Problem Statement:

A used car dealership specializes in selling cars from various brands. They would like to know if the mileage of these cars is a good predictor of their sale prices, and if the slopes and intercepts differ when comparing mileage and price for different brands of cars. What other factors might play a role and how in deciding the price that a customer might be willing to pay. As a data expert, the company relies on your expert analysis and recommendations to increase their profitability by setting the right pricing for their car sales business, such that it delights the customers and gains the company positive feedback/reviews so that their traction in the market increases and they can become one of the key players.

Dataset Source :

https://drive.google.com/file/d/1Li-1EvAmOW8qznQTQKHy3nueOYSf855V/view?usp=sharing

Tasks:

1. Collect and preprocess the data: Collect data on the mileage and sale prices of used cars. Preprocess the data by cleaning it, handling missing values, and transforming it as necessary.

2. Conduct exploratory data analysis (EDA): Conduct EDA to identify patterns and relationships between mileage and sale prices for each brand of the car.

3. Split the data: Split the data into training and testing sets. Use the training set to build the regression model and the testing set to evaluate its performance.

4. Choose a regression model: Choose an appropriate regression model to use based on the characteristics of the data and research question. Common types of regression models used for this problem include linear regression, polynomial regression, and multiple regression.

5. Evaluate the regression model: Evaluate the performance of the regression model using various metrics, such as R-squared, mean squared error (MSE), or root mean squared error (RMSE).

6. Interpret the coefficients: Interpret the coefficients of the regression model to determine if mileage is a good predictor of sale prices for each brand of car, and if the slopes and intercepts differ between the brands.

7. Make predictions: Once the regression model is trained and evaluated, use it to make predictions on new data. These predictions can be used to estimate the sale prices of luxury cars based on their mileage.

8. Provide recommendations: Based on the analysis, provide recommendations to the used car dealership on how to best price their cars to achieve their business objectives.

Apart from your notebook, you also need to create the following as described:

Slide deck(PPT)

Create a slide deck presenting the main findings of your analysis. The slide deck should have no more than 6 content slides + 1 title slide. Here is a suggested outline as you think through the slides; you do not have to use this exact format for the 6 slides.

Title Slide
Slide 1: Introduce the topic and motivation
Slide 2: Introduce the data
Slide 3: Highlights from EDA
Slide 4: Final model
Slide 5: Interesting findings from the model
Slide 6: Conclusions + future work

You can use the software of your choice to create your slide deck. Save your slide deck as PDF or provide a link to view your slides online (e.g. in Google Slides). Be sure you grant the correct access permissions, we have access to your slides.Present your findings in the form of a PPT.

You also need to prepare a write up for the project which gives insight into your methodology and process. Use the following sections to help organize your write up(Documentation/Presentation):

Section 1: Introduction

This section includes an introduction to the project motivation, data, and research question. Describe the data and definitions of key variables. It should also include some exploratory data analysis. All of the EDA won’t fit in the paper, so focus on the EDA for the response variable and a few other interesting variables and relationships.

Section 2: Regression Analysis

This section includes a brief description of your modeling process. Describe how you chose the modeling approach, how you conducted model selection, interactions you considered, and any variable transformations. This is also where you will output the final model and include a brief discussion of the model assumptions, diagnostics, and any model fit statistics (e.g. R2 , AUC, etc.)

Section 3: Discussion

This section includes any relevant prediction and/or conclusions from your model. This should not just be a list of coefficient interpretations but rather use the interpretations from the model to support your conclusions. Remember, you are sharing a narrative with a business or research colleague who could potentially use your model to make decisions. They want to understand the practical conclusions and implications of the model results.

Section 4: Limitations

This section includes discussion about issues pertaining to the reliability and validity of your data or appropriateness of the regression analysis. This can be 1 - 2 paragraphs on what you would do differently if you were able to redo the project or next steps if you could continue working on the project.

Section 5: Conclusion

This section includes a summary of your key results and any final points you want the reader to learn. It can also include ideas about future work.

Section 6: Additional Work

This section includes anything that is not included in the main body of the paper. This could be additional EDA, other models you've tried, additional analysis, etc. There is no page limit on the additional work, but it should still be neatly organized and easy for the reader to navigate.

Grading Criteria:

The project will be graded based on the following criteria:

- Consistency: Did you clearly answer the question of interest?
- Clarity: Can the audience easily understand your analysis process and any sort of conclusions/arguments you make?
- Relevancy: Did you use the appropriate statistical techniques to address your question? Was your analysis thorough (e.g. did you consider interactions in addition to main effects?)?
- Interest: Did you attempt to answer a challenging and interesting question rather than just calculating a lot of descriptive statistics and simple linear regression models?
- Organization: Is your write up and presentation organized in a way that is neat and clear for the audience to understand?

A general breakdown of scoring is as follows:

- 90%-100%: Outstanding effort. Students understand how to apply all statistical concepts, can put the results into a cogent argument, can identify weaknesses in the argument, and can clearly communicate the results to others.

- 80%-89%: Good effort. Students understand most of the concepts, put together an adequate argument, identify some weaknesses of their argument, and communicate most results clearly to others.
- 70%-79%: Passing effort. Students have misunderstandings of concepts in several areas, have some trouble putting results together in a cogent argument, and communication of results is sometimes unclear.
- 60%-69%: Struggling effort. Student is making some effort, but has a misunderstanding of many concepts and is unable to put together a cogent argument. Communication of results is unclear.
- Below 60%: Students are not making a sufficient effort.