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? (3 marks)

Answer:

- Count of Rental Bikes is more during the fall season and very low in spring.
- Count of Rental Bikes is more during 2019 year
- As per monthly data, count of Rentals are varying more.
- There is no much change in the Rentals based on Weekdays and Working Days.
- People are preferring to rent a bike if the weather is clear and less preferring during rain.
- People are using Bikes during not a holiday.

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

Answer:

- We will use drop_first=True during dummy variable creation to avoid multicollinear affect.
- First column can be explained with the help of other created dummy variables. So we can avoid redundancy.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

Temperature

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

I have validated the Linear Regression Model based on P-Values in statsmodel Linear Regression model summary

- First I considered all 27 features and evaluated R2 Score on Test data
- Then I reduced to 14 features based on P-value > 0.1 and evaluated R2 on Test data.
- Then I reduced to 10 features based on P-value > 0.005 and evaluated R2 on Test data.

There is no much difference for R2 score between for 10 and 14 features model. So I selected the 10 features model.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Temperature, Weather as light rain, year as 2019

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

- Linear regression is a supervised machine learning algorithm that draws linear relationship between one or more independent features of data (X) and dependent feature (y) of data.
- It aims to predict the value of Y based on the values of X by fitting a linear equation to the data. This equation is represented as $Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betanXn$, where $\beta0$ is the intercept and $\betai$ (i = 1, 2, ..., n) are the regression coefficients.
- The goal is to find the values of $\betai$ that minimize the sum of squared errors between the predicted and actual values of Y. This process is known as least squares optimization.
- We follow gradient descent to find the coefficients.
- We use sklearn and statsmodel popular python libraries to build linear regression model

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet is a collection of four datasets that share the same mean, variance, linear regression line, and correlation coefficient, but have very different appearances when plotted. This illustrates the importance of visualizing data before relying on summary statistics alone.

Example: four groups of students with the same average score, standard deviation, and linear regression line for study hours and grades. However, when you plot the student data, you notice that one group has a curved relationship, another has a cluster around the regression line, another has an outlier, and the last has no visible relationship. This highlights the limitations of summary statistics and the need for visual inspection of data.

3. What is Pearson's R? (3 marks)

Answer:

- Pearson's R(Pearson correlation coefficient) is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.
- Its value ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. Positive correlation means that as one variable increases also other variable increases. Negative correlation means that as one variable increases, the other variable tends to decrease.
- Pearson's R is a widely used data analysis.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

- Scaling refers to the process of transforming the numerical features in a dataset to a similar range.
- It is performed to make sure that all features contribute equally to machine learning model and prevent features with larger magnitudes from dominating the learning process.
- Normalized scaling transforms features to have a mean of 0 and a standard deviation of 1, while standardized scaling transforms features to range between 0 and 1 or -1 and 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

Infinite VIF (Variance Inflation Factor) indicates that there is perfect multicollinearity present in the data which means one variable can be perfectly predicted by linear combination of the other variables. So variance of the coefficient estimate for that variable becomes infinitely large, leading to infinite VIF value.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

- Q-Q (quantile-quantile) plot, is graphical way to compare quantiles of two datasets to assess whether they come from the same distribution.
- In linear regression, Q-Q plots are specifically used to check if the standardized residuals follow a normal distribution, one of the key assumptions of linear regression. If the points on the Q-Q plot fall approximately along a straight line, it suggests that the residuals are normally distributed.
- Deviations from the line indicate departures from normality.