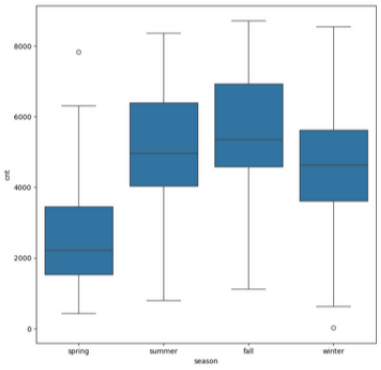
Assignment-based Subjective Questions

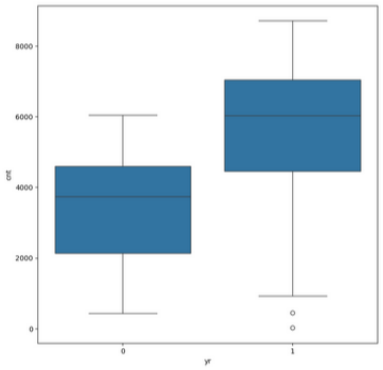
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

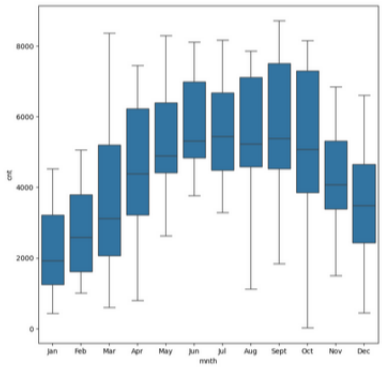
Total Marks: 3 marks (Do not edit)

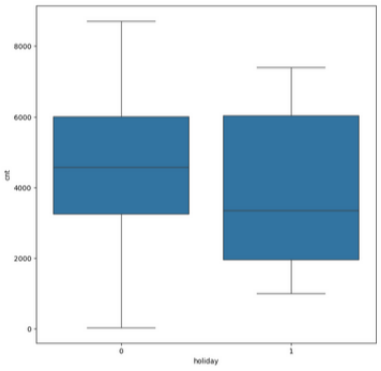
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

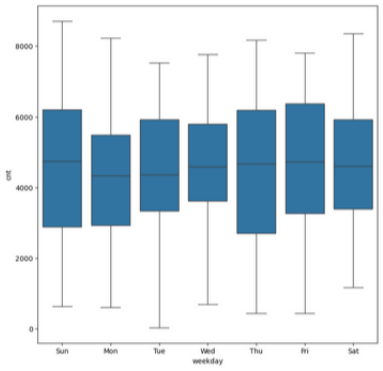
- The median bike rentals are more in fall, followed by summer, winter and spring.
- There are significantly more bike rentals in the year 2019 than in 2018.
- The median rentals have marginally increased from January to June, stayed almost stable till October, followed by marginal decrease till December.
- The month of January has lower median rental than any other months.
- The rentals are less on holidays compared to working days.
- The medium rentals are almost same on all the days.
- There are no rentals during Rain or Thunderstorms (4) and more rentals in 1 (Clear & Few Clouds) followed by 2 (Mist & Cloudy) and fewer rentals for - 3 (Light Snow)

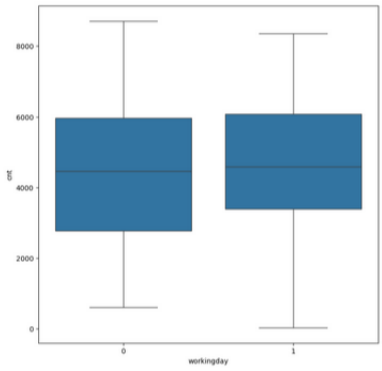


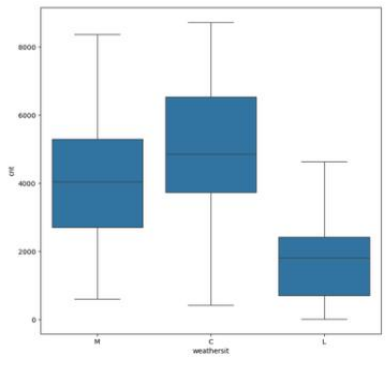


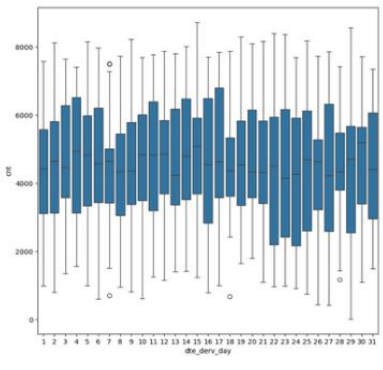












Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

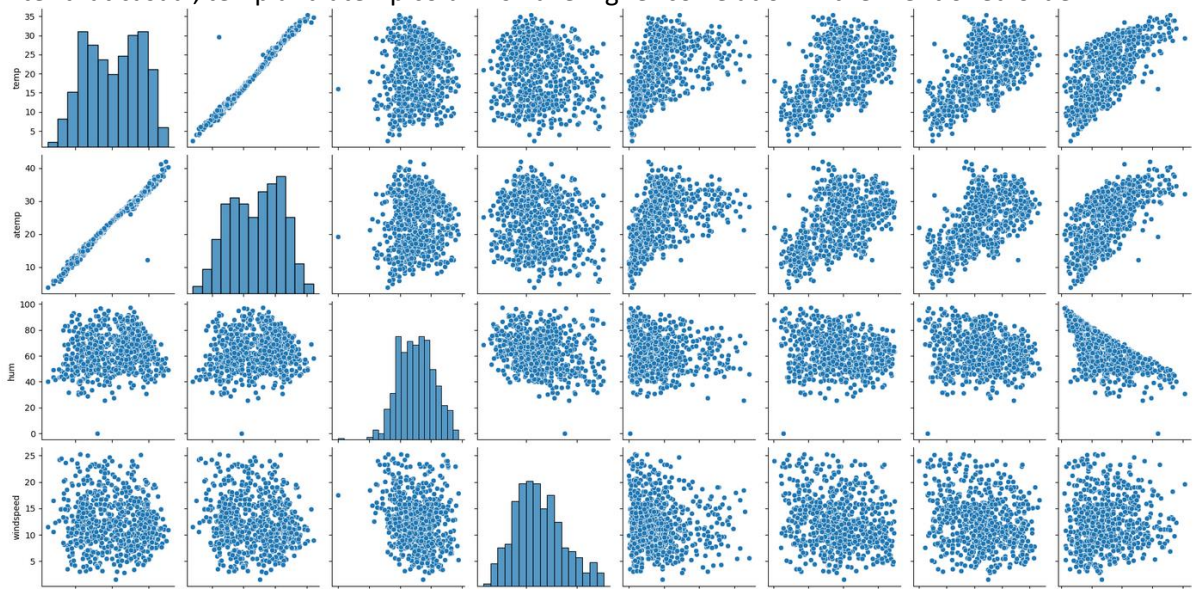
We convert categorical variables into dummy variables. The idea of dummy variable creation is to build 'n-1' variables for a categorical variable with say 'n' levels.

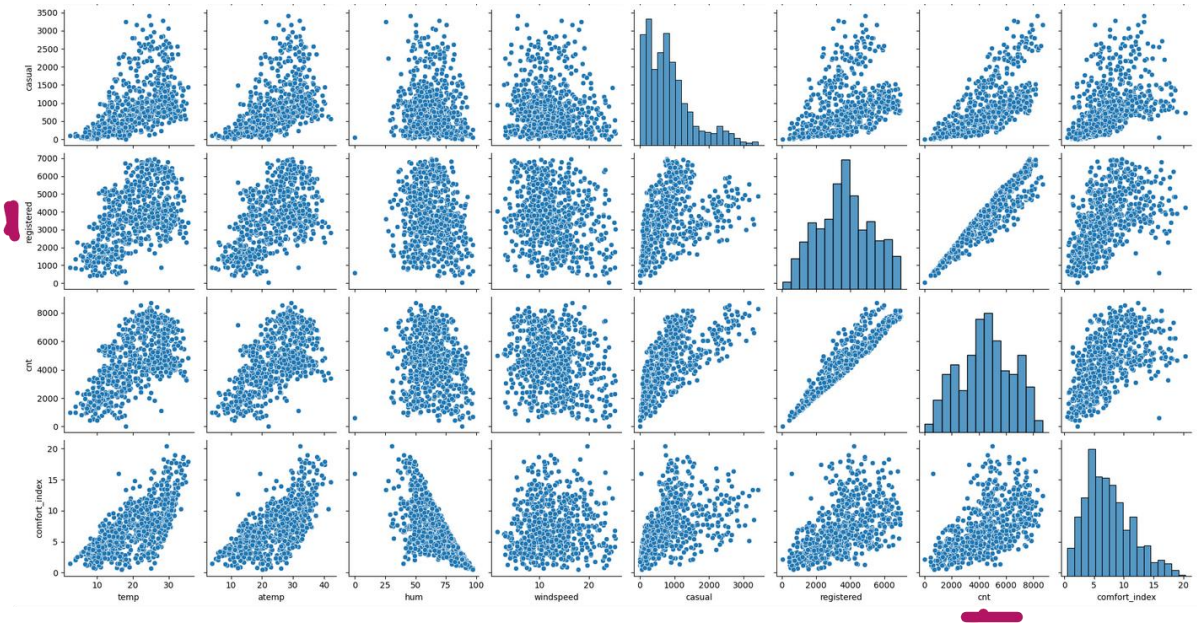
While converting, if we do not drop one of the columns it will lead to Multicollinearity, which is high correlation between independent variables. It will also be difficult in interpreting coefficients as it is challenging to understand the impact of each category variable on the dependent variable.

Suppose we have a Gender column with Male and Female values in it. Creating dummy variable will create two columns Gender_Male, Gender_Female for the same, and we can drop either of the 2 variables.

drop_first=True will drop the first column and by dropping the first category, we avoid redundant information and improve the model's interpretability and stability.

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)
Registered variable has the highest correlation with cnt.
After that casual, temp and atemp columns have higher correlation in the mentioned order.





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)

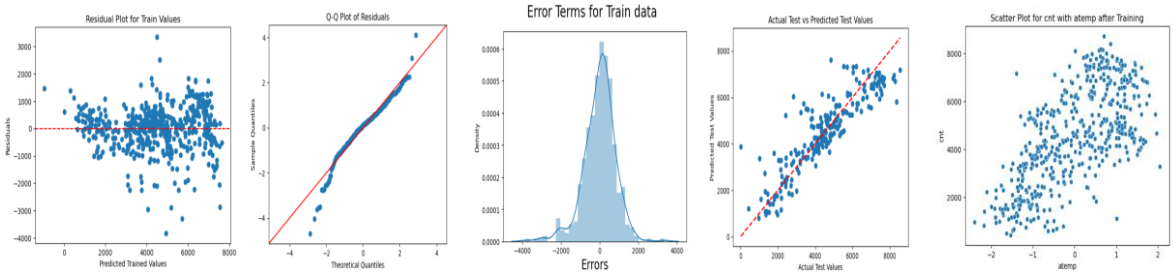
Linearity: Verify that the relationship between the predictors and the response variable is linear. The only numerical predictor variable (atemp) in the final multiple regression model has linear relationship with cnt. Check the scatter plot below for more details.

Independence: Check that the residuals are independent. Check below the residual plot which indicates the same.

Homoscedasticity: Ensure that the residuals have constant variance. You can plot the residuals against the fitted values and look for patterns, and homoscedasticity means there's no clear pattern. Please find below residual plot vs y-train predicted values.

Normality of Residuals: Verify that the residuals are normally distributed. This can be assessed using a Q-Q plot or a histogram of the residuals. Please find above Q-Q plot and hist plot for residuals.

Multicollinearity: The VIF (Variation Inflation Factor) summary indicates that there are no variables with VIF more than 5. Please refer the notebook model4 VIF details.



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)
As per the final model developed, we have 8 features which are used in the projection. The top 3 by correlation coefficient are -

1) yr_1 - Year 2019 (1043.7114) - Rental growth has increased significantly from 2018 to 2019.
2) atemp - Feeling Temperature (869.1712) - It has very high impact on the bike rentals as many people tend to rent bikes when the feeling temperature is good and not too cold or too hot. This suggests that people are more likely to rent bikes when the feeling temperature is good to the human body, which aligns with the expected seasonality.
3) season_spring (-414.8227) - Rentals decrease significantly in the spring probably because of rains.

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)

Linear Regression is a statistical method used to model the relationship between a dependent variable (y) and one or more independent variables (x). The goal is to find the best-fitting line that minimizes the difference between the predicted values and the actual values.

Simple linear regression model with one independent variable can be represented as:

$y = c + mx + e$ where:

y : Dependent variable (Target variable)
x : Independent variable (Predictor variable)
c : Intercept
m: Slope coefficient
e : Error term (residual)

Multiple Linear Regression: For multiple independent variables, the model extends to:

$y = c + m_1x_1 + m_2x_2 + ... + m_nx_n + e$

The two popular methods to fit the best-fit line in linear regression are

1) OLS (Ordinary Least Squares)
2) Gradient Descent

The goal is to minimize the Residual Sum of Squares, which is calculated as below:

a) Squaring each residual (to eliminate negative values).
b) Summing up all the squared residuals.
The lower the RSS, the better the model fits the data. A perfect fit would have an RSS of 0, meaning the model's predictions are exactly equal to the actual values.

Model Evaluation - Once the model is trained, it's essential to evaluate its performance. Common

evaluation metrics include:

R-squared: Represents the proportion of the variance in the dependent variable that is explained by the independent variables.
Mean Squared Error (MSE): Measures the average squared difference between the predicted and actual values.
Root Mean Squared Error (RMSE): The square root of MSE, providing a measure of error in the same units as the dependent variable.
Mean Absolute Error (MAE): Measures the average absolute difference between the predicted and actual values.

Assumptions of Linear Regression :

Linearity: The relationship between the predictors and the dependent variable should be linear. This means that changes in the dependent variable should be proportional to changes in the predictors.
Independence: Observations should be independent of each other. This is particularly crucial in time-series data where there is a risk of serial correlation.
Homoscedasticity: The residuals should have constant variance. If the variance changes (heteroscedasticity), it can lead to inefficient estimates and affect the model's reliability.
Normality of Residuals: Residuals should be normally distributed, especially important for hypothesis testing. This can be checked using a Q-Q plot or a histogram of residuals.
No Multicollinearity: Independent variables should not be highly correlated with each other. Multicollinearity can make it difficult to determine the individual effect of each predictor on the dependent variable.
The VIF (Variation Inflation Factor) can be used to detect Multicollinearity

---

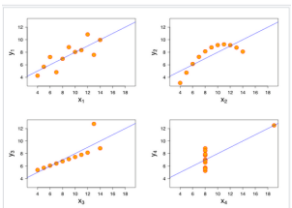
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a set of four datasets, each with 11 (x, y) points. These datasets were created by statistician Francis Anscombe to demonstrate the importance of visualizing data before analyzing it, and highlights how statistics alone can sometimes be misleading.

Key Points about Anscombe's Quartet:

Identical Summary Statistics: All four datasets have nearly identical summary statistics, including mean, variance, correlation, and regression line. This makes it seem like they represent the same underlying relationship.
Visually Distinct: When plotted, the four datasets reveal very different patterns:

Dataset plots from Wikipedia



Dataset 1: Appears to be a simple linear relationship when plotted.
Dataset 2: Shows a clear curve and a non-linear pattern.
Dataset 3: A perfect linear relationship with an outlier.
Dataset 4: Displays a vertical line with one distinct outlier that distorts the correlation.

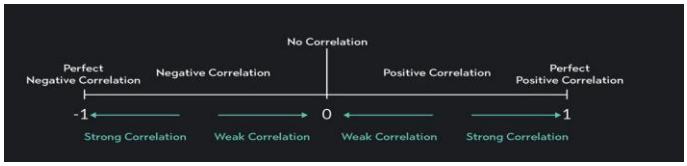
Key Takeaways:

Visualization Matters: Numerical summaries alone may not reveal the full picture of the data.
Outliers: Can significantly affect statistical measures and the fit of models.
Pattern Recognition: Visual inspection can help identify patterns and deviations that numerical analysis might miss.
Data Integrity: Reminds us to scrutinize data both visually and numerically to ensure accurate interpretations.

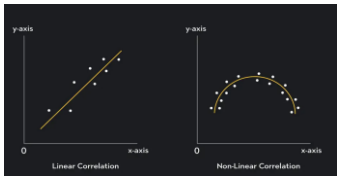
Anscombe's Quartet underscores the essential practice of graphing data. By looking at the plots, we gain insights that numbers might not reveal.

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)

1) The Pearson correlation coefficient is one of the most common methods for measuring correlation, and it measures a linear correlation's direction and magnitude.
2) A linear correlation is strictly positive or negative, whereas a non-linear correlation can change with the values of x and y.
3) Pearson's r ranges from -1 to 1, where -1 represents a perfect negative correlation, and 1 represents a perfect positive correlation.
4) a. Perfect Negative Correlation (r=-1)
   b. Negative Correlation (-1≤ r <0)
   c. Zero Correlation (r=0)
   d. Positive Correlation (0<r≤1)
   e. Perfect Positive Correlation (r=1)



5) Graph showing a linear correlation and a non-linear correlation



6) Scatter plots are a useful way of visualizing correlations
7) You can use the correlation when there is linear relation ship, variables are quantitative, and there is no outliers.

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)

Scaling adjusts the range of feature values in your dataset to ensure they are on a similar scale. It can improve the performance of machine learning algorithms

Scaling is performed for the below reasons:

1) Performance of the algorithm is improved
2) Ensures that no single feature dominates the model due to its magnitude.
3) Scaling data is more standardized.

Min-Max Scaling:

1. Scales data to a specific range, typically between 0 and 1

Standardisation Scaling:

1. Shifts the data to have a mean of 0 and a standard deviation of 1.
2. Assumes normal distribution.

The advantage of Standardization over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there are outliers.

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)

Variance Inflation Factor (VIF) value can become infinite when there is perfect multicollinearity in the dataset. The formula for calculating VIF is 1/(1-R Square)

R-Square becomes 1 when there is a perfect multicollinearity, and this makes the denominator as zero (0) , which makes the VIF as infinite. It is a clear indication that the predictors are redundant and one or more predictors need to be removed or adjusted.

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 (Quantile-Quantile) plot is used to assess if a dataset follows a specific distribution, most commonly a normal distribution. It compares the quantiles of the data to the quantiles of a theoretical distribution.

1) Perfect Fit: If the data follows the theoretical distribution, the points on the Q-Q plot will form a straight line.
2) Deviation from Normality: Deviations from the straight line indicate departures from normality.
3) S-shaped curve: The data might be skewed.
4) Tails deviating from the line: The data might have heavier or lighter tails than a normal distribution.

<u>Use and Importance in Linear Regression</u>

Assessing Normality of Residuals**:** In linear regression, there is an assumption that the residuals should be normalized. A Q-Q plot will help check this assumption.

Deviations: If the residuals deviate from the normality assumption, it can indicate potential issues with the model, such as the presence of outliers, skewness, or heteroscedasticity.
