**Building a Predictive Model for Automobile Prices in USD**

# **Abstract:**

“A new car is second only to a home as the most expensive purchase many consumers make” (Herbert, Amy, et. al). When dealing in such a transaction, being a small percentage off in a price estimate could result in thousands of dollars being left on the table. To estimate the value of a car is tricky, there are a lot of different variables to take into play. Additionally, the value of a car isn’t linear - meaning a car won’t depreciate the same amount each year - so even if the price was estimated a year ago, that number would be outdated today. The purpose of our prediction model is designed to estimate a car’s price based on many features, resulting in a fair and accurate assessment that will be current with the market at the time of the valuation.

# **Authors:**

Ryan Geier, geierr@uw.edu

Swapnil Kumari, swapkum@uw.edu

Gene Park, [genepark@uw.edu](mailto:genepark@uw.edu)

Radha Priyanka Jaggumantri, rpj9@uw.edu

William Froelich, willfr@uw.edu

Last updated: 07/20/2022

**Contents:**

**Abstract:** 1

**Authors:** 1

**Initial Data Collection Report:** 3

Data Source:3

**Data Description Report:** 3

Dictionary: 3 to 4

Univariate properties: 4 to 5

Univariate Visualization: 5 to 8

**Data Exploration Report**: 9

Mileage: 9 to 10

Cylinder:10 to 11

Airbags: 11 to 12

Manufacturer: 12 to 14

Category: 14 to 15

Color: 15 to 16

Leather interior: 17 to 18

Gear Box type: 18

Fuel type: 19

Turbo: 19 to 20

**Data Quality Report**: 20

Missing data: 20

Outlier: 21

Distribution:21

Numeric features: 21

Categorical features:21

**Inferences and business decision**:22

**References:** 22

# **Initial Data Collection Report**

## **Data source**

1. Price Data: Free Data source Webpage

· Format: Microsoft Excel CSV

· URL: https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge

· Cost and future availability: Subject to Kaggle.com

# **Data Description Report**

## **Dictionary**

* ID: This is the unique ID for each Car
* Levy: This is the imposition of tax by authority
* Manufacturer: Automobile producers having both small and large car manufacturers
* Model: Name used by manufacturer for the purpose of sale
* Country: Manufacturer hub location
* Prod. Year: The year when car was produced
* Category: This is classification based on built of the car
* Leather Interior: This is the leather seats in the car.
* Fuel Type: Petrol, Diesel, Hybrid, Electric, LPG, others
* Engine: This is a machine which is part of a car supporting in motion
* Turbo: A feature used to enhance car’s speed and power
* Mileage: Number of miles covered/traveled
* Cylinders: The number indicated the number of cylinders used in a car to produce power
* Gear Box: Transmissions used in automobile. The column indicates the type of gear box used
* Drive Wheels: Number of wheels used for driving
* Doors: Number of doors in the car
* Wheel: Number of wheels in a car
* Color: This is the different color the car is available in
* Airbags: A protection unit in the event of accident or crash to minimize damage to the driver
* Price: The estimated cost of the car

## 

## **Univariate Properties**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Type** | **Count** | **Missing** | **%** | **Unique** | **Min** | **Q1** | **Med** | **Q3** | **Max** | **Mean** | **SD** | **Skew** | **Kurt** |
| Manufacturer | String | 19237 | 0 | 0 | 65 |  |  |  |  |  |  |  |  |  |
| Model | String | 19237 | 0 | 0 | 1590 |  |  |  |  |  |  |  |  |  |
| Prod. year | Integer | 19237 | 0 | 0 | 54 | 1939 | 2009 | 2012 | 2015 | 2020 | 2010.91 | 5.67 | -2.08 | 11.33 |
| Category | String | 19237 | 0 | 0 | 11 |  |  |  |  |  |  |  |  |  |
| Leather interior | String | 19237 | 0 | 0 | 2 |  |  |  |  |  |  |  |  |  |
| Fuel type | String | 19237 | 0 | 0 | 7 |  |  |  |  |  |  |  |  |  |
| Engine | Float | 19237 | 0 | 0 | 65 | 0 | 1.8 | 2 | 2.5 | 20 | 2.31 | 0.88 | 2.19 | 19.22 |
| Turbo | String | 19237 | 0 | 0 | 2 |  |  |  |  |  |  |  |  |  |
| Mileage | Integer | 19237 | 0 | 0 | 7687 | 0 | 70139 | 126000 | 188888 | 2147483647 | 1532235.69 | 48403869.38 | 38.90 | 1598.76 |
| Cylinders | Integer | 19237 | 0 | 0 | 13 | 1 | 4 | 4 | 4 | 16 | 4.58 | 1.20 | 2.09 | 6.49 |
| Gear box type | String | 19237 | 0 | 0 | 4 |  |  |  |  |  |  |  |  |  |
| Color | String | 19237 | 0 | 0 | 16 |  |  |  |  |  |  |  |  |  |
| Airbags | Integer | 19237 | 0 | 0 | 17 | 0 | 4 | 6 | 12 | 16 | 6.58 | 4.32 | 0.08 | -1.33 |
| Price | Integer | 19237 | 0 | 0 | 2315 | 1 | 5331 | 13172 | 22075 | 26307500 | 18555.93 | 190581.27 | 136.47 | 18824.52 |

**Univariate Visualizations**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

# 

# **Data Exploration Report**

This report details the relationship between each potential feature with the label “Price.”

Summary Table

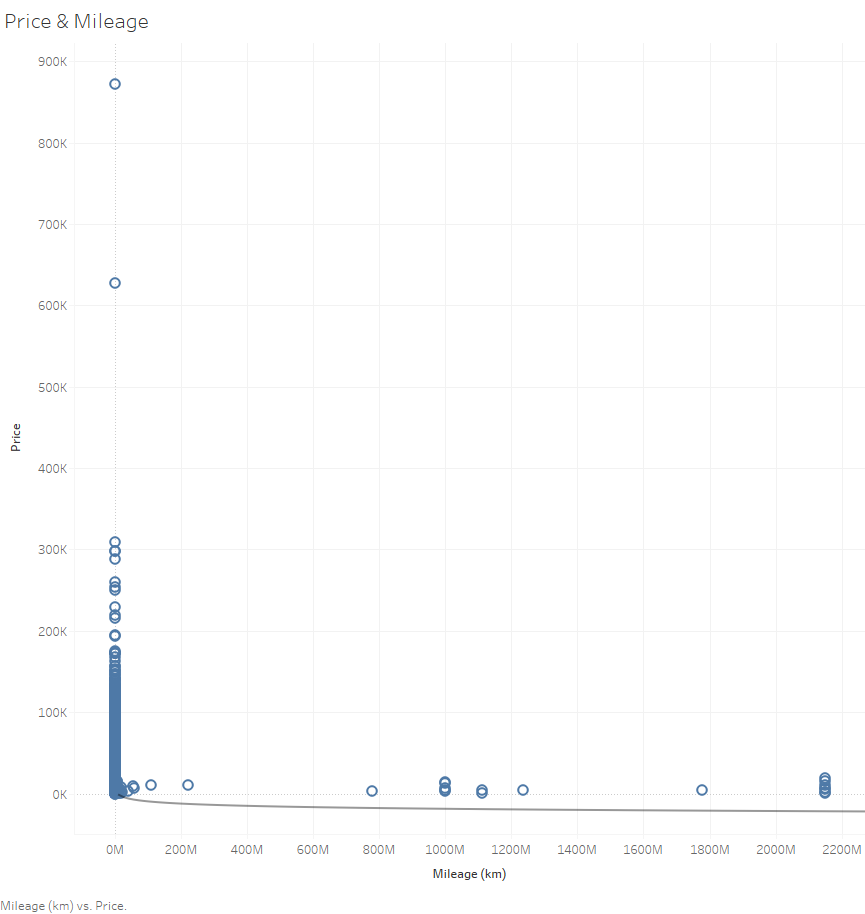
|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Analysis** | **Effect size** | **P-value** |
| Mileage | R squared | 0.044087 | < 0.0001 |
| Cylinders | R squared | 0.0150115 | < 0.0001 |
| Airbags | R squared | 0.0544687 | < 0.0001 |
| Manufacturer | *F* Stat | 70.83 | < 0.0001 |
| Category | *F* Stat | 106.05 | < 0.0001 |
| Leather Interior | *F* Stat | 278.8014457 | < 0.0001 |
| Fuel Type | *F* Stat | 176.0992026 | < 0.0001 |
| Turbo | *F* Stat | 182.36333904037 | < 0.0001 |
| Gearbox type | *F* Stat | 311.5590674 | < 0.0001 |
| Color | *F* Stat | 16.2 | < 0.0001 |

The remainder of the report includes greater details on each relationship. We find that all six features are worth including during the modeling phase.

## **Mileage**

**H10**: We expect Mileage to have a negative correlation with Price because consensus in used car condition is that the higher mileage a car has, the lower the car price will be due to higher maintenance cost.

**H11**: There is no relationship between Mileage and Car Price



**Equation: Price = -3976.77\*ln(Mileage (km)) + 63624**

R-Squared: 0.044087

Standard error: 18988

p-value (significance): < 0.0001

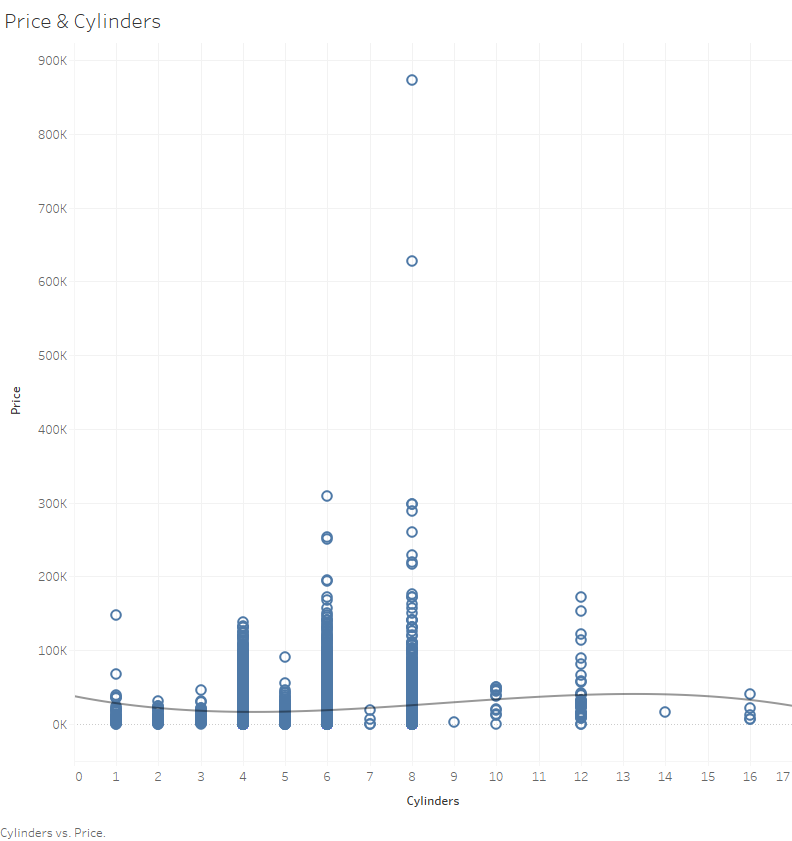
**Summary:** Mileage has a medium effect on Car Price because the R square value is not too high, however it is a very reliable result based on the P-value. An exponential formula is used to calculate the best R square value. The result suggests that there’s indeed a negative correlation between Mileage and Price because the formula shows when Mileage(x value) increases, Price(y-value) decreases.

## 

## **Cylinders**

**H20**: We expect Cylinders to have a positive correlation with Car Price as more cylinders are used for cars equipped with stronger engines for speed or heavier duty purpose.

**H21**: There’s no relationship between Car Price and Cylinders



**Equation: Price** = -66.5233\*Cylinders^3 + 1749.54\*Cylinders^2 + -11283.4\*Cylinders + 37595.7

**R-Squared**: 0.0150115

**Standard error**: 19638.8

**p-value (significance)**: < 0.0001

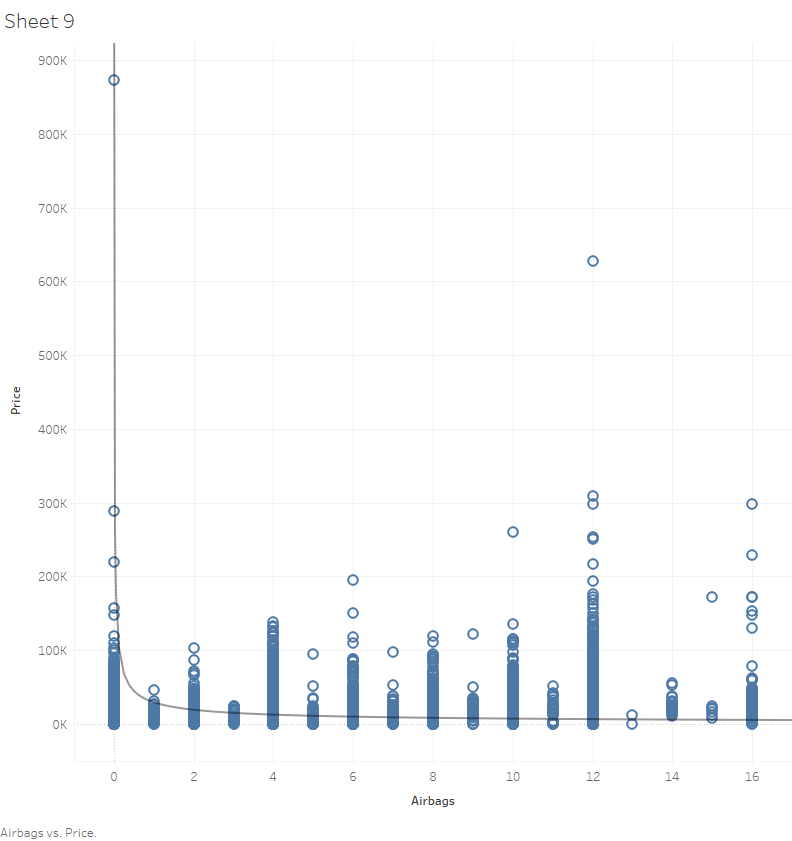
## 

**Summary:** Cylinder has a medium effect on Car Price as R square value is not too high, yet it is a very reliable result as the P value is <0.0001. Polynomial formula is used to calculate the best R square score. However, based on the result, it is difficult to determine whether a cylinder has positive or negative correlation with the Car Price due to the complexity of the formula that can yield both positive or negative results. Further research is needed to determine the relationship.

## **Airbags**

**H30**: We expect that Airbags will have a positive effect on Car Price as it is a desirable safety feature for many people. Thus, higher count of Airbags will increase Car Price

**H31:** There’s no relationship between Airbags and Car Price.



Equation: ln(Price) = -0.613678\*ln(Airbags) + 10.2815

R-Squared: 0.0544687

Standard error: 1.47735

p-value (significance): < 0.0001

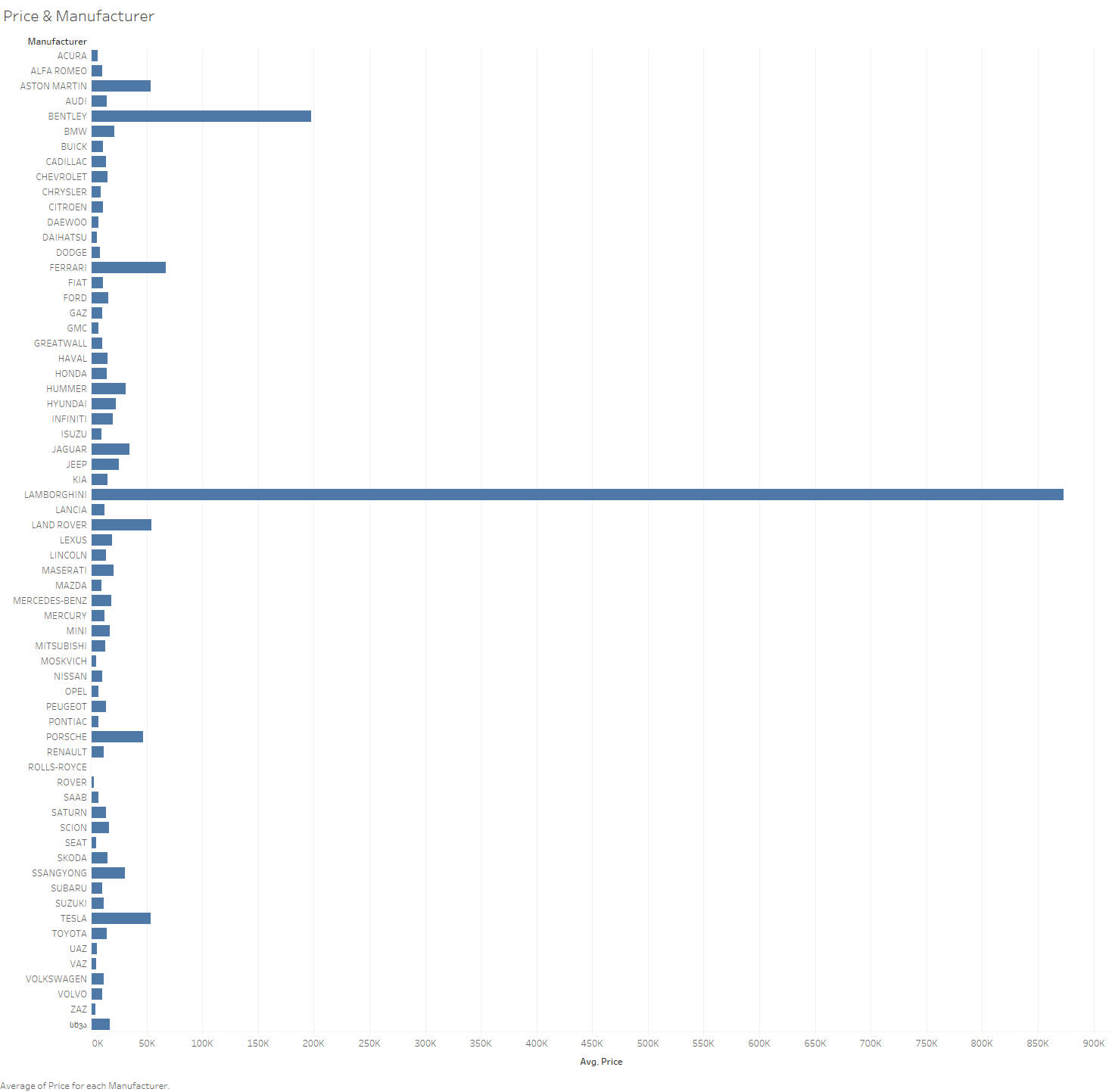
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Panes** |  | **Line** |  | **Coefficients** | |  |  |  |
| **Row** | **Column** | **p-value** | **DF** | **Term** | **Value** | **StdErr** | **t-value** | **p-value** |
| Price | Airbags | < 0.0001 | 16830 | ln(Airbags) | -0.61368 | 0.019709 | -31.1371 | < 0.0001 |
|  |  |  |  | intercept | 10.2815 | 0.03853 | 266.842 | < 0.0001 |

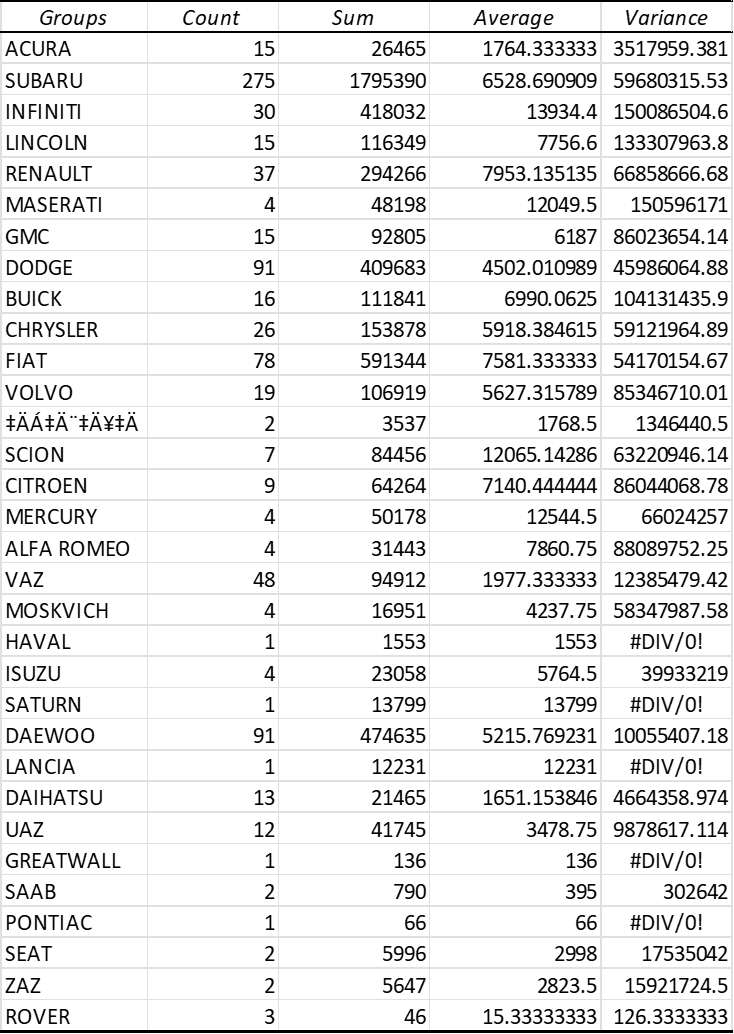
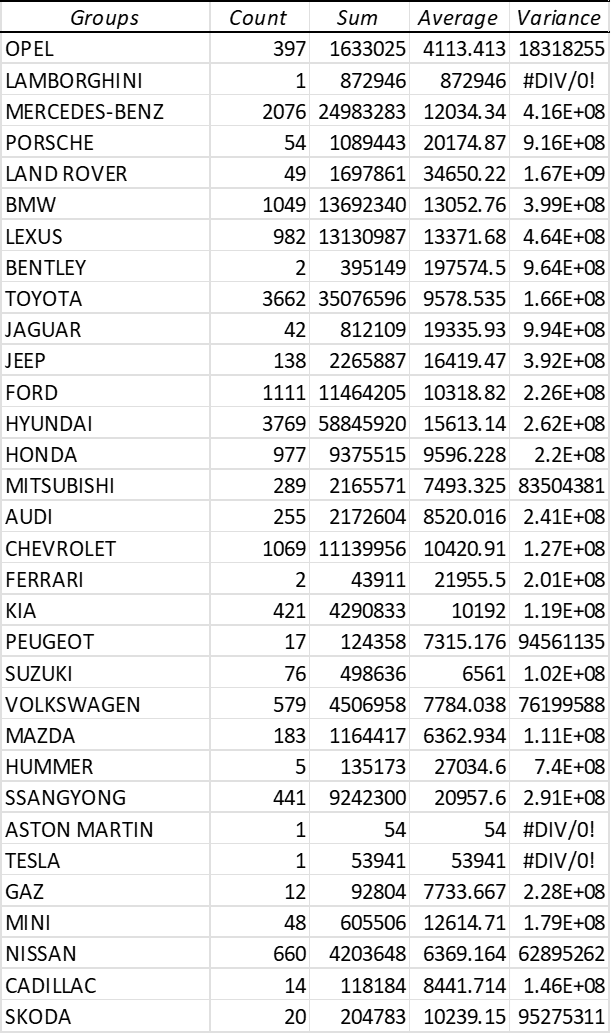
**Summary:** Airbags have a medium effect on Car Price as R square value is not too high with very reliable P value. Power formula was used to calculate the best R square value. Contrary to our belief, the result indicates there may be a negative correlation between Airbags and Car Price. Further research is required to to understand the relationship as there may be other variables affecting the results.

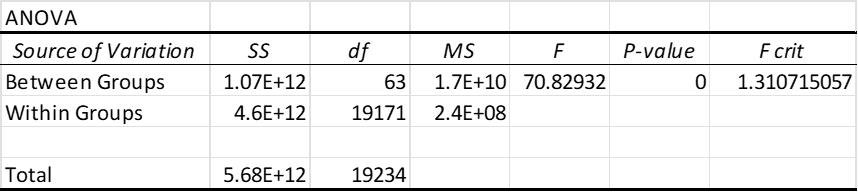
**C2N**

## **Manufacturer**

**H40:** We expect there's a strong correlation between Manufacturer and Car Price based on research that illustrates the brand name and its effect on car price. (Baltas, George, and Charalampos Saridakis, 2010)

****

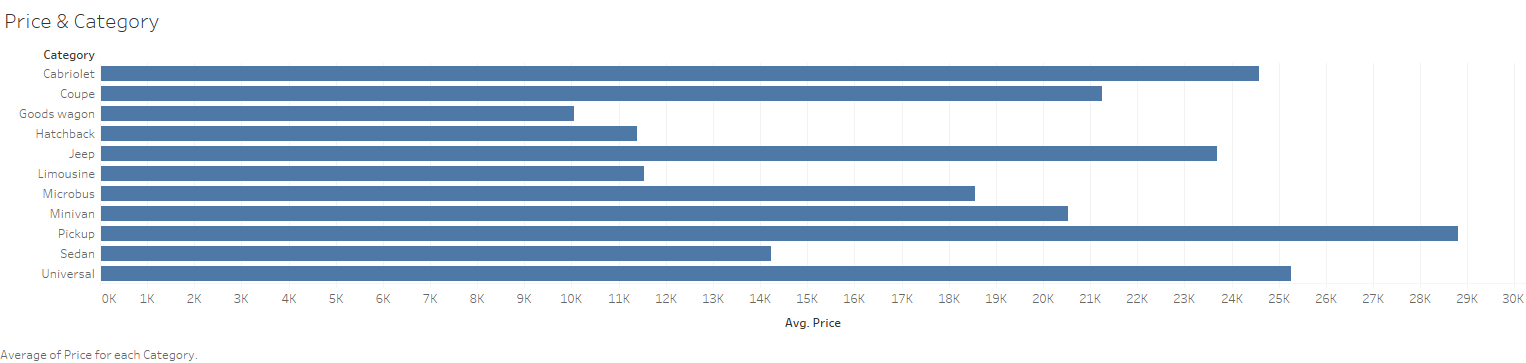
****

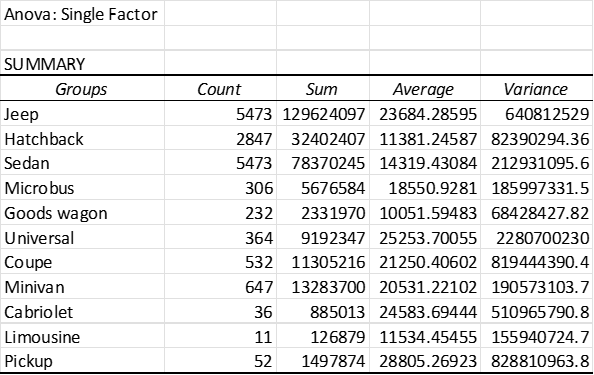
****

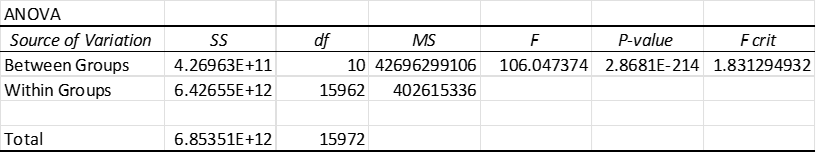
**Summary:** The F value suggests that Manufacturer has a medium effect on Car Price. P-value is >0.0001 which indicates it is a significant value. However, it is not as high a contributing factor as we expected, therefore we suspect there are more variables that contribute to the price than just manufacturers. Further research is needed to understand the relationship.

## **Category**

**H50**: We suspect Category has some correlation with Car Price as certain car categories cost more to manufacture and target different market segments.



****

****

**Summary:** The F value suggests that there’s some correlation between Category and Car Price. However, further analysis is required to understand the relationship as there may be other variables affecting the price more so than category.

## **Color**

There is no clear theory or any available resources to provide details on the relationship between color and price of a new car.

Hypothesis

**H60:** We expect that there’s no relationship between color and Car Price for a new car as the market is more driven by availability and other technical factors

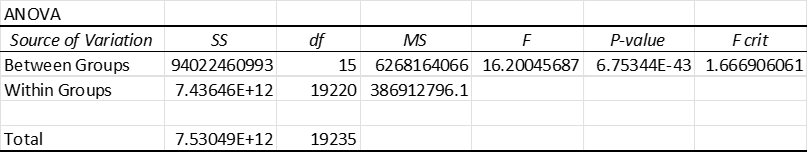
**H61**: Color will have a positive effect on Car Price

Chart, bar chart

Description automatically generated

Diagram

Description automatically generated



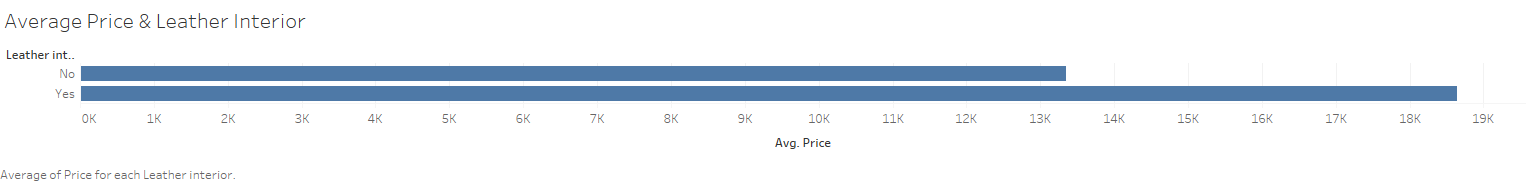
Summary: The F value suggests that there is some positive but low effect of car color on its price in the case of a new car. The p-value suggests that the color is a significant feature in determining the car price.

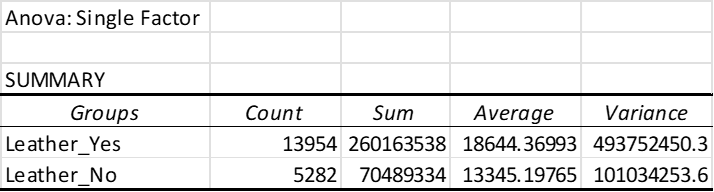
## **Leather Interior**

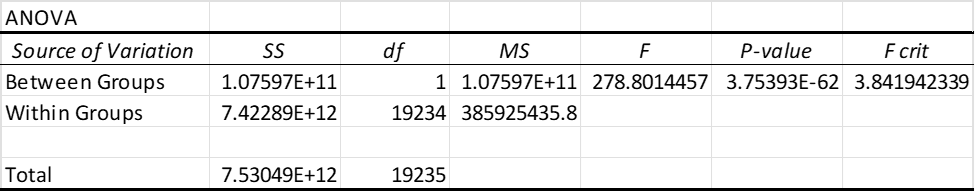
Hypothesis

**H70**: We expect that Leather Interior will have a positive effect on Car Price as it is perceived as a luxury option in the interior trim. Thus, such custom features will increase Car Price. (ELL, NBPOW, 2008)

**H71:** There’s no relationship between Leather Interior and Car Price.







Summary:

The F value suggests that there is a good positive effect of leather interior on car price. The p-value suggests that it is a significant feature in determining the car price as it is a custom feature and it is not usually available with all the cars. Hence, consumers expect vehicles with durable and attractive interiors which will increase the car price.

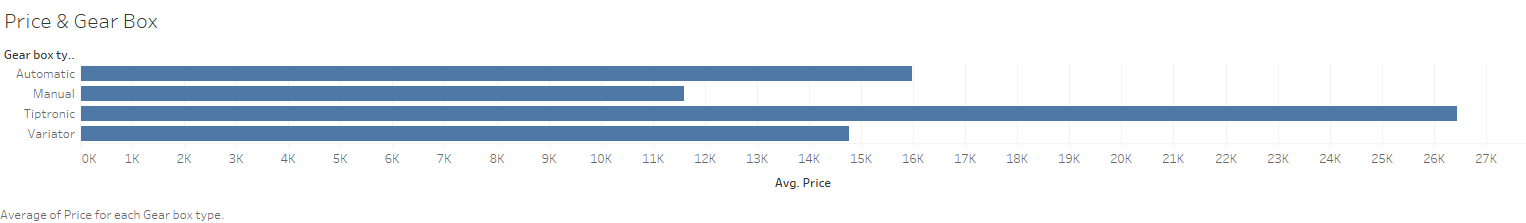
## 

## **Gear box type**

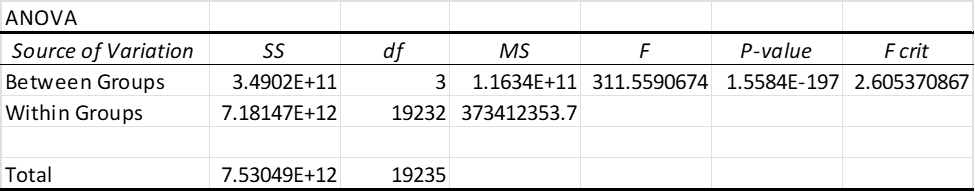
Hypothesis

**H80**: We expect that gearbox type will have a positive effect on Car Price as it is one of the most important factors which directly affects the driving experience. Thus, such features will increase Car Price (Proshchyna, Tetiana,2020)

**H81:** There’s no relationship between gearbox type and Car Price.







Summary:

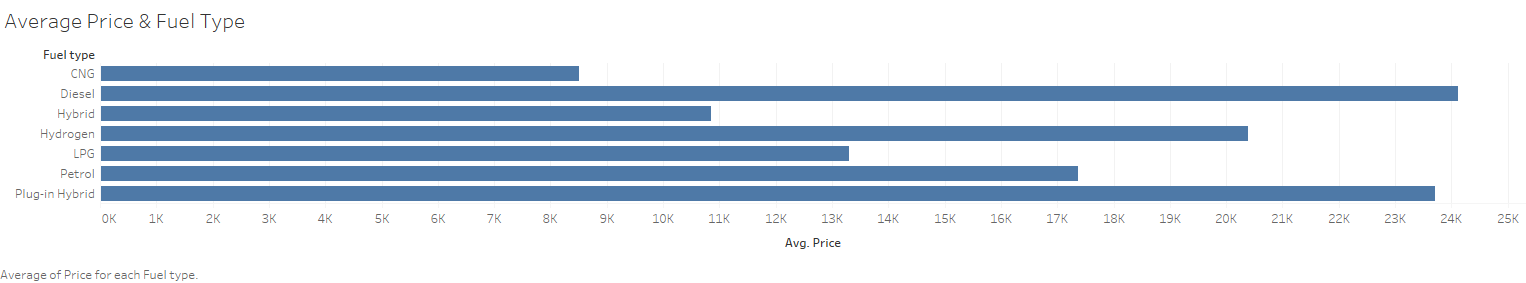
The F value suggests that there is a good positive effect of gearbox type on car price. The p-value suggests that the color is a significant feature in determining the car price as it has a direct impact on driving experience.

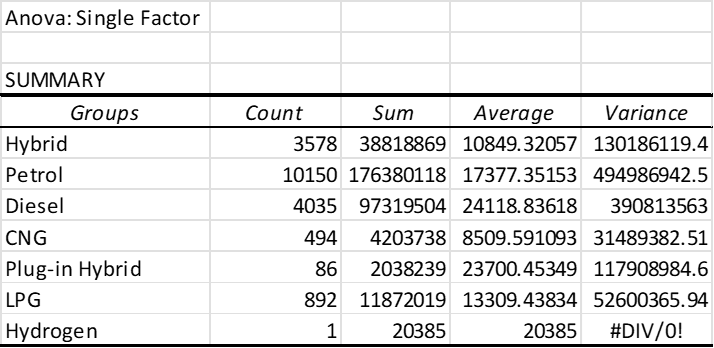
**Fuel Type**

Hypothesis:

**H90**: We expect that Fuel type will have a positive effect on Car Price as it is directly linked to the availability and cost of the fuel. This also has an impact on the environment. Thus, such features will have an impact Car Price (Van Vliet, Oscar PR, 2010)

**H91:** There’s no relationship between fuel type and Car Price.





Summary:

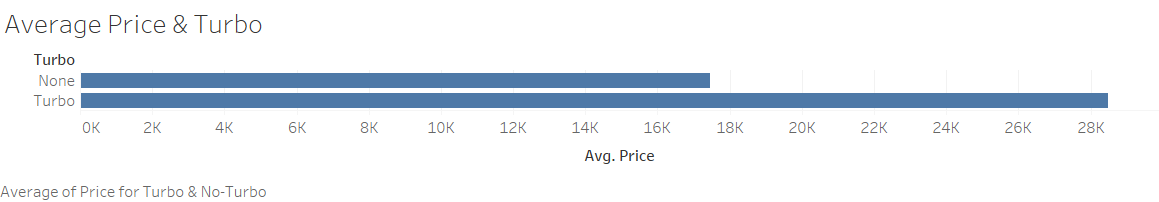
The F value suggests that there is a positive effect of fuel type on car price. The p-value suggests that the color is a significant feature in determining the car price. As the fuel costs are increasing and the impact of global warming is also increasing, people and governments are looking to shift to pocket friendly and environmentally friendly alternatives.

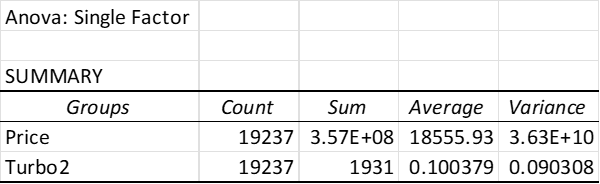
## **Turbo**

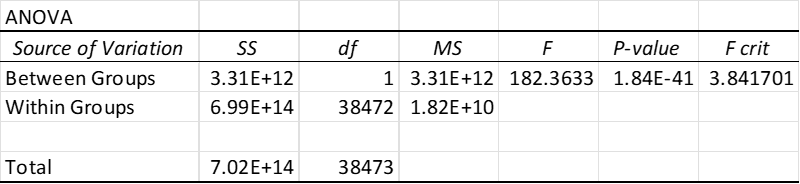
Hypothesis:

**H100**: We expect that Turbo will have a positive effect on Car Price as it is linked to the performance of the vehicle.

**H101:** There’s no relationship between Turbo and Car Price







Summary:

The F value suggests that there is a positive effect of Turbo on Car Price. The P-value suggests that Turbo is a significant feature in determining the Car Price. Whether or not a car has turbo appears to improve performance of the vehicle, which raises the price in which it demands.

# **Data Quality Report**

## **Missing Data**

Levy was missing around 80% of its data and had a very poor description of the feature from the source so it was removed. One other issue that we found was that the model names sometimes are described differently from model to model with difference due to language and different model trims. Although that may have been the case the data still had a higher permutation feature importance than that of standardized models so the varied model was kept. The rest of the data set was missing no data whatsoever so no further actions were taken to 1) delete cases/features, 2) replace missing values with a constant, or 3) impute missing values.

## **Outliers**

We encountered some outliers that were affecting our statistical analysis heavily. One example was a car with a $26 Million dollars price which had no clear explanation. As most price data points were in between $10k to $80k, it did not serve our purpose of business understanding. Therefore, we omitted it accordingly by using ‘Clip Value’ Pill in Azure and removed it directly from the data set in Excel for our Anova test.

## **Distributions**

Our initial data set of Price(label) showed much skewed distribution of 136.47. We decided to use Cube Root transformation to normalize the distribution which resulted in skewness value 1.32. We also used ln+1 transformation on Mileage(feature) to improve the distribution skewness from 38.9 to -3.63

### **Numeric Features**

Prod. Year, Engine, Mileage, cylinders, doors, airbags

The distributions of **Prod Year,Engine,cylinders,doors and airbags**  were not of any concern during analysis

Although the **Mileage** feature skewness was 38.9, the histogram revealed a level skewness that may benefit from being addressed in the Data Preparation Phase:

Based on shape of skewness we took natural ln+1 for the transformation and improved the skewness to -3.63

No concern on outliers in this dataset based on our tests and experiments during data understanding and preparation phase were observed.

## **Categorical Features**

Categorical features in the data set include **Manufacturers, Model, country, category, fuel type, Leather interior, turbo, gearbox type, Drive wheels, wheel and color**

Invalid records: Manufacturer and model columns had invalid data. However, it was 2 valid records and 9 invalid records respectively, so not an area of concern Therefore, we do not recommend any form of deleting, grouping or adjustments.

**Inferences and Business Decisions**

As mentioned above, as a business decision we decided to remove Levy from the model as it has more than 80% missing values. Throughout the exploratory data analysis and feature selection, we found that all other features were much more apt to predicting the price than Levy - additionally, the feature wasn’t given a definition by the creator of the dataset.

Our final model selection for this data set experiment to predict car prices was Boosted decision tree.

As per the results we were able to identify a few features which are critical to predict if a customer will buy a car. Few of the relevant features are listed below:

**Airbags, Prod. Year, Gear box type, engine, mileage, leather interior, turbo, manufacturer**

While our model tested with a relatively high coefficient of determination and low p-value, it can be suggested that not all relevant features needed to precisely predict a car’s price were present in our dataset - other models that predict a car’s value, such as Kelly Blue Book, appear to be much more accurate in their predictions. Further data discovery would be necessary to compete. However, depending on the goal of the end-user, our model would be a fast and easy prediction calculator that could be used to understand how different features interact and affect the price of a car.

# **References:**

Avorn, Jerry. "Benefit and cost analysis in geriatric care: turning age discrimination into health policy." *New England Journal of Medicine* 310.20 (1984): 1294-1301.

Baltas, George, and Charalampos Saridakis. "Measuring brand equity in the car market: a hedonic price analysis." *Journal of the Operational Research Society* 61.2 (2010): 284-293.

ELL, NBPOW. "Design of automotive interior textiles." *Textile advances in the automotive industry* (2008): 113.

Hebert, Amy, et al. “Buying a New Car.” *Consumer Advice*, 7 Mar. 2022, https://consumer.ftc.gov/articles/buying-new-car.

Van Vliet, Oscar PR, et al. "Techno-economic comparison of series hybrid, plug-in hybrid, fuel cell and regular cars." *Journal of power sources* 195.19 (2010): 6570-6585.

Proshchyna, Tetiana. *Approved by \_*. Diss. Kyiv School of Economics, 2020.