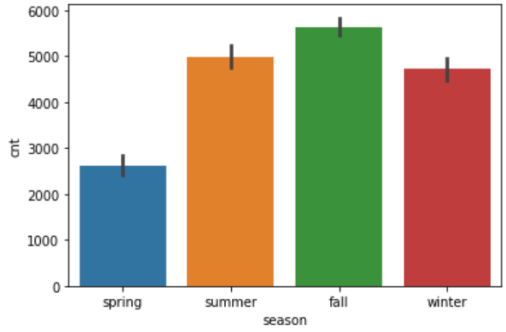
Assignment-based Subjective Questions

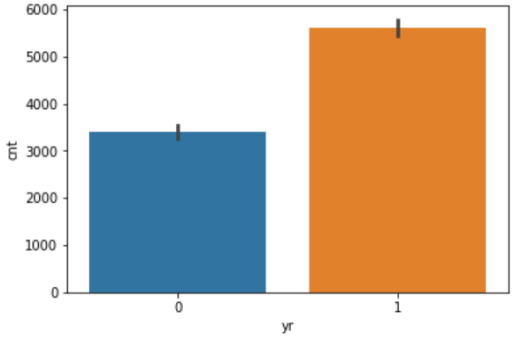
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

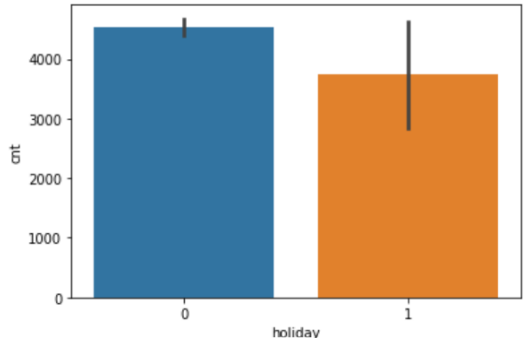
Answer: <Your answer for Question 1 goes below this line> (Do not edit)



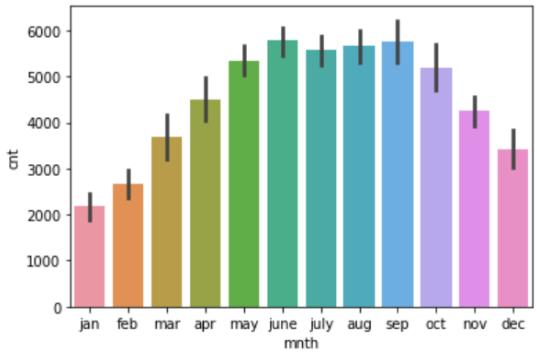
>>the count is largest in fall season which means when the weather is good outside.



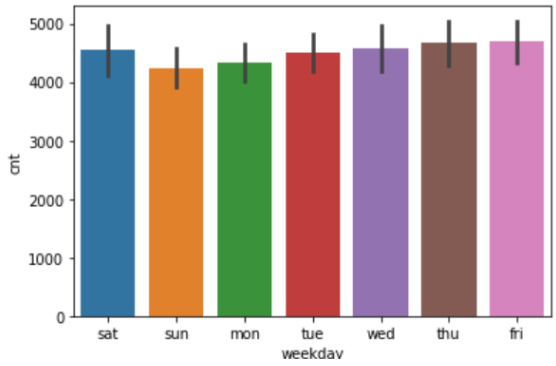
>>Count has increased in 2019 in comparison to 2018.



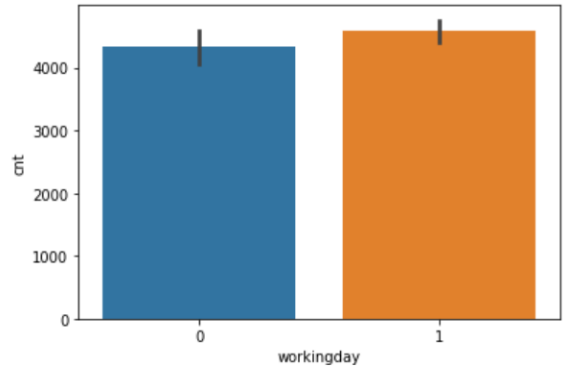
>> When it is not a holiday people travel more to work & thus demand is high.



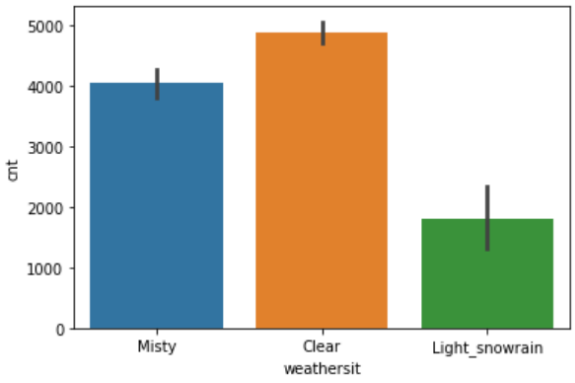
>> It is more starting from June to september when fall season is present.



>>This is least on sunday when offices are closed on weekends. Other days is comparatively the same.



>> If it is a working day then the demand is slightly more



>>When weather is clear then it is highly in 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)

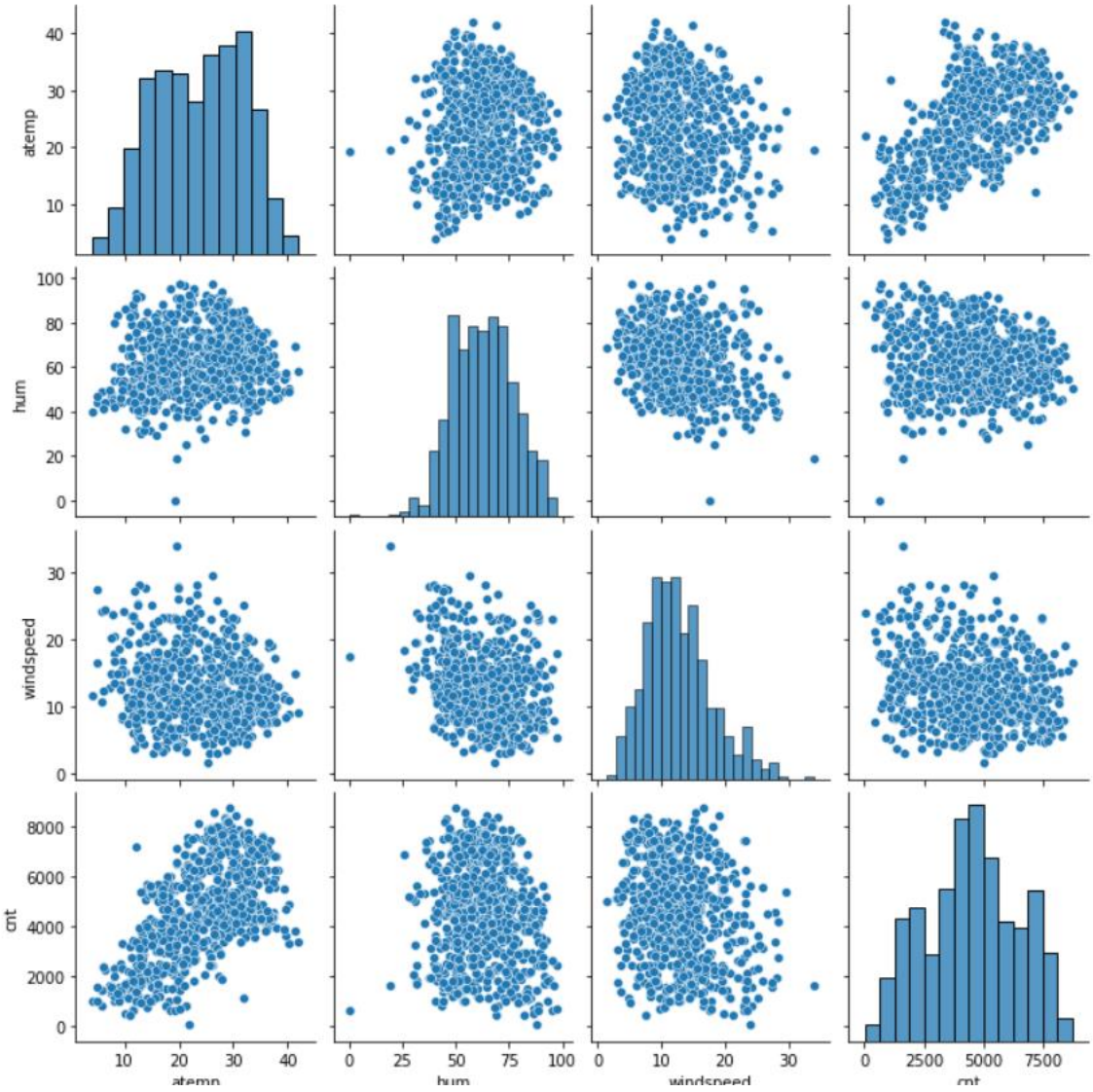
Because in dummy variable creation using pandas all N values are created & introduces redundancy in the dataframe, so to remove 1 newly created dummy column the command is used drop_first=True so as to make it N-1 dummy variable creation. Thus reducing the complexity & multicollinearity. Eg: For column season there are 4 values 1,2,3 & 4. So after dummy creation there would be 2,3 & 4 dummy columns only.

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)

Temp & Atemp has the highest correlation with CNT.



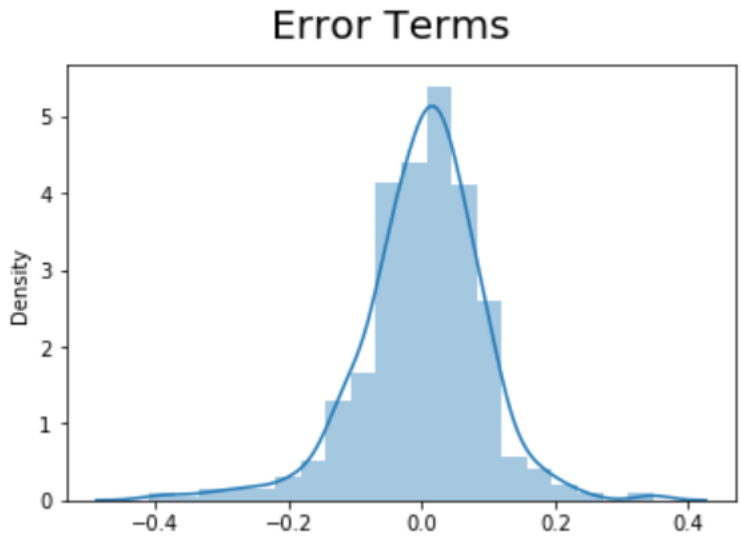
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

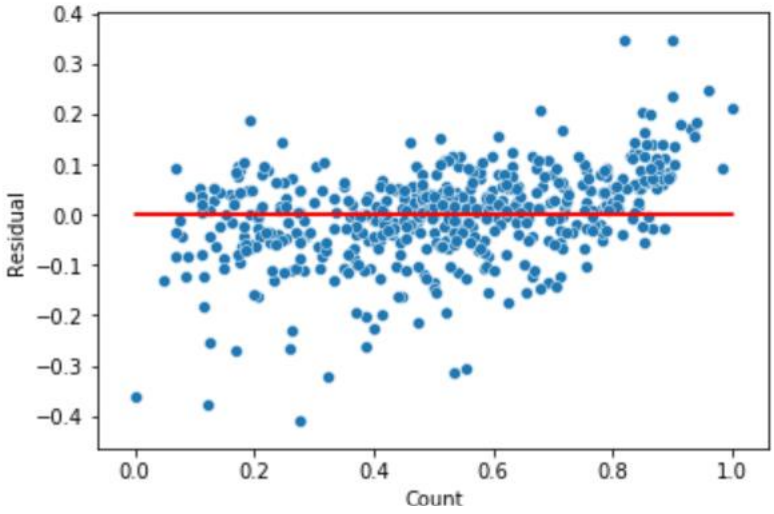
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

By using the same predictors of the model on the test set to predict the values of the target. Then check for residual errors using y_test from data & calculated y_test_pred from model.

Also check for assumptions:



Normal distribution of residuals



homosedacity – no pattern

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)
Top 3 features are:

1. Atemp – direct relationship
2. Year – direct relationship
3. Light_snowrain – Indirect relationship

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 to model the relationship between the predictor & target variables.

The generic equation can be represented by

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots\ldots + \beta nXn + e$

There are 2 types of LR – Simple & multiple

Assumption taken in the process:

- Linearity – the relationship between target & predictor should be linear/hyper lane
- Normal distribution – The residual should have mean at 0 & should be normally distributed
- Homoscedasticity – showing constant variance
- No visible patter with the residuals

Steps to build & evaluate the LR model

1. Data preparation
2. Split the data into train & test set
3. Scaling the data
4. Build the model with predictors
5. Evaluate the model on test set - OLS
6. Calculate the value of R^2
7. Check for aforementioned assumptions

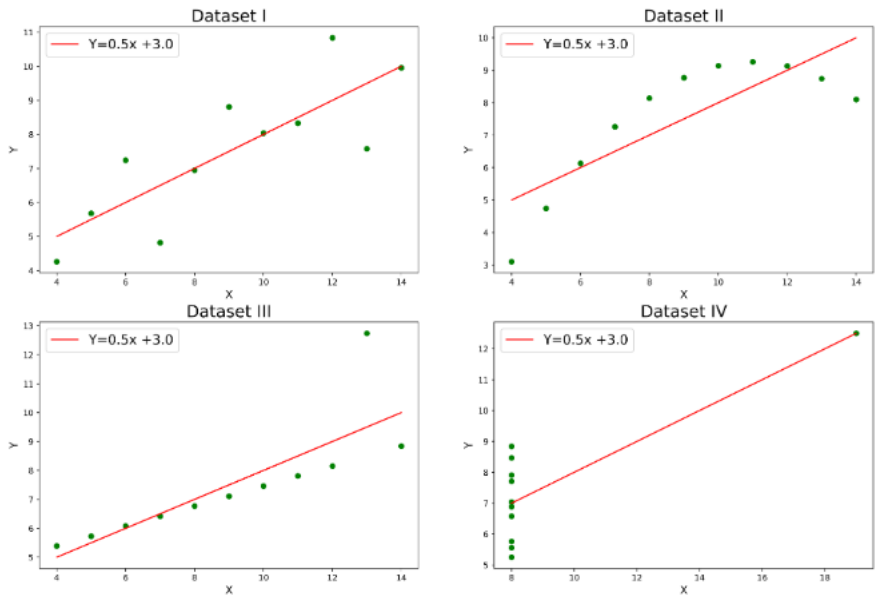
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Anscombe's quartet demonstrates that datasets with identical summary statistics can exhibit vastly different distributions and relationships. It underscores the critical need for visual data exploration t o detect patterns, outliers, and anomalies that summary statistics may obscure.

This visualization highlights the distinctive nature of each dataset, showcasing the power of graphica l analysis in understanding data beyond mere numbers.



Anscombe's quartet Plot

Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Conclusion

While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

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, also known as Pearson's correlation coefficient, is a measure of the linear relationship b etween two variables. It quantifies how closely related two sets of data are and ranges from -1 to 1.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Value Range:

- +1: Perfect positive linear relationship.

- 0: No linear relationship.

-1: Perfect negative linear relationship.

Pearson's R is a powerful statistical tool that helps in determining the strength and direction of a line ar relationship between two continuous variables.

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 transforming the range of data to ensure that all features contribute equall y to the model. It adjusts the values of the features to a common scale without distorting differenc es in the ranges of values.

Scaling is performed because

- Improves Convergence: Many machine learning algorithms, like gradient descent, perform better and converge faster when the features are on a similar scale.
- Equal Contribution: Features with larger ranges could dominate the learning process, leadi ng to biased models. Scaling ensures all features contribute equally.
- Algorithm Requirement: Some algorithms, such as K-Nearest Neighbors (KNN) and Principal Component Analysis (PCA), require scaled data to fu nction correctly.

Normalized Scaling:

- Definition: Normalization rescales the data to a fixed range, usually [0, 1].
- Formula: $$ X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} $$
- Use Case: Used when the distribution of data does not follow a Gaussian distribution and when you want to scale the data within a specific range.

Standardized Scaling:

- Definition: Standardization transforms the data to have a mean of 0 and a standard deviati on of 1.
- Formula: $$ X' = \frac{X - \mu}{\sigma} $$ Where μ is the mean and σ is the standard deviati on of the feature.
- Use Case: Used when the data follows a Gaussian distribution or when algorithms assume standard normal distribution (e.g., Linear Regression, Logistic Regression).

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>

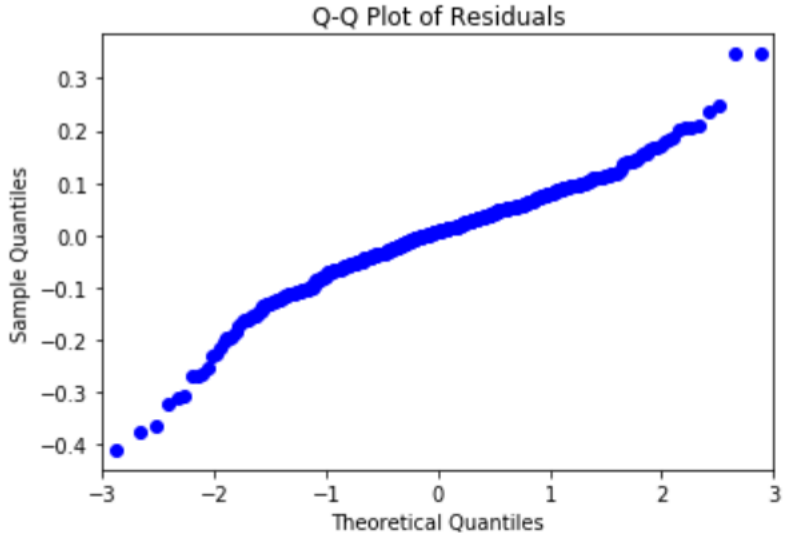
It happens because when 1 predictor is perfectly in multicollinearity with one or more another

predictor variable.
Multicollinearity affects the stability and reliability of the regression coefficients, making it crucial to detect and address. It happens when R2=1 & VIF=1/1-R2 so it gives infinite.

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)

A Q-Q plot, or Quantile-
Quantile plot, is a graphical tool used to assess whether a set of data follows a particular distributio n. It plots the quantiles of the data against the quantiles of a theoretical distribution, typically the normal distribution.
Use in Linear Regression:

- Assessing Normality of Residuals:
  - Linear regression assumes that the residuals (errors) are normally distributed.
  - A Q-Q plot helps check this assumption by plotting the residuals against a normal distri bution.
  - If the residuals are normally distributed, the points will lie approximately along the 45-degree reference line.
- Identifying Deviations:
  - Deviations from the reference line indicate departures from normality.
  - Patterns in the Q-Q plot can reveal skewness, kurtosis, or outliers in the residuals.
  - This helps in diagnosing potential issues with the model.



Importance in Linear Regression:

- **Validation of Assumptions**: Ensures the assumption of normality of residuals, which is cruci al for valid statistical inference.
- **Model Diagnostics**: Helps identify potential problems like outliers or non-normality that can affect the model's performance and validity.
- **Improving Model Fit**: By diagnosing issues with residuals, Q-Q plots can guide adjustments to the model or the choice of a more appropriate model.