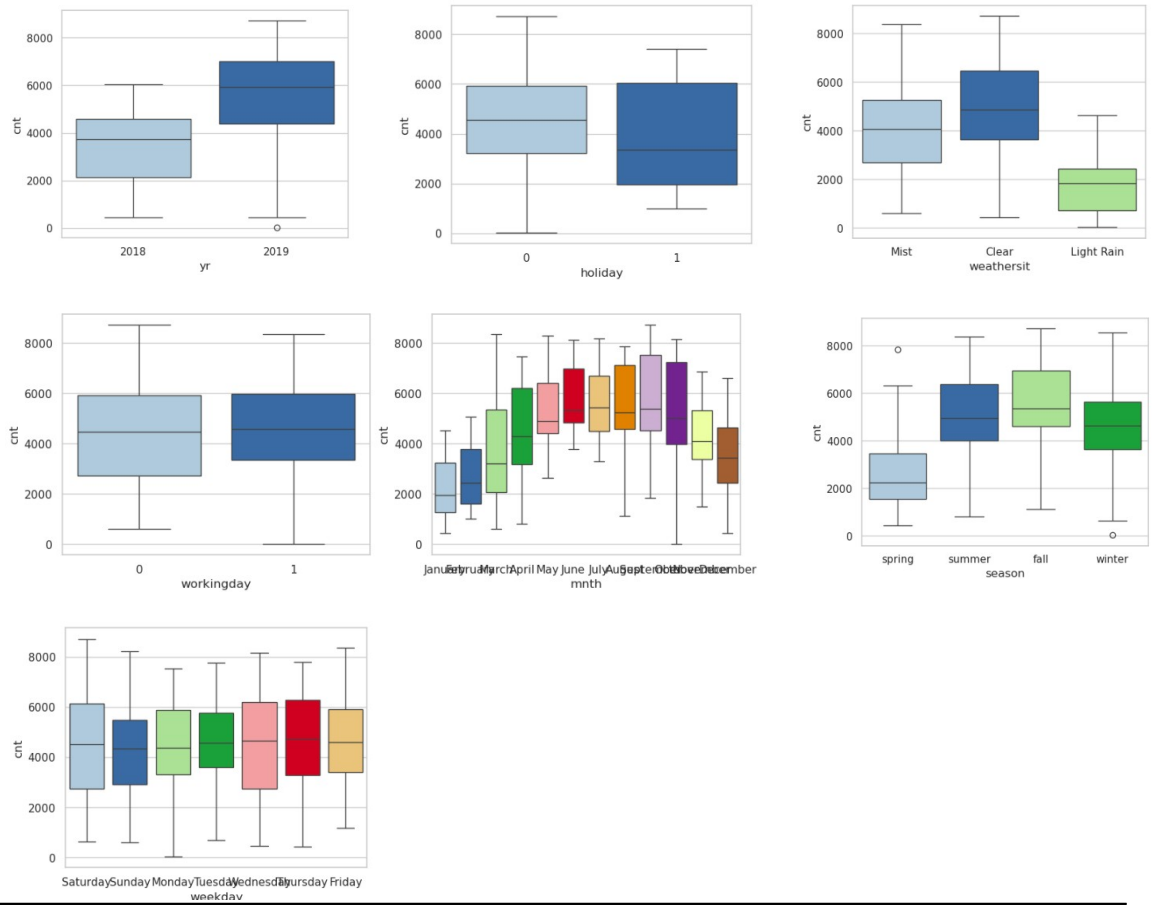
Assignment-based Subjective Questions

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)



Categorical variables are yr, mnth, season, workingday, weekday, weathersit have major effect on the dependent variable “cnt”. This variables are visualized using both box plot and bar plot.

As shown in above fig, we can see that how season, weather, working day, weekday, month influenced the bike rentals.

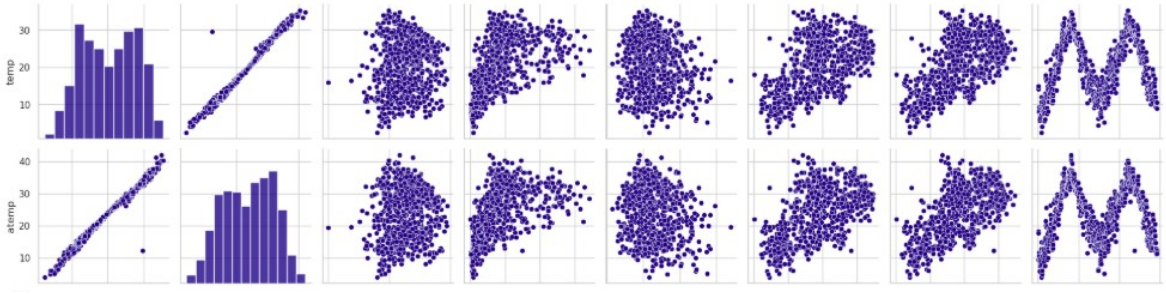
- **Season:** Different seasons likely influence bike rentals because of varying weather conditions and people's willingness to ride bikes. Bike Rentals are more during the fall season and then in summer.
- **Working days:** On working days, bike rentals may be influenced by commuting patterns, while non-working days may have more recreational rides.
- **Month:** Different months will have varying effects on rental counts, likely correlating with temperature, weather, and seasonality. For instance, bike rentals peak during warmer months (spring and summer) and drop during colder months (fall and winter).
- **Weather:** Weather conditions can have a strong impact on bike rentals. Clear weather might encourage more rentals, while rainy or snowy conditions could reduce them.
- **Weekday:** Weekdays vs. weekends typically show different bike rental patterns due to working schedules. Weekdays may see lower rentals compared to weekends when more people might be free to ride.

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)

When creating dummy variables it is important to avoid **multicollinearity** while using drop_first=True. By dropping one category, we eliminate this issue, making the dropped category the reference and simplifying model interpretation. Without this, including all categories leads to perfect correlation between the variables, causing redundancy and unstable model estimates. Ensuring model to be both efficient and reliable.

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)

Looking at the pair-plot among the numerical variables, temperature 'temp' and 'atemp' has the highest correlation with the target variable 'cnt'. Bike rentals are observed at high temperatures.



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)

To validate the assumptions of Linear Regression after bulding the model on the training set, first check the **Linearity** by plotting residuals against fitted values to ensure there are no patterns. Then check residuals vs. time for autocorrelation. Then plot residuals vs. fitted values to ensure constant variance. Check if residuals are normally distributed or not. Check for **multicollinearity** by examining the **Variance Inflation Factor (VIF)** or correlation matrix to ensure predictors aren't highly correlated. It ensures a reliable and well-fitting regression model.

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)

Based on the final model, temp, yr and June are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

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 supervised machine learning method which finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. It establishes a relationship between a dependent variable (target) and one or more independent variables (predictors) using a straight line. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

When there's one independent variable then it's simple linear regression such as house prices predictions acc. to the area.

When there's more than one independent variable then it's multiple linear regression such as demand of bike rentals acc. to the season, working days, temperature or more.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. If the dependent variable expands on the Y-axis and the independent variable progress on X-axis then it's positive linear relationship. If the dependent variable decreases on the Y-axis while the independent variable increases on the X-axis then its negative linear relationship.

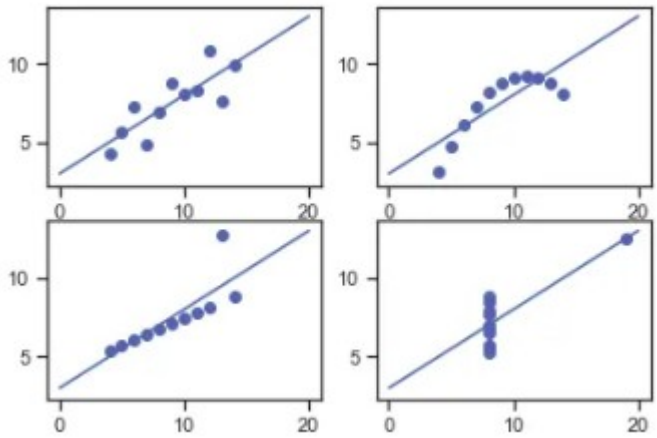
The goal of the linear regression algorithm is to find the best values to find the best fit line and the best fit line should have the least error.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

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 shows the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.



- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.

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 correlation coefficient is a statistical measure that evaluates the strength and direction of the relationship between two continuous variables. It is considered the most effective method for assessing associations due to its reliance on covariance. This coefficient not only reveals the magnitude of the correlation but also its direction.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}}$$

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 is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like F-statistic, p-values, R-squared, etc.

Normalized scaling brings all of the data in the range of 0 and 1. It loses some of the info in the data set during outliers.

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.

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>

VIF formula is: 'i' refers to the ith variable.

$$VIF_i = \frac{1}{1-R_i^2}$$

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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>

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.