Assignment-based Subjective Questions

Q1. 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.

- **Season** :Fall season has the most bookings.
- **Weather situation** : Clear weather has more bike rentals.
- **Weather situation** :The no. of booking goes down when the weather situation is Light snow rain
- **Year**: For year 2019 their is a significant increase in the booking as compared to 2018
- **Holiday** :The median value of the rentals are significantly higher when there is no holiday.
- **Month**:Mid year specially may, june, july, aug, sep and oct experience maximum riding.
- **Working day** :Booking seems almost equal irrespective of the working day or not.
- **Holiday:** When it's holiday, booking seems less in number which seems reasonable. Maybe people want to stay at home and relax.
- **Weekday** : Wed ,Thu, Fir, Sat have more bookings as compared to the start of the week.

Q1. 2. Why is it important to use drop_first=True during dummy variable creation?

Ans.

It is important to use drop_first:

- to avoid multicollinearity, and capture the necessary information for regression analysis.
- The dummy variable trap, also known as the "dummy variable multicollinearity" problem, occurs when we include all levels of a categorical variable as dummy variables in a regression model. Including all levels introduces perfect multicollinearity because the sum of the dummy variables for a particular observation will always be 1.
- the no. of variables decreases thereby ensuring efficient parameter estimation
- there is no loss of information and the information can be interpreted easily

---

Q1. 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans.

Atemp and temp has the highest correlation with the target variable ("cnt")

Q1. 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans.

We did the residual analysis of the error terms :

- Normality -> plotted the histogram of the error terms and saw it was normally distributed (approximately)
  Independence -> There should be no systematic patterns or correlations in the residuals. We check this visually by plotting the residuals against the predicted values or the independent variables.
  Homoscedasticity -> The residuals should have constant variance across all levels of the independent variables. We assess this by plotting the residuals against the predicted values or the independent variables.
- Calculated the VIF and checked for multicollinearity . Since no significant amount of multicollinearity was present in model 6 hence its was fairly good to go.

Q 1. 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans.

The top 3 features contributing significantly towards explaining the demand of the shared bikes :

1. Temp : temperature
2. Yr : year
3. Winter

General Subjective Questions.

Q2 .1. Explain the linear regression algorithm in detail.

Ans.

Linear regression is a algorithm used for predicting a continuous outcome variable based on one or more independent variables. It assumes a linear relationship between the independent variables and the dependent variable.

The linear regression algorithm aims to find the best-fit line that represents the relationship between the independent variables and the dependent variable.

Assumptions:

- Linear relationship: The relationship between the independent variables and the dependent variable is assumed to be linear.
- Independence: The observations are assumed to be independent of each other.
- Homoscedasticity: The variance of the errors (residuals) is constant across all levels of the independent variables.
- Normality: The errors are assumed to follow a normal distribution.
- No multicollinearity: The independent variables should not be highly correlated with each other.

Equations:

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

The most common method for estimating the coefficients is called Ordinary Least Squares (OLS), which finds the values of the coefficients that minimize the sum of squared residuals

Linear regression allows us to interpret the coefficients to understand the impact of each independent variable on the dependent variable. The coefficients (β_1, β_2, ..., β_n) represent the change in the dependent variable associated with a one-unit change in the corresponding independent variable, assuming all other variables remain constant.

Q2 .2. Explain the Anscombe’s quartet in detail.

Ans.

Anscombe's quartet is a valuable reminder of the importance of visualizing data and understanding its patterns and relationships beyond summary statistics. It highlights the potential pitfalls of relying solely on statistical measures and encourages data analysts to explore their data graphically for a more comprehensive understanding.

Anscombe's quartet is a set of four datasets that have identical statistical properties, including means, variances, correlations, and linear regression coefficients. They were created by the statistician Francis Anscombe in 1973 to illustrate the importance of data visualization and the limitations of relying solely on summary statistics. Despite the datasets having similar statistical summaries, they exhibit distinct patterns and relationships when graphically represented. Here's a detailed explanation of Anscombe's quartet.

Key Insights:

- Anscombe's quartet emphasizes the limitations of relying solely on summary statistics, as the datasets can have significantly different patterns and relationships despite similar statistical properties.
- It underscores the importance of data visualization in understanding and interpreting data accurately.
- The quartet showcases the impact of outliers on statistical analysis, regression lines, and correlation coefficients.
- It serves as a reminder that summary statistics can provide an incomplete picture and that graphical exploration can reveal nuances and insights not captured by numbers alone.

Q.2.3 What is Pearson's R?

Ans.

Pearson's correlation coefficient, commonly denoted as Pearson's R is a statistical measure that quantifies the linear relationship between two continuous variables used to assess the strength and direction of the linear association between two variables.

- Pearson's R ranges between -1 and 1.
- A value of 1 indicates a perfect positive linear relationship, where the variables increase proportionally.
- A value of -1 indicates a perfect negative linear relationship, where one variable decreases proportionally as the other increases.
- A value of 0 indicates no linear relationship between the variables

The sign of Pearson's R indicates the direction of the linear relationship. A positive R suggests a positive association, where both variables tend to increase together, while a negative R suggests a negative association, where one variable tends to decrease as the other increases.

Q 2.4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans.

Scaling, in the context of data preprocessing, refers to the process of transforming the numerical features of a dataset to a consistent scale. It involves adjusting the values of the variables to a specific range or distribution. Scaling is performed to ensure that all features have a comparable influence on the analysis or model and to avoid issues caused by different scales or units.

Scaling is performed to:

- Avoid bias: Variables with larger scales may dominate or bias the analysis or model more than those with smaller scales, leading to misleading results.
- Enable comparison: Scaling allows for meaningful comparisons between variables by putting them on a consistent scale.
- Facilitate model convergence: Scaling can help certain algorithms converge more quickly and improve their performance.

- Improve interpretation: Scaling can make the coefficients or weights in a model more interpretable and comparable.

Normalized Scaling:
Normalized scaling, also known as feature scaling or min-max scaling, transforms the values of the variables to a specified range, usually between 0 and 1.

The formula for normalized scaling is:

X_scaled = (X - X_min) / (X_max - X_min)

Normalized scaling preserves the shape of the distribution, but it may be sensitive to outliers.

Standardized Scaling:
Standardized scaling, also known as z-score normalization or standardization, transforms the values of the variables to have a mean of 0 and a standard deviation of 1

- The formula for standardized scaling is:
    - X_standardized = (X - μ) / σ
    - X_standardized represents the standardized value, X is the original value of the variable, μ is the mean of the variable, and σ is the standard deviation of the variable.
- Standardized scaling centers the distribution around 0 and adjusts the spread of the values. It is less sensitive to outliers compared to normalized scaling.
- Standardized scaling is useful when the distribution of the variable is not known or when the algorithm or model assumes normally distributed variables.

Q.2.5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans.

The occurrence of infinite Variance Inflation Factor (VIF) values typically indicates a problem known as perfect multicollinearity. Perfect multicollinearity arises when one or more independent variables in a regression model can be perfectly predicted by a linear combination of other independent variables. In such cases, the VIF calculation breaks down, resulting in an infinite value.

- For example, if you have two independent variables, X1 and X2, and X2 can be expressed as a linear combination of X1, such as X2 = 2 * X1, perfect multicollinearity exists.
- In this case, the regression model cannot distinguish the individual effects of X1 and X2, leading to an infinite number of solutions.

When perfect multicollinearity exists, the regression model fails to calculate the VIF correctly.

Q.2.6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans.

- A Q-Q plot compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically the standard normal distribution (mean = 0, standard deviation = 1).
- The Q-Q plot is constructed by plotting the observed quantiles on the x-axis and the corresponding theoretical quantiles on the y-axis.
- If the data follows the theoretical distribution closely, the points in the Q-Q plot will fall approximately on a straight line.

2. Purpose and Use:

- The primary purpose of a Q-Q plot is to assess whether the distribution of the observed data deviates from a specific theoretical distribution.
- In linear regression, Q-Q plots are commonly used to examine the normality assumption of the residuals.
- The residuals are the differences between the observed dependent variable values and the predicted values from the regression model.
- If the residuals are normally distributed, the Q-Q plot of the residuals against the standard normal distribution should show a roughly straight line.
- Deviations from a straight line in the Q-Q plot can indicate departures from normality, suggesting potential issues with the linear regression assumptions.

Interpretation:

- In a Q-Q plot, if the points closely follow a straight line, it suggests that the data distribution is similar to the theoretical distribution being compared (e.g., the residuals are normally distributed).
- If the points deviate from the straight line, it indicates departures from the assumed distribution

The normality assumption of the residuals is crucial in linear regression as it affects the validity of statistical inference, hypothesis testing, and confidence intervals.
Departures from normality can lead to biased parameter estimates, incorrect p-values, and unreliable predictions.
