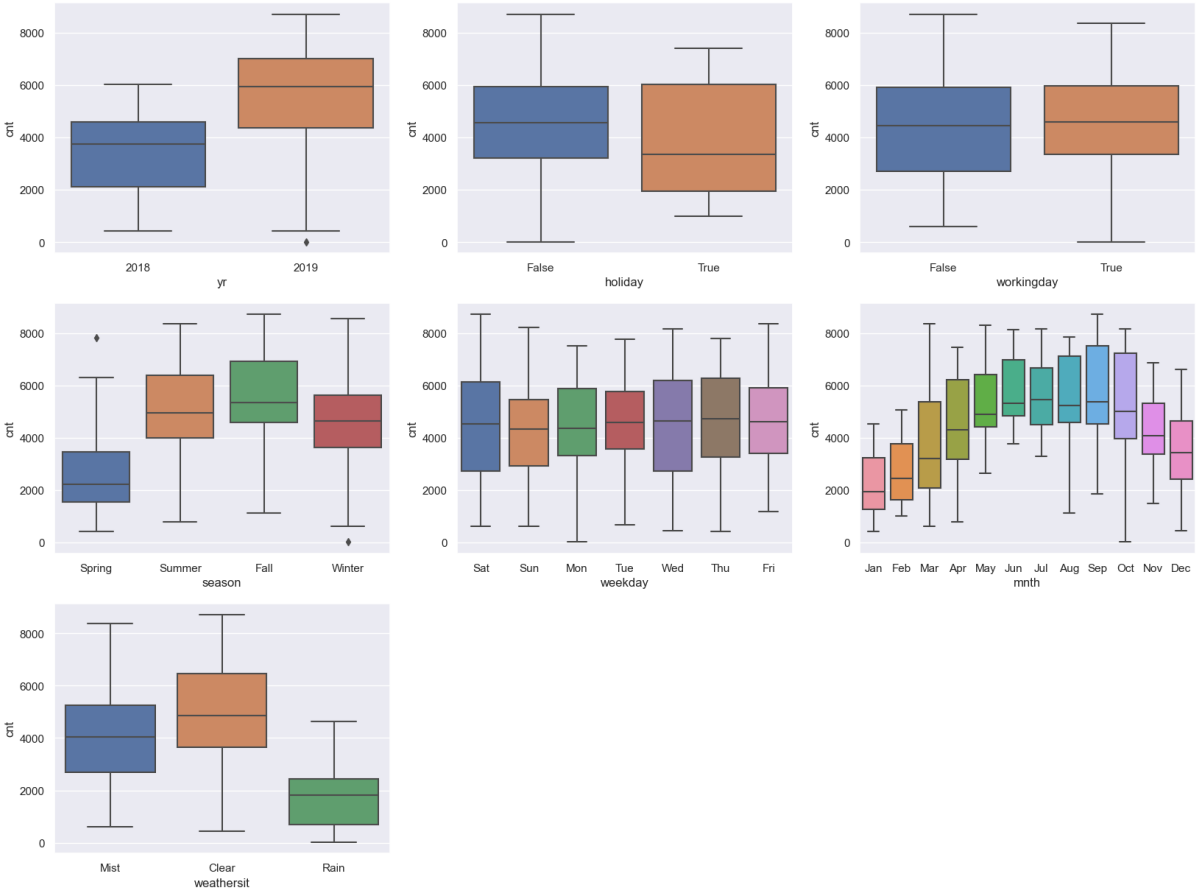
Assignment-based Subjective Questions

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



The categorical variable in the dataset were yr, holiday, workingday, season, weekday, mnth and weathersit. These were visualized using a boxplot. These variables had the following effect on our dependent variable:

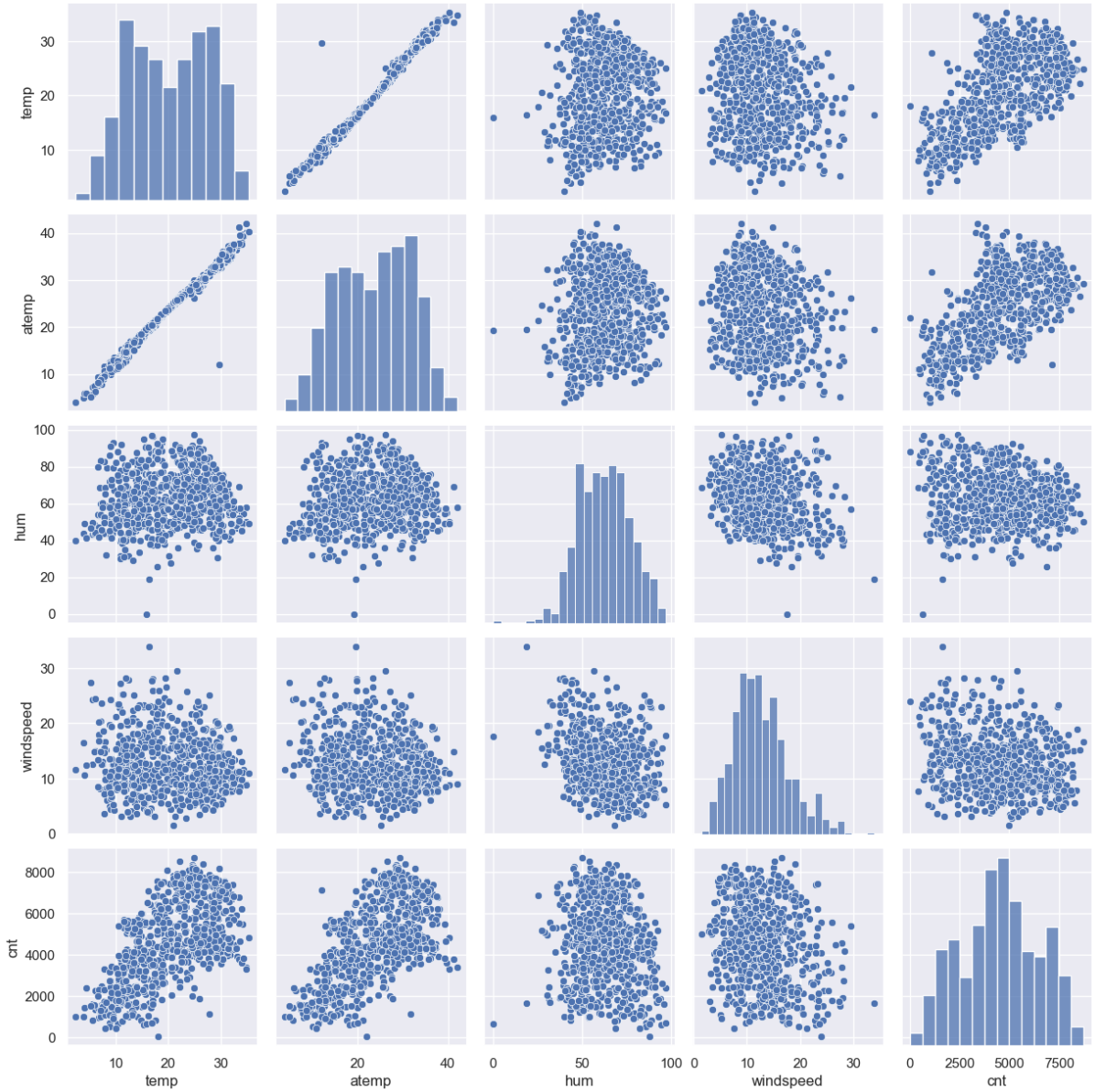
- Yr - The number of rentals in 2019 was more than 2018.
- Holiday – The number of rentals reduced during holiday.
- Workingday – The median count of users is not changed much almost throughout the week.
- Season - The boxplot showed that spring season had least value of cnt whereas fall season had maximum value of cnt. Summer and winter seasons had intermediate values of cnt.
- Mnth - September saw highest no of rentals while December saw least. This observation is in accordance with the observation made in weathersit. The weather situation in December is usually heavy snow due to which the rentals might have been down.
- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable. Highest count was seen when the weathersit was ‘Clear, Partly Cloudy’.

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

Because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Using the below pairplot it can be seen that, “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).

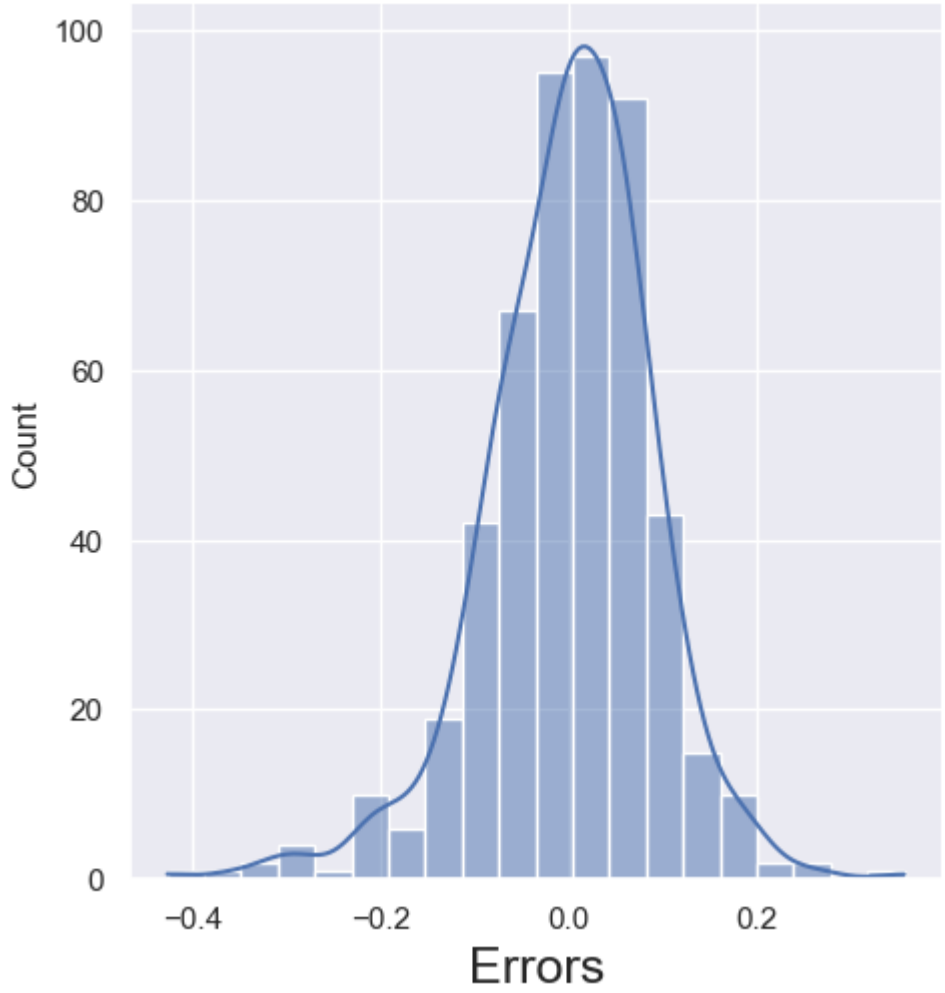


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

There exists a linear relationship between the independent variables as x and the dependent variable as y. This means that as the value of x changes, the value of y changes proportionally. This can be checked by plotting a scatter plot of x and y and looking for a

straight line pattern. If the relationship is not linear, a transformation of the variables or a non-linear model may be needed.

Residuals distribution should follow normal distribution and central around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.



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

The top 3 features are:

- temp - coefficient : 0.433502
- season_Winter - coefficient: 0.043219
- mnth_Sep - coefficient: 0.058635

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.

The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variables.

Linear regression computes the linear relationship between a dependent variable and one or more independent variables. The dependent variable is also known as the response or the target, and the independent variables are also known as the explanatory or the predictors.

The equation of a simple linear regression model with one independent variable is:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Where:

- y is the predicted value of the dependent variable y for given value of independent variable value x
- B0 is the intercept, the predicted value of y when x is 0
- B1 is the regression coefficient - how much we expect y to change as x changes
- x is the independent variable
- e is the error of the estimate or how much variation there in our estimate of the regression coefficient.

The equation of multiple linear regression model with more independent variables is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Where:

- B1 = coefficient for X1 variable
- B2 = coefficient for X2 variable
- B3 = coefficient for X3 variable and so on…
- B0 is the intercept (constant term)

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

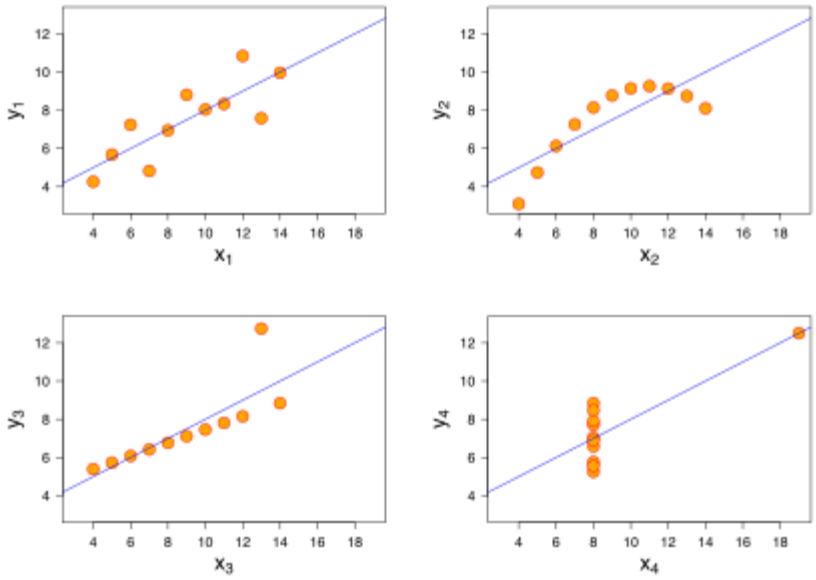
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics,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

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. **What is Pearson's R?**

Pearson's R is a measure of linear correlation between two sets of data. It is also known as Pearson's correlation coefficient, Pearson product-moment correlation coefficient, or simply the correlation coefficient.

Pearson's R is a number between -1 and 1 that indicates how strongly and in what direction two variables are related. A positive value of R means that the variables tend to increase or decrease together, while a negative value means that they tend to move in opposite directions. A value of 0 means that there is no linear relationship between the variables.

Pearson's R can be calculated using the following formula:

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

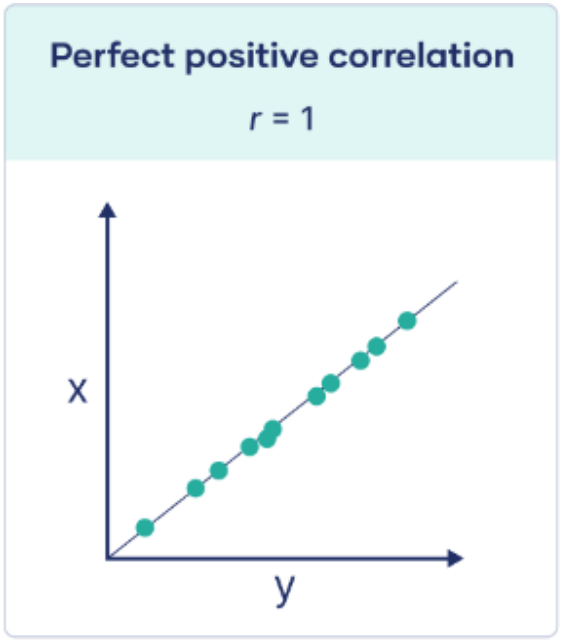
y_i = values of the y-variable in a sample

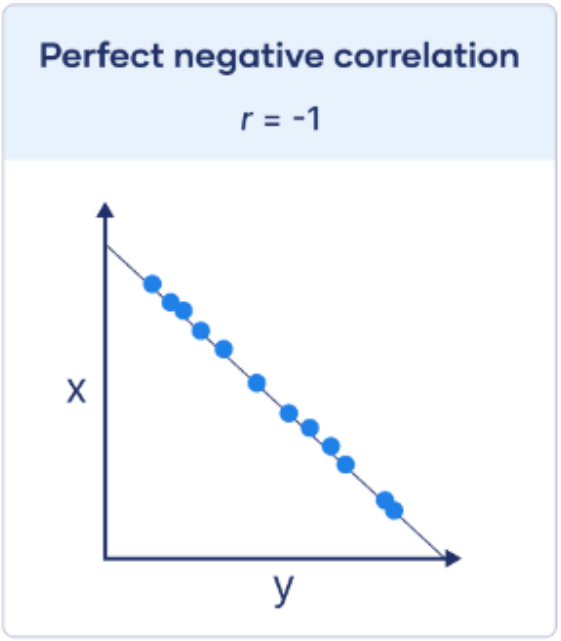
$\bar{y}$ = mean of the values of the y-variable

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

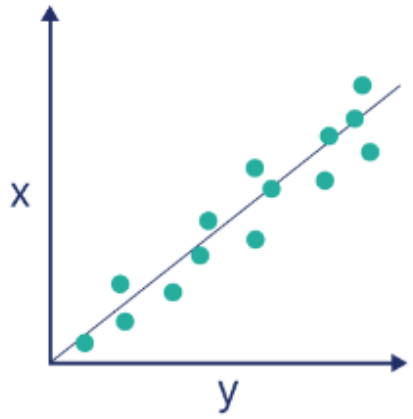
As can be seen from the graph below





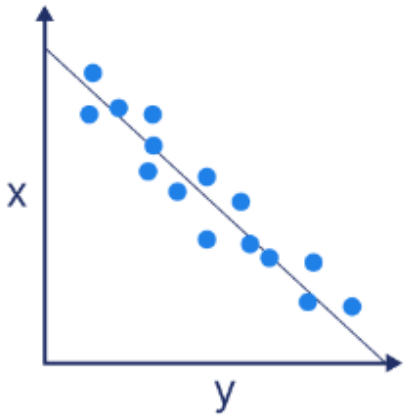
Strong positive correlation

$r > .5$



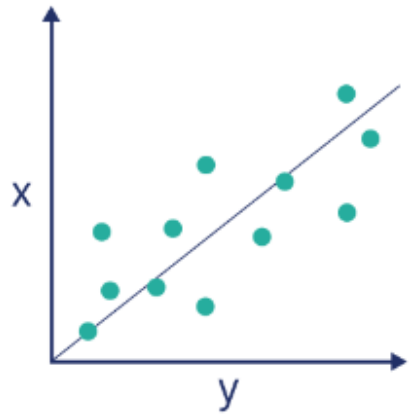
Strong negative correlation

$r < -.5$



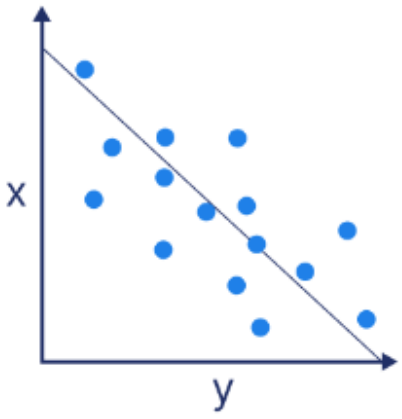
Weak positive correlation

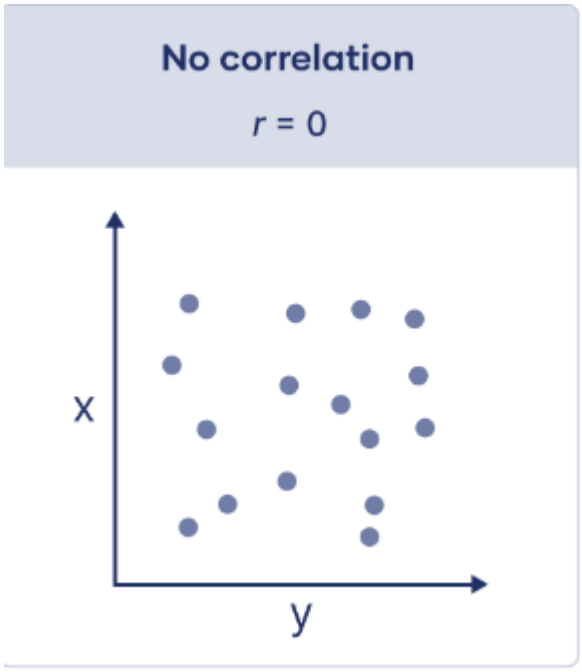
$.3 > r > 0$



Weak negative correlation

$0 > r > -.3$





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

Scaling is a data preprocessing technique that involves transforming the values of features or variables in a dataset to a similar scale. This is done to ensure that all features contribute equally to the model and to prevent features with larger values from dominating the model.

Scaling is performed when working with datasets where the features have different ranges, units of measurement, or orders of magnitude. Scaling can help to improve model performance, reduce the impact of outliers, and ensure that the data is on the same scale.

There are different types of scaling techniques, such as normalization and standardization. Normalization and standardization are both methods of rescaling the data, but they have different formulas and effects.

Normalization is a scaling technique that changes the range of the data to [0, 1] or [-1, 1]. It preserves the shape of the original distribution and does not reduce the importance of outliers. Normalization is useful when the data has a known or fixed range, and when preserving the original distribution is important.

Standardization is a scaling technique that changes the mean and standard deviation of the data to 0 and 1, respectively. It shifts the distribution to have a zero mean and unit variance. It reduces the effect of outliers and makes the data more comparable. Standardization is useful when the data has an unknown or varying range, and when reducing the influence of outliers is important.

The following table summarizes some of the differences between normalization and standardization:

Normalization	Standardization
Rescales the data to [0, 1] or [-1, 1]	Rescales the data to have zero mean and unit variance
Preserves the shape of the original distribution	Shifts the distribution to be more normal-like
Does not reduce the importance of outliers	Reduces the effect of outliers
Useful for data with known or fixed range	Useful for data with unknown or varying range
Formula: X' = (X – min(X))/(max(X) – min(X))	Formula: X' = (X – mean(X))/sd(X)

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

VIF stands for variance inflation factor, which is a measure of multicollinearity among the independent variables in a regression model. VIF quantifies how much the variance of a regression coefficient is inflated due to the presence of multicollinearity.

The value of VIF is calculated by the following formula:

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

The value of VIF can range from 1 to infinity. A value of 1 means that there is no multicollinearity among the variables. A value greater than 1 means that there is some degree of multicollinearity, and a value greater than 10 indicates a serious multicollinearity problem.

An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model. This happens when R-squared approaches 1, which means that the regression of the variable on the other variables has a perfect fit. This implies that there is a perfect linear relationship or correlation among some of the 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, short for quantile-quantile plot, is a type of plot that we can use to determine whether or not the residuals of a model follow a normal distribution. If the points on the plot roughly form a straight diagonal line, then the normality assumption is met.

A Q-Q plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).

A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q-Q plots can be used to compare collections of data, or theoretical distributions.

A Q-Q plot is useful in linear regression because it can help us check one of the four assumptions of linear regression: normality. This assumption states that the residuals of the model should be normally distributed. If this assumption is violated, then the confidence intervals and significance tests for the model may not be reliable.