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?

const	*season*	*yr*	*mnth*	*holiday*	*weekday*	*workingday*	*weathersit*	*temp*	*atemp*	*hum*	*windspeed*
1458.492	*508.1607*	*2047.063*	*-39.7384*	*-518.4792*	*69.1898*	*124.1337*	*-610.4661*	*48.4537*	*72.2032*	*-9.8794*	*-38.2316*

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

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

Year and weather had highest correlation with target variable.

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

By plot residuals with each independent variable, like temperature. Performing a residual analysis and checking if residuals are normally distributed with mean=0. However, serial correlation among residual does exist.

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

Bike demand was mostly explained by Year, season, weather.

season (1:spring, 2:summer, 3:fall, 4:winter). Coefficient is 508.1607, hence every unit increase in season then demand increases by 508 units

year (0: 2018, 1:2019). Coefficient is 2047.063, hence with every unit increase in year demand increases. In other words, sales in 2019 demand were higher than 2018 or presence of yearly trend can be there.

weathersit : 1= Clear, 2= Mist + Cloudy, 3= Light Snow, 4= Heavy Rain. Coefficient is (-) 610.4661 and indicates lower demand as weather becomes snowy or heavy raining.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators *that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).*

2. Explain the Anscombe's quartet in detail.

3. What is Pearson's R?

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

and standardized 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. Most of the times, collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared etc.

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

It occurs when multicollinearity exist among >=2 independent variables

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