# **Introduction**

For this study, a dataset containing secondhand Volkswagen car sales collected from the UK has been used. This analytical analysis will investigate what drives the change in a cars price and by which direction each variable influences the price. Focusing on Engine Size, Miles per gallon (MPG), Fuel Type, Tax, Mileage and Transmission type, this analysis will investigate how variables influence differs between the car models selected. Additionally, an appendix slide included in the PowerBi document demonstrates how predictive analytics can be used with multiple linear regression to estimate a car’s price given inputs. This analysis focuses on the top 3 models with the greatest data points. The models chosen for this analysis are ‘Golf’, ‘Polo’ and ‘Tiguan’. These models have a total 9,915 data points with Year ranges between of 13 to 20, providing the highest variance in data.

# **Summary Of The Dataset**

The entire dataset contains 15,157 rows made up of 9 variables, these variables are described below:

|  |  |
| --- | --- |
| Variable | Description |
| 1. Model | Vehicles Model/Name |
| 1. Year | Year of Manufacture |
| 1. Price | The price of sale in UK pounds |
| 1. Transmission | Type of transmission the car is equipped with |
| 1. Mileage | The Odometer Reading |
| 1. Fuel Type | The Type of Fuel the car uses |
| 1. Tax | Tax paid on sale of vehicle |
| 1. MPG | Miles per gallon |
| 1. Engine Size | Displacement of internal combustion engine in liters |

Out 27 different car models 3 have been chosen. The models chosen held the top 3 observation counts in the dataset Golf had 4,863 observations with a 20-year range of data, Polo had 3,287 observations with a 16-year range of data, Tiguan had 1,765 observations with a 13-year range of data, (Figure 1.1).

For each model in this analysis Price is the dependent variable with the remaining 7 variables being the independent variables (Model excluded). The numerical variables are Price, Year, Mileage, Tax, MPG and Engine size. The Categorical variables are Fuel Type and Transmission Type. Prices in the dataset range from $1.25k to $69.99k, Year ranges from 2000 to 2020, Mileage ranges from 1 to 212,000, tax ranges from $0 to $580, MPG ranges from 0.3 to 188.30 and Engine size ranges from 0 to 3.20L.

The dashboard shows both a descriptive and diagnostic summary of the data. Figure 1.2 shows a correlation plot of the numerical variables. The yellow squares show the positively correlated variables and purple show the negatively correlated variables. Figure 1.3 shows the median price of a model split up by year. Figure 1.4 show the distribution of price for each of the models. Figure 1.5 shows categorical boxplots for Transmission Type and Fuel type against Price. Figure 1.6 shows scatterplots for numeric variables against Price. The dashboard is adaptive with multiple filter selections available. Years can be filtered by dragging the bar under figure 1.2 to select a year range. This allows a drill down feature on specific ranges of years. Below this is the model selection allowing drill down on any of the 3 models, by default this has been set to all models. Users are also able to change the numeric scatterplots by selecting the numeric X axis variable and separate this by the changing categorical variables as the legend, (Figure 1). Please ensure that field parameters are turned on for selection to work (Go to File > Options and settings > Options > Preview features).

*Figure 1: Dashboard*

Graphical user interface, chart

Description automatically generated

Observing the price distribution in figure 1.4 Golf and Polo are right-skewed (the mean of the price is large than the median of the price), where Tiguan has a more symmetrical distribution. The median price of Polo and Tiguan is around $10,000 while the median price of Tiguan is around $22,000.

Golf appears to be the oldest model based on figure 1.3. It can be observed that Golf frequently had the most expensive model up until about 2008. Post, the model Tiguan (Yellow) can be seen to frequently hold the highest median price for each year with Golf (Blue) second and Polo (Purple) third.

Looking at the categorical variables in figure 1.5, Fuel Type Diesel cars show the highest range with petrol second. Petrol cars also show the highest number of outliers. Notably the only model with Hybrid cars is Golf. For Transmission type, Automatic cars have the highest spread with Semi-Auto second. Manual cars show the highest number of outliers.

# **Analysis Undertaken**

*Figure 2: Data Acquisition*

* 1. Table

     Description automatically generated***Data acquisition:***

Graphical user interface, text, application

Description automatically generatedThe data is download from the server and save as a csv file. This file then is import to PowerBI report using “Get Data” and connect. After we have a preview, then we comes to next step “Transform”.

* 1. ***Data preparation:***

|  |  |
| --- | --- |
| In Power Query Editor, first we check the input type to confirm it’s suitable with the data description | Table  Description automatically generated |
| Since the format of Model columns has a space before model name, then we remove that space by trimming that column. Select only Golf, Polo and Tiguan through the dropdown filter. | Graphical user interface, application, table  Description automatically generated |
| View the list of categorical variables to check if it’s similar to data description. |  |
| Duplicate the table and create a separate dataset containing the counts for each model for each year. |  |
| Create a relationship between the two tables using “Manage Relationships” where the year columns in both tables equal each other. |  |

* 1. ***Parameter Setup:***

|  |  |
| --- | --- |
| Go to File > Options and settings > Options > Preview features. Select the field parameters checkbox. This will allow us to create adaptable variable parameters for the scatterplots. |  |
| Create two field parameters, one containing the categorical variables and the other one containing the numerical variables (excluding price). |  |

***4. Data Exploration, Visualization and Analysis***

Data exploration has been undertaken by exploring meaningful visuals which provide insightful descriptive and prescriptive statistics about each of variables. These visualizations have been built both through the local PowerBi environment as well as through using the PowerBi R Script visual.

The numerical variables were plotted using the R library “corrplot” creating a color-coded correlation matrix showing the direction of relationships between variables. The dark purple shows the negatively correlated variables and the yellow shows positively correlated variables. The correlation coefficients are shown in white. (See Figure 3)

Chart

Description automatically generated*Figure 3: Correlation Plot*

Box plots were used to show the distribution of price against the categorical variables Transmission Type and Fuel Type. These were also completed in the R Script Visual using the R library “ggplot”. These plots show the range, median and number of outliers for price against these variables. (See Figure 4)

*Figure 4: Boxplots*

Chart, box and whisker chart

Description automatically generated*Text, email

Description automatically generated*

To explore data trends, scatter plots were used. Field parameters have been created (See details in 3), allowing the user to select which variables and legend they want to view against price. Users can see the relationship of each variable with price and drill down on specific categories or models. (See Figure 5)

*Figure 5: Scatterplots*

Field parameters created above have been inserted here.

Graphical user interface

Description automatically generated***Graphical user interface, application

Description automatically generated***

The row counts and the Median of price by year and model visuals have been built using the PowerBi internal charts, while the price distribution model was built using ggplot. These provide the descriptive statistics for this analysis. (See Figure 6)

*Figure 6: Descriptive Visualizations*

*Chart, bar chart, histogram

Description automatically generatedChart, histogram

Description automatically generated*

*Chart, funnel chart

Description automatically generated*

**Results**

Analysis of the relationships between price and the other numerical variables is undertaken using the correlation plot, bar chart and scatter plot. The price for Golf shows positive correlation with year, tax and engineSize, while negative correlation with mileage and mpg. Year shows the strongest positive correlation to price at 0.70 while mileage has the strongest negative correlation at -0.67. The relationship with year is confirmed by the bar chart showing the steep slope of rising median price except for the 2003 which appears as an outlier. The scatter plot for Golf with mileage and mpg (mpg < 100) in association with price confirm the correlation direction seen in the correlation plot, while plots for engine size and tax do not show any definitive trends associated with price. (See Figure 7)

*Figure 7: Golf*

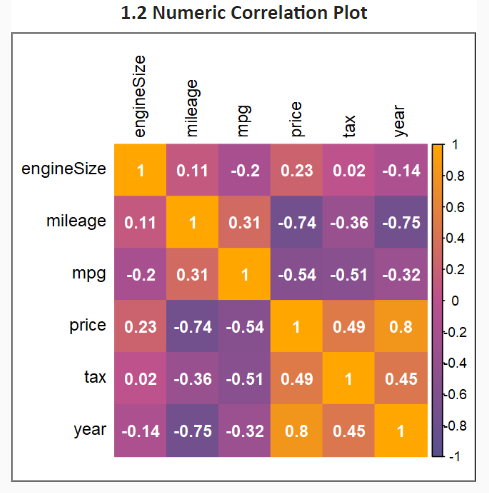
Text

Description automatically generated with low confidenceChart

Description automatically generated

For both Polo (Figure 8) and Tiguan (Figure 9), Year has the strongest positive correlation while mileage has the strongest negative correlation to price. This is confirmed again by the bar plot and scatter plot for mileage. Tiguan has the strongest negative correlation with mileage at -0.77 and the strongest positive correlation with year at 0.83. The scatterplot with MPG for both Polo and Tiguan does not show any obvious trends in price decreasing, compared to the relationship found with Golf. There appears to be no relationship between engine size and tax with price for all models.

*Figure 8: Polo*

Chart, scatter chart

Description automatically generated

*Figure 9: Tiguan*

Chart

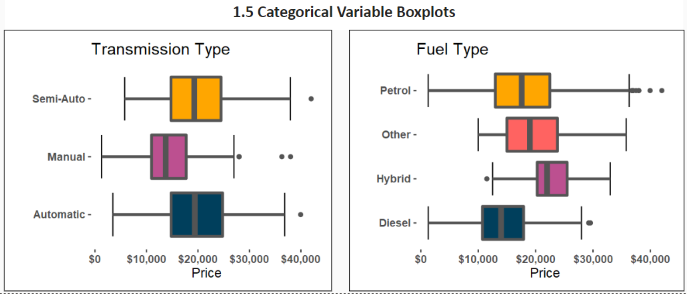
Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

Looking at the boxplots of transmission types, a Golf (Figure 10) with Manual transmission shows has a lower median price and range than all other transmission types. The same is held true for Polo and Tiguan models. While the price of Semi-Auto and Auto transmissions for Golf and Tiguan aren’t significantly different, a Polo model with Semi-Auto transmission seems to hold a higher price range than Auto. Looking at the fuel type, notably only Golf has the Hybrid fuel type which also holds the highest median price, while Polo and Tiguan have ‘other’ fuel type which has the highest median price. Golf with Diesel fuel show a lower price on average than other types, while Polo and Tiguan car have differences in price range between Diesel and Petrol the median value is relatively similar in both cases.

*Figure 10: Golf Box Plots*

**Chart, box and whisker chart

Description automatically generated**

*Figure 11: Polo Box Plots*

*Figure 12: Tiguan Box Plots*

**Chart, box and whisker chart

Description automatically generated**

**Recommendations and conclusions**

***Conclusion:***

Variables are assessed as holding importance in influencing price if they hold either a strong positive or negative correlation. A strong correlation is defined as having a coefficient of between 0.70 and 1, moderate correlation is between 0.40 and 0.7 and weak to no correlation is below 0.40. Based on this definition, all 3 models show a strong positive correlation between year and price, with Polo having the weakest influence of the 3. The second influential variable mileage shows a negative relationship across all models, being strong on Polo and Tiguan, while only a moderately negative on Golf.

As mpg increases for Polo and Tiguan models, the price will tend to decrease, while for Golf this trend is only exists for cars which have mpg less than 100. According to correlation plot, tax and engine size is moderately correlated to price for Golf and Polo’s. However, this is not supported by the scatter plot. Since the correlation plot only shows the linear relationship between a pair of variables, it is not always suitable with real-life data modelling. Hence, here the scatter plot shows a better view of the actual relationships and we can conclude that those variables have low impact on the price for all 3 car models.

Transmission and Fuel type were also found to affect the price of the 3 models. The price of Manual cars for all models is clearly lower than other types, while the difference in price for the other type (Semi-Auto and Auto) for Golf and Tiguan is not considerable. For fuel type, as mentioned above Polo and Tiguan with fuel type “Other” have the highest median price while for Polo with Hybrid has the highest median price.

***Recommendation:***

In the framework of this project, we used fundamental analysis techniques with correlation analysis, visualization with bar chart, box plot, scatter plot, and drill up and down to get the details for each car model. It was found that some variables can have impact on price however they can have a non-linear relationship (for example mpg < 100 for Golf car). Attached to the PowerBi document is a price prediction model based on multiple linear regression for each model. A user can investigate by adjusting each of the variables and see how price moves. A linear model has been created separately for each car model with the coefficients built into the dashboards by R script. As this is based on a linear model it does not capture some of the nonlinear relationships found above. In future, we could extend this analysis to incorporate different machine learning models (i.e decision tree, random forest) to investigate these non-linear relationships. In addition, we could collect other potential variables such as color, interior detail, times of repairing to create better prediction accuracy for the estimation of a cars price.