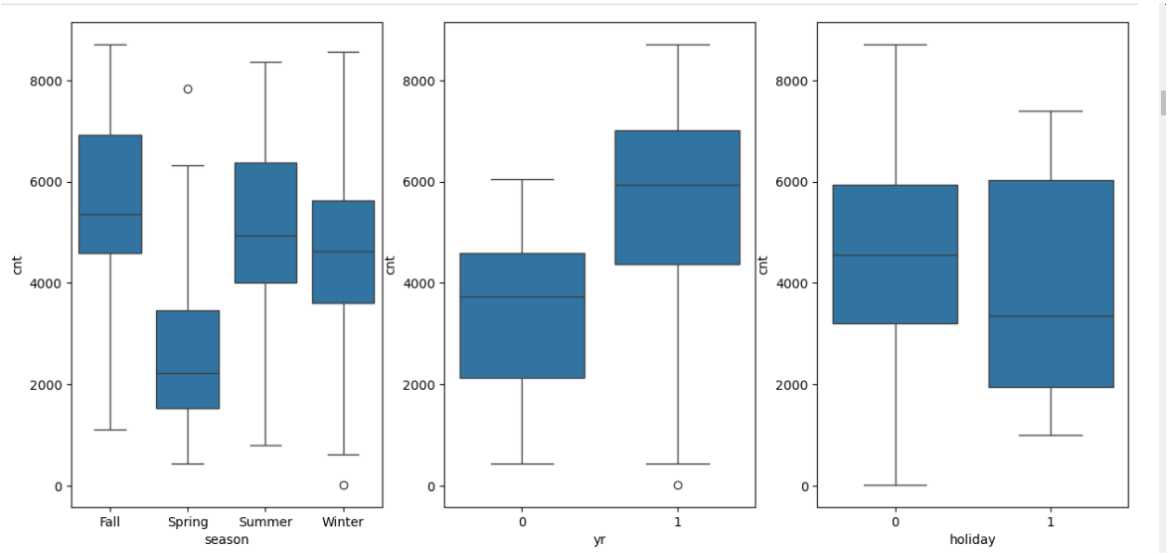
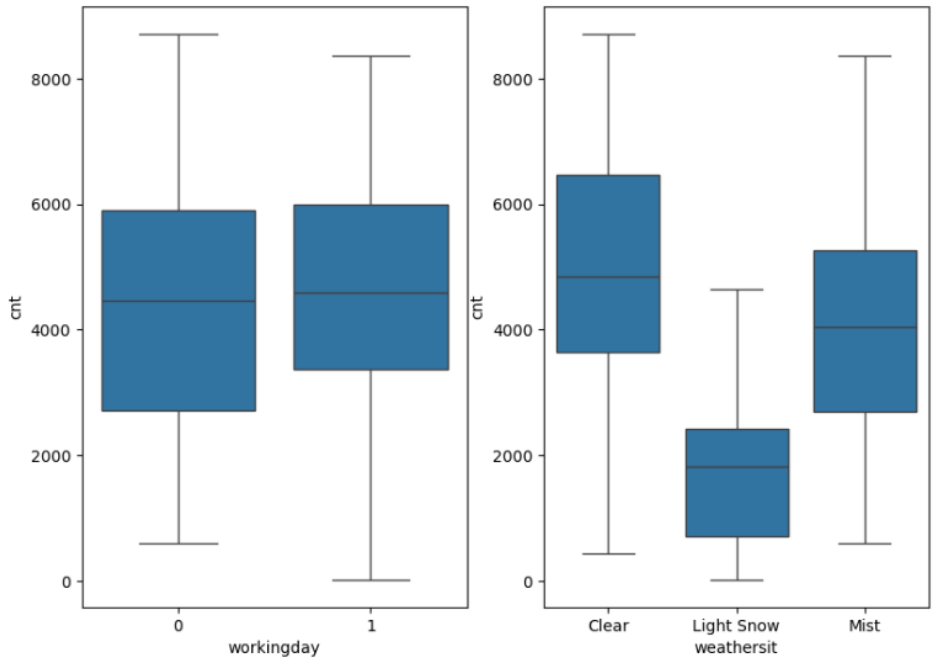
Assignment-based Subjective Questions

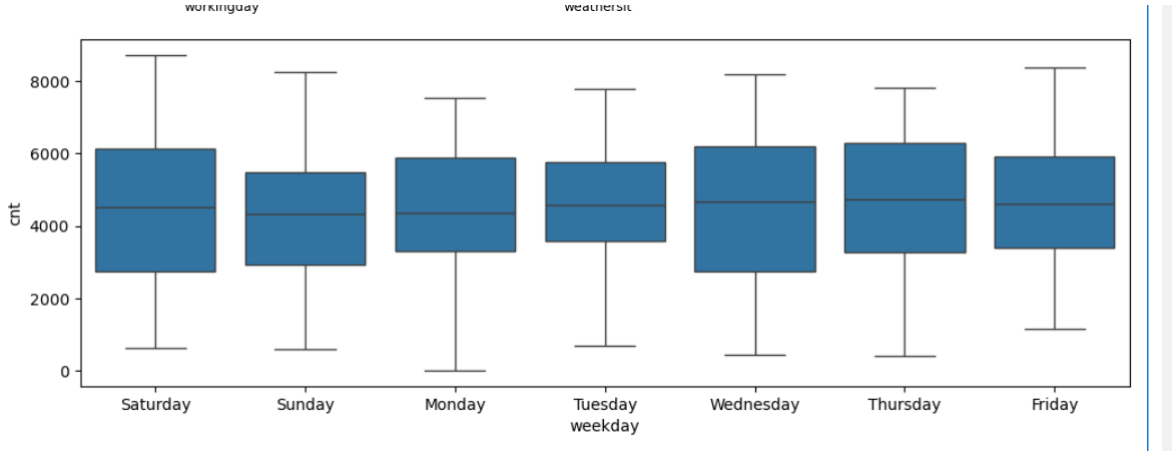
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**
   Answer:



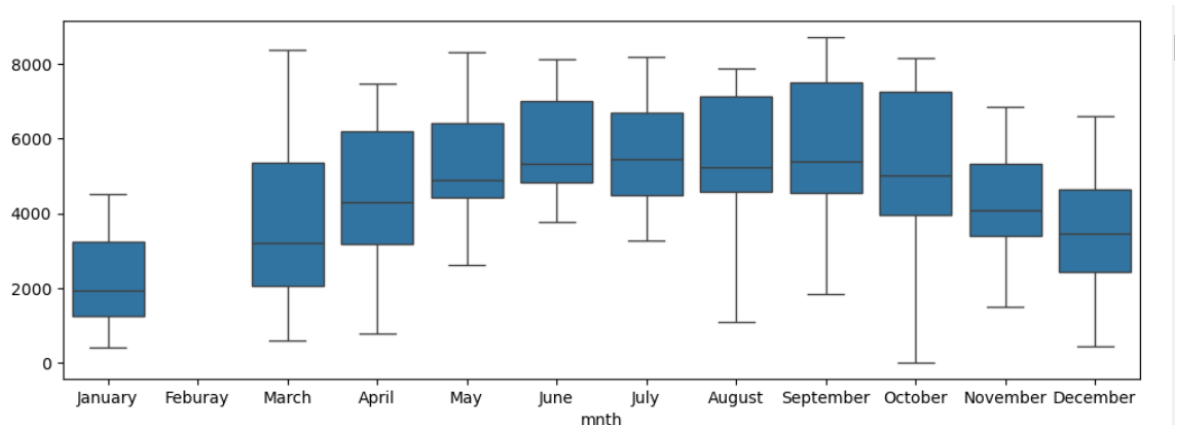
- For Season variable- Fall is the highest median so demand for bike rental was quite high. Its inverse in Spring and Winters
- For Year variable- Year 2019 has high demand due to a greater number of people than year 2018
- For Holiday variable- Rental reduces during holiday



- Workingday: Maximum booking happened between 4000 to 6000, median is constant throughout week so there is not much difference working days or non-working days.
- Weathersit: High user when Light Snow (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) while no user during Clear(Clear, Few clouds, Partly cloudy, Partly cloudy weather)
-



- Weekday: Rentals are almost equal so much difference.



Month variable (mnt): Rentals were increased from Jan to September, and was on peak in month of September month, then lowered until reach December.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   Answer:

   It helps to avoid multicollinearity in the dataset during dummy variable creation.

   Multicollinearity occurs when tow or more predictor variables in regression mode are highly correlated.

   It helps to improve the interpretation and stability of the regression model.

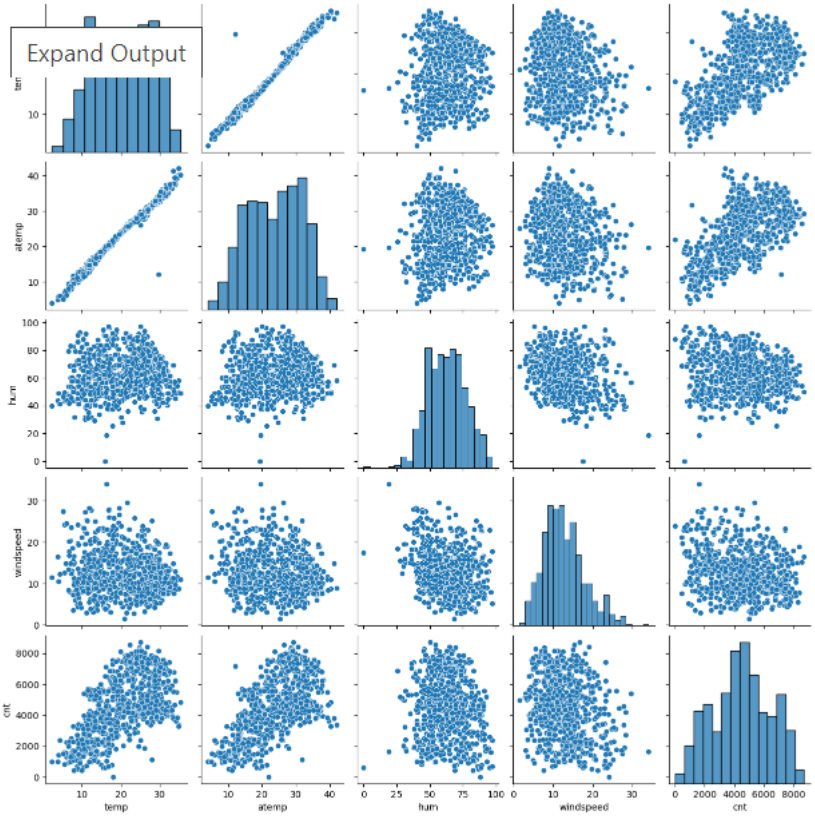
   Syntax: drop_first=True

   Example: data = pd.get_dummies(data['season'],drop_first=True)

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer:

“temp” and “atemp” has the highest correlation with the target variable “cnt”.



4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Answer:

Validation of Linear Regression after building the model based on training set.

- Residual analysis: Difference between observed and predicted values for randomness and constant variance.
- Normality of error terms: Error term should be normally distributed.
- Multicollinearity check: There should be insignificant multi collinearity between variables. VIF can identify.
- Linearity: By plotting observed vs predicted graph or by examine residual plots
- Homoscedasticity: Plot residual against predicators. There spread should be ideally consistent.
- Independence of residuals: Plot residual against time or pattern. There should be no apparent patterns or correlation.
- Linear relationship validation: Linearity should be visible among variables.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Answer:

1. Temperature: coefficient 0.446163 says increase in temp will increase in bike rentals.
1. Weathersit: Coefficient 0.297004 says Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds has more rentals.
2. Year: Coefficient 0.231425 says Yearwise rentals increased.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

   Answer:

   Linear regression is a statistical method used to model the relationship between one or more independent variables (often denoted as xx) and a dependent variable (often denoted as yy). The goal of linear regression is to find the best-fitting straight line that describes the relationship between the independent and dependent variables.

   1. Assumptions:

   Linearity: The relationship between the independent and dependent variables is linear.

   Independence: Observations are independent of each other.

   Homoscedasticity: The variance of the residuals (the differences between the observed and predicted values) is constant across all levels of the independent variable.

   Normality: The residuals are normally distributed.

   2. **Model Representation:** The linear regression model can be represented as:

   $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\varepsilon y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\varepsilon$ where:

   yy is the dependent variable,

   x1,x2,...,xnx1,x2,...,xn are the independent variables,

   β0β0 is the intercept (the value of yy when all independent variables are zero),

   β1,β2,...,βnβ1,β2,...,βn are the coefficients (the slopes of the regression line),

   εε is the error term.

   3. Objective:

The objective of linear regression is to find the values of the coefficients (β0,β1,...,βnβ0,β1 ,...,βn) that minimize the sum of the squared differences between the observed and predicted values (the residuals). This is known as the least squares method.

Fitting the Model: The model is fitted to the data by estimating the values of the coefficients. This is typically done using optimization techniques such as gradient descent or analytical methods such as the normal equation.

Model Evaluation:
Once the model is fitted, it is evaluated to assess its performance and determine how well it captures the relationship between the variables. Common evaluation metrics include:
R-squared: The proportion of the variance in the dependent variable that is explained by the independent variables.
Mean squared error (MSE): The average of the squared differences between the observed and predicted values.
Prediction:

Overall, linear regression is a simple yet powerful method for modeling the relationship between variables and making predictions based on that relationship.

4. Explain the Anscombe's quartet in detail. (3 marks)

Answer:
Anscombe's quartet is set of four datasets that have nearly identical simple descriptive statistics but differ greatly when graphed or analyzed further. It was created by the statistician Francis Anscombe to demonstrate the importance of visualizing date before analyzing it and to highlight the limitation of summary statistics alone.
Each dataset consists of 11 points with x and y coordinates, shares the mean, variance, correlation coefficient and linear regression line. However, when plotted, they reveal different relationships between the variables such as linear, or no relationship at all. This highlights danger of relying solely on summary statistics without examining the underlying data distribution. Anscombe's quartet is often used to emphasize the importance of exploratory data analysis and the potential pitfalls of relying solely on summary statistics.

5. What is Pearson's R? (3 marks)

Answer:

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.

Pearson correlation	**Coefficient (r)**	**Correlation type Interpretation** (When one variable change, other variable changes in)
Between 0 and 1	Positive correlation	Same direction.
0	No correlation	There is no relationship between the variables.
Between 0 and –1	Negative correlation	Opposite direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is preprocessing technique used in machine learning to standardize the range of independent variables or features. The goal of scaling is to bring all variables to a similar scale, typically between 0 and 1 or with a mean of 0 and standard deviations of 1. This ensures that no single feature dominates the others, and that the algorithm can converge faster and perform better.

Scaling performed to address differences in the units or magnitudes of different features, which can otherwise affect the performance of machine learning algos, especially those sensitive, such as K-nearest neighbors and support vector machines.

Main difference between normalized scaling and standardized scaling lies in the method used to scale the features:

1. Normalized Scaling (Min-Max scaling):
    a) It rescales features to fixed range, typically between 0 and 1.
    b) Useful when the distribution of the data is unknown or not Gaussian.
    c) Sensitive to outliers
    d) Retains the shape of the original distribution.
    e) May not preserve the relationships between the data points.
    f) Equation: (x – min)/(max – min)

2. Standardized Scaling (Z-score normalization) :
    a) standardizes the feature to have mean 0 and standard deviation of 1.
    b) Useful when the distribution of the data is Gaussian or unknown.
    c) Less sensitive to outliers
    d) Changes the shape of the original distribution.
    e) Preserves the relationships between the data points.
    f) Equation: (x – mean)/standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)

Answer:

VIF (Variance Inflation Factor) can become infinite when one of the independent variables in regression model is linear combination of other independent variables.

This perfect multicollinearity leads to an inflated VIF, as variance of the coefficient estimate become infinitely large.

Practically, infinite VIF indicates that one or more predictors in the model are redundant or highly correlated with each other, making it difficult to estimate their individual effect accurately.

In such cases, it's necessary to identify and address the multicollinearity issue, often by removing one of the correlated variables or using techniques like regularization to stabilize the coefficient estimates.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

Answer:

Q-Q plot, Quantile-Quantile plot is graphical tool used to compare two probability distributions by plotting their quantiles against each other.

In the context of linear regression, Q-Q plot is often used to assess whether the residuals (difference between observed and predicted values) follow normal distribution.

Importance of Q-Q plot in linear regression lies in its ability to check the assumption of normality of residuals. If the residuals are normally distributed, the points in the q-Q plot will approx. fall along straight line. Deviation from straight line indicate depatures from normality, which can affect the validity of statistical inferences drawn from the regression model.

Q-Q plot helps to verify whether the residuals meet the assumptions required for linear regression analysis, which is crucial for making accurate predictions and reliable inferences.