Experian –Case Study

The purpose of this case study is the understand your Python coding and data analysis & modelling skills. You don't need to create a super accurately performing model, and instead show us that you know your fundamentals when it comes to data analysis and modelling!

We ask you to deliver a Jupyter notebook containing your work (including your codes, graphs, answers, and comments). Good luck!

You are working as a Data Scientist in *Carsperian*, a car manufacturing company. The company has recently decided to enter the US market and needs to plan a marketing / pricing strategy.

As the Data Analytics team you are asked to:

- Analyze which car models/specifications are more suitable for the market

- Give a reasonable estimate price with regards to the current market prices

Your manager has tasked you with this project and eager to impress your manager, you decided to analyze the existing competitors and collected all available market data. The dataset contains detailed information regarding the cars available on the market (such as its brand, model, horsepower, etc.) as well as their price. You will use this dataset to answer the questions below.

Question 1: Exploratory Data Analysis

Before diving into the modeling, you need to conduct data analysis to get a general view of the dataset

a) What does the dataset look like? How many records & features are there?
b) Which data types are available?
c) What are the summary statistics for the numerical variables?
   (You don't need to go into a detailed explanation, you can present the basic summary including mean, median, min, max, Q1 & Q3 percentiles etc.)
d) How many car brands are there within the dataset?
e) Are there any missing values? If there were, what methods would you used to fix this issue?
   (You don't need to go into detail, you can explain it in 2-3 sentences)

Question 2: Data Visualization

As you need to present your findings to all business stakeholders, you decided to create visuals that best explains the current car specifications available on the market. Using the variables available, create 3-5 plots that are interesting to you, that gives insight on what types of cars we should be introducing to the market? (Hint: Using different types of charts would be a plus!)

Question 3: Data Analysis & Modelling

Moving on to the modelling, you decided to use a Linear Regression model to estimate the price. (Due to company policy, you don't need to use any categorical variables for modelling)

a) For feature selection, plot a correlation matrix and comment on your findings
b) The threshold correlation value is selected as 0.75, which variables will you include in the model? Are there any variable pairs that might be affected by multicollinearity? If so, how will you deal with this issue?
c) Now that you are done with feature selection, prepare the features for the model, (Hint: you need to "scale" your variables) and split the dataset into test-train sets (30% and 70% respectively)
d) Perform multiple linear regression? What are the model results? Are there any variables with p-values larger than 0.05? If so, how should you proceed? Please comment on your findings.
e) To automize the model process for later use, can you create a model pipeline that includes the following steps:
    a. A class that performs all the necessary feature engineering & selection
    b. Scales the data accordingly
    c. Gets model output

    How would you fit the training dataset and then get a prediction for test dataset using this pipeline?

Question 4: Uploading to AWS

You are done with the study and generated great insights for the business. Impressed with your work, your manager asked you to store the project files (dataset and model) in the company AWS S3 server. You are provided with the following information:

```
ACCESS_KEY = "my_access_key"
SECRET_KEY = "super_secrey_key"
SESSION_TOKEN = "session_token_123"
BUCKET_NAME = "carsperian-market-price-analysis-2106",
```

a) Convert your model into a pickle file
b) Using the information given above, how would you upload the model and the dataset to company S3 environment? (Hint: You need to use an external Python package to do this. You don't need to specify an "OBEJECT_NAME" as just use the files names instead. As you might not be able to test your code, just create a script that you think that should work)