Assignment-based Subjective Questions

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

- Fall season has more bookings compare to other seasons.
- 2019 has more no of bookings compared to 2018,
- June and Sephas more booking compared to other months however may, July, Aug, and sep also having good no of bookings
- There is not much difference between the day's of the week
- Clear weather has attracted more no of bookings compare to other weathers
- The no of bookings are less for holidays compared to not holiday
- There is no much difference between the working and non-working days

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

drop_first=True is important to use while creating dummy variable is because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

temp variables is having the highest correlation with the target variable

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

I have validated the assumptions of linear regression based on three assumptions which are Normality of error terms, Multilinearity check, and linear relationship validation

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

temp, winter, and sep features are most significant towards explaining the demand of shared bikes

General Subjective Questions Answers:

1). Explain the linear regression algorithm in detail?

Linear regression is a simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the

independent variable(predictive variables) and the dependent variable(target variables). Further, it is classified as Simple linear regression and multiple linear regression. If there is a single input variable, such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

- We plot a graph between the variables which best fit the given data points.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).
- To calculate best-fit line linear regression uses a traditional slope-intercept form. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.
- The linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error. In Linear Regression, Mean Squared Error (MSE) cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE)

2). Explain the Anscombe's quartet in detail?

There are different data sets that are nearly identical in simple descriptive statistics like mean, median. And variance etc when we observe the datasets. But the data sets have very different distributions when we visualise the data on scatter plots. Anscombe's quartet illustrate the importance of plotting data before you analyze it and build your model.

It suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3). What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1.

Pearson r Formula

R = (the covariance of two variables)/(the product of their standard deviations)

- R = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- R = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- R = 0 means there is no linear association
- 0 > R< 5 means there is a weak association
- 5>R < 8 means there is a moderate association
- R > 8 means there is a strong association

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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why scaling is important because of, Most of the times, collected data set contains features highly varying in magnitudes, units and range because of different type of variables. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Actually, Scaling only affects the coefficients but not the other important parameters like P-values and R-square.

Normalisation or Min-Max scaling brings all of the data in the range of 0 and 1 where as Standardization replaces the values by their Z scores and it brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

This is happens because of the multicollinearity. Which means a few variables may be expressed exactly by a linear combination of other variables. In those cases R2=1 which lead to VIF = 1/(1-R2) is infinity. To

solve this problem, we should drop one of the variable from the data set which is causing the perfect multicollinearity.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the

Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.