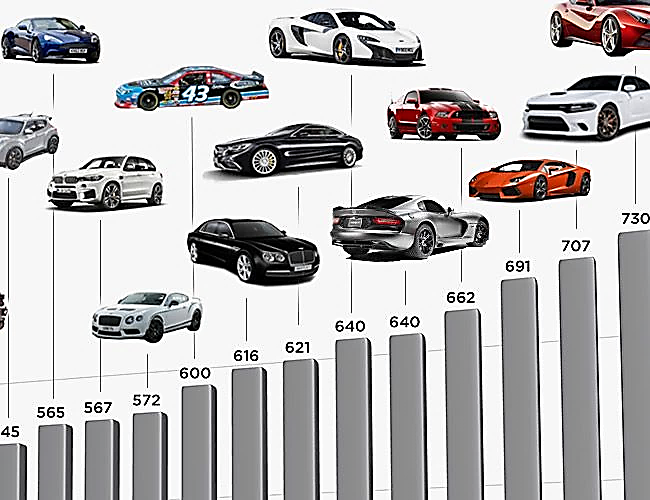
Research investigation on finding a correlation between price and horsepower of a Car

Module: Team Research and Development Project, 7COM1079



January 2021, University of Hertfordshire

Group No. 18

Team SRN’s

19021342 | 19024901 | 19025931 | 19054390 | 20006670

Module Leader

Dr. John Noll

**Table of Contents**

[Introduction 1](#_Toc61011794)

[Literature Review 1](#_Toc61011795)

[Research Question 2](#_Toc61011796)

[Analysis 2](#_Toc61011797)

[Descriptive statistics 2](#_Toc61011798)

[Inferential statistics 3](#_Toc61011799)

[Non-parametric analysis 3](#_Toc61011800)

[Data Visualization 4](#_Toc61011801)

[Conclusion 6](#_Toc61011802)

[References 7](#_Toc61011803)

[Table 1 Attributes types in data analysis 2](#_Toc61024587)

[Table 2, Descriptive Statistics 3](#_Toc61024588)

[Table 3, T-test statistics 3](#_Toc61024589)

[Table 4, Correlation Methods 4](#_Toc61024590)

[Figure 1, Histogram of the frequency distribution of dependent variable price 4](#_Toc61608789)

[Figure 2, Histogram of the frequency distribution of independent variable horsepower 5](#_Toc61608790)

[Figure 3, scatter plot between price and horsepower of a car 5](#_Toc61608791)

# **Introduction**

(Enis Gegic, 2019) , A car price forecast was a high-interest research area, as it requires significant effort and expertise from the field expert. In this report research is conducted into finding a relationship between the horsepower and the price of a car, to achieve this, descriptive statistics and reasonable data analysis are used. Various factors that influence the relationship between horsepower and the price of a car are also studied. Predicting the price of cars by using the available independent features like horsepower, engine base, wheel size, etc., which eventually help the manufacturing companies to build accordingly and manipulate the design of the car, business strategy, etc., to meet certain price levels. So further the model will be good to understand the pricing dynamics of the new market. In this report, we use horsepower and the price of a car to find the correlation between these independent and dependent variables.

# **Literature Review**

**Journal Article: Car price prediction using machine learning techniques**

(Enis Gegic, 2019), the research investigates predicting the price of used cars by apply supervised machine learning algorithms like linear regression, naïve Bayes, etc., It is trivial to find the price of a car since there are many factors that impact the dependent variable such as mileage, horsepower, etc., Due to rising fuel prices, fuel economy is also an important factor in influencing the price of a car. So different techniques are used to forecast the price of the car and need to compare the performance of each one and go with the technique which has high accuracy.

**Journal Article: Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy**

(Stefan Lessmann, 2017)**,** To predict the resale prices of used vehicles, the paper is concerned with statistical models. To investigate how various degrees of freedom in the modeling process lead to forecast accuracy, an empirical study is performed. First, a comparative study of alternative prediction methods shows that in resale price forecasting, random forest regression is especially effective. It is also shown that it is necessary to avoid the use of linear regression, the dominant approach in previous work. Second, empirical findings demonstrate the existence of heterogeneity in the forecast of resale prices and define techniques that can automatically resolve its detrimental impact on resale prices.

**Journal Article: How much is my car worth? A methodology for predicting used cars prices using Random Forest**

(Nabarun Pal, 2018), A significant and much-needed issue to be discussed is forecasting the prices of used vehicles. Customers can be abused thoroughly by repairing The prices of used cars are unsustainable and many fall into the trap. The absolute requirement of a used car price increases, thus, Method of prediction to assess effectively the worthiness of There are a number of characteristics of a vehicle, such as the age of the car, its make, the origin of the car (the original country), Of the producer), its km and its horsepower. Other considerations, such as The type of fuel that it uses, its style, its braking mechanism, its volume (measured in cc) cylinders, acceleration, number of doors.

# **Research Question**

***Question: Is there any correlation between price and horsepower of a car.***

***Null hypothesis: There is no correlation between price and horsepower of a car.***

***Alternative hypothesis: There is a correlation between price and horsepower of a car.***

Am investigating car price prediction, features which are horsepower and price which impact the cost of a car. So here the technique which is used is the T-test and correlation technique to analyze the car data set which is taken from Kaggle (Kumar, 2019). So, there are 26 features that are available in the dataset but here we are using only 2 features which are horsepower(independent) and price(dependent). We have 206 observations with 26 features which are quite fewer numbers but still, the data is enough to proceed the further investigation.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Type of variable | Nature of variable | Measurement of variable |
| price | Dependent variable | Numerical | Interval |
| horsepower | Independent variable | Numerical | Interval |

Table 1 Attributes types in data analysis

(Zimmerman, 2012) The t-test is used to determine whether the coefficient of correlation varies significantly from zero and, thus, there is proof of a relationship between the two variables. And there is the implicit premise that the data is randomly sampled from a normal distribution. A p-value of less than 0.05 (typically ≤ 0.05) is statistically significant. A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis.

(Schober, Boer, & Schwarte, 2018) In data analysis and modeling, correlations can be useful to better understand the relationships between variables. The association of the statistical relationship between two variables is called correlation. A correlation may be positive(+1), meaning that both variables move in the same direction, and negative(-1), meaning that the values of the other variable decrease when the value of one variable increases. There may also be neutral or zero correlation(0), which indicates that the variables are unrelated.

# **Analysis**

## **Descriptive statistics**

In order to make it easier to grasp the results, Descriptive Statistics summarizes the data at hand by numbers such as mean, median, mode, etc. Beyond what is available, it does not require any generalization or inference. This implies that only the representation of the data (sample) available is descriptive statistics and not based on any probability theory.

In table 2. Provides the descriptive statistics of dependent and independent variable so-called price and horsepower of a car.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | N | Mean | Standard deviation |
| horsepower | 205 | 104.117 | 39.4 |
| price | 205 | 13276.7106 | 7970 |

Table 2, Descriptive Statistics

## **Inferential statistics**

Two sample T-Test statistics:

In statistical hypothesis testing, a two-sample test is a test carried out on the basis of data from two random samples, each of which is independently collected from a separate population. The aim of the test is to decide if the difference is statistically significant between these two populations. performed a T-test on the dataset on features horsepower and price and found that p-value has 2.2e-16(which is a very small number 2.2 \* 10^-16) which means it is smaller than 0.05 p-value. So that we reject the null hypothesis. That means alternative hypothesis: true difference in means is not equal to 0.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test | t | df | p | Mean difference in horsepower | Mean difference  of price | |
| Two sample T-test | 23.608 | 205 | 2.2e-16 | 104.1171 | | 13276.7106 |

Table 3, T-test statistics

Analysis of correlation is a statistical methodology used to measure the intensity of the association between two quantitative variables. A high correlation means that there is a close association between two or more variables, while a weak correlation means the variables are barely related.

## **Non-parametric analysis**

Since the data is not a normal distribution, so we need to use non-parametric analysis and we use the 'spearman correlation technique' to find the correlation between two features. I have done a correlation between the price and horsepower of a car and got a value of 0.85. And this means it is strongly correlated with each other and has a positive correlation. It means one variable increases other variables increases. In Table 4 we calculated 3 types of correlation methods such as spearman (Wikipedia, n.d.), Kendall, Pearson.

So as the p-value is less than the significance value it clearly states that there is a significant correlation between the price and horsepower of a car.

Where spearman and Kendall are non-parametric analysis methods work for generally not normal distribution data or non-linear data, whereas Pearson used for normal distribution data or Gaussian distribution.

|  |  |
| --- | --- |
| Correlation Methods | |
| Spearman | 0.85 |
| Kendall | 0.67 |
| Pearson | 0.80 |

Table 4, Correlation Methods

# **Data Visualization**

(Tableau, n.d.) The graphical representation of information and data is the visualization of data. Data visualization tools offer an accessible way of seeing and interpreting trends, outliers, and patterns in data by using visual elements such as charts, graphs, and maps.

Chart, histogram

Description automatically generated

Figure 1, Histogram of the frequency distribution of dependent variable price

Figure 1 Represents the histogram of the dependent variable price. The curve above shows the data is not a normal distribution. Also, around 0-10000 $ range the frequency is a bit high than other price ranges. And the frequency is very low when the price is increasing. In the above graph, we can clearly state that people are interested in buying an economic price car.

Chart, histogram

Description automatically generated

Figure 2, Histogram of the frequency distribution of independent variable horsepower

The above histogram represents the independent variable horsepower. It shows that at around 50-100 range horsepower the frequency is a bit high compared to other horsepower. When horsepower increases the frequency is gradually decreasing which means people are looking for economic price ranges.

Chart, scatter chart

Description automatically generated

Figure 3, scatter plot between price and horsepower of a car

The above Figure 3 explains the correlation between price and horsepower of a car. The linear line shows that if horsepower increases then the price also increases. So, there is a positive correlation between them. We can also see some outliers in the data where horsepower in range 150-200 the price is around 15000-20000 dollars but for the same horsepower, few observations seem the price is around 40000 dollars.

# **Conclusion**

To validate the research of correlation between price and horsepower of a car, I have performed few tests namely the T-test and Spearman correlation test and established the p-value which is less than 0.05 so here we reject the null hypothesis and accept the alternative hypothesis. And also got the Spearman correlation value as 0.85 which means there is a strong positive correlation between the features price and horsepower of the car. I overviewed these factors but other several factors also impact the price of the car such as mileage, specifications, etc. that aren't reviewed in this research. Going forward stronger relationships can be predicted using other factors and analyze properly using machine learning techniques.

# **References**

Eduardo Gonzálezn, A. C. (2015). How car dealers adjust prices to reach the product efficiency frontier in the Spanish automobile market. *Omega*, 38-48.

Enis Gegic, B. I. (2019). Car price prediction using machine learning techniques. *researchgate*.

Kumar, M. (2019). *kaggle*. Retrieved from Kaggle: https://www.kaggle.com/hellbuoy/car-price-prediction

Nabarun Pal, P. A. (2018). How much is my car worth? A methodology for predicting used car prices using Random Forest. *Future of Information and Communications Conference (FICC)*.

Schober, P. M., Boer, C. P., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *International Anesthesia Research Society*.

Stefan Lessmann, S. V. (2017). Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *researchgate*.

*Tableau*. (n.d.). Retrieved from Tableau: https://www.tableau.com/en-gb/learn/articles/data-visualization

*Wikipedia*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Spearman%27s\_rank\_correlation\_coefficient

Zimmerman, D. W. (2012). Correcting Two-Sample z and t Tests for Correlation:An Alternative to One-Sample Tests on Difference. *Psicológica*.