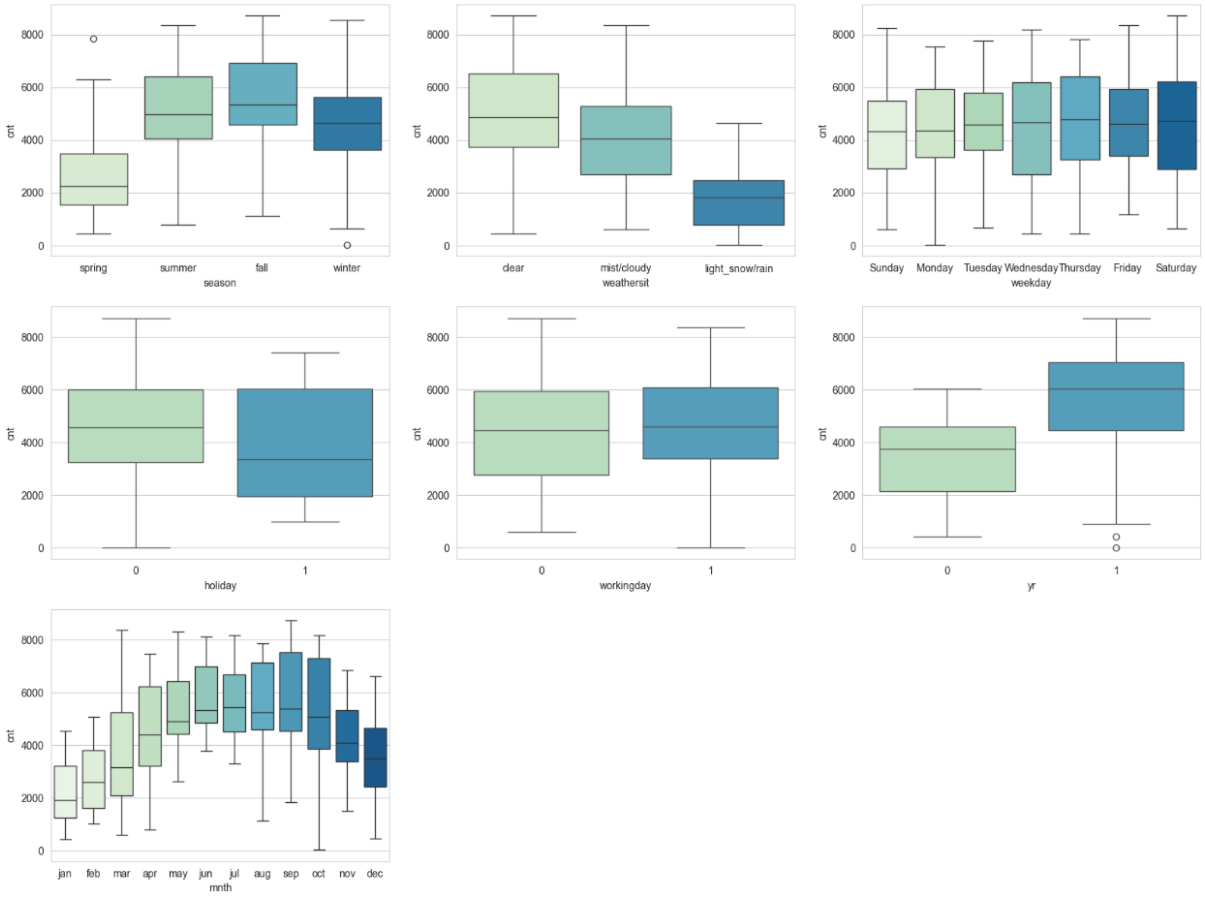
Assignment-based Subjective Questions

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)

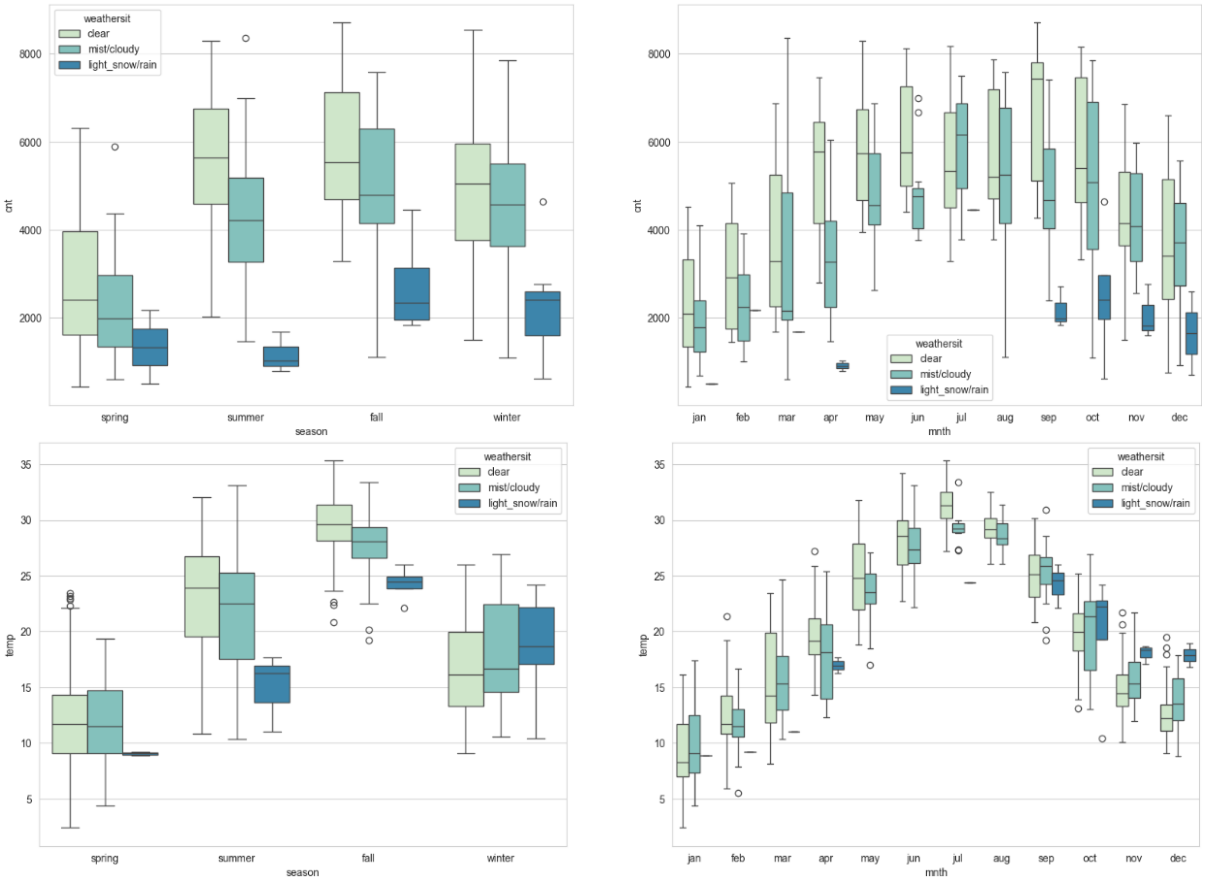
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Analysis and inferences of Categorical variables are as follows:



Inferences:

- There are few outliers for cnt values observed in the dataset. For example, for spring and winter season and for year 2019.
- The bike rental is highest in the fall, followed by summer, winter and spring.
- Weather plays an important role. Bike rental is highest in clear weather and decreases as the weather worsen.
- Day of the week doesnt seem to make much of a difference.
- Holidays seems to have less demand
- workingday seems to have very less impact.
- Business have grown significantly over the year from 2018 to 2019.
- Bike rental distribution in months is similar to seasons.



Inferences:

- Weather is generally clear or cloudy from January till August
- Generally, there is no rain or snow in May, June, August.
- Average temperature is highest in the fall and lowest in the spring. July having the highest temperature.
- Average number of rentals per day is also highest in the fall and least in the spring. This is displaying the high correlation between temp and cnt.

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)

Using *drop_first=True* ensures that the dummy variables remain independent, preventing multicollinearity issues. When *drop_first=True* is used, the omitted category serves as the baseline. The coefficients of the other dummy variables represent the difference in the dependent variable between that category and the baseline category.
For example, the Categorical variables season has four categories as follows:

Season	Spring	Summer	Fall	Winter
Spring	1	0	0	0
Summer	0	1	0	0
Fall	0	0	1	0

Winter	0	0	0	1

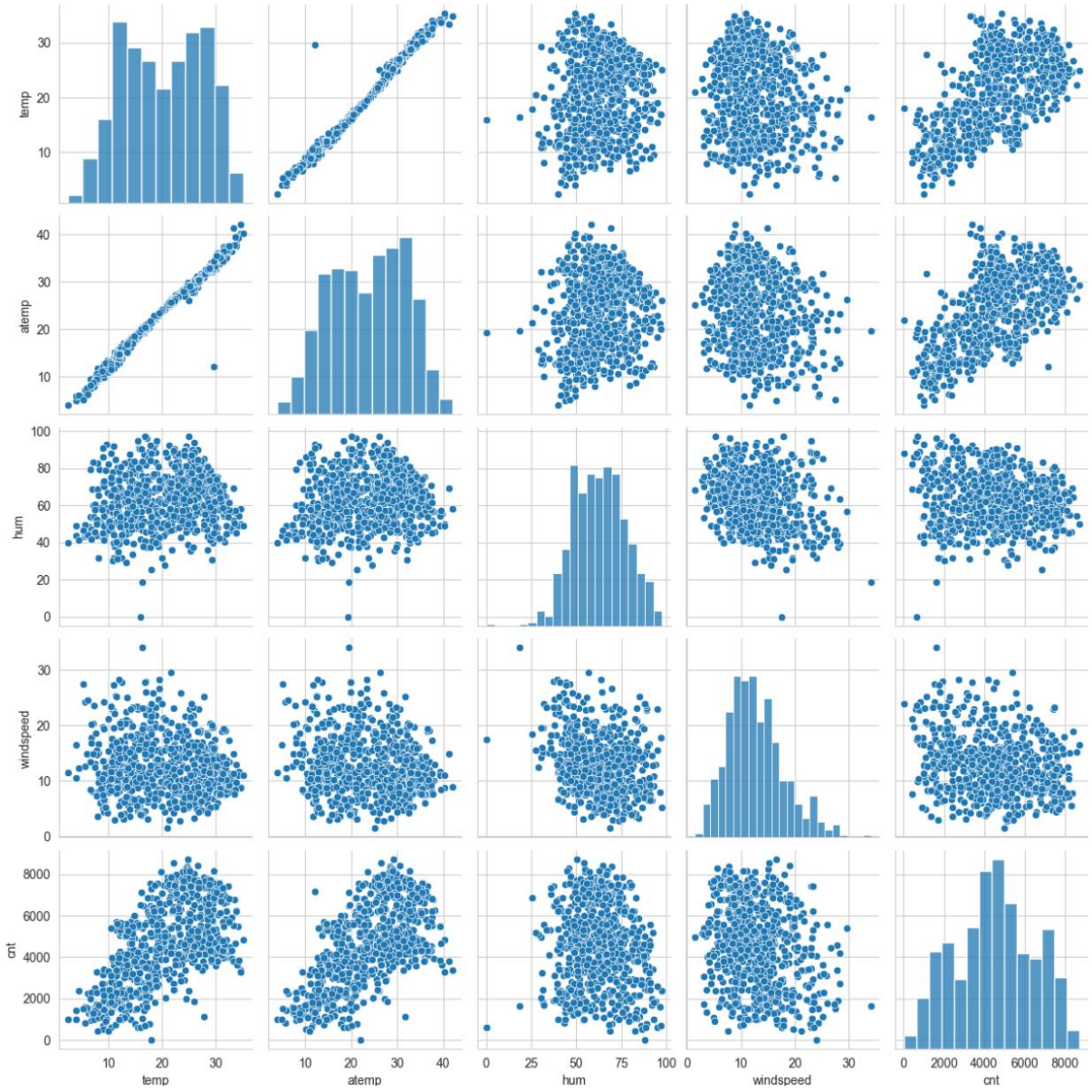
If four dummy variables are used, we have multicollinearity because the fourth can be determined if any three is known.
When we use *drop_first=True*, one dummy variable is dropped (example: fall in this case):

Season	Spring	Summer	Winter
Spring	1	0	0
Summer	0	1	0
Fall	0	0	0
Winter	0	0	1

Now the Linear regression model can still interpret the 'fall' variable when the other three variables are '0'. Thus, multicollinearity is prevented.

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)

The Pair-plot for continuous variables is as follows:

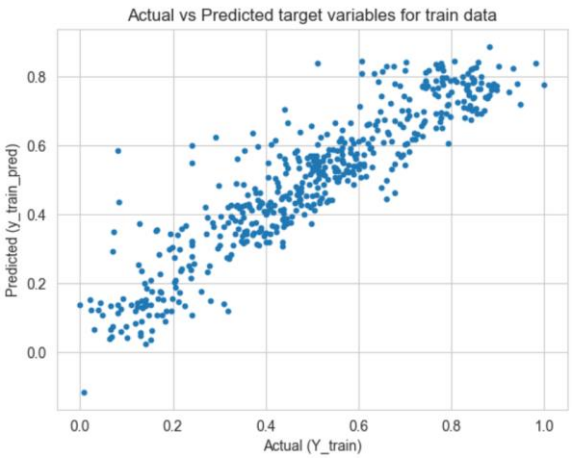


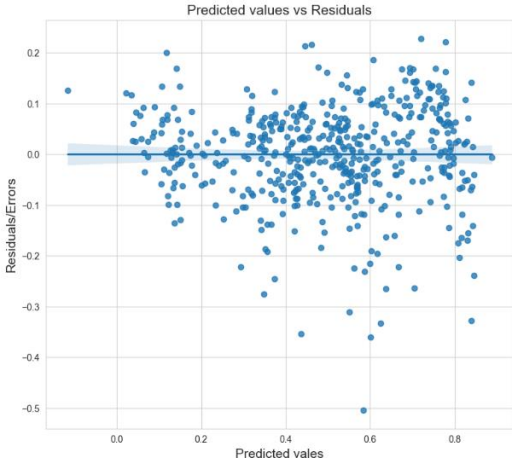
Looking at the pair-plot it is evident that `temp` and `atemp` has similar and highest correlation with the target variable `cnt`. But, since we are dropping the `atemp` to avoid multi collinearity, we can say that `temp` has highest correlation with the target variable.

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)

The assumptions of Linear Regression model and their validations are as follows:

1. **Linearity**





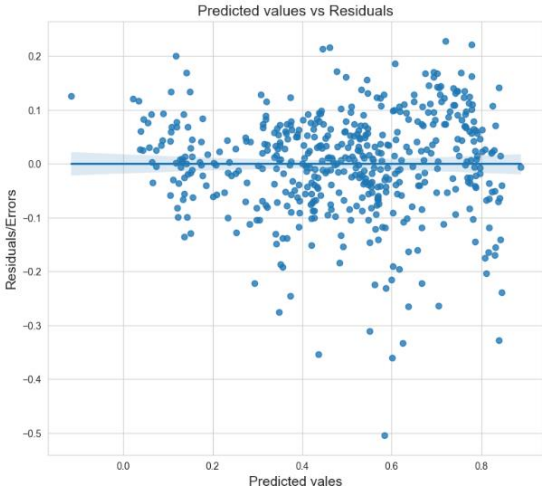
- Scatter plot of actual vs predicted values for target variable follows a linear pattern.
- In Predicted values vs residual plot, residuals are randomly scattered around zero without any clear pattern.

2. No Multicollinearity

```
------------------- VIF results -------------------
                     Features   VIF
3              season_winter  2.55
1                       temp  2.44
0                         yr  2.08
5                   mnth_nov  1.86
8      weathersit_mist/cloudy  1.54
4                   mnth_dec  1.43
2              season_spring  1.27
6                   mnth_sep  1.18
7  weathersit_light_snow/rain  1.06
```

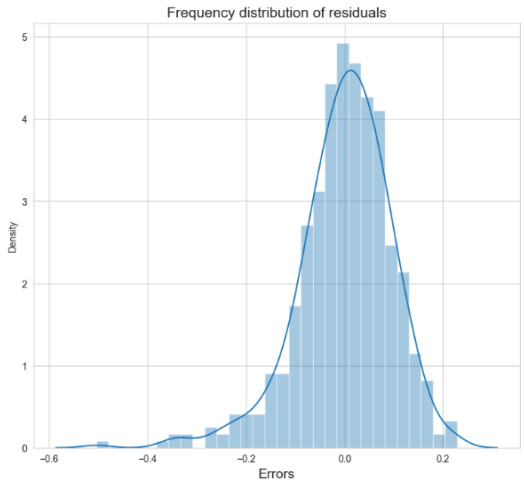
- The VIF values of the independent variables of the final model is <5.

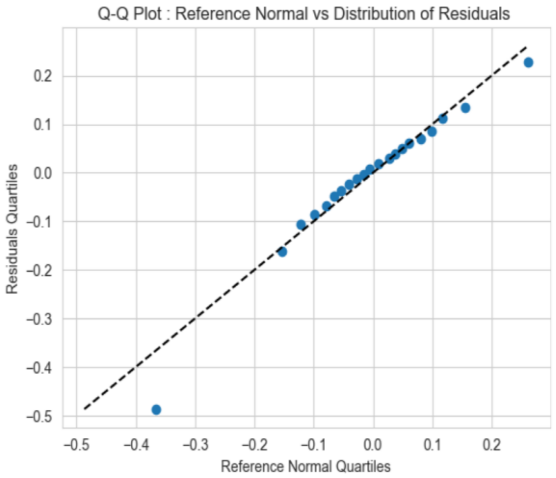
3. **Homoscedasticity:** Constant variance of Residuals



- In residual vs Predicted values plot, the residuals are randomly scattered around zero and the variance looks constand across the range.

4. Normality of Residuals





- The residuals follow a near normal distribution.

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)

	coef	std err	t	P>|t|
const	0.1880	0.023	8.159	0.000
yr	0.2447	0.009	28.064	0.000
temp	0.4152	0.030	13.755	0.000
season_spring	-0.1245	0.016	-7.587	0.000
season_winter	0.0958	0.015	6.379	0.000
mnth_dec	-0.0571	0.018	-3.110	0.002
mnth_nov	-0.0772	0.019	-3.963	0.000
mnth_sep	0.0528	0.016	3.386	0.001
weathersit_light_snow/rain	-0.2651	0.028	-9.609	0.000
weathersit_mist/cloudy	-0.0823	0.009	-9.023	0.000

From the final model, top three features contributing significantly towards explaining the demand of shared bikes are:

1. temp: positive correlation
2. yr: positive correlation
3. weathersit-Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds: negative correlation

---

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 algorithm is a statistical method to understand and predict a continuous dependent variable (target) based on one or more independent variables (features). It assumes a linear relationship between dependent and independent variables.

A linear regression model can be formulated using the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Where:

Y = dependent or target variable

$X_1, X_2, \ldots, X_n$ are independent variables

$\beta_1, \beta_2, \ldots, \beta_n$ are coefficients representing the relationship between each independent variable and the dependent variable

β_0 = intercept (the value of Y when each of the independent variables are zero.

When number of independent variables is one, it is called a Simple Linear Regression model.

When we have multiple independent variables, it is called Multiple Linear Regression Model.

Assumptions of Linear Regression:

1. Linearity: The relationship between independent and dependent variables is linear.
2. Independence (no Multicollinearity): Independent variables should not have high correlation with each other.
3. Homoscedasticity: The variance of residuals should remain constant across all values of the independent variable
4. Normality of Residuals: Residuals should be normally distributed
5. No Autocorrelation: Residuals should not show any patters.

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>

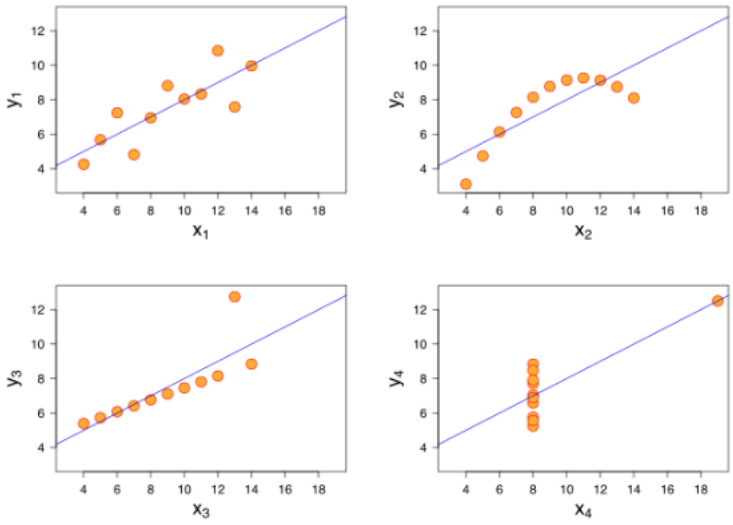
Anscombe's quartet was created by statistician Francis Anscombe in 1973. It comprises of a set of four datasets having identical statistical properties in terms of mean, variance, correlations, R-squared and linear regression line, but having different representation when plotted using scatter plot.

Each of the four datasets comprises of 11 x-y pair of data.

The four datasets of 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

Their scatter plot and linear regression line:



Explanation:

- The top left one seems to have a linear relationship between x and y.
- In the top right one, there is a non-linear relationship between x and y.
- In the bottom left one, there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the bottom right has a completely different distribution.

Conclusion:

The descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns. So, it is important to visualize and understand the data trend before using it in Linear regression models.

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)

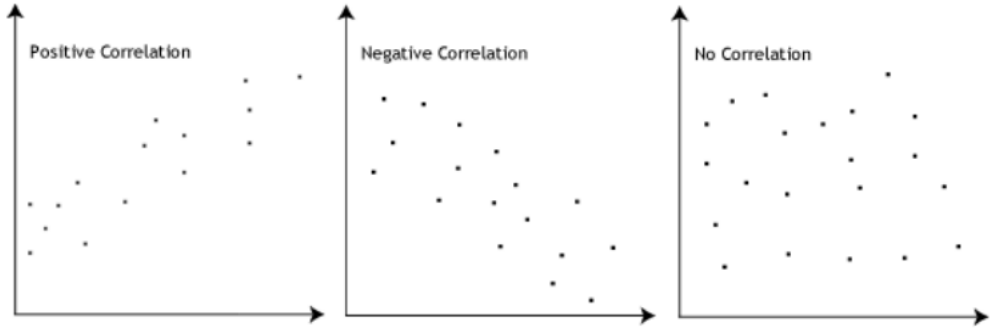
<Your answer for Question 8 goes here>

The Pearson's Correlation Coefficient (r) measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of the standard deviations. The result always has a value between -1 and 1. It can only reflect linear correlation of variables and ignores other type of correlation.

It can be represented by the following formula:

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

The Pearson correlation coefficient can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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>

What is Scaling?
Scaling in data analysis refers to the process of adjusting numerical features in a dataset to have a similar range or scale, ensuring that no single feature dominates the analysis due to vastly different value ranges, while still preserving the relative differences within each feature.

For Example, dataset with people's ages (ranging from 0 to 100) and their incomes (ranging from $20,000 to $200,000). These features have vastly different scales. Scaling brings these features to a similar range, like 0 to 1, while preserving the relative differences between the values.

Why is scaling performed?

- Preventing feature dominance: Features with larger magnitudes dominate those with smaller magnitudes, leading to biased models. Scaling ensures that all features contribute equally.
- Improving Model Performance: Many ML algorithm works better work better when numerical values are on a similar scale.

Difference between normalized scaling and standardized scaling?

- Range: Normalized scaling typically results in values between 0 and 1, while standardized scaling can result in values with a wider range, including negative values.
- Outliers: Normalized scaling is more sensitive to outliers, as it uses the minimum and maximum values. Standardized scaling is less sensitive, as it uses the mean and standard deviation.
- Distribution: Standardized scaling can help to make the data more normally distributed, while normalized scaling does not change the shape of the distribution.

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>

The Variance Inflation Factor (VIF) is used to measure multicollinearity in a linear regression model. The formula for VIF is:

VIF = 1 / (1 - R^2)

Where R^2 is the coefficient of determination from regressing one independent variable against all the other independent variables in your model.

For VIF to become infinite, R^2 must be 1.

If R^2 is 1, that means that independent variable can be perfectly predicted using other independent variables in the model. Which means, there is a **perfect multicollinearity**.

Causes for R^2 =1:

- **Duplicate Variables:** can be present in the dataset or be introduced by some transformation
- **Dummy variable trap**: While creating dummy variables for a categorical variable if one category is not dropped.

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>

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, such as the normal distribution. It compares the quantiles of your data to the quantiles of the theoretical distribution.

Q-Q plot works as follows:

- X-axis: Theoretical quantiles (expected values from a normal distribution)
- Y-axis: Sample quantiles (actual values from the dataset)
- If the data is normally distributed, points align along a straight diagonal line.
- Deviations from this line indicate departures from normality.

Use and importance of Q-Q plot in linear regression:

- In linear regression, one of the key assumptions is that residuals (errors) should be normally distributed. A Q-Q plot is used to check this assumption.
- when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not
- can help identify outliers by revealing data points that fall far from the expected pattern of the distribution.
