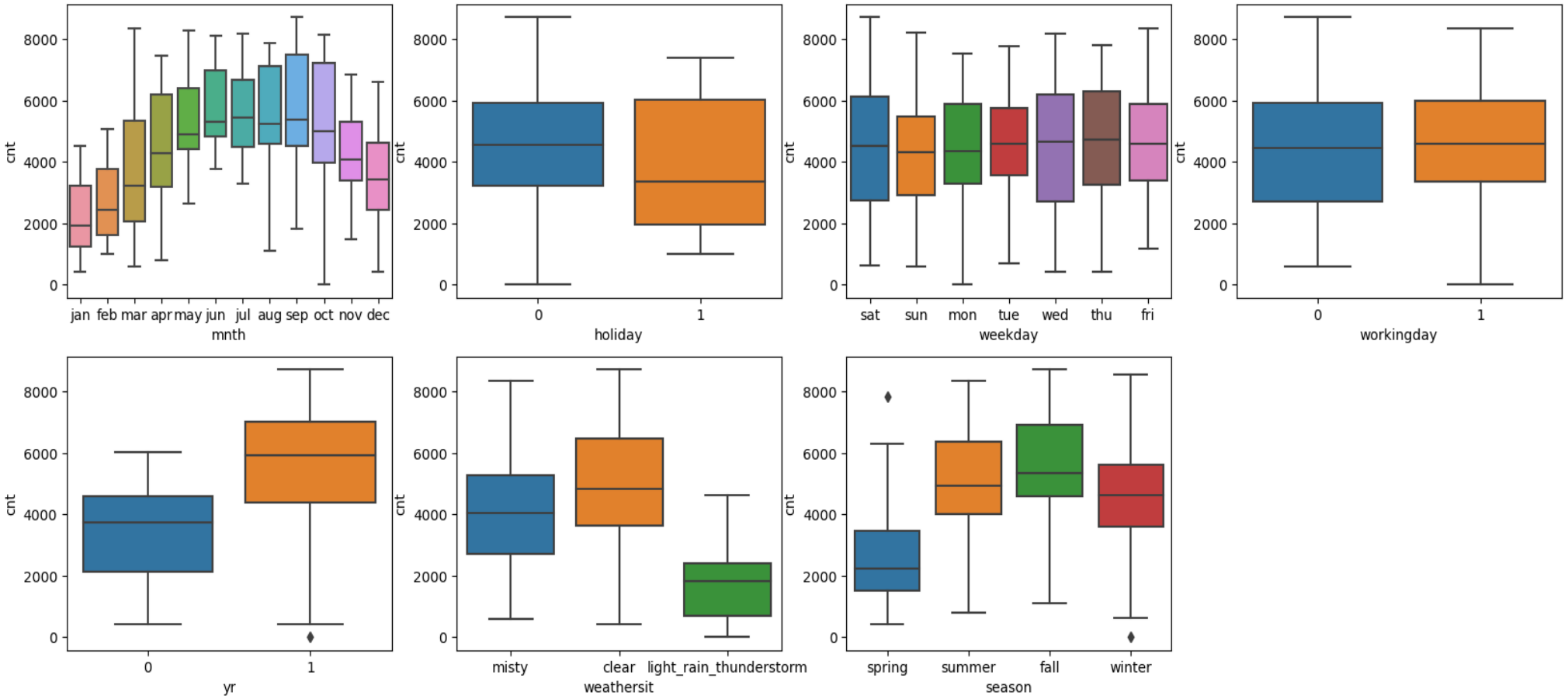
## Assignment-based Subjective Questions

**1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?** ? (3 marks)

Answer:

- Count is High in September month, 2019, On Thursday's with clear weather condition during fall season.
- Count is Low in January month, 2018, On Sunday's with light rain and thunderstorm weather condition.



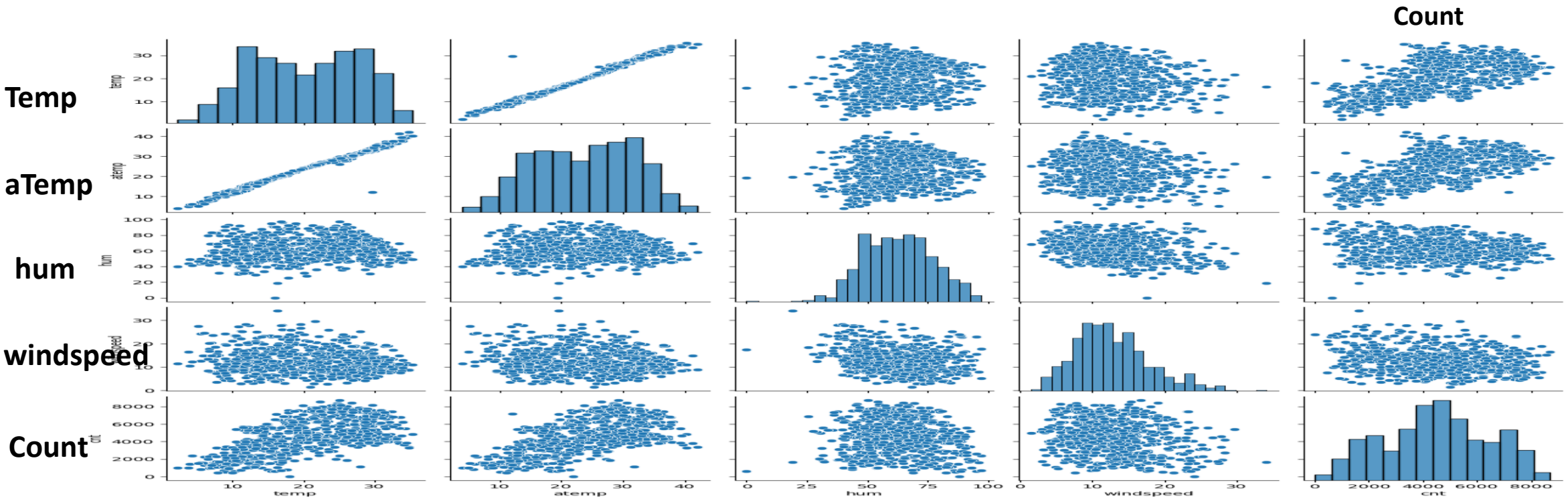
**2. Why is it important to use drop_first=True during dummy variable creation?** (2 mark)

Answer:

For every n classification of a column, there must be n-1 dummy variables. So, drop_first = True, drops the first dummy variable and remaining n-1 dummy variables are present. Dropping first dummy variable is important because, It helps in reducing extra column during dummy variable creation, Hence it reduces the correlation created among dummy variables.

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

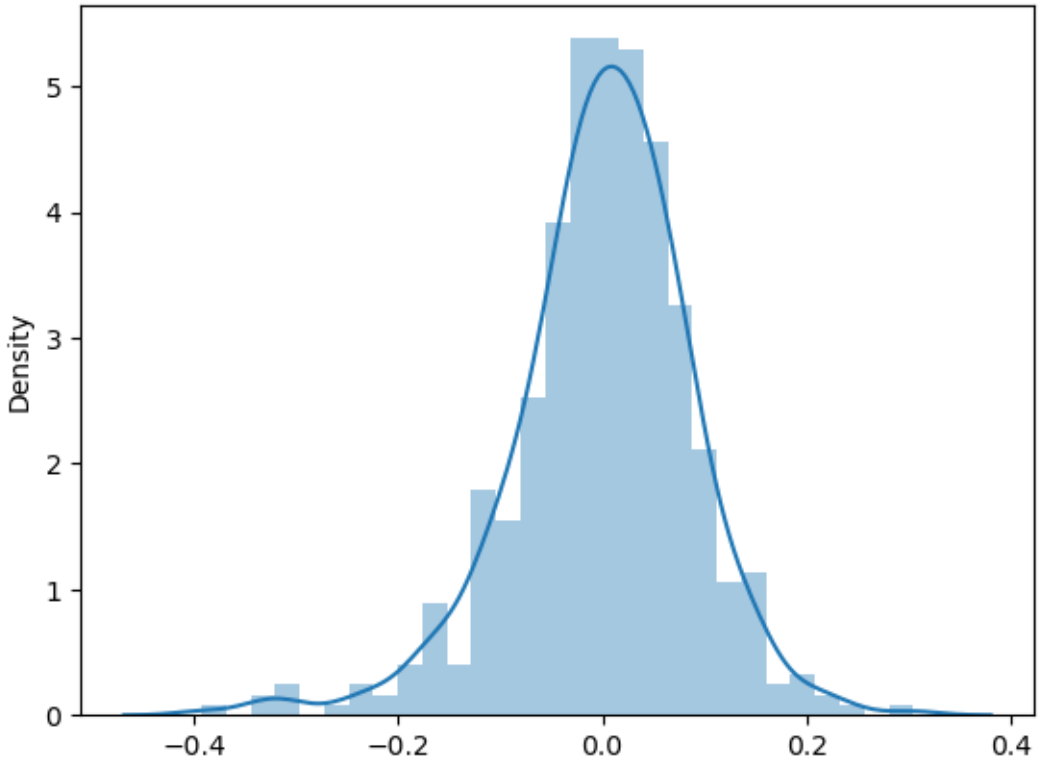
Answer: Temperature is Highly correlated with Count (1 mark)



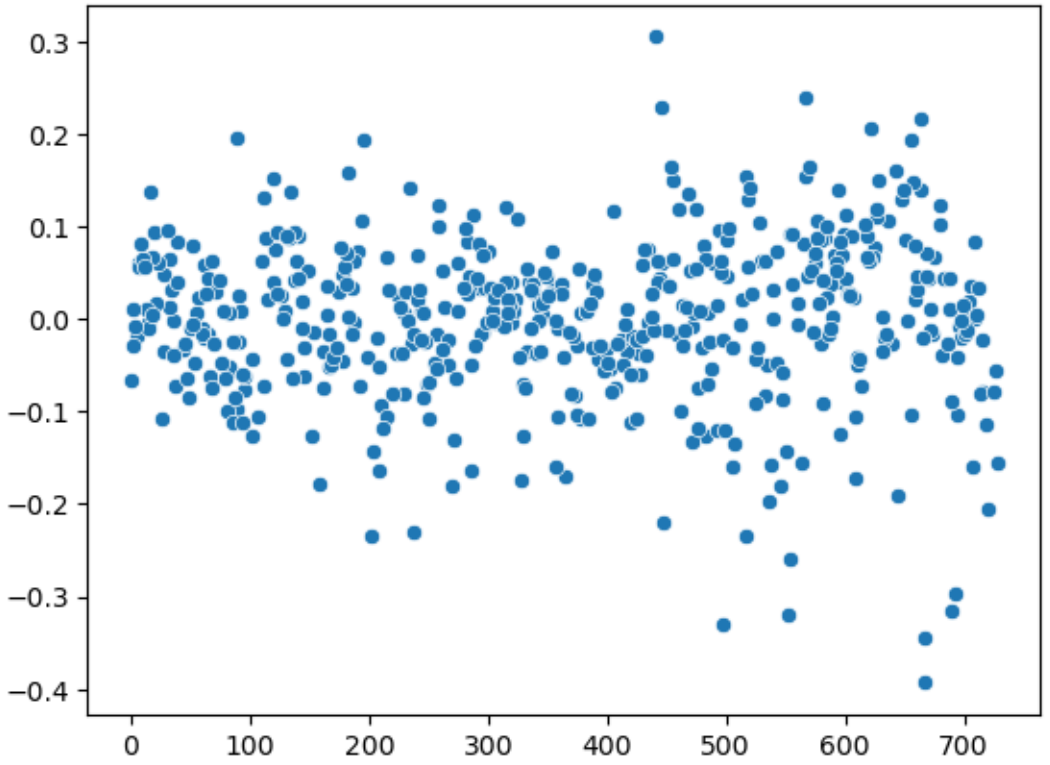
**4. How did you validate the assumptions of Linear Regression after building the model on the training set?** (3 marks)

Answer:

- Error terms normally distributed with mean zero.
- Error terms independent of each other
- No visible patterns are seen
- Variance is constant for all error terms(Homoscedasticity)




**Residuals distribution**




**Residuals Scatterplot**

**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?** (2 marks)

Answer:

Temperature, Light rain and thunderstorm weather situation and year.

1. *Temperature:* For unit increase in temperature, Count increases by 0.567
2. *Weather situation light rain and thunderstorm*: Weather situation as light rain and thunderstorm, Decreases the count by 0.243
3. *Year:* From 2018 to 2019 count increases by 0.229

Linear Equation:

cnt = 0.229*yr - 0.059*holiday + 0.043*workingday + 0.567*temp - 0.165*hum - 0.193*windspeed + 0.0757season_summer + 0.125*season_winter - 0.039*mnth_jan - 0.044*mnth_jul + 0.092*mnth_sep + 0.053*weekday_sat - 0.243*weathersit_light_rain_thunderstor-0.054*weathersit_misty

## General Subjective Questions

**1. Explain the linear regression algorithm in detail.** (4 marks)

Answer:

Regression model Predicts a continuous variable. Such as predicting sales in a day or predicting temperature in a city using past data. Linear Regression model fits a straight line through data points to get best fit line. Having an outliers may disrupt linear regression model. Linear Regression is comprehensive and transparent, Hence any one can understand. An Algorithm that provides linear relationship between independent and dependent variables. To predict future events.

*Where is Linear Regression used:*

Lr used in real time business applications, such as

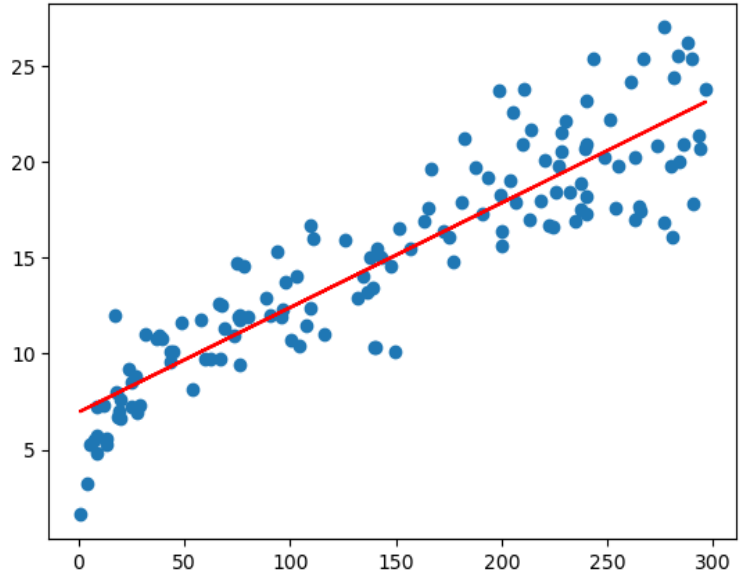
- A web series in various Ott platform having good number of viewership and suddenly viewership dropped. The features affecting the views count and be predicted using linear regression algorithm.
- A Bike sharing company, Which allows users to rent a bike for short periods. increasing its demand for shared bikes day by day. Suddenly due to pandemic demand got turn down. Linear regression helps is predicting various features that impacts the demand for shared bikes like weather situation, season, temperature etc.

*Understanding Linear Regression algorithm:*

- First scatter plot all data points, with dependent variable on y-axis and Independent variable on x-axis, LR fits a straight line through data points and By calculating equation of a line, one can find slope(change in y/change in x), Intercept(point where line intercepts with y-axis)

*Finding Best Fit line:*

The Goal of Linear Regression is to obtain best values for slope and intercept to find best fit line. LR performs Ordinary least squares or Mean square error or cost function to find best fit line with manual error.




*What is R-squared:*

R-squared is the statistical measure of how close the data are to the fitted regression line. R-squared value is nothing but square of Correlation Coefficient. Higher R squared means good fit model. Whereas lower R squared indicates that the model is not fitted to data points.

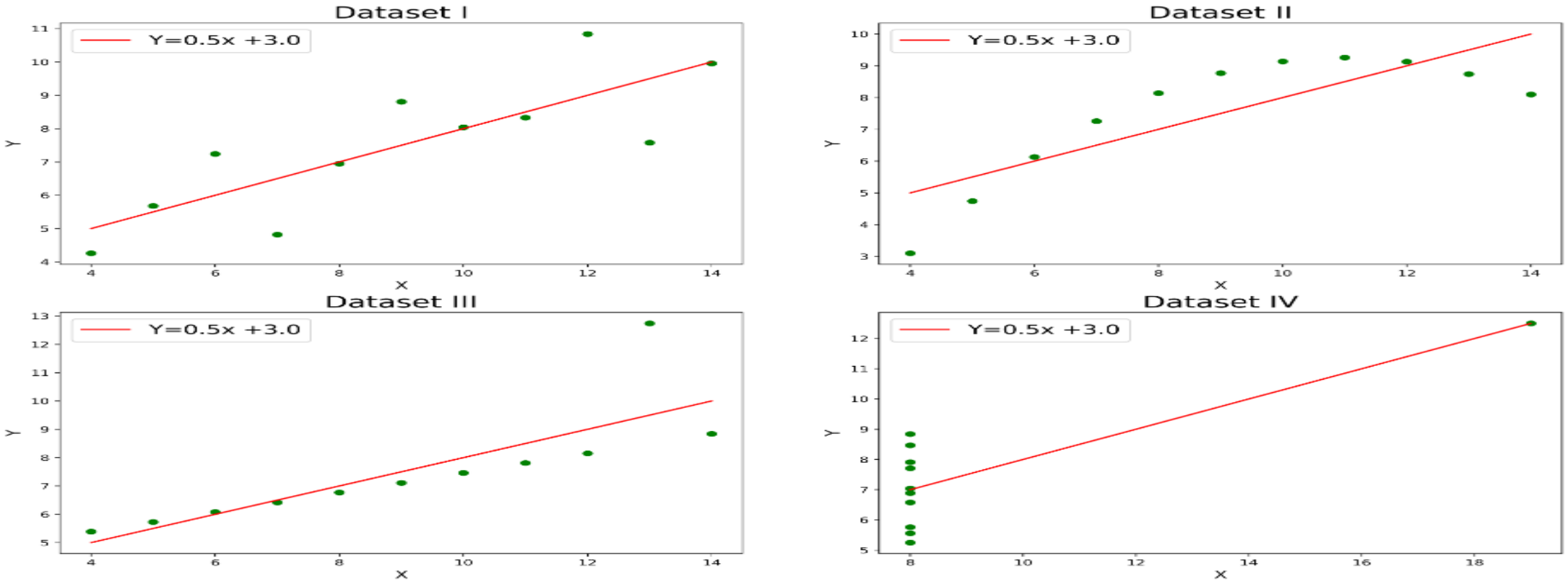
**2. Explain the Anscombe's quartet in detail.** (3 marks)

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

When we take four different datasets, with x,y pairs and Calculate these statistical measures for all four. These looks similar But while plotting them, each plot looks completely different

| | I | II | III | IV |
|---|---|---|---|---|
| Mean_x | 9.000000 | 9.000000 | 9.000000 | 9.000000 |
| Variance_x | 11.000000 | 11.000000 | 11.000000 | 11.000000 |
| Mean_y | 7.500909 | 7.500909 | 7.500000 | 7.500909 |
| Variance_y | 4.127269 | 4.127629 | 4.122620 | 4.123249 |
| Correlation | 0.816421 | 0.816237 | 0.816287 | 0.816521 |
| Linear Regression slope | 0.500091 | 0.500000 | 0.499727 | 0.499909 |
| Linear Regression intercept | 3.000091 | 3.000909 | 3.002455 | 3.001727 |

*Scatter plot*: Even though statistical coefficients same for all four datasets, there is difference in distribution of data points.




*Conclusion:*

While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

**3. What is Pearson's R?** (3 marks)

The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

| Pearson's r | Strength | Direction |
|---|---|---|
| Greater than 0.5 | Strong | Positive |
| Between 0.3 to 0.5 | Moderate | Positive |
| Between 0 to 0.3 | Weak | Positive |
| 0 | No correlation | None |
| Between 0 to -0.3 | Weak | Negative |
| Between -0.3 to -0.5 | Moderate | Negative |
| Less than -0.5 | Strong | Negative |

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?** ? (3 marks)

Scaling:

Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Why scaling is performed:

In a dataset all features may not have same range of values, Scaling facilitates meaningful comparisons between features, improves model convergence, and prevents certain features from overshadowing others based solely on their magnitude.

Difference between Normalized and standardized scaling:

Two popular rescaling methods- Min-Max scaling or Normalized scaling and Standardization (mean=0 and sigma=1). 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 extreme data point (outlier).

*Normalized:* min:0 and max:1, No outliers as it comprises all data points between 0 and 1. Mostly used.
*Standardization:* mean:0 and standard deviation:1, outliers still exist. Used when there is an extreme oulier.

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?** (3 marks)

VIF(Variance Inflation Factor) describes the correlation between independent variables. If there is perfect correlation between independent variables, then VIF = infinity,

- *VIF above 10 :* Definitely High VIF and variable should be eliminated as it causes multicollinearity
- *VIF greater than 5:* Should not ignore, As there is a chance of multicollinearity
- *VIF less than 5:* Good VIF value. No need to eliminate this as it is significant for predicting.

If VIF = 5,
5 = 1/1-R**2
R**2 = 0.8 (Good Model)

*Formula:*
VIF = 1/1-R**2, If R**2(correlation) is more then VIF is less and vice versa.

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.** (3 marks)

Answer:

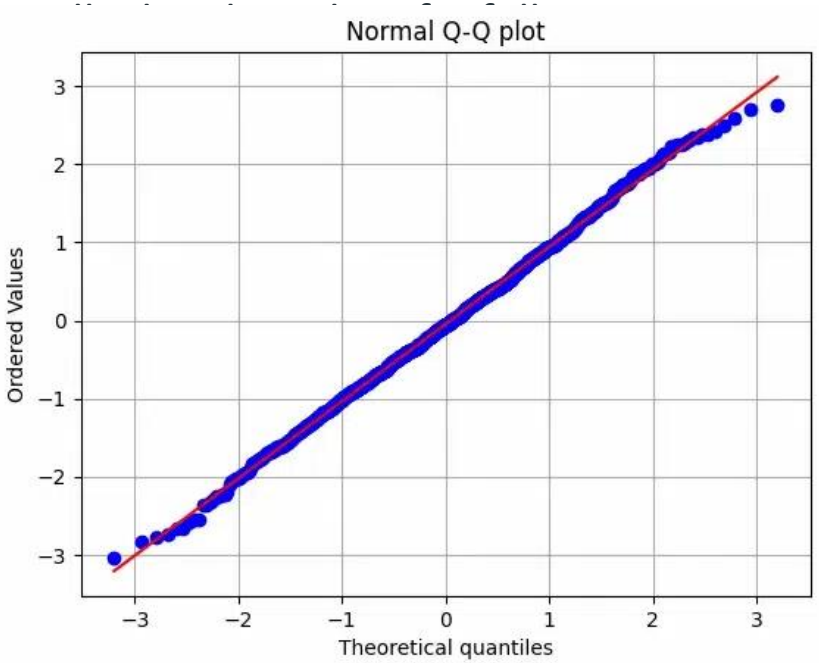
The quantile-quantile( q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not.

A point on the plot corresponds to one of the quantiles of the second distribution plotted against the same quantile of the first distribution.

Use of Q-Q plot:

Q-Q plots are particularly useful for assessing whether a dataset is no other known distribution. They are commonly used in statistics, data analysis, a and identify departures from expected distributions.

Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data



Importance:

Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.

Q-Q plots can help identify outliers by revealing data points that fall far from the expected pattern of the distribution.

- Assessing Distributional Assumptions
- Detecting Outliers
- Comparing Distributions
- Assessing Normality
- Model Validation
- Quality Control