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?

- Bike demand in the fall is the highest followed by Summer , which means good season / weather is a driving factor.
- Bike demand takes a dip in spring.
- Bike demand in year 2019 is has boomed significantly when compared to 2018, which could mean Bike rentals are getting popular.
- Overall Bike demand is high during Fall in the months from May to October.
- Bike demand is high if weather is clear (good) or with mist cloudy(moderate) as temperature is optimal for Bike riding. While it is low when there is light rain or light snow.
- The demand of bike is almost similar throughout the weekdays.
- Bike demand doesn't change whether day is working day or not.

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

- Python libraries such as Pandas and scikit-learn have parameters built in to encode categorical data, one hot encoding, where a dummy variable is to be created for each discrete categorical variable for a feature.
- This can be done by using pandas.get_dummies() which will return dummy-coded data. Here we use parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.
- If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.
- **Reducing Redundancy:** By discarding one dummy variable, we effectively remove the redundancy in the model caused by perfect multicollinearity. This allows the regression model to estimate the coefficients of the remaining dummy variables accurately and interpret their effects on the outcome variable independently.

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

- atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

- Validated the assumptions of linear regression by checking Less Multi-collinearity between features (Low VIF) , error distribution of residuals and linear relationship between the dependent variable and a feature variable.

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

- Based on the final model, those will be : 'Temp', 'Year' (positively influencing) and 'snowy and rainy weather(weathersit_Bad)' (negatively influencing).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a form of supervised machine learning algorithm, where the target variable is continuous. It estimates the relationship between a target variable and one or more predictor variables.

This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is equation of linear regression:

$$y= m1x1 + m2x2 + m3x3 + ... + m(n)\ x(n) + c\ .$$

Where y is target variable and x1, x2 , x3 … xn are predictor variables . And we have two unknowns, m, and c, and we need to choose those values of m and c, which provides us with the minimum error.

There are two types of linear regression- simple linear regression and multiple linear regression. Simple linear regression is used when a single independent variable is used to predict the value of the target variable. Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable. A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give an accurate representation of two datasets being compared.

3. What is Pearson's R?

Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between "-1 "and "+1".

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

Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model. The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R-squared value is 1 for this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped to define a working model for regression.

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

The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.