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?
   Ans: The categorical variables are :
   Season -
   The highest bike rentals are done in the fall season and the lowest are done in the spring season.

   month –
   The month is aligned with the season, the highest bike rentals are done in Jul, aug and sep months and the lowest bike rentals are done in Jan, Feb and Mar months.

   holiday –
   The bike rentals are less on holidays when compared to the non-holidays

   weekday –
   The average bike rentals are around the same on all the days. There is no relation with weekdays.

   workingday -
   The average bike rentals are around the same on all the working days and non-working days. There is no relation with working days column.

   weathersit –
   The bike rentals are high on clear weather days and low on the snow weather days.

2) Why is it important to use drop_first=True during dummy variable creation?
   Dummy variable is created for a categorical column. While creating dummy variables, drop_first is set to true to drop the first level of a column. In a regression model, there is a chance of independent variables are highly correlated. While creating dummy variable of having n levels, we create n-1 dummy variables. We will consider one of the level as a reference or baseline to compare the other levels.

3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   The registered variable has the highest correlation with target variable. The temp and atemp has the 2^{nd} highest correlation with the target variable.

4) How did you validate the assumptions of Linear Regression after building the model on the training set?
   once the features coefficients are having lesser p-values than 0.05 and VIF below 5, the model is finalised. The assumptions are :
   1) The error terms are normally distributed.
   2) The error terms are independent to each other.

5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Temp - the coefficient of 0.4279 indicates that 0.4279 units of bike rentals increases for every 1 unit of temp increases.
weather_snow - the coefficient of 0.2494 indicates that 0.2494 units of bike rentals decreases for every 1 unit of snow weather increases.
year - the coefficient of 0.2360 indicates that 0.2360 units of bike rentals increases for every 1 unit of year increases.

General Subjective Questions

1) Explain the linear regression algorithm in detail.
Linear regression is an algorithm used to predict a target value based on the input given. This algorithm is categorized as one of the supervised learning methods. This method is used to model the relationship between the target variable and one or more independent variables. There are some steps to be followed to build a model using linear regression. They are:
1) Data cleaning and preparation – In this step, we look at the data imputation if required and also we check for multicollinearity, having a strong correlation between the independent variables.
2) Split the train and test data
3) Building Model – we build the model by using the training data to fit into a straight line equation. Here we find the values of coefficients to predict the data using this model.
Y = b0 + b1X1 + b2X2 + ... + bn*Xn
    i) Where:
    ii) Y is the dependent variable
    iii) b0 is the intercept or constant term
    iv) b1, b2, ..., bn are the coefficients or slopes of the independent variables
    v) X1, X2, ..., Xn are the independent variables

4) Model Evaluation – after the model is trained with the training data, we use this model to predict the values from the test data.

2) Explain the Anscombe's quartet in detail.
Anscombe's quartet consists of four datasets which have same summary statistics(mean, standard deviation and the correlation) but when they are visualised, the four datasets are very different to each other.

3) What is Pearson's R?
Pearson's R is a measure of the linear relationship between two continuous variables. It is a measure of the strength and direction of the association between two variables. Pearson's R can range from -1 to 1, where a value of -1 indicates a perfect negative linear relationship, a value of 0 indicates no linear relationship, and a value of 1 indicates a perfect positive linear relationship.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Scaling is a step of pre-processing which is applied to the numeric variables to normalize the

data within the particular range. Most of the times, the dataset contains columns with different units. If they are considered to build a model, the algorithm takes only the magnitude of the coefficients but not the units which can result in a wrong model. It is important to note that the scaling affects the coefficients but not the t-statistic, F-statistic, p-values, R-squared, etc. There are 2 types of scaling:

Min-Max scaling- This brings the data between 0 and 1

Standardization scaling: This brings the data into a standard normal distribution with 0 mean and 1 standard deviation.

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

The VIF is the measure of multicollinearity in a regression model between the predictor variables. If the VIF is infinity which means there is a perfect correlation. The formula for VIF is

VIF=1/(1-R^2)

This happens when there is perfect multicollinearity in the regression model, which means that one or more of the independent variables can be expressed as a linear combination of the other independent variables. When there is perfect multicollinearity, the regression model cannot be estimated because the coefficients are not unique. In other words, the model cannot distinguish the effects of the collinear variables on the dependent variable. This results in an infinite VIF because the estimated variance of the regression coefficients becomes infinitely large.

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

A Q-Q plot are plots of two quantiles against each other. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. In linear regression, Q-Q plots are useful for checking the assumption of normality of the residuals, which is a crucial assumption for the validity of the regression model. The residuals are the differences between the actual values of the dependent variable and the predicted values of the dependent variable based on the regression model. If the residuals are normally distributed, then the Q-Q plot of the residuals will show a straight line. However, if the residuals are not normally distributed, then the Q-Q plot will show deviations from a straight line, indicating that the residuals are not normally distributed.