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)

Ans – The categorical variables are season, yr, mnth, holiday, weekday, weathersit are categorical variables in the dataset.

- The bike demand is significantly less in SPRING and more active in FALL.
- The Bike demand is active from June to September.
- Due to unfavorable winter weather, the demand drops i.e. months of November, December and January.
- The demand for bikes is more for Pleasant/ Clear weather and on Weekdays.
- There is less demand for bikes when there are holidays/weekends.

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

Ans- 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? (1 mark)

Ans- ‘atemp’ and ‘temp’ are highly correlated to each other.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans – The model fitted on training displayed that there is linear relationship between target and independent variables using model summary (good r2 score and fitted the line equation giving linear relation), the error terms are normally distributed and independent based on plots/calculations and there is a constant variance when predicted using training set. All this says that assumptions of linear regression have been validated.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans-

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Temp = 0.452098
2. Year = 0.234489
3. Weather
(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) = - 0.282231

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Ans - The Linear regression algorithm is a regression method that computes a linear relationship between variables and further to predict continuous values, such as age, salary. The variable you want to predict is called the dependent/target variable. The variable you are using to predict the other variable's value is called the independent variable. Typically the algorithm can be Simple linear regression (Having 1 independent and 1 dependent variable) or Multiple linear regression (Having many independent and 1 dependent variable). The main ideology is to find and fit the best straight line equation(y=mx+c), based on independent variables, considering the error/residuals as near to 0. Each

independent variable is associated with coefficient that fits it on plane. Along with it, the linear regression follows some assumptions as below:

- There must be linear relation between independent variables and dependent variable.
- The error terms must be normally distributed and center to 0.
- There should be constant variance.

2. Explain the Anscombe's quartet in detail. (3 marks)
Ans- Anscombe's Quartet consists of 4 data sets with same descriptive statistics. The data inside may be different, but the descriptive statistics come out similar. This causes an illusion that all 4 datasets are similar in nature and may further lead to wrong assumptions while performing analytical solution. This is where Data visualization comes into picture. Data visualization sheds light on areas where calculated summary stats fails. The data has different distributions and appear differently when plotted using any plot/graph and gives more insights. The Anscombe's Quartet tells us that we should not completely rely on the summary/descriptive statistics and must use Data visualizations to make it clear and understandable.

3. What is Pearson's R? (3 marks)
Ans- The Pearson's R Correlation is used to check the linear strength or linear correlation between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation. For example; in most cases, universally, the income of a person increases as his/her experience increases.

We should use the Pearson correlation coefficient when the relationship is linear, both variables are continuous in nature and have no outliers present.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling? (3 marks)
Ans- Scaling is process of converting/manipulating different variables with

different units on to a same plane for better model understanding. Since different variables have different units, it can be very difficult to work with them. Often variables with large scale units turn a model unstable since larger weights are being used/learned by model. This may degrade the prediction power or even its learning ability to find patterns in dataset. So to get the optimal weights with minimal error rate, we need to scale dataset.

Normalization refers to rescaling values of original variables to range of 0 and 1. The data is fitted using $y = (x - min) / (max - min)$ formula. Here the data is not heavily manipulated and works best when outliers are present. Example: MinMax Scaler.

Standardization refers to rescaling of distribution of original variables using mean=0 and standard deviation=1. The data is rescaled by $y = (x - mean) / standard\_deviation$. Here the data is heavily manipulated.

Example: Standard Scaler

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

Ans- **VIF** measures the strength of the correlation between the independent variables and this detects multi-collinearity issue. The value of VIF can be infinite when the variables are perfectly collinear or correlated to each other.

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

Ans- The Q-Q plot shows how the quantiles of two distributions line up, with our theoretical distribution (e.g., the normal distribution) as the x variable and our model residuals as the y variable. It is used to determine whether two sets of data come from the same statistical distribution. It is very useful in linear regression when we have different testing and training datasets. In this case, it is critical to ensure that both sets of data come from the same source in order to keep the model sane.