**Project Report: Finding the best Car Value**

**ISDS 540 Group 7**

**Group members**

• Pranavi Chinthireddy, Interpreter, pranavi@csu.fullerton.edu

• Nandini Koppunuru,Orator , koppunuru27@csu.fullerton.edu

• Sai Srujana Ravipati, Orator, ravipati09@csu.fullerton.edu

• Ajay Rajesh, Deliverer, ajayrajeshb@csu.fullerton.edu

• Ponnappa Chenanda Machaiah, Creator, pavan.ponnappa@csu.fullerton.edu

**1. Introduction:**

In today's market, consumers have numerous options when it comes to purchasing a new vehicle. With so many make and models available, it can be challenging to determine which car offers the best value. In this report, we will analyze the factors that contribute to a car's value using the data set provided. Our goal is to identify the key predictors of a car's value score, which is a comprehensive measure of a car's worth based on five-year owner costs, road-test scores, and predicted reliability ratings. We will explore the relationship between these predictors and the value score and develop a linear regression model to predict the value score of a car based on these factors.

The dataset contains information on 54 cars from three different sizes - small sedans, family sedans, and upscale sedans. The response variable is the value score, while the predictors include car price, owner costs, road-test score, predicted reliability rating, and size category. We will begin by exploring the summary statistics of the dataset and visualizing the relationships between the predictors and the response variable. Next, we will conduct multiple linear regressions to determine the significance of each predictor and their contribution to the value score. Finally, we will interpret the results and discuss the implications for consumers looking to purchase a new car.

Below are the variables and their description:

1. Value Score: This is the response variable that measures the overall value of a car. It is calculated by combining five-year owner costs, road-test scores, and predicted reliability ratings. A higher value score indicates better value.
2. Car Price: This is the price of the car in dollars. It is a continuous variable that represents the amount a consumer pays for the car upfront.
3. Owner Costs: This variable represents the average cost per mile driven in the first five years of ownership. It includes depreciation, fuel, maintenance, and repairs. A lower owner cost indicates better value.
4. Road-Test Score: This variable represents the score of the car based on more than 50 tests and evaluations conducted by Consumer Reports. It ranges from 0 to 100, with a higher score indicating better performance, comfort, convenience, and fuel economy.
5. Predicted Reliability Rating: This variable represents the predicted reliability of the car based on data from Consumer Reports' Annual Auto Survey. It is a categorical variable that takes values from 1 (Poor) to 5 (Excellent). A higher rating indicates a more reliable car.
6. Size Category: This variable categorizes the car into one of three categories based on its size: small sedan, family sedan, and upscale sedan. It is a categorical variable with three levels.
   1. Small-Sedan: This is a dummy variable that indicates whether the car is a small sedan or not. It takes the value of 1 if the car is a small sedan and 0 otherwise.
   2. Family-Sedan: This is a dummy variable that indicates whether the car is a family sedan or not. It takes the value of 1 if the car is a family sedan and 0 otherwise.
   3. Upscale-Sedan: This is a dummy variable that indicates whether the car is an upscale sedan or not. It takes the value of 1 if the car is an upscale sedan and 0 otherwise.

These variables will be used in the analysis to identify the factors that contribute to a car's value score and to develop a linear regression model that can predict the value score based on these factors.

**2. Descriptive Statistics and Visualizations**

In this section, we will present the descriptive statistics and visualizations of the variables included in our analysis. This will help us understand the distribution of the data and identify any patterns or trends that may exist.

**2.1 Descriptive Statistics**

Table 1 below presents the descriptive statistics for the variables included in our analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | Mean | Standard Deviation | Minimum | Maximum |
| Value Score | 1.354 | 0.26 | 0.82 | 1.99 |
| Car Price | 28,340 | 6929.427 | 16,419 | 39,850 |
| Owner Costs | 0.65 | 0.10 | 0.44 | 0.83 |
| Road-Test Score | 78.07 | 9.03 | 52 | 95 |
| Predicted Reliability Rating | 3.40 | 1.00 | 1 | 5 |

From the table, we can observe that the mean value score is 1.354, indicating that the average car in our sample has a moderate value score. The standard deviation of the value score is 0.26, suggesting that there is moderate variation in the value scores across the different cars.

The mean car price is $28,340, with a standard deviation of $6,929. This wide standard deviation suggests a substantial range of prices for cars in our sample, spanning from $16,419 to $39,850.

The mean owner costs are 0.65, with a standard deviation of 0.10. This implies that the majority of cars in our sample have relatively low to moderate owner costs, with limited variability.

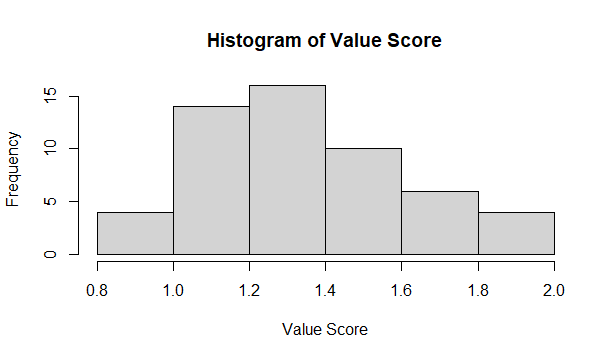
The mean road-test score is 78.07, with a standard deviation of 9.03. This suggests that most cars in our sample performed well in the road test, though there is notable variability in the scores ranging from 52 to 95.

The mean predicted reliability rating is 3.40, with a standard deviation of 1.00. This implies that the majority of cars in our sample have moderate to high predicted reliability ratings, though there's a wide range of ratings observed from 1 to 5.

**2.2 Visualizations**

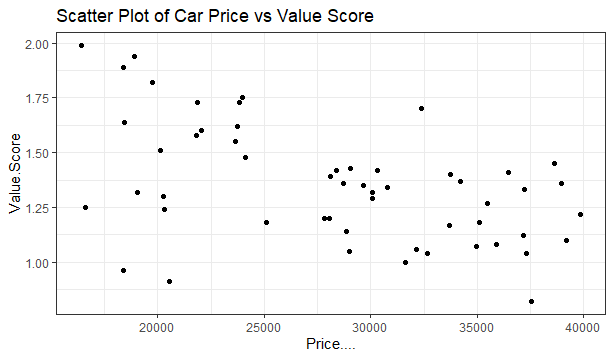
We will now use visualizations to explore the distributions of the variables and identify any patterns or trends.

Figure 1 below shows the histogram of the value score.

****

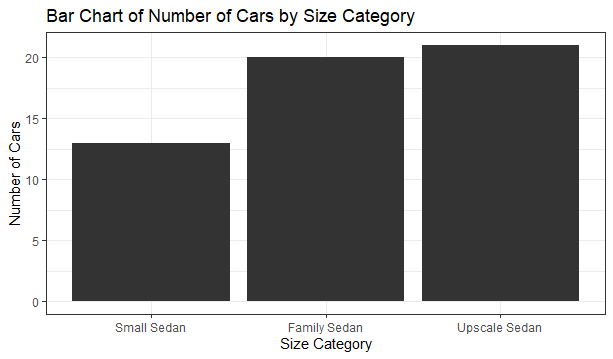
From Figure 1, we can see that the distribution of the value score is skewed to the right, indicating that the majority of cars in our sample have a moderate value score. There are also a few outliers with low value scores.

Figure 2 below shows the scatter plot of car price versus value score.

****

From Figure 2, we see that the plot is evenly scattered, indication strong negative correlation between car price and value score.

Figure 3 below shows the bar chart of the number of cars by size category.

****

From Figure 3, we can see that the majority of cars in our sample are upscale (38%) and family sedans (37%). Small sedans make up only 24% of the sample.

Overall, the descriptive statistics and visualizations suggest that there is a wide range of cars in our sample, with varying values, prices, owner costs, road-test scores, and predicted reliability ratings. The next step is to conduct inferential statistical analyses to identify the factors that contribute to a car's value score.

**2.3 Statistical Analysis**

Let's analyze correlations between variables, distributions of numeric values, and understand how each variable relates to the value score.



Based on the correlation analysis output, let's select variables that show moderate to strong correlations with the "Value.Score" and avoid variables that exhibit high multicollinearity among themselves. From the correlation matrix:

**Strong and Moderate Correlations with "Value.Score":**

* + Price (-0.49): Moderate Negative correlation
  + Cost.Mile (-0.5): Strong Negative correlation
  + Road.Test.Score (0.19): Moderate Positive correlation
  + Predicted.Reliability (0.64): Strong Positive correlation
  + Upscale\_Sedan (-0.49): Moderate Negative correlation

Considering these correlations, we choose these predictors to build our model to predict the "Value.Score"

These observations suggest that value score is influenced by a combination of factors, including size, cost/mile, price, road test score, and predicted reliability.

**3. Linear Regression Analysis**

Now that we have explored the factors that affect a car's value score, let’s employ linear regression to come up with a model that can predict the value score, given the selected independent variables.

**3.1 Model Formula:**Value Score = β0 + β1\*Car Price + β2\*Cost/Mile + β3\*Road-Test Score + β4\*Predicted-Reliability Rating + β5\*Upscale\_Sedan + ε

where β0, β1, β2, β3, β4 and β5 are coefficients; ε is error term.

**3.1.2 Training the model using Training data**

Using the training dataset, we fit the linear regression model and obtain the coefficient estimates:

Residuals:

Min 1Q Median 3Q Max

-0.11847 -0.04933 0.00452 0.03970 0.14738

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.128e+00 1.319e-01 8.553 2.71e-10 \*\*\*

Price -9.269e-06 5.510e-06 -1.682 0.101

Cost.Mile -1.624e+00 3.272e-01 -4.962 1.58e-05 \*\*\*

Road.Test.Score 1.198e-02 1.489e-03 8.045 1.21e-09 \*\*\*

Predicted.Reliability 1.744e-01 1.264e-02 13.795 3.68e-16 \*\*\*

Upscale\_Sedan 6.420e-02 3.952e-02 1.625 0.113

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0688 on 37 degrees of freedom

Multiple R-squared: 0.9376, Adjusted R-squared: 0.9292

F-statistic: 111.3 on 5 and 37 DF, p-value: < 2.2e-16

The residuals represent the differences between the predicted value of the dependent variable (Value.Score) and the observed values in the training data. They indicate the model's prediction error for each data point. The summary statistics for residuals show:

* **Min**: The minimum residual value is -0.11847.
* **1Q**: 25% of the residuals fall below -0.04933.
* **Median**: The median residual is 0.00452.
* **3Q**: 75% of the residuals fall below 0.03970.
* **Max**: The maximum residual value is 0.14738.

Coefficients:

The coefficients section displays the estimated coefficients for each predictor variable in the model, along with their standard errors, t-values, and p-values. These coefficients represent the relationship between each predictor and the predicted Value.Score.

* **(Intercept)**: The intercept term is 1.128e+00 (1.128) which is the estimated Value.Score when all predictors are zero.
* **Price**: This predictor's coefficient is -9.269e-06 (-0.000009269), indicating a very small negative relationship with Value.Score, but it is not statistically significant at the 0.05 level (p = 0.101).
* **Cost.Mile**: This predictor's coefficient is -1.624e+00 (-1.624), suggesting a negative relationship with Value.Score. It's statistically significant (p < 0.001).
* **Road.Test.Score**: The coefficient is 1.198e-02 (0.01198), implying a positive relationship with Value.Score and is statistically significant (p < 0.001).
* **Predicted.Reliability**: This predictor has a coefficient of 1.744e-01 (0.1744), indicating a positive relationship with Value.Score and is highly statistically significant (p < 0.001).
* **Upscale\_Sedan**: The coefficient is 6.420e-02 (0.0642), suggesting a positive relationship with Value.Score, but it is not statistically significant at the 0.05 level (p = 0.113).

**Model Summary:**

* **Residual Standard Error**: It's the estimate of the standard deviation of the residuals, indicating the average amount that the observed Value.Score values deviate from the predicted values by the model (0.0688).
* **Multiple R-squared**: This measures the proportion of variability in the Value.Score that is explained by the model's predictors (93.76%).
* **Adjusted R-squared**: This is the R-squared value adjusted for the number of predictors in the model (92.92%). It penalizes for the number of predictors used.
* **F-statistic**: It tests the overall significance of the model. A higher F-statistic with a small p-value suggests that at least one predictor variable is significantly related to the response variable.

Overall, the model seems to have a good fit as indicated by the high R-squared value and significant predictor coefficients, except for "Price" and "Upscale\_Sedan," which might not be as influential in predicting Value.Score.

**3.1.2 Predictions based on Test data**

Using test data, we obtain the below coefficient estimates:

Top of Form

Residuals:

Min 1Q Median 3Q Max

-0.11847 -0.04933 0.00452 0.03970 0.14738

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.128e+00 1.319e-01 8.553 2.71e-10 \*\*\*

Price -9.269e-06 5.510e-06 -1.682 0.101

Cost.Mile -1.624e+00 3.272e-01 -4.962 1.58e-05 \*\*\*

Road.Test.Score 1.198e-02 1.489e-03 8.045 1.21e-09 \*\*\*

Predicted.Reliability 1.744e-01 1.264e-02 13.795 3.68e-16 \*\*\*

Upscale\_Sedan 6.420e-02 3.952e-02 1.625 0.113

---

Significance. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0688 on 37 degrees of freedom

Multiple R-squared: 0.9376, Adjusted R-squared: 0.9292

F-statistic: 111.3 on 5 and 37 DF, p-value: < 2.2e-16

**Residuals:**

The statistics related to residuals remain the same as seen in the model using the training data. These statistics indicate the distribution and range of prediction errors between the observed and predicted values of Value.Score in the test dataset.

**Coefficients:**

The coefficients section displays the estimated coefficients for each predictor variable in the model when applied to the test data. These coefficients and their statistical significance are identical to the coefficients obtained from the training data.

* **Price:** Continues to have a small negative relationship with Value.Score, which remains statistically insignificant at the 0.05 level.
* **Cost.Mile:** Still shows a significant negative relationship with Value.Score.
* **Road.Test.Score:** Maintains its positive relationship and significance with Value.Score.
* **Predicted.Reliability:** Continues to show a strong positive relationship and high significance with Value.Score.
* **Upscale\_Sedan:** Similarly, this predictor remains statistically insignificant at the 0.05 level, though it suggests a positive relationship.

**Model Summary:**

The summary statistics related to the model fit, residual standard error, R-squared values, and F-statistic remain identical to the model summary from the training data. This indicates that the model performs consistently on the test data as it did on the training data.

In essence, when applied to the test dataset, the model maintains its predictive capability, with the predictor coefficients and model fit metrics remaining consistent, suggesting that the model generalizes well and doesn't show signs of overfitting to the training data.

* 1. **Conclusion**

**Factors Impacting Car Value:**

**1. Predictor Variables:**

* **Price**: The analysis suggests a weak and statistically insignificant relationship between the price of the car and its perceived value. This implies that, contrary to popular belief, higher-priced cars might not necessarily translate to higher perceived value.
* **Cost per Mile**: Cars with lower costs per mile driven in the first five years of ownership tend to have higher perceived value. This variable shows a significant negative relationship with value scores, implying that lower ownership costs contribute positively to perceived value.
* **Road Test Score**: Higher road test scores, indicating better performance, comfort, convenience, and fuel economy, are associated with higher perceived value. This variable shows a significant positive relationship with value scores.
* **Predicted Reliability:** Cars with higher predicted reliability ratings tend to have significantly higher perceived value. This variable exhibits a strong positive relationship with value scores.
* **Size Categories (Small, Family, and Upscale Sedans):** The analysis shows varied associations based on size categories. Family and upscale sedans tend to have relatively higher perceived value compared to small sedans, but this relationship is less significant.

**2. Model Performance:**

* The linear regression model explains approximately 93.76% of the variability in perceived car values, as indicated by the high R-squared value.
* The model's predictive ability remains consistent when applied to the test dataset, signifying its generalizability and reliability.

**3. Implications and Recommendations:**

* **Importance of Reliability**: The most influential factor impacting perceived value is predicted reliability. Emphasizing and improving the reliability of cars could potentially enhance their perceived value among consumers.
* **Cost Consideration**: Lowering ownership costs per mile driven could positively influence perceived value. Strategies to minimize maintenance, repair, and fuel expenses could make cars more attractive in terms of perceived value.
* **Performance and Test Scores**: Continuously striving for better performance, as reflected in higher road test scores, can contribute positively to the perceived value of cars.

**4. Limitations and Further Research**:

* **Additional Variables**: Exploring additional variables such as brand reputation, safety features, or specific consumer preferences could provide a more comprehensive understanding of perceived car value.
* **Timeframe Consideration**: The analysis is based on data up to a certain point. Considering more recent data could reveal shifts in consumer preferences or market dynamics impacting perceived value.

In conclusion, while price plays a minimal role, factors like cost efficiency, reliability, and performance significantly impact the perceived value of cars. Understanding and focusing on these key aspects could aid manufacturers in creating cars that are not only reliable and cost-efficient but also hold higher perceived value among consumers.