Assignment Based Subjective Questions

1. From the analysis we can infer that most of the independent variables doesn't affect the dependent variables. There are only a few which have good correlation.

2. Dummy variables are created in the form of binary, so if there are 3 dummy variables, then by only knowing the value of 2 dummy variables 3 would be known. So, we can always exclude one.

3. temp and atemp had the highest correlation

4. R-square talks about how much of data variation can be explained by the model. Having a good R square value for a model is necessary. Also, for MLR it is better to know the Adjusted R-square value as it penalised with respect to the number of features. With that the overall significance of a model can be explained by F-Static. Higher the value of F-static the more significant the model would be.

5. The top feature that explains demand of bikes are temp/atemp, weathersit and windspeed.

General Subjective Questions

1. In regression modelling technique the output variable is a continuous variable. So, the linear regression explains the relationship between predicted variable and the independent variable by a straight line. a. Simple Linear Regression b. Multiple Linear Regression c. Simple Linear Regression - SLR shows the relationship between predicted variable and one independent variable whereas in MLR the independent variables are more than one. The strength of LR is assessed by R square technique or Residual standard error.

2. Anscombe's quartet has four nearly similar descriptive statistics, but they have very different distributions and look very different on graphs. It states that while comparing the datasets, statistics might not give the entire picture, they can be very different when they are plotted on graphs.

3. Pearson's R - Pearson Correlation Coefficient measures the strength of relationship between two variables. It has a value between [-1,1]. -1 means complete negative correlation, 0 means no correlation and 1 means complete positive correlation.

4. Scaling - Scaling is a feature to convert independent variables into a similar scale for better interpretation. When there are many independent variables, they can vary to different scales and lead to weird coefficients which are harder to interpret. So, for better interpretation and faster convergence scaling is used. Standardized scaling - the mean of the independent variable is 0 and SD is 1 whereas in normalized scaling all the values lie between 0 and 1.

5. Infinite VIF means the perfect correlation between 2 independent variables. In case of perfect correlation R square is 1 which leads to infinite VIF. If we drop one of the variables from the dataset which is causing perfect multicollinearity, it can be resolved.

6. It compares 2 probability distributions by plotting their quantiles against each other. If the two distributions are exactly equal, then the graph would be a straight line. It answers the most fundamental question if the curve is Normally Distributed.