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?

Answer:

- Positive coefficients like year,temp,season_summer,season_winter and mnth_Sep indicative that increase in these fields contribute to increase in value count.
- Negative coefficients indicative that increase in these values will cause decrease in value of count.
- temp,season_winter,season_summer has large coefficients and indicative the importance of climate.

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

Answer:

drop_first=**True** helps in reducing the extra column created during dummy variable creation, hence it is very important to use.

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

Answer:

count of registered users

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

Answer:

- R Square value verification
- Checking VIF of model
- Residual Analysis of trained data

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

Answer:

Top 3 contributors are:

- Count of Registered user
- Temperature in Celsius
- Weather conditions

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is the simple form of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$y = c + mx$

or

$y = beta(zero) + beta(1)*x$

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet consists of four datasets having nearly identical simple statistical properties, yet appear very different when plotted or visualized.

Each dataset consists of particular number of (x,y) points (generally 11). They demonstrate both the importance of plotting data before analysis it and the effect of outliers on statistical properties.

3. What is Pearson's R?

Answer:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

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

Answer:

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

  Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. And helps to implement normalization in python.

  Standardization Scaling:
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (zero) and standard deviation (one).

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

Answer:
It indicates perfect correlation, when VIF = infinity, between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) = infinity.
To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:
Q-Q Plots (Quantile-Quantile plots) are graph of two quantiles against each other. A quantile is a fraction where certain values fall below that particular quantile.
The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.