**STATISTICS FOR DATA SCIENCE**

**Assignment-I**

**“ Luxury Car Performance**

**And**

**Price Analysis “**

Submitted by:

Name & PRN of student

*22030121023*

*[ANANYA PATTNAIK]*

BCA Students, Sem 3

Symbiosis Institute of Computer Studies & Research

Symbiosis International (Deemed) University

Model colony, Atur Centre,

PUNE - 411016

2022-25 Batch

**ACKNOWLEDGMENT**

I would like to extend my sincere gratitude to the following individuals and entities for their invaluable support and contributions to this project:

I express my heartfelt appreciation to ***Dr. Shilpa Mujumdar*** for providing the necessary infrastructure and resources for this project.

My gratitude goes to the *Director of the Symbiosis Institute of Computer Studies and Research (SICSR) and the* *Deputy Director* for their guidance and encouragement throughout the course of this project.

I would like to thank the *Admin Officer* for their assistance in administrative matters related to this project.

A special thanks to the *course in charge* for their guidance and mentorship during the project.

I acknowledge the sources of data used in this project, including websites, books, magazines, and other references that have been cited appropriately. The data from these sources was instrumental in shaping the content and analysis presented in this project.

**I** would also like to thank any other individuals or organizations who have directly or indirectly contributed to the successful completion of this project.

This project would not have been possible without the collective support and collaboration of all these individuals and entities. Thank you for your unwavering support and encouragement.

[ANANYA PATTNAIK]

**INDEX**

|  |  |  |
| --- | --- | --- |
| SR NO | CONTENTS | PAGE NO |
| 1 | Abstract | 4 |
| 2 | Introduction | 5 |
| 3 | Methodology | 7 |
| 4 | Data Analysis | 8 |
| 5 | Conclusions | 18 |
| 6 | References | 19 |
| 7 | Annexure | 20 |

**Abstract**

Statistical tools are used to analyze the data on the “Luxury Cars Price and Performance Analysis”. We have used used excel methods like bar chart, line graph, etc.. , Arithmetic Mean, Geometric Mean, Harmonic Mean, Standard Deviation, Variance, Coefficient of Variation, Coefficient of kurtosis, Correlation, for data analysis,.

We found that The dataset is NON NORMAL because it lacks symmetry, has have extreme values, or has a steeper “dome” other than a typical bell. It is leptokurtic(more than 7) and has extreme outliers.

Skewness for horsepower, Torque and Price is positive(distribution of hp,torque and price is positively skewed) while the same for year is negative (the distribution of year is negatively skewed).

The entire distribution of all the columns is Leptokurtic which implies that the kurtosis for all the four columns is greater than zero.

As mentioned in the above (iii) point table there is a high positive correlation between horsepower and torque(0.9377…),Moderate positive or considerable correlation between horsepower and price(0.7855…), torque and price(0.7624…),poor or negligible correlation between year and horsepower(0.0123..),year and torque(0.0343…) but there is a negative correlation between year and price(-0.0125..)

Slope is POSITIVE(1.139…) which shows a direct relation between torque and horsepower i.e. if torque increases(x) then horsepower(y) also increases .

**Introduction**

This mini-project serves as a critical component of our Elements of Statistics course, providing a practical application of statistical analysis techniques. The selection of this topic, "Luxury and Exotic Car Performance and Pricing " stems from a collective interest in the automotive industry and its intriguing interplay between performance metrics and pricing.

**Objective of the Project:** The primary objective of this project is to conduct a comprehensive statistical analysis of a dataset containing essential attributes of luxury and exotic cars. By examining variables such as year, horsepower, torque, and price, we aim to derive valuable insights into the factors that influence the pricing of these high-end automobiles.

**Data Collection:** The dataset used in this project was primarily collected from www.kaggle.com The data encompasses the period between 2021 and 2022, offering a snapshot of the latest models in the luxury car market.

**Variables Considered:** The dataset incorporates the following variables:

1. **Car Make-Model**: This column represents the make and model of each car.
2. **Year**: This column indicates the manufacturing year of the car.
3. **Horsepower**: This is the measure of the engine's power output in horsepower (HP), which is an important factor in determining a car's performance.
4. **Torque (lb-ft)**: Torque is the rotational force produced by the engine, measured in pound-feet (lb-ft). It's also a significant factor in a car's performance, especially in acceleration.
5. **Price (in USD)**: This column represents the price of each car in US dollars, which is a crucial factor for potential buyers.

**Population and Sample:** Our population of interest comprises luxury and exotic cars available in the market during the specified period. As obtaining data for the entire population is impractical, we opted for a **non-random sample** consisting of select high-end car models to perform our analysis.

**Sampling Method:** The method of sampling used is **Simple Random Sampling.** The selection of car models for our dataset was purposive, focusing on renowned luxury brands and high-performance vehicles.

**Type of Data:** The dataset is considered relatively small in scale, given the limited number of observations(60 records only) , which allows for detailed and focused analysis.

**Related Work:** In the same domain, we've consulted research articles on EBSCO Host to uncover vital insights into the connection between performance attributes and pricing strategies in the luxury and exotic car market. These findings have shaped our project, highlighting:

* High-performance features justify premium prices.
* Luxury car brands emphasize exclusivity and craftsmanship.
* Evolving market dynamics influence pricing and marketing strategies.

These articles have offered valuable insights and formed the basis for our project's methodology and approach.

With these considerations, our project aims to understand why luxury and exotic cars cost so much by looking at how their performance features, like speed and power, are connected to their high prices. We want to figure out what makes these cars so expensive.

**Methodology for data analysis**:

In our data analysis process, we employed a combination of tools and techniques to gain meaningful insights from the dataset. The primary tool used for data analysis was Microsoft Excel, which allowed us to efficiently perform various calculations and visualize trends. Here's a breakdown of the tools and methods utilized:

**Calculation of Measures:**

* + **Mean (Geometric and Harmonic Mean):** We calculated the geometric mean (GM) and harmonic mean (HM) to understand the central tendency of the data.
  + **Standard Deviation:** Excel was used to compute standard deviations, helping us assess the dispersion or variability in the data.

**Statistical Software:**

* + Microsoft Excel was used as our statistical software of choice due to its versatility and user-friendly interface. It allowed us to perform calculations efficiently and accurately.

**Correlation and Covariance Analysis:**

* + We utilized Excel's built-in functions to calculate the correlation coefficient, which helped us assess the degree of linear association between variables.

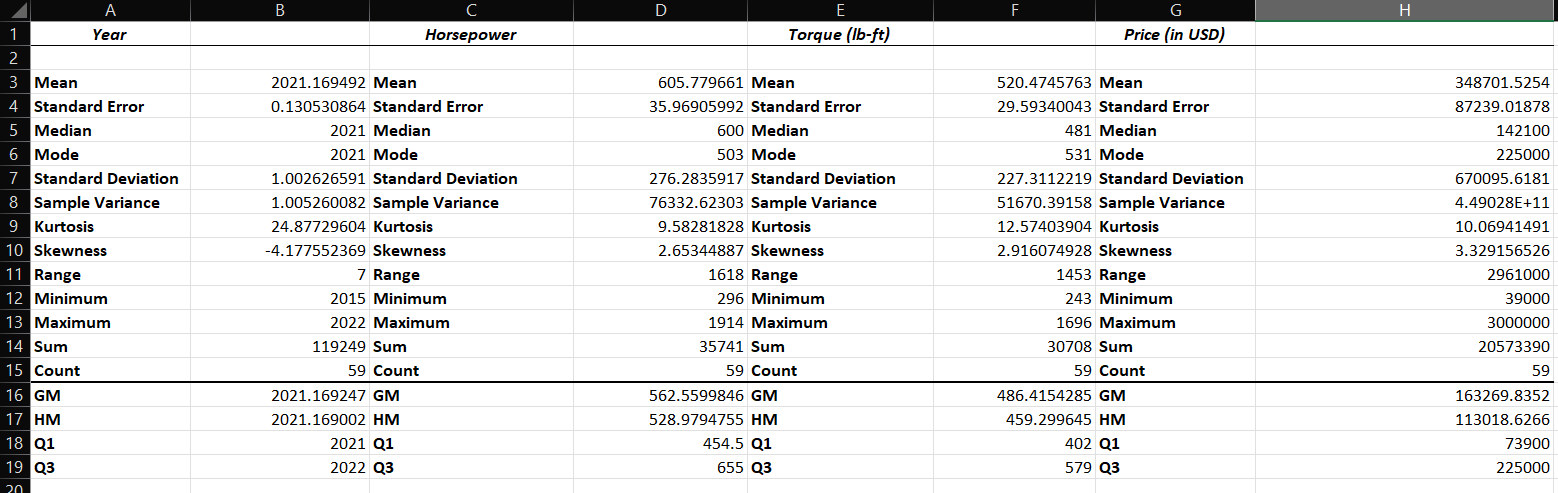
**Graphical Representation:**

* + We created various graphs and charts in Excel, including scatter plots, histograms, and bar charts, to visually represent and interpret the data. These visuals enhanced our understanding of trends and relationships within the dataset.

Overall, Microsoft Excel played a pivotal role in our data analysis process, allowing us to perform calculations, visualize patterns, and draw meaningful conclusions from the dataset efficiently.

**Data Analysis**

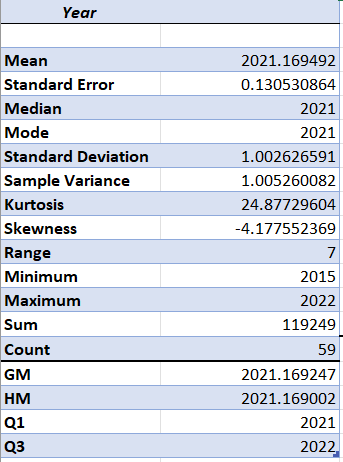
In this step, we create graphs and summary tables for the dataset as provided which involves using data analysis tools like Microsoft Excel



Mentioned above are the summary tables of averages, dispersion, skewness, kurtosis, correlation etc. This table provides a comprehensive overview of the statistics for the four variables, helping to understand the central tendency, variability, and distribution of the data for each variable.

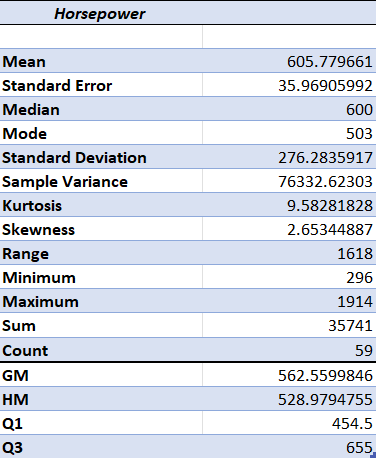
***DETAILED ANALYSIS***

**1.FOR YEAR COLUMN:**



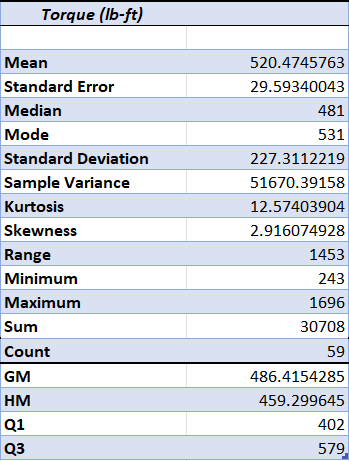
1. Mean (Average): The average year in the dataset is approximately 2021.17. This suggests that the dataset is centered around the year 2021.
2. Median: The median year is 2021, which aligns with the mean, indicating that the data is not heavily skewed by extreme values.
3. Mode: The mode is 2021, which means that 2021 is the year that appears most frequently in the dataset.
4. Standard Deviation: The standard deviation (approximately 1.00) measures the dispersion or spread of the data around the mean. In this case, it indicates that the years are relatively close to the mean year of 2021.
5. Sample Variance: The sample variance (approximately 1.01) quantifies the variability in the data. It's slightly larger than the square of the standard deviation, reinforcing the notion of some variability in the years.
6. Kurtosis: **(LEPTOKURTIC (KURTOSIS>0))**  The kurtosis value (24.88) is quite high, suggesting that the distribution of years has heavy tails and more extreme values than a normal distribution. This indicates the presence of outliers or years that deviate significantly from the mean.
7. Skewness: The negative skewness value (-4.18) indicates that the distribution is skewed to the left, meaning it has a longer tail on the left side. This aligns with the presence of outliers or earlier years in the dataset.
8. Geometric Mean (GM): The geometric mean (2021.17) is similar to the arithmetic mean, indicating that the data does not exhibit strong skewness.
9. Harmonic Mean (HM): The harmonic mean (2021.17) is also close to the arithmetic mean, indicating a relatively balanced distribution.
10. Q1 and Q3: The first quartile (Q1) is 2021(which means 25 percent of cars are made in or before 2021), and the third quartile (Q3) is 2022(75% are made before 2022). This shows that the middle 50% of the data falls within this year range.

**2.FOR HORSEPOWER COLUMN:**



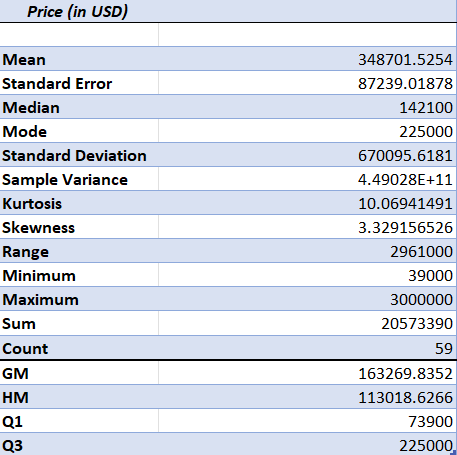
1. Mean (Average): The mean horsepower in the dataset is approximately 605.78, indicating that, on average, the cars in the dataset have around 605.78 horsepower.
2. Median: The median horsepower is 600, which closely aligns with the mean. This indicates that the distribution of horsepower is relatively symmetric and not heavily influenced by extreme values.
3. Mode: The mode is 503, indicating that 503 horsepower is the most frequently occurring value in the dataset.
4. Standard Deviation: The standard deviation (approximately 276.28) measures the spread or variability of horsepower values around the mean. The relatively high standard deviation suggests that there is significant variability in the horsepower of the cars in the dataset.
5. Sample Variance: The sample variance (approximately 76,332.62) quantifies the extent of variability in the dataset. The large variance value reinforces the presence of significant differences in horsepower among the cars.
6. Kurtosis: **(LEPTOKURTIC (KURTOSIS>0))** The high kurtosis value (approximately 9.58) indicates that the distribution of horsepower values has heavy tails and more extreme values than a normal distribution. This suggests the presence of outliers or cars with exceptionally high or low horsepower.
7. Skewness: The positive skewness value (approximately 2.65) indicates that the distribution is skewed to the right, meaning it has a longer tail on the right side. This suggests that there are cars with higher-than-average horsepower values.
8. Geometric Mean (GM): The geometric mean (approximately 562.56) provides an alternative measure of central tendency, considering the multiplicative nature of data.
9. Harmonic Mean (HM): The harmonic mean (approximately 528.98) is another measure of central tendency that is sensitive to outliers.
10. Q1 and Q3: The first quartile (Q1) is 454.5, and the third quartile (Q3) is 655, indicating the range within which the middle 50% of the data falls.

**3.FOR Torque (lb-ft) COLUMN:**



1. **Mean (Average):** The mean torque is approximately 520.47 lb-ft, indicating that, on average, the cars in the dataset have around 520.47 lb-ft of torque.
2. **Median:** The median torque is 481 lb-ft, which is slightly lower than the mean. This indicates that the distribution of torque values is slightly right-skewed, with some higher values pulling the mean upward.
3. **Mode:** The mode is 531 lb-ft, indicating that 531 lb-ft of torque is the most frequently occurring value in the dataset.
4. **Standard Deviation:** The standard deviation (approximately 227.31) measures the spread or variability of torque values around the mean. The standard deviation suggests that there is significant variability in torque among the cars in the dataset.
5. **Sample Variance:** The sample variance (approximately 51,670.39) quantifies the extent of variability in the dataset. The large variance value reinforces the presence of significant differences in torque among the cars.
6. **Kurtosis:** The high kurtosis value (approximately 12.57) indicates that the data is **LEPTOKURTIC (KURTOSIS>0)** and distribution of torque values has heavy tails and more extreme values than a normal distribution. This suggests the presence of outliers or cars with exceptionally high or low torque.
7. **Skewness:** The positive skewness value (approximately 2.92) indicates that the distribution is skewed to the right, meaning it has a longer tail on the right side. This suggests that there are cars with higher-than-average torque values.
8. **Geometric Mean (GM):** The geometric mean (approximately 486.42) provides an alternative measure of central tendency, considering the multiplicative nature of data.
9. **Harmonic Mean (HM):** The harmonic mean (approximately 459.30) is another measure of central tendency that is sensitive to outliers.
10. **Q1 and Q3:** The first quartile (Q1) is 402 lb-ft, and the third quartile (Q3) is 579 lb-ft, indicating the range within which the middle 50% of the data falls.

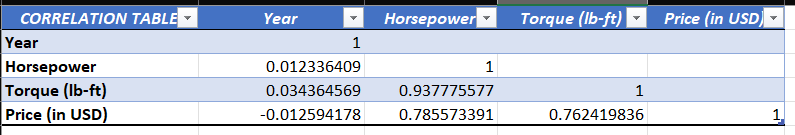
**4.FOR Price (in USD) COLUMN:**

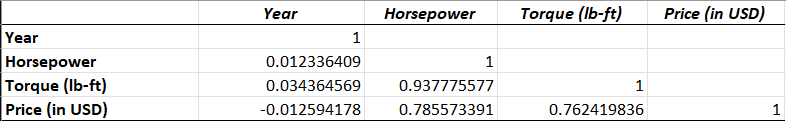


1. Mean (Average): The mean price is approximately $348,701.53, indicating that, on average, the cars in the dataset have a price of approximately $348,701.53.
2. Median: The median price is $142,100, which is considerably lower than the mean. This indicates that the distribution of prices is right-skewed, with some significantly higher-priced cars pulling the mean upward.
3. Mode: The mode is $225,000, indicating that $225,000 is the most frequently occurring price in the dataset.
4. Standard Deviation: The standard deviation (approximately $670,095.62) measures the spread or variability of price values around the mean. The high standard deviation suggests that there is substantial variability in car prices among the cars in the dataset.
5. Sample Variance: The sample variance (approximately $449,028,000,000) quantifies the extent of variability in the dataset. The large variance value indicates significant differences in car prices.
6. Kurtosis: The high kurtosis value (approximately 10.07) indicates that the data is ***leptokurtic*** and distribution of price values has heavy tails and more extreme values than a normal distribution. This suggests the presence of outliers or cars with exceptionally high prices.
7. Skewness: The positive skewness value (approximately 3.33) indicates that the distribution is strongly skewed to the right, meaning it has a longer tail on the right side. This suggests that there are cars with much higher-than-average prices.
8. Geometric Mean (GM): The geometric mean (approximately $163,269.84) provides an alternative measure of central tendency, considering the multiplicative nature of data.
9. Harmonic Mean (HM): The harmonic mean (approximately $113,018.63) is another measure of central tendency that is sensitive to outliers.
10. Q1 and Q3: The first quartile (Q1) is $73,900, and the third quartile (Q3) is $225,000, indicating the range within which the middle 50% of the data falls.

Above mentioned are all the possible conclusions that can be drawn from the tables of each column along with the stated summary of findings.

***ii.) CORRELATION:***

******



**1. Year and Horsepower:**

* The correlation coefficient between "Year" and "Horsepower" is very close to zero (approximately 0.0123).
* This suggests a **very weak correlation** between the model year of the cars and their horsepower.

**2. Year and Torque (lb-ft):**

* The correlation coefficient between "Year" and "Torque (lb-ft)" is also very close to zero (approximately 0.0344).
* This indicates a **very weak, nearly non-existent linear relationship** between the model year of the cars and their torque values.

**3. Year and Price (in USD):**

* The correlation coefficient between "Year" and "Price (in USD)" is negative (approximately -0.0126).
* This suggests a **very weak, negative correlation** between the model year of the cars and their prices.

**4. Horsepower and Torque (lb-ft):**

* The correlation coefficient between "Horsepower" and "Torque (lb-ft)" is relatively strong (approximately 0.9378).
* This indicates a **strong positive correlation** between a car's horsepower and torque.

**5. Horsepower and Price (in USD):**

* The correlation coefficient between "Horsepower" and "Price (in USD)" is positive but not extremely strong (approximately 0.7856).
* This suggests **a moderate positive linear relationship** between a car's horsepower and its price.

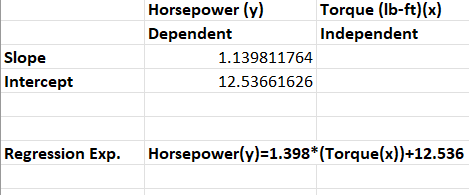
**6. Torque (lb-ft) and Price (in USD):**

* The correlation coefficient between "Torque (lb-ft)" and "Price (in USD)" is positive and moderately strong (approximately 0.7624).
* This indicates a **moderate positive linear relationship** between a car's torque and its price.

iii.) SLOPE AND INTERCEPT:

**Slope (Coefficient for Torque):** The slope of approximately 1.1398 represents the change in the dependent variable (Horsepower) for a one-unit change in the independent variable (Torque).

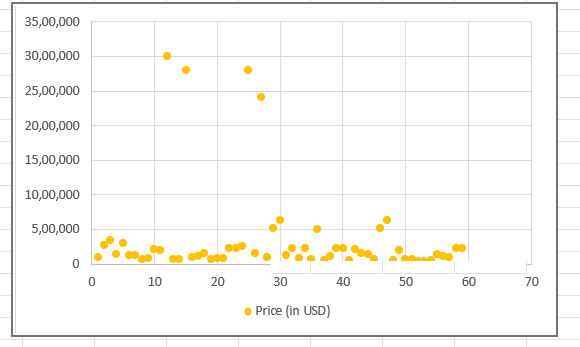
POSITIVE SLOPE INDICATES DIRECT RELATION BETWEEN HORSEPOWER AND TORQUE.



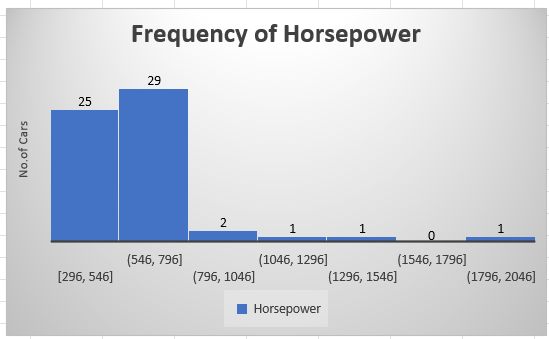
**Intercept:** The intercept of approximately 12.5366 represents the estimated value of the dependent variable (Horsepower) when the independent variable (Torque) is equal to zero.

THE VALUE OF DEPENDENT VARIABLE (x) IS DETERMINED FROM THE INDEPENDENT VARIABLE(y).

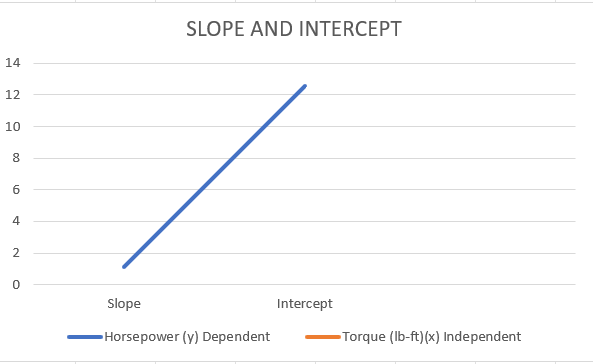
**GRAPHS:**

****

**1.)The price of almost all the cars lie around the range of 50K usd except 4 car models.This tells us that the average luxurious car costs around 50K usd.**

****

**2.) Most of the horsepower of the cars lie between the range of [546,796],Hence making it the modal range.**

****

**3.) Graph has a POSITIVE slope indicates DIRECT RELATIONSHIP between horsepower and Torque.**

**CONCLUSIONS:**

* 1. The dataset is NON NORMAL because it lacks symmetry, has have extreme values, or has a steeper “dome” other than a typical bell. It is leptokurtic(more than 7) and has extreme outliers.
  2. Skewness for horsepower, Torque and Price is positive(distribution of hp,torque and price is positively skewed) while the same for year is negative (the distribution of year is negatively skewed).
  3. The entire distribution of all the columns is Leptokurtic which implies that the kurtosis for all the four columns is greater than zero.
  4. As mentioned in the above (iii) point table there is a high positive correlation between horsepower and torque(0.9377…),Moderate positive or considerable correlation between horsepower and price(0.7855…), torque and price(0.7624…),poor or negligible correlation between year and horsepower(0.0123..),year and torque(0.0343…) but there is a negative correlation between year and price(-0.0125..)
  5. Slope is POSITIVE(1.139…) which shows a direct relation between torque and horsepower i.e. if torque increases(x) then horsepower(y) also increases .

**FOR GRAPHS:**

The dataset suggests that the majority of cars are priced near 50,000 USD, indicating an average luxury car price; most cars have horsepower in the range of 546 to 796, the modal range; and a positive-sloping graph signifies a direct relationship between horsepower and torque

References:

<https://www.kaggle.com/>

<https://www.scribbr.com/methodology/simple-random-sampling/#:~:text=Simple%20random%20sampling%20is%20a,equal%20chance%20of%20being%20selected>.

Annexure:

