Global Scientific JOURNALS

GSJ: Volume 12, Issue 3, March 2024, Online: ISSN 2320-9186
www.globalscientificjournal.com

ROLE OF STATISTICS IN DATA MINING OPERATIONS: ANALYZING THE PATTERNS IN A USED CAR DATASET

Michelle Rehman

KeyWords
Correlation Analysis, Data Mining, Data Visualization, Descriptive Statistics, Linear Regression Model, Market Trends, Data Analysis, Statistical Techniques.

ABSTRACT

This paper portrays the strong interplay between statistics and data mining by using a dataset that contains information regarding used cars. The connection between the two concepts showed extreme correlation when uncovering valuable insights regarding market trends, pricing dynamics, and customer preferences. This research outlines the crucial role of statistical methodologies in easing out the complexities of used car data, highlighting the importance and significance in directing decision-making processes within the automotive industry.

The data set includes various features about used cars such as their make, model, manufacturing year, price, and geographical location. Techniques like descriptive statistics, correlation analysis, and regression modeling were used to detect patterns and relationships, and also predict factors that could provide meaningful trends and patterns.

After adopting these methods, some meaningful insights were revealed. Descriptive statistics revealed key trends in car prices, mileage distributions, and popular car models. Correlation analysis revealed the relationships between a car's price with features such as age of car and mileage. Regression analysis was employed to predict the probability of a car's re-sale and this information was used to provide valuable insights for buyers, sellers, and industry stakeholders.

In conclusion, the results showed a deep connection between statistics and data mining when it came to extracting meaningful insights from a used car dataset. By employing statistical methodologies, stakeholders in the automotive industry can obtain a valuable comprehension regarding dynamics in pricing, customer behavior, and latest market trends. This promotes the idea of making more well-informed decisions and driving innovation in the automotive industry.

1. Introduction

The goal of this section is to highlight data mining as a key milestone that can be defined as an essential procedure which extracts meaningful information from very large datasets and then uproots particular trends and patterns, turning those findings into significant insights and predictions for various businesses. According to Parzen E. **[1]**, the name “statistical methods mining” was coined for the process of developing and marketing various maps of the world of statistical knowledge to apply to data mining the grand statistical knowledge that exists in literature.

1.1 Overview of Data Mining Operations

Data Mining operations can be classified by the type of function they were initially designed to perform. These include (but are not limited to) predictive modeling, segmentation, dependency modeling and probabilistic graphical modeling. There is a certain methodology that should be followed while using a data mining operation. It includes the following stages: 1) identify the question or goal, 2) collect data samples, 3) prepare and refine data, 4) activate data mining algorithm(s), and 5) analyze the algorithm’s results.

1.2 Importance of Statistics in Data Mining

It is quite evident that statistical techniques make up most of the crucial parts in a data mining process with a specific focus on patterns analysis. Descriptive statistics are used to gather insights into the distribution depth and variability of data. On the other hand, inferential statistics are employed to gather predictions for certain possibilities with estimated probabilities e.g., regressions are perhaps the single most widely used statistic in commercial predictive data mining applications according to **[2]**.

1.3 Focus of the Research: Analyzing Patterns in a Used Car Dataset

This research paper's main focus is on the symbiotic relationship between statistics and data mining operations, with a specific kingpin on analyzing patterns within a dataset of used cars. By using a comprehensive analysis with the help of statistical techniques and methodologies, the aim is to uncover valuable insights into the dynamics of the used car market, shedding light on factors influencing car prices, market trends, and consumer preferences.

2. Materials and Methods

For this section of the paper, a database called Cars4U provided on Kraggle **[3]** is used that displays all the necessary information regarding all the used cars across all the major cities of India. This dataset will be utilized to showcase the key interplay between statistics and some data mining operations.

2.1 Description of Used Car Dataset

Python programming language was used to assess this dataset that was initially in a CSV (Comma Separated Values) format. There was a comprehensive collection of attributes that defined each used car in extreme depth, these included (but are not limited to) the manufacture year, the kilometers driven, geographical location, transmission type, mileage, engine, initial price, new price, and some more. This dataset offers a rich source of information encapsulating various aspects of the used car market, providing a rich platform for exploring patterns and trends within the automotive industry.

2.2 Attributes Captured in the Dataset

This research is inspired by and is an extension to the performance comparison of data mining algorithms for car evaluation in **[4]**. After using Python's *shape()* function, it was seen that the dataset carried 7253 distinct observations with 13 variables taken as attributes. The data type and further information regarding the used car dataset was deeply studied using the *data.info()* function. The following output was displayed:

Table 1: Attributes for Each Observation with Data Type

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7253 entries, 0 to 7252
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Name               7253 non-null   object
 1   Location           7253 non-null   object
 2   Year               7253 non-null   int64
 3   Kilometers_Driven  7253 non-null   int64
 4   Fuel_Type          7253 non-null   object
 5   Transmission       7253 non-null   object
 6   Owner_Type         7253 non-null   object
 7   Mileage            7251 non-null   float64
 8   Engine             7207 non-null   float64
 9   Power              7078 non-null   float64
 10  Seats              7200 non-null   float64
 11  New_price          1006 non-null   float64
 12  Price              6019 non-null   float64
dtypes: float64(6), int64(2), object(5)
memory usage: 736.8+ KB
```

Table 1 shows that the variables like *Mileage, Engine, Power, Seats, New_Price*, and *Price* have missing values. Numeric variables like *Mileage* and *Power* are of datatype as float64 and int64 respectively. Categorical variables like Location, Fuel_Type, Transmission, and Owner_Type are of object data type. The following table summarizes the attributes captured in this dataset:

Table 2: Data Summary

Dataset Characteristics:	Multivariate	**Number of Instances:**	7253
Attribute Characteristics:	Categorical	**Number of Attributes:**	13
Associated Tasks:	Classification	**Missing Values?**	Yes

2.3 Statistical Techniques Applied

For a thorough analysis of this dataset, the following statistical techniques were employed: descriptive statistics, correlation analysis, and finally regression modeling. Not only do these methods help us gain insights into the data at hand, but also signify the analyst about any and every kind of anomaly present within the dataset.

2.3.1 Descriptive Statistics

Statistics can be utilized to demonstrate the characteristics of a group of observations; this is called descriptive statistics **[5]**. Descriptive measures like mean, median, standard deviation, and range offer insights into the distribution and variability of the data.

2.3.2 Correlation Analysis

A correlation analysis is conducted to explore the relationship between two variables within a given dataset. This study is fundamentally based on the assumption of a straight line (linear) relationship between the quantitative variables as stated in **[6]**. The strength and direction of the relationship between variables that define the price, mileage, and age of car have been taken into account.

2.3.3 Regression Analysis

According to Ij, statistics draws population inferences from a sample **[7]**. The regression model was employed to assess the relationships between dependent variables, in this case, the price of car, with independent variables such as the mileage and age of the used car. The linear regression model was mapped so as to estimate the effect of predictor variables on the target variable and predictions were made based on the model.

3. Results

The results that were derived from the techniques mentioned above provided valuable insights about the used cars dataset, uncovering patterns, relationships, and predictive factors. The analysis provides a comprehensive understanding of the factors influencing car prices and market trends in the used car industry.

3.1 Descriptive Statistics Findings

Python's built-in function for summary statistics (*describe()*) was used to gain a deeper understanding of the overall trends and patterns within the dataset. The following information was gathered:

Table 3: Summary Statistics Overview

	count	mean	std	min	max
Year	7253	2013.36	3.25	1996.00	2019.00
Kilometers_Driven	7253	58699.06	84427.72	171.00	6500000.00
Mileage	7251	18.14	4.56	0.00	33.54
Engine	7207	1616.57	595.28	72.00	5998.00
Power	7078	112.76	53.49	34.20	616.00
Seats	7200	5.28	0.80	2.00	10.00
New_price	1006	22.77	27.75	3.91	375.00
Price	6019	9.47	11.18	0.44	160.00

From the above table, the following information was gathered:

1. From the *Year* variable's row, we can observe through the min and max columns that the range is 1996 – 2019 which means that both latest and old models are popular choices when it comes to buying used cars.
2. The average for *Kilometers_Driven* is ≈58k KM. The range displayed through the min and max columns shows discrepancy as the max value is given to be 650k KM. This is proof of the presence of an outlier. This record was removed.
3. A data entry issue is foreseen as the *Mileage* variable shows that 0 cars would be sold with 0 mileage.
4. Further presence of outliers can be witnessed in *Engine* and *Power* variables as this makes the data appear to be rightskewed.
5. *Seats* variable is an important factor when considering the price of a car and according to the summarized statistical data, the average number of seats in a car appears to be 5.

A further analysis was done by developing a new variable *Car_Age* with the intention of making it a contributing factor to a car's price. Furthermore, the use of statistical summary to understand the most popular car in the user market was employed by

utilizing Python's *split()* function with the *Name* variable. This information was given to the Matplotlib and Seaborn libraries of Python to produce the following graphical facts:

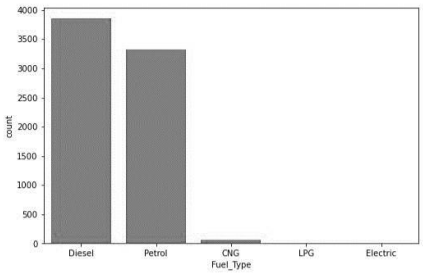

Figure 1. Count Graph for Popular Car Based on Fuel

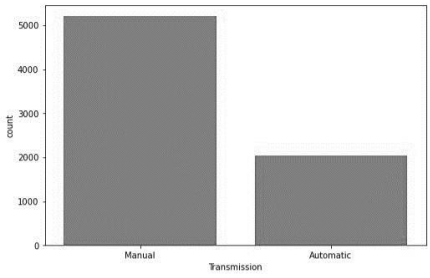

Figure 2. Count Graph for Popular Transmission Choice

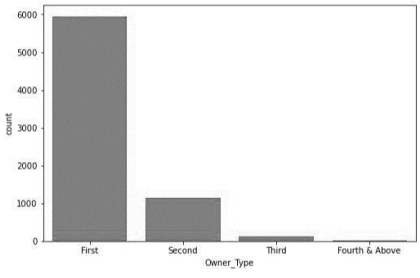

Figure 3. Count Graph for Preference of Overall Used and Unused Cars

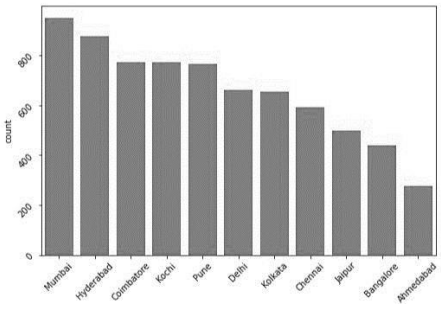

Figure 4. Count Graph for Cities with the Most Car Sales

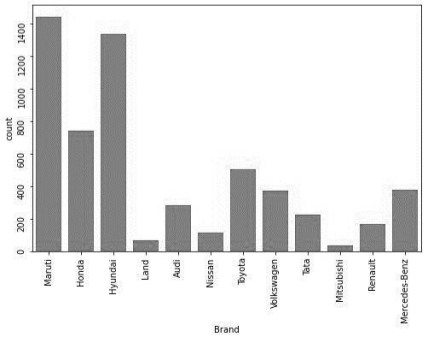

Figure 5. Count Graph for Popular Car Brand

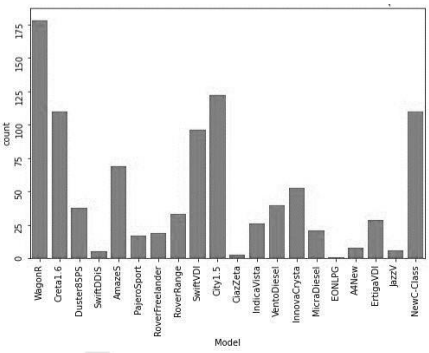

Figure 6. Count Graph for Popular Car Model

From the above graphical representations, the following information was inferred:

1. Almost 53% of the cars are of the *Fuel_Type* Diesel which means that most customers prefer diesel cars. Since diesel is cheaper than petrol, this could be one of the reasons for this fuel type to be popular.
2. Most customers prefer manual drive over automatic drive as about 73% of cars have manual transmission.
3. About 82% of the cars show ownership as the first-owned type. This can help in understanding that most buyers prefer unused cars.
4. Mumbai city has the highest number of cars available for purchase. Ahmedabad has the lowest.
5. Almost 20% of the cars belong to the brand Maruti, making it the most in-demand brand available.
6. The model 'WagonR' ranks first among all models which are available for purchase, making it the most popular choice.

3.2 Correlation Analysis Results

It is to be noted that the correlation coefficient only provides a measure of the linear association between the variables **[8]**. The analysis assesses the strength and direction of the relationship between variables. Through a small sample scatterplot representation, it was observed that there was a strong negative correlation between the age of the car with its price i.e. the older the car, the lower the price.

Furthermore, another correlation relation was observed but this time with a strong positive effect. This can be seen between the variables that explain the mileage and car's age. The following scatterplots show the negative and positive correlations between the variables mentioned above:

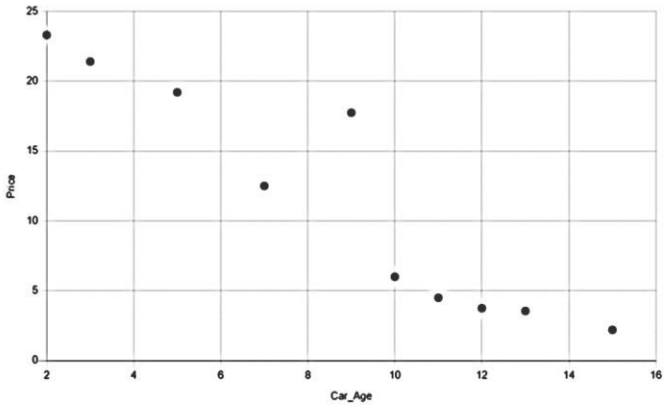

Figure 7. Negative Correlation Between Age of Car and Its Mileage

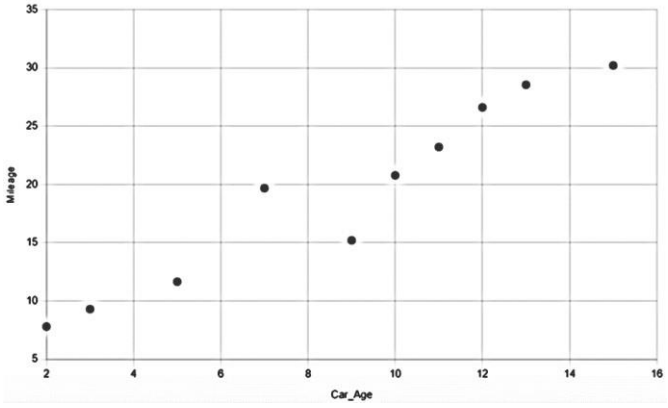

Figure 8. Positive Correlation Between Age of Car and Its Price

Through the above analysis, it was concluded that the lesser the price of a car, the more its age and the more its age, the more mileage the car possesses. By combining the above analysis it could hence be concluded that the more mileage on a car, the less its price will be.

3.3 Regression Modeling Insights

According to Aalen O. O., a linear model is suggested for the influence of covariates on the intensity of function **[9]**. For this reason, the relationship between the price variable in the existing dataset is regressed over the mileage of the cars. The regression analysis showed the line-of-best-fit which is produced after minimizing the sum of squared error terms (the squared difference between the data points and the line). The following scatterplot was obtained after inputting a small sample data to gather insights and make meaningful predictions:

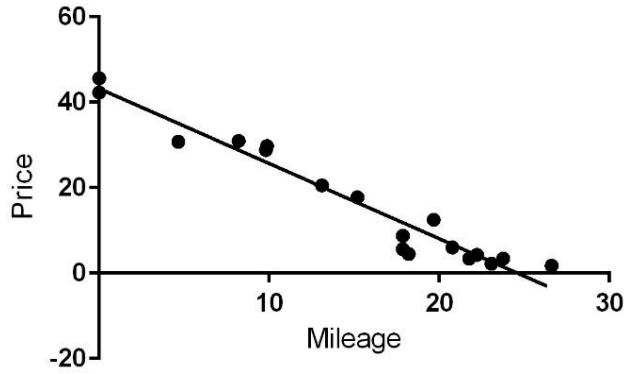

Figure 9. Price Negatively Regressed on Mileage

The regression equation, which is given as $\hat{y}$

$$= \beta 0 + \beta 1 x + \varepsilon$$

was calculated to be

$$\text{Price} = 43.22 + (-1.756) * \text{Mileage} + \varepsilon$$

Hence from the declining line-of-best-fit and the negative value of the coefficient for the independent variable Mileage, it can be witnessed that mileage of a car has a significant negative effect of 1.756 on unit car price, with newer cars being more popular despite having higher prices. We can assume that cars with less mileage have higher chances of being re-sold.

4. Discussion

This analysis of a used car dataset provides valuable insights into market dynamics, including trends in car prices and relationships between variables that define the mileage of a car with the price of a car. Kuonen D. summarized the concept by implying how important it is to note that data mining can learn from statistics – that, to a large extent, statistics is fundamental to what data mining is really trying to achieve **[10]**.

4.1 Interpretation of Findings

The interpretation of findings from this study of used car dataset unveils many valuable patterns. Descriptive statistics revealed important trends and distributions within the data, such as the average price of used cars and the typical mileage of vehicles. This method was also deployed to filter out the popular brands, models and car types in the consumer market. Correlation analysis was employed to understand the relationship between certain variables and figure out their dependency type, highlighting factors that significantly influence car prices. Lastly, regression modeling was used to offer predictive capabilities, enabling the estimation of car prices and identify key drivers that contribute to market dynamics.

4.2 Implications for Buyers, Sellers, and Industry Stakeholders

The implication of this study extends to various stakeholders within the automotive industry. For buyers, this analysis provides valuable guidance for making informed purchasing decisions. By understanding the factors that influence car prices, buyers can negotiate better deals and identify value propositions in the market. For sellers, this analysis offers insights into pricing strategies and market positioning. By leveraging information about market trends and consumer preferences, sellers can optimize pricing strategies to maximize profits and attract potential buyers. Moreover, industry stakeholders, including dealerships, manufacturers, and policymakers, can benefit from this study by gaining a deeper understanding of market dynamics and consumer behavior. These findings provide valuable intelligence for shaping business strategies, allocating resources effectively, and driving innovation within the automotive industry.

5. Conclusion

The purpose of this research was to provide proof of the interchangeable dependencies between statistics and data mining. Throughout the process of analyzing the used cars dataset, it was made evident that statistics play a significant role in data mining operations, and further extends the operations to detect trends and patterns while making meaningful insights. Thus, this concludes the roles of statistics in data mining operation using a used car dataset.

6. References

[1] E. Parzen, "Data mining, statistical methods mining, and history of statistics," Computing Science and Statistics, pp 365-374, 1998.

[2] P. Smyth, "Data mining at the interface of computer science and statistics," Data Mining for Scientific and Engineering Applications, pp 3561, Jan. 2001.

[3] S. Bedi, "Cars4U dataset," Kaggle, Oct. 2020. https://www.kaggle.com/datasets/sukhmanibedi/cars4u/data

[4] J. Awwalu, A. Ghazvini, and A. A. Bakar, "Performance comparison of data mining algorithms: a case study on car evaluation dataset," International Journal of Computer Trends and Technology, vol. 13, no. 2, pp 78-82, 2014.

[5] G. Marshall and L. Jonker, "An introduction to descriptive statistics: a review and practical guide," Radiography, vol 16, no. 4, pp 1-7, 2010.

[6] N. J. Gogtay and U. M. Thatte, "Principles of correlation analysis," Journal of the Association of Physicians of India, vol. 65, no. 3, pp 78-81, 2017.

[7] H. Ij, "Statistics versus machine learning," Nat Methods, vol 15, no. 4, 233, 2018.

[8] D. Weenink, "Canonical correlation analysis," In Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, vol. 25, pp 81-99, Aug. 2003.

[9] O. O. Aalen, "A linear regression model for the analysis of life times," Statistics in Medicine, vol. 8, no. 8, pp 907-925, 1989.

[10] D. Kuonen, "Data mining and statistics: what is the connection?," The Data Administration Newsletter, vol. 30, pp 1-6, 2004.