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? (3 marks)
    a. MistOrBroken_clouds and Clear_few/partly_Clouds weather have high bookings.
    b. Clearly number of bookings are increased in 2019
    c. And mistBroken_clouds and Clear_Few/Partly_Clouds weathers have seen more bookings in 2018 and 2019.
    d. Summer, fall and winter have many bookings compared to spring.
    e. Friday, Saturday, and Thursday have a slightly greater number of bookings.
    f. Mondays have more bookings on holidays.
    g. The second and third weeks of the month seem to have more bookings.
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

    When we create dummy variables (o's and 1's) we just need n-1 dummy variables (n is number of unique values). Ex : if we are creating a variable with values Yes and No. Yes means 1 and No means 0. We just need one dummy variable to present Yes (1) and No (0). It will reduce the number of dummy variable columns. Hence, it is important to use drop_first=True.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

    Temp and atemp variables have highest correlation with target(cnt). Registered and casual variables also have the highest correlation but since they are also count of the bike bookings, we are ignoring them.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?(3 marks)
    a. Linear relation between independent variables and target variable.
    b. Error term should be normally distributed.
    c. Multicollinearity should be insignificant. Independent variables in the model should not be correlated with each other.
    d. Homoscedasticity: there should be no clear pattern in residual terms.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

    Temp, yr and Snow_Rain_ScatteredClouds weather are the three variables contributing significantly towards explaining the demand of bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.(4 marks)

    Linear regression attempts to explain the relationship between a dependent and an independent variable using a straight line. The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables.

Linear regression is classified into two types based on the number of independent variables.

1. Simple linear regression
1. Multiple linear regression

Mathematically equation for the relation between independent variable and target(dependent) variable is $Y = \beta_0 + \beta_1 X_1$

Y is target(dependent) variable.

X is an independent or predictor variable.

B_0 is constant.

Error $e_i = Y_i - Y_{Predicted}$

Formula for RSS Residual sum of squares(Ordinary least squares method)

$(e_1)^2+(e_2)^2+.....+(e_n)^2$ = RSS(Residual Sum of Squares)

$RSS = (Y_1- \beta_0 - \beta_1 X_1)^2+ (Y_2 – \beta_0-\beta_1 X_2)^2+...+ (Y_n – \beta_0-\beta_n X_n)^2$

When we have more than one independent variable then its called multiple linear regression $Y=\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n+\epsilon$

Assumptions for linear regression: -

Relationship between variables – Linear regression model assumes that the relationship between response and feature variables to be linear.

Multi-collinearity – Linear regression model assumes that there is little or no multi-collinearity in the data. Multi-collinearity occurs when the independent variables or features have a dependency on each other.

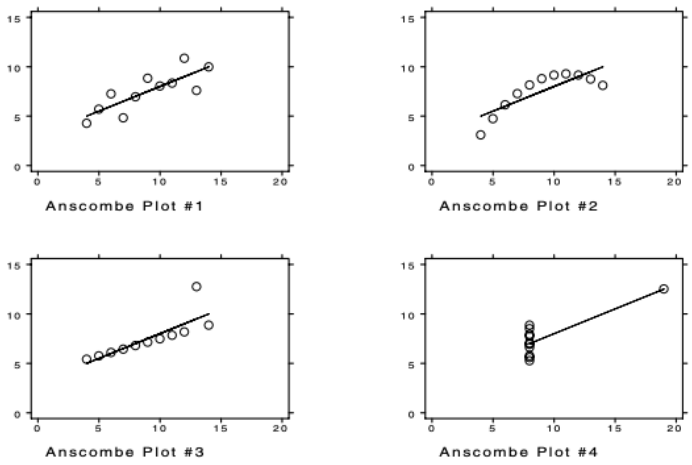
Normality of error terms – Error terms should be normally distributed

Homoscedasticity – There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. It can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear

differently when plotted on scatter plots as shown in below fig.



3. What is Pearson's R?(3 marks)

Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between the different variables and their relationships. Pearson's correlation coefficient returns a value between -1 and 1.

The interpretation of the correlation coefficient:

- If the correlation coefficient is -1, it indicates a strong negative relationship. It implies a perfect negative relationship between the variables.
- If the correlation coefficient is 0, it indicates no relationship.
- If the correlation coefficient is 1, it indicates a strong positive relationship. It implies a perfect positive relationship between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

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

Scaling Methods:-

Standardization(subtracting mean and dividing by standard deviation such that it is centered at zero and has standard deviation 1)

Standardized scaling is not much effected by outliers

Normalization(MinMax scaling)(getting the value between 1 and 0 or –1 and 1)

Minmax scaler is effected by outliers

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

   Q-Q plots are also known as Quantile-Quantile plots. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

   A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.