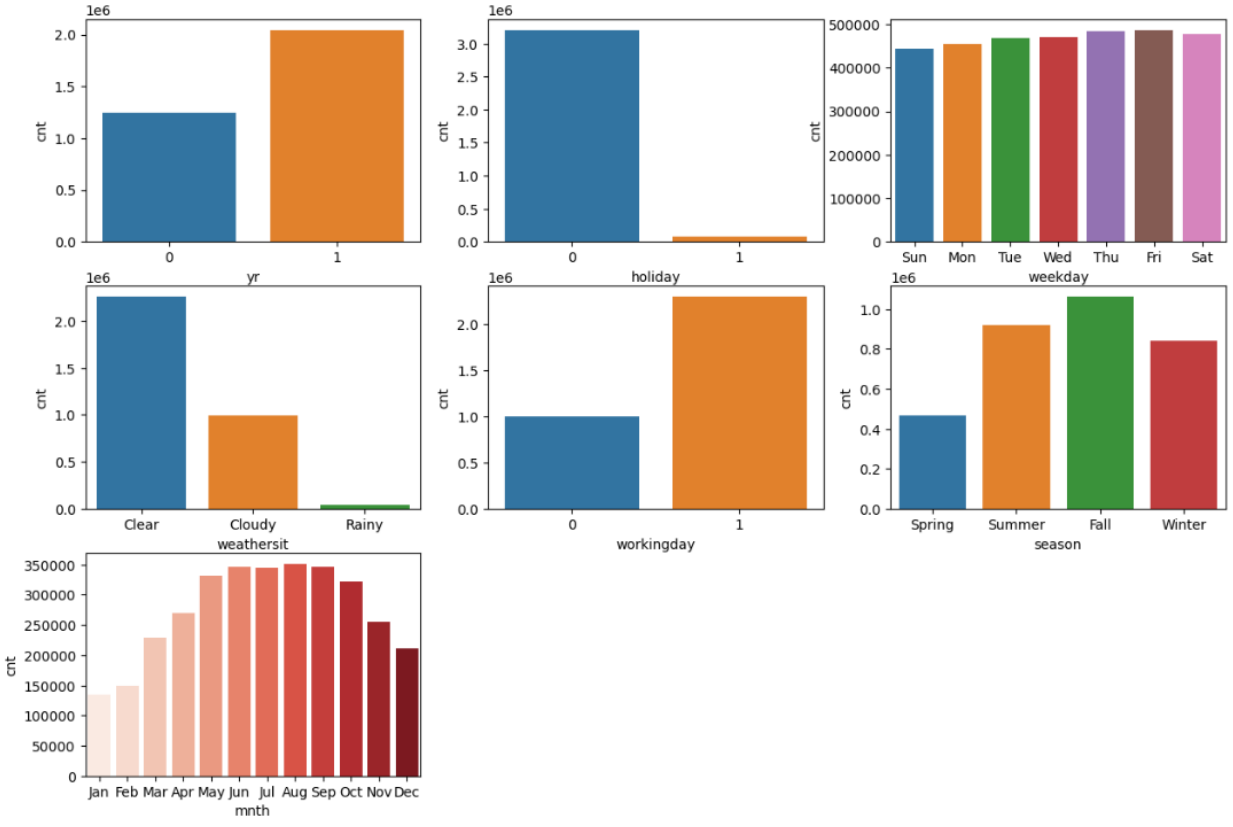
Assignment-based Subjective Questions

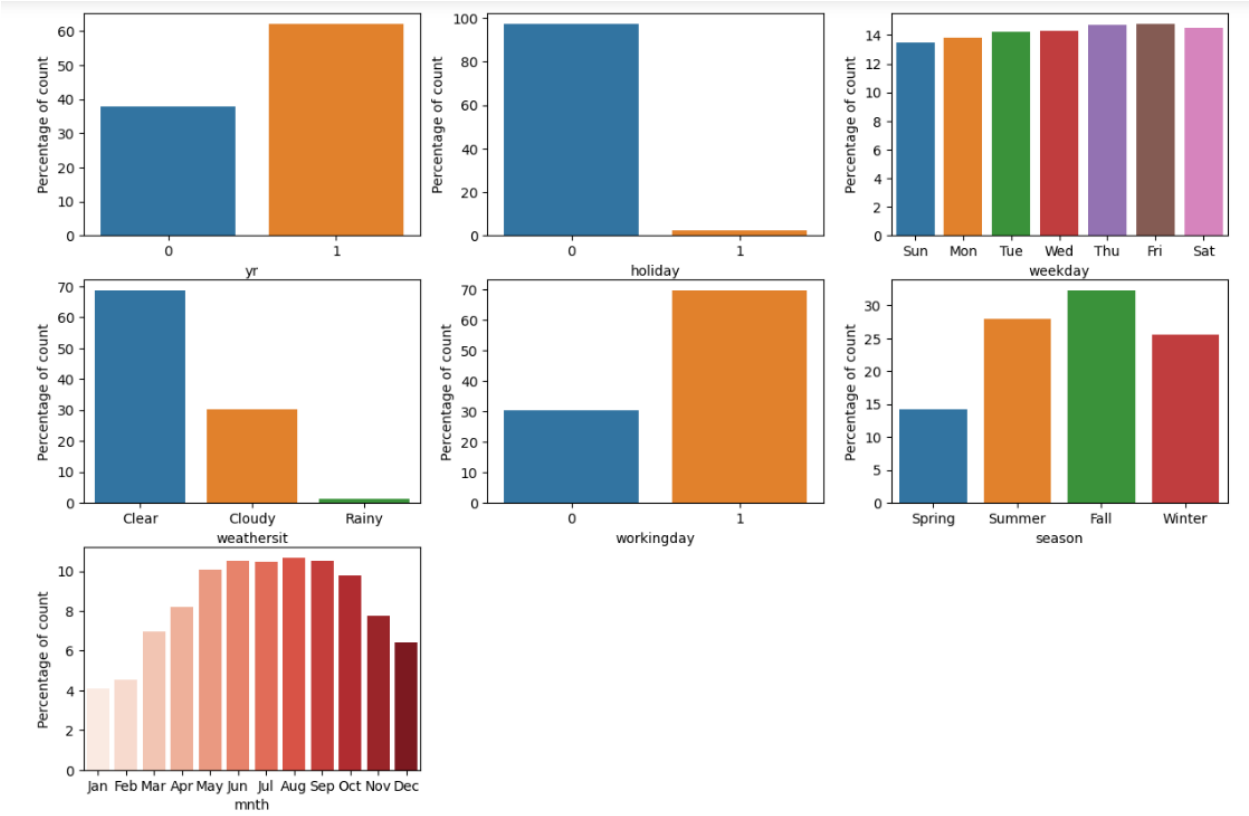
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   - Year(yr) - The rental count in 2019 is around **20% more** than in 2018
   - Holiday – Bikes are rented mostly on non holiday (**around 97%).**
   - Weekdays - The rental count during weekdays is almost the same.
   - Weather(weathersit) – Around **70%** bikes are rented on a clear day and least on snowy or rainy day
   - Working day - The rental count during working day is more than on non working day.
   - Season - Maximum number of bikes are rented in Fall, followed by Summer, Winter, Spring.

   Below are the graphs with count and all categorical variables



Below are the graphs with count percentage and all categorical variables



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

For a categorical column having k categories, it is sufficient to have k-1 variables to represent it. For example for category season having value spring, fall, summer and winter, we can use three dummy variables to represent values.

winter	summer	fall	spring
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

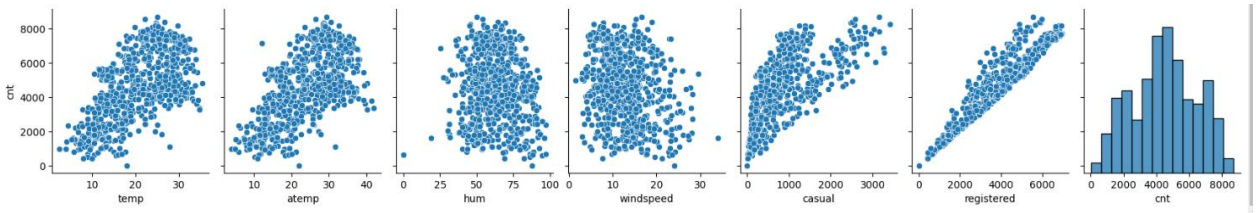
The season spring will be considered when values for all season related dummy categories are zero.

When performing linear regression with categorical variables using dummy variables, dropping the kth dummy variable (where k is the number of categories - 1) is essential to avoid multicollinearity issues. If you don't drop the kth dummy variable, you'll likely encounter problems in the interpretation of your regression results and potentially lead to unreliable model estimates.

To avoid these issues, it's standard practice to drop one dummy variable (usually the baseline category) when fitting a linear regression model with categorical variables. This choice does not affect the quality of the model or the accuracy of predictions, as long as you interpret the coefficients correctly in relation to the baseline category.

Also, keeping the kth dummy variable in a linear regression model can increase the complexity of the model, specifically in terms of the number of predictors. This is because including the kth dummy variable introduces an additional predictor variable to the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

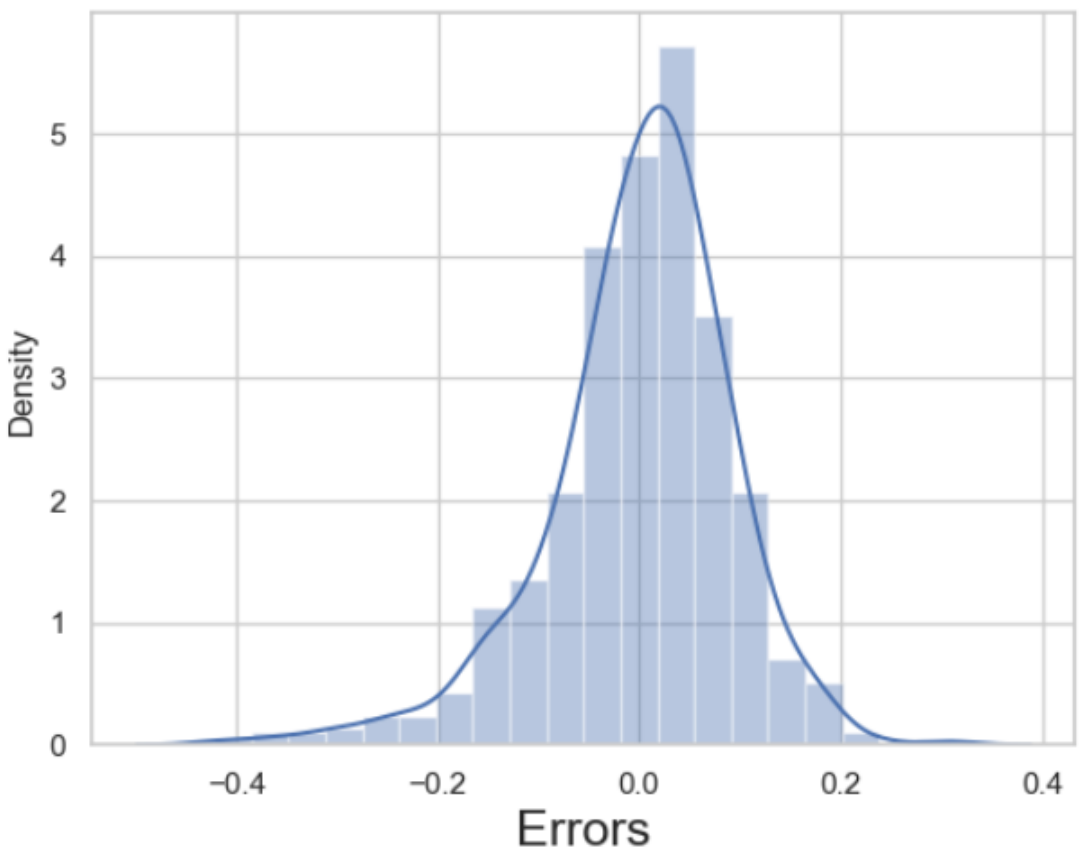


We can observe a linear relationship of cnt(target variable) with temp, atemp, casual and registered. Casual and registered variable are hightly correlated with cnt but it is because they are subset of cnt hence we need not consider it. Apart from this we can observe correlation with temp and atemp as well.

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

Examining Normality:

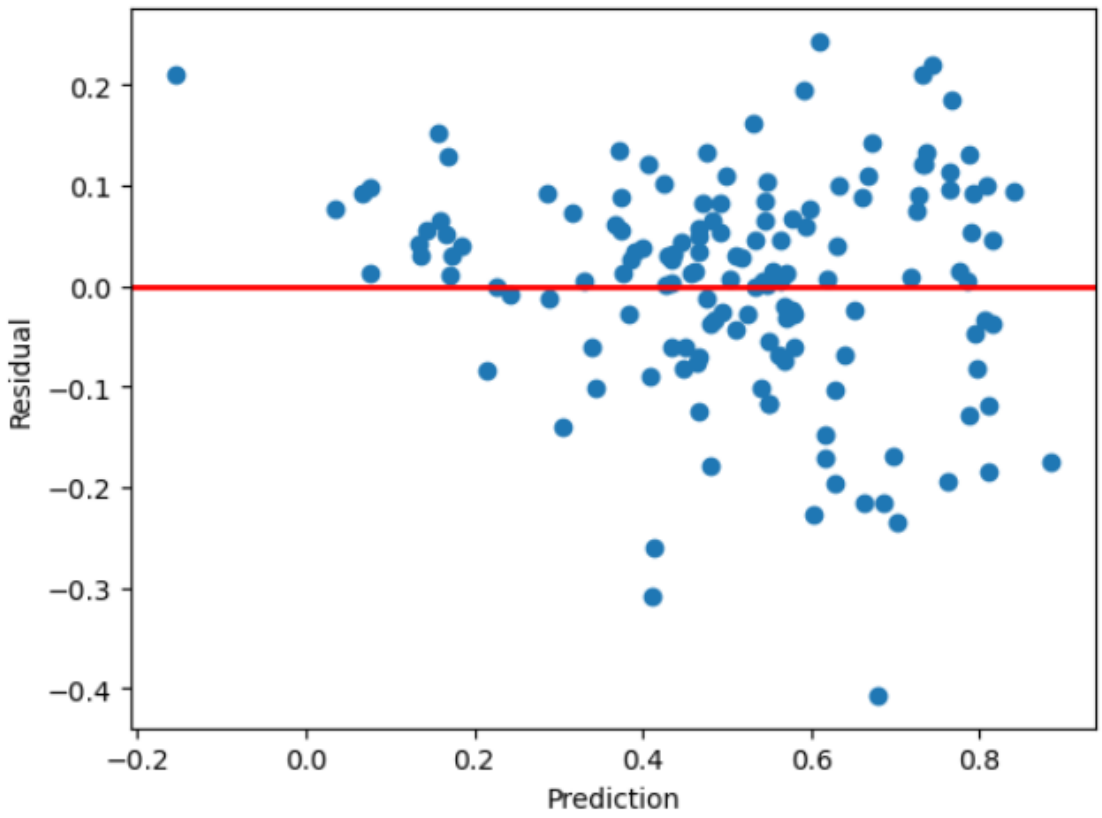
Histogram of Residuals: Plot a histogram of the residuals and compare it to a normal distribution. Deviations from normality may indicate issues with the assumption of normality. Following is the graph for Model 3



Linearity Assumption:

Residuals vs. Fitted Values Plot: Plot the residuals (the differences between the observed and predicted values) against the predicted values. Look for a random scatter of points around the zero line, indicating that the relationship between the predictors and the response is approximately linear.

Below screenshot is from Model 4:



Multicollinearity Assumption:

Correlation Matrix and VIF: Calculate the correlation matrix among predictor variables and compute the Variance Inflation Factor (VIF) to identify potential multicollinearity. VIF values above 5 or 10 are often considered indicative of multicollinearity.
Below is screenshot for Model 2

	Features	VIF
2	windspeed	3.91
0	year_2019	1.96
3	cloudy	1.52
10	april	1.46
9	march	1.46
8	february	1.39
16	october	1.36
11	may	1.33
14	august	1.33
18	december	1.32
17	november	1.31
12	june	1.28
15	september	1.28
7	saturday	1.27
5	wednesday	1.26
13	july	1.24
6	friday	1.22
4	rainy	1.10
1	holiday	1.06

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

Below features are common in all the models created for Bike Rental agency

- Year(yr) –major contribution
- Windspeed
- Rainy season

Hence, these can be considered as top 3 features

General Subjective Questions

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

A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X.

- The dependent variable is also called the **response variable**.
- Independent variables are also called **explanatory** or **predictor variables**.
- Continuous predictor variables are also called **covariates**.
- Categorical predictor variables are also called **factors**.

A multiple linear regression model is

$$y_i = \beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}+\varepsilon_i, \quad i=1,\cdots,n,$$

where

- n is the number of observations.
- yi is the ith response.
- βk is the kth coefficient, where β0 is the constant term in the model. Sometimes, design matrices might include information about the constant term.
- Xij is the ith observation on the jth predictor variable, j = 1, ..., p.
- εi is the ith noise term, that is, random error.

Some examples of linear models are:

$$y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}+\varepsilon_i$$

$$y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{3i1}+\beta_4 X_{2i2}+\varepsilon_i$$

$$y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i1}X_{i2}+\beta_4 \log X_{i3}+\varepsilon_i$$

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

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on 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.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

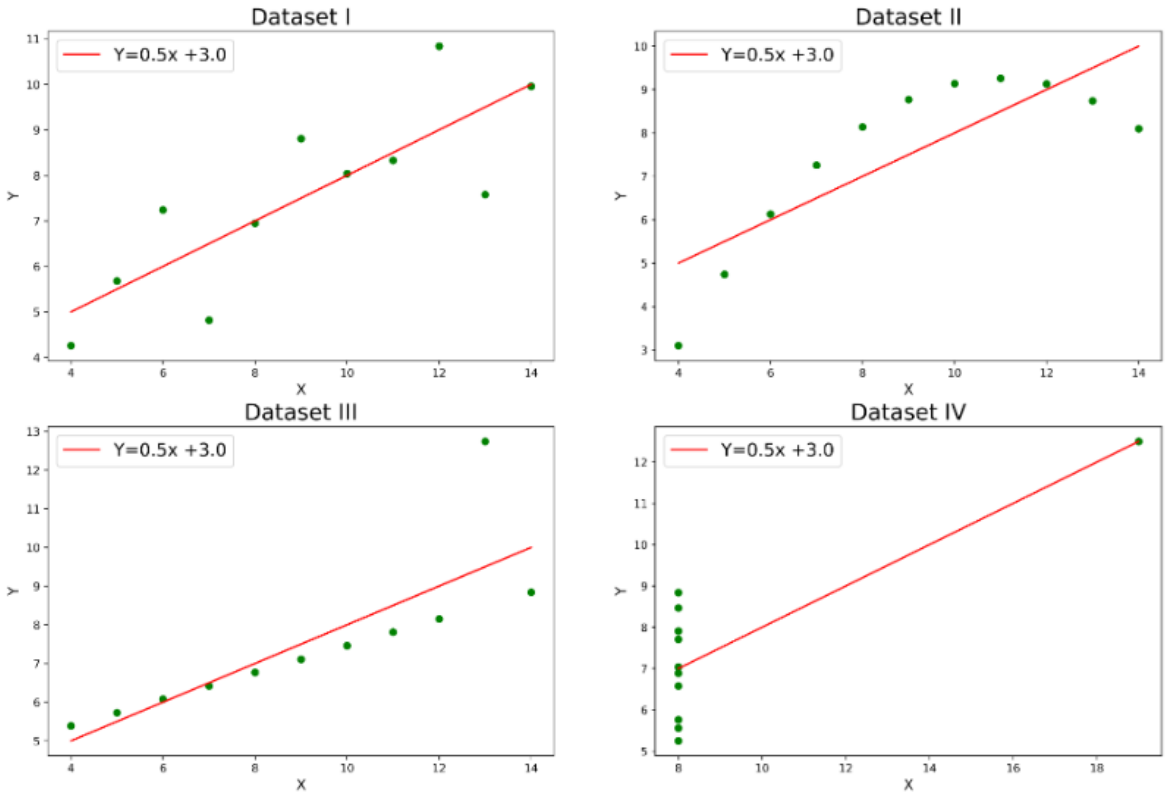
Data set

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

Mean, Variance, Correlation, Linear regression slop and intercept of Data set

	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

Graphical representation of Data set



The quartet highlights the principle that even though two datasets might have the same means, variances, and correlations, they can exhibit entirely different structures. This underscores the significance of exploratory data analysis and the power of data visualization in gaining a deeper understanding of data.

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

Pearson's correlation coefficient, often denoted as Pearson's "r," is a statistical measure that quantifies the linear relationship between two continuous variables. It measures the strength and direction of the linear association between the variables. Pearson's correlation coefficient ranges from -1 to 1, where:

- A value of 1 indicates a perfect positive linear correlation, meaning that as one variable increases, the other variable also increases proportionally.
- A value of -1 indicates a perfect negative linear correlation, meaning that as one variable increases, the other variable decreases proportionally.

- A value of 0 indicates no linear correlation, implying that there is no consistent linear relationship between the variables.

Pearson's correlation coefficient is calculated using the following formula:

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

Where:

- n is the number of data points (observations).
- xi and yi are the individual data values of the two variables.
- $\bar{x}$ and $\bar{y}$ are the means (averages) of the x and y values, respectively.

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

Scaling is a preprocessing step in data preparation where the numerical values of features (variables) in a dataset are transformed to fit within a specific range or distribution.

The goal of scaling is to bring all the features to a comparable level, which can improve the performance and convergence of various machine learning algorithms.

There are two common types of scaling: normalized scaling and standardized scaling.

<u>Normalized Scaling (Min-Max Scaling):</u>

- In normalized scaling, also known as min-max scaling, the values of the features are linearly transformed to fit within a specific range, typically [0, 1].
- The formula for normalized scaling of a feature x is:
  x_scaled = (x - x_min) / (x_max - x_min)
- This scaling maintains the original distribution of the data and is useful when you want to preserve the relationships between the data points.

<u>Standardized Scaling (Z-score Scaling):</u>

- In standardized scaling, the values of the features are transformed such that the resulting distribution has a mean of 0 and a standard deviation of 1.
- The formula for standardized scaling of a feature x is:

x_scaled = (x - mean) / standard_deviation

- This scaling centers the data around zero and adjusts the spread of the data. It does not preserve the original distribution but is often useful when you want to give equal weight to all features regardless of their original units.

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

A situation where the VIF is calculated to be infinite usually occurs when perfect multicollinearity is present among the predictor variables. Perfect multicollinearity happens when one predictor variable can be exactly predicted from a linear combination of other predictor variables.

For example, consider a simple case where you have three predictor variables, and one of them is a constant multiple of the sum of the other two variables:

$X = 2Y+Z$

In this case, you can perfectly predict the value of X, using the linear combination of $2Y$ and Z. When you compute the VIF for X , it will be infinite because the formula for VIF involves dividing the variance of the coefficient estimate by its minimum possible value, which is zero due to perfect multicollinearity.

The formula for calculating VIF for a predictor variable X_j is:

$VIF = 1 / (1 - R^2)$

Where R^2 is the coefficient of determination obtained when regressing the independent variable in question against all the other independent variables in the model. VIF quantifies how much the variance of the estimated regression coefficient is increased due to multicollinearity.

Now, when the correlation between two or more independent variables is extremely high (close to perfect correlation), it can lead to a situation where the R^2 value is very close to 1. This makes the denominator of the VIF formula very close to 0 ($1 - R^2 \approx 0$), which leads to an extremely large value for VIF. In the mathematical limit, if two or more variables are perfectly correlated ($R^2 = 1$), the VIF becomes infinite.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Here's how a Q-Q plot works:

- **Quantiles**: Quantiles are points in a dataset that divide the data into equal-sized groups. For example, the median is a quantile that divides the data into two equal halves. Other quantiles include quartiles (dividing the data into four parts) and percentiles (dividing the data into 100 parts).
- **Creating the Q-Q Plot**: To create a Q-Q plot, you start by sorting the observed data in ascending order. Then, for each data point, you find the corresponding quantile in the theoretical distribution. For example, if you're testing for normality, you would find the expected quantiles from a standard normal distribution.
- **Plotting**: You then create a scatter plot with the observed data's quantiles on the x-axis and the expected quantiles from the theoretical distribution on the y-axis. If the data follows the theoretical distribution closely, the points on the plot will roughly form a straight line. Deviations from a straight line indicate departures from the theoretical distribution.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.