ss

Assignment-based Subjective Questions

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)

1. In the season of fall, the number of bike rentals were the highest as compared to other season's.
2. In the year 2019, bike rentals reached their peak.
3. Months of June and September saw the highest number of bike rentals.
4. As per the data, the highest number of bike rentals were done during working days.
5. There is an increasing trend in bike rentals from Sunday to Friday.
6. Bikes were rented more frequently on working days than on holidays or weekends.
7. Clear weather conditions were the most favorable for bike rentals.

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)

- Using drop_first=True is useful for creating k – 1 dummy variables out of k categories. This helps avoid multicollinearity, which occurs when a variable is predicted by other variables.
- In the bike-sharing dataset, we have a column season that contains four categories: spring, summer, fall, and winter. By dropping the spring category (using drop_first=True), we represent it implicitly with all 0's, while the other seasons are represented by 1's in their respective columns.

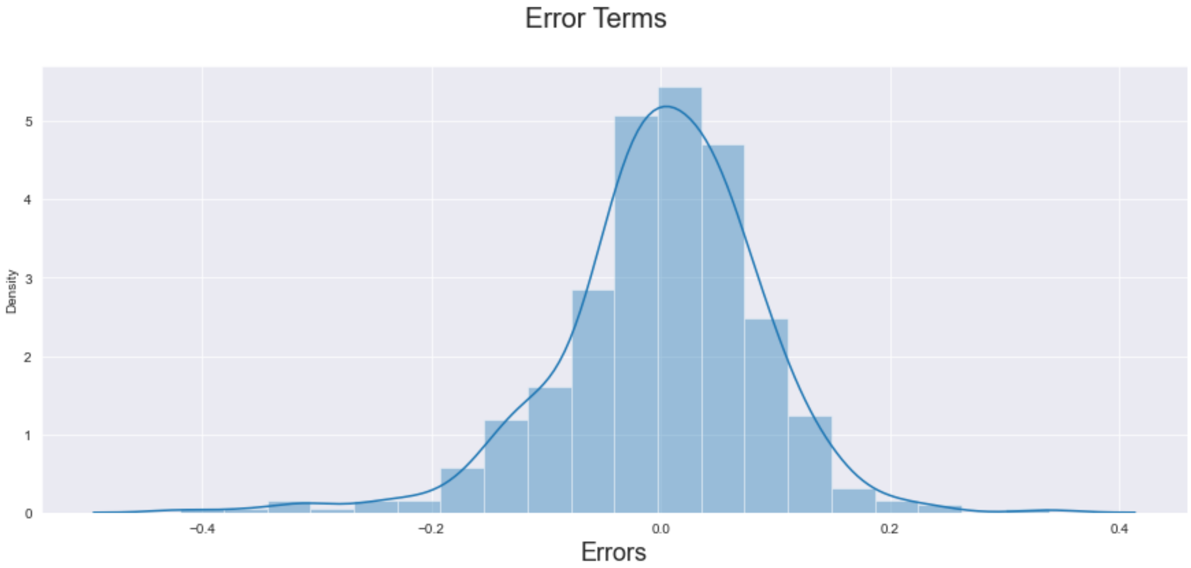
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Temperature has the highest correlation with the target variable bike rental count (cnt).

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)

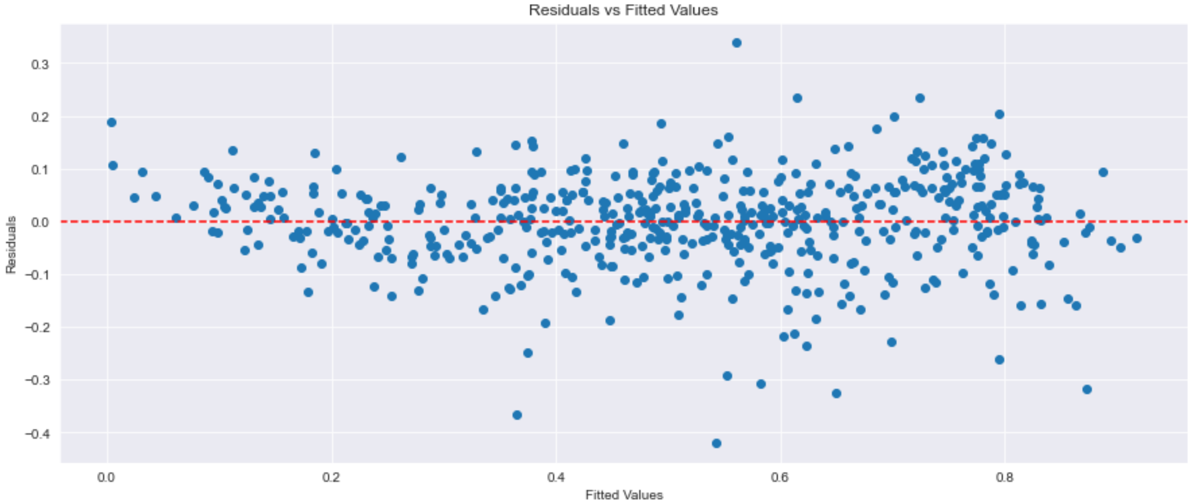
1) To check whether the error terms are normally distributed, we used a distribution plot(distplot) to plot the residuals. The below plot shows the error terms are centered around the mean indicating a normal distribution.

ss



3) To check whether the error terms have constant variance (homoscedasticity), we used a scatter plot and in the below plot you can see that random scatter of data points around the zero line explains the assumption of homoscedasticity being respected.

Additionally, since the residuals are scattered randomly without any visible trend, it suggests that the relationship between the predictor variables and the target variable is linear, which aligns with the assumptions of linear regression.



4) To check for multicollinearity, we examined the Variance Inflation Factor (VIF) for all independent variables. Since the VIF values were all below 5, this suggests that there is no significant multicollinearity among the variables.

5) To check whether the error terms are independent of each other, we used the Durbin-Watson statistic, which tests for autocorrelation among the residuals. The statistic ranges from 0 to 4, where a value of 2 indicates no autocorrelation. Since the value of the Durbin-Watson statistic is 2.0595 (acc. bike sharing dataset), which is close to 2, we can conclude that there is almost no autocorrelation, and therefore, the error terms are independent of each other.
Durbin-Watson statistic: 2.0595

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)

Based on the final model, below are the top 3 features that have contributed significantly towards explaining the demand for shared bikes:

1)Temperature – Temperature is the most significant feature as it has a strong positive correlation with bike rental count(cnt).

2) Year – The data indicates that the year 2019 saw a peak in bike rentals, resulting in a significant increase in the bike rental count.

3) Winter—During winter, bike rentals were higher, making this a significant feature for Boombikes's business.

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)

Linear Regression model is a supervised learning algorithm used for predicting a continuous target variable based on one or more independent variables. It establishes a linear relationship between the dependent variable and independent variables.

Types of linear regression model : .

1) Simple Linear Regression Model -

In this model, there is only one independent variable(X) and one dependent variable(Y).

The relationship between them is represented as : **y = B0 + B1X** (B0 is Beta 0 and B1 is Beta 1) Where y = Dependent Variable, X = Independent Variable, B0 = Constant/Intercept,

B1 = Coefficient(slope)

The goal is to find the best Beta 1 values that minimize the difference between actual and predicted values.

2) Multiple Linear Regression Model –

This model involves multiple independent variables and a single dependent variable

Equation is : **y = W1X1 + W2X2 + ….. + WnXn**

Where y = Dependent Variable, X1,X2…Xn = Features, W1,W2,…Wn = Coefficients(Weights), Wn+1 = contsant

In multiple linear regression, all variables are represented as vectors:

X = Feature matrix containing all independent variables.

y = Target vector containing all the predicted values.

W = Coefficient vector (weights)

In a regression model, we use the Ordinary Least Squares (OLS) method to estimate coefficients by minimizing the sum of squared residuals, where residuals are the difference between actual and predicted values**.**

The Mean Squared Error (MSE) is a commonly used metric that calculates the average squared residuals**.**

Additionally, Gradient descent is an optimization algorithm that updates model parameters step by step to minimize the cost function. It is especially useful for large datasets where OLS is computationally expensive.

We have two feature or coefficient selection approaches such as
RFE – Recursive feature elimination (automated approach) which automatically selects the number of features that are provided to the model.
Manual Approach – In this approach we can either select all the variables and drop one variable at a time also called as Backward Elimination approach. Additionally, we can select one variable and keep on adding other variables also called as Forward Selection Approach.

While doing feature selection we take into consideration additional metrics such as below :

1) p-values for independent variables should be less than 0.05 to indicate statistical significance.
2) Variance Inflation Factor (VIF) should be less than 5 to avoid multicollinearity issues.
3) A small p-value for the F-statistic indicates that the overall model is significant.
4) R^2 (coefficient of determination) measures how well the independent variables explain the dependent variable's variability.

There are very important assumptions for linear regression model which are as follows:

1) Error terms should be normally distributed
2) Error terms should be independent of each other
3) Error terms should have constant variance(homoscedasticity)
4) Linear Relationship between X and Y
5) No Multicollinearity.

In short, the goal of linear regression is to determine the best-fitting coefficients to minimize errors and make accurate predictions on test data, while selecting the most relevant features using approaches like RFE and manual selection.

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)

Anscombe's Quartet is a group of four datasets that have nearly identical statistical properties (such as mean, variance, correlation, and regression line) but look very different when plotted on a graph.

This demonstrates the importance of data visualization before applying statistical models, as relying only on numerical summaries can be misleading.

For example, consider four datasets with the same statistical measures. Despite these similarities, when we plot them, we observe that
1) One dataset follows a normal linear trend.
2) Another dataset shows a curved (non-linear) pattern.
3) A third dataset appears linear, but one outlier significantly affects the trend.
4) The fourth dataset has one extreme outlier that distorts the regression line.

Anscombe's Quartet proves that statistics alone are not enough—it is important to visualize data to understand its true pattern before applying any algorithms.

ss

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)

Pearson correlation coefficient denoted as r is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1 and is widely used in linear regression models and statistical analysis. It also helps to determine the strength of the relationship among 2 variables.

- If r = 1, there is a perfect positive correlation, meaning that as one variable increases, the other also increases proportionally.
  For eg: Restaurants with great food quality and better customer service tend to attract more customers, leading to higher revenue.

- If the r = 0, there is no correlation, indicating that changes in one variable do not predict changes in the other.
  For eg : In a consulting firm managing multiple projects, the success of one project for a client has no correlation with the success of a project for a different client as they are independent of each other.

- If the r = -1, there is a perfect negative correlation, meaning that as one variable increases, the other decreases proportionally.
  For eg: As the number of employees decreases in a startup project, the turnaround time (TAT) for a project increases.

Note - Correlation does not imply causation, meaning that even if two variables are correlated, one does not necessarily cause the other to change.

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)

Scaling is the process of transforming variables or features into a specific range so that each variable is on the same scale or has similar magnitudes ensuring uniformity across the dataset.

Scaling is performed to ensure that all variables are on similar scale, which helps improve model performance as well helps in better interpretation. When features are on the same scale, optimization algorithms such as (Gradient descent) can converge faster towards the minimum, leading to more stable model training. Additionally, scaling helps prevent overfitting, as features with larger magnitudes can disproportionately influence the model, leading it to overfit to those features.

The differences between normalized scaling and standardized scaling are as follows:

1) In normalized scaling also know as minmaxscaling, the variables are scaled in such a way that all the values lie between 0 and 1 using the maximum and minimum values in the data. The formula is : x = x-min(x)/max(x) – min(x)

2) In standardized scaling, the variables are scaled in such a way that their mean is 0 and standard deviation is 1.
   The formula is : x = x – mean(x)/sd(x)
3) Normalized scaling is sensitive to outliers because it transforms all the variables into a specified range (typically 0 to 1 or -1 to 1), and any outliers can distort this range, pulling other values closer to the boundaries.
4) In contrast, standardized scaling is less affected by outliers because it uses the mean and standard deviation, which are less influenced by extreme values compared to the minimum and maximum.

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)

- VIF (Variable Inflation factor) is a measure used to detect multicollinearity among the independent variables in a regression model.
- The VIF becomes infinite when there is perfect correlation (correlation of 1 or -1) among the independent variables. This means that one independent variable can be exactly predicted from another, causing perfect multicollinearity. As a result, the independent variables become redundant, making it impossible to distinguish between their individual effects on the dependent variable.
- Generally, VIF values > 10 have high multicollinearity issues suggesting that there is a strong correlation among the independent variables or predictors.

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) is a statistical tool that is used to compare the quantiles of the dataset against a theoretical distribution typically a normal distribution.

- QQ Plot helps us to determine whether the error terms (residuals) are normally distributed which is a very important assumption for a regression model.

- If the residuals are normally distributed, the points in the QQ plot will closely follow the 45 degree diagonal reference line. If the data points deviate from the line, especially at the tails, then it can indicate that error terms may not be normally distributed which can impact the regression model's validity. There can also be possibility of these deviations due to the existence of outliers.
