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)

Ans: All of the categorical variables have an impact on the dependent variable, here are the conclusions for each:

1. Season: Count increases in fall and summer while it decreases in winter and its the least in spring.
2. Year: Count is increased in 2019
3. Month: This follows the same trend as that of seasons.
4. Holiday: Count is more during non-holidays, this has a negative impact.
5. Day of the Week: This is the only categorical variable which has almost negligible impact on the dependent variable.
6. Working Day: This has minor impact, count is a bit lower on the non working days.
7. Weather: This also has an impact, during snow, the count decreases while its the max when the weather is clear.

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

Ans: If a categorical variable has k levels, then we can easily infer all levels using (k-1) binary columns.Its important so that we can avoid having a variable with a high VIF and correlation with other vars and rather than removing it in the end using either RFE or manually, we should remove it before hand.

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

Ans: As per the pair-plot, atemp(0.65) or temp(0.64) are highly correlated with the target var cnt.

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

Ans: For verification of the assumptions, I calculated the Residuals by comparing y_pred with y_train and plotted them. I got a **normal distribution** for these error terms. Apart from this, the adjusted R squared is 0.829 and the p score is 0 or close to 0 for all the independent variables.

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

Ans: Here are the top 3 features along with their coefficients:

1. Temp = 0.4949
2. Snow = -0.249, ie. negative correlation.
3. Year = 0.231039, this is mainly because of better business in 2019.

General Subjective Questions

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

Ans: Linear regression is a fundamental statistical and machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It is often employed for predicting a continuous numeric outcome. Here's a detailed explanation of the linear regression algorithm:

1. **Objective**:
Linear regression aims to find the linear relationship between the independent variables and the dependent variable. It seeks to create a linear equation that best fits the data, allowing us to make predictions based on this equation.

2**. Linear Model**:
The core of linear regression is a linear model, expressed as:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$

Y is the dependent variable.
$X_1, X_2, ..., X_n$ are the independent variables.
$\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the coefficients (parameters) that we want to estimate.
ε represents the error term, which accounts for the variability not explained by the model.

3. **Estimating Coefficients:**
The goal is to determine the coefficients ($\beta_0, \beta_1, \beta_2, ...$) that minimize the sum of squared differences between the predicted values ($\hat{Y}$) and the actual values (Y) in the training dataset. This is usually done using a method like Ordinary Least Squares (OLS).

4. **Assumptions**:
Linear regression relies on several key assumptions, including:

Linearity: The relationship between independent and dependent variables is assumed to be linear.
Independence: Errors (ε) are independent of each other.
Homoscedasticity: The variance of errors is constant across all values of the independent variables.
Normality: The errors follow a normal distribution.

5. **Types of Linear Regression**:

Simple Linear Regression: Involves a single independent variable.

Multiple Linear Regression: Uses multiple independent variables.

6. **Model Evaluation:**

Various metrics like Mean Squared Error (MSE), R-squared (R^2), and adjusted R-squared are used to evaluate the performance of the linear regression model. These metrics help assess the goodness of fit and predictive accuracy.

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

Ans: Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they exhibit substantially different properties when graphed or analyzed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and not relying solely on summary statistics. Here's a more detailed explanation:

1. The Four Datasets:

Anscombe's Quartet consists of four data sets, each containing 11 data points. In each set, there are two variables: X and Y.

2. Deceptively Similar Summary Statistics:

The striking aspect of Anscombe's Quartet is that when you calculate summary statistics such as mean, variance, and correlation for each dataset, they are remarkably similar. This might lead one to believe that the datasets are similar in nature.

3. Key Differences Revealed through Visualization:

Despite their similar summary statistics, the datasets are very different when visualized. The quartet demonstrates that relying solely on statistics can be misleading. Here are the key characteristics of each dataset:

- **Dataset I:** It forms a simple linear relationship.
- **Dataset II:** It also shows a linear relationship but with an outlier that affects the regression line.
- **Dataset III:** This dataset has a non-linear relationship.
- **Dataset IV:** It contains an outlier that strongly influences the correlation and regression line.

4. Implications:

Anscombe's Quartet underscores the importance of data visualization in understanding data. Even when summary statistics appear to be similar, the underlying data can have significantly different patterns. It emphasizes that graphical analysis can reveal relationships, outliers, and nuances that might not be apparent from mere numerical summaries.

In summary, Anscombe's Quartet serves as a powerful reminder that statistics alone can be deceptive, and the visual exploration of data is essential to gain a comprehensive understanding of its characteristics. It is often used in statistical education to highlight the value of data visualization and the limitations of relying solely on summary statistics.

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

Ans: Pearson's correlation coefficient, often denoted as "Pearson's R," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1 and is widely used to assess the degree of association between two variables. Here's a brief explanation:

1. Calculation: Pearson's R is computed by taking the covariance of the two variables and dividing it by the product of their standard deviations. The formula for Pearson's R is:

$$R = (\Sigma[(X - \bar{X})(Y - \bar{Y})]) / [n * \sigma X * \sigma Y]$$

- **X** and **Y** are the data points.
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y.
- σX and σY are the standard deviations of X and Y.
- **n** is the number of data points.

2. Interpretation: Pearson's R produces a value between -1 and 1:
- R = 1 indicates a perfect positive linear relationship.
- R = -1 indicates a perfect negative linear relationship.
- R ≈ 0 indicates a weak or no linear relationship.
- The sign of R (positive or negative) indicates the direction of the relationship.

3. Use: Pearson's R is employed in various fields, including statistics, social sciences, economics, and data analysis. It helps determine how strongly two variables are associated and in which direction they tend to move together or apart. It is a valuable tool for data exploration, hypothesis testing, and making predictions.

In summary, Pearson's R is a statistical measure used to assess the strength and direction of the linear relationship between two continuous variables. It is a fundamental tool for understanding and quantifying associations in data analysis.

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

Ans: Scaling is the process of adjusting numerical features in a dataset to a common scale. Scaling is done to ensure fairness among features and improve the performance of various algorithms. It is done so that **model coefficients can be compared**.

Normalized Scaling brings values within the range of [0, 1], preserving the original distribution. It's used when there are no significant outliers.

Standardized Scaling centers data around a mean of 0 and a standard deviation of 1. It's robust to outliers and appropriate when data distribution isn't uniform.

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

Ans: Sometimes, the value of VIF (Variance Inflation Factor) can become infinite. This occurs when one of the independent variables in a regression analysis is a perfect linear combination of one or more other independent variables. When variables are perfectly correlated, the mathematical calculation of VIF involves dividing by zero, resulting in an infinite VIF value. This is problematic because a high VIF suggests a high degree of multicollinearity (correlation between independent variables), which can undermine the reliability of regression results and make it difficult to discern the individual contributions of correlated variables to the dependent variable.

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

Ans: A Q-Q plot is a scatterplot used in statistics to check if a dataset follows a specific theoretical distribution, like the normal distribution. In linear regression, it's crucial for evaluating the normality of residuals (differences between observed and predicted values). A straight Q-Q plot line indicates well-behaved residuals, confirming that key assumptions of linear regression are met. Deviations may prompt further investigation or the need for alternative regression methods.