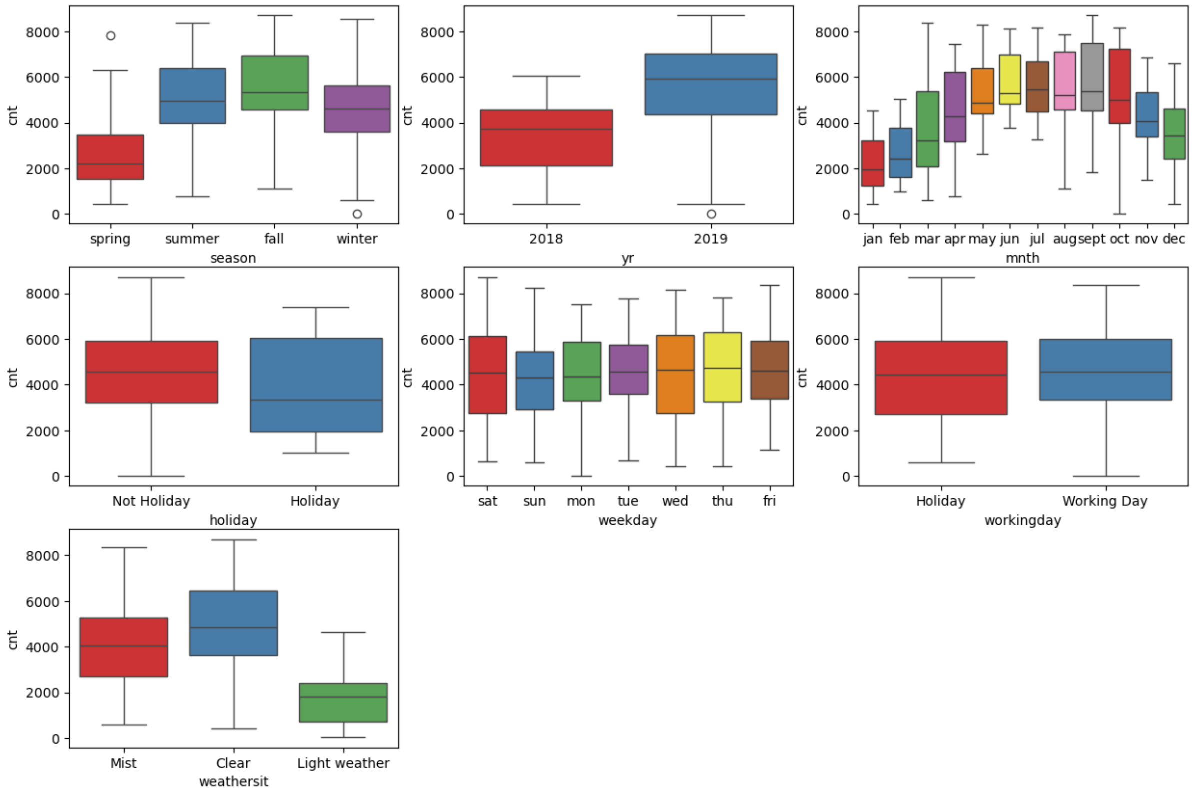
Assignment-based Subjective Questions

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)



We can observe the following from above box plots

- Fall season has more count compared to other seasons.
- Compared to 2018 we have the count increased in 2019.
- May to October has more count and we can see pattern starts to increase from start of year to mid and then declines till the end of the year.
- We can see the count is more on non-holidays and less on holidays.
- Since the median varies only less amount, we can say count is almost equal on all days of the week.
- Booking seems to be almost equal whether it’s a working day or not.
- Clear weather attracts more bookings.

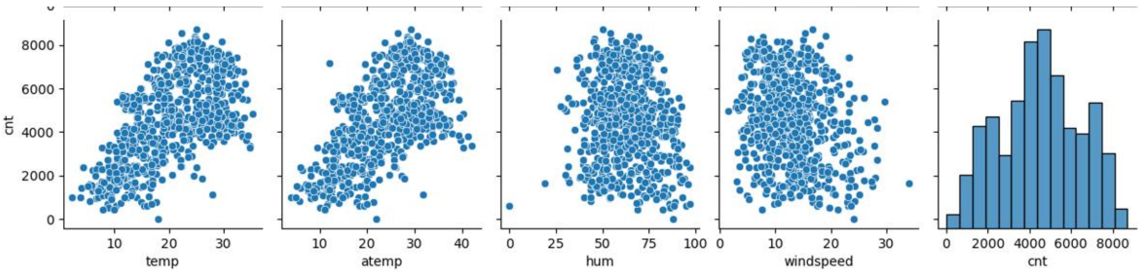
Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using **drop_first=True** prevents **multicollinearity** and simplifies interpretation by **avoiding redundancy** in the model and treating one category as the baseline. It's a common best practice, especially when working with regression models, to ensure more stable and interpretable results.

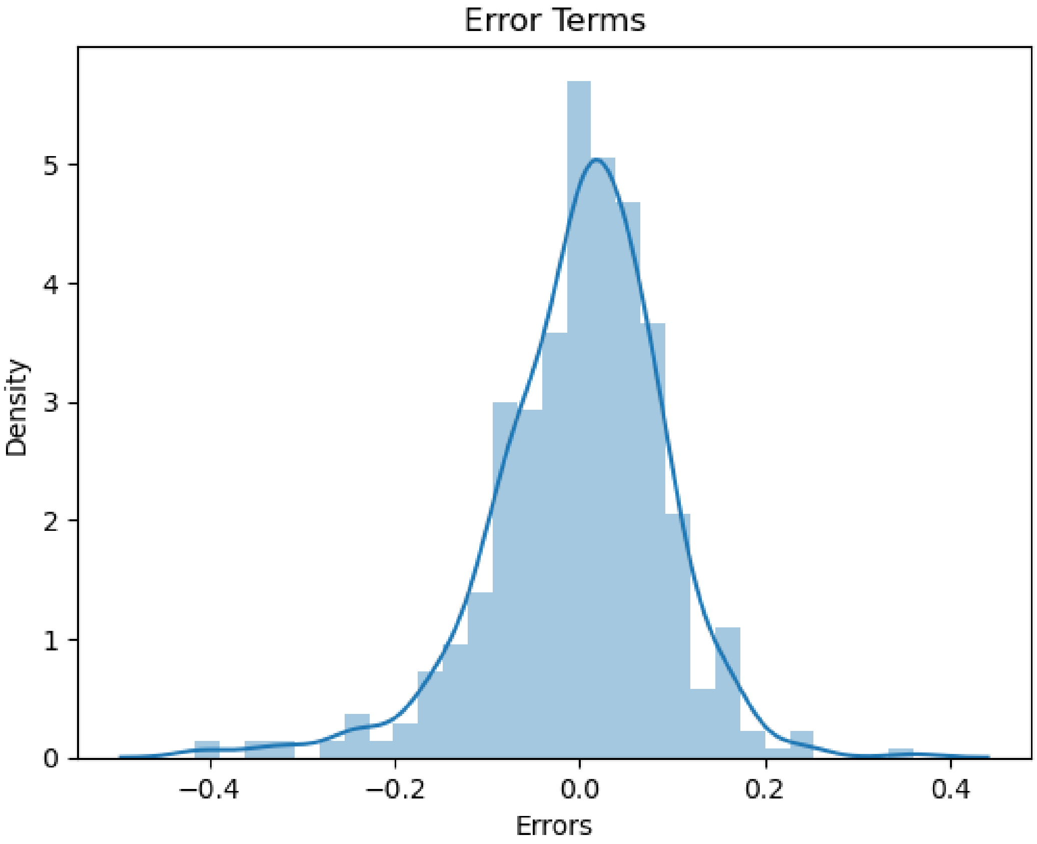
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

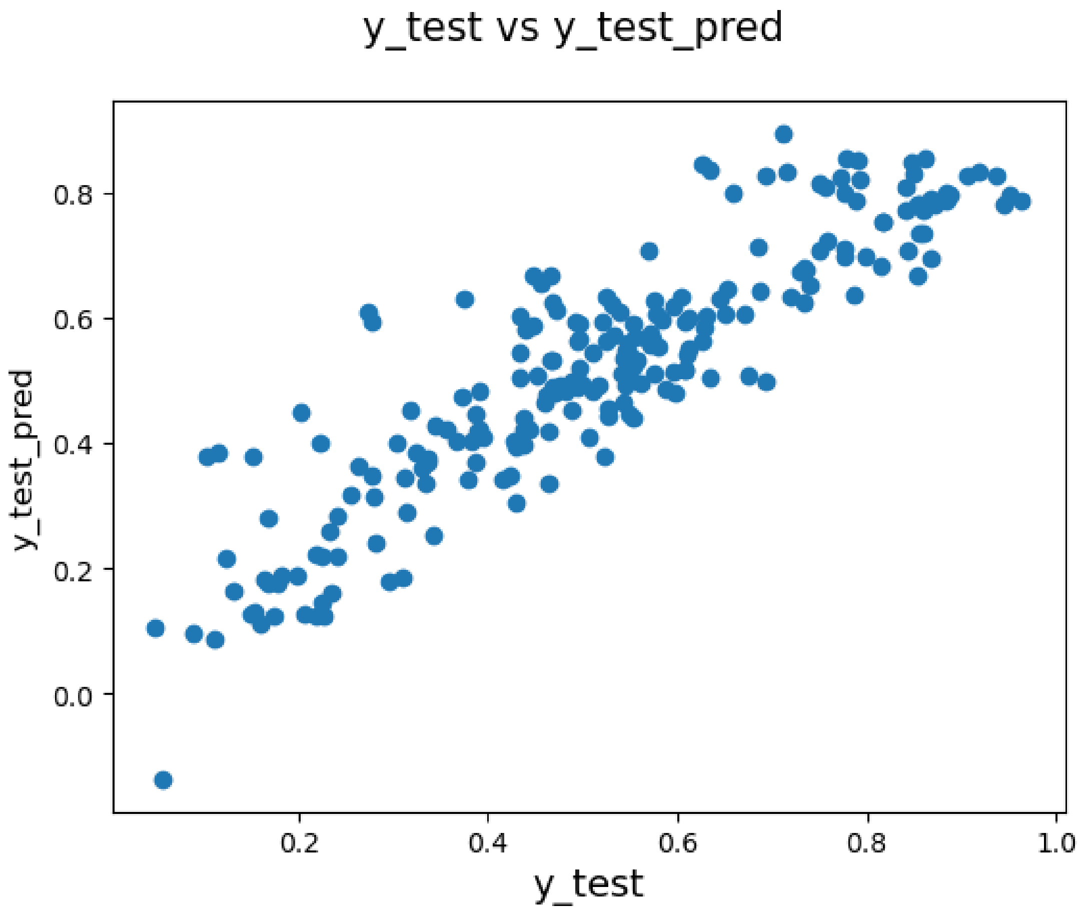


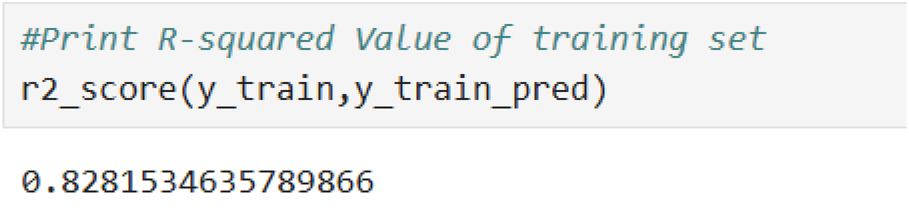
From the above pair-plot we can see **temp** and **atemp** have high correlation with target variable "**cnt**" since from heatmap we got **0.63** value for both temp and atemp.

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)

Residual analysis on the error terms so that we can check whether it follows a normal distribution with mean is centered around zero and we're able to see this in our model.



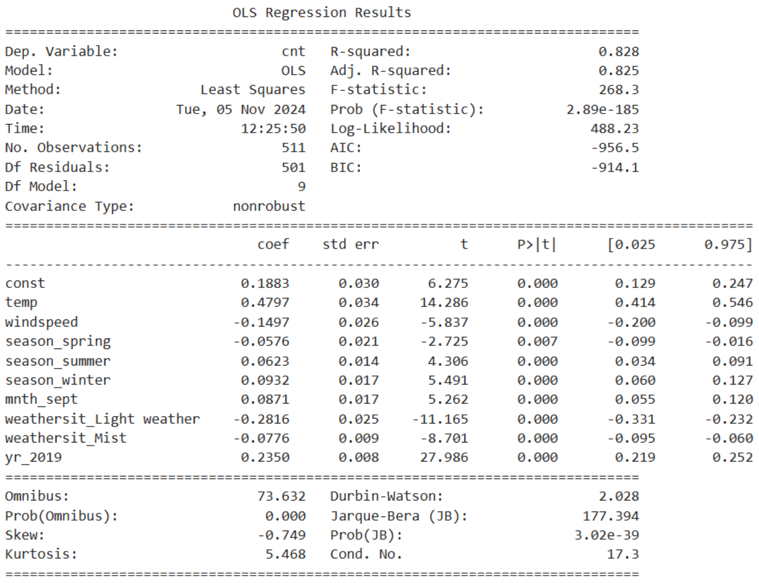




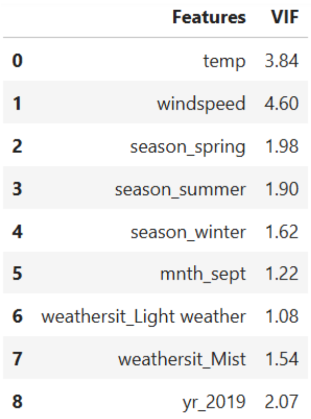
```
#Print R-squared Value of training set
r2_score(y_train,y_train_pred)
```

```
0.8281534635789866
```

Calculating the R-squared value and it’s above 80%(0.80) and check if the VIF is less than 5 and P value are less than 0.5. we’re able to see our model has 0.82 R squared which is greater than .80 and VIF values are less than 5 and P values are all almost near or equal to zero.

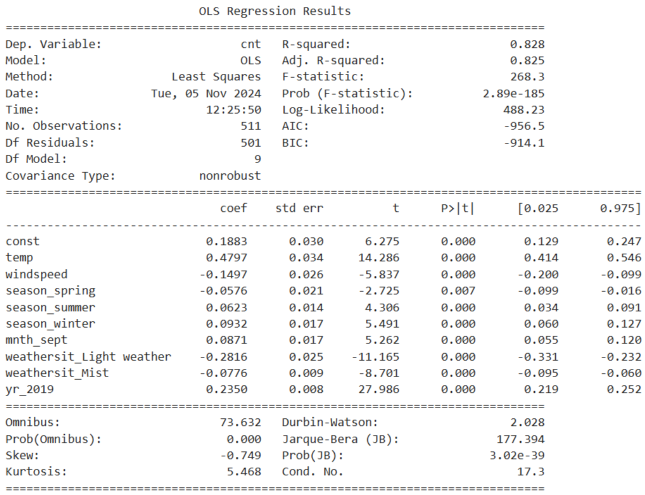


```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.828
Model:                            OLS   Adj. R-squared:                  0.825
Method:                 Least Squares   F-statistic:                     268.3
Date:                Tue, 05 Nov 2024   Prob (F-statistic):          2.89e-185
Time:                        12:25:50   Log-Likelihood:                 488.23
No. Observations:                 511   AIC:                            -956.5
Df Residuals:                     501   BIC:                            -914.1
Df Model:                           9                                         
Covariance Type:            nonrobust                                         
==============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                        0.1883      0.030      6.275      0.000       0.129       0.247
temp                         0.4797      0.034     14.286      0.000       0.414       0.546
windspeed                   -0.1497      0.026     -5.837      0.000      -0.200      -0.099
season_spring               -0.0576      0.021     -2.725      0.007      -0.099      -0.016
season_summer                0.0623      0.014      4.306      0.000       0.034       0.091
season_winter                0.0932      0.017      5.491      0.000       0.060       0.127
mnth_sept                    0.0871      0.017      5.262      0.000       0.055       0.120
weathersit_Light weather    -0.2816      0.025    -11.165      0.000      -0.331      -0.232
weathersit_Mist             -0.0776      0.009     -8.701      0.000      -0.095      -0.060
yr_2019                      0.2350      0.008     27.986      0.000       0.219       0.252
==============================================================================
Omnibus:                       73.632   Durbin-Watson:                   2.028
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              177.394
Skew:                          -0.749   Prob(JB):                     3.02e-39
Kurtosis:                       5.468   Cond. No.                         17.3
==============================================================================
```



	Features	VIF
0	temp	3.84
1	windspeed	4.60
2	season_spring	1.98
3	season_summer	1.90
4	season_winter	1.62
5	mnth_sept	1.22
6	weathersit_Light weather	1.08
7	weathersit_Mist	1.54
8	yr_2019	2.07

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)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.828
Model:                            OLS   Adj. R-squared:                  0.825
Method:                 Least Squares   F-statistic:                     268.3
Date:                Tue, 05 Nov 2024   Prob (F-statistic):          2.89e-185
Time:                        12:25:50   Log-Likelihood:                 488.23
No. Observations:                 511   AIC:                            -956.5
Df Residuals:                     501   BIC:                            -914.1
Df Model:                           9
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                        0.1883      0.030      6.275      0.000       0.129       0.247
temp                         0.4797      0.034     14.286      0.000       0.414       0.546
windspeed                   -0.1497      0.026     -5.837      0.000      -0.200      -0.099
season_spring               -0.0576      0.021     -2.725      0.007      -0.099      -0.016
season_summer                0.0623      0.014      4.306      0.000       0.034       0.091
season_winter                0.0932      0.017      5.491      0.000       0.060       0.127
mnth_sept                    0.0871      0.017      5.262      0.000       0.055       0.120
weathersit_Light weather    -0.2816      0.025    -11.165      0.000      -0.331      -0.232
weathersit_Mist             -0.0776      0.009     -8.701      0.000      -0.095      -0.060
yr_2019                      0.2350      0.008     27.986      0.000       0.219       0.252
==============================================================================
Omnibus:                       73.632   Durbin-Watson:                   2.028
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              177.394
Skew:                          -0.749   Prob(JB):                     3.02e-39
Kurtosis:                       5.468   Cond. No.                         17.3
==============================================================================
```

From the above co-efficient values we can say the top 3 features are:
1st – temp , 2nd – light weather(negative) , 3rd – year or windspeed

---

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 fundamental statistical and machine learning technique used to model the relationship between a dependent variable (often referred to as the target or output) and one or more independent variables (features or predictors). The objective of linear regression is to find a linear relationship that best fits the data points. Below is a detailed explanation of the linear regression algorithm, including its types, mathematical formulation, assumptions, implementation, and applications.

Types:

Simple Linear regression (SLR):

This involves one dependent variable and one independent variable. The relationship is modeled using a straight line.

The equation of the line is given by:

$$y=\beta0+\beta1x+\epsilon$$

where:

y is the dependent variable,
x is the independent variable,
$\beta0$ is the intercept,
$\beta1$ is the slope of the line,
ϵ is the error term (the difference between observed and predicted values).

Multiple Linear Regression (MLR):

This involves one dependent variable and multiple independent variables. The model can capture more complex relationships.

The equation of the line is given by:

$$y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon$$

where $x1,x2,...xn$ are the independent variables.

In order to find best fit line, A measure called sum of squared errors iscommonly used which is a cost function. To minimize or maximize this cost function, Gradient descent optimization is used.

The strength of linear regression model can be assessed using 2 metrics:

1. R2
2.Residual standard error(RSE)

Steps involving in creating a linear regression model:

1.Data Collection
2.Data Preprocessing
3.EDA
4.Train Test Split
5.Training the model
6.Evaluate the model
7.Interpret the model

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 is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly in their distributions and relationships between the variables. Created by the statistician Francis Anscombe in 1973, this quartet serves as a compelling illustration of the importance of visualizing data before drawing conclusions from statistical analyses. Each dataset in Anscombe's quartet contains 11 pairs of xxx and yyy values, allowing for the analysis of linear regression and correlation.

The Datasets

The quartet consists of four distinct datasets, labeled I, II, III, and IV. Here's a summary of each dataset:

1. **Dataset I:**
   - Values:
     x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
     y: 8, 6, 7, 9, 11, 12, 4, 3, 10, 6, 4
   - This dataset represents a linear relationship between xxx and yyy.
2. **Dataset II:**
   - Values:
     x: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8
     y: 6, 5, 7, 9, 11, 12, 4, 3, 10, 6, 4
   - Here, all xxx values are the same (8), resulting in a vertical line with varying yyy values.
3. **Dataset III:**
   - Values:
     x: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8
     y: 12, 10, 14, 9, 11, 12, 4, 3, 10, 6, 4
   - This dataset has a non-linear relationship, featuring a parabolic shape.
4. **Dataset IV:**
   - Values:
     x: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8
     y: 4, 6, 7, 9, 11, 12, 4, 3, 10, 6, 4
   - This dataset also has a distinct pattern, but it is characterized by outliers that can significantly affect regression results.

Statistical Properties

Despite their differences, the datasets share similar statistical properties:

- Mean of xxx: 999
- Mean of yyy: 7.57.57.5
- Variance of xxx: 111111
- Variance of yyy: 4.1254.1254.125
- Correlation coefficient (rrr): Approximately 0.8160.8160.816 for all datasets
- Linear regression line: All datasets yield similar regression lines with the same slope and intercept.

Visualizing Anscombe's Quartet

To fully appreciate the differences among these datasets, visualization is crucial. When plotted, the datasets look quite different despite their statistical similarities:

1. Dataset I shows a clear linear relationship.
2. Dataset II displays a cluster of points with a vertical distribution.
3. Dataset III reveals a parabolic curve.
4. Dataset IV contains an outlier that skews the visual representation.

Importance of Anscombe's Quartet

Anscombe's quartet emphasizes several key principles in data analysis:

1. **Data Visualization**: Statistical summaries (mean, variance, correlation) can be misleading. Visualizing the data helps uncover underlying patterns, trends, and relationships.
2. **Modeling Assumptions**: Different datasets may fit a linear regression model similarly but may represent different underlying relationships. It's important to check assumptions (e.g., linearity, homoscedasticity).
3. **Outliers and Influential Points**: The presence of outliers can significantly affect regression

results, leading to erroneous interpretations. It is crucial to identify and analyze the impact of outliers in datasets.

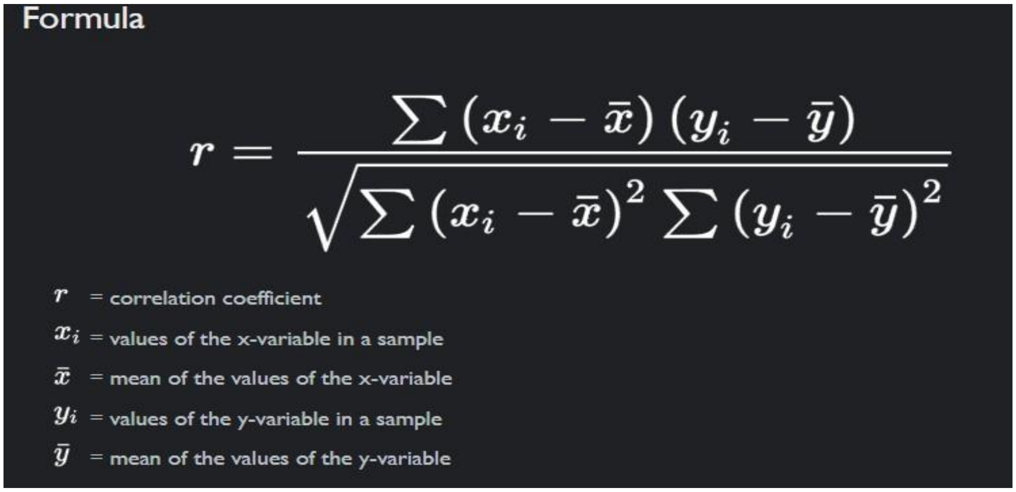
4. **Exploratory Data Analysis (EDA):** Anscombe's quartet illustrates the necessity of EDA to understand data thoroughly before applying statistical models.

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 correlation coefficient (*r*) is the most common way of measuring a linear correlation only. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It summarizes the characteristics of a dataset. It is widely used in various fields, including statistics, psychology, economics, and biology, to analyze the strength and direction of relationships between variables.

- 0 < r <= 1 indicates a positive linear relationship
- r = 0 indicates no correlation
- -1 <= r < 0 indicates a negative linear relationship



Formula

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

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

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 the process to normalize the data within the particular range. There are two ways to scale the data. They are

- Normalization – Min Max Scaling – It scales the values between 0 and 1
  - X = (x – xmin) / (xmax – xmin)

- Standardization – It scales the value that it has a mean of 0 and standard deviation of 1.
    - X = (x – mean) / standard deviation

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(i) = 1 / (1 – Ri^2)

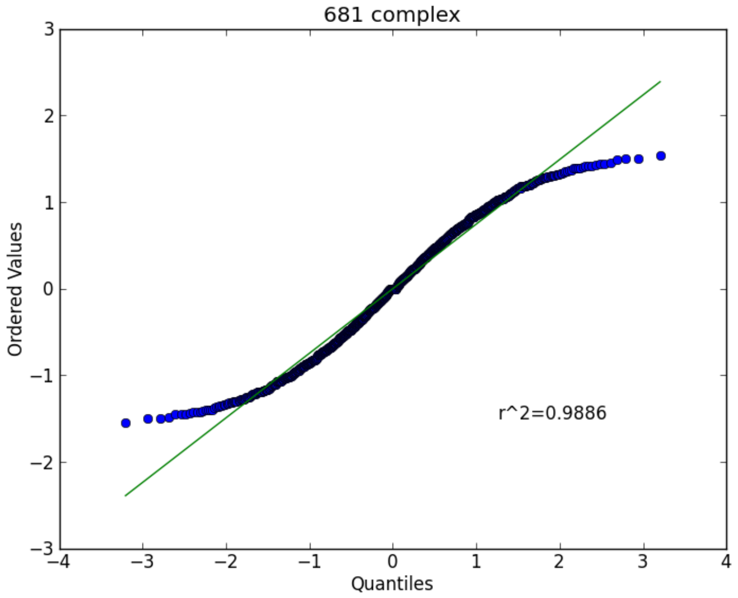
This is the formula to calculate VIF. The reason for the value of VIF to be infinite is that the R Squared values is equal to 1 and the denominator equals to 0. This denotes perfect correlation between variables.

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>

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions .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. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?


681 complex
Ordered Values
Quantiles
r^2=0.9886
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4