Assignment-based Subjective Questions

Submitted by Battini_lokesh

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

I have analyzed the relationship between the categorical variables and the target variable using boxplots and derived the following insights:

- Season: Fall (Season 3) shows the highest demand for rental bikes.
- Year: There is a noticeable increase in demand in 2019.
- Month: Demand consistently grows each month until June, with September experiencing the highest demand. After September, the demand decreases.
- Holiday: When it is not a holiday, the number of bookings tends to be lower. This seems reasonable as, on holidays, people may prefer to stay at home and spend time with family.
- Weekday: The data does not provide a clear trend for demand based on weekdays.
- Weather Condition: The "Clear" weather condition has the highest demand.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

"drop_first=True" is important to use as it helps reduce the extra column created during dummy variable creation. This reduces the correlation among dummy variables and avoids redundancy in the dataset.

If we do not drop one of the dummy variables from a categorical variable, it leads to a situation where one variable can be perfectly predicted from the others. This redundancy, combined with the constant term (intercept) in the model, causes a multicollinearity issue that can negatively impact the model's stability and interpretability.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Based on the pair-plot of numerical variables, the feature "temp" shows the highest correlation with the target variable "cnt". It exhibits a strong linear relationship with the target.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I validated the assumptions of Linear Regression by checking the following:

- Normality of Error Terms: Verified that the residuals are normally distributed with a mean of 0.
- Homoscedasticity: Ensured that the error terms do not exhibit any specific pattern (constant variance across predictions).
- Multicollinearity: Used Variance Inflation Factor (VIF) to identify and address multicollinearity issues.
- Linearity: Confirmed a linear relationship between the predictors and the target variable.
- Overfitting Check: Compared the R^2 and Adjusted R^2 values to ensure the model is not overfitting.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top 3 variables with the highest correlation values are:

1. `temp` (0.4789)
2. `year` (0.2342)
3. `winter` (0.0968)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to the observed data. The goal is to find the coefficients (weights) that minimize the difference between the predicted and actual values. In simple linear regression, the equation is of the form:

$$Y=\beta 0+\beta 1X+\epsilon$$

Where **Y** is the target, **X** is the predictor, **β0** is the intercept, and **β1** is the coefficient. The error term **∈** represents the difference between predicted and actual values. In multiple linear regression, there are multiple predictors.

The process of training a linear regression model involves finding the best-fitting coefficients that minimize the Mean Squared Error (MSE), which measures the average squared differences between actual and predicted values. Optimization methods like Gradient Descent or the Normal Equation can be used to find these optimal coefficients.

Once the model is trained, it can make predictions by applying the linear equation to new data. The performance of the model is evaluated using metrics like R-squared (R^2), which measures how well the model explains the variance in the target variable, and Mean Absolute Error (MAE), which measures the average error.

Linear regression is simple and easy to interpret, but it assumes a linear relationship between the variables, which may not always hold true in real-world data. It is also sensitive to outliers and relies on assumptions like homoscedasticity (constant variance of errors) and normality of residuals. Extensions of linear regression, such as Ridge and Lasso Regression, help address some of its limitations by adding regularization to avoid overfitting.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>
Anscombe's Quartet is a set of four datasets created by statistician Francis Anscombe in 1973 to show the importance of visualizing data. Although the datasets have nearly identical statistical properties, they reveal very different patterns when plotted.

Each dataset in Anscombe's Quartet has:
- The same mean for both X and Y.
- Identical variances for both X and Y.
- The same correlation coefficient of 0.82 between X and Y.
- The same linear regression equation.

Despite these similarities, when you plot the datasets, the differences become clear:

1. Dataset 1 shows a perfect linear relationship between X and Y. The data points lie along a straight line, and the regression model fits the data well.

2. Dataset 2 has a non-linear relationship. Although the summary statistics suggest a strong linear correlation, the data follows a curve rather than a straight line, making the linear regression model unsuitable.

3. Dataset 3 contains an outlier that skews the regression line. The summary statistics still suggest a linear relationship, but the outlier distorts the data and the regression results.

4. Dataset 4 has most points aligned on a vertical line, with one outlier far from the cluster. The linear regression model doesn't fit well, and the data doesn't suggest a meaningful relationship despite similar statistical summaries.

The key takeaway from Anscombe's Quartet is that statistical summaries like mean, variance, and correlation can be misleading. Visualization is crucial for understanding the true relationships in data, as outliers and non-linearities can significantly impact statistical analyses.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's **r** is a numerical summary of the strength and direction of the linear relationship between two variables. If the variables tend to increase and decrease together in a consistent way, the correlation coefficient will be positive. On the other hand, if one variable increase while the other decreases, the correlation will be negative.

Pearson's **r** always lies between -1 and 1. A value of **r** = 1 indicates a perfect positive linear relationship, where both variables increase together. A value of **r** = -1 indicates a perfect negative linear relationship, where one variable increase while the other decreases.

A positive correlation means that as one variable increases, the other also increases, or as one decreases, the other decreases. In contrast, a negative correlation means that as one variable increases, the other decreases, and vice versa.

In summary, Pearson's **r** measures the strength and direction of the linear relationship between two variables, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a method used to normalize the range of independent variables, ensuring that all features are on the same scale, particularly in regression. If scaling is not performed, the algorithm may incorrectly consider larger values as more important than smaller ones. This can lead to biased predictions.

For example, if the weight of a device is 500 grams and the weight of another is 5 kg, the algorithm might treat 500 grams as a larger value, which is incorrect. Machine learning algorithms work with numbers, not units, so scaling is crucial for accurate predictions.

Scaling affects only the coefficients in a regression model and does not impact other statistical parameters like t-statistics, p-values, or R-squared.

There are two common methods for scaling:
1. Normalization: Scales data to a range between 0 and 1.
2. Standardization: Transforms data to have a mean of 0 and a standard deviation of 1.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The value of Variance Inflation Factor (VIF) can become infinite when there is a perfect linear relationship between one independent variable and other independent variables in the model. This occurs when one variable can be exactly predicted by a linear combination of the others. In such cases, the variance of the estimated coefficients for that variable becomes infinitely large, leading to an infinite VIF.

To explain further:

- VIF = Infinity: If a variable is perfectly correlated with other independent variables (i.e., there is perfect multicollinearity), the VIF becomes infinite because the variance of the coefficient for that variable cannot be estimated reliably.
- VIF = 1.0: If the independent variables are orthogonal (uncorrelated) to each other, then each variable has a VIF of 1.0, indicating no multicollinearity.

Thus, infinite VIF signifies perfect multicollinearity, making it impossible to separate the individual effects of the correlated variables in the model.

Question 11. What is a **Q-Q** plot? Explain the use and importance of a **Q-Q** plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A **Q-Q** (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, typically the normal distribution. It compares the quantiles of the observed data with the quantiles of the theoretical distribution. If the data points lie approximately along a straight line, it suggests that the data follows the expected distribution. If the points deviate significantly from the line, it indicates that the data does not follow the specified distribution.

Use and Importance of a **Q-Q** Plot in Linear Regression:

1. Normality of Residuals: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A **Q-Q** plot helps in visually checking this assumption. If the residuals follow a normal distribution, the points in the **Q-Q** plot will lie close to the straight line.

2. Model Diagnostics: A **Q-Q** plot is useful for diagnosing model fit. If the plot shows deviations from the line, it may indicate problems like non-normal residuals, suggesting that the model might not be appropriate or that there may be issues like outliers, skewness, or heteroscedasticity.

3. Assumption Checking: Linear regression assumes that errors are normally distributed. The **Q-Q** plot provides a simple and effective way to verify this assumption. If the residuals deviate significantly from the line, transformations or different modeling techniques might be needed.

In summary, a **Q-Q** plot is a valuable diagnostic tool in linear regression to ensure the normality of residuals and to check the validity of model assumptions.