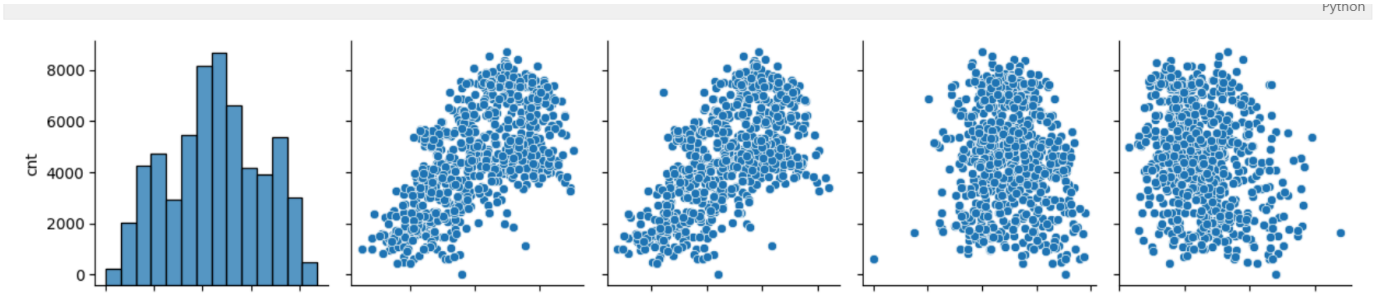
BIKE SHARING ASSIGNMENT

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?**
    - The categorical variables available in the dataset given are season, workingday, weathersit, weekday, yr, holiday, and mnth.
    - **Season:** Most favourable season for bike sharing is Fall, and the least favourable is spring.
    - **Workingday:** Registered users seem to rent the bikes on the working days and casual users prefer the biking on holidays. When seen in total there is not much variation in no. of bikes renting on a working day and a non-working day.
    - **Weathersit:** The most favourable weather condition is the clean/few clouds days.
    - **Yr:** Increase in no. of bikes from 2018 to 2019.
    - **Mnth:** Bike rentals are more in June, July, August, September and October months.

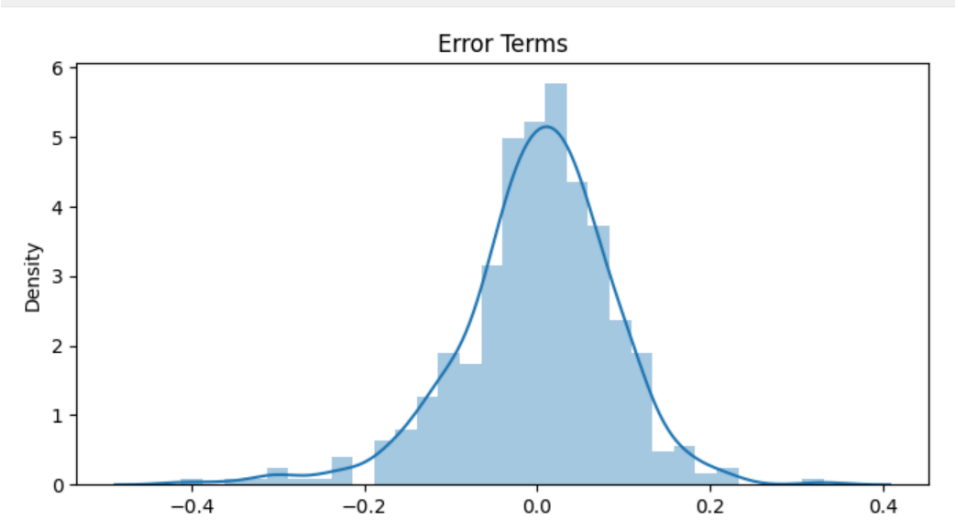
2. **Why is it important to use drop_first=True during dummy variable creation?**
    - Using drop_first = True when creating dummy variables in pandas to avoid the dummy variable trap, which occurs when one of the dummy variables is mostly collinear with the other dummy variables.
    - When dummy variables are created for categorical variables, each category is represented using 0 or 1.
    - If all the variables are included, one of the variables may be well predicted by the others, resulting in multicollinearity. When drop_first = True, one category is dropped. So when all other variables are 0 it represents the dropped category.
    - Also it reduces the no. of features in the dataset with no loss of information.

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

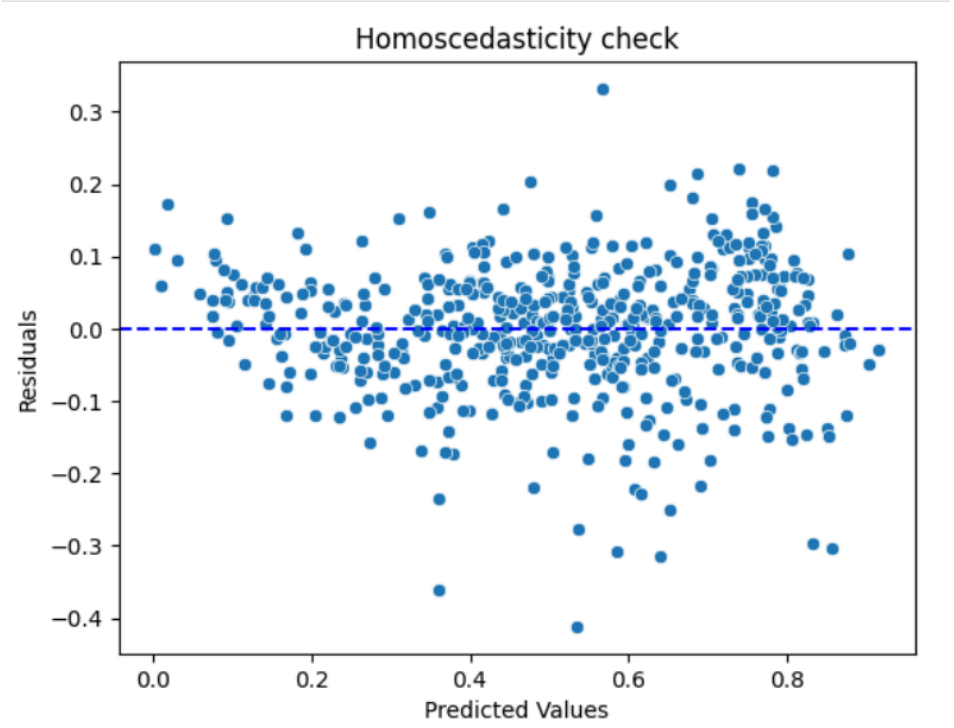


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - **Residual Analysis**

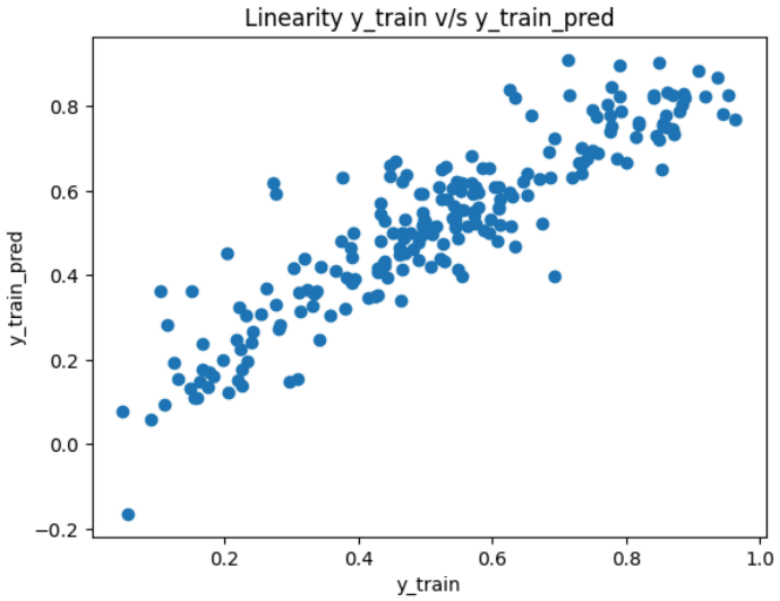
- Checking the residuals ie; the differences between observed and predicted values
- The residuals should be approximately normal distributed and there should be no obvious pattern in the residuals.



- **Homoscedasticity:** When residuals are plotted against predicted values, the spread has to be nearly constant across all levels of predicted values.



- **Linearity:** When a scatter plot is made for actual and predicted target variables points should fall approximately along a diagonal line.



- **Independence of Residuals:** There has to be no specific pattern in residuals when plotted against variables.
- **Multicollinearity:** VIF values have to be below a threshold value, generally VIF below 5.
- Model performance on new data has to be consistent and generalised, ensuring no overfitting.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - **Temp:** temp is positively correlated to target variable cnt.
   - **Yr:** No. of bikes increases from 2018 to 2019.
   - **Season:** No. of bike rides are less in Rainy.
   - Equation for best fit line: `0.54 * temp + 0.23 * const + 0.23 * yr + 0.14 * winter + 0.13 * Sep + 0.1 * summer + 0.06 * Aug + 0.04 * Oct + 0.02 * Mar -0.05 * Mist -0.09 * holiday -0.18 * hum -0.19 * windspeed -0.24 * Light`

General Subjective Questions

1. **Explain the linear regression algorithm in detail**
   - **Linear Regression:** Linear Regression mainly aims in finding a linear equation that best predicts the dependent variable from the independent variables.
   - There are 2 types of linear regression algorithms
     - Simple Linear Regression – Dependent on a single independent variable $y=\beta0+\beta1x+\epsilon$
     - Multiple Linear Regression – Dependent on multiple independent variables. $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$
   - **Assumptions of Linear Regression**
     1. **Linearity**: The relationship between the dependent and independent variables is linear.

2. **Normal distribution of Residuals**: The residuals should follow a normal distribution.
3. **Independence**: The residuals are independent of each other
4. **Homoscedasticity**: The variance of the residuals is constant across all levels of the independent variables.

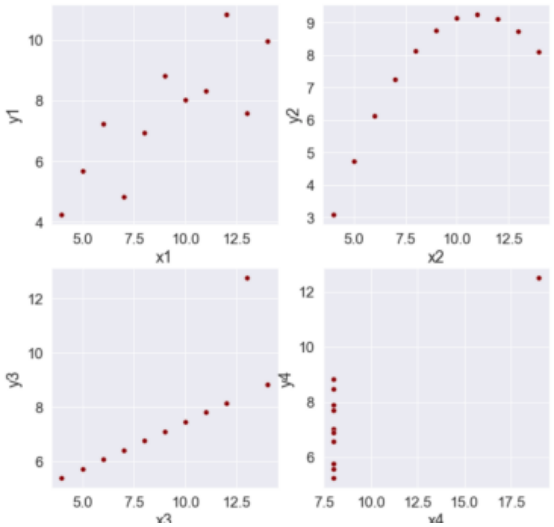
- **Objective Function:** The goal is to find the values of (bo, b1, b2,..., b) that minimises the sum of the squared differences between the observed and predicted values.

  MSE = Σ(yi − pi)2

- The model is trained on a dataset where the algorithm learns the coefficient values that best fit the data. This involves feeding output pairs into the algorithm and adjusting the coefficients until the model produces predictions that are close to reality.
- Once the model is trained, it can be used to make predictions about new, unseen features. The predicted values are obtained by adding the new input to the regression equation.
- Model's performance is calculated based on R-square, MSE, F-statistic etc..

2. Explain the Anscombe's quartet in detail.

- Anscombe's quatrain is a group of four records that have almost the same basic description but are very different in plotting. This highlights the importance of visualisation and the limitations of relying solely on written records.
- This quatrain introduces the idea that data sets with similar values can show different patterns when plotted. It was created by author Francis Anscombe in 1973 to highlight the importance of analysing data and the impact of outliers and other useful observations on statistical properties.



Dataset is as follows:

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.040000	9.140000	7.460000	6.580000
1	8	8	8	8	6.950000	8.140000	6.770000	5.760000
2	13	13	13	8	7.580000	8.740000	12.740000	7.710000
3	9	9	9	8	8.810000	8.770000	7.110000	8.840000
4	11	11	11	8	8.330000	9.260000	7.810000	8.470000
5	14	14	14	8	9.960000	8.100000	8.840000	7.040000
6	6	6	6	8	7.240000	6.130000	6.080000	5.250000
7	4	4	4	19	4.260000	3.100000	5.390000	12.500000
8	12	12	12	8	10.840000	9.130000	8.150000	5.560000
9	7	7	7	8	4.820000	7.260000	6.420000	7.910000
10	5	5	5	8	5.680000	4.740000	5.730000	6.890000

Descriptive statistics

	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	9.000000	9.000000	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.000000	4.000000	8.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.500000	6.500000	8.000000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	9.000000	9.000000	8.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	11.500000	11.500000	8.000000	8.570000	8.950000	7.980000	8.190000
max	14.000000	14.000000	14.000000	19.000000	10.840000	9.260000	12.740000	12.500000

- All four groups are similar when analysed using simple statistics, but different when shown graphically.
- The first scatter plot shows a slight linear relationship.
- The second plot, the relationship between the variables is non-linear.
- In the third plot, the relationship is linear, but there is a distinct regression line.
- The fourth plot illustrates the situation where high correlation can lead to a correlation coefficient even if other data points do not show correlation between the variables.

3. What is Pearson's R?

- Pearson's r, also known as the Pearson correlation coefficient (PCC), is a parameter used to measure the strength and direction of the relationship between two continuous variables. This is simply called "correlation."
- It indicates whether an increase in one variable is associated with an increase (positive correlation) or decrease (negative correlation) in other variables, or no correlation at all (no correlation).
- The formula for calculating Pearson's r is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

n is the number of data points.

x and y are the two variables being compared.

Range: The Pearson correlation coefficient, r, ranges from -1 to +1:

- ➔ **+1**: A perfect positive linear relationship. As one variable increases, the other increases proportionally.
- ➔ **-1**: A perfect negative linear relationship. As one variable increases, the other decreases proportionally.
- ➔ **0**: No linear relationship between the variables.

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

- Scaling is a data preprocessing technique used to adjust the values in a dataset so that they are consistent. This is especially true when the magnitudes of the eigenvalues are very different.
- Scaling brings all the features into comparison by modelling them better and more accurately, especially those that are sensitive to large values.
- **Why is scaling performed**
- Avoid coefficient bias: In linear regression, the coefficients represent how much the variable changes with a unit change in the independent variable. If the values of the characteristics are very different, the model will assign larger coefficients to the characteristics with larger values, even if they are not significant. This can lead to negative results.
- In gradient based optimization method, the model adjusts the weight back. If the features are not scaled, the convergence will be slower because gradient descent cannot take small steps in size with small values and larger steps in size with me. This may lead to performance degradation. Scaling the features ensures that gradient descent updates all parameters at a comparable rate, speeding up convergence.
- Easy to interpret coefficient change in features.
- **Difference between normalized scaling and standardized scaling**
- **Normalized Scaling:** rescales the data to a fixed range, lie between **0 and 1**, or between **-1 and 1**.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Makes feature values comparable in a bounded range.
- No assumption of normality.
- **Standardized scaling:** Standardization transforms the data such that each feature has a mean of 0 and a standard deviation of 1.

$$Z = (x - \mu)/\sigma$$

- Makes feature values comparable but without bounding them.
- Assumes data may have a normal distribution.

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

- The Variance Inflation Factor (VIF) measures the degree of multicollinearity in a set of independent variables in a regression model.
- Higher VIF indicates a higher degree of multicollinearity.

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Infinite VIF usually occurs when there is significant multicollinearity between one independent variable and another variable in the sample. This means that an independent variable is the optimal combination of one or more independent variables.

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

- A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically a normal distribution in the context of linear regression. It compares the quantiles of the observed data against the quantiles of a specified theoretical distribution.

Importance of a Q-Q plot in linear regression

- **Normality of Residuals:** One of the assumptions in linear regression is that the residuals are normally distributed. The Q-Q plot is used to check whether the residuals follow a normal distribution. If the residuals do not appear normal, it suggests that the model may not be well-fitted, or that the assumptions of linear regression are violated.
- **Identifying Outliers:** The Q-Q plot can help in detecting outliers by showing whether certain data points deviate significantly from the expected quantiles.
- **Improving Model Fit:** If the Q-Q plot shows that the residuals are not normally distributed, one might apply transformations to improve the model fit.