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)

i)Season - Season 1 is low in performance , season 2 and 3 are similar with season 4 catching up behind them

ii)Holiday - There is significant difference in mean of holiday and non-holiday; non-holiday gets more counts

iii)Weekday variable - The count remains almost stable across all days, with minor fluctuations

iv) Workingday – Working day and non-working day donot vary much except marginal increase for working day

v)Year- The latter year is having more counts and showing increase in popularity as the organisation ages

vi)Month - Counts increase steadily from January to peak around July–September

---

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 during dummy variable creation prevents multicollinearity by removing one reference category, which serves as a baseline. This ensures the model doesn't include redundant information, improving interpretability and avoiding overfitting. It simplifies the regression model by reducing the number of predictors.

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)

Variable temp

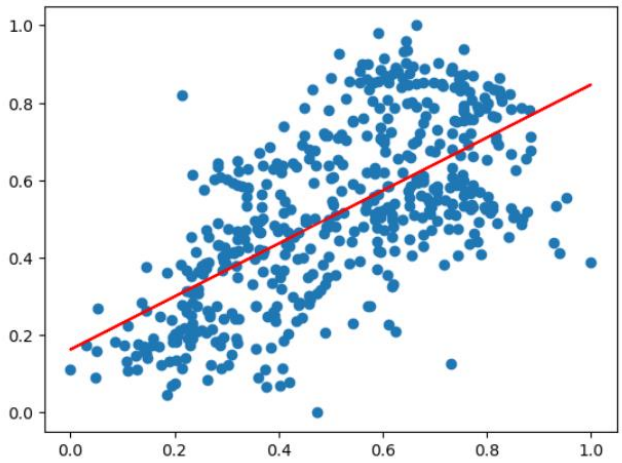
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

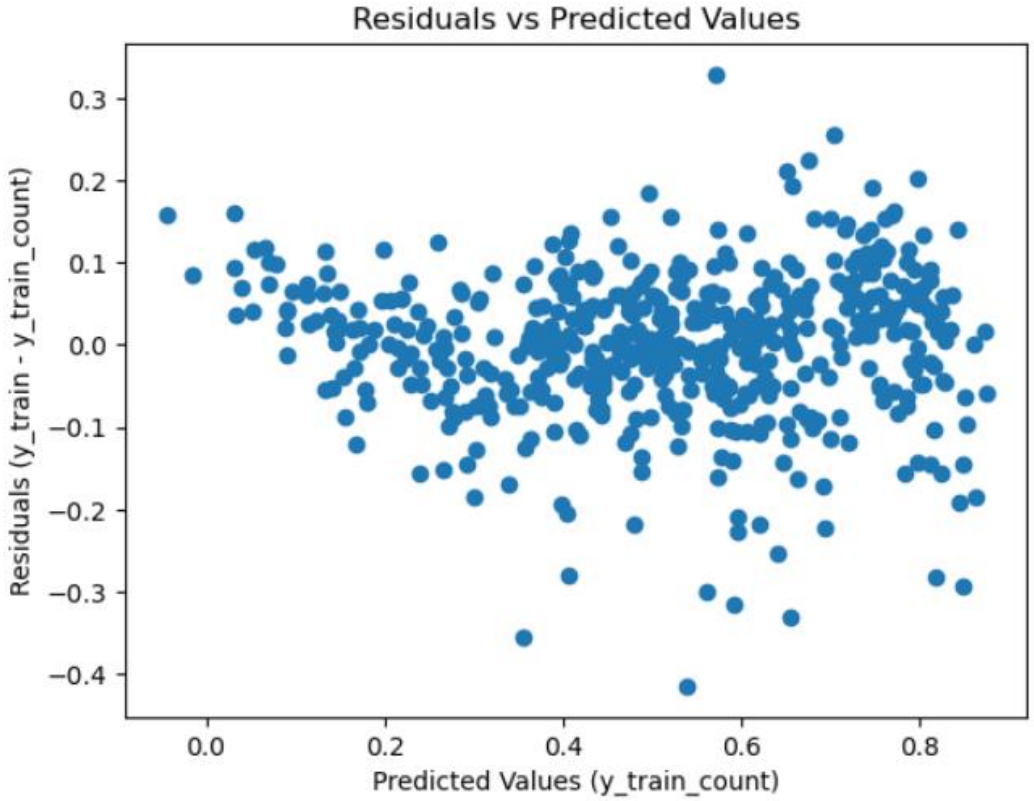
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. **Linearity** : **There is a linear relationship between X and Y** – The manual approach used in the beginning found 'atemp' having a strong linear relationship. The below scatter plot is for count = 0.163 + 0.684*atemp

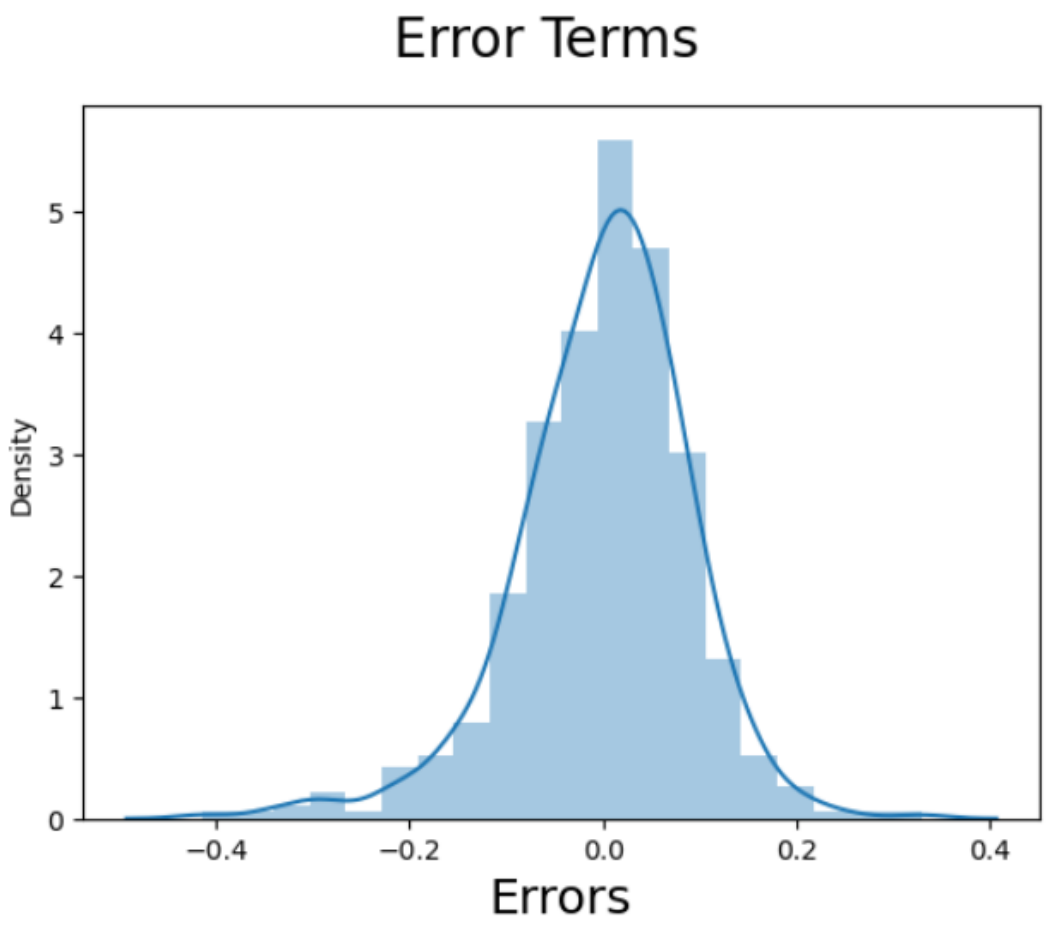
---



2. **Error terms are *independent* of each other** - Below scatter plot shows residual errors independent without any pattern



3. **Error terms are *normally distributed* with mean zero** - The residual errors are uniformly distributed



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)

1.Temp: Coefficient = 0.4741 (highest positive impact).
2. Rain Snow: Coefficient = -0.2869 (highest negative impact).
3. Year: Coefficient = 0.2346 (second highest positive impact).

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>

Overview : Linear regression is a supervised learning algorithm used to predict a continuous target variable based on one or more input features (independent variables). Linear regression finds a straight-line relationship between the features and the target. The line is represented as $y=b0+b1x1+b2x2+\cdots+bnxn$

where y is target variable , b0 intercept(value of y when all x are zero)

b1, b2,…,bn : Coefficients showing the effect of each feature x1,x2,…,xn

How it is done : Minimize the difference (errors) between the predicted values of y and the actual target values of y. This is done by minimizing the **sum of squared errors (SSE)** using a method called **Ordinary Least Squares (OLS)**.

Assumptions:

- Linearity: The relationship between features and target is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Variance of errors is constant across all levels of input.
- Normality: Errors (residuals) are normally distributed.

Output: The algorithm gives coefficients for each feature. These coefficients help explain how much the target changes when a feature changes, keeping others constant.

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 Francis Anscombe in 1973. These datasets are special because they have almost identical numerical statistics, like the mean, variance, correlation, and regression line. However, when plotted on a graph, they look very different from each other.

The first dataset shows a normal linear relationship, where the data points fit the regression line well. The second dataset has a curved pattern, showing a non-linear relationship. In the third dataset, most points follow a linear trend, but a single outlier influences the regression line heavily. The fourth dataset forms a vertical cluster, with one point completely affecting the regression.

So numbers alone can be misleading and to truly understand data, it is important to visualize it. This ensures that patterns, relationships, and outliers are clearly seen and not hidden by similar statistics.

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 measures the strength and direction of the linear relationship between two variables. It is also called the correlation coefficient. It ranges from −1 to +1, where +1 means a perfect positive linear relationship, −1 means a perfect negative linear relationship, and 0 means no linear relationship.

For example, if one variable increases as the other increases, R will be positive. If one variable decreases as the other increases, R will be negative.

Pearson's R helps in understanding how strongly two variables are related and whether the relationship is direct or inverse. It assumes the data is linear and normally distributed.

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 the process of adjusting the range of data features so that they are on a comparable scale. It is best practice used in multiple linear regression to ensure that no single feature dominates others due to differences in their magnitude. For example, features like age (in years) and income (in dollars) might have vastly different scales and could affect the model's performance if left unscaled.

Scaling is performed to improve the efficiency and accuracy of machine learning algorithms, especially those that rely on distance calculations like k-Nearest Neighbors or gradient-based algorithms like logistic regression and neural networks. It ensures that all features contribute equally to the model and prevents numerical instability during computations. Moreover, it significantly reduces algorithm execution time.

The difference between normalized and standardized scaling lies in how the data is adjusted. Normalization scales the data to a fixed range, typically between 0 and 1, using (x-min)/(max-in). Standardization transforms the data to have a mean of 0 and a standard deviation of 1 using the formula (x-mean/sigma). Normalization is useful for maintaining the original data distribution, while standardization is better when data follows a normal distribution or when no fixed range is required.

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 Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity, which means one predictor variable is an exact combination of one or more other predictors. This happens when the data has features that are completely dependent on each other.

VIF measures how much a variable's importance is distorted because of its relationship with other variables. If one variable is perfectly predictable by others, the calculations for VIF break down, and its value becomes infinite.

An infinite VIF shows a major issue in the data, and it can be fixed by removing or combining the variables that are too closely related. This helps make the model more stable and reliable.

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 plot, or quantile-quantile plot, is a graph used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It plots the quantiles of the dataset on the x-axis and the quantiles of the theoretical distribution on the y-axis. If the points in the plot form a straight line, it means the dataset follows the theoretical distribution.

In linear regression, a Q-Q plot is used to check whether the residuals (errors) follow a normal distribution, which is one of the key assumptions of linear regression. If the residuals are normally distributed, the plot will show a straight diagonal line.

The Q-Q plot helps in diagnosing potential problems in the regression model. If the points deviate significantly from the straight line, it suggests that the residuals are not normal, which can affect the reliability of hypothesis tests and confidence intervals in the model.
