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?

From the analysis of the categorical values from the datasheet, we could infer

- Demand for bike rentals increased from 2018 to 2019.
- Fall Season has the highest rental rates.
- September month have the highest demand for bike rentals.
- There is very less / no demand in severe weather conditions.

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

When we create dummy variables from categorical data we convert that into numbers. But this can cause multicollinearity and redundancy. To avoid that we drop one dummy variable from each category.

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

Temp and atemp have the highest correlation with the target variable.

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

We can check the VIF and the error distribution of the residuals. The residuals should a normal distribution with mean=0. We plot a distribution of the residuals to check if this is happening. If the curve is centred around zero, it means our model is aligned with the training data.

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

The top 3 features are temperature, humidity and year

General Subjective Questions.

1. Explain the linear regression algorithm in detail.

Linear regression is one of the simplest and most widely used statistical techniques in machine learning and data science. It's used to predict a continuous target variable based on one or more input features.

Linear regression assumes a linear relationship between the input features (independent variables) and the target (dependent variable). The goal is to find the best-fitting straight line that predicts the output values within a range.

Advantages:

1. Simple to understand and implement.
2. Requires less computational resources.
3. Provides a clear relationship between input and output variables.

Disadvantages:

1. Assumes a linear relationship between variables, which might not always hold true.
2. Can be sensitive to outliers.
3. Doesn't handle non-linear relationships well.

In practice, it's essential to check the assumptions of linear regression, visualize the data and residuals, and possibly consider more complex models if linear regression doesn't provide a satisfactory fit.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of four data sets that provide a useful caution against applying individual statistical methods to data without first graphing them. They have identical statistical properties, but look total different when graphed.

3. What is Pearson's R?

The relation coefficient doesn't just tell us whether two variables move in the same or opposite direction like the covariance, it also indicates how strong the relationship is and its value range from -1 to 1. R = covariance / (std.deviation of X * std.deviation of Y)

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

Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It helps in speeding up the calculations in an algorithm.

The difference between normalization and standardization is that while normalization helps you to scale down the feature between 0 to 1, where as standarized scaling helps to scale down the feature based on the standard normal distribution.

5. You might have observed that sometimes the value of VIF infinite. Why does this happen ? Variance inflation factor(VIF), (1 / (1 – R2)) is a measure of multicollinearity in the set of multiple regression variable. Multicollinearity occurs when the x variables are themselves related.The value of

VIF is infinite when there is a perfect correleation between two independent variables. We need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Q-Q plot is a probability plot, to visualize how close a sample distribution is to a normal distribution. This helps us to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. It also helps to find out if the error in the dataset are normal in nature or not.