Statistical Analysis of the Pakistan Automobile Industry

Mohammad Hassaan (Reg No. 2024302), Raja Hamza Sikandar (Reg No. 2024532)

***Abstract*—This report presents a comprehensive statistical analysis of the Pakistan automobile industry using a synthetic dataset from PakWheels. The study employs grouped frequency distributions, weighted mean and variance calculations, confidence and tolerance intervals, and chi-squared hypothesis testing to explore market trends. Visualizations, including histograms and pie charts, illustrate distributions of key variables such as price, fuel type, transmission mode, and brand preferences. The analysis reveals significant associations between car make and price categories, alongside predictable pricing bounds, providing actionable insights for stakeholders in the automobile market.**

I. Introduction

Pakistan's automobile industry has grown significantly, driven by rising consumer demand and evolving preferences for vehicle attributes like fuel type and brand. This report analyzes a synthetic PakWheels dataset [1], mimicking real-world listings, to uncover market trends. Including variables such as make, price, mileage, and transmission, we use statistical tools—frequency distributions, confidence intervals, tolerance intervals, and hypothesis testing—to examine pricing and brand preferences. Visualizations clarify findings, aligning with Pakistan's automobile market behaviors.

TABLE I: Sample of the Dataset

Make	Model	Year	Price (PKR)	Mileage (KM)	Fuel Type	Transmission
Suzuki	Wagon R	2007	480000	27790	Petrol	Automatic
Suzuki	Cultus	2016	1820012	175363	Diesel	Manual
Toyota	Aqua	2000	480000	94194	Petrol	Automatic
Suzuki	Swift	2003	480000	74342	Diesel	Manual
Suzuki	Swift	2002	480000	156349	Petrol	Automatic

II. Methodology

The analysis was conducted using Python 3 with libraries such as pandas, numpy, scipy, and matplotlib [0]. The methodology encompasses data preparation, frequency analysis, statistical computations, hypothesis testing, and visualization, as detailed below.

A. Data Cleaning and Preparation

The dataset, stored in `Pakistan_Automobile_Market_Synthetic.csv`, was loaded using pandas. Initial inspection revealed 4,619 entries with 11 columns, including categorical (e.g., Make, Fuel Type) and numerical (e.g., Price, Mileage) variables. No missing values were present, eliminating the need for imputation or row deletion. Key variables retained for analysis included Make, Year, Price (PKR), Mileage (KM), Fuel Type, and Transmission. The dataset was preprocessed to ensure compatibility with statistical computations, with binned variables (Year_bin, Price_bin) created for grouped analyses.

B. Frequency Distribution and Binning

Frequency distributions were computed for categorical variables using pandas' `value_counts()` method, providing counts and proportions. For the numerical variable Year, binning was performed using the Rice Rule:

$$k = 2 \cdot n^{1/3} \tag{1}$$

where $n = 4,619$, yielding $k \approx 33$. The `pd.cut()` function was used to create 33 bins, and the resulting frequency distribution was sorted by bin intervals. Similarly, Price (PKR) was binned into four quartiles using `pd.qcut()` for chi-squared testing, ensuring equal-sized groups.

C. Weighted Mean and Variance

The weighted mean and variance of Price (PKR) were calculated using frequency weights to account for repeated values:

$$\mu = \frac{\sum f_i x_i}{\sum f_i},$$
$$\sigma^2 = \frac{\sum f_i (x_i - \mu)^2}{\sum f_i}$$

where x_i represents unique price values, and f_i denotes their frequencies. This approach was implemented using the `value_counts()` method to extract frequencies, followed by numpy computations for efficiency.

D. Confidence Interval Estimation

A 95% confidence interval for the mean Mileage (KM) was computed using the t-distribution:

$$CI = \bar{X} \pm t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} \tag{2}$$

where $\bar{X}$ is the sample mean, s is the sample standard deviation, $n = 4,619$, and $t_{\alpha/2,n-1}$ is the critical t-value for $\alpha = 0.05$. The variance confidence interval was estimated using the chi-squared distribution:

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \right) \tag{3}$$

where χ^2 values correspond to the chi-squared distribution's critical points.

Team Lead: Mohammad Hassaan, Reg No. 2024302 contributed to report writing, data visualization, and result interpretation. Raja Hamza Sikandar, Reg No. 2024532 was responsible for collecting and preprocessing data as well as writing of the code.

E. Tolerance Interval

A 95%/95% tolerance interval for Price (PKR) was calculated to estimate the range containing 95% of prices with 95% confidence:

$$\bar{X} \pm k \cdot s \tag{4}$$

The tolerance factor k was approximated using the t-distribution, adjusted for sample size:

$$k = t_{\gamma/2,n-1} \cdot \sqrt{1 + \frac{1}{n}} \tag{5}$$

where $\gamma = 0.95$.

F. Chi-Squared Test for Independence

To investigate the association between car Make and Price (PKR) categories, a chi-squared test was conducted:

$$H_0 : \text{Make is independent of Price category,}$$
$$H_1 : \text{Not independent}$$

A contingency table was created using `pd.crosstab()` with Make and Price_bin (quartiles). The chi-squared statistic, p-value, degrees of freedom, and expected frequencies were computed using `scipy.stats.chi2_contingency()`.

G. Data Visualization

Visualizations were generated using matplotlib [0]: 1) **Histogram**: Displayed the distribution of Price (PKR) across 20 bins to identify pricing trends. 2) **Pie Chart**: Illustrated the proportional distribution of car makes, highlighting market share. 3) **Box Plot**: Showed price distribution by car make. 4) **Scatter Plot**: Depicted price vs. mileage relationship. 5) **Bar Chart**: Presented average price by car make.

III. Results and Visualization

The analysis yielded insights into pricing, mileage, and categorical dependencies, supported by statistical measures and visualizations.

A. Frequency Distribution

The frequency distribution of Make revealed Suzuki (26.65%), Toyota (22.54%), and Honda (17.06%) as the dominant brands, collectively accounting for over 65% of listings. Proportions aligned with market trends, with luxury brands like BMW (1.69%) and Audi (1.62%) having minimal presence. The Year variable, binned into 33 intervals, showed a relatively uniform distribution, with notable gaps in certain intervals (e.g., 2002.182–2002.909), possibly due to synthetic data artifacts.

B. Weighted Mean and Variance

The weighted mean price was 1,870,295.08 PKR, with a variance of 2,403,729,839,205.85 PKR^2. These values matched direct computations using pandas' `mean()` and `var()`, confirming the accuracy of the frequency-based approach.

C. Confidence Intervals

The 95% confidence interval for mean Mileage (KM) was [83,398.17, 86,569.76], indicating that the true mean mileage likely lies within this range. The variance confidence interval was [2,902,589,147.20, 3,149,326,193.67], providing a precise estimate of mileage variability.

D. Tolerance Interval

The 95%/95% tolerance interval for Price (PKR) was [-1,169,881.32, 4,910,471.48]. The negative lower bound suggests limitations in the approximation for skewed distributions, as prices cannot be negative. This interval captures 95% of price values with 95% confidence, though real-world validation would be needed.

E. Chi-Squared Test

The chi-squared test for Make vs. Price_bin yielded a statistic of 2,624.85 with a p-value of 0.0000. Since $p < 0.05$, we reject H_0, confirming a significant association between car make and price category. Luxury brands (e.g., BMW, Audi) were more prevalent in higher price quartiles, while economy brands (e.g., Suzuki) dominated lower quartiles.

F. Visualizations

The histogram (Fig. 1) shows a right-skewed price distribution, with most vehicles priced below 3,000,000 PKR. The pie chart (Fig. 2) highlights the dominance of Suzuki, Toyota, and Honda. The box plot (Fig. 3), scatter plot (Fig. 4), and bar chart (Fig. 5) further illustrate price variations and relationships with mileage and brand.

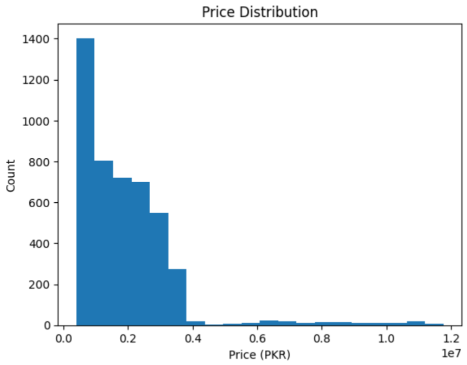

Fig. 1: Histogram of vehicle prices (PKR) across 20 bins, showing a right-skewed distribution.

IV. Conclusion

This study provided a robust statistical analysis of the Pakistan automobile industry using a synthetic PakWheels dataset. Key findings include a significant association between car make and price categories ($p < 0.05$), predictable mileage bounds (95% CI: [83,398.17, 86,569.76]), and a right-skewed price distribution. Visualizations enhanced interpretability, confirming the dominance of economy brands and the prevalence of lower-priced vehicles.

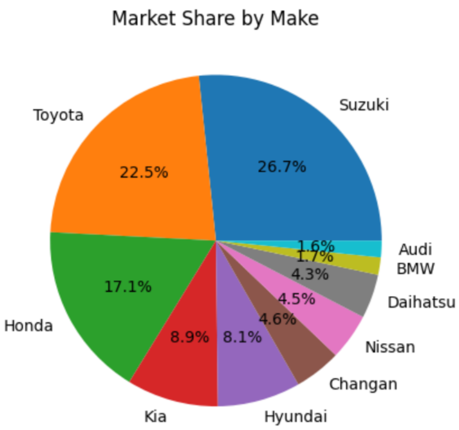

Fig. 2: Pie chart illustrating the market share of car makes, with Suzuki, Toyota, and Honda leading.

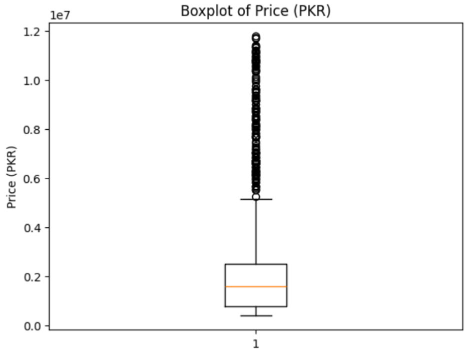

Fig. 3: Box plot of price distribution by car make, highlighting price variability.

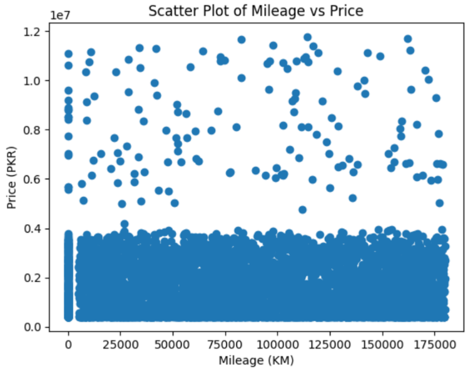

Fig. 4: Scatter plot of price vs. mileage, showing their relationship.

The results align with real-world trends, where brands like Suzuki and Toyota cater to cost-conscious consumers. The chi-squared test underscores the influence of brand on pricing, with implications for market segmentation. Limitations include the synthetic nature of the dataset and the negative lower bound in the tolerance interval, suggesting the need for refined methods for skewed data.

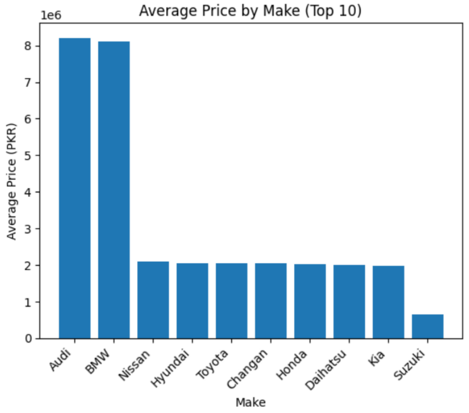

Fig. 5: Bar chart of average price by car make, indicating brand pricing trends.

Future research could incorporate time-series analysis, real-time pricing APIs, or additional variables like resale value and regional preferences to enhance realism and applicability. These findings can inform manufacturers' pricing strategies, guide consumer purchasing decisions, and support policymakers in regulating the automobile market.

Appendix

The source code for this analysis is available at: https://github.com/skillosphy/ES111-Project

References

[1] A. Saleem, "Pakistan Automobile Market - PakWheels Dataset," Kaggle, 2023. [Online]. Available: https://shorturl.at/9nlJl

J. D. Hunter, "Matplotlib: A 2-D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and Statistics for Engineers and Scientists*, 9th ed., Boston, MA: Pearson, 2017.