Assignment-based Subjective Questions

Question: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer:

1. Linear relationship between **cnt** and other independent variables (**atemp**, **temp**, **casual**, **registered**) .
2. Second year had more **cnt** (count of total rental bikes).
3. **cnt** tend to increase from 5th month to 10th month, it is sort of a normal distribution.
4. workingday 0 have more **cnt**
5. Non Holiday have more **cnt** (when holiday value is 0)

Question: Why is it important to use drop_first=True during dummy variable creation?
Answer: To reduce the **multicollinearity**

Question: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: **registered** column

Question: How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer:

1. Checking for VIF
    a. If there are predictors having the VIF > 5, there is a chance of collinearity.
2. Plotting residuals against the predicted value, it should not change.
3. Plotting Q-Q plot for the residuals, it should follow the normal distribution.

Question: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bike
Answer**:**

1. **Casual**
2. **registered**
3. **atemp**

General Subjective Questions

Question: Explain the linear regression algorithm in detail ?
Answer:

In words

1. You start by guessing a line. It could be any line, but you'll eventually adjust it to make it better.
2. Then, you measure how far each data point is from the line. These distances are called residuals.
3. You square each of these distances (to make negative distances positive and to give more weight to larger distances), and then add them all up.
4. Your goal is to find the line that makes this sum of squared distances as small as possible. This line is called the "best-fit" line or the "regression" line.
5. Once you have this best-fit line, you can use it to make predictions. If you have a new house size, you can plug it into the equation of the line to estimate its price.

How we can achieve using sklearn, statsmodel and other libraries.

1. Get the dataset, and find the relationship of the variables, with target variable
2. Remove the very least correlated variables.
3. Test train split
4. Transform the variables
    a. Encode the categorical variables.
    b. Min Max scaling of the train split.
5. Create linear regression model
6. Import RFE from **sklearn**, add the model to it, train data
7. Select top K features, having low p value, and low VIF
8. Create linear regression model
9. Fit X_train
10. Checkout the model summary
    a. Look at the p-value, f-statistics
    b. Calculate the VIF of the predictors those having the high p-value.
11. Predict the **y_test** and calculate the **r2_score**

Question: Explain the Anscombe’s quartet in detail.
Answer: Anscombe's quartet is a collection of four datasets that have nearly identical statistical properties (e.g., mean, variance, correlation), but exhibit vastly different

patterns when graphed. This example highlights the importance of visualizing data and the limitations of relying solely on summary statistics. Despite their similar statistical summaries, the datasets in Anscombe's quartet demonstrate the need for

Question: What is Pearson's R?
Answer:

1. R=1, indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally.
2. R=−1, indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
3. R=0, indicates no linear relationship between the variables.

Question: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer: Taking reference of the min-max scaling and standardization

1. Min-Max scaling, where values are scaled to a specific range (e.g. 0, 1), and standardization, where values are scaled to have a mean of 0 and a standard deviation of 1.
2. Overall, feature scaling is an important preprocessing step in machine learning to ensure that all features contribute equally to the model's performance and to prevent issues such as features dominating due to their larger scale.

Question: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: It indicates perfect multicollinearity between the independent variables in the regression mode

Question: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: Q-Q plot (Quantile-Quantile plot) is a data visualization tool, which is used to check whether a set of data follows a particular probability distribution (e.g. Normal distribution). It compares the quantiles.