General Subjective Questions

1. **Explain the linear regression algorithm in detail**

   Linear Regression algorithm is a machine learning model which gives a continuous output variable. It is a supervised learning method in which we have previous data and based on that data we build the model. It is a predictive modelling technique used to establish relationship between dependent variable with independent variable.

   Steps involved in Linear Regression Algorithm:

   - **Data Visualization**

     Here Exploratory Data Analysis(EDA) needs to be performed on the data to understand various variables of the data. Target variable needs to be fixed in this phase. Data cleaning needs to be done here. Then correlation between different variables needs to be found out. Correlations can be used to detect multicollinearity.

   - **Data Preparation**

     Categorical features needs to be identified in the data. Dummy variables needs to be created for categorical features which has more than binary values. Scaling needs to be done so that all the variables will have equal importance with respect to target variable. Scaling can be done either by standardizing or Minmax scaling. In standardizing mean will be zero and standard deviation is one where as in minmax scaling variables are scaled in such a way that all values are between zero and one. Variables needs to be divided into dependent and independent variables. Variables that are highly related to others should be removed. Variables that doesn't have any business relation needs to be removed.

   - **Data Modelling & Evaluation**

     Data needs to be divided into test and train sets. R square and p values needs to be checked to evaluate the models.

2. **Explain the Anscombe's quartet in detail**

   In 1973 statistician Francis Anscombe demonstrated that analyzing data only by descriptive analytics is not enough and graphing the data is quite important. He did this by making 4 sets of data consisting of 11 biaxial points each. When the data is analysed we can see that the mean, standard deviation, correlation of x, y points of each data is same but when we plot the same data we found that they are completely different. First data scatter plot seems like there is a linear relationship between x and y. Second data set scatter plot seems like there is a non linear relationship between x and y. Third data set scatter plot seems like there is a perfect linear relationship between x and y except one point which might be a outlier. Fourth data set scatter plot shows anexample where one high leverage point is enough to produce a high correlation coefficient.

3. **What is Pearson's R?**

   Pearson's R also known as pearson's correlation coefficient or bivariate correlation or pearson product moment correlation coefficient or the correlation coefficient is used for measuring a linear correlation. It gives us the strength and direction of the linear relationship between the two variables. Its value lies between +1 and -1. When its value is positive i.e. between +1 and 0 we can find a positive linear relationship between the two variables, when its value is negative i.e. between -1 and 0 we can find a negative linear relationship between the two variables, when its values is 0 there is no correlation between the two variables. It also shows how distant are the points from the best fit line. If its value is +1 or -1 then then all the points are on the best fit line. It also tells us the slope of the best fit line. Pearson's R is better used when the variables are quantitative, normally distributed, have no outliers and have preferably a linear relationship.

4. **What is Scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   In a data different independent variables has different set of values which will affect the importance the algorithm gives to the variables because of their higher or lower values. To neutralize it the values of these independent variables will be scaled to a certain range of values so that all the variables will be given equal importance. Scaling is performed for the ease of interpretation and faster convergence for gradient descent methods.

   In normalized scaling, scaling is performed using mean and standard deviation to have the scaled values mean as zero and standard deviation as one where as in standardized scaling, scaling is performed using maximum and minimum values to bring the scaled values to the range of [0,1].

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

   Variation Inflation Factor(VIF) is used to check relationship of one independent variables with all the other independent variable. AS the value of VIF increases multi collinearity increases. Usually VIF<5 is preferred to avoid multi collinearity. The value of VIF will be infinite when there is perfect correlation between the variables.

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

   Q-Q plot is Quantile-Quantile plot where quantile of sample distribution will be plotted and compared with the standard distributions like normal, uniform and exponential to check if the sample distribution is following any of the theoretical subscription. Understanding which distribution help it follows helps us in determining the best model for the analysis in linear regression. It also tells us the skewness of the distribution. It is used to find out if two data sets come from populations with a common distribution or

have a similar distributional shapes or similar tail behavior or have common location and scale.If all the quantiles lie on the 45 degree angle from x-axis then it is similar population.

Assignment based Subjective Questions

1. **Why is it important to use drop_first=True during dummy variable creation?**
   It is used to avoid the extra column during dummy variable creation which reduces the correlation between dummy variables.
2. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Temp
3. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   By checking the R square value and p value
4. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Temp, season and month
5. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   They are showing both positive and negative correlation with the dependent variable