Assignment-based Subjective Questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: a) Business was gaining popularity over year there is increase in % of bike booking in 2019 than in 2018.

b) Maximum no of bookings were in fall season, and in spring it was quite less that means people prefer to use bike share in case of fall summer and winter.

c) From month 1-12 middle moths have maximum no of bookings , distribution follows bell curve

d) On working day people use more bikes than none working day may be they prefer to commute using bike share

e) Holiday has less no of bookings

d) In extreme weather condition people will not book bikes, maximum no of bikes booked in clear season and very less bikes booked in thunderstorm

Q2) Why is it important to use drop_first=True during dummy variable creation?

Ans) Using this we can avoid creation of multiple dummy variables, it helps in extra column created during execution of command

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans) 'Registered' is having highest correlation, this could be because if person it registering than he/she is intended to use the bike and hence count of booking increases

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans) Below are the validations:

1) Drew correlation matrix and found that none of the variable was dependent with target variable as 1 correlation factor
2) During pair plotting could see linear dependency of variables with the Target variable
3) Checked model prediction against the actual test Y data to see if there is linear relation .
4) Checked the residuals having mean around 0 and were normally distributed .

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1) Registered
2) Temperature
3) Month

General Subjective Questions

1) **Explain the linear regression algorithm in detail**

   Ans: Linear regression is ML algorithm where we have set of data input and output both some part of it will be used build the model and rest for evaluating the model , since it uses output too it is supervised machine learning algorithm

   There are certain assumptions to keep in mind when it comes to Linear Regression

   1) Relationship between the final/target variable is linear with respect to independent variables or predictors
   2) Errors are normally distributed
   3) In case of multiple predictors, predictors should be independent of each other

      Etc

Below are the steps performed in case of linear regression

1) First of all we need to understand data well, transform data when needed, check for wrong data etc
2) Next would be to find correlation between variables for this we can use correlation matrix
3) In given dataset then we need to reserve amount for Training data and Test data , we can follow any rule for the same most commons ones are 70:30 or 80:20 while segregating the data we can do it randomly . There are inbuilt functions in python libraries which can be used for the same
4) Next is to use training data and predict the coefficients , linear regression follows straight line equation and hence we will try to find best suited value of m and c in below linear regression equation

$$Y = m X + C$$

5) In given data set first we start with the single variate analysis and try to train the model one of the way could be use OLS (ordinary least squares)
6) After which we check the model parameters and summary to check significance and prediction percentage
7) Like wise we can use different variables to find the model
8) In case we don’t get suitable result we build model using multiple variables
9) Process is same as single variable . There would be multiple variables in training and test data which will be used to predict the target variable
10) In case of multivariate things like scaling would be of importance there are multiple ways to do the same like min max, normalization
11) After building multiple models we need to also evaluate the model
12) For this we will use R square method where first we will predict Y values from test set using model and then calculate R square with actual value
13) We will compare multiple R squares to get best value of model or best suited model.

Q2) Explain the Anscombe’s quartet in detail.

Ans: It can be defined as four data sets which are almost identical but they have very different distributions when we draw graph (scatter)

This demonstrates importance of analysing or doing EDA before performing Linear regression or other algorithm for model formation

This suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.

Data-set II — shows a curve shape but doesn’t show a linear relationship (might be quadratic?).

Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.

Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

Q3) What is Pearson’s R?

Pearson’s R is also termed as Pearson Correlation Coefficient . It is most common way Linear Correlation measurement

This is ultimately number between -1 and 1 that measures strength and direction of relationship between two variables

If value of it is between 0 and 1 that means positive correlation if one variable increases other also increases and vice versa for example as temperature increases demand of water also increases means temperature and water demand are positively corelated

If value is between -1 and 0 it means one variable is negatively correlated with another variable as the age of car increase its net worth decreases

0 means there is no relation ship between two variables, two variables are independent of each other

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

In general terms scaling is process of enlargement or diminishment of a factor. In terms of machine learning when we modify set of values or a factor to fall with in range it is called scaling

We need it for :

1)Ease of interpretation, when we have multiple variables which has hugely different ranges it becomes very difficult to compare it . Hence we scale such variables

2) Faster convergence of gradient descendent method

There are majorly two ways for scaling

1) Standardization: Subtracting by mean and dividing by standard deviation such that it is centred at 0 and has standard deviation of 1.

2) MinMax scaling : Is getting the value between 0 and 1 it is calculated using Xmin and Xmax for feature X

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF is variance inflation factor, large value of VIF indicates that there is high correlation between the variables. If the VIF is 8, this means that the variance of the model coefficient is inflated by a factor of 8 due to the presence of multicollinearity. It could happen that two variables are completely correlated if this happens that means change in value of one variable will have definite effect on other variable. If there is perfect correlation then value of VIF will be infinity

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.:

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.