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 1:

Season: There’s a noticeable pattern with the number of rentals. Rentals are higher in summer and fall, while they are lower in spring.

Year: The data shows more rentals in 2019 compared to 2018. This could be due to increased awareness of the service or changes like the impact of COVID-19.

Month: Rentals vary across the months, generally increasing during mid-year (May to October) when the weather is more pleasant, and decreasing in other months.

Holiday: There are fewer rentals on holidays. This suggests that people prefer to stay home on holidays, whereas they rent more on regular days, likely to commute to work.

Weather: Clear weather leads to higher rentals. There are fewer rentals when it’s lightly snowing, and no rentals during heavy snow, indicating that weather conditions significantly impact rental activity.

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

Answer 2:

When you create dummy variables, using drop_first=True is important because it helps prevent "multicollinearity." Multicollinearity happens when you have too many variables that are very similar to each other, which can mess up your analysis.

For example, if you have a categorical variable like "color" with values "red," "blue," and "green," creating dummy variables without drop_first=True would give you three columns: "color_red," "color_blue," and "color_green." This can cause problems because these three columns add up to 1, meaning one of them is always predictable from the others.

By using drop_first=True, you drop one of the dummy columns (like "color_red"). This way, you only have "color_blue" and "color_green," and you avoid the multicollinearity issue. It makes your model simpler and easier to interpret.

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

with the target variable? (1 mark)

Answer 3:

Looking at the pair-plot temp and atemp have highest correlation. It has positive relation with the target variable

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

Answer 4:

- **No Multicollinearity**: I checked the Variance Inflation Factor (VIF) for each independent variable to see if there was any multicollinearity

- **Linearity Between X and Y**: I checked if there is a linear relationship between the independent variables and the dependent variable by plotting residuals (the differences between actual and predicted values) against the predicted values. If the plot showed no clear pattern and looked random, it indicated that the linearity assumption was met.

- **Normality of Residuals**: I plotted a histogram of the residuals to see if they followed a normal distribution, it indicated that this assumption was satisfied.

- **Homoscedasticity**: I looked at the residual plot (residuals vs. predicted values) again to check for homoscedasticity. If the spread of the residuals looked consistent and didn't show any obvious patterns like funnel shapes it suggested that the homoscedasticity assumption was valid.

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 5:

Two features significantly impacting the demand for shared bikes are temperature (Temp) and weather condition with light snow (weathersit_lightSnow). Here's how each feature affects the bike demand:

- Temperature (Temp) has a coefficient of +4095.62. This means that for every 1 unit increase in temperature, the bike demand increases by 4095 units, assuming all other factors remain constant.
- The weather condition weathersit_lightSnow has a coefficient of -2634.20. If there is light snow (weathersit_lightSnow = 1), keeping all other variables constant, the bike demand decreases by 2634 units.

Therefore, these coefficients provide insights into how changes in temperature and weather conditions impact the demand for shared bikes, helping to understand the variables' contributions in the model.

General Subjective Questions

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

Answer 1:

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. Here’s how it works in detail:

1. **Basic Concept**:
   - Linear regression finds the best-fitting line through the data points.
   - The line is represented by the equation $y=mx+ c$ for simple linear regression and $Y= m1x1+m2x2\ldots\ldots+mnxn+c$. where x1..xn are features of the data set and Y is dependent variable. m1 …mn are called coefficients c is the intercept
2. **Assumptions**:
   - **Linearity**: The relationship between the dependent and independent variables is linear.
   - **Independence**: The observations are independent of each other.
   - **Homoscedasticity**: Constant variance for the residuals for all levels of the independent variables.
   - **Normality**: The residuals (differences between observed and predicted values) are normally distributed.
3. **Steps in Linear Regression**:
   - **Collect Data**: Gather data with the dependent variable and one or more independent variables.
   - **Split Data**: Split the data into training and testing sets.
   - **Train the Model**: Use the training set to fit the linear regression model by finding the best coefficients that minimize the sum of squared errors.
   - **Make Predictions**: Use the model to predict the dependent variable values in the test set.
   - **Evaluate the Model**: Check how well the model performs using metrics like R-squared and Mean Squared Error (MSE).

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

Answer 2:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. It was created by the statistician Francis Anscombe to illustrate the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Here are the details of each dataset in the quartet:

1. **Dataset 1**:
   - It looks like a straight-line relationship between xxx and yyy.
   - The data points are distributed closely along a straight line.
2. **Dataset 2**:
   - This also shows a linear relationship but with a curve.
   - The data points form a parabolic shape, indicating a non-linear relationship.
3. **Dataset 3**:
   - It includes an outlier that significantly affects the linear relationship.
   - Most of the data points fall on a straight line, but one outlier pulls the regression line away from the main cluster of points.
4. **Dataset 4**:

  - This dataset has a single high-leverage point.
  - The data points are mostly scattered, with one point that strongly influences the slope of the regression line.

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

Answer :

Pearson's RRR, or Pearson correlation coefficient, is a measure of the linear correlation between two variables, typically denoted as X and Y. It quantifies the strength and direction of the linear relationship between the variables. Here's how it is defined and used:

- **Definition**: Pearson's R is calculated as the covariance of X and Y divided by the product of their standard deviations. Mathematically, it is represented as:
- Pearson's R ranges from -1 to +1:
- The magnitude of RRR indicates the strength of the linear relationship:
  - close to 1 indicates a strong linear relationship.
  - close to 0 indicates a weak linear relationship.

Pearson's RRR is widely used in statistics and data analysis to assess relationships between variables, but it assumes that the variables are normally distributed and that there is a linear relationship between them.

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

and standardized scaling?

Answer 4:

Scaling refers to the process of transforming numerical columns to a standard scale, typically to have a consistent range. It is performed to ensure that variables with different scales and units contribute equally to the analysis and model training process.

Reasons for Scaling:

- **Equal Contribution**: Variables with larger ranges can dominate those with smaller ranges, affecting model performance.
- **Algorithm Requirements**: Many machine learning algorithms perform better or converge faster when features are on a similar scale.

Difference between Normalized and Standardized Scaling:

- **Normalized Scaling**: Scales the data to a [0, 1] range. It's useful when the distribution of data is not Gaussian (non-normal distribution).

- **Standardized Scaling**: Scales the data to have a mean of 0 and a standard deviation of 1. It assumes a Gaussian distribution and is often used when features have different units or scales

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

Answer 5:

The occurrence of an infinite VIF (Variance Inflation Factor) typically happens when there is perfect multicollinearity among the independent variables in the dataset. Perfect multicollinearity means that one or more variables can be exactly predicted using a linear combination of the others. This leads to a situation where the VIF calculation involves division by zero or very small numbers, resulting in an infinite VIF value. In the given data set when I created dummy variable to Year column the year column has infinite VIF with this dummy variable

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

Answer 6:

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess whether a given dataset follows a particular distribution, such as a normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution (like a normal distribution).

Use and Importance in Linear Regression:

- **Distribution Assessment**: It helps check if the in linear regression follow a normal distribution. Which is one the assumption for linear regression
- **Identifying Outliers**: Q-Q plots can also reveal outliers or deviations from expected patterns in the data distribution.These tools are fundamental in ensuring that linear regression models are appropriate for the data at hand, facilitating accurate interpretation and reliable conclusions.