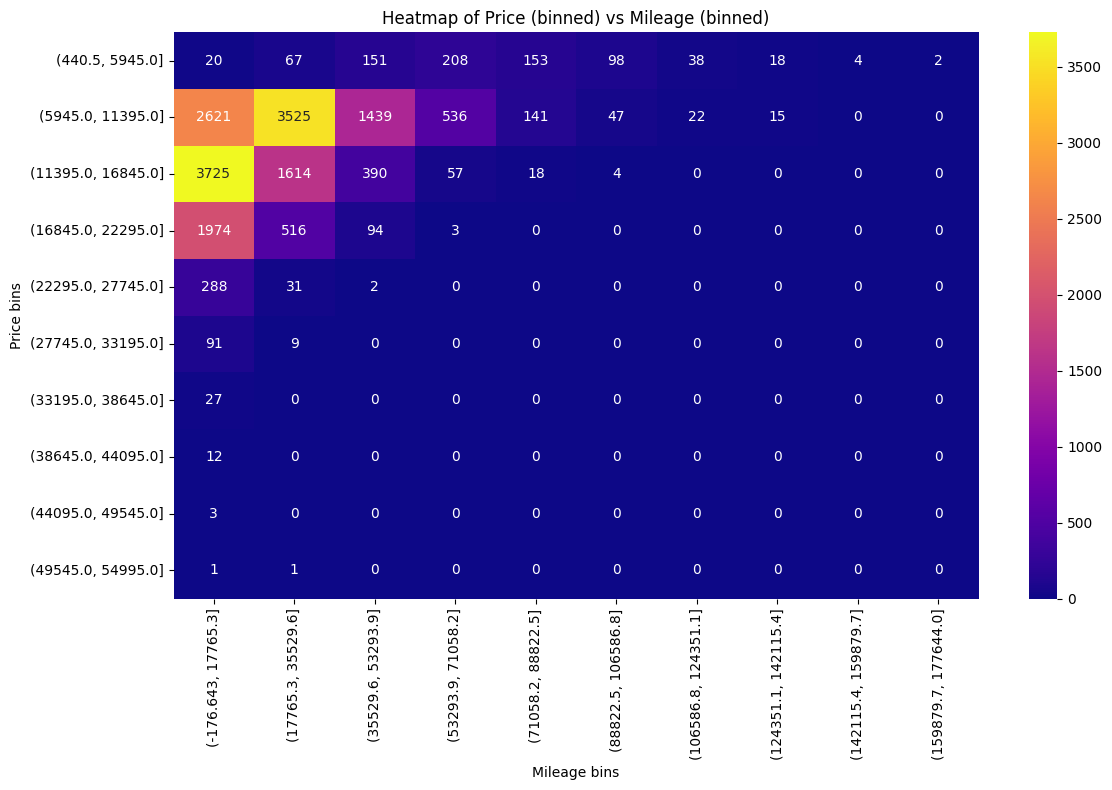


We also conducted a correlation analysis to explore the relationships between key numerical features in the dataset. Using a correlation heatmap, we identified how different variables such as mileage, year, engine size, and MPG are associated with car price. Notably, we observed a negative correlation between mileage and price, indicating that as a car’s mileage increases, its market value tends to decrease — a logical relationship reflecting wear and Depreciation.

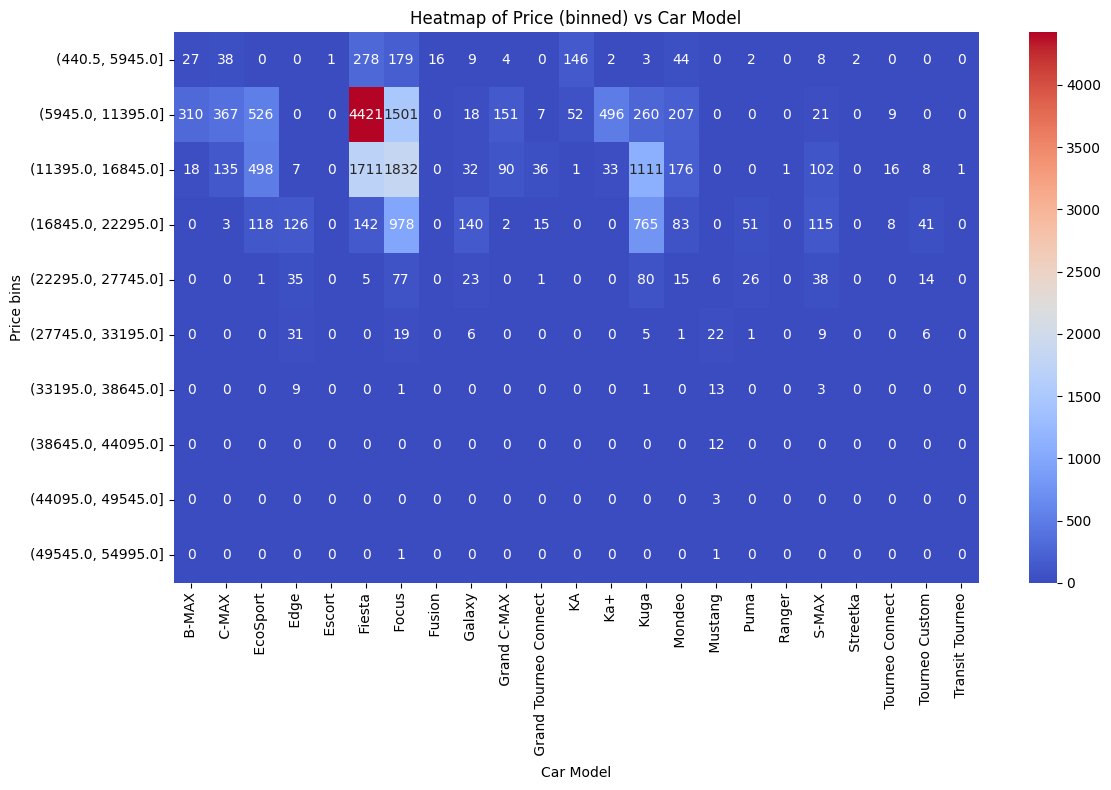
This heatmap visualizes the distribution of Ford used car listings by grouping prices into bins on the y-axis and model years on the x-axis, with the color intensity representing the count of listings in each combination. The heatmap reveals that most listings concentrate in mid-price ranges, aligning with mid-range car models, while newer model years tend to cluster in higher price bins and older models in lower price bins. This pattern not only highlights typical market trends and depreciation effects but also helps identify market hotspots, offering valuable insights into pricing strategies and demand segmentation in the used car market.



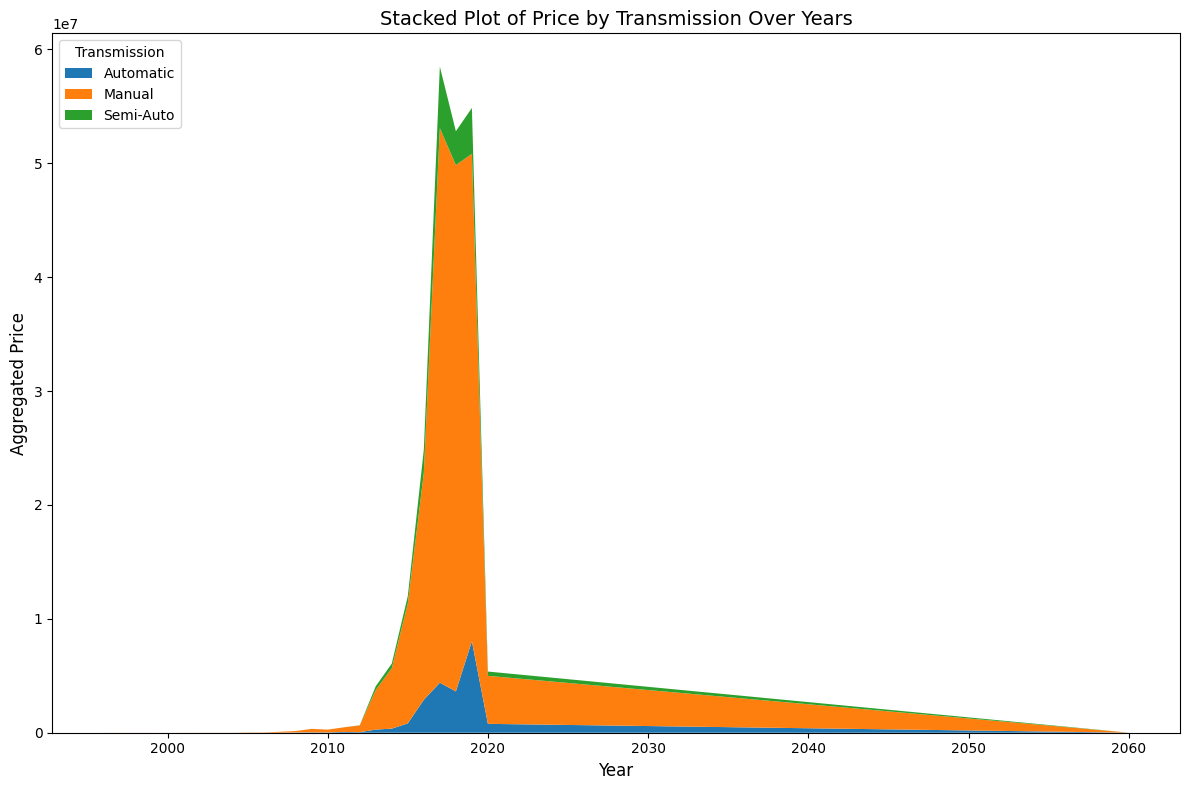
This heatmap displays how used Ford car listings are distributed across binned mileage (x-axis) and binned price (y-axis), with each cell’s color intensity reflecting the count of listings in that specific mileage-price range. The visualization reveals a notable clustering of mid-range mileage vehicles in moderate price bins, indicating a common balance between usage and value. Conversely, higher mileage cars often fall into the lower price bins, showcasing depreciation’s effect on market desirability. Overall, this heatmap underscores the interplay between mileage and price, helping buyers and sellers evaluate how vehicle usage impacts market value.



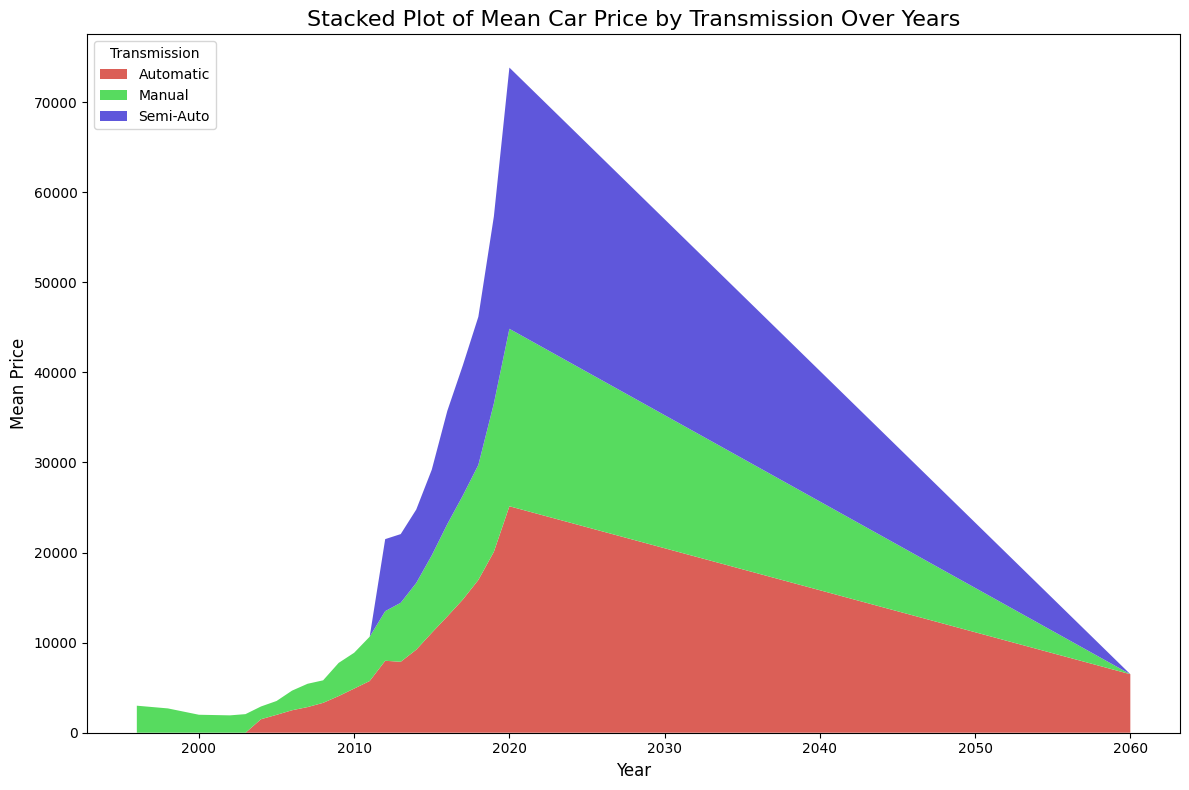
This heatmap presents binned price ranges on the y-axis against different Ford car models on the x-axis, with color intensity indicating the count of listings in each model-price combination. As seen, certain models dominate the higher price bins, suggesting they retain more market value or feature premium characteristics, while others cluster in the lower bins, highlighting affordability or older model variations. By revealing which models are most frequent in each price bracket, this visualization provides useful insights into market positioning and helps buyers or sellers understand how specific models align with certain price ranges.



This stacked plot illustrates the aggregated prices of used Ford cars over time, broken down by transmission type (Automatic, Manual, Semi-Auto). The notable peak around the late 2010s suggests a significant concentration of listings or higher total value during that period, while the distribution of stacked segments reveals which transmission categories contribute most to overall pricing. Manual transmissions appear to comprise the largest share in terms of total price, followed by Automatic and then Semi-Auto. Post-2020, the aggregated prices taper off, potentially reflecting fewer listings or shifts in market demand. Overall, this visualization highlights how transmission types vary in total value across different production years.



This stacked area chart shows how the **mean price** of used Ford cars evolves over the years for each transmission type (Automatic, Manual, Semi-Auto). From around the early 2000s to the late 2010s, all three categories generally rise, with **Semi-Auto** exhibiting the highest average prices, followed by **Manual** and then **Automatic**. After peaking, the chart tapers off—likely reflecting fewer listings beyond those model years. By illustrating how the typical selling price varies by transmission type over time, this visualization highlights differences in market demand and the enduring value of certain transmission categories.



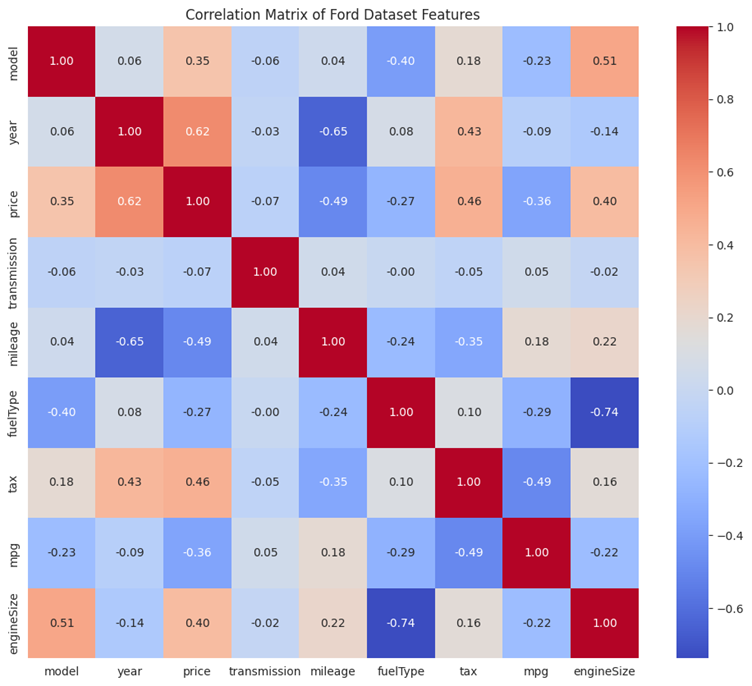
Both charts visualize how used Ford car prices vary over time by transmission type, but **they measure price differently**. In the **first chart**, the y-axis shows the **mean (average) price**, revealing the typical value of cars each year. This view indicates that Semi-Auto cars reach the highest average prices, followed by Manual and then Automatic, suggesting Semi-Auto listings generally cost more per car. In contrast, the **second chart** uses the **summed (aggregated) price**, illustrating the total monetary value across all cars in each year-transmission group. Here, Manual transmissions dominate—reflecting not necessarily a higher individual price, but rather a **larger overall volume** of Manual cars that inflates the total. Consequently, the **mean plot** highlights which transmission types are more expensive on average, while the **sum plot** emphasizes how many cars (and thus total market value) each transmission type contributes in a given year.

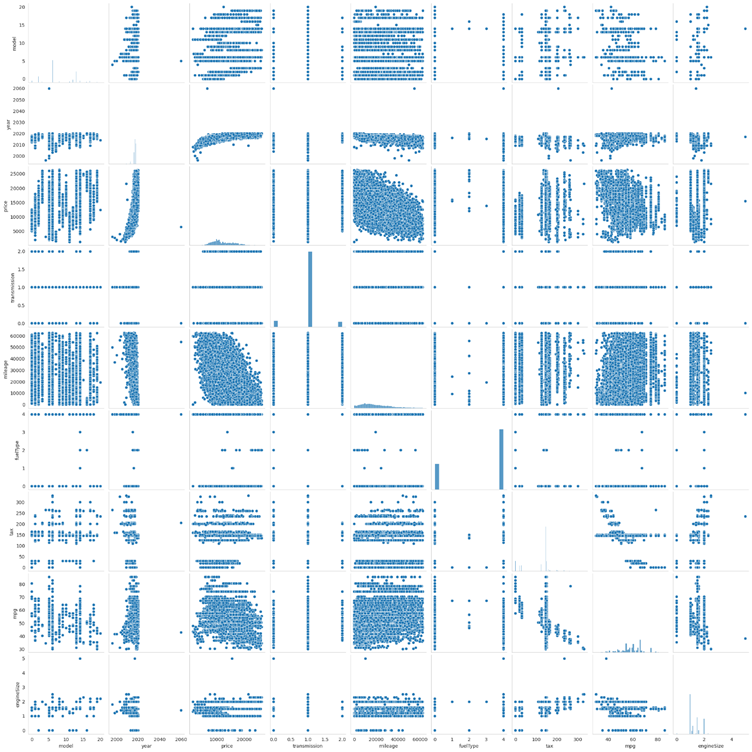


***Machine Learning(Abdullah Kheshfa):***

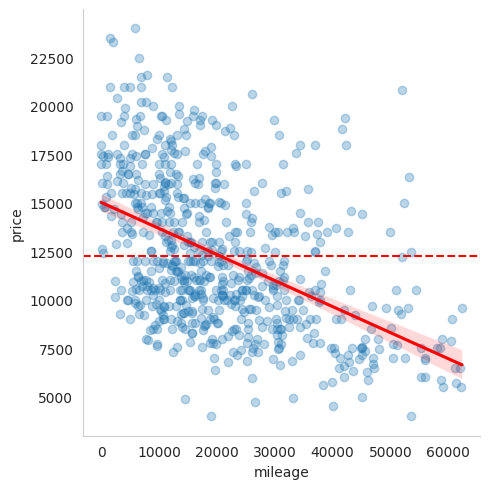
**Visualising the relations:**

Modeling: I used linear regression to predict the price of the car. I built different models on different number of features and I used the correlation matrix with the pair plot to understand what features affect the price, the feature with the strongest impact on price is mileage so we use scatterplot to understand the relation in a bigger image.

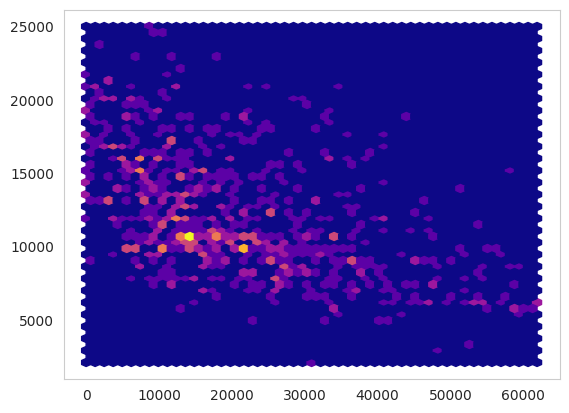




We can see there’s a negative relation between the two but the data is too clustered, so we use a sample to make scatter more visually appealer I took a sample of the population , plotted the mean and the regression line:



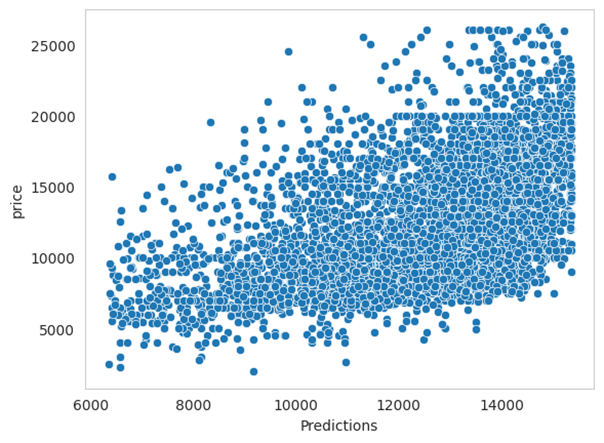
I also used the hexbin to observe the distribution more clearly:



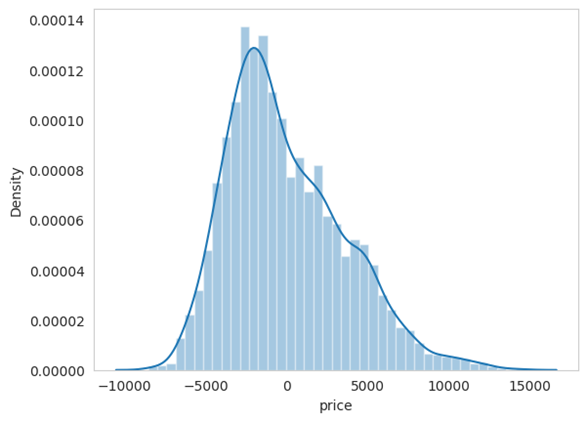
**Modeling:**

1st model:

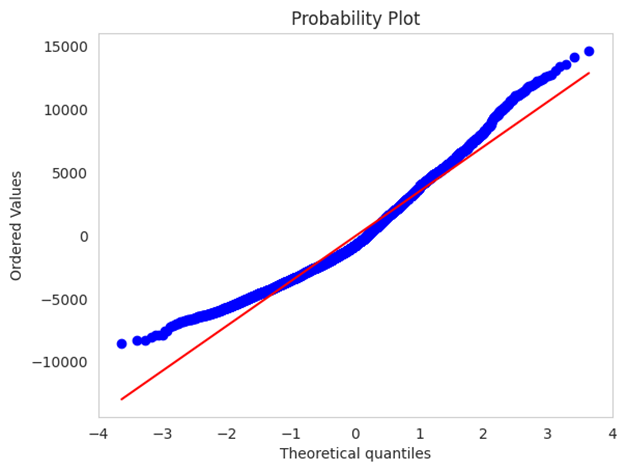
The first model gave me a accuracy of 23% using the r2-score which is low and even when observing the predictions and the actual values we can see a pattern but we can also observe how a lot of datapoints are not following this pattern.



observing our residuals distribution we find that its right skewed and has a mean of almost negative 5000 which tells us that our model predictions and actual values have large differences between them which also tells us that our model isn’t fully capturing the pattern and our predictions are incorrect, and that indicates that our model is bad.



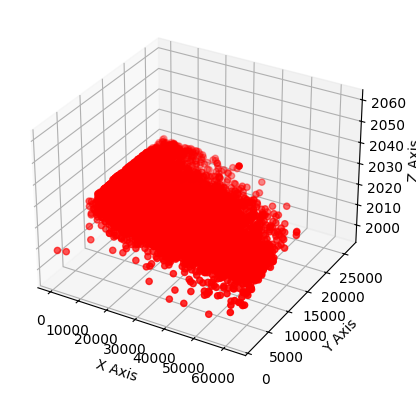
Also the Q-Q plot shows us how the residuals are not fitting on the line perfectly which also confirms that there’s a skewed distribution and both tails are heavy due to outliers. The axis indicates the quantiles and the y axis the ordered values(residuals).



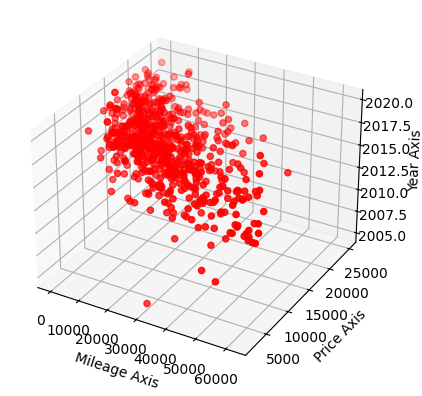
2nd model:

I added the year column which could help us find a better accuracy since if we go back to the corr matrix we would find that it has a strong relation with the price. We can in the 3D plot that there’s a relation between the 3.

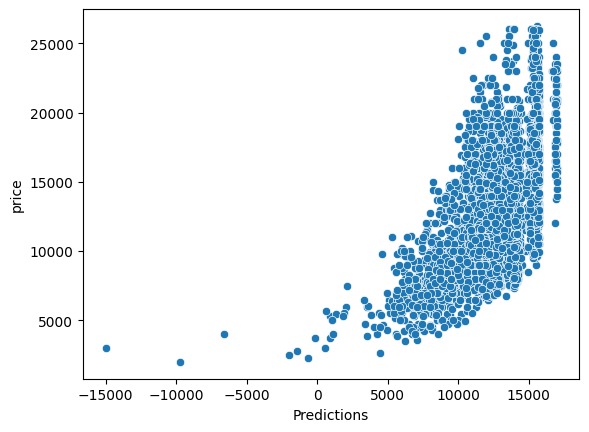
The population:



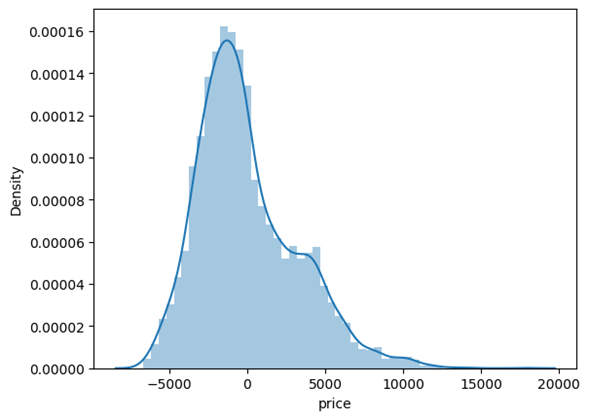
The sample:



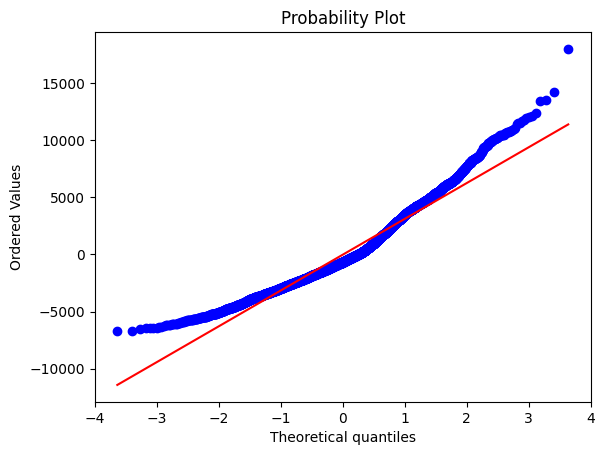
We can also see the predictions and the actual values have a pattern similar to the previous but they are more cluttered than before which is good, but the pattern is diagonal which we are trying to avoid.



The residual distribution is still skewed but the mean is now closer to 0 than before:



The Q-Q plot also tells us that the distribution is close to normal since the datapoints are more fitted on the line than before and the tails aren’t heavy as before:

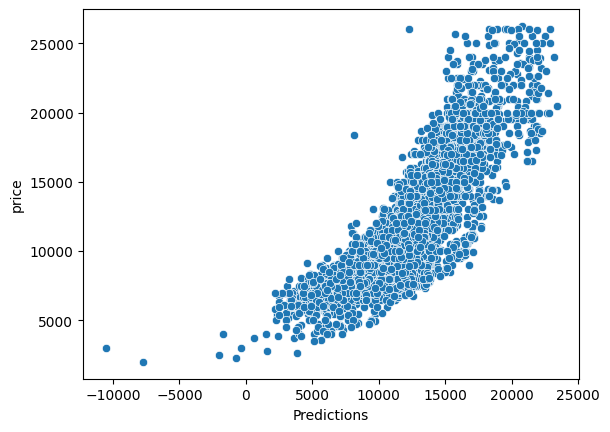


Overall our model had a better accuracy than before which is 40%, but we still need to try to achieve a better accuracy.

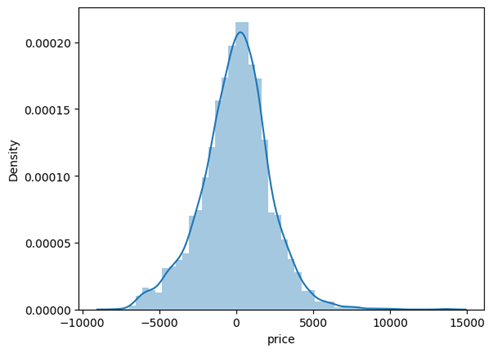
3rd model:

This time we will select all the features(because they are all numbers which the machine can read) and this time we will applying MinMax scaling.

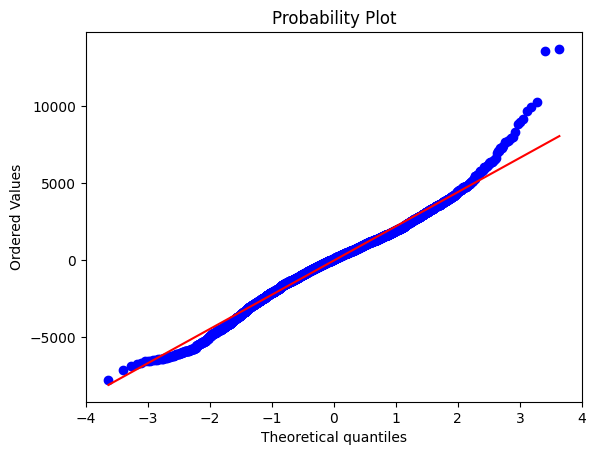
Our predictions and actual values have a more clear pattern than before and is more vertical than before and more cluttered. This indicates that our model is capturing the price in a very good manner:



Our residuals distribution also follows a normal distribution with very small skewness but still better than the previous models:



Also The Q-Q data points are more fitted than before which indicates a normal distribution with a heavy tail on the upper end:



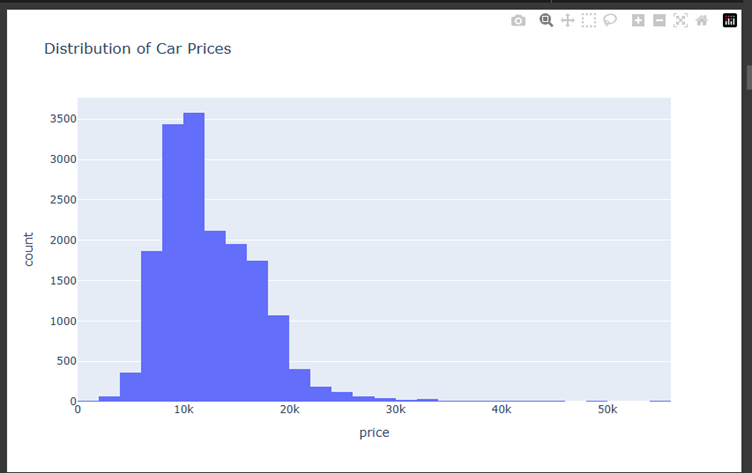
We got a accuracy of 71%

4th model:

For the last model I tried filtering the rows by a certain years which is all years above and equal to 2000, but our accuracy didn’t change much, and the visualizations aren’t much different.

We used 4 types of graphs to asses our model: scatterplots(to find the relation), correlation matrix(to understand the correlation between different features), scatterplot(to see if our data is capturing the pattern, we can understand this by seeing if our predictions and actual values are going in the same direction in a diagonal way, residual distribution(helps us understand the distribution of our residuals and the mean of them which can determine if the model is accurate or note) and finally q-q plot(gives us better detection and visualization of deviations in the tails of the distribution than the residuals distplot(histogram).

***Interactive visualizations(Habeeb and Abdul Rahman al-hussieni):***



The dataset's distribution of automobile prices is depicted by this histogram. The y-axis displays the quantity of automobiles available at each price range, while the x-axis displays the car costs (in GBP).

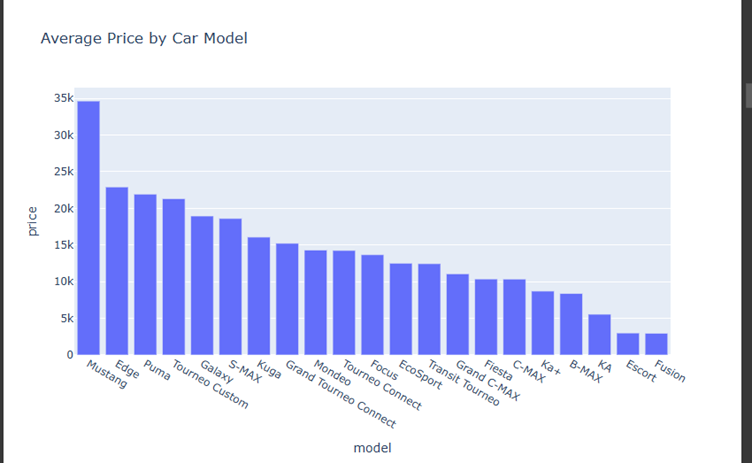
With a distinct peak in the center of the chart, most automobiles are priced between £8,000 and £15,000.

Because of the right-skewed distribution, fewer automobiles in the higher price categories (over £25,000) are available.

With fewer luxury or high-end models, this trend suggests that the majority of the automobiles in the sample are within the mid-range affordability range.

For modeling, determining this distribution is essential since it provides insight into the behavior of the target variable and guides any transformation or normalization tactics.

Price clusters, market emphasis, and possible outliers may all be found using this visual aid.

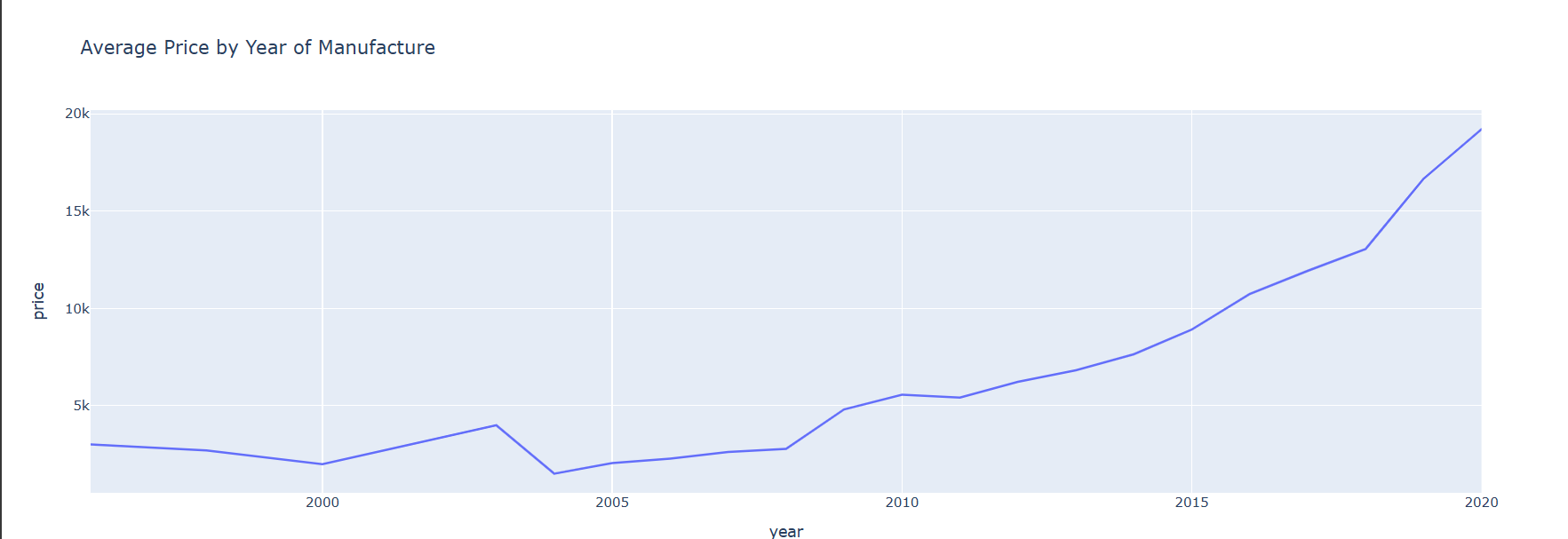


This bar chart displays the average price of Ford cars based on their model. It is clear that the Mustang model has the highest average price, reaching around £35,000. Models such as Edge, Puma, and Tourneo Custom also exhibit high prices. On the other hand, models like Fusion and Escort show significantly lower average prices. This visualization highlights the importance of the "model" feature in price prediction.

A graph showing a number of fuel prices

AI-generated content may be incorrect.

This scatter plot explores the relationship between mileage and price, categorized by fuel type. A clear negative correlation is observed: as mileage increases, price generally decreases. Petrol and Diesel vehicles dominate the dataset, while Hybrid and Electric vehicles are relatively scarce. The chart shows how different fuel types affect price trends across mileage ranges.



The average cost of vehicles by year of manufacture is shown in this line graph. The average price shows a steady growth from 2010 to 2020. The graph shows a forecasted pattern of a slow drop after 2020. According to this pattern, newer automobiles are more expensive, and over time, vehicle depreciation significantly reduces value.A screenshot of a graph

AI-generated content may be incorrect.

This box plot compares the distribution of car prices across different transmission types. Automatic and Semi-Auto vehicles generally have higher price medians compared to Manual ones. The presence of many outliers, especially in Automatic transmissions, indicates variability in high-end vehicle pricing.

A pie chart with a red and blue circle

AI-generated content may be incorrect.

The distribution of automobiles by fuel type is displayed in this pie chart. The bulk of automobiles (69.3%) are fueled by gasoline, with diesel coming in second (30.6%). The low adoption rate of alternative fuel cars in this sample is highlighted by the fact that hybrid, electric, and other fuel types together make up less than 1% of the dataset.

A graph of a graph

AI-generated content may be incorrect.

This scatter plot shows the relationship between engine size and fuel efficiency (measured in miles per gallon). It reveals that larger engine sizes generally correspond to lower fuel efficiency. Hybrid cars show notably high mpg values, indicating better fuel performance. The chart reinforces the trade-off between engine performance and fuel consumption.

A graph of a number of blue and white bars

AI-generated content may be incorrect.

This bar chart presents the ten most common Ford models in the dataset. The Fiesta model leads by a significant margin, followed by Focus and Kuga. These popular models reflect market demand and could influence supply-based pricing patterns.

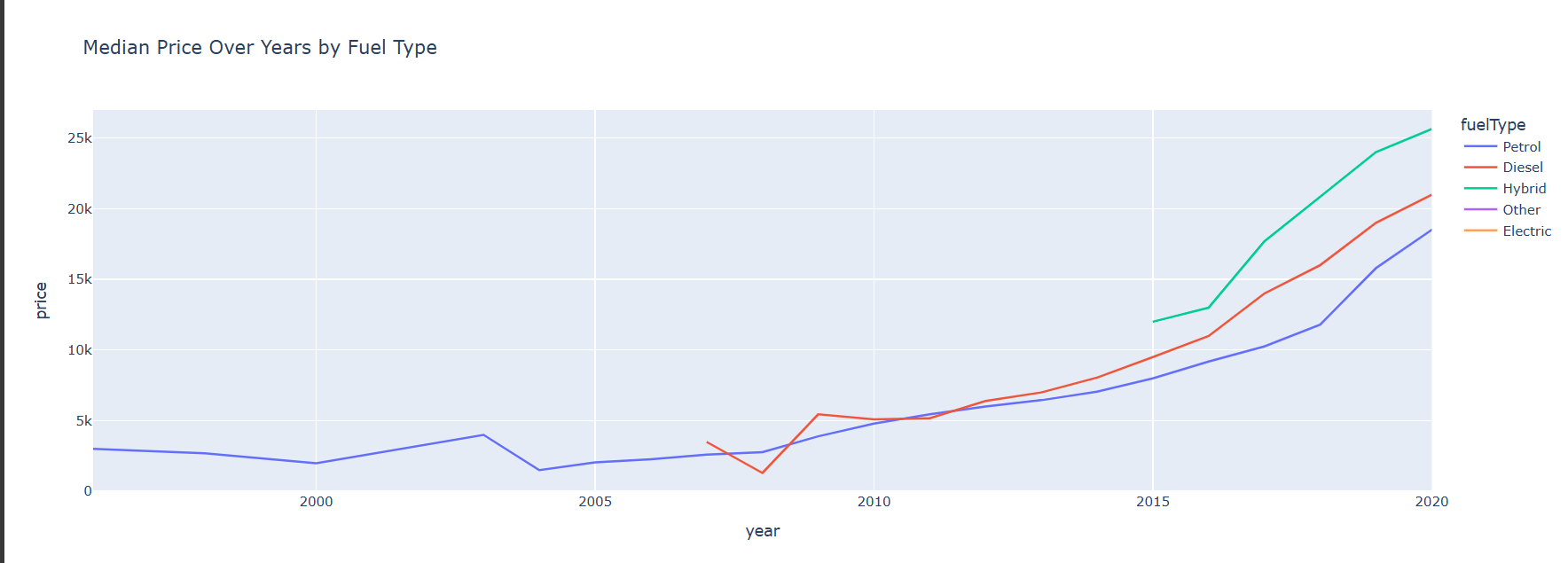
A screenshot of a graph

AI-generated content may be incorrect.

This box plot compares fuel efficiency (mpg) across fuel types. Hybrid vehicles show the highest and most variable mpg values, while Diesel performs slightly better than Petrol. The data for Electric and Other types appears limited and less informative due to small sample sizes.

A graph of different colored dots

AI-generated content may be incorrect.This multifaceted scatter plot illustrates how the relationship between mileage and price differs by transmission type. All three panels (Automatic, Manual, and Semi-Auto) reflect a similar trend: price decreases as mileage increases. However, Automatic vehicles appear to maintain higher price levels at comparable mileage ranges.



This line chart shows the median car prices over the years for different fuel types. Hybrid cars show the steepest rise in median price, indicating increased market value. Diesel and Petrol follow a similar upward trend until 2020, after which a slight decline appears. The chart emphasizes how fuel type impacts pricing trends ov

***Static Visualizations(Fathy Al ghandoor):***