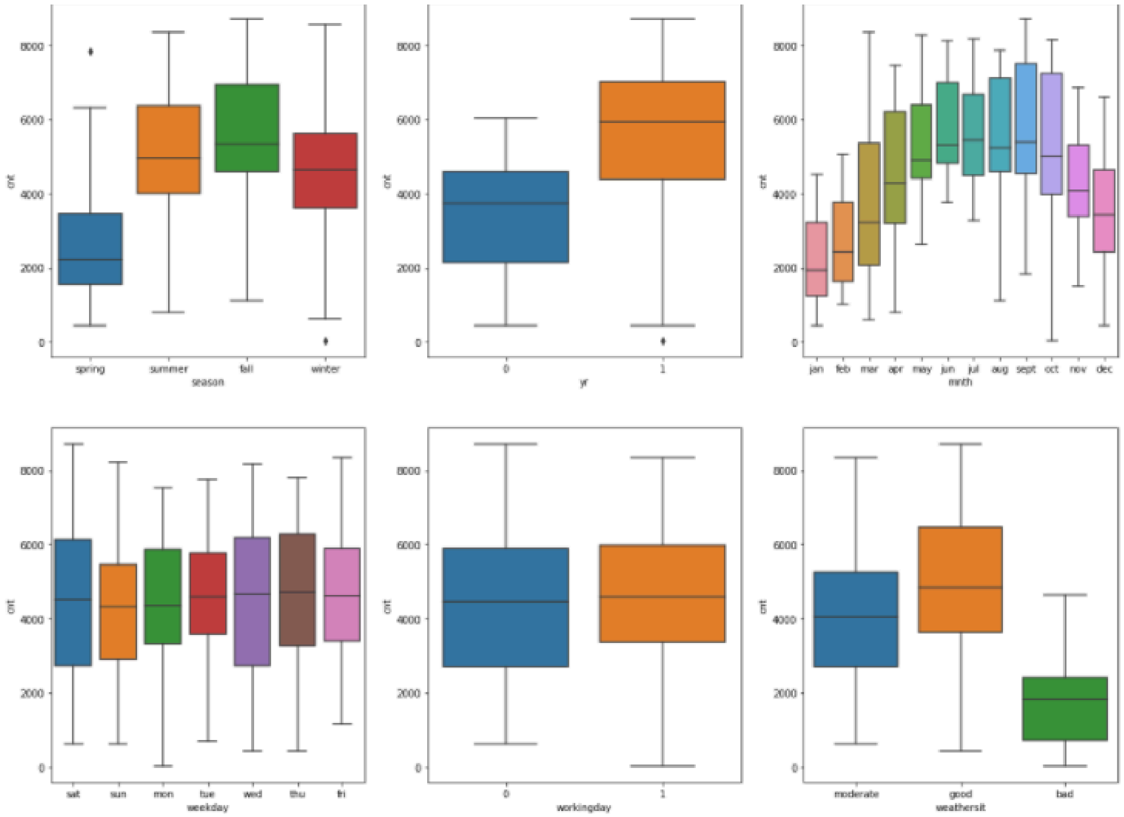
Assignment-based Subjective Questions

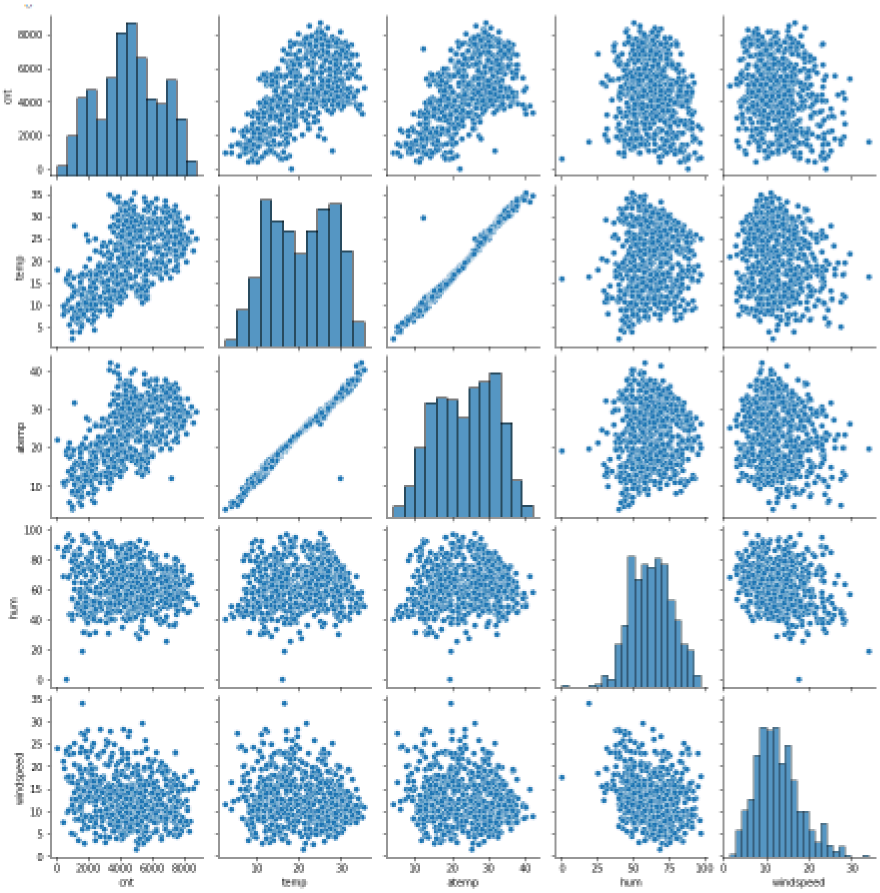
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?   (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)



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 avoids the dummy variable trap, which is multicollinearity caused by including all dummy variables. It also simplifies the model by using one category as a reference, making coefficients interpretable and reducing redundancy.

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)



The ‘temp’ and ‘atemp’ variables have highest correlation when compared to the rest with target variable as ‘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)

Linear Regression models are validated based on Linearity,No auto-correlation,Normality of error,Homoscedasticity, Multicollinearity

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 that has significant impact towards explaining the demand of the shared bikes are temperature,year and season

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 predictive modeling technique used to identify the relationship between a dependent variable (target) and one or more independent variables (predictors). It models this relationship as a straight line that best fits the data points, showing how the target variable changes with changes in the predictors.

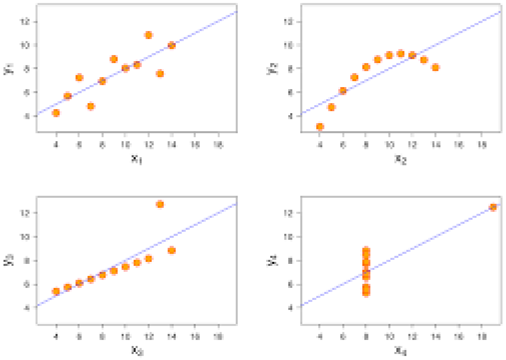
Types:

Simple Linear Regression: Involves one independent variable.
Multiple Linear Regression: Involves more than one independent variable.
Objective: The goal is to determine the optimal values for the intercept and slope, for simple regression or coefficients (for multiple regression). These values define the best-fit line, which minimizes prediction errors.
Regression Line:
Positive Relationship: As the predictor increases, the target also increases.
Negative Relationship: As the predictor increases, the target decreases.
Error Minimization: To find the best-fit line, the algorithm uses a cost function, such as the Mean Squared Error (MSE), which measures the average squared difference between actual and predicted values. Techniques like optimization algorithms or Recursive Feature Elimination (RFE) can be used to improve the model and identify relevant predictors.

Linear regression strives to model the relationship in the most accurate way by minimizing errors and ensuring the line fits the data points effectively.

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 that have nearly identical simple statistical properties but are strikingly different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. The quartet highlights how relying solely on summary statistics can lead to misleading conclusions.

The Four Datasets

Dataset 1 (Linear Relationship):

Follows a nearly perfect linear relationship.

The data points lie close to the regression line.

Dataset 2 (Non-linear Relationship):

Shows a clear non-linear pattern.

The data is best fit by a curve, not a straight line.

Dataset 3 (Influence of an Outlier):

The majority of the data points lie on a straight line.

A single outlier heavily influences the statistical measures and the regression line.

Dataset 4 (Vertical Alignment with Outlier):

Most data points share the same
x-coordinate, resulting in a vertical alignment.

A single outlier significantly affects the regression line and the correlation.

Dataset 1: A linear trend.

Dataset 2: A clear parabolic curve.

Dataset 3: A straight line disrupted by an outlier.

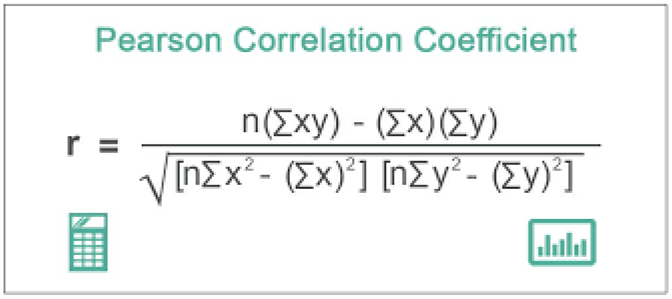
Dataset 4: A nearly vertical distribution with one outlier far from the group.

Anscombe's Quartet continues to be used in teaching statistics to stress:

- The importance of exploratory data analysis (EDA).
- The necessity of graphing data as a part of the analytical process.
- The dangers of over-relying on automated tools that summarize data without visual inspection.

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 Correlation Coefficient

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

Pearson's rrr, also known as the Pearson Correlation Coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson and is widely used in statistics and data analysis.
Applications of Pearson’s r:
Data Analysis:
To understand relationships between variables, such as height and weight or study hours and grades.
Regression Analysis:
To test the strength of linear dependence before fitting a regression model.
Feature Selection:
In machine learning, it is used to identify strongly correlated features.
Hypothesis Testing:
To test whether a linear relationship exists between two variables (using r-values and p-values).
Limitations of Pearson’s r:
Sensitivity to Outliers:
Outliers can significantly distort the correlation coefficient.
Linear Relationships Only:
Pearson’s r does not capture non-linear relationships.
Does Not Imply Causation:
A high correlation does not mean one variable causes the other.

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 preprocessing technique in data analysis and machine learning used to adjust the range of independent variables (features) so that they contribute equally to the analysis or model. This is particularly important in algorithms that are sensitive to the magnitude of feature values.
Why is Scaling Performed?
Algorithm Sensitivity
Faster Convergence
Equal Contribution of Features
Improved Accuracy
Consistent Interpretation

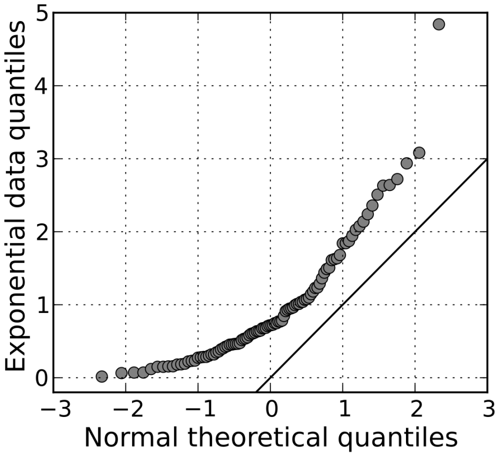
Aspect	Normalized Scaling	Standardized Scaling
Purpose	Rescales data to a fixed range (e.g., [0, 1])	Centers data to mean 0 and scales to unit variance.
Range	[0, 1] or [-1, 1], depending on scaling.	No fixed range; depends on the dataset.
Outlier Sensitivity	Highly sensitive to outliers.	Less sensitive to outliers.
Application	Use when range constraints are critical (e.g., pixel intensity).	Use when features follow a normal distribution or for algorithms requiring mean-centered data.
Examples of Algorithms	k-NN, min-max constraints in neural networks.	PCA, logistic regression, linear 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>
The Variance Inflation Factor (VIF) is a statistical measure used to detect multicollinearity in regression analysis. It quantifies how much the variance of a regression coefficient is inflated due to the presence of multicollinearity among independent variables.
A VIF value becomes infinite when:
Perfect Multicollinearity:
This leads to the denominator in the VIF formula (1 - ri^2) becoming zero, resulting in an infinite value.
Exact Linear Dependency
the regression matrix becomes singular and cannot be inverted, making it impossible to estimate the regression coefficients reliably.

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 (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, most commonly the normal distribution. It compares the quantiles of the dataset against the quantiles of a reference distribution. If the data matches the distribution, the points on the Q-Q plot will lie approximately along a straight line.

Importance of Q-Q Plot in Linear Regression

Validation of Assumptions:

Linear regression relies on the assumption that residuals are normally distributed. A Q-Q plot provides a quick visual way to validate this assumption.

Improves Model Reliability:

Checking for normality ensures that statistical tests (e.g., hypothesis testing for coefficients) produce valid results.

Detecting Issues with the Model:

If residuals deviate significantly from normality, it may indicate:

Outliers or influential points.

Incorrect functional form of the model.

Heteroscedasticity or non-constant variance.

Guiding Transformations:

If normality is violated, a Q-Q plot helps identify whether transformations (e.g., log, square root) can improve the model's performance.