Question 3:
Intent: Check you problem solving approach on machine learning

Consider the dataset on Question 2. Now,
A client has solicited your services to develop a machine-learning model that can forecast the approximate value of their customers' used cars. The objective is to provide accurate quotations to customers on the price to offer for the purchase of their used cars. You have been furnished with a dataset of used cars, and your task is to:

1. conduct exploratory data analysis to identify crucial features that will be utilized in the model.
2. Please justify the selection of these features and aim to incorporate as many as possible.
3. kindly identify any potential challenges or limitations you anticipate/encounter during the feature selection process. (if any)
4. (Bonus) Try to propose a good model you feel would be able to best fit the features you have selected to make predictions.

I could not find any solution to this question as most of the terminologies felt unfamiliar and I was not sure as to which procedure to use so as to find solutions to the above mentioned queries

I will certainly go through with the question again and try to come up with a solution !

The steps I took up to understand the problem statement!

- Firstly, as Machine Learning all together was a new challenge for me, I started off by understanding what exactly machine learning is, why is it used, how the process is done etc.
- Leaving behind all the complexity, ML is a way to be able to draw conclusion from data (a way of data analysis)
- Let’s say we have been given few fields, e.g.- price and quantity of a commodity. We can easily predict that price increases with quantity, but this increase need not be linear. So what needs to be done here is **find a pattern that explains how exactly they’re related, express that as a model and use the model for prediction of unknown data.**
- But the given dataset, there are parameters and much more permutation and combinations that will be difficult for us to carry forward, that is why we take computer’s help to figure out a pattern based on some rules that we set!

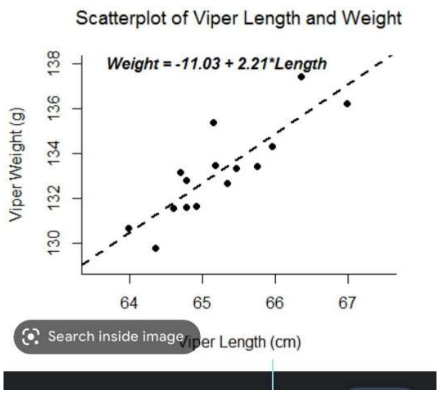
- step 1 - conduct exploratory data analysis to identify crucial features utilised in the project.
  - ➢ Step 1 is to identify fields like vehicleType, brand, kilometre etc, which may pay an important role in solving the objective **[The objective is to provide accurate quotations to customers on the price to offer for the purchase of their used cars]** i.e., Used car as in....2nd hand car. We need to analyse different features with keeping price as the dependent variable and conduct analysis.

https://colab.research.google.com/drive/1mRQsGEgkopliFOCArdplLiIypv-BXlBu?usp=sharing

has the solution for my trial

I came across the concept of linear regression where we plot our dependent variable (price) on the y-axis against the independent variables.
Here the independent variable can be brand, model, kilometre, yearOfRegistration, vehicleType etc
Based on these points the quotation can be calculated.
But the drawback of linear regression model is that we can plot against only 2 numerical values and here as we can see there are many string data type based attributes, so plotting the graph dint give accurate results.

During linear regression we get a plot something similar to this



The points here are the data points
The distance of each data point from the line is called error
When data points are far from the line error is high, that is, there’s a weak correlation between the 2 chosen parameters
When data points are near the line error is less, that is, there’s a strong correlation between the chosen parameters

Now what is Correlation?
it shows whether and how much the changes in one variable are related to changes in another variable.
It can range between +1 and -1
+1 indicates there is positive correlation between the parameters and when one increases the other value also increases
-1 means there is negative correlation between the parameters and when one increases the other decreases
0 indicates no correlation between the two.

However correlation being positive need not always imply that the parameter is the best choice for the crucial feature list.

LIMITATION

- It looks like there are few **unnecessary data** values in the dataset so initially we will have to clean the data set
- Another issue is **missing data,** it can be seen that few data in the data set are missing, it may lead to wrong predictions.

To find the best feature we need to make the data set undergo many iterations.
So we can see that its not a one step process
A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.