**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)

Working days, season time and year has a positive effect on target variable.

Weather situation, holiday, wind speed has a negative effect on target variable.

1. Why is it important to use **drop\_first=True** during dummy variable creation? (2 mark)

Drop\_first = True helps to eliminate 1 column during dummy variable creation. For example, if a categorical column has 4 variables, then drop first feature will create 3 columns (n-1) without losing any information.

1. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

Temperature (feels like) has the highest correlation of 0.63

1. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

by performing prediction on validation data and checking metrics like r2\_score, rmse.

1. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Year (2019 was a better year) > Feels like Temperature, Season, Working days

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning model that target prediction values of target variable based on independent variables. It is used to find out relationship between variables and forecasting.

Linear Model = Bo + B1X1 + B2X2……… BnXn

1. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet is made up of 4 data sets that have similar descriptive statistics but when plotted on graph, they appear very different due to effect of outliers and other factors. It demonstrates the importance of visualizing data.

1. What is Pearson’s R?

The Pearson’s R is a method to measure a linear correlation between different variables. It is a number between -1 and 1 that measures the strength and direction of relationship between 2 variables. 0 to -1 correspond to negative correlation with -1 being the most negative correlation and vice versa for 0 to 1.

1. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling? (3 marks)

Scaling is a preprocessing step to scale the data on same scale in a certain range. It helps to speed up analysis on huge dataset and maintain same scale across various variables.

Normalized scaling also known as min-max scaling scales the data in the range of 0 to 1 while Standardize scaling replaces the values by their Z scores by bringing all data into a standard normal distribution with a mean of zero and standard deviation of 1.

MinMaxScaler is better for dataset with a dataset containing extreme values.

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

(3 marks)

VIF is infinite when there is perfect correlation between 2 variables. An infinite VIF indicates that corresponding variable may be expressed exactly by a linear combination of other variables.

1. 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 called as Quantile-Quantile plots. They plot 2 quantiles against each other.

The use of Q-Q plots is to find out if two sets of data come from same distribution. If 2 data sets come from a common distribution, the points will fall on the reference line.