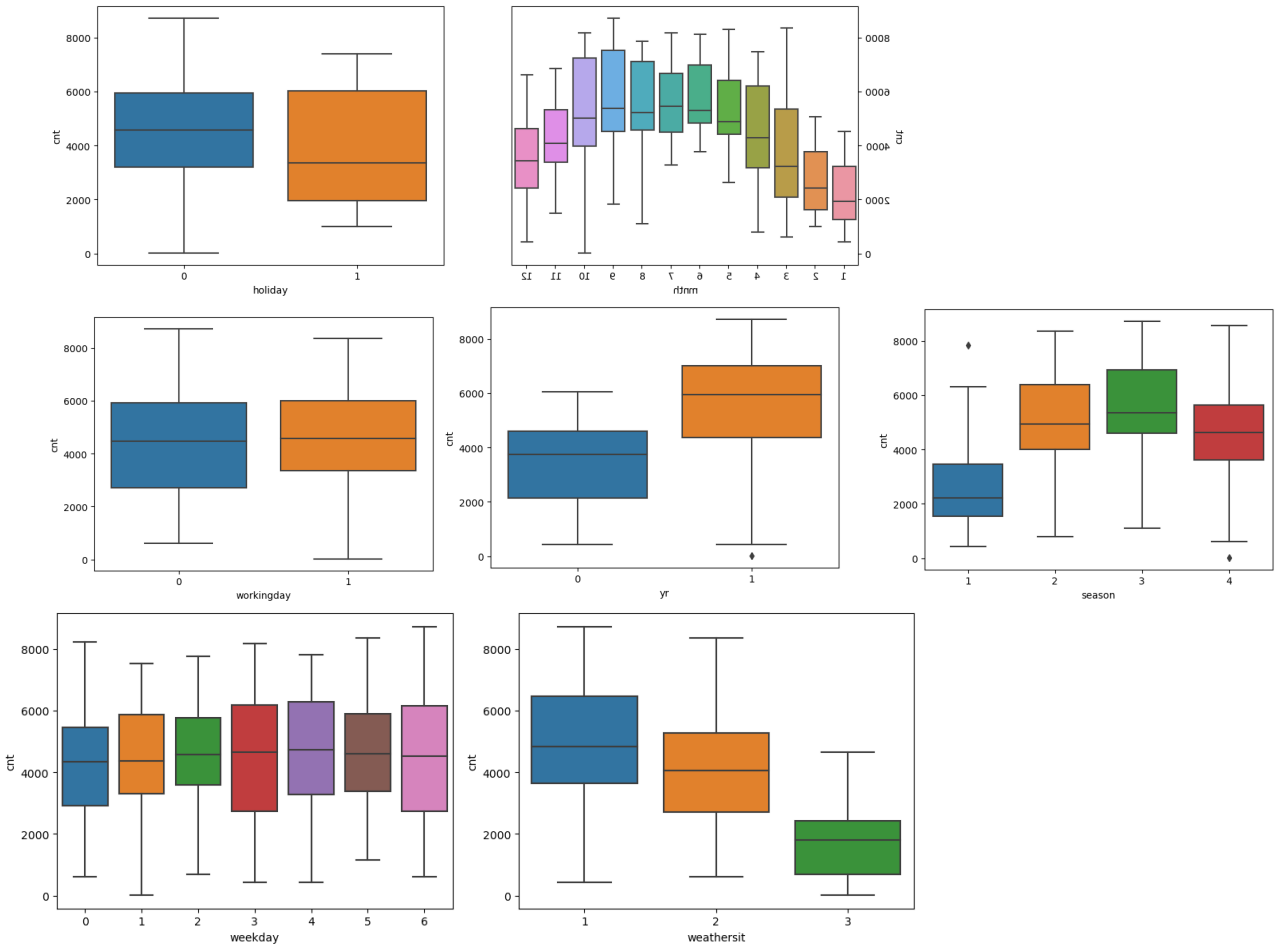
Assignment-based Subjective Questions

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

Ans.

- Number of bikes hired was far greater in 2019 as compared to 2018.
- The demand for bikes is highest during clear weather, remains relatively steady during misty weather, but drops significantly during light snow.
- The median values of bikes hired on weekdays are almost the same. This implies that, on average, the demand for bikes is consistent throughout the weekdays. The medians being similar suggest a stable and steady demand during the weekdays but The maximum number of bikes hired occurs on weekday_6
- Overall distribution of bikes rented on holidays is lower than the working days.



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

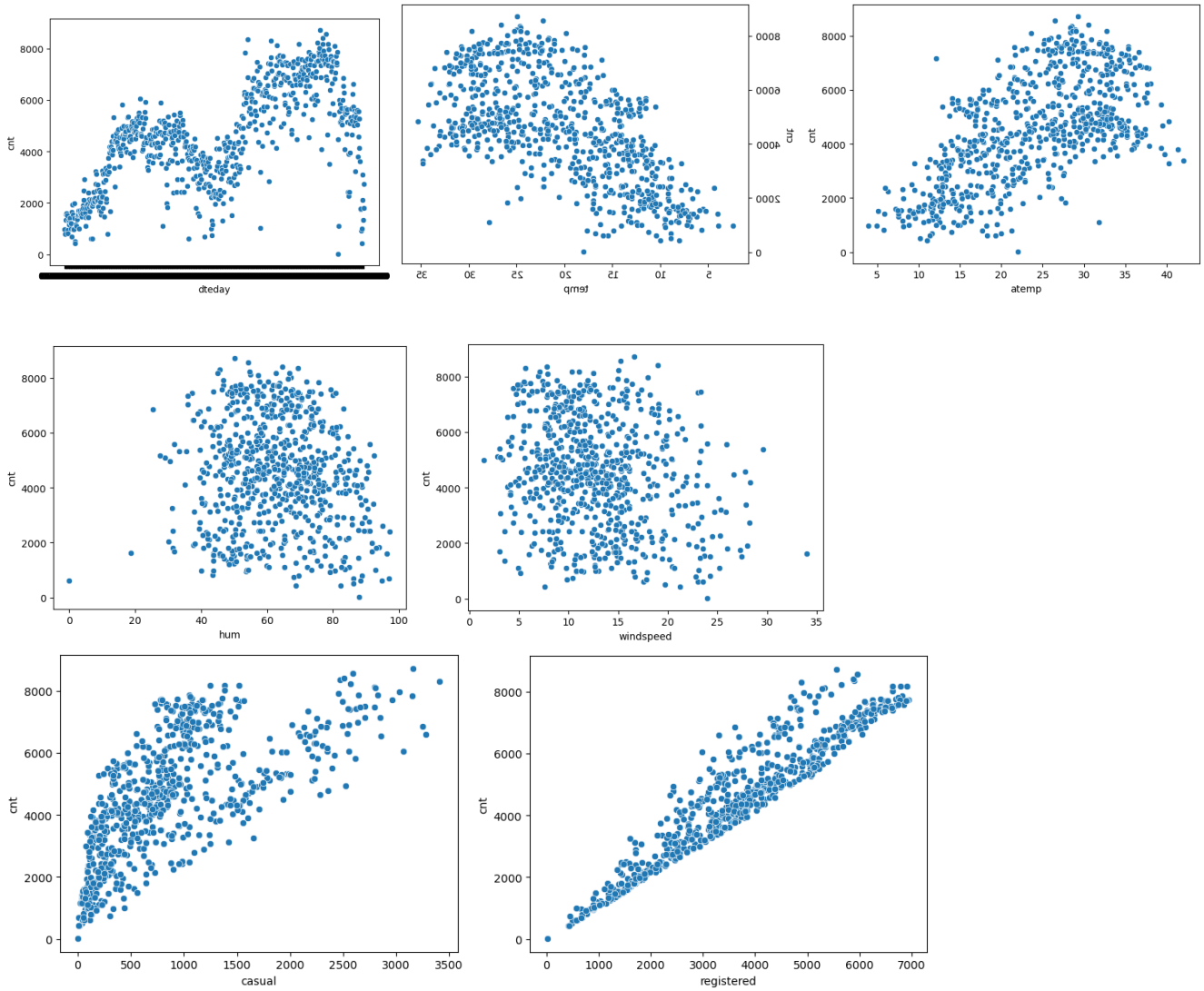
Ans.

- When using drop_first=True during the creation of dummy variables, the first categorical variable is indeed dropped, and (n-1) categorical variables are used.

- "drop_first=True" is used to address issues related to multicollinearity and to simplify the model by using fewer dummy variables while still preserving the necessary information about the categorical variable.

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

Ans. 'atemp' has highest correlation with the target variable named as cnt.



4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans.

The following assumptions are considered:

- Linear Relationship between dependent and independent variables::
  For this we can create a scatter plot between the independent and dependent variables.

- Normality of the residuals:
  We can plot the residual values using qq-plot. A straight line will show the normality. Else, distplot can also be used.

- No or little multicollinearity: We can check the VIF values of the independent variables.

- Homoscedasticity:
  It is to check if error is constant along values of the dependent variable. A scatter plot and a constant line from 0 in y-axis is drawn and the deviation of error from zero-line is checked.

- All independent variables are uncorrelated with error terms: Scatter plot is drawn between independent and residuals.

- Observations of the error terms are uncorrelated with each other:
  It is to check whether there is a correlation inside the observations of the error term. A line graph of residuals is plotted.

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

Ans.

- Temperature
- Year-2019
- Weekday

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression is a statistical method used to model the relationship between a dependent variable(target) and one or more independent variable(Predict).

The standard equation that explains the model is Y=mx+c

Assumptions:

- Y and X are linearly related
- Errors/Residuals are normally distributed and independent of each other
- Errors should have constant variance (homoscedasticity)

The idea is to fit the line on the datapoints and understand if the line fits the dataset "significantly". This is achieved used NULL Hypothesis on the beta coefficient. Usually P-value method is used to find the significance of the fit and therefore the variables.

Features are also scaled using Min-Max Scaling to ensure that coefficients of multiple dependent variables are comparable and portray the right information.

2. Explain the Anscombe's quartet in detail

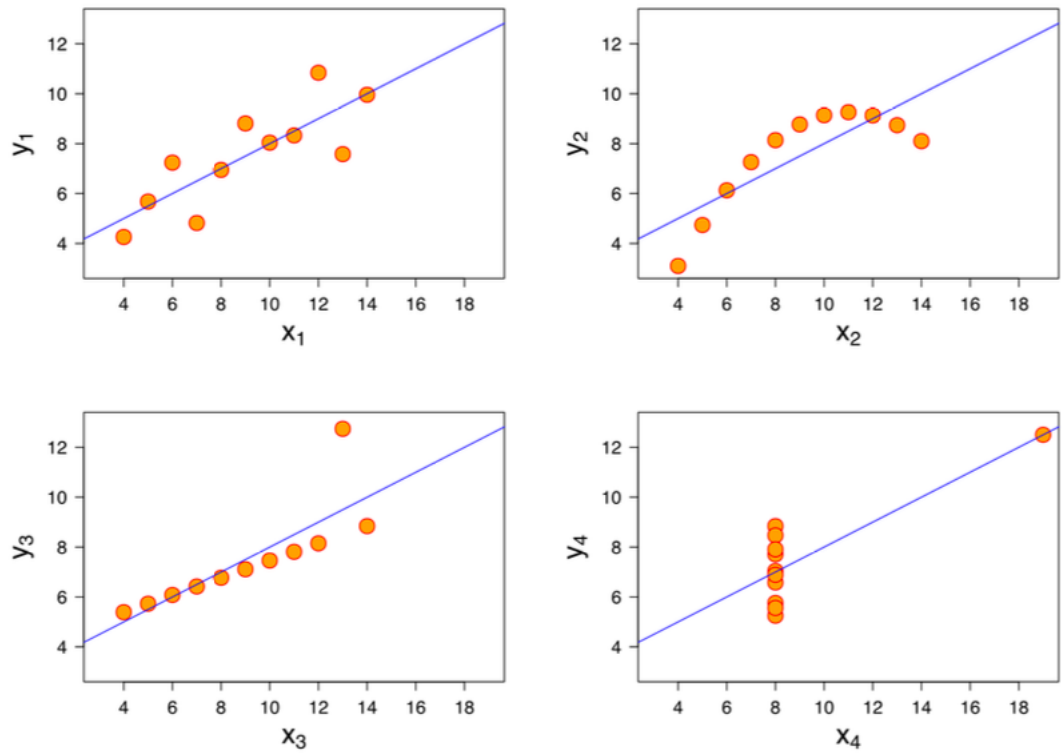
Ans. It consists of four graphs which have similar descriptive statistics like mean, variance, standard deviation etc. but if we plot the graphs for checking the distribution of the data, we see it is quite different.

Dataset:

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Graphical Representation of Data:



Observations:

Graph I:
A simple linear relationship with a strong positive correlation.

Graph II:
Similar to Dataset I but with one outlier that drastically affects the linear regression line.

Graph III:
Two distinct groups, each with a linear relationship.

Graph IV:
No clear linear relationship when plotted.

3. What is Pearson's R?

Ans. Pearson correlation coefficient known as Pearson's r, the Pearson product-moment correlation coefficient, is a measure of linear correlation between two sets of data.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is the process of resetting the variables on a comparable scale especially when lot of independent variables are involved.
Scaling is performed mainly for the below reasons:

- Ease of interpretation
- Faster convergence for gradient descent method

Standardized Scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.

Normalized Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

Ans. VIF = 1/(1-r^2)
Now, a VIF of infinity means r^2 is 1. It means correlation coefficient can be +1 or -1. This suggests that there is perfect correlation between the dependent and independent variable.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Ans.

A Q-Q plot is a valuable tool in linear regression for checking the normality assumption and identifying potential issues with the distribution of residuals. It provides a visual assessment of how well the observed data aligns with the theoretical distribution, helping researchers make informed decisions about the validity of their regression model.