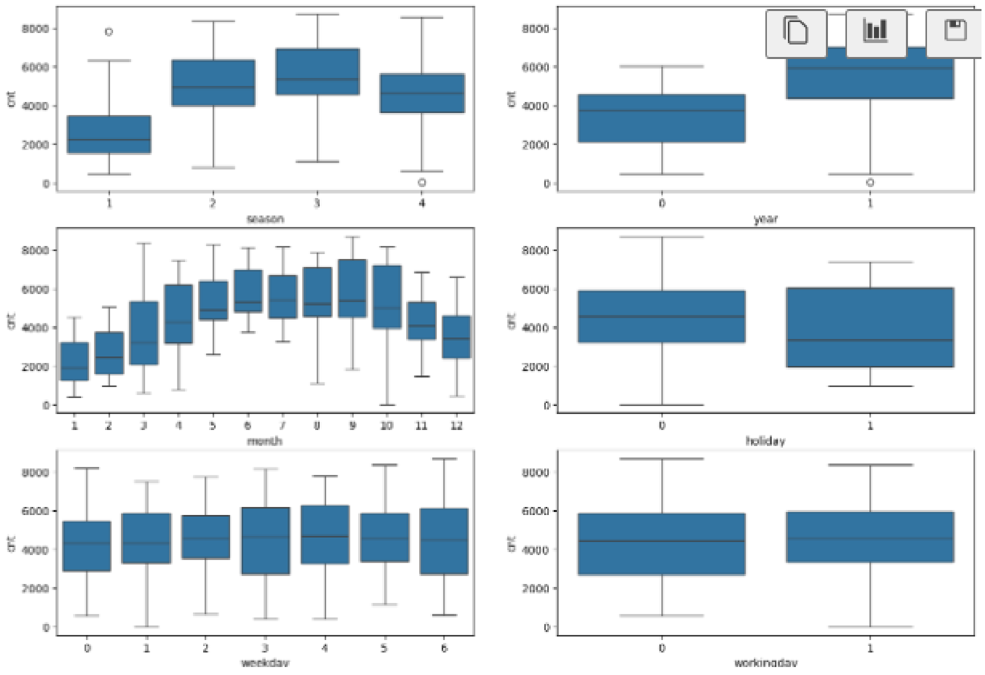
Assignment-based Subjective Questions

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)

There are the categorical variables season, year, month, holiday, weekday and workingday which impact the target variable cnt as depicted below:

1. Fall season has the highest bike rentals followed by, Summer, Winter and Spring.
   As spring has less booking, this time can be utilized for repairing the bikes or for any other maintenance activity.
2. June, August, September, and October have more demand as compared to other months of the year, whereas January, Feb, December being the low demand months
3. Wednesday and Thursdays have more demand.
4. 2019 has a sudden increase in the demand for bike rentals as compared to 2018.
5. Working days and holidays don’t have much impact on the bike rentals demands.



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)

The drop_first=True parameter removes the one dummy variable to avoid redundancy and multicollinearity during the dummy variable creation. It keeps k-1 categories for a variable with k unique values, making model simple and more stable.

Benefits/ importance of drop_first=True

1. **Avoiding Multicollinearity:**

   - If you include dummy variables for all categories (e.g., season _spring, season _summer, season _fall, season _winter), one category becomes redundant. This is because the presence of the information about any three categories is enough to predict the presence of the fourth (due to the constant sum constraint).
   - Including all dummy variables can lead to multicollinearity, where one predictor variable can be linearly predicted from the others with a substantial degree of accuracy.

2. **Interpreting Coefficients:**

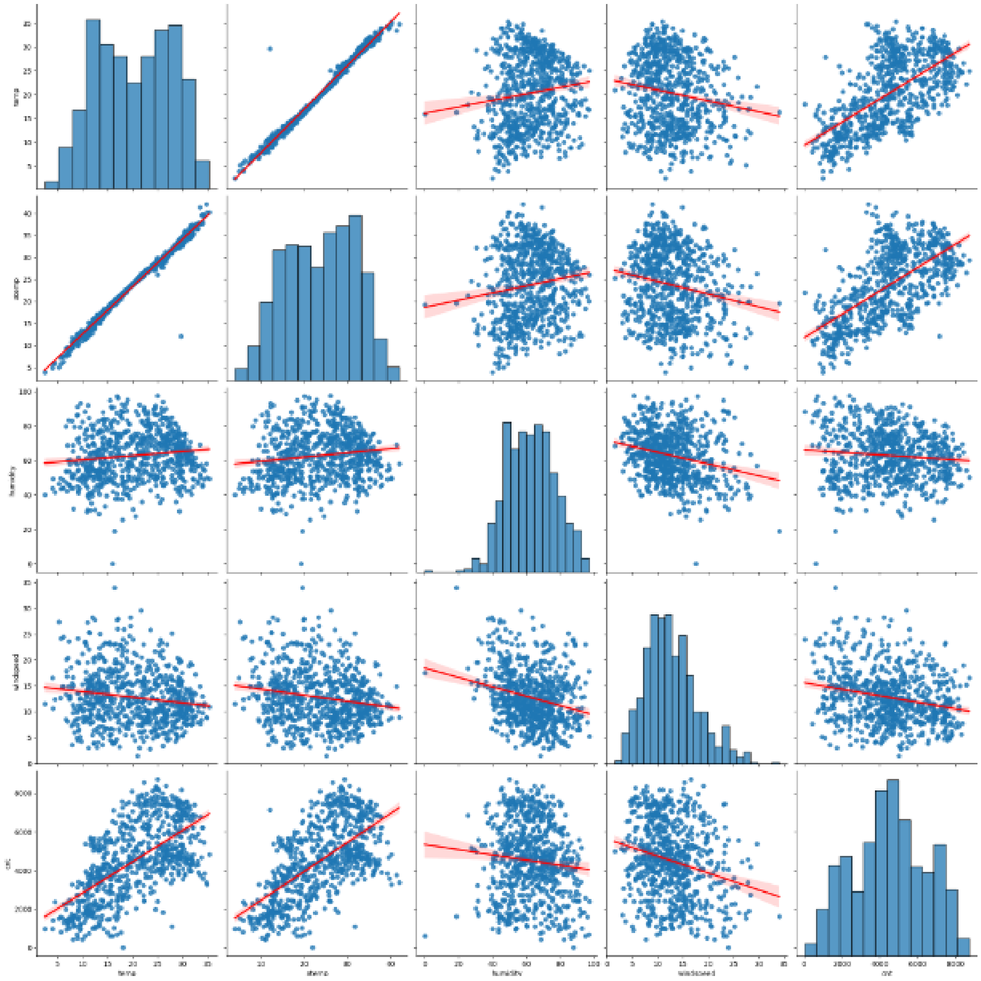
   - In regression models (like linear regression), coefficients of dummy variables represent the change in the dependent variable associated with a particular category relative to the baseline category (the category omitted).
   - When drop_first=True, one dummy variable is dropped (typically the one representing the baseline category). This ensures that each coefficient represents the effect of being in that category compared to the baseline category.

3. **Improving Model Performance:**

   - By reducing multicollinearity, models can perform better because they avoid the issues of unstable coefficients and inflated standard errors that can arise when predictors are highly correlated.

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)

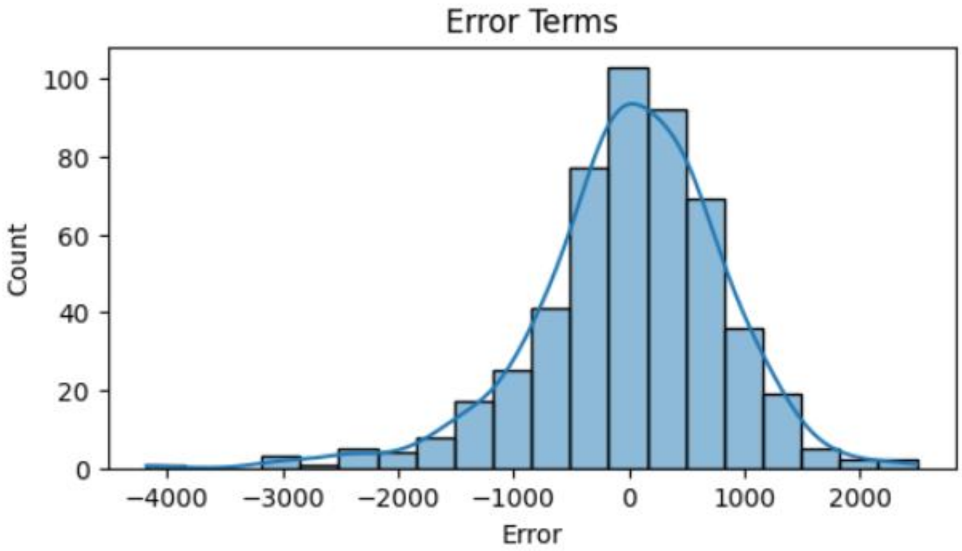
Per analysis, “temp” and “atemp” independent features have the highest correlation to the target variable “cnt”. Both of these independent variables have positive correlation with target variable, as compared to other independent variables. Below is the snapshot for the same.



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)

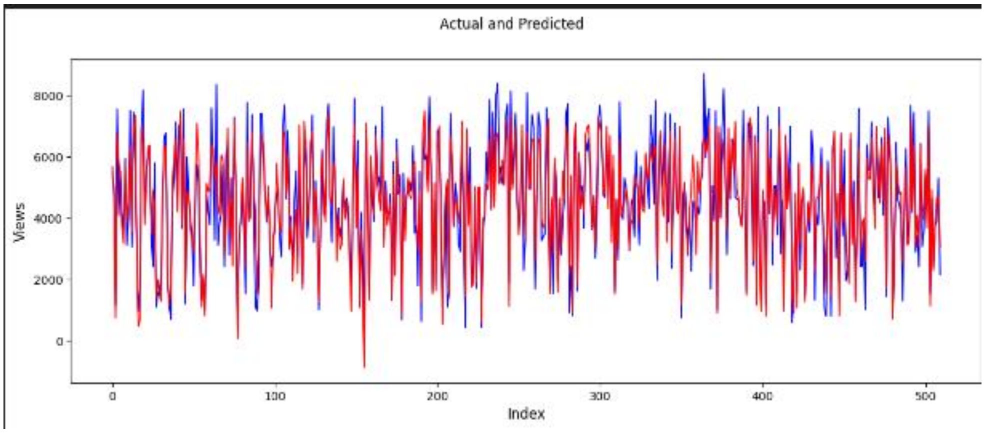
Multiple Linear Regression models are validated based upon fulfilling the criterion specified/based on Linearity of relationships, Independence of errors, Normality of error terms, homoscedasticity (constant variance of error) and No multicollinearity (variance inflation factor).

Residual analysis of the train data:



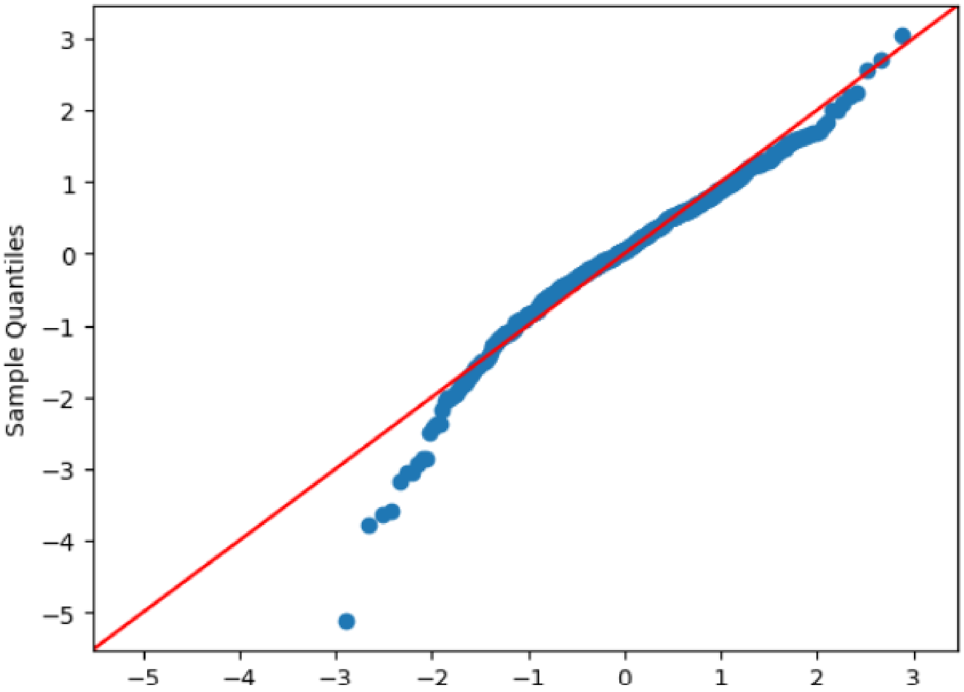
Error terms are normally distributed as per the above histogram.

Test vs Predicted:



Actual and predicated values are overlapping, so accuracy looks good.

Cross verifying the above conclusion with qq-plot



Here we could see most of the data points lie in on the straight line, and which indicates that the error terms are normally 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" or "atemp" (one of both)
2. "season"
3. "year"

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 used for predictive modeling used to deduce the linear relationship between a set of independent variables and one dependent variable aka target variable. The relationship is defined by the equation y=mx+c:
Y is the target variable, x is the independent variable, and m is the slope/coefficient, and the c is y intercept, this equation can be used to describe simple linear regression, multi linear regression is when there are multiple independent variables. c is also called constant.
Y= m1.x1+m2.x2+…..+mn.xn+c
These expressions provide the values for the coefficients which describe the relationship between

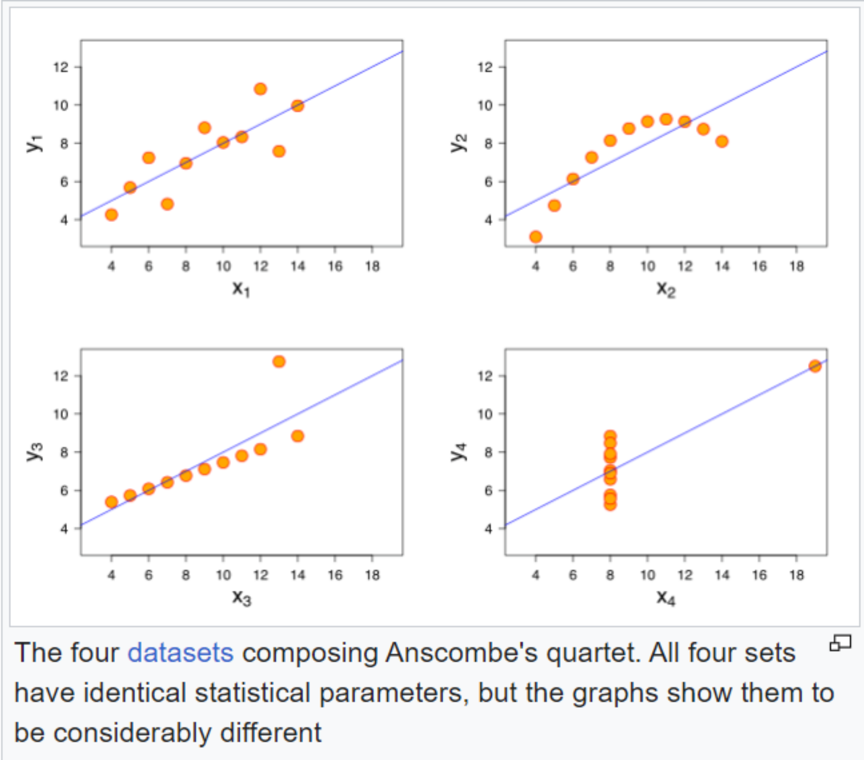
the independent and the dependent variables using the cost function and RSS, TSS, R^2 and adjusted R^2, either positive or negative as the slope of the line in 2D space of x and y coordinates.

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 group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is also used sometimes to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic **properties.**

Anscombe's quartet is a famous data set in statistics and data visualization that illustrates the importance of graphing data before analyzing it and highlights the pitfalls of relying solely on summary statistics. It consists of four small datasets that have nearly identical statistical properties when examined using simple summary statistics (mean, variance, correlation) but appear very different when plotted.



The four datasets composing Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

---

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>
The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It

is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

Below is the formula for calculating the pearson’s R:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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 in machine learning, also known as feature scaling or data normalization, it is the process of transforming a dataset's numerical features to a common scale or range. This is done to ensure that all features contribute equally to the model, which can improve the algorithm's effectiveness and speed up processing.
In both Normalized and standard scaling, one is transforming the values of numeric variables so that the transformed data points have specific helpful properties. The main difference is that, in scaling, you're changing the range of your data, while in normalization, you're changing the shape of the distribution of your data.

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>

This happens during the get dummies where some redundant or collinear data is written as dummy columns e.g. is drop_first is not used, all the features with multiple categories must be converted one by one and then concatenated to form one data frame otherwise it would result in inf VIF. Secondly if there are columns/features used which are highly correlated and high collinearity is introduced.

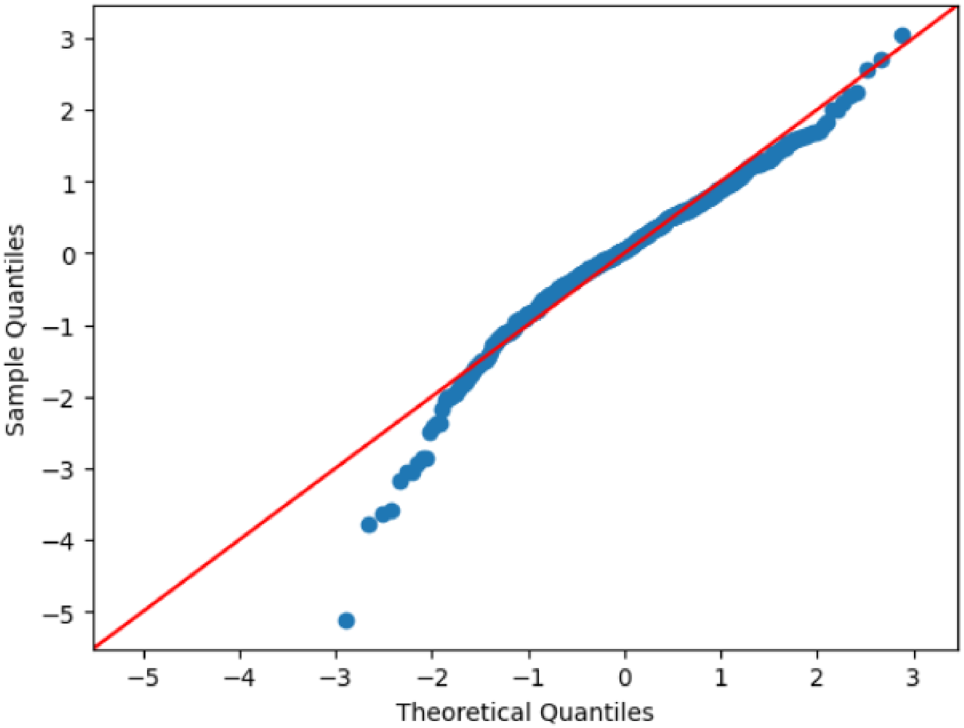
Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

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 graphical tool that compares two sets of data to determine if they come from the same distribution. It's a scatterplot that plots the quantiles of one data set against the quantiles of the other.


Sample Quantiles
Theoretical Quantiles