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:

18 of the independent attributes are sufficient to determine accuracy. Following are the attributes –

- holiday
- windspeed
- registered
- mnth_dec
- mnth_feb
- mnth_jan
- mnth_jul
- mnth_mar
- mnth_may
- mnth_nov
- mnth_oct
- mnth_sept
- weekday_mon
- weekday_sat
- weekday_sun
- weekday_wed
- weathersit_bad
- weathersit_good

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

Answer: drop_first=True is necessary since the information is captured by other columns.

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

Answer: temp and atemp has higher correlation of 0.99 followed by registered and cnt with 0.95.

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

Answer:

1. VIF for the independent attributes are not infinity thus indicating no correlation between attributes.
2. Scatter plot of errors are evenly distributed.
3. Error Terms lie around mean.

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

Answer: Following are the list of top 3 features contributing significantly towards explaining the demand of the shared bikes

1. Holiday
2. Windspeed
3. Registered

General Subjective Questions

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

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

Answer: Anscombe's Quartet is a famous set of four datasets created by statistician Francis Anscombe in 1973 to demonstrate a crucial principle in data analysis: the importance of visualizing data rather than relying solely on summary statistics. The key insight is that very different datasets can share identical statistical properties, but, when plotted, shows diametrically different relationship.

It is used in Data Analysis, Statistical Communication & Quality Control.

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

Answer: Pearson's R, also known as Pearson's correlation coefficient or Pearson product-moment correlation, is a statistical measure that indicates the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, where:

- +1 indicates a perfect positive correlation
- -1 indicates a perfect negative correlation
- 0 indicates no linear correlation

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

Answer: Scaling is a data preprocessing technique used in machine learning and data analysis to adjust the range of features in a dataset. The goal is to ensure that all features contribute equally to the model's performance and to improve the algorithm's convergence during training.

Scaling is performed to

1. **Improves Model Performance** by having algorithms converge faster by ensuring that all features are on a similar scale.
2. **Prevents Dominance of one feature over another**: To prevent attributes with larger ranges dominate those with smaller ranges, potentially skewing the model.
3. **Enhances Interpretability**: Scaled features make the coefficients of linear models more interpretable.

- Normalization rescales the feature values to a range between 0 and 1 or -1 to 1. Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.

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

Answer:

VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity in the data. This can happen due to the following reasons - Duplicate columns in the dataset, Variables that are linear combinations of others, Including both a variable and its percentage/proportion & Dummy variable trap (including all dummy variables).

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

Answer: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. In linear regression, Q-Q plots are especially important for checking the normality assumption of residuals.