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:
We can see we used several categorial variables in our model. Let's see the significance of these variables.

Year: So, we can see year is impacting the demand in a positive way. As we go to the next year the demand for the bikes increases.

Season: We can see in spring there were less people who were renting the bikes. In fall people rented bikes more.

Weather: on cloudy and stormy weather conditions people took less bikes which is obvious. On clear days the count was high.

Holiday: During holidays the demand for these rental bikes are low.

Month: September stood out with the highest rental counts.

Overall these all the factors have some sort of influence on the demand of the bikes be it positive or negative. All these insights are important for the business to take correct and informed decisions and to formulate their strategies accordingly.

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

Ans: The drop_first = True is used to drop the column in the dummy variable creation process because if we do not drop the columns this may lead to multicollinearity of the model. The first column is dropped as it is a standard column across the domain.

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

Ans: Temperature and a emp has the highest correlation with cnt variable.

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

Ans: The distribution of residuals terms should be normal, with the mean centred around 0. We can check by plotting the distplot of the residual terms.

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

Ans: The top 3 features that contribute significantly are temp which has coeff of 0.5029. That means for every unit increase in temp the demand increases by 0.5029 keeping rest of the variables constant.

The second feature is year which has a coeff of 0.2326 which means for every increase of years will lead to 0.2326 increase in the demand assuming rest of the variables are constant.

The third feature is weather situation light rain. This is getting decreased by 0.2989 for each unit increase in days saying the demand decreases.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a statistical model which is used to establish the connection between one or more independent variables and a dependant Variable. The relation is established on the idea that the variables have linear relationship, which states that variations in the independent variable or variables have a linear relationship with variations in the dependant variables.

Assumptions:

1. The relationship between the independent and the dependant variables should be linear.
2. The observations are independent of each other
3. The variance across the error should be constant.
4. The residuals are normally distributed
5. The independent variables should not be highly correlated with each other.

Types:

There are two types of linear regression one is Simple linear regression which has one dependant and one independent variable and the other is multiple linear regression which has more than one independent variable. Each model is represented by a straight-line equation

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p ,$$

Where y is the dependant variable and x1,x1,,xp are the independent variables and b1,b2 being the coefficients and b0 is he intercept.

The main idea of the linear regression is to find the best fit line through the points which minimises the square of the error terms.

Each coefficient gives the change in dependant variable when one unit of the independent variable is changed.

The model is evaluated by the r2 and the adjusted r2 which means the variance in the dependant variable that is explained by the model.

Residual analysis is conducted to check the assumptions are met for the model

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics such as mean, variance, correlation and regression lines but appear very different when visualized on plots. Each data set has different patterns with X and Y points plotted. By this we can say that we need to visually explore the datasets as statistics may not fully reveal the relation.

3. What is Pearson's R?

Ans: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Pearson's R helps to quantify the strength and direction of the association between two variables, making it a valuable tool in statistical analysis for understanding relationships in data. Below is a formula for calculating the Pearson correlation coefficient (r):

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

and standardized scaling?

Ans: Scaling is a preprocessing technique used in data analysis and machine learning to standardize or normalize the values in the independent variables. The main goal of scaling is to bring all the variables to a similar scale to prevent the model from giving huge differences in the coefficients of the dependant variables.

Scaling helps in increasing the model performance ensuring all the features are in a similar range

Normalized scaling usually scales the features between 0-1. Also, the outliers are handled into this 0-1 range.

Standardized scaling is done in such a way that the mean of the data is 0 and the standard deviation of 1. By standardizing the data shape is maintained unlike in normalizing.

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

Ans: We can sometime observe the VIF of some features o be infinite. This implies that this feature is highly explained by all other features. So, the information given by this feature is given by all other features. Hence, we can remove this feature from the model.

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

The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption. it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.