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 analysis of the categorical variables from the dataset it is very clear that there is a certain rise in demands from the month number 4 to month number 9. Then there is a subtle decrease. It is also clear that the demands rise in clear weather conditions.

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

drop_first=True means while getting the dummy variables, drop the first column. This is used to avoid the problem of multicollinearity, which is a situation where two or more predictor variables in a regression model are highly correlated with each other. If T is the total number of dummy columns created, they all can be successfully represented by T-1 columns. Therefore, we use a **drop_first=True** to drop first column straight away.

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

If we look at the pair-plot among the numerical variables carefully, we will observe that 'atemp' has 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 do the Residual analysis to check our model's accuracy. And then we plot a Distplot to check the distribution to residuals. A normal distribution is a positive indicator for a our model.

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

1. 'atemp' - It positively affects the demand.
2. 'winter' - It positively affects the demand.
3. 'windspeed'- It negatively affects demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a powerful algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (also called independent variables). In other words, linear regression tries to find the best linear relationship between the predictor variables and the outcome variable.

Steps included in working of Linear Regression are:

1. Data Preparation- This includes EDA and scaling of numerical data.
2. Select the predictor variables- Choosing the predictor variables that you believe are most important for predicting the outcome variable.
3. Build the model- You can build a model by using Statsmodel or SKlearn.
4. Evaluate the model- Evaluation can be done by creating residuals and then calculating R2 value.
5. Use the model for prediction- Predictions can be performed on different dataset or Test dataset
6. Interpret the results- Interpreting the final results by also considering business aim.

2. Explain the Anscombe's quartet in detail.

The quartet consists of four datasets, each containing eleven (x,y) pairs of data, and have identical summary statistics such as mean, variance, and correlation coefficients. However, upon visual inspection, each dataset is unique and has a distinct pattern of relationships between the two variables.

The purpose of the quartet is to demonstrate the importance of visualizing data and not relying solely on summary statistics. It illustrates the limitations of summary statistics in describing a dataset and highlights the need for graphical exploration of data to gain a better understanding of its underlying patterns and relationships.

3. What is Pearson's R?

Pearson's r is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It is also known as the Pearson correlation coefficient

Pearson's r ranges from -1 to +1, where -1 indicates a perfect negative linear relationship, +1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship between the two variables.

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

Scaling is a data preprocessing technique used to transform variables so that they have a comparable range of values. In other words, scaling is the process of transforming data in such a way that it falls within a specific range or scale. The goal of scaling is to make sure that no variable dominates over others in the analysis and to make comparisons between variables meaningful.

Scaling is performed to address the issue of different scales and units of measurement in different variables. Scaling is used to standardize the data to a common scale, making the analysis more reliable.

There are two main types of scaling: normalized scaling and standardized scaling. Normalized scaling, also known as min-max scaling, is a technique that scales the data to a fixed range of values, typically between 0 and 1. Standardized scaling, also known as z-score scaling, is a technique that transforms the data so that it has a mean of zero and a standard deviation of one.

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

In some cases, the value of VIF may be infinite. This happens when one or more of the predictor variables in the regression model are perfectly collinear, which means that they are linearly dependent on each other. Perfect collinearity occurs when one variable is a perfect linear function of another variable or a combination of other variables in the model.

When perfect collinearity exists, the regression model cannot be estimated, and the VIF for the collinear variables becomes infinite. This is because the variance of the regression coefficient for the collinear variables cannot be calculated.

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

Q-Q stands for Qualtile-Quantile. A Q-Q plot is a graphical technique used to compare the distribution of a sample data to a theoretical distribution. The Q-Q plot is a scatter plot of the quantiles of the sample data against the corresponding quantiles of the theoretical distribution. A Q-Q plot can be used to assess whether the sample data follows a particular theoretical distribution, such as a normal distribution or an exponential distribution.

In linear regression, a Q-Q plot is often used to check whether the residuals are normally distributed. A Q-Q plot of the residuals is created by plotting the quantiles of the residuals against the expected quantiles of a normal distribution. If the residuals are normally distributed, the points on the Q-Q plot will lie approximately on a straight line. If the residuals are not normally distributed, the points on the Q-Q plot will deviate from a straight line.