**LUIS ALBERTO MERCADO RODRIGUEZ – PST107 - A00147277**

**Statistical Analysis and Insights from the Dataset**

It is essential to comprehend a dataset in order to derive significant findings and make well-informed choices. The purpose of the study is to determine and measure the variables affecting used automobile pricing in Australia. The effects of characteristics like brand, gearbox type, and body design are still unclear, despite the fact that conventional criteria like age and miles are widely documented. Does the brand, fuel type, gearbox, or body design affect the price? Using statistical techniques and visualizations, this paper offers a data-driven approach to comprehending these price factors in order to identify patterns, connections, and insights. Parameter estimation, hypothesis testing, and univariate, bivariate, and distribution analyses are all included in the analysis.

**Description of the dataset**  
The dataset used for this study includes 17,952 rows and 16 columns, containing information on used cars in Australia. Key variables include:

* **ID:** Unique identifier for the car
* **Name:** Full name of the car
* **Price:** Cost of the car
* **Brand:** Manufacturer of the car
* **Model:** Year of the car model
* **Variant:** Specific version or trim of the car
* **Series:** Car's series or generation
* **Year:** Year the car was manufactured
* **Gearbox:** Transmission type (e.g., automatic, manual)
* **Type:** Body style (e.g., sedan, SUV, hatchback)
* **Fuel:** Type of fuel used (e.g., petrol, diesel, electric)
* **Status:** Current condition (e.g., new, used)
* **Kilometers:** Total distance traveled by the car
* **CC:** Engine displacement in cubic centimeters
* **Color:** Exterior color of the car
* **Seating Capacity:** Number of seats available

This dataset was sourced from Kaggle (Australia Car Market Data, 2024)

**1. Univariate Analysis and Mean, Median and Standard Deviation**

The California State University, consider that “Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own” (Univariate Statistics, n.d.). Univariate analysis focuses on summarizing and understanding each variable independently. For continuous variables like Price, Kilometers, and Year, we used measures of central tendency (mean, median) and dispersion (standard deviation) to describe the data's spread and center.

**Analysis of Price**

The Price column represents the target variable.

* **Mean Price**: 36,761
* **Median Price**: 35,000
* **Standard Deviation**: 30,334

The median is less than the mean, indicating a left-skewed distribution. This skewness suggests that while most cars are mid-range in price, a small number of higher-priced models skew the average upward, indicating potential for targeted high-end marketing strategies. Additionally, because of a large standard deviation, we can deduce that there is a combination of both high-end and budget-friendly vehicles.A graph of a distribution

Description automatically generated

**Analysis of Kilometers**

* Mean: 103,442 km
* Std Dev: 80,386 km

Due to high variability, we can infer that used and nearly new cars are included in the dataset. So we can infer that the dataset contains a wide variety of cars in terms of year of manufacturing.

A graph of a number of miles

Description automatically generated

**Analysis of Year**

* **Year:**
  + Mean: 2015
  + Median: 2016, is consistent with the mean.

A graph of a number of years

Description automatically generated

This variable suggests that the majority of vehicles are less than 10 years old, indicating that newer vehicles are trending in the market.

**Categorical Features**

For categorical variables, which means, fixed variables with a limited number of values like Brand, Fuel, and Gearbox, bar plots show the frequency distribution of each category. For example:

A graph of a number of brand

Description automatically generated

* Brand has several entries, with the top brands being Toyota, Holden, Ford, Mazda, and Hyundai.

A graph of fuel type distribution

Description automatically generated

* Fuel shows a higher frequency for unleaded petrol over diesel but their values are very similar.

A graph with different colored bars

Description automatically generated

* Gearbox demonstrates that automatic vehicles dominate the Australian market, indicating a potential focus for dealerships to increase inventory and marketing efforts on high-demand automatic models, which could boost sales in this segment. And low existence of special gearbox types like AWD, Front, and Rear.

A graph showing body type popularity

Description automatically generated

* Body type demonstrates a big dominance of Wagon body type in the market but also the medium-low availability of hatchback and sedan cars.

These insights lay the foundation for comparing these features with the target variable.

**2. Bivariate Analysis**

The bivariate analysis examines the relationships between pairs of variables, as defined by Kenneth C Wright “Bivariate analysis. The analysis of two variables simultaneously for the purpose of determining the empirical relationship between them (eg, computation of a simple correlation coefficient)”(Bivariate Analysis, 2002) particularly between the predictors and the target variable Price.

**Price vs. Categorical Variables**

Our goal is to assess the influence of categorical variables like Brand, Fuel type, Gearbox type, and Body design on a customer's car purchasing decision. Box plots illustrate the variation in Price across different categories.A graph with blue dots and white text

Description automatically generated

* Brand: Luxury brands like BMW and Mercedes have significantly higher prices compared to economy brands like Kia and Chevrolet.

A graph of a graph showing the price of fuel

Description automatically generated

* Fuel: Diesel cars show higher prices on average, likely due to their efficiency and higher resale value. Having outliers in the premium unleaded petrol may be related to the existence of luxury cars in the dataset.

A graph with blue dots

Description automatically generated

* Gearbox: Automatic vehicles have a higher median price compared to manual vehicles.

A graph of different colored objects

Description automatically generated with medium confidence

* Body Design: Although wagon cars are the most common in the market, coupe body styles are typically the priciest, especially when considering luxury car exceptions.

These connections assist us in comprehending how various characteristics affect car prices.

**3. Distribution Analysis**

Distribution analysis helps describe how data points are spread across the range of a variable. For the office of the impact analysis of the Australian Government, “Distributional analysis furthers the informative value of a Cost-Benefit Analysis (CBA) by articulating how costs and benefits are distributed across different cohorts of a population. ”(Distributionaal Analysis, 2023) This is critical for identifying skewness, kurtosis, and potential outliers.

* **Price Distribution**: A histogram and kernel density estimate (KDE) show a right-skewed distribution. The majority of cars are priced between 10,000 and 60,000, indicating a broad middle-market focus. The long tail of higher prices highlights a niche market for luxury cars, presenting an opportunity for premium dealerships to target high-income customers.
* **Kilometers Distribution**: Most cars have mileage under 100,000, but a few outliers extend beyond 200,000.
* **CC Distribution**: Engine sizes are concentrated around standard values like 1,200cc and 2,000cc.

**4. Parameter Estimation and Inference**

Using regression models RandomForestRegressor and applying the method of feature impotance which is defined as “a method to help us interpret a predictive model. The values indicate the relative contribution of the dataset features to the predictions generated by the model.”(What are feature importance values?, n.d.), we estimated the importance of features in predicting Price. The model determined the following graph:

A graph with blue bars

Description automatically generated

**Feature Importance**:

* **Brand:** With a contribution of 8% to price variation, brand stands out as a key determinant of vehicle pricing. This strong influence highlights the potential for dealerships to enhance their sales strategies by emphasizing well-known, high-demand brands in marketing campaigns, particularly for premium or luxury models where brand perception plays a crucial role.
* **Fuel Type:** Contributing 7% to price variation, fuel type is another critical factor, reflecting consumer preferences for fuel efficiency and long-term cost savings. Dealerships could leverage this insight by promoting fuel-efficient models or offering targeted incentives for vehicles with alternative fuel types, such as hybrids or electric cars.
* **Body Type:** Accounting for approximately 3% of the price variation, body type influences pricing to a lesser extent but still presents opportunities for segmentation. For instance, highlighting the premium nature of coupes or the practicality of SUVs in marketing could attract specific customer groups looking for style or utility.
* **Gearbox:** With a minimal contribution of 0.6%, gearbox type has a relatively small impact on price. However, emphasizing the convenience and widespread preference for automatic transmissions in marketing materials could still resonate with a broader audience, particularly in urban markets.

These insights assist us in developing effective predictive models.

**5. Correlation Analysis**

Correlation analysis quantifies the strength and direction of relationships between variables, “Correlation Analysis is statistical method that is used to discover if there is a relationship between two variables/datasets, and how strong that relationship may be.” According to Emily James in FlexMR (What is Correlation Analysis? A Definition and Explanation, 2022):

A screenshot of a chart

Description automatically generated

**Correlation Between Numerical Variables**

Using a heatmap to visualize correlation:

* Kilometers (r=−0.4) shows a moderate negative correlation with Price, indicating that cars with higher mileage tend to cost less.
* CC (r=0.3) has a positive correlation, suggesting that cars with larger engines are generally more expensive.
* Positive correlation: Variables increase together such as CC and Price.
* Negative correlation: One variable increases while the other decreases like Kilometers and Price.

The correlation matrix revealed that CC and Year are strong predictors, confirming their importance in the regression models.

**6. Hypothesis Testing**

A black screen with white text

Description automatically generatedTo validate relationships, we performed ANOVA Tests for hypothesis testing, “ANOVA, which stands for Analysis of Variance, is a statistical test used to analyze the difference between the means of more than two groups.”:

* **Brand:** Luxury brands have a significant impact on pricing.
* **Type:** The type of car, such as a wagon or hatchback, influences the price.
* **Status:** New cars available in stock command much higher prices than used or demo models.
* **Gearbox:** There are notable price differences associated with transmission types.

**7. Visualization**

Visualizations play a crucial role in making data more accessible and engaging:

* **Histograms and KDEs**: Highlight the distribution of continuous variables.
* **Box Plots**: Compare medians across categories.
* **Scatter Plots**: Display relationships between numerical predictors and Price.
* **Heatmaps**: Show correlations among variables.

These visualizations help communicate the data story effectively.

**8. Model Performance**

Using various regression models, the dataset's predictive capabilities were evaluated:

A screenshot of a graph

Description automatically generated

* **Top Performers**: ExtraTreesRegressor and CatBoostRegressor achieved the highest R² score on the test set (0.847) and recorded the lowest mean squared error.
  + - * **Low-Performers:** SGDRegressor and SVR resulted in negative R² values, signaling inadequate fit and ineffectiveness in modeling this dataset.

**Conclusion**

After considering more traditional factors such as age and mileage, brand reputation (which varies by 8%) and fuel type (which varies by 7%) emerge as the leading indicators of pricing, based on an analysis of 17,952 used cars in Australia. Although luxury brands tend to be pricier, Toyota and Holden command a significant share of the market, with 85% of listings featuring automatic transmissions. Machine learning algorithms provide valuable insights to market participants by revealing that transmission type (0.6%) and body style (3%) influence pricing less than was previously understood.

**References**

*univariate statistics*. (n.d.). <https://home.csulb.edu/~msaintg/ppa696/696uni.htm#:~:text=Univariate%20analysis%20explores%20each%20variable,each%20variable%20on%20its%20own>.

Wright, K. C. (2002). The language of research. *Academic Radiology*, *9*(10), 1194–1200. <https://doi.org/10.1016/s1076-6332(03)80521-4>

*Distributional analysis | The Office of Impact Analysis*. (2023, August 3). <https://oia.pmc.gov.au/resources/guidance-assessing-impacts/distributional-analysis>

*What are feature importance values?* (n.d.). <https://www.pi.exchange/knowledgehub/what-are-feature-importance-values>

James, E. (2022, April 8). What is Correlation Analysis? A Definition and Explanation. *FlexMR*. <https://blog.flexmr.net/correlation-analysis-definition-exploration>

Bevans, R. (2024, May 9). *One-way ANOVA | When and how to use it (With Examples)*. Scribbr. <https://www.scribbr.com/statistics/one-way-anova/>