FINAL REPORT OF AUTOPRICING INTELLEGENCE

DSBA – Aug 22’ Group 4

**Mentor** : Jainesh Garg

**Team** : Aditiya Raman

Ajai Subramanian K

Hassain B

Rohith VK

Vignesh PB

**Table of Contents**

[Introduction 4](#_Tocijkcfkifuzw5)

[Project 4](#_Toc23edlrmt75cb)

[Industry 4](#_Tocafqkytay03ux)

[Company 4](#_Toc5mrqce6sgo7w)

[Need for the present study 5](#_Tocifw2zxpsa0om)

[Core problem statement and Objective 5](#_Tocnqgeyft8kber)

[Core Problem Statement: 5](#_Toczunnyjlvzize)

[Objectives of the Project: 5](#_Tocse1l0erwhaw7)

[Data source and Description 6](#_Tocjybih78wm5yq)

[Data source 6](#_Tocnugr7u4rlen8)

[Number of rows and columns: 6](#_Toc8g15lj5cansc)

[Exploratory Data Analysis 6](#_Toc3jcdl3fx8340)

[Understanding my Data 6](#_Toc4u5nrl5yajtu)

[Data Info 7](#_Toc3gj9sgvghc3h)

[Cleaning Data 7](#_Toctfcoutspf8i7)

[Filling nulls with 'unknown' 8](#_Tocb8oqei1pyu21)

[Distribution Plots 8](#_Tocmgme2b3ilv57)

[Frequency Plots 9](#_Tockzas4h9mn3ad)

[Further Cleaning & Preprocessing 10](#_Toct3sdw55m8qw8)

[Distribution of Horsepower per Car Maker/Country 16](#_Tocflw6ohpsj0df)

[Horizontal bar plot to visualize the number of cars produced by different car manufacturers 19](#_Toc4uy7xekh5s0j)

[Machine Learning Models 20](#_Tocdnpmi35lw29m)

[Decision Tree Regressor 20](#_Tocvgasiswg6n1k)

[Model Comparison 24](#_Toc1f1lnfy4nin8)

[Based on these metrics: 28](#_Tocaxc08dc2e9b5)

[Profit calculation 29](#_Tocwgvptmwetpep)

[Visualization 11](#_Toc1j37atygdm35)

[Distribution of Price per Year 11](#_Toc664m1elzs0zi)

[Distribution of Price per Country 12](#_Tocbln8qpebf131)

[Distribution of Mileage per Car Maker/Country 13](#_Tocdd4zq5wxyopo)

[Cars that are most frequent in the dataset 18](#_Tocxkq42n4x84mm)

[Correlation coefficients of the numeric variables 19](#_Tocwu7yl998z9dq)

[Linear Regression Model 21](#_Tocgpmzq12ppfw7)

[KNN Regressor 22](#_Toc5j0t8xcaq3bp)

[Random Forest Regressor 22](#_Tocydgey8lehocp)

[XgBoost Regressor 23](#_Toc6dmmmt8h5ync)

[Residual Plot 26](#_Tocftieg14begqa)

[Recommendation 30](#_Toc6s2c51osz09u)

### Introduction

Selling a pre-owned vehicle can be a strategic financial decision, and understanding the factors that influence the price is crucial. Every car has a unique value, and as a seller, you want to extract the highest possible price while providing transparency and a fair deal to the buyer. To achieve this, assessing the various determinants that impact the resale price, such as the car's age, mileage, condition, and additional features, is essential.

Sellers also need to consider market dynamics, seasonal trends, and the local demand for specific car models. Leveraging data-driven techniques and predictive models can aid sellers in accurately pricing their vehicles, thereby optimizing their profit margins. By analyzing historical data and market insights, sellers can make informed decisions that benefit both their financial interests and the buyer's expectations. In this context, the ability to predict and understand the resale value of used cars plays a pivotal role in the seller's success and profitability.

### Project

The project aims to develop a predictive model for used car data that can help buyers, sellers, and dealers to estimate the fair value of a used car. The model will use historical data on used car sales, including factors such as the car's age, mileage, make and model, and condition, to predict the market value of a particular used car. Along with the price we calculate the profit based on the selling price, year of manufacture and original car price. Calculated with depreciation factor.

### Industry

The used car market is a significant and growing industry, with millions of used cars sold every year worldwide. The industry is highly competitive, with many players, including individual sellers, dealerships, and online marketplaces. In recent years, there has been a growing demand for accurate and reliable used car valuations, as buyers and sellers seek to make informed decisions about pricing and negotiations.

### Company

The company involved in the project is a technology-driven firm that provides solutions for the automotive industry. The company offers a range of products and services, including online marketplaces, automotive data analytics, and predictive models for various automotive-related applications.

### Need for the present study

The present study is essential to address the challenges faced by buyers and sellers in the used car market. Currently, determining the fair value of a used car is a time-consuming and challenging process, as there are many factors to consider, such as the car's age, mileage, make and model, and condition. By developing a predictive model for used car data, buyers and sellers can quickly estimate the market value of a used car, which can help them make informed decisions about pricing and negotiations. Additionally, the model can benefit dealerships and other automotive businesses by providing them with accurate and reliable valuations for trade-ins and resale vehicles.

### Core problem statement and Objective

Core Problem Statement: The core problem statement for this project is to develop a predictive model for used car data that accurately estimates the fair market value of a particular used car based on historical sales data and various relevant factors such as age, mileage, make and model, and condition. The objective is to provide buyers, sellers, and dealers with an easy-to-use tool for estimating the value of a used car, which can facilitate fair and efficient transactions in the used car market.

#### Objectives of the Project:

Develop a predictive model for used car data that can accurately estimate the market value of a particular used car based on historical sales data and relevant factors such as age, mileage, make and model, and condition.

* Collect and analyze a large dataset of historical used car sales data to train the predictive model.
* Evaluate the performance of the predictive model using various metrics such as accuracy, precision, recall, and F1-score.
* Develop a user-friendly interface for the predictive model that can be easily accessed by buyers, sellers, and dealers.
* Test the model on a range of different make and model vehicles and evaluate its performance in different market conditions.
* Provide recommendations for future improvements to the predictive model based on feedback from users and market trends.
* These objectives will help to achieve the main goal of the project, which is to develop a reliable and accurate tool for estimating the market value of used cars, which can benefit buyers, sellers, and dealers in the used car market.
* Calculated the profit based on the selling price, year of manufacture and original car price.

### Data source and Description

Data source

The source of the dataset is Kaggle. Our goal is to build a model that can accurately predict the price of new used cars based on their features. Although similar models are available for this dataset, our model will be built in such a way that it not only predicts the market price of the used car but also predicts the profit a trader will get for each car that is being sold.

#### Number of rows and columns:

The model will be trained on a dataset of 404026 used cars with known prices and their associated features. There are 404026 rows and 18 columns.

### Exploratory Data Analysis

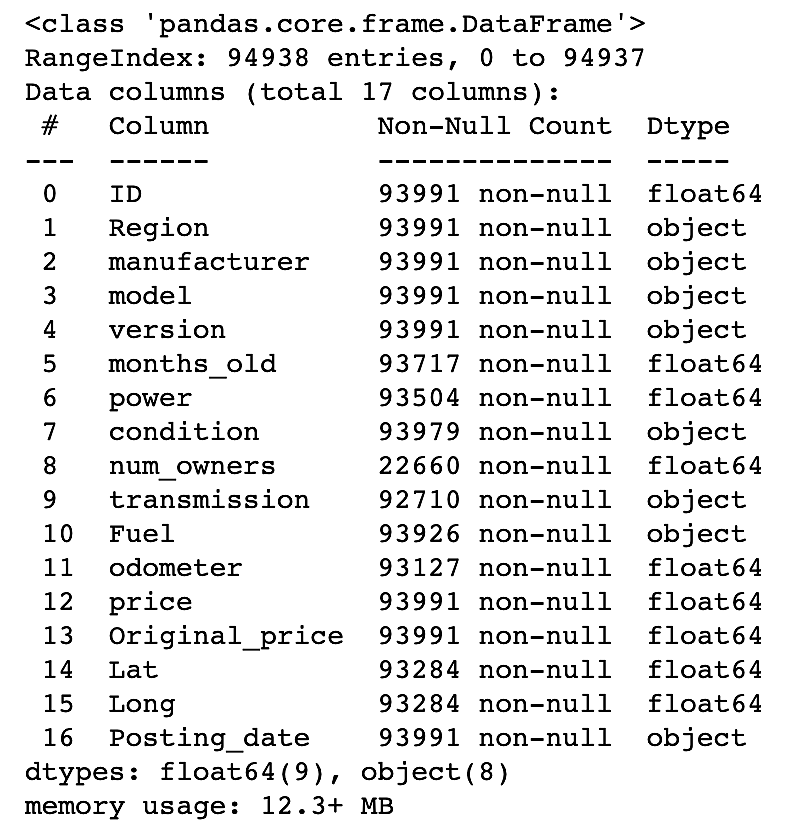
#### Understanding my Data

In this section, we've made several data transformations to enhance its interpretability. Firstly, we converted the "months\_old" column into a more understandable format, creating an "age" column that now represents the model year of the car. For instance, if you're selling a 2006 Honda, the "age" column will display "2006" instead of indicating that the car is 144 months old.

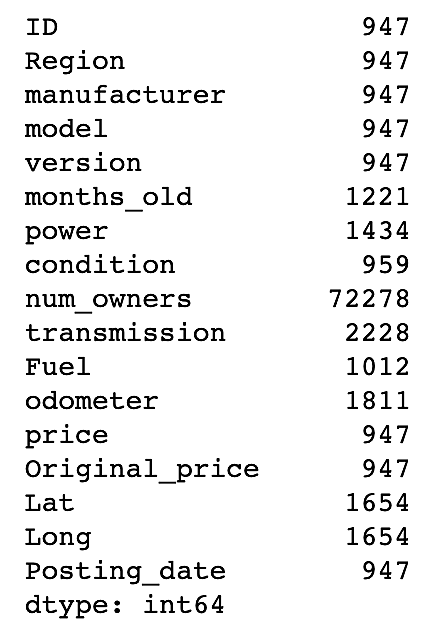
Additionally, we conducted data cleaning by removing outliers. Specifically, we eliminated outliers like brand new cars with zero odometer mileage. It's important to note that we chose not to remove outliers in columns such as price and others. This decision aligns with our project's objective, which is to predict the price and profit of any car, enabling the model to generalize effectively on various types of unseen data.

Following these data transformations, we proceeded to visualize the distributions of the data. This step helped us gain a deeper understanding of the underlying data distribution and guided our approach to feature engineering for building our machine learning models.

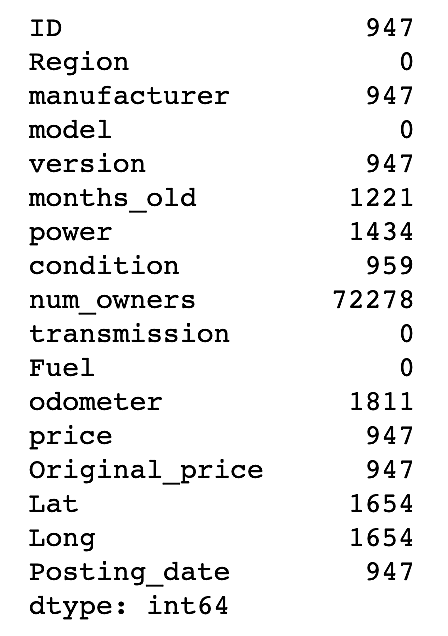
#### Data Info



#### Cleaning Data

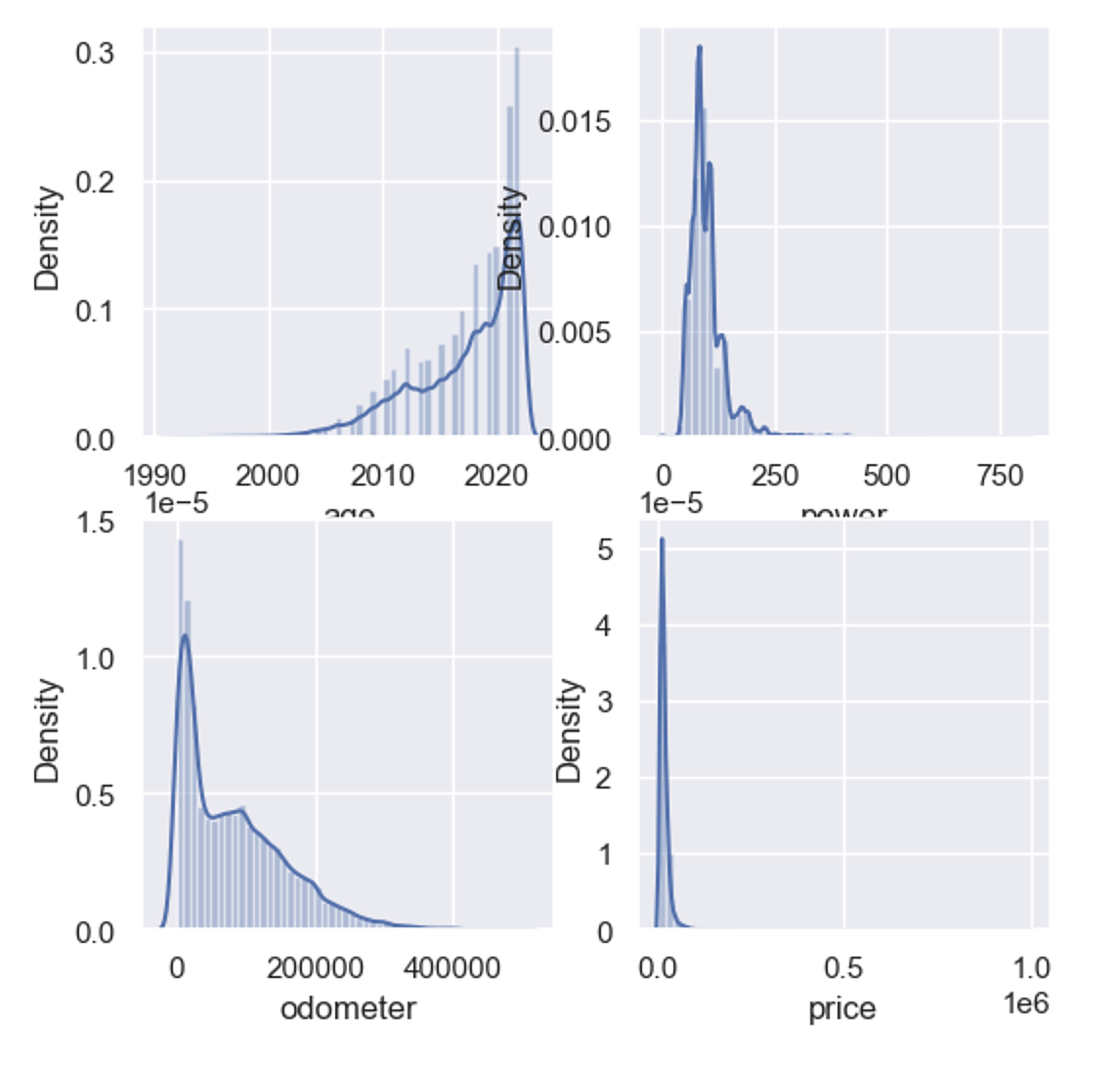
Cleaning out nulls and duplicates from data  
  


#### Filling nulls with 'unknown'



#### Distribution Plots

In this section, we are evaluating the distributions and characteristics of our dataset's "numeric" or "continuous" independent variables, which are typically quantitative variables. Additionally, we are examining the frequency or occurrence of different categories within our categorical variables, which are variables that represent distinct groups or labels.



#### Frequency Plots

Frequency of the different categorical variables to get a better understanding of their distribution and potentially drop some variables that may act as outliers.

|  |  |
| --- | --- |
|  |  |
|  |  |

#### 

#### Further Cleaning & Preprocessing

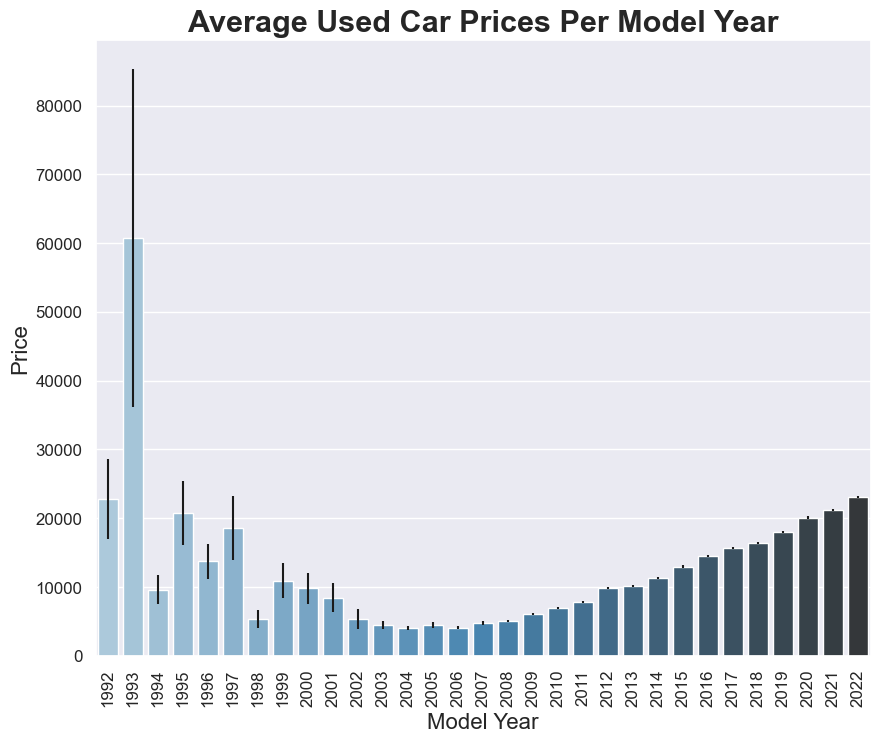
* Based off the above frequency plots, some categories may act "outliers" due to very low counts.
* Drop two fuel\_type categories, etanol & CNG. (low frequency, very imbalanced)
* Kept all of the categories within gear type.
* Then clean the outcome variable column 'price', from any outliers such as cars priced over 400,000 euros or less than 0 euros.

|  |  |
| --- | --- |
| Cleaing Fuel type | Cleaning Gear type |
|  |  |

### Visualization

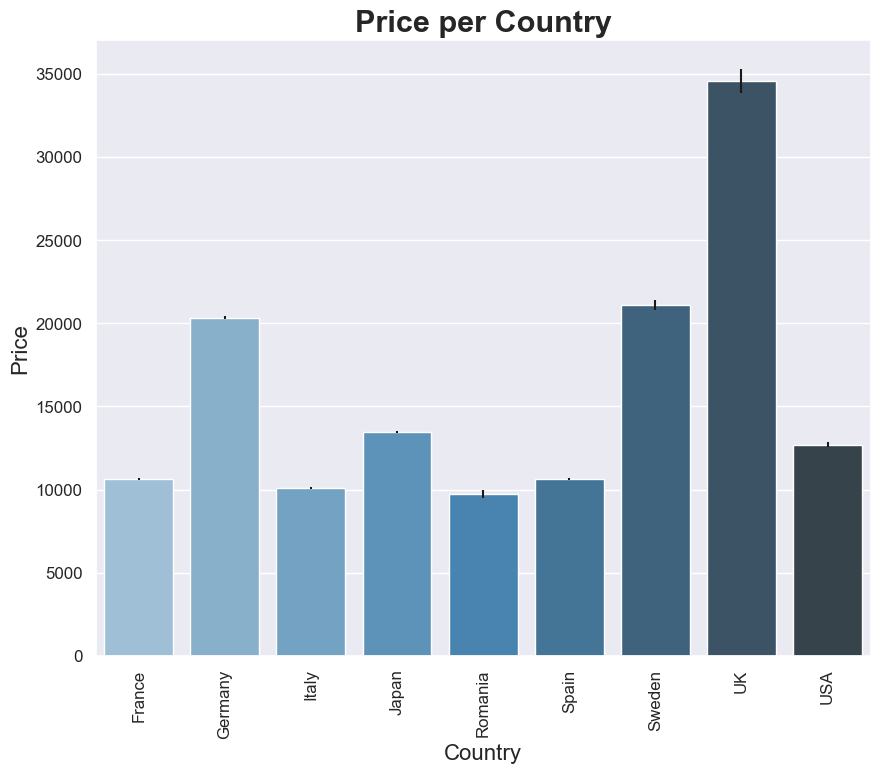
#### Distribution of Price per Year

Age of the car affects the price, where newer cars are higher in price and older cars tend to be lower in price. An interesting point to note is that cars before 1990 are really high in price, which may suggest the existence of classic cars (note: error bars are really high, very few old cars in the dataset with high variance in prices).

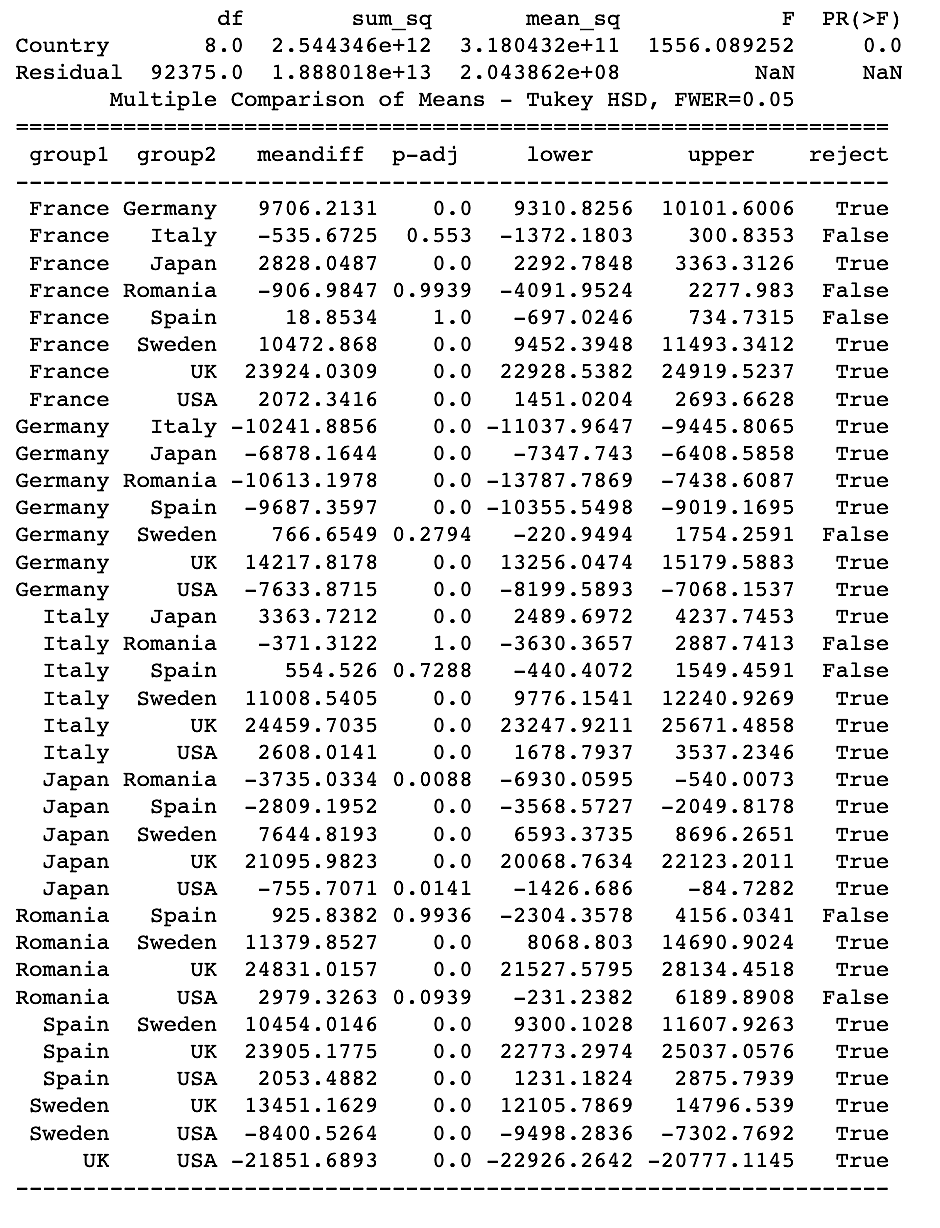


#### Distribution of Price per Country

* Country Level Analysis  
    
  Add the country in which each car was manufactured to my dataset to try to pull out some interesting results. Repeat the distribution of Price but instead use country to see if cars manufactured in different countries are more expensive than cars from other countries.

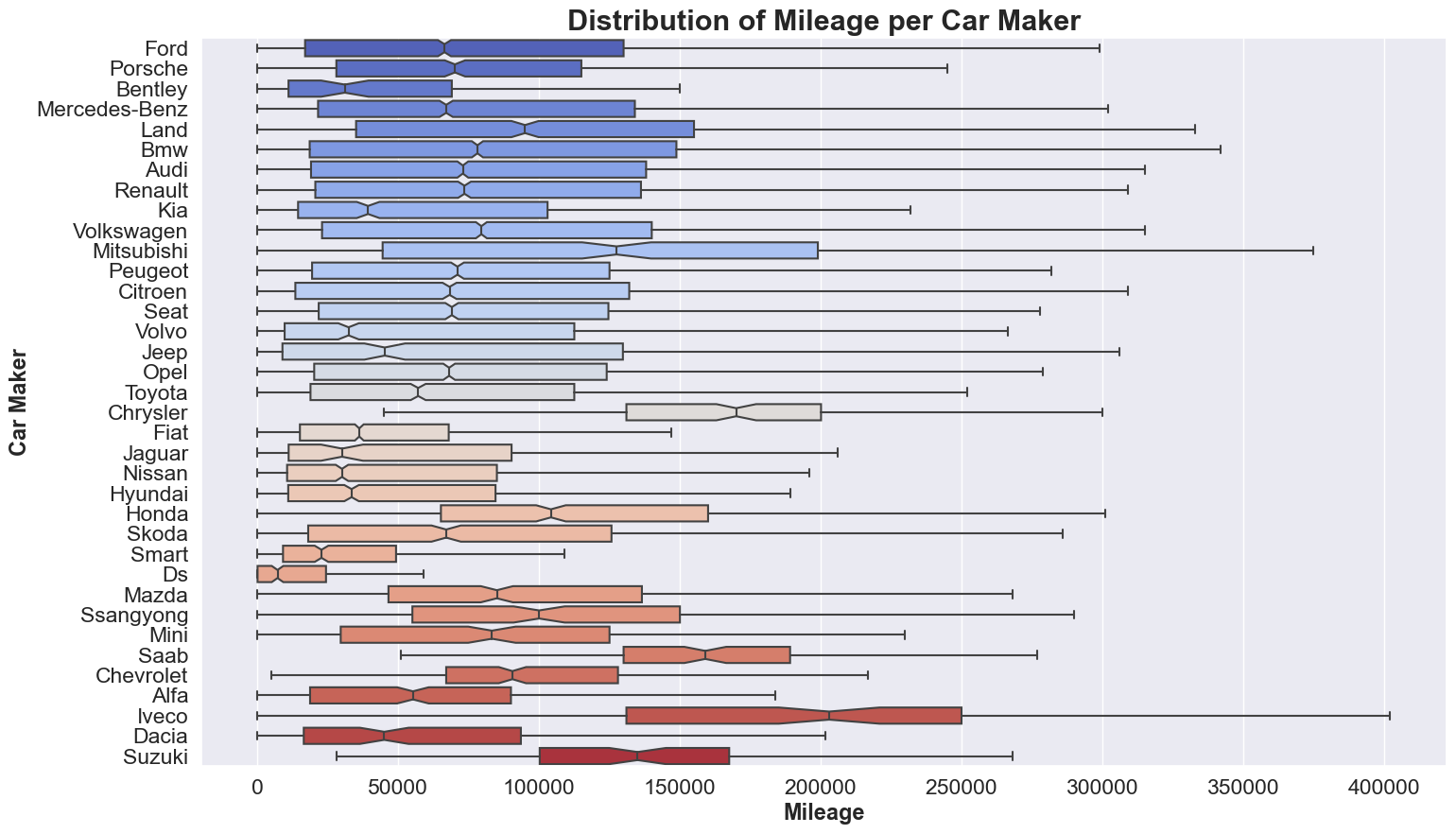


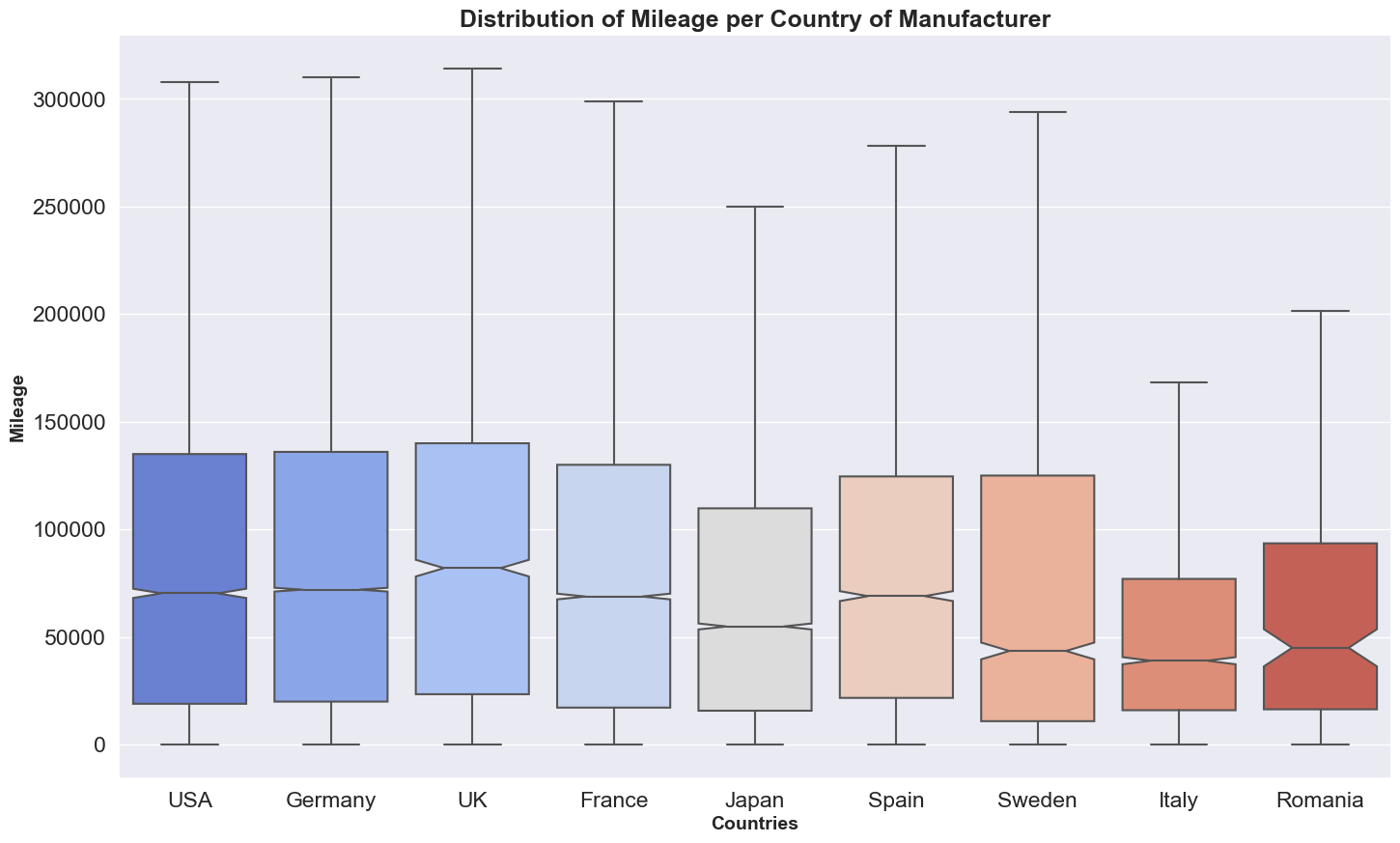
* Statistical Analysis  
    
  We chose to do a one-way ANOVA as working with the price means for each year/sample. Followed this up with Tukey's honest significance test to find which specific pairwise means are signifcantly different from each other (adjusts p\_values).

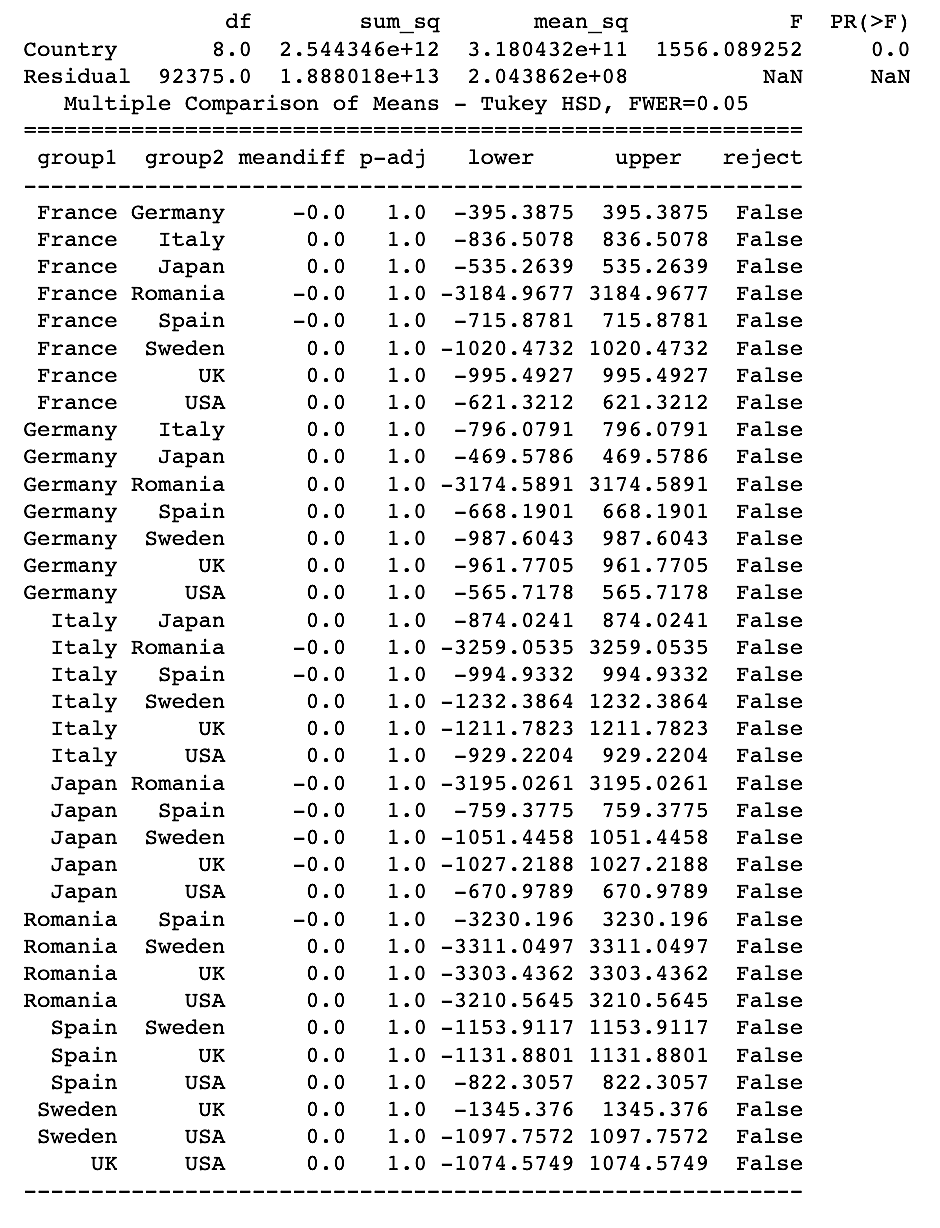


#### Distribution of Mileage per Car Maker/Country

* Statistical Analysis  
    
  Please note that we only highlight one car maker of interest as we use that as an example in the presentation. Followed up post-hoc analysis for small things that were cool to add in the presentation. It is interesting to find a bus manufacturer (Iveco) has siginificantly higher mileage relative to all the other car makers.



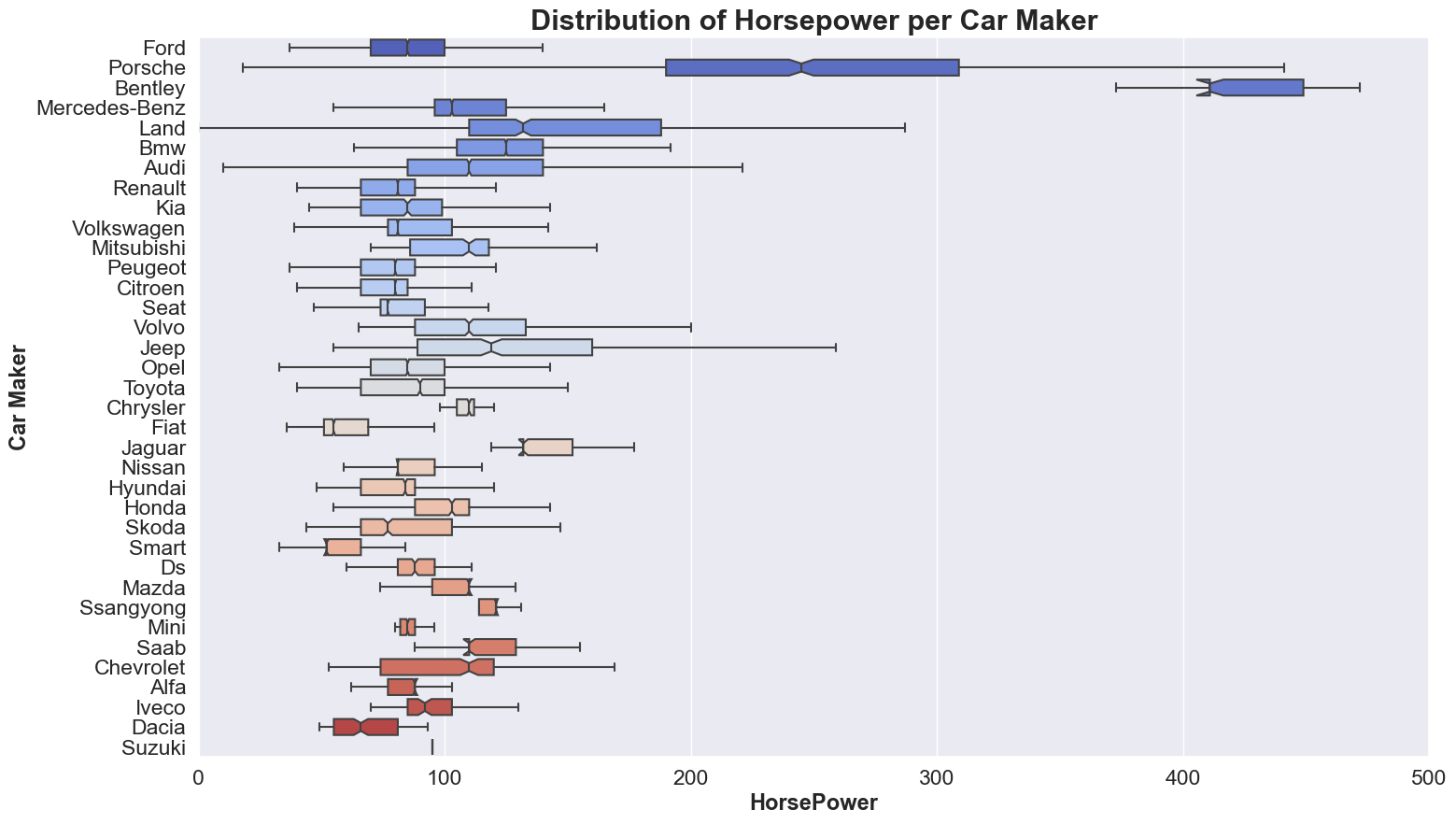


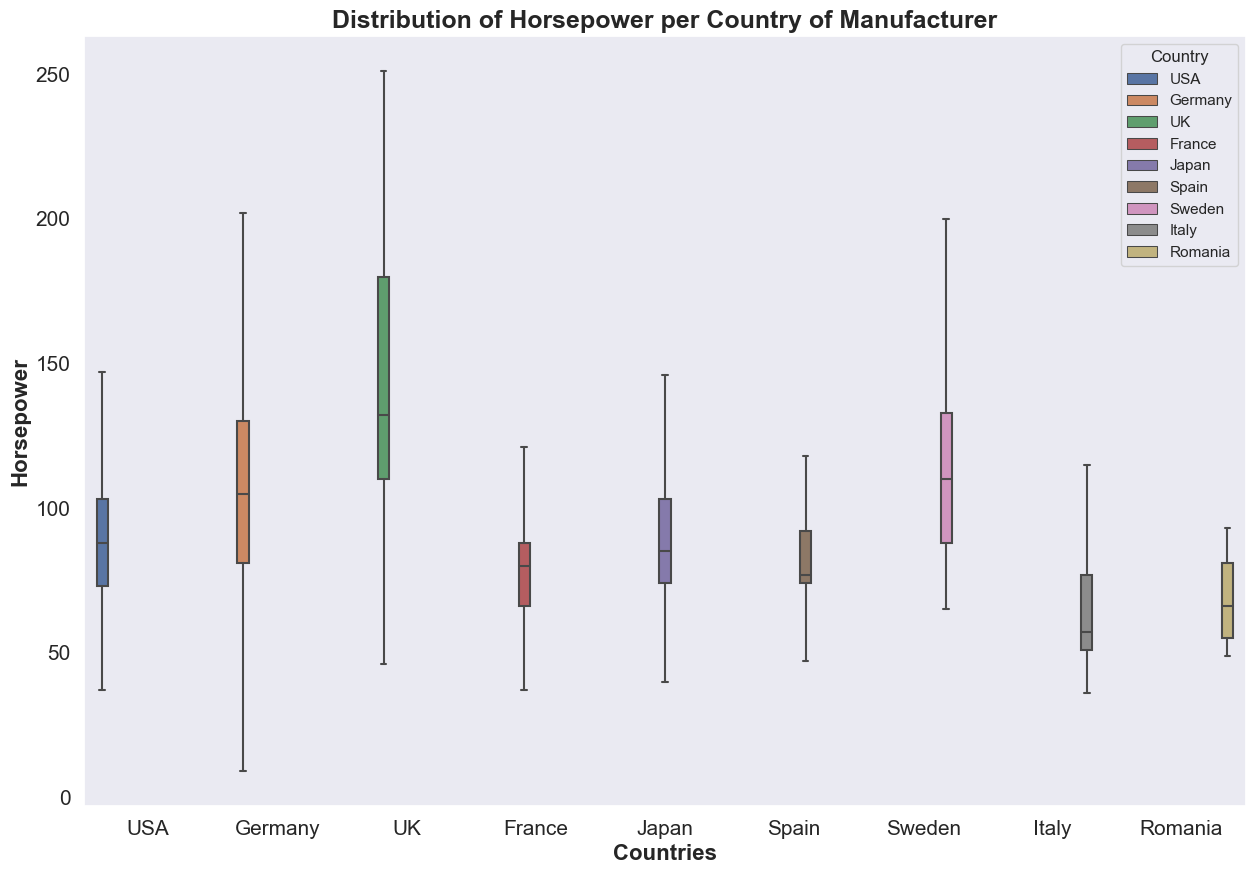
ANOVA  


#### 

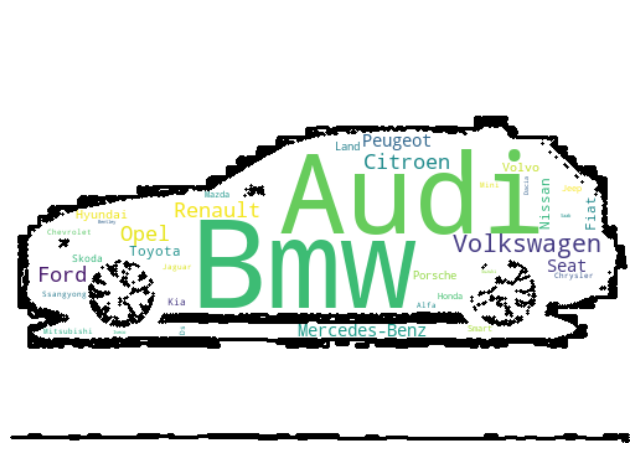
#### Distribution of Horsepower per Car Maker/Country

Statistical Analysis (Country)  
  
Please note that we only highlight one country of interest as we use that as an example in my presentation. It is interesting to see that the UK has significantly higher horsepower cars relative to the other countries. The cars from the UK in this dataset are relatively luxury (Jaguar, Bentley and Land Rover). We would've expected Italy because of cars like Ferrari and Lamborghini, but these car makers are not in the dataset.



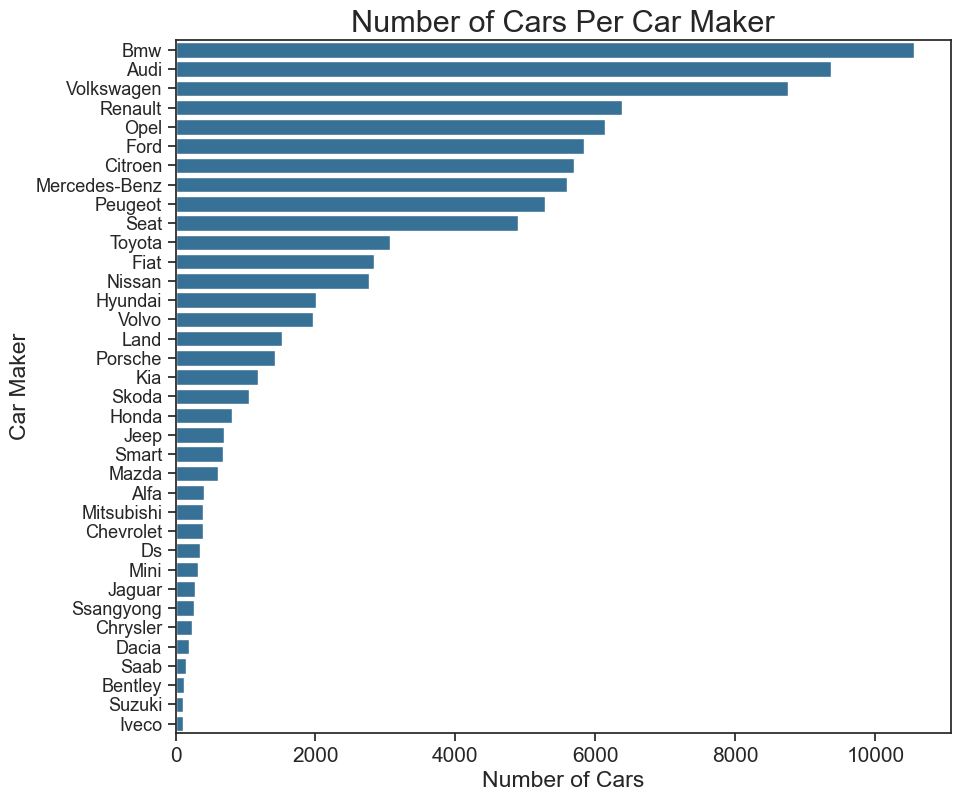


Cars that are most frequent in the dataset  
  
Visualization in a word cloud where the size of each manufacturer's name is determined by its frequency in the dataset, and the shape of the cloud follows the silhouette of the loaded image

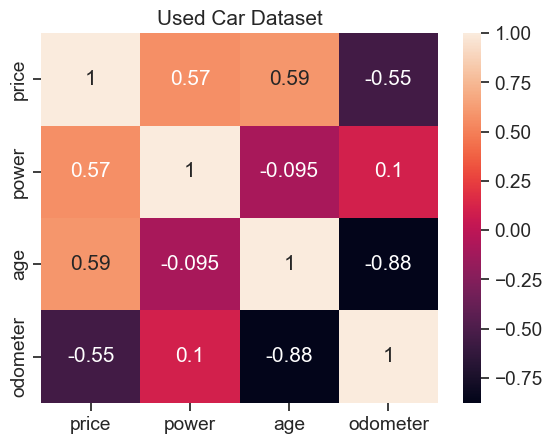


#### 

#### Horizontal bar plot to visualize the number of cars produced by different car manufacturers



#### Correlation coefficients of the numeric variables



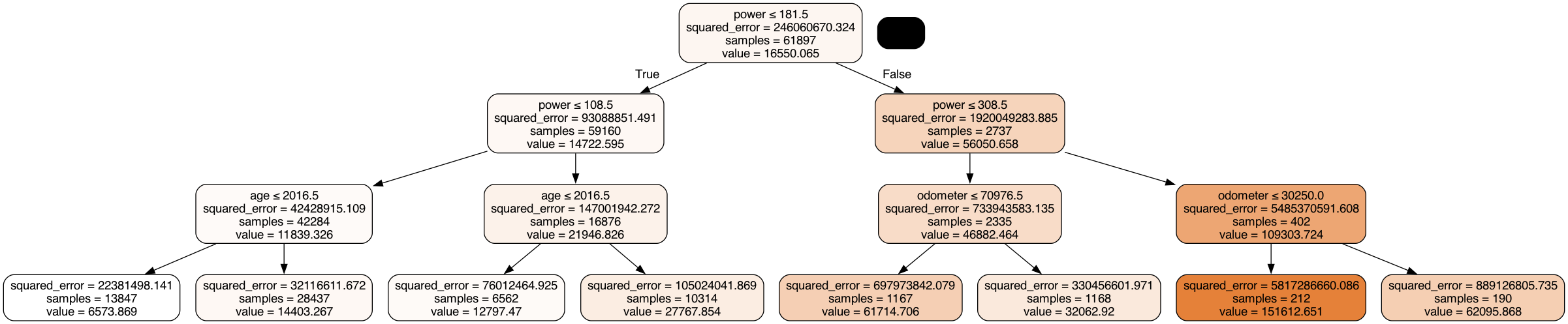
Machine Learning Models

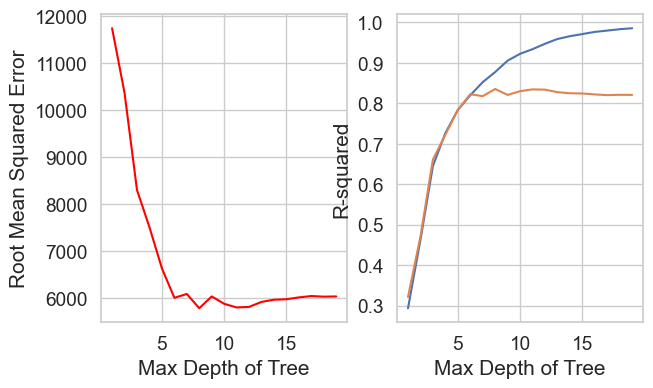
In this section, we test out 5 different regression models on the dataset.

* We don't do K fold or any standard form of cross validation for the first 3 models as it would take a lot of time and processing power.
* For the Decision tree Regressor and the KNN regressor, we just test out a range of values for the max\_depth and number of neighbors and plot out the scores for the train/test values.
* we select the hyperparameter that don't result in overfitting and that return a good score on the test set. (small differences between training and testing sets --> better at generalizing) (essentially finding the right bias variance tradeoff)
* For the Xgboost and Random Forest Regressor, we optimized the hyperparameters using a randomized search cross validation and each took about one to two days to run.
* Note: the Xgboost was a lot slower to optimize hence why there are less parameters in the randomized search specific to the Xgboost regressor.
* Due to a lot of the heavy processing, we had to restart my kernel a couple times and didn't rerun the randomized search CVs but we saved the that the CV returned.
* Metrics: We use the R-Squared to determine which model we will use for deployment.
* We also use the Root Mean Squared error as it reflects the units of the outcome variable (price).

#### 

#### Decision Tree Regressor

The first regression tree was just a proof of concept, to see which feature minimizes confusion for the tree most. (which feature is the most important in terms of mean squared impurity)  
  




* The Score on the training set with a decision tree regressor is: 0.8772708804008376
* The Score on the test set with a decision tree regressor is: 0.8353329544337932
* Mean squared error: 5783.88

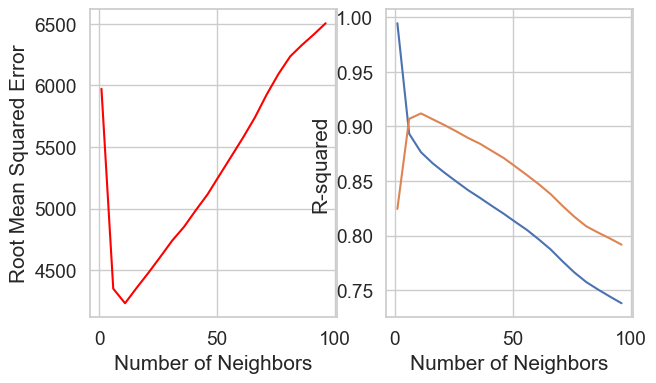
#### Linear Regression Model

We have pre processed the data, build a linear regression model, made predictions, and evaluate its performance. The key goal is to understand how well the model can predict target values based on the features.

* The Score on the test set with a linear regression is: 0.7833717157006708
* Root Mean Squared Error: 6633.97

#### KNN Regressor

This took a significant execution time delays because it involves a loop that traverses a range of n\_neighbors values in a KNeighborsRegressor model, calculating different scores and root mean squared errors. Specifically, it used a range defined as np.arange(1, 100, 5), which implies that the loop covered values from 1 to 99, incrementing by 5 at each step.



* The Score on the test set with a KNN regressor is: 0.9068651883541627
* Mean squared error: 4349.83

Random Forest Regressor  
  
When the RFR performed on the train data, we got the result below

* The Score on the test set with a random forest regressor is: 0.9073880007288752
* Mean squared error: 4337.60

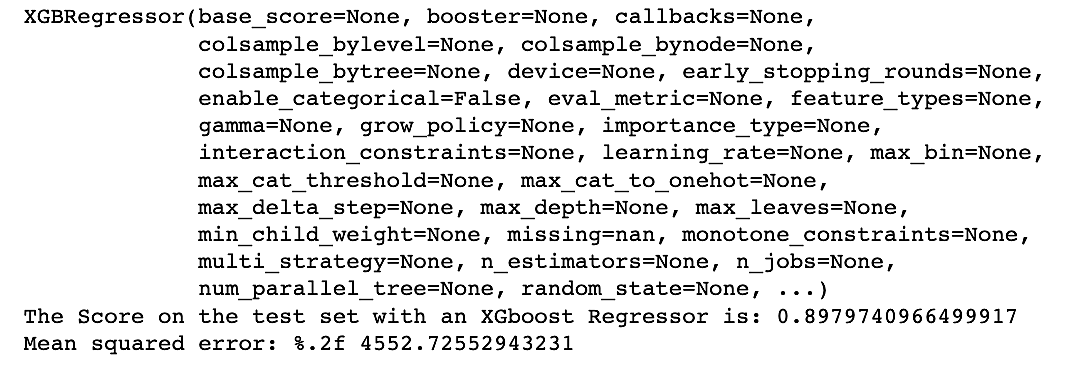
Them, We tried to optimized the RFR previously and we got the paremeters below

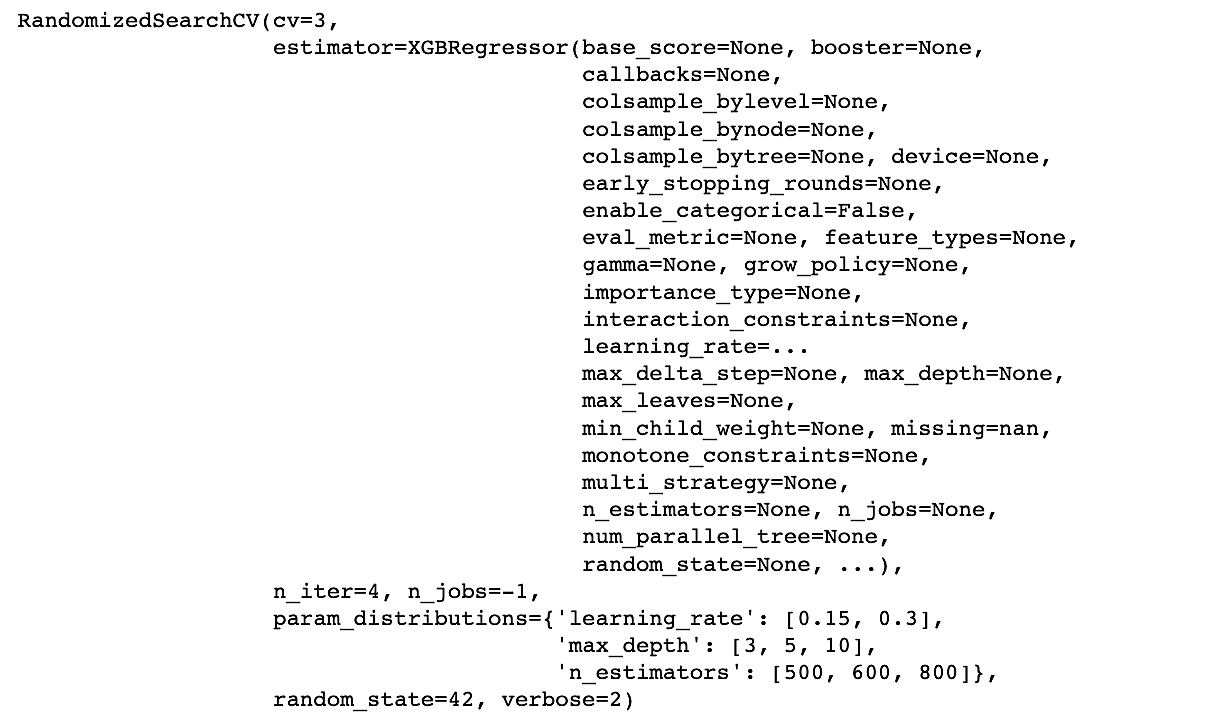
* These are the best parameters from the randomized search:{max\_depth: 40, n\_estimators: 600, min\_samples\_split: 10, min\_samples\_leaf: 1, bootstrap: False, max\_features: sqrt}

The part ran several hours since it is performing a randomized search for hyperparameter tuning on a RandomForestRegressor model. It's attempting to find the best combination of   
hyperparameters from a defined search space. Remember that hyperparameter tuning can be a time-consuming process, especially when dealing with complex models and large datasets.

* The Score on the train set with a hyperparameter optimized random forest regressor is: 0.9447691922364366
* The Score on the test set with a hyperparameter optimized random forest regressor is: 0.9306952666875629
* Mean squared error: 3752.30

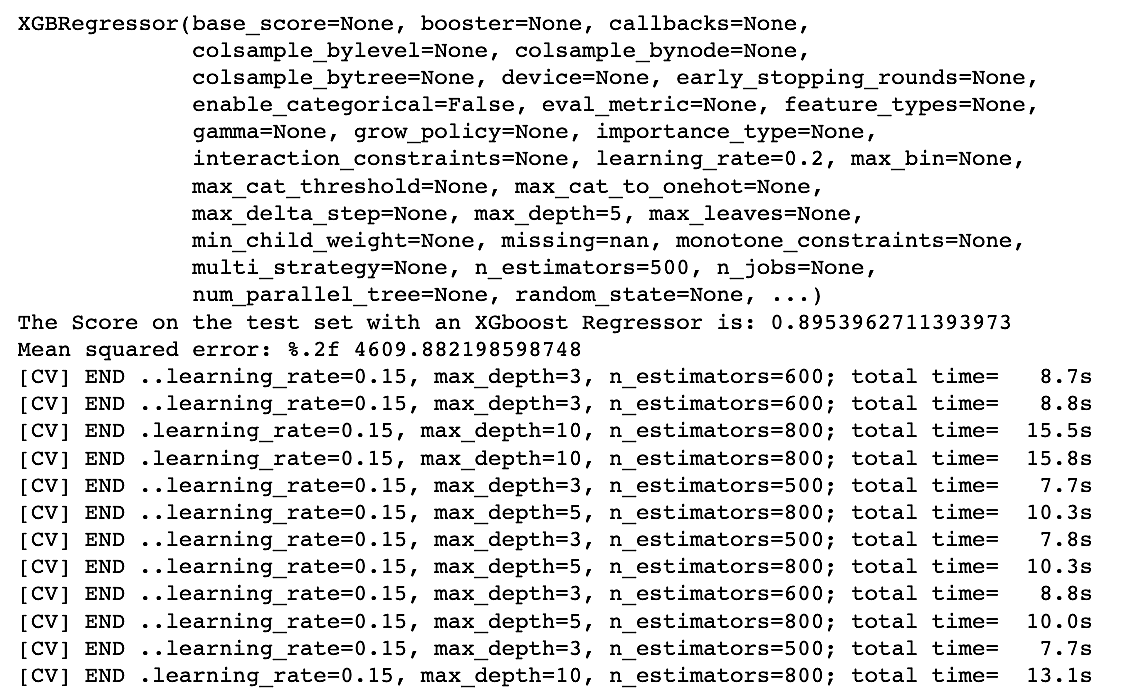
###### XgBoost Regressor

We hace used this model on training data. After training the model, it's using the test data (X\_test) to make predictions, and then it's evaluating the model's performance.  
  


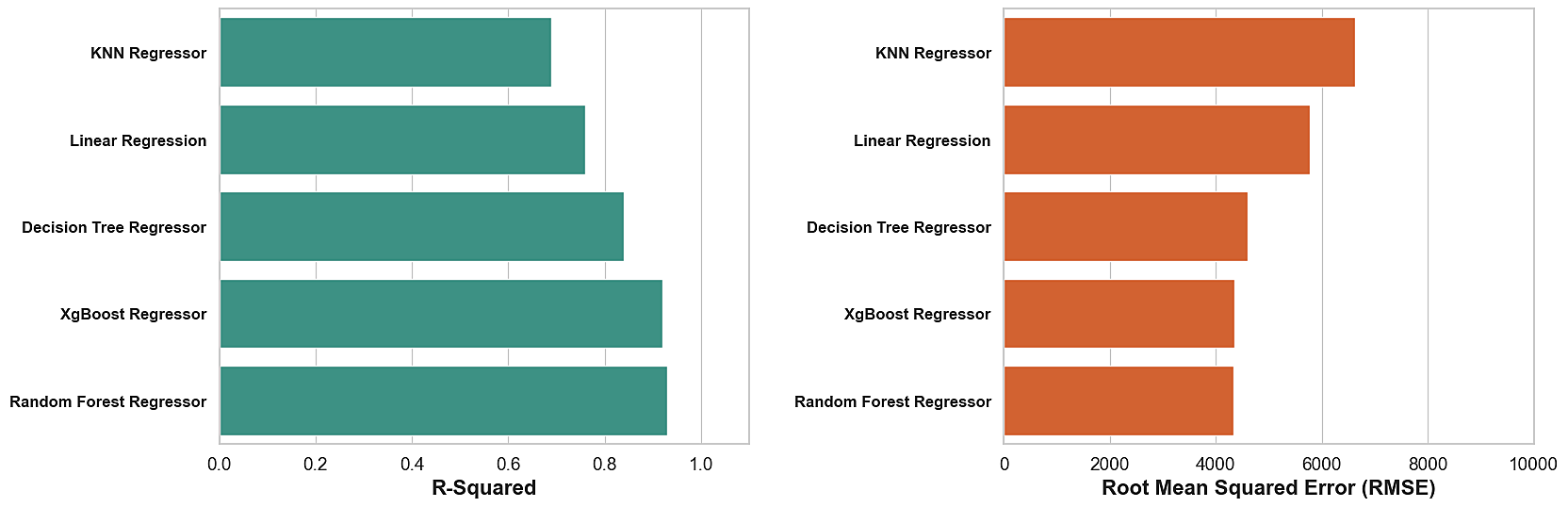
Them, We tried to optimized the XgBoost and performed a randomized search   
  


These are the best parameters from the randomized search (Xgboost) :{max\_depth: 5, n\_estimators: 500, min\_samples\_split: 10, learning\_rate=0.2}

Then we used the best parameters from randomized search and the output is below



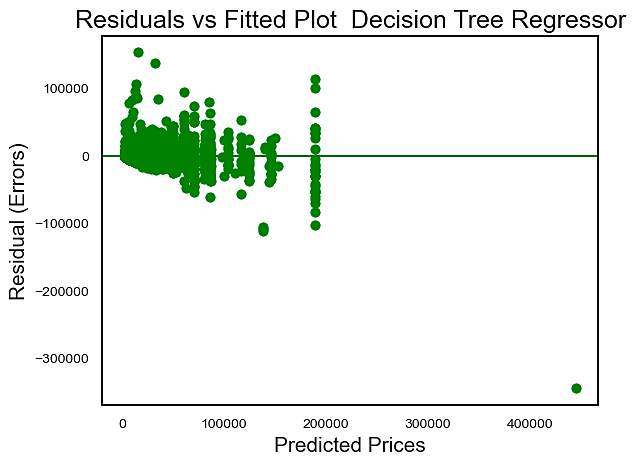
Model Comparison

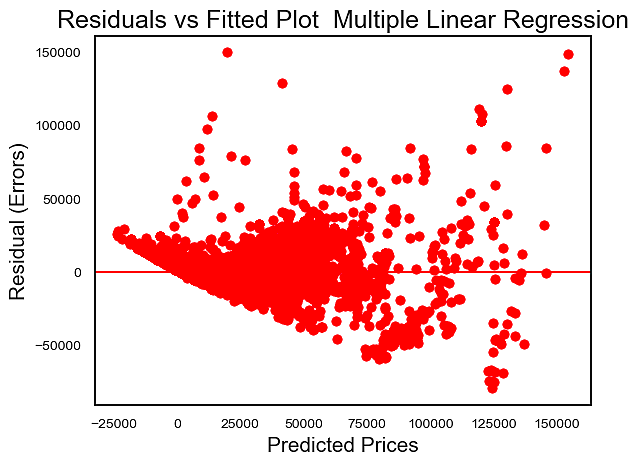


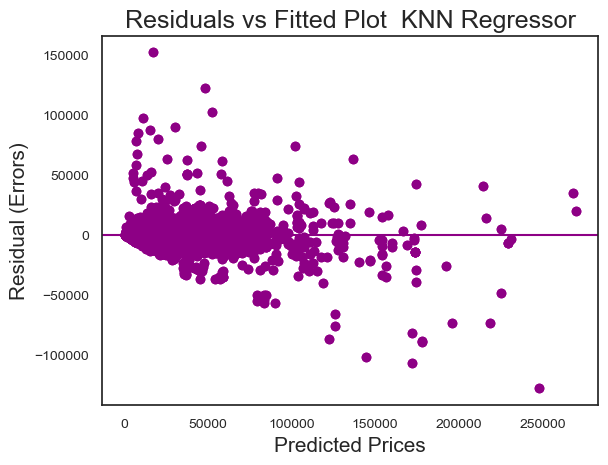
|  |  |
| --- | --- |
| **1. Decision Tree Regressor**  Training Set R-squared: 0.8773  Test Set R-squared: 0.8353  Mean Squared Error (MSE): 5783.88 | **2. Linear Regression**  Test Set R-squared: 0.7834  Root Mean Squared Error (RMSE): 6633.97 |
| **3. KNN Regressor**  Test Set R-squared: 0.9069  MSE: 4349.83 | ***4. Random Forest Regressor (Optimized)***  ***Training Set R-squared: 0.9448***  ***Test Set R-squared: 0.9307***  MSE: 3752.30 |
| ***5. XGBoost Regressor***  ***Test Set R-squared: 0.8954***  MSE: 4609.88 |  |

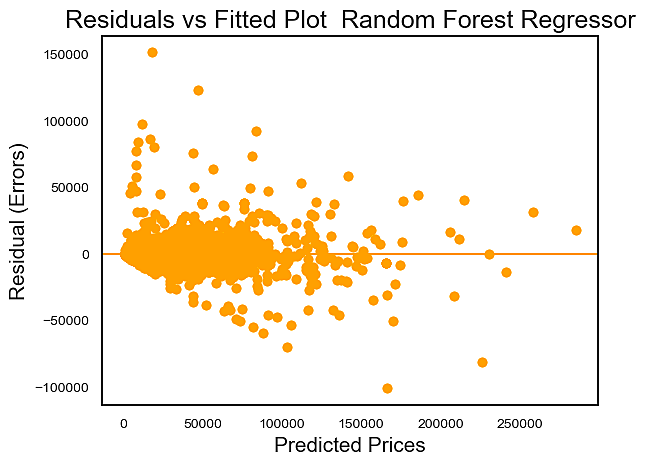
### Residual Plot

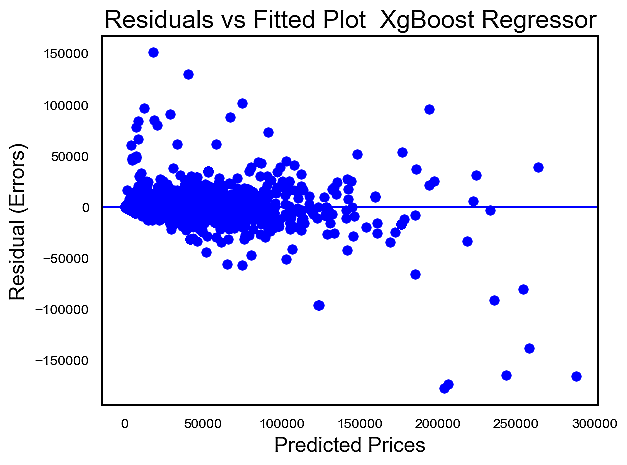
Using this we can identify potential issues with the model and decide whether any model modifications or transformations are needed. A well-behaved residual plot, with residuals randomly scattered around zero without any discernible patterns, suggests that the model is a good fit for the data. If issues are detected, further analysis or model adjustments may be necessary to improve the model's performance.











#### 

#### 

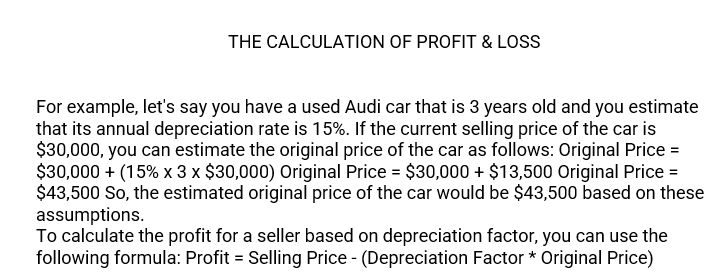
#### Based on these metrics:

* The Random Forest Regressor with hyperparameter optimization appears to have performed the best. It has a high R-squared score on both the training and test sets, indicating good explanatory power and generalization capability. The MSE is also relatively low, suggesting accurate predictions.
* The KNN Regressor also performed well, with a high R-squared score on the test set and a low MSE.
* The Decision Tree Regressor performed decently but had a slightly lower R-squared score on the test set compared to the other models, and a higher MSE.
* The Linear Regression model had the lowest R-squared score on the test set and the highest RMSE, indicating that it may not fit the data as well as the other models.
* The XGBoost Regressor performed well but had a slightly lower R-squared score and a higher MSE compared to the Random Forest and KNN models.

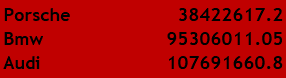
**The Random Forest Regressor with hyperparameter optimization would be a strong candidate**.

Profit calculation

Profit earned by car sellers based on the depreciation factor of the car manufacturers.  
 **Profit = Selling Price - (Depreciation Factor \* Original Price)**



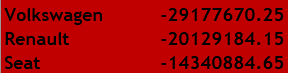
We found that the top three manufacturers of selling a car are the below:



The below has the highest profits.



We found that the top three manufacturers who do not make a good profit



Summary of the Profit & Loss:

* + We recommend considering the purchase or emphasis on used cars from the three manufacturers: Porsche, BMW, and Audi. These manufacturers have consistently demonstrated the ability to yield some of the highest resale values and profits among others.
  + We advise against purchasing or prioritizing used cars from the three manufacturers: Volkswagen, Renault, and Seat. These brands typically offer lower resale values, resulting in reduced or no profit when compared to other manufacturers in the market
  + We recommend purchasing or directing your focus toward used cars from the following models: Tucson, Mat-su, and Birmingham, as they not only yield the highest profits but are also the top-selling models among the manufacturers.

### 

### Recommendation

* Based on Analysis, we can to divide our cars into 3 segment Low, Medium and High budget.
* We also aim to reduce the time required for the development of a precise random forest regression model. These enhancements will lead to an immediate and significant impact on our business.
* The next step post that would be to cluster different sets of data and see if we should make multiple models for different locations/car types.
* With Increasing petrol rates diesel car are in more demand in recent years, acquiring and selling them can high profits
* Our used car pricing model is designed to assist used car sellers in ensuring they receive fair compensation for their pre-owned vehicles.
* While our model currently provides accurate predictions and value analysis for used cars, there is still a margin of error. Minimizing this error is a top priority*.*
* Mileage is inversely correlated with Price. Generally, high mileage cars are the lower budget cars.
* Kilometres Driven have a negative relationship with the price which is intuitive. A car that has been driven more will have more wear and tear and hence sell at a lower price, everything else being 0.
* The categorical variables are a little hard to interpret. But it can be seen that all the car category variables in the dataset have a negative relationship with the Price and the magnitude of this negative relationship decrease as the brand category moves to lower brands
* To maximize profits for both used car sellers and our business, several key strategies include:
  + Increasing the availability of "low mileage" cars.
  + Expanding the inventory of vehicles with "automatic transmissions."
  + Adding more "diesel" cars to our listings.