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?

A. The demad of bike is less in the month of spring when compared with other seasons. The demand bike increased i n the year 2019 when compared with year 2018.

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

A.drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creati on

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

A.we see that no variable is highly correlated with another variable in any way. So, we can further proceed and chec k Multi-Collinearity while creating models itself

4.How did you validate the assumptions of Linear Regression after building the model on the training set?
A. After building model, we cannot finalise untill we proove the residual analysis wherein we check whether the dist ribution of Error is around 0 or not.
From the Model Summary Report we can say that all the p-values of respective features are well under control.P-Val ue of "weekday_saturday" feature is more than 0.05, which make its co-efficient insignificant.

5.Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
A.weathersit_Light_Snow(negative correlation).
yr_2019(Positive correlation).
temp(Positive correlation).

General Subjective Questions:

1.Explain the linear regression algorithm in detail.
A Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Re gression models a target prediction value based on independent variables. It is mostly used for finding out the relatio nship between variables and forecasting. Different regression models differ based on  the kind of relationship bet ween dependent and independent variables they are considering, and the number of independent variables getting us ed.

2.Explain the Anscombe s quartet in detail.
A.Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive sta tistics, 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.

3.What is Pearson s R?
A.is a measure of linear correlation between two sets of data

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
A scaling:It 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.

Normalized scaling:It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to im

plement normalization in python.
Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution whic h has mean ( □ ) zero and standard deviation one (s).

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?
A If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent varia bles.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
A.Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theore tical distribution such as a Normal, exponential or Uniform distribution.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certa in values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point a nd 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 4 5 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall o n that reference line.