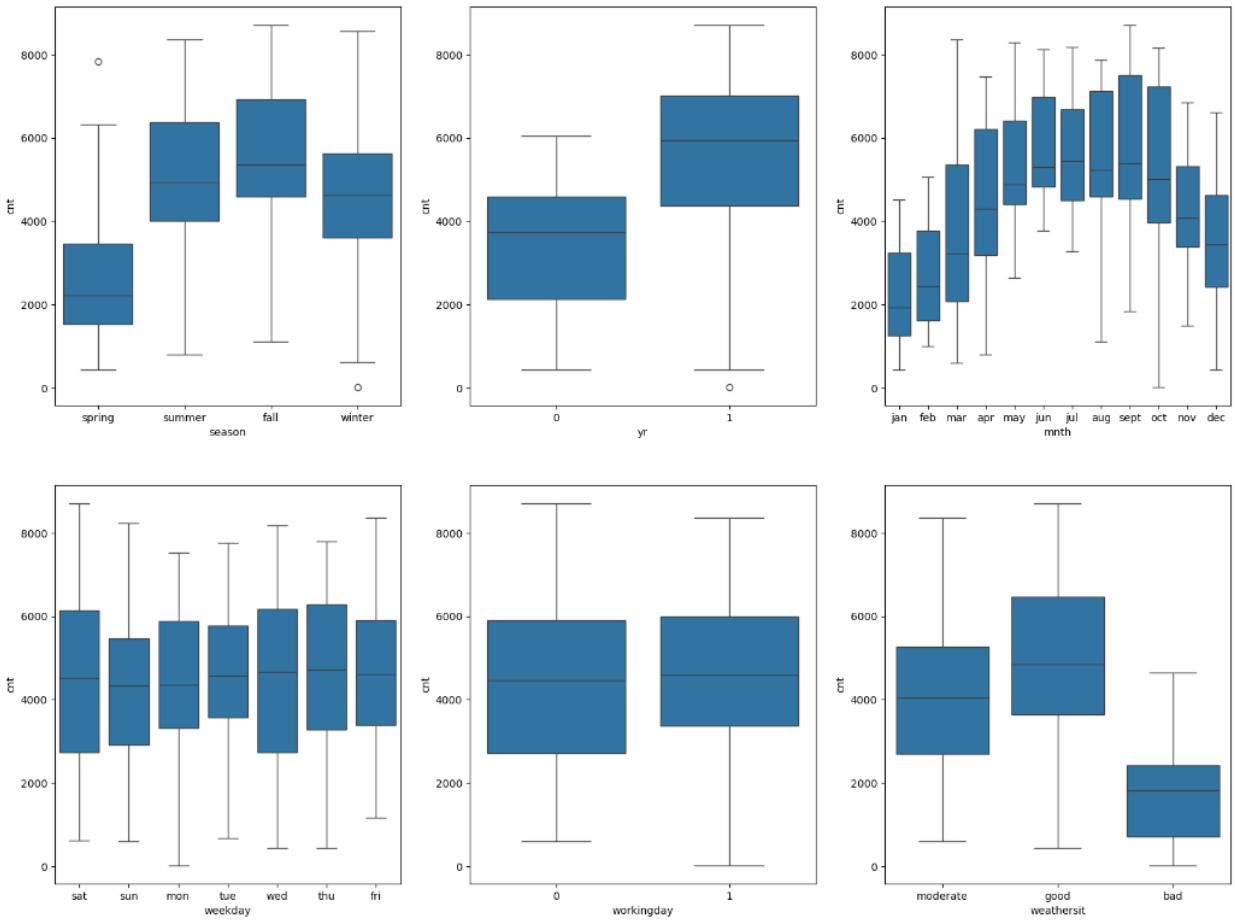
Assignment-based Subjective Questions

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

Answer: There are multiple categorical variables namely season,mnth,yr,weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'. I have created box plot and bar plot to understand this. The below fig shows the correlation among the same



- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of aug, sep and oct. Trend increased starting of the year till mid of the year and then flattened for couple of month before started decreasing towards the end of year.
- Good weathersit attracted more bookings.

- Wed, thu, Fir, and Sat have a greater number of bookings.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted more number of booking from the previous year, which shows business is growing.

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

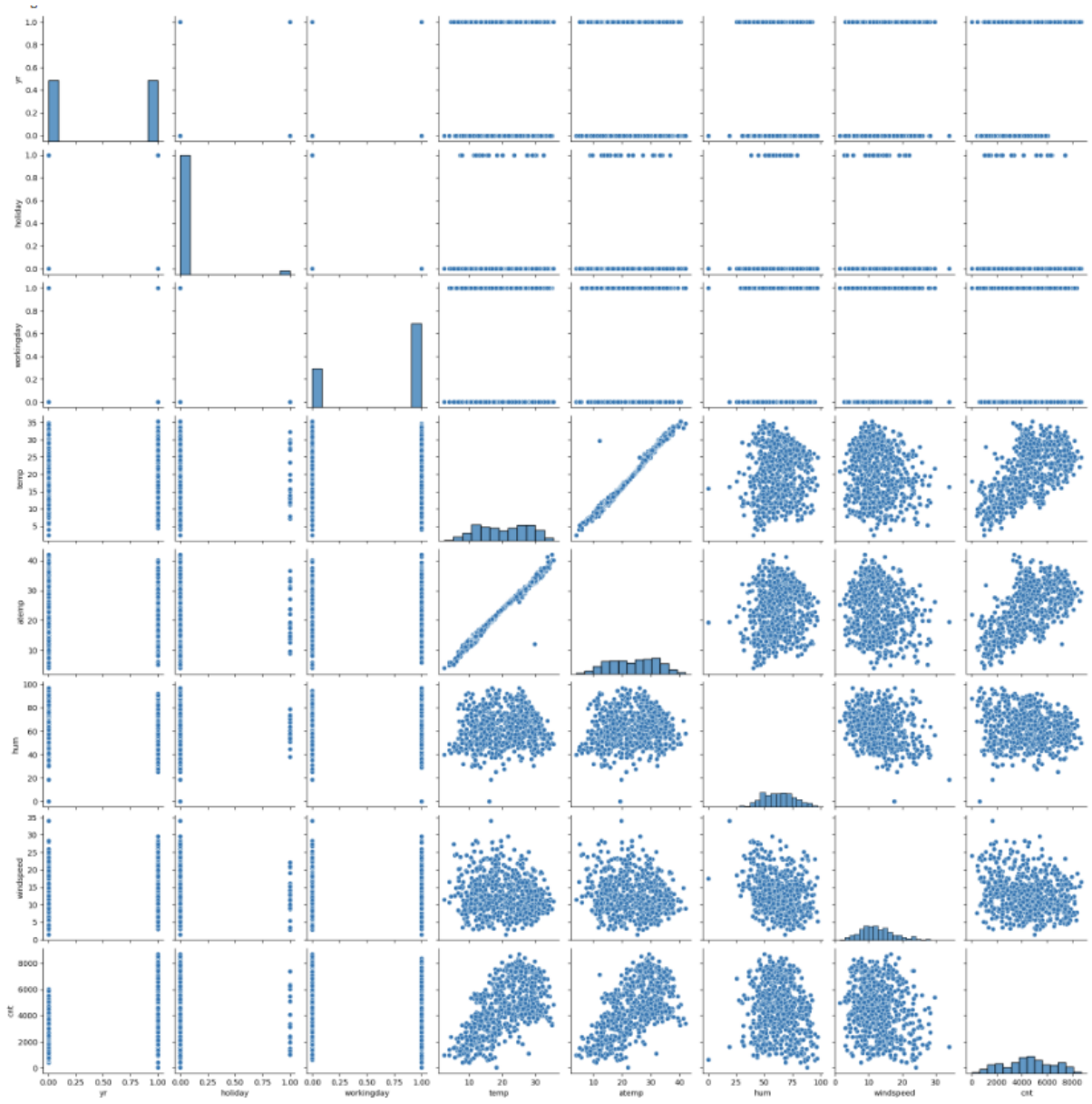
Answer: Dummy variables are used to represent categories as numbers (0 or 1). If a category has 'n' possible values, we only need to create 'n-1' columns to represent them. This is because the last category can be figured out from the rest. By using drop_first=True, we remove one of the dummy columns, which helps avoid extra correlation between the columns and makes the model work better.

Let's say we have 3 types of values in Categorical column and we want to create dummy
variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

Answer:

The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.



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

Answer: Linear Regression models are validated based on
Normality of error terms- Error terms should be normally distributed
Linear relationship validation- Linearity should be visible among variables
Homoscedasticity- There should be no visible pattern in residual values.
Multicollinearity check- There should be insignificant multicollinearity among variables.
Independence of residuals- No auto-correlation

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

Answer: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- Season

- months(July, September)

- Year

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Answer:
   Linear Regression is a statistical method used to analyze the relationship between one dependent variable and one or more independent variables. This method shows how the dependent variable changes when the independent variable(s) change.

   Basic Idea: Linear regression assumes a linear relationship, meaning that as one or more independent variables increase or decrease, the dependent variable also increases or decreases in a predictable manner.
   The mathematical formula for linear regression is: $Y=mX+c$
   Where:
   - **Y** is the dependent variable we are trying to predict.
   - **X** is the independent variable used for prediction.
   - **m** is the slope of the line (it shows the effect of X on Y).
   - **c** is the intercept (if X is 0, then Y will be equal to c).

   Positive vs. Negative Relationship:

   - **Positive Linear Relationship**: Both independent and dependent variables increase together.
   - **Negative Linear Relationship**: The independent variable increases, but the dependent variable decreases.

   Types of Linear Regression:

   - **Simple Linear Regression**: When there is one independent variable.
   - **Multiple Linear Regression**: When there are multiple independent variables.

Key Assumptions of Linear Regression:

1. **Multicollinearity**: Independent variables should not be highly related to each other.
2. **Autocorrelation**: There should be no dependency between the errors (residuals).
3. **Linear Relationship**: The relationship between the dependent and independent variables should be linear.
4. **Normal Distribution of Errors**: The error terms (difference between actual and predicted values) should follow a normal distribution.
5. **Homoscedasticity**: The residuals (errors) should not show any specific pattern.

2. Explain the Anscombe's quartet in detail.

Answer:

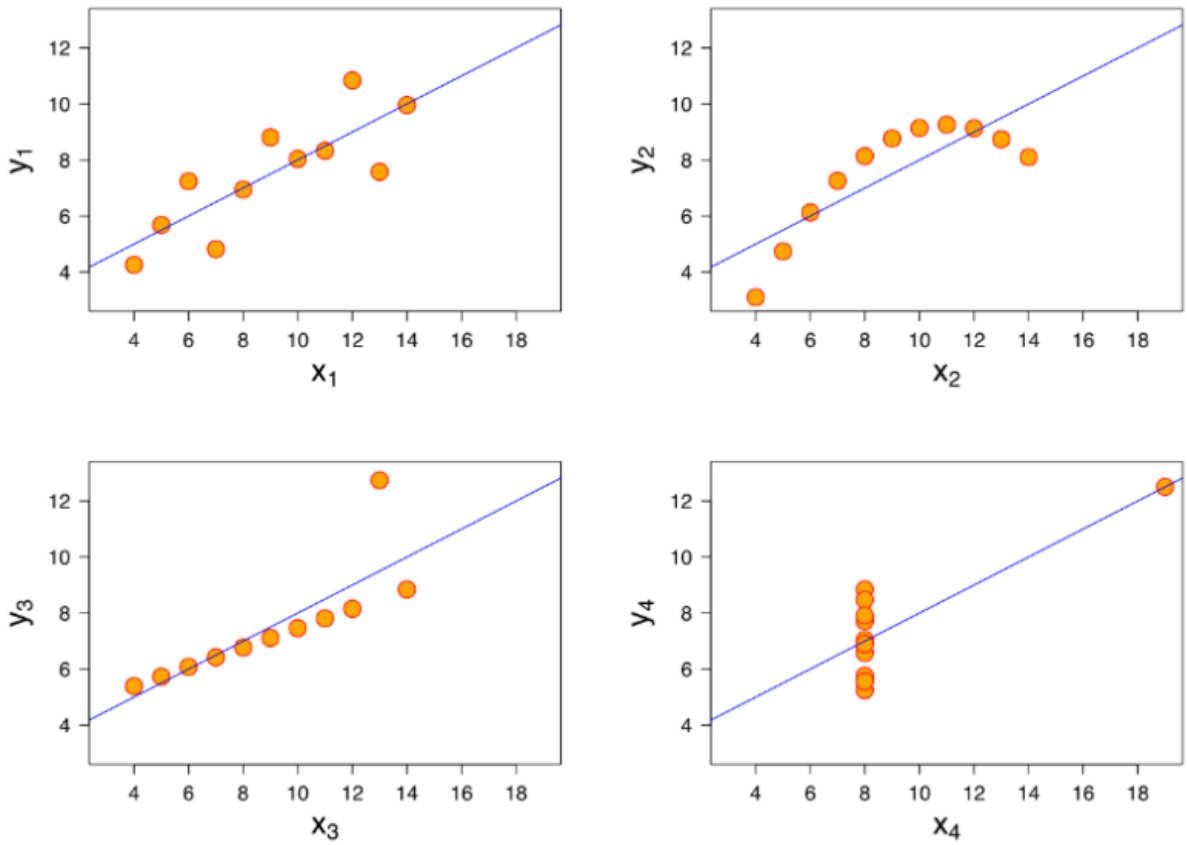
Anscombe's Quartet is a group of four datasets created by statistician Francis Anscombe to illustrate the importance of **visualizing data** before interpreting it. Each dataset in the quartet contains eleven (x, y) pairs, and despite having almost identical summary statistics, they look very different when graphed.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Key Features of Anscombe's Quartet:

- **Similar Summary Statistics**: Each dataset shares the same basic statistical properties:
  - Mean of x: 9
  - Mean of y: 7.50
  - Variance of x: 11
  - Variance of y: 4.13
  - Correlation between x and y: 0.816
  - Same linear regression line for all datasets

Despite these identical numbers, each dataset tells a **very different story** when plotted graphically.



Differences Between the Datasets:

1. **Dataset I**: Appears to have a well-fitting linear relationship. This dataset behaves as expected, where the data points fit a straight line.
2. **Dataset II**: Although the summary statistics suggest a linear trend, the data is non-linear. It forms a curved pattern, showing that the dataset is not suitable for a simple linear model.
3. **Dataset III**: This dataset is mostly linear, but one extreme outlier skews the regression line. The outlier significantly affects the accuracy of the model.
4. **Dataset IV**: All the data points are nearly constant except for one outlier, which gives the illusion of a strong correlation between x and y, even though the data doesn't form a meaningful pattern.

The Lesson:

Anscombe's Quartet highlights the **importance of data visualization** in understanding the true nature of a dataset. Summary statistics like means, variances, and correlations don't always tell the full story. By plotting the data, one can detect outliers, non-linear trends, or unusual patterns that may not be visible from the numbers alone.

The quartet is a reminder that relying solely on summary statistics without visualizing the data can lead to incorrect conclusions.

3. What is Pearson's R?

Answer:

Pearson's R, also known as the Pearson product-moment correlation coefficient (PPMCC), is a statistic used to measure the strength and direction of the linear relationship between two variables.

Key Points about Pearson's R:

- **Range**: Pearson's R ranges from -1 to 1.
  - A value of **1** indicates a perfect positive linear relationship.
  - A value of **-1** indicates a perfect negative linear relationship.
  - A value of **0** means no linear relationship exists between the two variables.
- **Linear Relationship**: Pearson's R specifically measures how well the relationship between two variables can be described with a straight line.
- **Interpretation**:
  - **Positive correlation** (r > 0): As one variable increases, the other variable also increases.
  - **Negative correlation** (r < 0): As one variable increases, the other variable decreases.
  - **No correlation** (r ≈ 0): There is no consistent linear relationship between the two variables.

Formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:

- **Xi and Yi -** These are the individual data points of the two variables x and y. i is the index of each observation, and n is the total number of observations.
- **Xbar and ybar -** These are the mean (average) values of the variables x and y, respectively.

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

Answer: Scaling refers to adjusting the range or values of features in the dataset to ensure they are on a consistent scale. This is important because machine learning models can be influenced by differences in the magnitude of features, especially when some features have much larger values than others.

Difference Between Normalized Scaling and Standardized Scaling

Feature	Normalized Scaling	Standardized Scaling
Definition	Scales features using min and max values	Scales features using mean and standard deviation
Range	Scales values between (0,1) or (-1,1)	Centers data around zero mean and unit variance
Effect of Outliers	Affected by outliers	Less affected by outliers
Use Case	Used when the scale of features is different	Used when data follows a normal distribution
Other Name	Min-Max Scaling	Z-Score Normalization
Distribution Assumption	No assumption about distribution	Assumes normal distribution

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

Answer: When the Variance Inflation Factor (VIF) is infinite, it indicates a perfect correlation between two independent variables. This happens because the R-squared value is 1, making the term 1/(1-R2) approach infinity. In such cases, one of the perfectly correlated variables should be removed from the dataset to address the multicollinearity issue.

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

Answer: A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. It helps assess whether a data set follows a specific theoretical distribution, such as Normal, Exponential, or Uniform.

Use of Q-Q Plot:

- **Comparing Distributions:** A Q-Q plot can compare two distributions to see if they are similar. If the data sets have similar distributions, the plot will be approximately linear.
- **Assessing Distribution Fit:** It helps determine if a data set follows a theoretical distribution by comparing its quantiles with those of the theoretical distribution.

Importance in Linear Regression:

- **Checking Normality:** In linear regression, residuals should ideally be normally distributed for valid statistical tests and confidence intervals. A Q-Q plot of residuals helps verify this normality assumption.
- **Comparing Datasets:** When using training and test datasets, a Q-Q plot can confirm if both sets come from populations with the same distribution, ensuring consistent data characteristics across sets.

Advantages:

- **Sample Size:** Q-Q plots can be used with various sample sizes.
- **Distributional Insights:** They reveal shifts in location and scale, changes in symmetry, and the presence of outliers.

In summary, Q-Q plots are essential for validating distribution assumptions in regression and comparing datasets to ensure consistency in analysis.