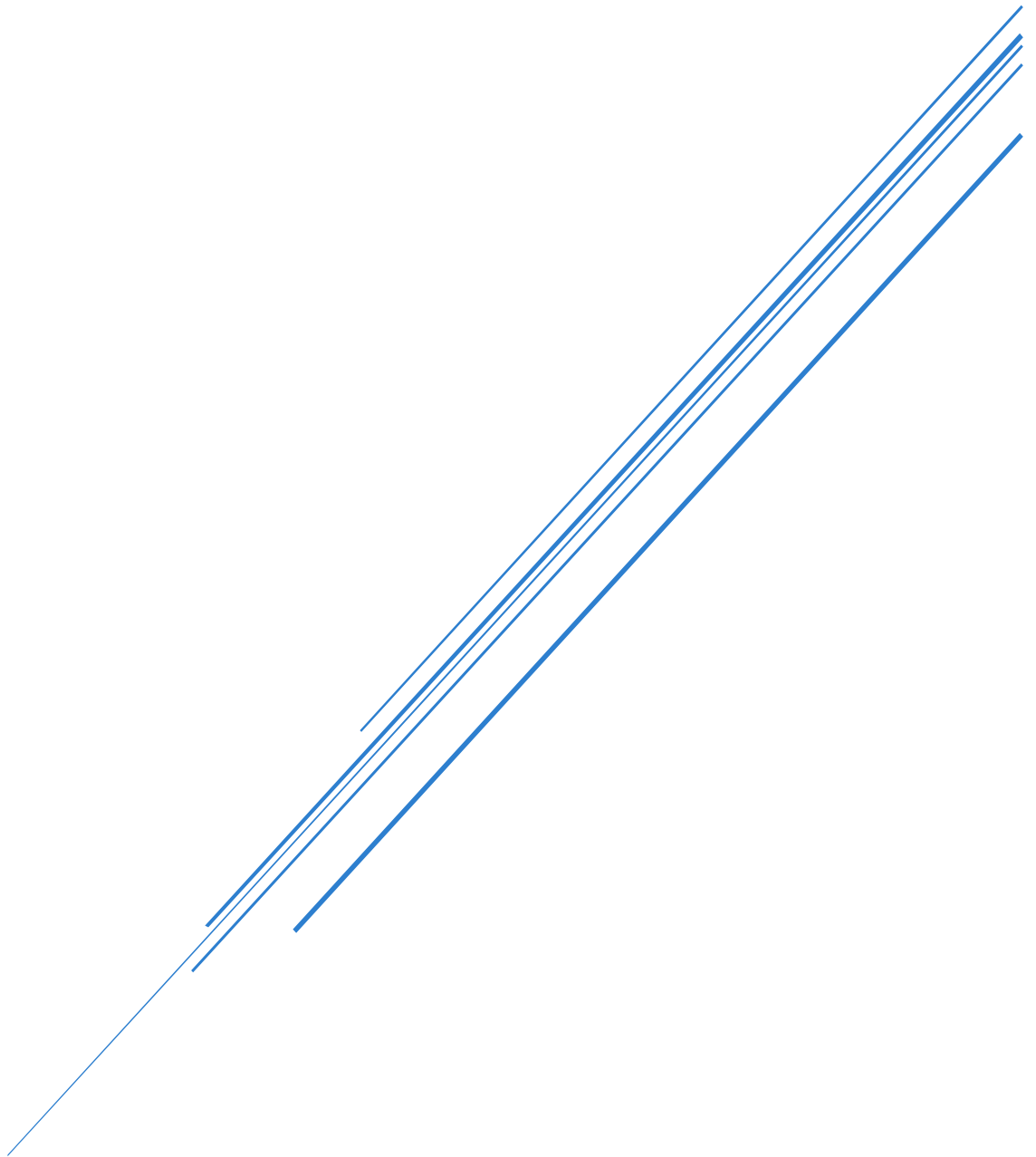


# Responses to Subjective Questions

Linear Regression Module



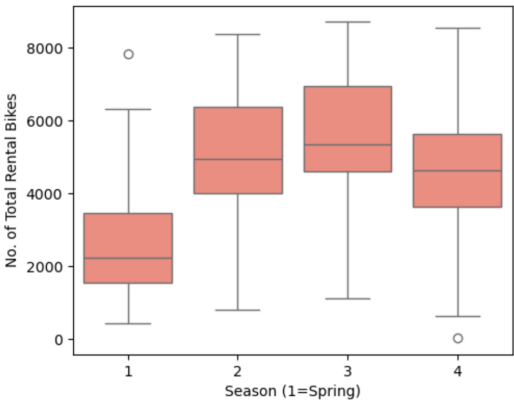
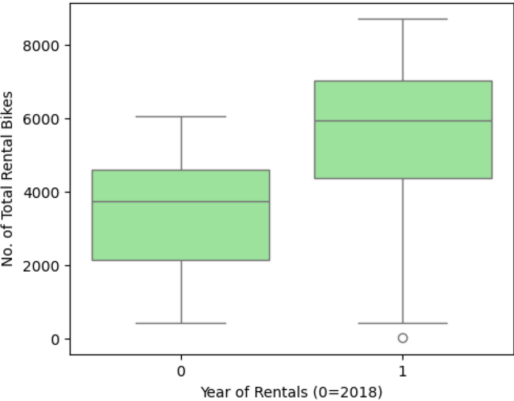
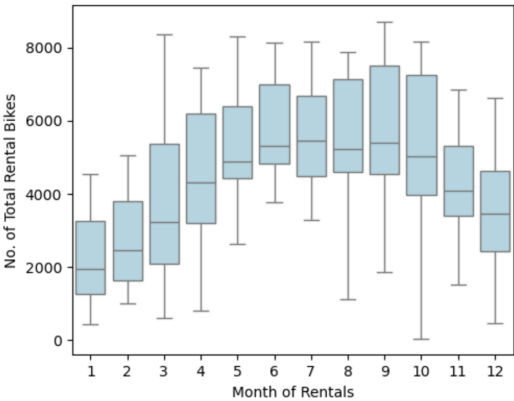
IIIT Bangalore  
Executive PG Programme in Machine Learning and AI

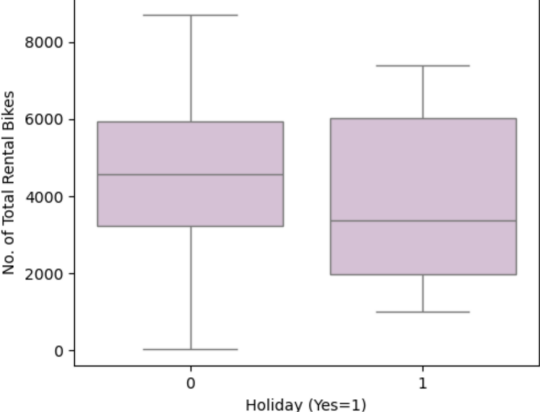
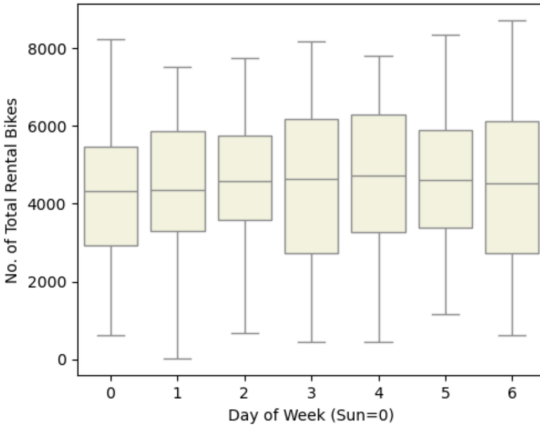
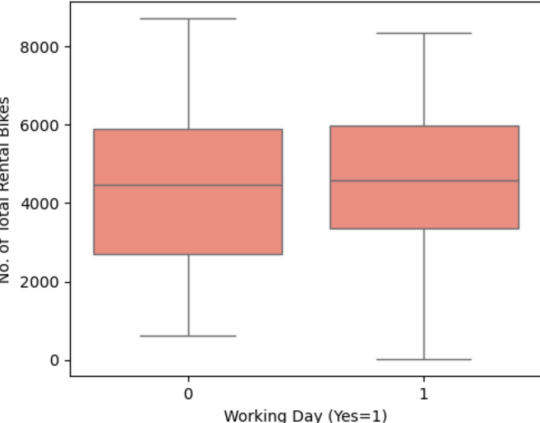
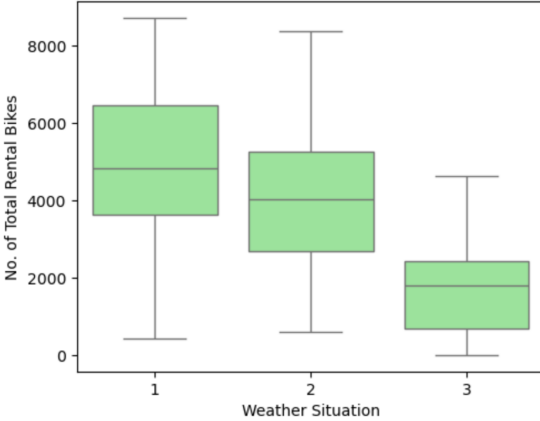
## Assignment-based Subjective Questions

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

Answer: Here are my inferences of the effect of categorical variables on dependent variable i.e. bike rental demand.

The categorical variables are: 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday' and 'weathersit'

Visualisation	Inference
	<p>1) Fall &amp; Summer seasons attract more number of users to rent bikes than Spring or Winter</p> <p><i>Legends: 1:spring, 2:summer, 3:fall, 4:winter</i></p>
	<p>2) Year 2019 saw more rentals than 2018 indicating a better awareness and hence usage of the rental facility year on year</p> <p><i>Legends: 0: 2018, 1:2019</i></p>
	<p>3) Month-wise, July and Sept saw higher demands while Jan &amp; Feb were lean months</p> <p><i>Legends: 1: Jan .... 12: Dec</i></p>

 <p>No. of Total Rental Bikes</p> <p>Holiday (Yes=1)</p>	<p>4) There are relatively less rentals on holidays than other days</p> <p><i>Legends: 0: Non-holiday, 1:Holiday</i></p>
 <p>No. of Total Rental Bikes</p> <p>Day of Week (Sun=0)</p>  <p>No. of Total Rental Bikes</p> <p>Working Day (Yes=1)</p>	<p>5) There were not much variation in demands across days of the week with working days getting slightly higher demands</p> <p><i>Legends: 0: Sun .... 6: Sat</i></p> <p><i>Legends: 0: Non-Working Day, 1: Working Day</i></p>
 <p>No. of Total Rental Bikes</p> <p>Weather Situation</p>	<p>6) Clear days attract more users to rent bikes than other weather situations</p> <p>7) There are no heavy rain or heavy snow days in the subject area</p> <p><i>Legends: 1: Clearday, 2: Cloudyday, 3: Wetday, 4: Rainyday</i></p>

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

**Answer:** Dummy variables are created corresponding to different unique values of a categorical variable in a dataset. Each dummy variable can have values of 0 or 1 indicating the absence or presence of that variable.

Typically if there are **n** unique values of a categorical variable, we need **n-1** dummy variables to represent them. We drop one dummy variable which is, by default, represented by the absence of other dummy variables i.e. when the values of all other dummy variables are zeros. This helps to improve the linear model by one degree of freedom. The use of **drop\_first=True** accomplishes this objective. Hence it's important to use it during dummy variable creation.

**Example:** In Bike Rentals dataset, we have a categorical variable called *weathersit*. It had 4 unique values:

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

The above 4 values were mapped to short names of Clearday, Cloudyday, Wetday and Rainyday respectively. In fact the dataset didn't have any observation with 4<sup>th</sup> weathersit i.e. Rainyday. Out of the 3 weathersit values that were present, the algorithm to create dummy variables generated only two of them using **drop\_first=True** clause. It dropped the first variable i.e. Clearday. The model interpreted an observation to be Clearday when values of both Cloudyday and Wetday were zero.

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

**Answer:** The target variable for the Bike Rentals model is *cnt* i.e. count of total rental bikes including both casual and registered users. The pair-plot was created using following numerical variables: *temp*, *atemp*, *hum*, *windspeed*, *casual*, *registered*, *cnt*.

Among these, 'registered' has the highest correlation with 'cnt'. But eventually, 'registered' and 'casual' variables were dropped from model building as they were included in target variable itself. Keeping aside these two variables, the next ones that have highest correlation is 'temp' and 'atemp' - both have similar correlation with 'cnt'.

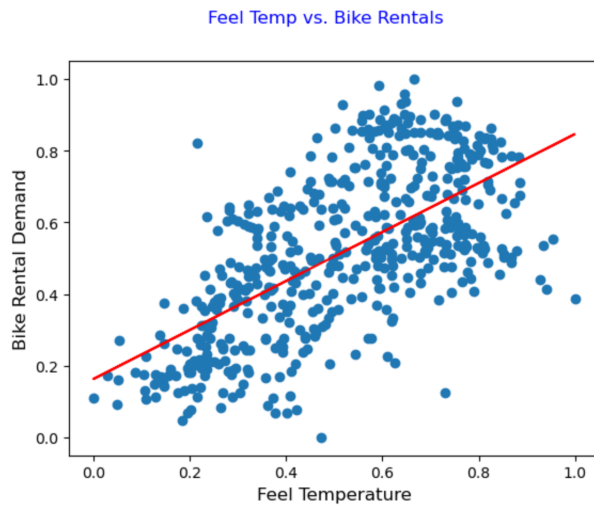
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

**Answer:** There are 4 assumptions in Linear Regression models. They are as follows:

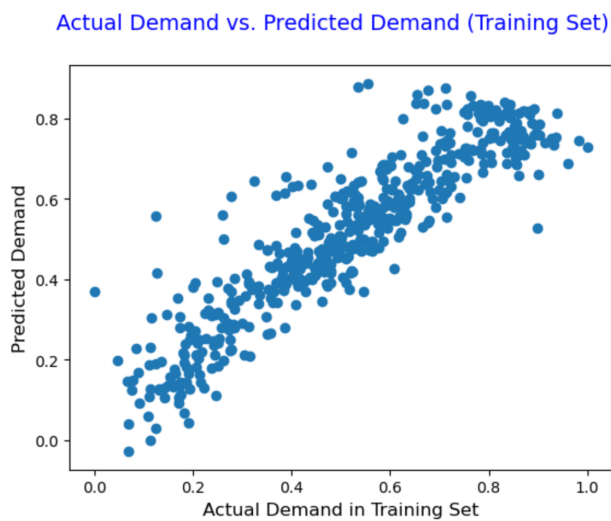
- 1) There is a linear relationship between independent variables and target variable
- 2) Error Terms (Residuals) are normally distributed with mean zero
- 3) Error terms are *independent* of each other

#### 4) Error terms have constant variance (homoscedasticity)

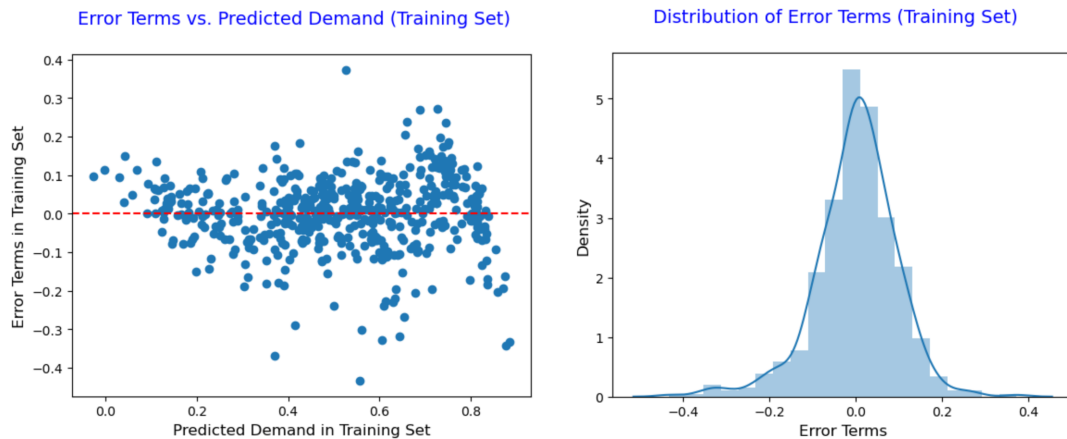
The **first** assumptions were validated by plotting the results of initial model using most prominent variable (atemp). The linear relationship was clear from the visualisation.



Subsequently, after finalising the model, we plotted actual vs. predicted values of target variable in a scatter plot (fig below). Here the points lie along a straight line with a slope of approximately 1. This further validated the **linear relationship among independent and target variables**.



The other three assumptions were validated by plotting the Error Terms in following two charts.



As can be seen from the above right chart, the Error Terms are normally distributed with mean zero thus validating **second** assumption.

The above left chart shows that the error terms are randomly distributed around  $y=0$  horizontal line and there are no patterns among them. This validates that they are independent which is the **third** assumption.

The above right chart further displays that the error terms are lying within a narrow range of mean  $\pm 0.2$  thereby validating that there exist constant variance among the error terms which is the **fourth** assumption (homoscedasticity).

With the above visualisations all the assumptions of linear regression models were validated.

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

Answer: The linear model equation for best fitted line came out to be as follows:

$$cnt = 0.268 + 0.235 \text{ yr} - 0.089 \text{ holiday} + 0.421 \text{ temp} - 0.145 \text{ windspeed} - 0.076 \text{ Cloudyday} - 0.282 \text{ Wetday} - 0.119 \text{ Spring} + 0.047 \text{ Winter}$$

Since each of the variables in the above equation was scaled using MinMax Scaler, their regression coefficients represent how much each of these variable contribute towards explaining the demands for shared bikes. By looking at the values of the coefficients, the top 3 contributing features are:

**temp** (Temperature in Degree Celsius) - As the temperature increases (in otherwise cold climate), bike demand will grow significantly

**Wetday** (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) – Demand will be less compared to Clearday on days with light rain/ snow.

**Yr** (year of demand) - As the year progresses, bike demand is expected to rise, because of increased awareness about the facility among the users

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

**Answer:**

Linear regression is one of the simplest and most commonly used algorithms in machine learning and statistics. It models the relationship between a dependent variable (response) and one or more independent variables (predictors) by fitting a linear equation to the observed data. Here is a detailed explanation of the linear regression algorithm:

### 1. Model Representation

In simple linear regression, the relationship between the dependent variable  $y$  and the independent variable  $x$  is modelled as a straight line:

$$[y = \beta_0 + \beta_1 x + \epsilon]$$

where:

- $y$  is the dependent variable.
- $x$  is the independent variable.
- $\beta_0$  is the intercept.
- $\beta_1$  is the slope of the line.
- $\epsilon$  is the error term, representing the difference between the observed and predicted values.

In multiple linear regression, the model includes multiple independent variables:

$$[y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon]$$

where  $(x_1, x_2, \dots, x_p)$  are the independent variables.

### 2. Assumptions

Linear regression relies on several key assumptions:

1. **Linearity:** The relationship between the dependent and independent variables is linear.
2. **Independence:** The residuals (errors) are independent.
3. **Homoscedasticity:** The residuals have constant variance at every level of  $x$ .
4. **Normality:** The residuals are normally distributed.

### 3. Estimation of Coefficients

The coefficients  $(\beta_0, \beta_1, \dots, \beta_p)$  are estimated using the method of least squares, which minimizes the sum of the squared differences between the observed and predicted values. This is often formulated as:

$$\min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where  $y_i$  is the observed value and  $\hat{y}_i$  is the predicted value for the  $i$ -th observation.

#### 4. Fitting the Model

To fit the linear regression model, we use the following steps:

1. **Formulate the Design Matrix  $X$ :**

- The design matrix  $X$  contains the values of the independent variables.
- For a dataset with  $n$  observations and  $p$  predictors,  $X$  is an  $n \times (p + 1)$  matrix (including a column of ones for the intercept).

2. **Calculate the Coefficients:**

- The coefficients  $\beta$  can be calculated using the Normal Equation:

$$[\beta = (X^T X)^{-1} X^T y]$$

Here,  $X^T$  is the transpose of the design matrix  $X$ , and  $y$  is the vector of observed values.

#### 5. Making Predictions

Once the coefficients are estimated, we can make predictions using the linear model:

$$[\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p]$$

where  $\hat{y}$  is the predicted value.

#### 6. Model Evaluation

To evaluate the performance of the linear regression model, we use metrics such as:

- **R-squared (Coefficient of Determination):** Measures the proportion of variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where  $\bar{y}$  is the mean of the observed values.

- **Mean Squared Error (MSE):** The average of the squared differences between observed and predicted values.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **Adjusted R-Squared Error:** Adjusted R-square accounts the number of predictors in the model and penalizes the model for including irrelevant predictors that don't contribute significantly to explain the variance in the dependent variables.



$$Adjusted R^2 = 1 - \left( \frac{(1 - R^2) \cdot (n - 1)}{n - k - 1} \right)$$

- $n$  is the number of observations
- $k$  is the number of predictors in the model
- $R^2$  is coefficient of determination

## 7. Handling Multiple Collinearity

In multiple linear regression, multicollinearity (high correlation between predictors) can be a problem. It can be detected using Variance Inflation Factor (VIF) and addressed by:

- Removing highly correlated predictors.
- Combining predictors (e.g., using Principal Component Analysis).
- Using regularization techniques like Ridge Regression or Lasso Regression.

VIF measures how much the variance of a regression coefficient is inflated due to collinearity with other predictors. It is calculated as:

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

where  $R_i^2$  is the R-squared value obtained by regressing the  $i^{\text{th}}$  predictor on all other predictors.

## 8. Conclusion

Linear regression is a powerful and interpretable algorithm for modelling relationships between variables. By understanding its assumptions, fitting the model, and evaluating its performance, we can make informed decisions and predictions based on data.

2. Explain the Anscombe's quartet in detail.

**Answer:**

**Objective:** 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.

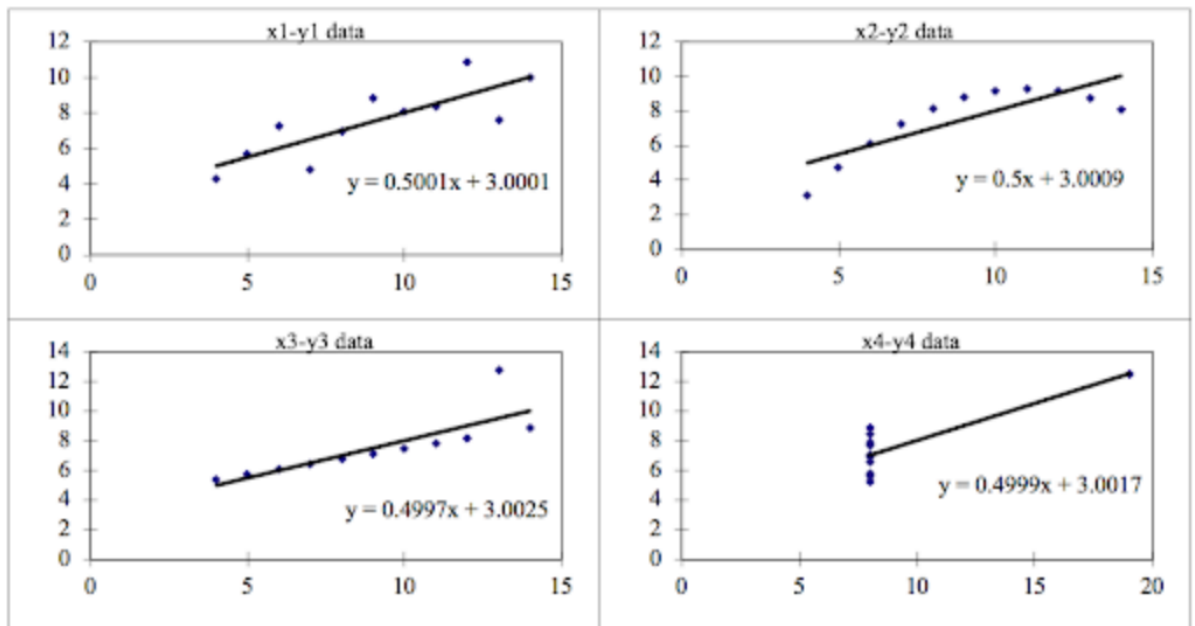
**What is it:** Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe which consists of a group of four data sets that have nearly the same statistical observations, involving variance and mean. But when the same data is plotted, the difference in their spread becomes quite obvious.

**The dataset and Summary Statistics:** The four data sets and their summary information is as below:

Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

As can be observed above, the mean, standard deviation and correlation coefficient for all the four data sets seem identical.

**The Plots:** The below plots are arrived at when these data sets are visualised on scatter plots:



As can be seen from above plots, each of the data sets present a different pattern. The patterns can be described as below:

- **Data Set 1:** Fits the linear regression model well with some variance
- **Data Set 2:** Data is non-linear, having the shape of a curve
- **Data Set 3:** Tight linear relationship between x and y, except for one large outlier
- **Data Set 4:** Value of x remains constant, except for one outlier as well

**Conclusion:** Data-sets may have identical number of statistical properties, yet produce dissimilar graphs. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

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

**Answer:**



Pearson's  $r$ , also known as Pearson's correlation coefficient or the Pearson product-moment correlation coefficient (PPMCC), is a statistical measurement of the strength and direction of a linear relationship between two quantitative variables. It's the most common way to measure linear correlation, say between a person's age and salary.

#### Strength & Direction of Correlation:

With a correlation analysis we can determine:

- How strong the correlation is
- In which direction the correlation goes.

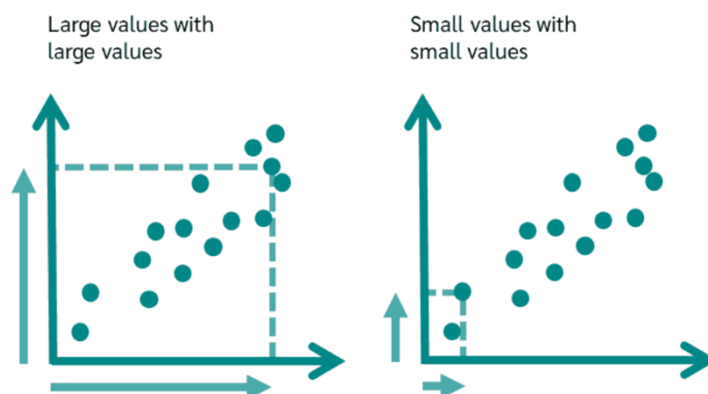
Amount of $r$	Strength of correlation
$0.0 < 0.1$	no correlation
$0.1 < 0.3$	low correlation
$0.3 < 0.5$	medium correlation
$0.5 < 0.7$	high correlation
$0.7 < 1$	very high correlation

From Kuckartz et al.: Statistik, Eine verständliche Einführung, 2013, p. 213

We can read the strength and direction of the correlation in the Pearson correlation coefficient  $r$ , whose value varies between -1 and 1.

An  $r$  between 0 and 0.1 indicates no correlation. An amount of  $r$  between 0.7 and 1 indicates a very strong correlation.

#### Positive correlation

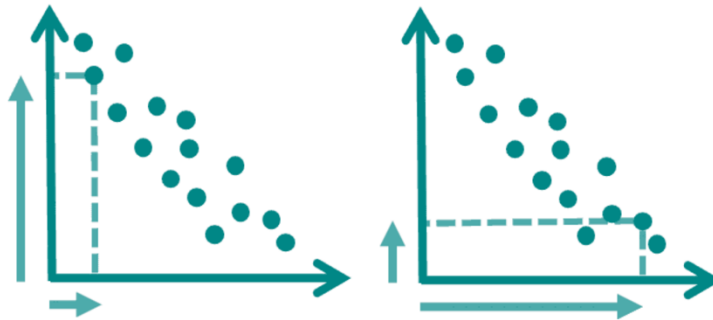


A positive relationship or correlation exists when large values of one variable are associated with large values of the other variable, or when small values of one variable are associated with small values of the other variable.

A positive correlation results, for example, between height and shoe size, in a positive correlation coefficient ( $r > 0$ ).

## Negative correlation

Large values with small values



A negative correlation is usually found between product price and sales volume. This results in a negative correlation coefficient ( $r < 0$ ).

## Computing Pearson's Correlation:

The Pearson correlation coefficient is calculated using the following equation.

$r$  = Pearson correlation coefficient

$x_i$  = individual values of one variable e.g. age

$y_i$  = Individual values of the other variable e.g. salary

$\bar{x}$  and  $\bar{y}$  = Mean values of the two variables respectively.

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

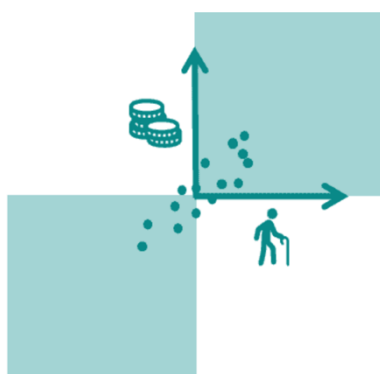
Where  $r$  is the Pearson correlation coefficient,

$x_i$  are the individual values of one variable e.g. age

$y_i$  are the individual values of the other variable e.g. salary

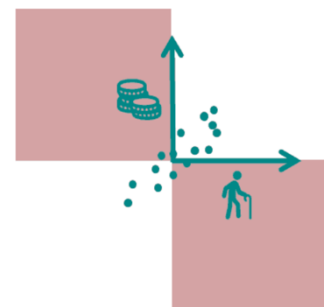
$\bar{x}$  and  $\bar{y}$  are respectively the mean values of the two variables.

So in our example, we calculate the mean values of age and salary. We then subtract the mean values from each of age and salary. We then multiply both values. We then sum up the individual results of the multiplication. The expression in the denominator ensures that the correlation coefficient is scaled between -1 and 1.



If we now multiply two positive values we get a positive value. If we multiply two negative values we also get a positive value (minus times minus is plus). So all values that lie in these ranges have a positive influence on the correlation coefficient.

If we multiply a positive value and a negative value we get a negative value (minus times plus is minus). So all values that are in these ranges have a negative influence on the correlation coefficient.



Therefore, if our values are predominantly in the two green areas from previous two figures, we get a positive correlation coefficient and therefore a positive correlation.

If our scores are predominantly in the two red areas from the figures, we get a negative correlation coefficient and thus a negative correlation.

If the points are distributed over all four areas, the positive terms and the negative terms cancel each other out and we might end up with a very small or no correlation.

### Testing Correlation Coefficient for Significance:

The Pearson correlation coefficient can also be used to test whether the relationship between two variables is significant. For this, we test whether the correlation coefficient in the sample is statistically significantly different from zero.

The null hypothesis and the alternative hypothesis in Pearson correlation are thus:

**Null hypothesis:** The correlation coefficient is not significantly different from zero (There is no linear relationship).

**Alternative hypothesis:** The correlation coefficient deviates significantly from zero (there is a linear correlation).

It is always tested whether the null hypothesis is rejected or not rejected.

Whether the Pearson correlation coefficient is significantly different from zero based on the sample surveyed can be checked using a t-test.

$r$  is the correlation coefficient

and  $n$  is the sample size

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

A *p-value* can then be calculated from the test statistic *t*. If the *p-value* is smaller than the specified significance level, which is usually 5%, then the null hypothesis is rejected, otherwise it is not.

### When to use Pearson's Correlation Coefficient:

The Pearson correlation coefficient is a good choice when **all** of the following are true:

- Both variables are quantitative
- The variables are normally distributed
- The data have no outliers
- The relationship is linear

If the assumptions are not met, other types of correlation coefficients can be used.

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

**Answer:**

Scaling is a technique to resize the independent variables present in a dataset to a fixed range in order to bring all values to same magnitudes. Generally performed during the data pre-processing step. **Used in Linear Regression, K-means, KNN, PCA, Gradient Descent etc.**

#### Why is scaling performed?

In machine learning, feature scaling is employed for a number of reasons:

- **Magnitude Comparability:** Scaling guarantees that all features are on a comparable scale and have comparable ranges. The magnitude of the features has an impact on many machine learning techniques. Larger scale features may dominate the learning process and have an excessive impact on the outcomes. We can ensure that each feature contributes equally to the learning process by scaling the features.
- **Algorithm performance improvement:** When the features are scaled, several machine learning methods, including gradient descent-based algorithms, perform better or converge more quickly.
- **Numerical Stability:** Numerical instability can be prevented by avoiding significant scale disparities between features. E.g. in matrix operations, having features with radically differing scales can result in numerical overflow or underflow problems.

Essentially, scaling features removes the bias of bigger features dominating the learning, producing skewed outcomes and that each feature contributes fairly to model predictions.

#### Difference between normalized scaling and standardized scaling

Feature	Normalized Scaling	Standardized Scaling
Definition	Scales the data to a fixed range, typically [0, 1] or [-1, 1].	Centers the data around the mean with a unit standard deviation.
Formula	$X' = (X - X_{\min}) / (X_{\max} - X_{\min})$ for [0, 1] range.	$X' = (X - \mu) / \sigma$ , where $\mu$ is the mean and $\sigma$ is the standard deviation.
Range	The range is user-defined, commonly [0, 1].	The transformed data will have a mean of 0 and a standard deviation of 1.
Sensitivity to Outliers	Highly sensitive, as it scales based on the min and max values.	Less sensitive compared to normalization, but still influenced by outliers.
Use Cases	Used when we need data to fit within a certain range, such as in neural networks or image processing.	Used when the data is assumed to be normally distributed, such as in algorithms like linear regression.
Example Scenarios	Min-Max Scaling for pixel values in image processing.	Z-score scaling for features in linear regression.

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

**Answer:**

VIF measures how much the variance of a regression coefficient is inflated due to collinearity with other predictors. It is calculated as:

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

where  $R_i^2$  is the R-squared value obtained by regressing the  $i^{\text{th}}$  predictor on all other predictors.

#### **Infinite VIF:**

An infinite VIF occurs when the  $R_i^2$  value is exactly 1. This indicates perfect multicollinearity, meaning that the  $i^{\text{th}}$  predictor is perfectly linearly dependent on one or more of the other predictors. In practical terms, it means we can express the  $i^{\text{th}}$  predictor as an exact linear combination of the other predictors.

#### **Reasons for Infinite VIF:**

1. **Perfect Collinearity:** If one of the independent variables is an exact linear combination of the others,  $R_i^2$  will be 1, resulting in a division by zero in the VIF formula.
2. **Duplicated Variables:** Including the same variable more than once or including variables that are perfectly correlated with each other (e.g., one variable is a scaled version of another) can cause perfect collinearity.
3. **Dummy Variable Trap:** In the context of categorical variables, if we include dummy variables for all categories, one of the dummy variables will be perfectly collinear with the others. For example, if we have a categorical variable with three levels and create dummy variables for all three, their sum will always be 1. Hence, we should use  $k-1$  dummy variables for  $k$  categories.

#### **Conclusion:**

An infinite VIF indicates a serious multicollinearity problem in our regression model. It is essential to diagnose and address this issue to ensure the stability and interpretability of regression coefficients.



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

**Answer:**

**Quantile-Quantile (Q-Q) plot**, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. It is an essential diagnostic tool in statistics and is commonly used in the context of linear regression to check the assumptions of the model.

**Definition:**

A Q-Q plot compares the quantiles of a sample distribution with the quantiles of a specified theoretical distribution. If the sample comes from the specified distribution, the points in the Q-Q plot will approximately lie on a straight line.

**Use and Importance in Linear Regression:**

In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. The Q-Q plot is used to visually check this assumption. Here's why it is important and how it is used:

**1. Checking Normality of Residuals:**

- **Assumption Verification:** Linear regression assumes that the residuals are normally distributed. Deviations from normality can affect hypothesis tests and confidence intervals.
- **Graphical Check:** A Q-Q plot provides a visual way to assess if the residuals deviate from normality.

**2. Identifying Deviations:**

- **Heavy Tails:** If the points in the Q-Q plot diverge significantly from the straight line at the ends, it indicates heavy tails.
- **Skewness:** If the points curve away from the line, it indicates skewness in the data.
- **Outliers:** Points that are far from the line can indicate outliers in the data.

**3. Model Validation:**

- **Improving the Model:** If the residuals are not normally distributed, it suggests that the model might not be the best fit. This can lead to model refinement, such as transformation of variables or using a different modelling approach.

**Conclusion:**

The Q-Q plot is a powerful tool in the diagnostic arsenal for linear regression. It helps ensure that the assumptions of normality for residuals are met, which is crucial for making valid inferences from the model. Using Q-Q plots allows for a visual assessment and aids in identifying any underlying issues with the model's assumptions.