**Data Analysis**

**Data** might **not** be **useful** on its own. **However**, when it is analysed, the **information gained** from data can **help** the user make **better decisions.** **Data analysis includes collecting, cleaning, transforming, and processing data.** **This analysis** is done to get information that can be useful in decision-making and can help individuals and organizations **make sense of data**.

In this lesson, you will learn about various data analysis methods and how to use them to draw insights from collected data. Before completing this lesson, **you should have the following knowledge:**

* Understand how basic data variable types, structures, and categories are used in data analysis.
* Know how to import, clean, organize, and aggregate data using various statistical tools.

**Types of data analysis introduction**

**Data analysts use** various methods and **tools** to **make sense of data**. **Different** kinds of **data** and information need **different** **methods** of analysis.

For example,

* one analyst may be trying to understand how an ancient group of people migrated across countries in the distant past for a historian.
* Another might be predicting modern travel patterns for an airline or airport. These two analyses would require different data collection methods and analysis tools.

This skill covers how to:

* Perform descriptive analysis
* Perform diagnostic analysis
* Perform predictive analysis
* Perform prescriptive analysis
* Perform hypothesis testing

# Descriptive analysis

**Descriptive analysis** is used to find out **what happened**. It **uses statistical tools** with data to **produce summaries and conclusions**. Descriptive analysis is an **important first step** in making sense of a data set.

For example, a list of blood pressure readings of patients in a trial study might not be easily understood. Descriptive analysis can be used to pull together the overall information about individual measurements. This overall information is more easily understood, and the next steps can be settled on easily.

The following are examples of business questions that descriptive analysis addresses:

* What were the sales during each quarter of the year 2022?
* How do sales during the first quarter of the year 2022 compare with sales during the first quarter of 2021?
* Which product had the most sales?

**Descriptive analysis** is important because it **helps** you:

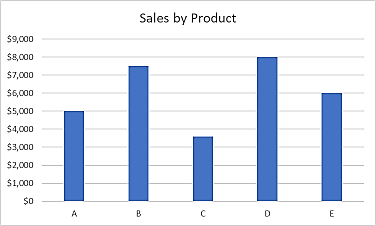
* Understand relationships between variables in a dataset
* Understand how the values of a variable are distributed
* Find errors in a dataset

Descriptive analysis can be categorized into the following types:

* Measures of frequency
* Measures of central tendency
* Measures of dispersion
* Measures of position

## Measures Of Frequency

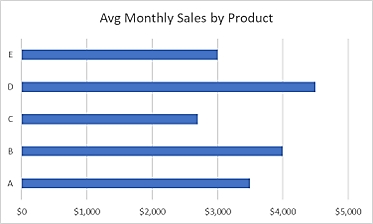
Measures of **frequency** provide information on the **number of times** a given **variable occurs**. They include **frequencies, ratios, and proportions**. In our “which product had the most sales” question, the analyst could provide a breakdown of sale amounts by product, as in **Figure 3-1**:



## Measures Of Central Tendency

Measures of **central tendency** show where the data is centred. Measures of central tendency indicate a kind of **middle of the distribution of data**. They **include** the **mean, median, and mode**. The relationships between different measures of central tendency can also give an analyst information about the spread of the data. An analyst looking at the measures of central tendency for product sales could tell what the average sales by month are for each item.

Graphical examples of measures of central tendency are shown in **Figure 3-2** with instructions provided for creation in later sections.



## Measures Of Dispersion

Measures of **dispersion** are used to describe the **spread of data**. It **helps** analysts **see** if there is a **wide variation** in data, **or** if it is **clustered** around one or more **specific values**. Measures of dispersion **include** the **range, variance, and standard deviation.** Someone reporting on product sales could give the range of sales, or the difference between the highest and the lowest sale numbers, as in **Figure 3-3**.



## Measures Of Position

Measures of **position** are used to determine **where** each **data point exists** in a given dataset. These measures **include percentiles and quartiles.** An analyst looking at product sales could arrange the items from least to most popular and then give a **percentile (ranking out of a hundred) or a quartile (ranking out of four)** to each product. Quartiles and percentiles can also be shown using the previous chart, **Figure 3-3**.

# Diagnostic analysis

Diagnostic analysis is used to **help** **explain** **why** data behaves the way it does. It helps explain the relationships between variables. This analysis is often **done after descriptive analysis** because it uses the results from descriptive analysis and looks for a cause. For example, a business owner can use diagnostic analysis to explain the reason for a sudden increase in sales. The following are examples of business questions addressed by this type of analysis:

* Why were sales in the first quarter of 2022 lower than sales in the first quarter of 2021?
* Why did marketing plan A outperform the other plans?
* Why did 10% of the customers leave?

Diagnostic analysis is important because it helps you:

* Determine the root cause of an event or trend
* Analyze factors that affect the performance of a business
* Better understand business data, allowing fast answers to crucial questions

## Steps To Follow When Conducting Diagnostic Analysis

Step 1 Use descriptive analysis to **identify events** that **require** further **investigation**. *In the question of why the first quarter of 2021 had more sales than the first quarter of 2022, an analyst would first identify means, medians, and modes for each quarter’s sales.* They would then identify the measures of spread to look for overlap in sales amount by quarter.

Step 2 **Identify** **data** that can be **useful** in **investigating** the events identified **in Step 1**. *After the analyst determines that Q1 in 2021 had greater sales than Q1 in 2022, they will look for internal and external reasons in related data.* Was there a change in product or in overall shopping patterns? What external factors can be included or eliminated (weather, economic changes, and social media trends to name a few)?

Step 3 Use the **data identified in Step 2** to **discover** **hidden relationships** that may have led to the events identified in Step 1. An analyst can perform a number of statistical tests to find relationships between data sets. Correlations, or specific relationship patterns, can be tested. For example, the analyst might find that average daily temperature is correlated, or related in a predictable way, to sales.

# Predictive analysis

Predictive analysis **uses** **current and historical data** to determine **what might happen in the future**. It can be used to answer the following business questions:

* What are our projected profits for the year?
* What is our projected employee turnover rate this year?
* How many new competitors are expected to enter our target market in the next 16 months?

Predictive analysis is important for many reasons, including:

* Helping businesses balance staffing needs across a region
* Predicting customer buying behavior
* Detecting and preventing fraud
* Targeting marketing campaigns to high-interest possible buyers

# Prescriptive analysis

Prescriptive analysis uses data to **recommend the best course of action**. It is used to help decide what should be done. Prescriptive analysis **tools** **help analyze data, determine possible action points, and make recommendations** on the next steps to follow. Predictive analysis can be used to answer the following business questions:

* Will sales increase if more advertising is conducted?
* Will our existing customer base be affected by an increase in product prices?
* Can influencer marketing help us to increase our customer base?

# Hypothesis testing

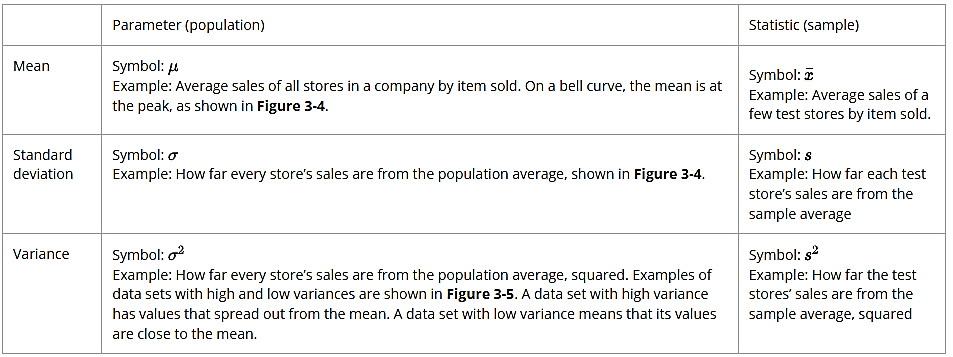
Hypothesis testing is a data analysis tool that uses data from a sample, then applies the test results to the whole group, or population.

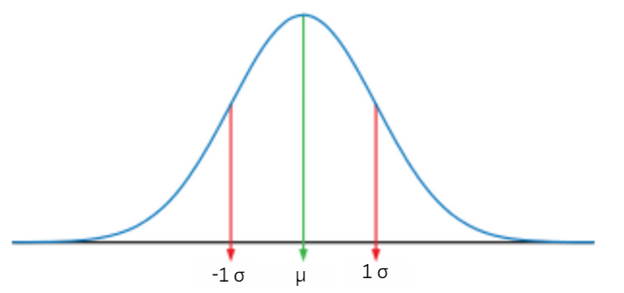
A **population** is a group where every member has something in common. Examples of populations include all registered voters in the US or all startups in Canada that failed before three years of operation.

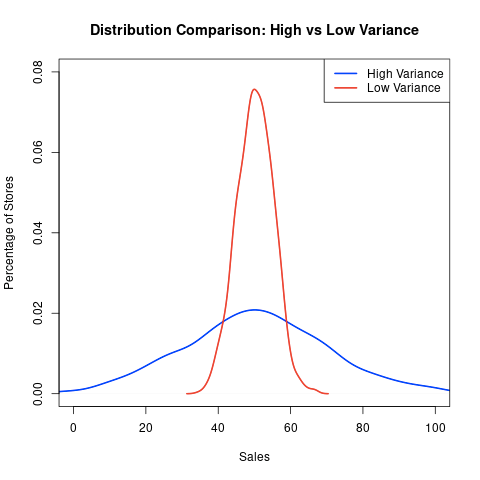
Often, analysts want to know things about a population but cannot collect data on all members of the population. Instead, they choose a part of the population, called a **sample**, and use it to draw conclusions about the population. A sample must be chosen carefully to make sure it doesn’t misrepresent the whole population. One key feature of a sample is that it must be random. In a random sample, every selection has an equal chance of getting selected.

A **parameter** is a characteristic of the population, while a **statistic** is a characteristic of the sample. **Table 3-1** illustrates some examples of parameters and statistics.

**Table 3-1**

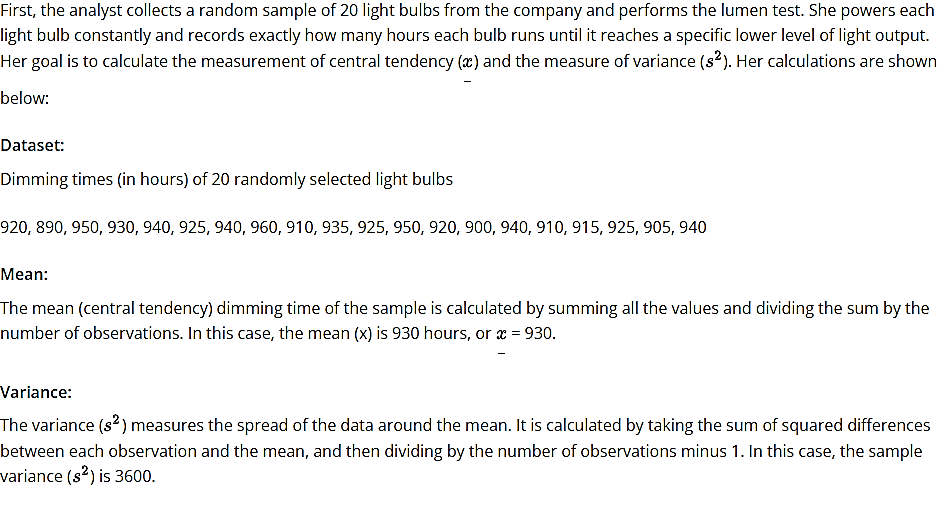


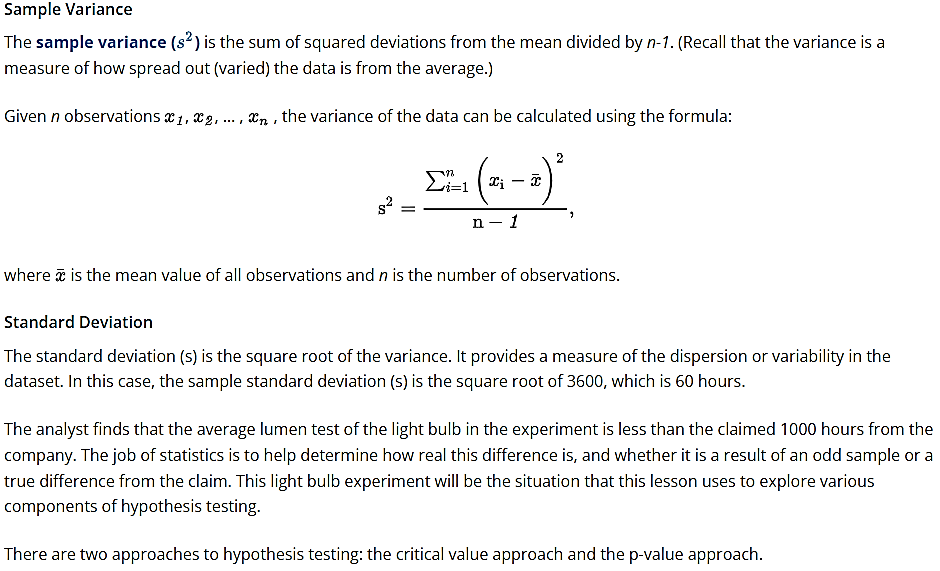


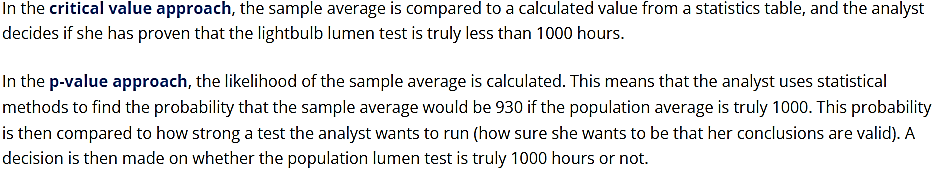


A **hypothesis** is a yes/no statement about a characteristic of a population. In the following example, the analyst is interested in a population average and writes a hypothesis about the average.

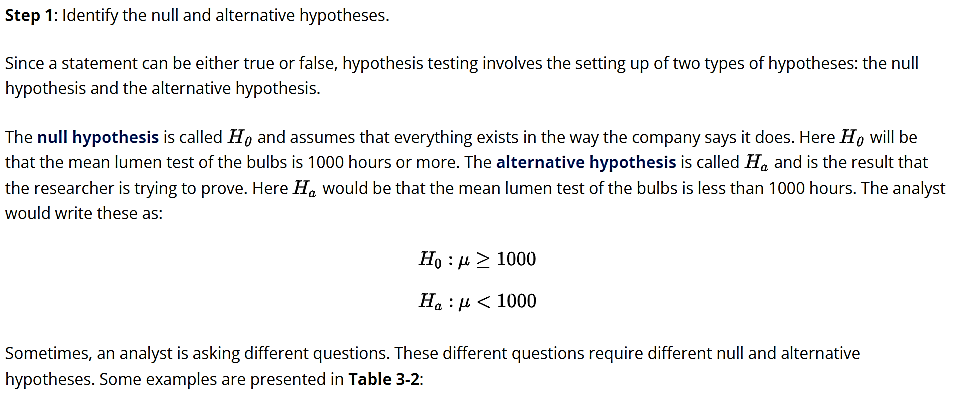
Light bulbs slowly get dimmer as they are used. A company that makes light bulbs claims that the mean (average) time its bulbs take to reach a specific level of dimness (called a lumen test) is 1000 hours. A consumer advocate would like to determine if the mean lumen test of the bulbs is actually less than 1000 hours.

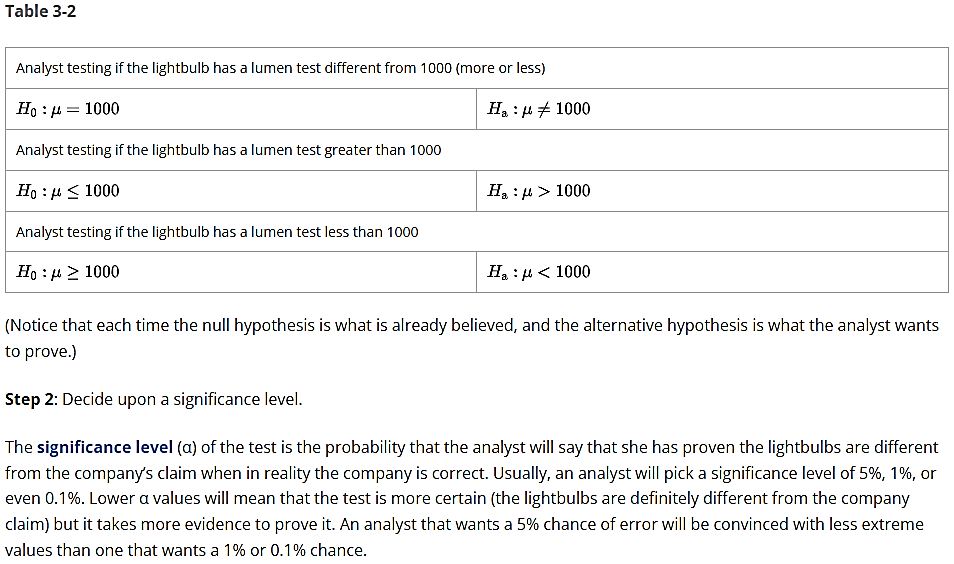


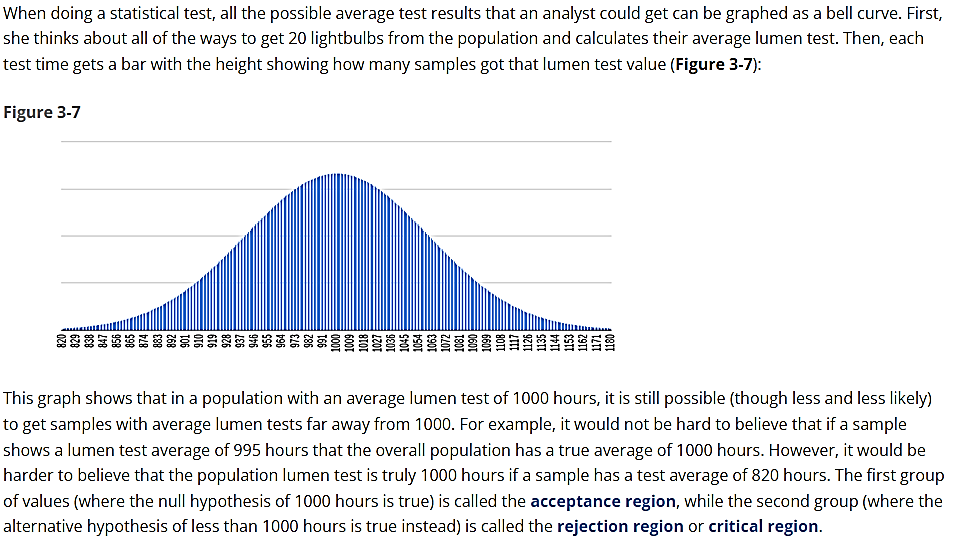
****

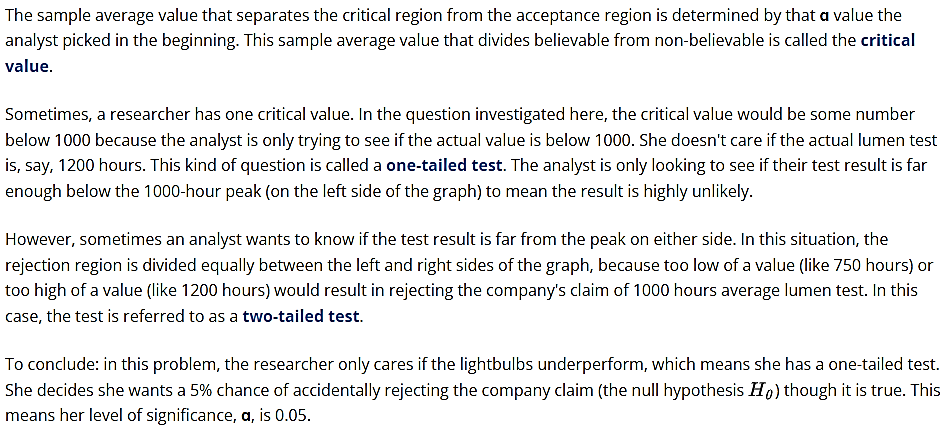
****

**Hypothesis testing using the critical value approach**

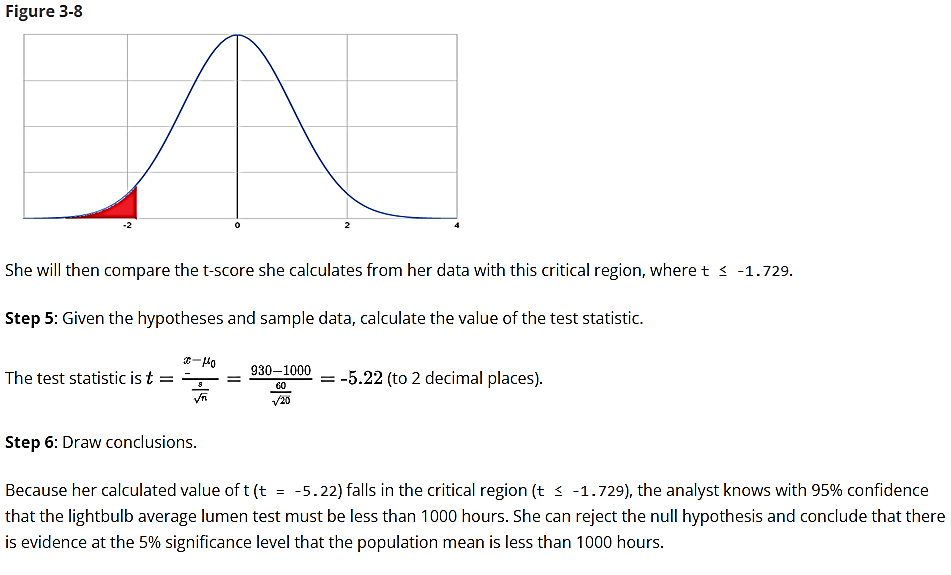












**Hypothesis testing using the p-value approach**

In the p-value approach, the analyst is still concerned with where her data falls on the bell curve, and if it is extreme enough to prove (with her desired level of accuracy) that the population mean has to be less than 1000. However, in this approach when she reaches **Step 4**, she won’t find the critical region. Instead, she is going to calculate the probability that her value is possible given a true population mean of 1000, and compare that to her desired confidence level of α = 0.05, or 5%.

If the probability she calculates, called the p-value of her test, is less than or equal to her chosen α-level, the null hypothesis is rejected. If it is greater than the chosen α-level, the null hypothesis is not rejected. For example, a sample with a calculated p-value of 0.15 or 15% would not allow the analyst to reject the null hypothesis and conclude the true population mean is less than 1000 hours. A sample with a calculated p-value of 0.02 or 2%, on the other hand, would allow her to reject the null hypothesis and conclude the true population mean is lower than 1000.

To calculate the p-value for the test in this problem, the analyst can run a bit of code in R. She calculates the test statistic as before to be -5.22. The degrees of freedom in a statistics problem is the size of the sample minus 1, and this one-tailed test is looking at the left, or lower, tail. In this code, she uses the pt function with the syntax:

**pt(test statistic, degrees of freedom, direction of the test)**

**pt(-5.22, 19, lower.tail = TRUE)**

The output is:

**[1] 2.437634e-05**

So the p-value of the analyst’s data is 0.0000244, or 0.0024%.

Because the p-value is less than 0.05, she can reject the null hypothesis and conclude that there is evidence at the 5% significance level that the population mean is less than 1000 hours.

Video: Data Description

# Data aggregation and interpretation metrics

**Data aggregation** is how an analyst **collects** **data** from **multiple sources** and **stores** it in an **understandable form**. For example, many rows of data in a spreadsheet can be summarized by the mean and variance of each row. Descriptive statistics are very helpful when performing data aggregation.

Data aggregation is another important part of the data exploration process. It helps analysts find trends in the data, make comparisons, and discover information that might have gotten lost in all the individual data points.

**Data interpretation** is the process of giving meaning to processed and analyzed data. It involves the following steps:

* Designing strong research questions
* Collecting data relevant to the questions you want to answer
* Analyzing collected data
* Summarizing the key findings of an analysis to answer research questions
* Reporting findings and conclusions

Data processing techniques, such as data filtering and searching, are essential in the process of data interpretation.

Data filtering involves splitting up the sample into groups to create new subsets to be analyzed.

Data searching helps to find specific records, e.g., unique values or specific strings, in a dataset.

Data is often aggregated using descriptive statistics, e.g., measures of frequency, central tendency, dispersion, and position.

In the following topics, we’ll examine some of the common statistical measures used to aggregate data.

## Count

The count or frequency of an item is the number of times the item occurs in a dataset.

**Example**

Find the number of dealerships that contain models of cars in each color in the following data:

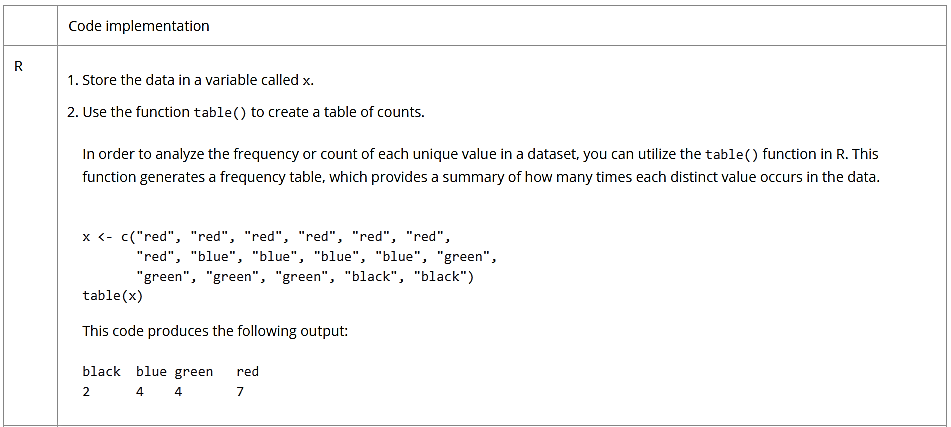
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dealership** | **Model** | **Color** | **Number in Stock** | **Miles Per Gallon (MPG)** |
| Velocity Motors | Corvette | Red | 2 | 19 |
| Elite Auto Group | Corvette | Red | 2 | 19 |
| Summit Motors | Model X | Red | 3 | 102 |
| Velocity Motors | GT-R | Blue | 1 | 16 |
| Precision Automotive | Civic | Blue | 3 | 31 |
| Elite Auto Group | Jetta | Green | 2 | 29 |
| Precision Automotive | Mustang | Green | 2 | 21 |
| Velocity Motors | Accord | Black | 2 | 30 |

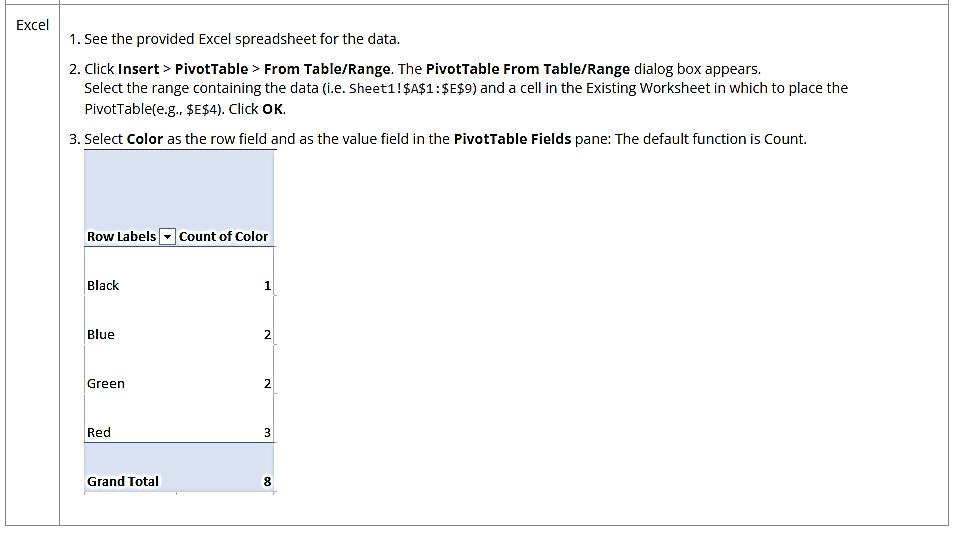
The data can be organized as in **Table 3-3**:

|  |  |
| --- | --- |
| Color | Count |
| black | 1 |
| blue | 2 |
| green | 2 |
| red | 3 |

The table gives the count or frequency of each car color in the data.

**Implementation In R And Excel**





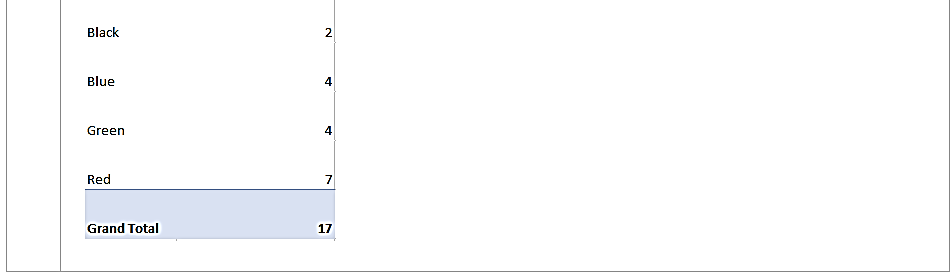
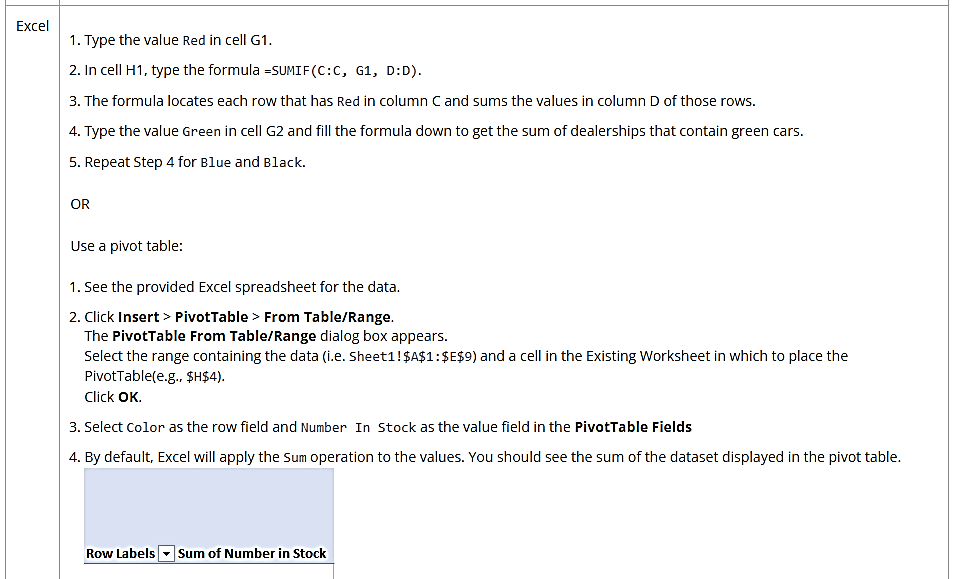
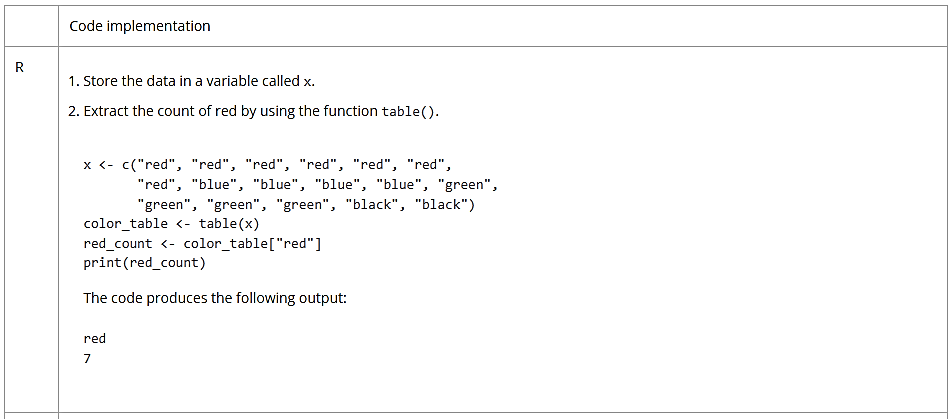
## Sum

The sum of the number of red cars in stock at each location in a dataset is the result of adding the values in the Number In Stock feature for all cars of a specific color.

**Example**

Find the sum of all the red cars in stock at all locations in the dataset.

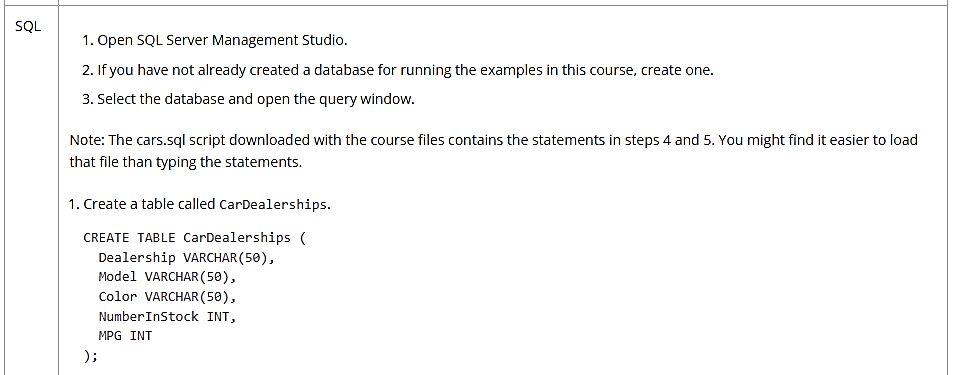
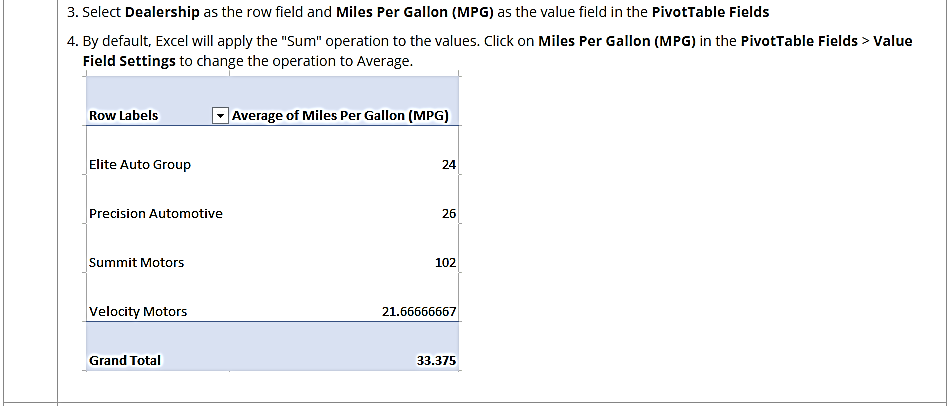
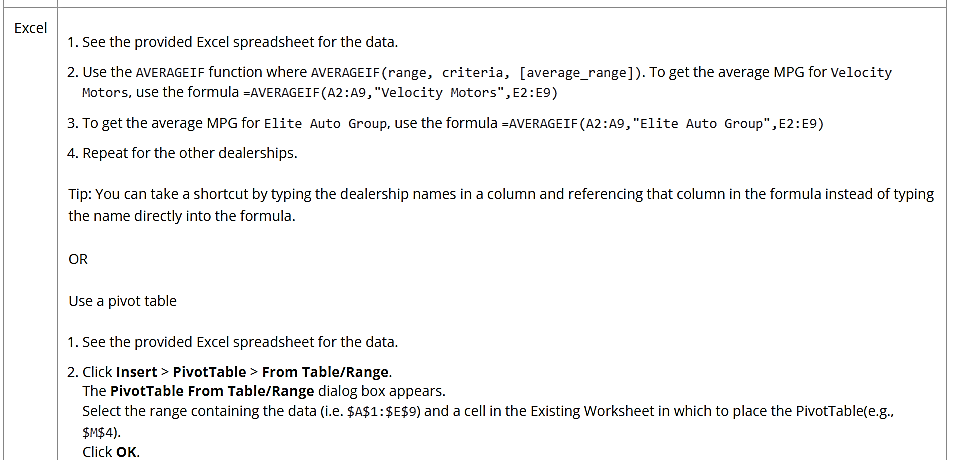
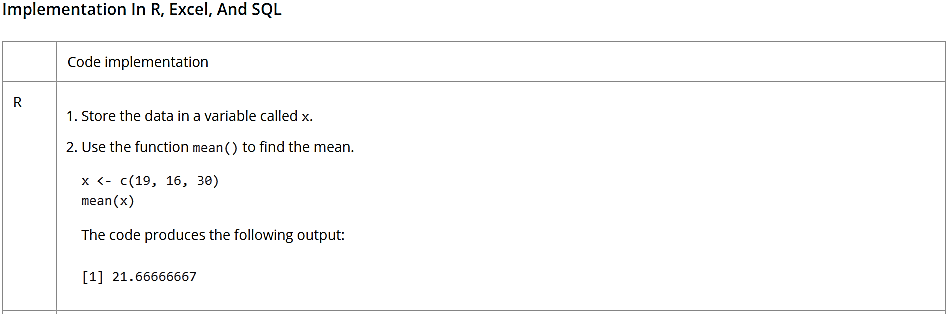
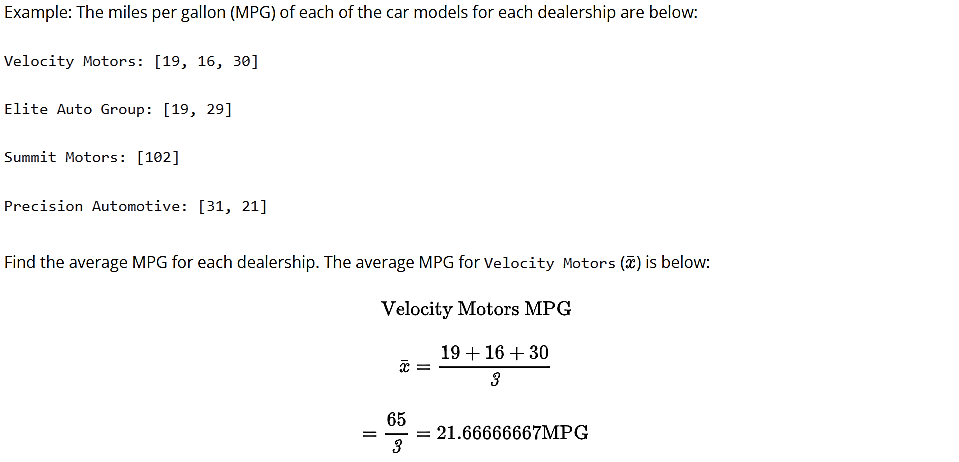
**Implementation In R And Excel**

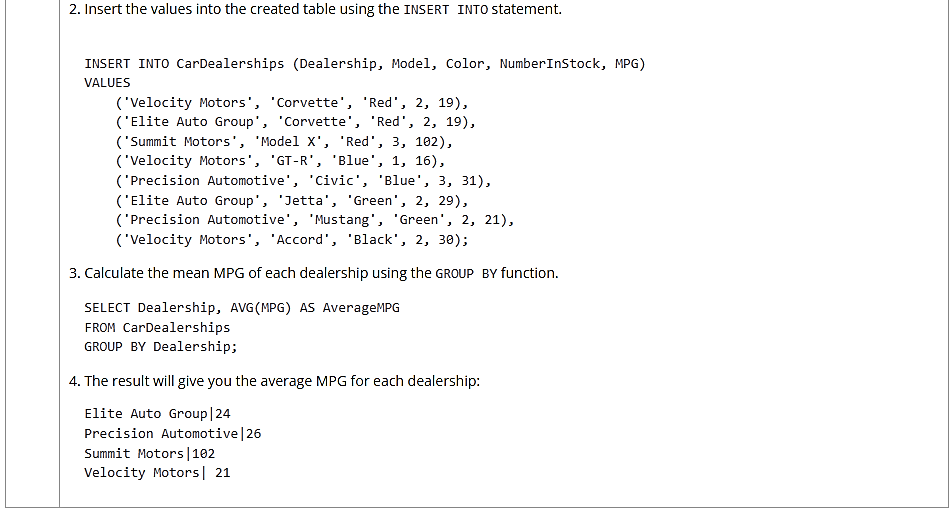
****

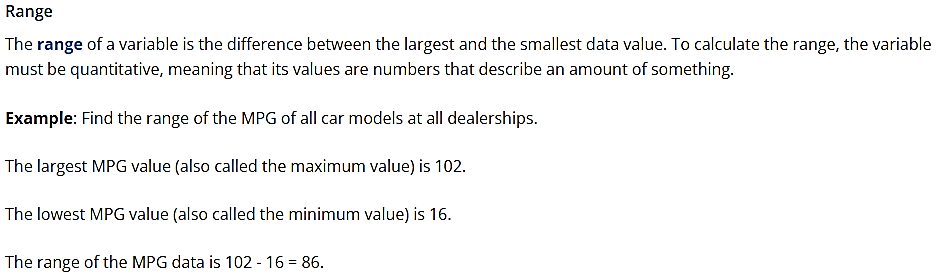
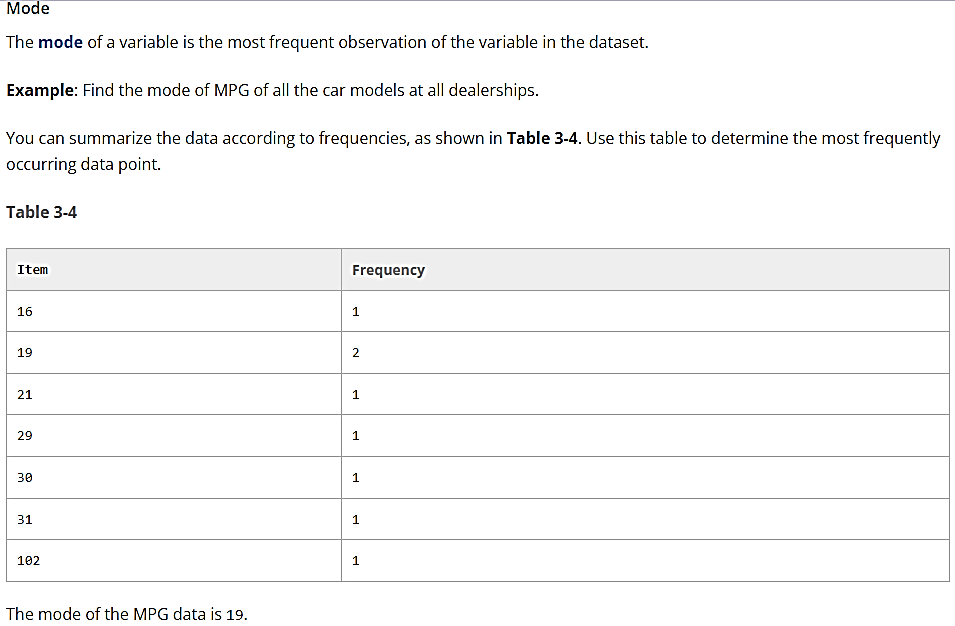
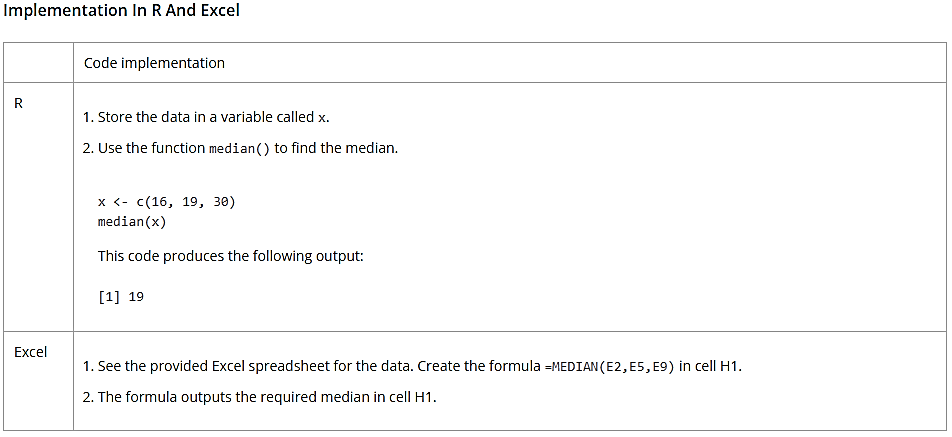
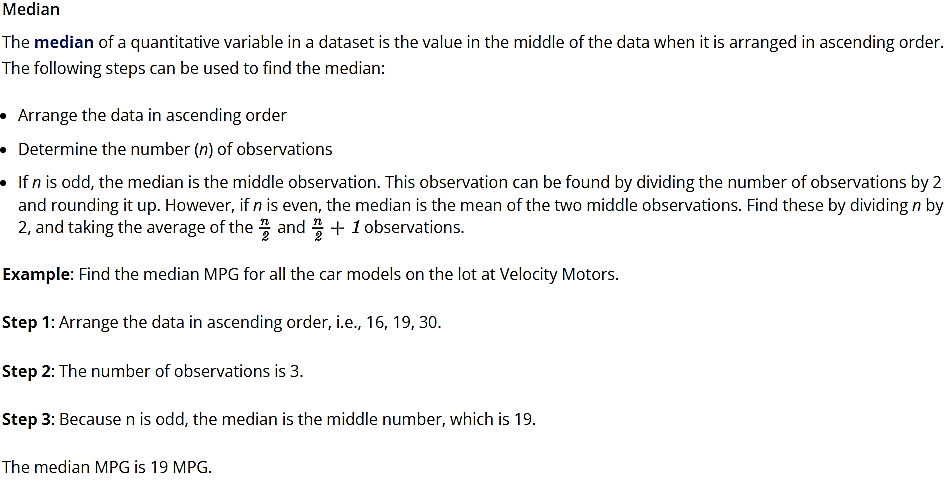
**Mean**

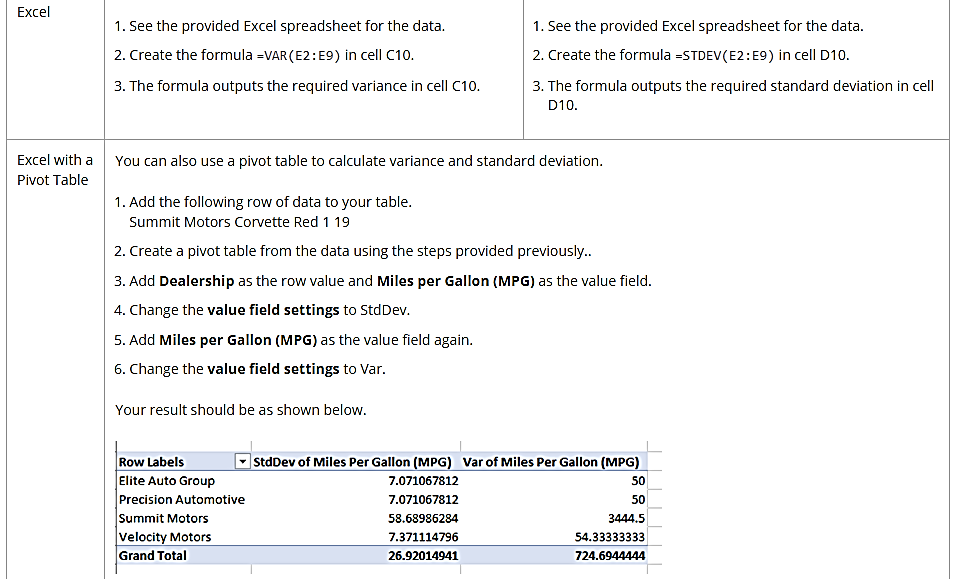
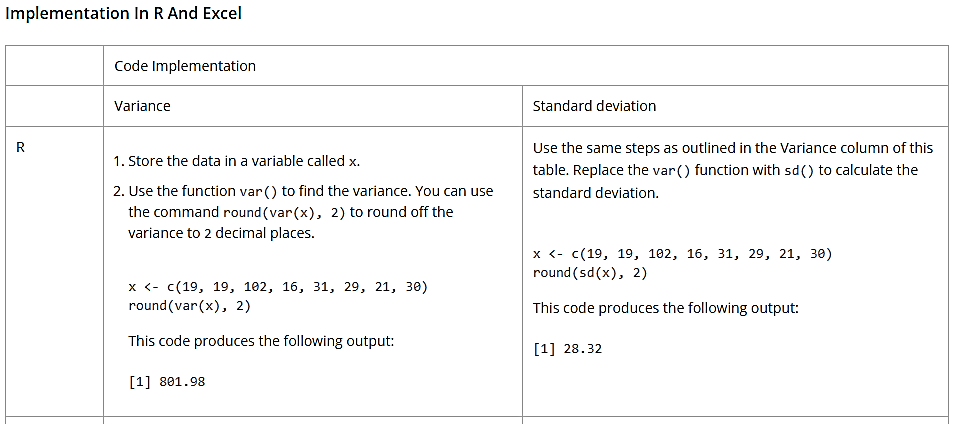
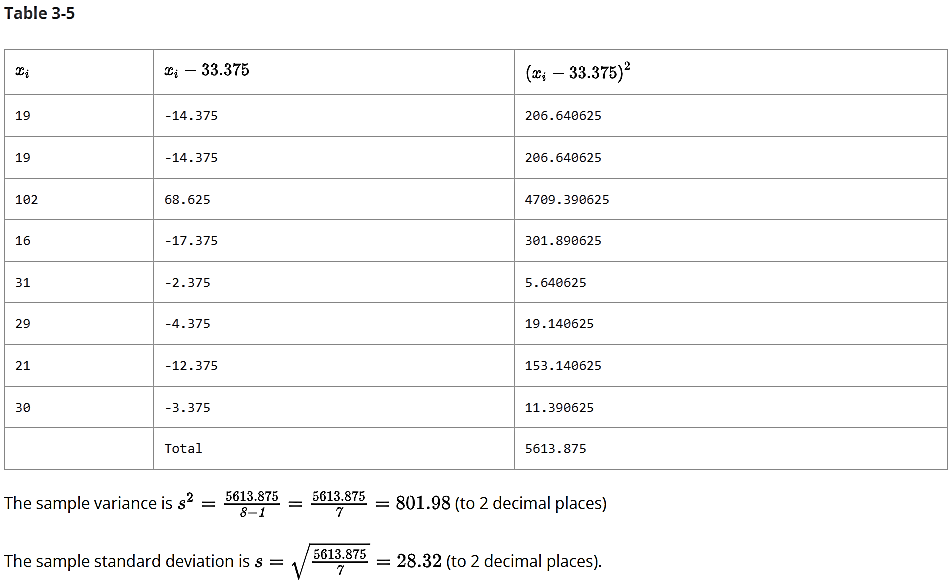
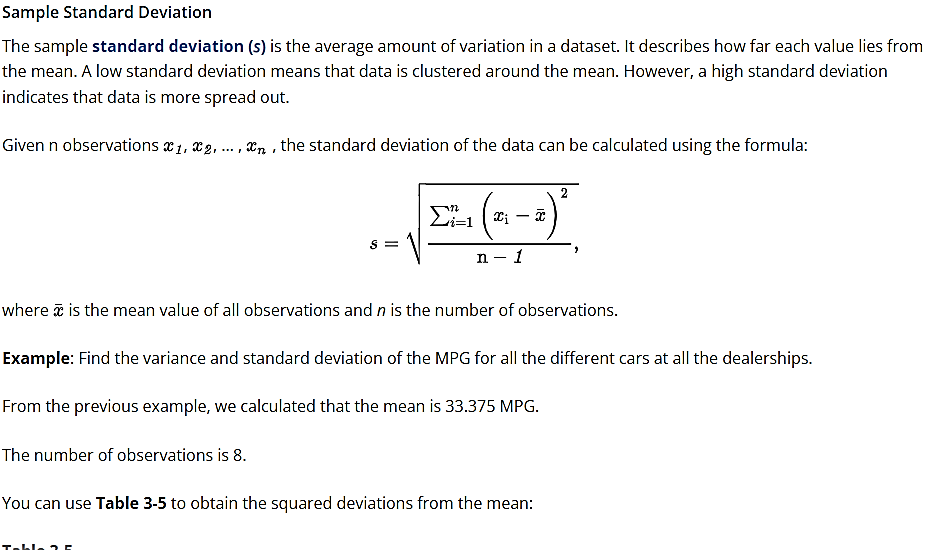
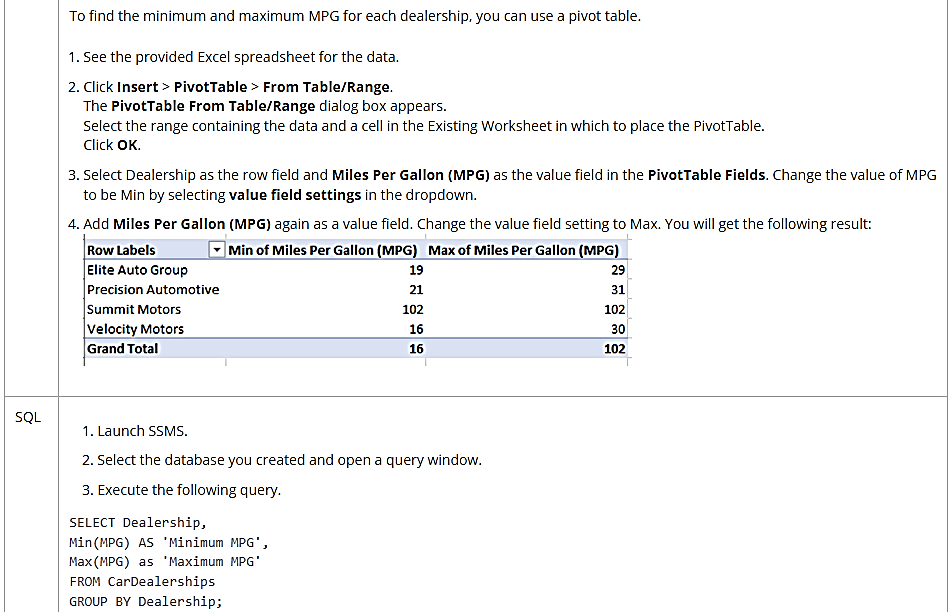
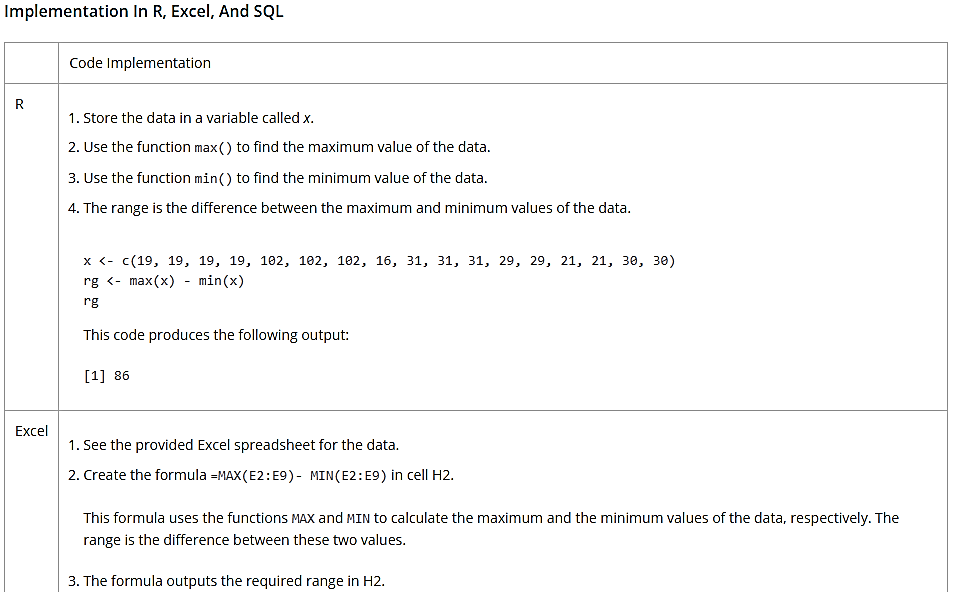
The **mean** of a variable in a dataset is calculated by adding all of the values of the variable in the dataset and dividing their sum by the number of observations.

**Video: How to calculate the mean**

****

****





**Data aggregation using a real dataset in R**

In this section, you will use the *chickwts* dataset to aggregate data in R.

The *chickwts* dataset is a built-in dataset in R. The data comes from an experiment in which newly hatched chickens were randomly divided into six groups with each of the groups receiving different feed supplements. The weights of the chickens were measured in grams after six weeks.

The dataset contains 71 observations on two variables, namely, weight and feed. In this dataset, weight denotes the weight of chickens in grams, while feed denotes the feed supplement type.

You can explore the dataset using the dim(), head(), and str() functions, which provide information on the dimensions, the first few rows, and the internal structure of the data frame, respectively.

**df = chickwts**

**dim(df)**

The code produces the following output:

[1] 71 2

**head(df)**

This code produces the following output:

weight feed

1 179 horsebean

2 160 horsebean

3 136 horsebean

4 227 horsebean

5 217 horsebean

6 168 horsebean

**str(df)**

This code produces the following output:

'data.frame': 71 obs. of 2 variables:

$ weight: num 179 160 136 227 217 168 108 124 143 140 ...

$ feed: Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...

You can save the *chickwts* dataset as *chickwts*.csv on your desktop to use later in the Excel and SQL environments using the following steps on Windows:

* Create a folder called DATA on your desktop.
* Determine the file path of the folder DAT.
* Open RStudio or any other R environment that supports access to the local file system on your computer.
* Copy the code snippet below into the editor.

**df = chickwts**

**write.csv(df, file = 'C:/Users/INSERT-YOUR-USER-NAME-HERE/Desktop/DATA/chickwts.csv', row.names = FALSE)**

The path of the file is written in R as follows: C:/Users/INSERT-YOUR-USER-NAME-HERE/Desktop/DATA/chickwts.csv (Replace “INSERT-YOUR-USER-NAME-HERE” with your actual username. It is important to remember to replace backslashes in file paths with forward slashes in R to avoid errors.)

* Select the entire code in the editor.
* Click the "Run" button, or use the keyboard shortcut (Ctrl + Enter on Windows) to execute the code.

**Count, Sum, And Mean Of Aggregated Data**

You may want to answer the following questions using the *chickwts* dataset:

1. How many chickens are in each group of feed supplements?
2. What is the total weight of chickens for each group of feed supplements?
3. What is the mean weight of chickens for each group of feed supplements?

To answer these questions in R, you can use the aggregate() function. The syntax of the aggregate() function is as follows:

**aggregate(quantitative\_variable, list("Group title" = categorical\_variable), function)**

In this case, quantitative\_variable is weight, categorical\_variable is feed, and function is the function you want to apply to the values in the grouped data (e.g., sum, mean, min, and max).

The following is an alternative syntax of the aggregate() function:

**aggregate(numerical\_variable~categorical\_variable, dataframe, function)**

In this case, dataframe is df (i.e., the name of the dataset you are using) and function is the function you want to apply to the values in the grouped data (e.g., sum, mean, or min).

|  |
| --- |
| Code implementation |
| Solution to Q. 1 (number of chickens fed each feed type) | Use the aggregate() function to group the data by feed. In this case, the function argument takes the value length as follows:  **df = chickwts**  **aggregate(df$weight, list("feed type"=df$feed),length)**  This code produces the following output:  feed type x  1 casein 12  2 horsebean 10  3 linseed 12  4 meatmeal 11  5 soybean 14  6 sunflower 12  Alternatively, you can use the table() function to obtain the counts for each group of feed supplements.  **table(df$feed)**  This code produces the following output:  casein horsebean linseed meatmeal soybean sunflower  12 10 12 11 14 12 |
| Solution to Q. 2 (total weight for each group of feed) | Use the aggregate() function to group the dataset by *feed* and find the sum of weights of the chickens for each group of feed supplements:  **df = chickwts**  **aggregate(df$weight, list("feed type"=df$feed),sum)**  This code produces the following output:  feed type x  1 casein 3883  2 horsebean 1602  3 linseed 2625  4 meatmeal 3046  5 soybean 3450  6 sunflower 3947 |
| Solution to Q. 3 (mean weight by feed type) | Use the aggregate() function in either of the following ways to group the dataset by feed and find the mean weight of each group of chicks based on the feed type they received:  **df = chickwts**  **aggregate(df$weight, list("feed type"=df$feed),mean)**  This code produces the following output:  feed type x  1 casein 323.5833  2 horsebean 160.2000  3 linseed 218.7500  4 meatmeal 276.9091  5 soybean 246.4286  6 sunflower 328.9167  OR  **df = chickwts**  **aggregate(weight~feed, df,mean)**  The code produces the following output:  feed weight  1 casein 323.5833  2 horsebean 160.2000  3 linseed 218.7500  4 meatmeal 276.9091  5 soybean 246.4286  6 sunflower 328.9167 |

**Data aggregation using a real dataset in SQL**

You will use the *chickwts* dataset to aggregate data in this section. Recall that you have saved this dataset in the folder named DATA as the file *chickwts.csv*.

Aggregate functions, commonly performed with the GROUP BY command, output a single value computed from a set of data. Examples of commonly used aggregate functions in SQL are COUNT(), SUM(), AVG(), MIN(), and MAX(). All aggregate functions in SQL ignore null values except for COUNT().

The GROUP BY clause groups rows with similar values into summary rows.

You will use *chickwts.csv* to answer the following questions:

1. How many chickens are in each group of feed supplements?
2. What is the mean weight of chickens for each group of feed supplements?
3. What is the variance for each group of feed supplements?

First, import the *chickwts.csv* dataset to your SQL database:

1. Open MicrosoftSQL Server Management Studio and connect to your SQL Server instance.
2. In the Object Explorer, expand the database where you want to import the dataset.
3. Right-click on the database, choose Tasks > Import Flat File and follow the instructions on screen to import the file.

**Implementation In SQL**

|  |  |
| --- | --- |
| Code solution to Q. 1  (Number of chickens fed each feed type) | Use the COUNT() function and the GROUP BY clause to calculate the number of chickens by groups of feed as follows:    SELECT COUNT(weight) As chickens, feed  FROM chickwts  GROUP BY feed; |
| Code solution to Q. 2  (Mean weight by feed type) | Use the AVG() function and the GROUP BY clause to calculate the mean weight of the chickens in each group of feed as follows:    SELECT AVG(weight) As avg\_weight, feed  FROM chickwts  GROUP BY feed; |
| Code solution to Q. 3 (variance by feed type) | Use the VAR() function and the GROUP BY clause to calculate the variance by groups of feed as follows:    SELECT VAR(weight) As variance, feed  FROM chickwts  GROUP BY feed; |

**Data aggregation using a real dataset in Excel**

In this section, you will use Power Query in Excel to aggregate data. You will also use *chickwts.csv* in the DATA folder.

First, you should import *chickwts.csv* into Power Query using the following steps.

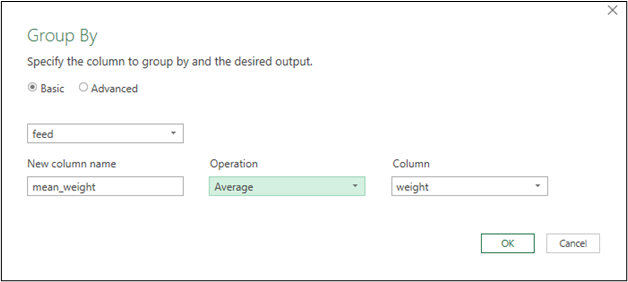
* Open a new Excel workbook.
  + In Excel 2016: Click **Data** > **New Query** > **From File** > **From CSV**
  + In Excel Office 365: Click **Data** > **From Text/CSV**
* The Import Data window opens.
* Navigate to the *chickwts.csv*file. Select it and click **Transform Data** in the window that pops up.

You would like to use *chickwts.csv* to answer the following question:

What is the mean weight of chickens for each group of feed supplements?

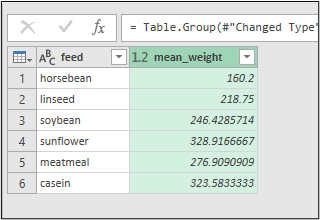
To calculate mean weight by groups of feed, click **Group By** in the **Home** tab.

In the **Group By** dialog box, choose the variables to group your data by, as shown in **Figure 3-9**.

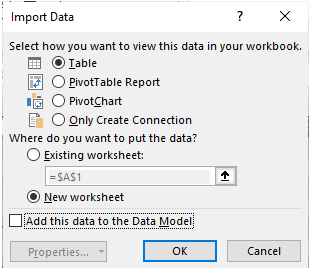
**Figure 3-9**

Group your data by feed, name the new column that will contain the aggregates mean\_weight; and apply the operation **Average** on the column weight.  
Click **OK**.

The actions above produce the aggregate mean table shown in **Figure 3-10**, and answers the questions posed in this section.

**Figure 3-10**

To move this result to the worksheet, click **Close & Load To …** in the **Home** tab of the query. Use the **Load To** dialog box that appears to choose how you would like to import the data and click **Load**, as shown in **Figure 3-11**.

**Figure 3-11**

# Data interpretation

While working with the *chickwts* dataset, you discovered that the mean weight of chickens fed casein was 323.5833. The mean weight of those fed soybeans was 246.4286. What do these numbers mean?

**Data interpretation** is how an analyst finds meaning in data and requires the analyst to judge the results of the analysis. The analyst relates the processed data to the research questions, explores relationships between the measurements, and draws inferences. They ask the question: What is the meaning of the pattern that the data displays?

The *chickwts* dataset statistics can help answer the following research questions:

* Does the mean weight of chickens fed casein at six weeks differ significantly from that of chickens fed soybean?
* Is there a significant relationship between chicken weight and feed type?

To answer these questions, the *chickwts* dataset must be analyzed using more advanced methods.

**Exploratory data analysis methods introduction**

**Exploratory data analysis** (EDA) is a critical first step in analyzing data for interpretation. It summarizes datasets by their main characteristics. Often data charting methods and summary statistics are used. EDA is useful in finding errors in the data, detecting outliers (weird data points), and analyzing relationships between variables.

This skill covers how to:

* Find relationships in a dataset
* Identify outliers in a dataset
* Drill a dataset
* Mine a dataset

**Finding relationships in data**

A critical part of EDA is finding relationships between variables. The analyst must understand what situation the data describes, the types of variables in the data, and if some variables influence the values of other variables.

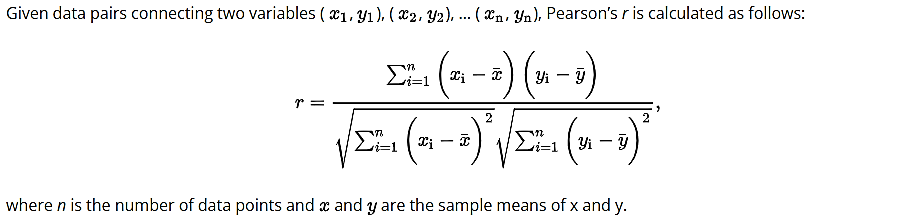
Lots of methods can be used to find relationships between variables. The method chosen will depend on what kind of data an analyst has. Is it numerical, where each value is represented by a number? Is it categorical, where each value is a category record (for example favorite color or hair color)? Is it mixed data, where some entries are numbers and some categories? Different kinds of datasets require different analysis methods.

**Correlation**

**Correlation** is a statistical measure that explains certain kinds of relationships between two variables. Specifically, correlation is a straight-line relationship; if one variable goes up, does another consistently go up or down? Correlation analysis is important because it helps an analyst pick factors for more investigation, and to include in mathematical models. It can be measured using the **correlation coefficient** (denoted by *r*). The value of r ranges from -1 to +1.

There are several types of correlation coefficients, depending on the data type. The most commonly used correlation coefficient is called **Pearson’s correlation** (also called Pearson’s *r*).

Pearson’s r measures the strength and direction of the linear relationship between two variables.



**Correlation Values**

* Positive correlations (an r-value between 0 and 1) show that the two variables change in the same direction. When one increases, so does the other. When one decreases, so does the second.
* Negative correlations (an r-value between -1 and 0) show two variables move in opposite directions. When one variable increases, the other variable decreases. When one decreases, the second increases.
* A zero correlation exists when there is no linear relationship between two variables.

Generally, you can use **Table 3-6** to help you interpret the correlation value:

**Table 3-6**

| **Absolute value of *r*** | **Strength of relationship** |
| --- | --- |
| Between 0 and 0.3 | Weak relationship |
| Between 0.3 and 0.7 | Moderate relationship |
| Between 0.7 and 1 | Strong relationship |

It is important to know that a strong linear correlation between two variables does not mean that one of the variables is causing a change in the other (causation). In other words, **correlation does not imply causation**. There is a positive correlation between the number of firefighters at a fire scene and the amount of damage caused by the fire. However, this does not mean that having more firefighters at the scene causes more damage. Rather, it is more likely that fires that are more severe and likely to cause more damage require more firefighters to respond to the scene. (Causation can only be determined from a properly designed, randomized, and controlled experiment and further analysis.)

When performing correlation analysis, analysts investigate how the data looks in a graph, in addition to looking at correlation values. Charting data points helps you better visualize and interpret correlation values. For example, it is easier to identify non-linear or curved relationships between variables by looking at a graph like a scatter plot.

Correlations are best visualized using **scatter plots**. A scatter plot shows the relationship between two numerical variables by plotting a dot for every observation. It allows you to identify overall patterns, directions, and strength of association.

**Video: Calculating correlations in R**

In this section, you will use the *marketing* dataset from the *datarium* package in R.

The *marketing* dataset contains data on the amount of money (in thousands of dollars) that a company is willing to set aside for advertising on three different media platforms (YouTube, Facebook, and newspaper) and the correlating effect on sales. It has 200 rows and 4 columns.

To obtain the *marketing* dataset, first install the datarium package using the following code:

**Note**: In some environments or installations of R, the datarium package may already be included by default. It is recommended that you check the installed packages in your specific R environment before installing it. You can do this by using the installed.packages() function or checking the package list in your IDE. If the datarium package is already present, there is no need to install it again.

install.packages("datarium")

Load the *marketing* dataset into the variable md using the following code:

require(datarium)

md <- marketing

To better understand the dataset, use the function dim() to determine the dimensions of the dataset, str() to obtain information about the rows and columns of the dataset, and head() to view the first few rows of the dataset. The code to view the dataset dimensions is as follows:

dim(md)

This code produces the following output:

[1] 200 4   
  
  
To display information about the rows and columns in the dataset, run the following code:

str(md)

This code produces the following output:

'data.frame': 200 obs. of 4 variables:

$ YouTube : num 276.1 53.4 20.6 181.8 217 ...

$ Facebook : num 45.4 47.2 55.1 49.6 13 ...

$ newspaper: num 83 54.1 83.2 70.2 70.1 ...

$ sales : num 26.5 12.5 11.2 22.2 15.5 ...

Run the following code to view the first six rows of the dataset:

head(md)

This code produces the following output:

YouTube Facebook newspaper sales

1 276.12 45.36 83.04 26.52

2 53.40 47.16 54.12 12.48

3 20.64 55.08 83.16 11.16

4 181.80 49.56 70.20 22.20

5 216.96 12.96 70.08 15.48

6 10.44 58.68 90.00 8.64

Missing values can be found in R using the function is.na(). Missing values in an R dataset are fields that contain NA. The is.na function checks each data point in a dataset and returns TRUE if it contains NA and FALSE if it does not contain NA. Tabulate these values using the table() function as follows:

table(is.na(md))

This code produces the following output:

FALSE

800

The code output above shows that none of the 800 data points in the dataset are missing values.

From the results received so far, the following statements can be made about the dataset.

* The dataset has 200 observations and 4 variables, namely, YouTube, Facebook, newspaper, and sales.
* All the values are numerical.
* The dataset does not have any missing values.

Descriptive statistics (e.g., averages) help an analyst better understand the data. The summary() function in R is used to compute summary statistics of data and models.

summary(md)

This code produces the following output:

YouTube Facebook newspaper sales

Min. : 0.84 Min. : 0.00 Min. : 0.36 Min. : 1.92

1st Qu.: 89.25 1st Qu.:11.97 1st Qu.: 15.30 1st Qu.:12.45

Median :179.70 Median :27.48 Median : 30.90 Median :15.48

Mean :176.45 Mean :27.92 Mean : 36.66 Mean :16.83

3rd Qu.:262.59 3rd Qu.:43.83 3rd Qu.: 54.12 3rd Qu.:20.88

Max. :355.68 Max. :59.52 Max. :136.80 Max. :32.40

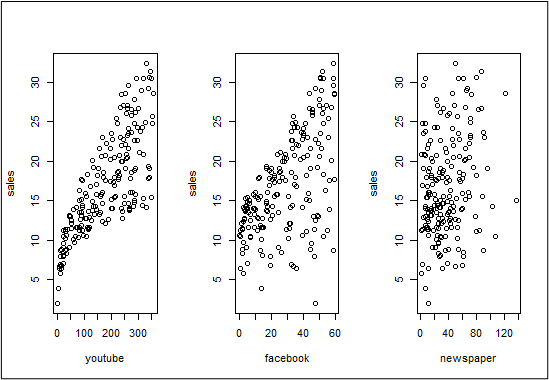
From the code output above, note that the company spends the most on YouTube advertising (mean budget is $176,450) and the least on Facebook advertising (mean budget is $27,920).

Scatter plots also help you visualize the relationship between sales and each explanatory variable (i.e., YouTube, Facebook, or newspaper). The function par(mfrow=c(nrows, ncols)) allows you to combine many plots in a single graph in R, i.e., a matrix of nrows by ncols plots. The mfrow argument specifies the dimensions of the grid, indicating the number of rows (nrows) and columns (ncols) of plots you want to arrange.

par(mfrow=c(1,3))

plot(md$youtube,md$sales,xlab = "youtube",ylab="sales")plot(md$facebook,md$sales,xlab = "facebook",ylab="sales")

plot(md$newspaper,md$sales,xlab = "newspaper",ylab="sales")

**Figure 3-12**

From **Figure 3-12**, the relationships between Facebook and sales, and YouTube and sales appear both positive and linear. The data is clumped in a line shape, rising from left to right. However, the relationship between YouTube and sales appears to be stronger than that between Facebook and sales because the line is clearer and the data is more tightly clustered. The relationship between newspaper and sales does not appear to be linear.

Compute the numerical value of the correlation between sales and each advertising medium can be done using the cor() function as follows:

cor(md)

This code produces the following output:

YouTube Facebook newspaper sales

YouTube 1.00000000 0.05480866 0.05664787 0.7822244

Facebook 0.05480866 1.00000000 0.35410375 0.5762226

newspaper 0.05664787 0.35410375 1.00000000 0.2282990

sales 0.78222442 0.57622257 0.22829903 1.0000000

From the last column of the R output, the correlation between sales and YouTube (0.78 to 2 decimal places) is shown to be stronger than that between sales and Facebook (0.58 to 2 decimal places). The correlation between sales and newspapers is weak (0.23 to 2 decimal places).

After the correlation analysis, further analysis of the relationship between sales and the explanatory variable YouTube is the next step because these two variables have a linear relationship and the strongest correlation.

Save the marketing data in the previously created DATA folder of your desktop as *marketing.csv* to use in later exercises of the lesson:

write.csv(md, file = 'C:/Users/INSERT-YOUR-USER-NAME-HERE/Desktop/DATA/marketing.csv', row.names = FALSE)

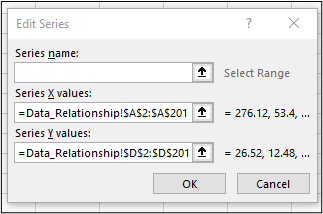
**Correlation analysis using a real dataset in Excel**

**Correlation analysis using a real dataset in Excel**

First, import the *marketing.csv* dataset to a new worksheet.

Create a scatter plot (or chart) to investigate the relationship between sales and YouTube using the following steps:

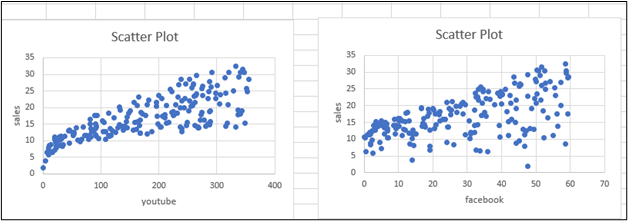
* Change the tab name to Data\_Relationship if it is not already.
* Click on the new worksheet anywhere far away from the data.
* Click **Insert** > **Scatter** (under the charts section) to create an empty scatter chart.
* Click **Chart Design Tools** > **Select Data** to open the **Select Data Source** dialog box.
* Click **Add** in **Legend Entries (Series)** to open the **Edit Series** dialog box.
* Use the **Edit Series** dialog box to choose the appropriate data ranges for each axis. X values should correspond to YouTube values (in the range A2:A201), and Y values should correspond to sales values (in the range D2:D201); refer to **Figure 3-13**.

**Figure 3-13**

* Click **OK**. You can then add a chart title and axes titles to the graph.

Repeat the steps above to construct scatter plots of sales against Facebook and sales against newspaper.

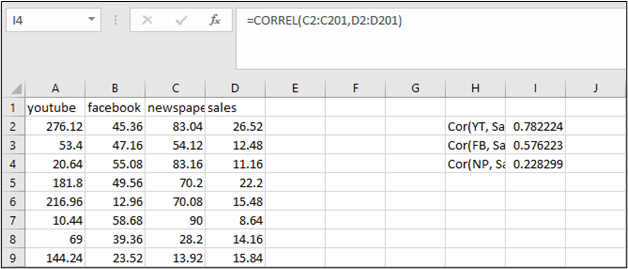
**Figure 3-14** shows scatter plots describing the relationships between YouTube and sales and Facebook and sales.

**Figure 3-14**

To calculate the Pearson correlation coefficient in Excel, use the function CORREL(range1, range2), where range1 and range 2 are the cell references containing the data.

Correlation formulas can be added in the cells I2, I3, and I4 of the worksheet for correlations between YouTube and sales, i.e., =CORREL(A2:A201,D2:D201), Facebook and sales, i.e., =CORREL(B2:B201,D2:D201), and newspaper and sales, i.e., =CORREL(C2:C201,D2:D201), respectively.

**Figure 3-15** shows the results of these computations in Excel.

**Figure 3-15**

**Correlation and scatter charts in Excel Part 1**

**Correlation analysis in Excel**

**Correlation and scatter charts in Excel Part 2**

**Cross tabulation**

**Cross tabulation** is a method used to analyze the relationship between two or more non-numerical variables. Cross tabulation involves grouping variables to determine the correlation between them. The process uses a table called a **crosstab** (also called a **contingency table** or a **two-way** table) to allow you to discover how often a combination of two variable values occurs.

**Cross Tabulation In R**

A food service worker at a local university wants to better understand the food preferences of the students served in the cafeteria. He does a brief survey of lunch students, asking them what food they would like to see added to the menu. He creates the dataset called ct containing two categorical variables food and gender. ct records the favorite foods and genders of 28 college students.

Create ct using the following R code:

food <- c(rep("sushi",13), rep("icecream",8), rep("pizza",7))

gender <- c(rep("Female",6),rep("Male",7),rep("Female",5), rep("Male",3),rep("Female",3),rep("Male",4))

ct <- data.frame(gender, food)

head(ct,5)

This code produces the following output:

gender food

1 Female sushi

2 Female sushi

3 Female sushi

4 Female sushi

5 Female sushi

Next, use ct and the function table() to create a crosstab containing frequencies based on gender and food as follows:

ct\_table <- table(ct$gender,ct$food)

ct\_table

This code produces the following output:

icecream pizza sushi

Female 5 3 6

Male 3 4 7

From the table, the frequencies of various pairs of characteristics are shown. For example, note that 6 females would like to see sushi added to the menu, while 5 females would like to see ice cream added to the menu.

A **proportions** table can also be created where the cell values are proportions of the total number of entries in the table (i.e., 28) using the following R code:

p\_table <- prop.table(ct\_table)

p\_table

This code produces the following output:

icecream pizza sushi

Female 0.1785714 0.1071429 0.2142857

Male 0.1071429 0.1428571 0.2500000

In this example, 25 percent of students surveyed are males who would like sushi added to the menu and 21 percent of students surveyed are females who would like sushi added to the menu.

Save the dataset *ct* as *ct.csv* in the previously created DATA folder to use in later exercises:

write.csv(ct, file = 'C:/Users/INSERT-YOUR-USER-NAME-HERE/Desktop/DATA/ct.csv', row.names = FALSE)

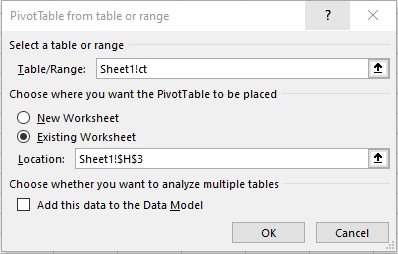
**Cross Tabulation In Excel**

Import ct.csv to a new worksheet:

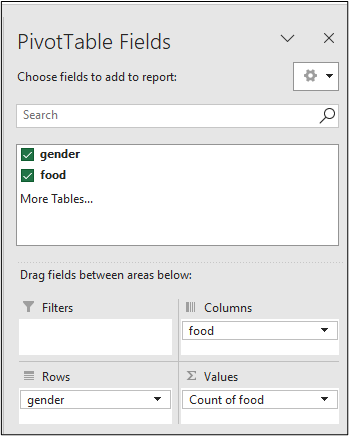
1. Create a new Excel workbook.
2. Click on the **Data** tab in the Excel ribbon.
3. Click the **Get Data/ From File / From Text/CSV** button, click on **Text/CSV**.
4. Click **Browse** and navigate to ct.csv and click **Import**.
5. At the next window, select **Next**.
6. At the next window, select **Transform Data**. This opens the Power Query Editor.
7. In the Power Query Editor, you will see a preview of your data in a table format.
8. Click on the **Load** button in the Power Query Editor. This will load the transformed data into your Excel worksheet. The data will be displayed in Excel with the first row as headers, as specified in the Power Query Editor.

Next, click **Insert** > **PivotTable** > **From Table/Range**.

The **PivotTable From Table/Range** dialog box appears. Select the range that contains the data and a cell in the **Existing Worksheet** in which to place the crosstab (e.g., cell H3), as shown in **Figure 3-16**. Click **OK**.

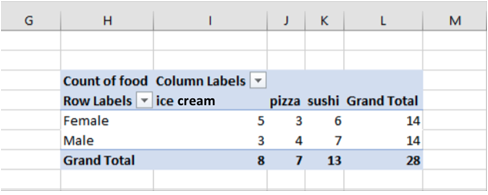
**Figure 3-16**

In the **PivotTable Fields** window that appears, drag the variable gender to the Rows area, the variable food to the Columns area, and the variable food again to the Values area, as shown in **Figure 3-17**.

**Figure 3-17**

The crosstab in **Figure 3-18** should appear after completing the steps outlined above.

**Figure 3-18**



**Identifying outliers in a dataset**

An **outlier** is a data point that is far from other points. In other words, it differs significantly from other data points.

Outliers in a dataset can be the result of measurement errors, data entry errors, or sampling problems. For example, a height of 155 m in a dataset containing human heights is obviously an error and the result of human error when inserted into the dataset.

Outliers can heavily influence statistical results, like the mean and standard deviation, resulting in misleading interpretations. Therefore, an analyst should identify any outliers present in a dataset and then decide what to do with them.

**How To Detect Outliers In A Dataset**

Outliers are best detected using the interquartile range (IQR) and visualizations, such as histograms.

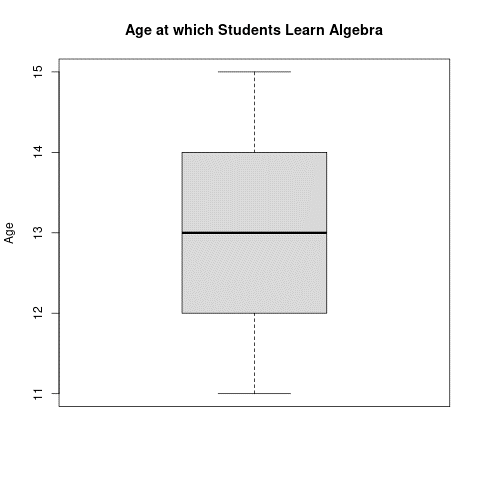
A **histogram** is a diagram used to study the distribution of numerical data. It is a type of bar plot in which every bar represents a class of data. The heights of the bars are the frequencies, the number of occurrences for each category, of the data classes. The bars have equal widths and touch each other. An example of a histogram is shown in **Figure 3-19**.

**Figure 3-19**

Typically, outliers will be found at the far left (extremely small values) or far right (extremely large values) of the histogram.

The **interquartile range** measures the spread of the middle half of a dataset. It is calculated using the formula IQR = *Q*3 – *Q*1, where *Q*3 is the third quartile (or upper quartile) and *Q*1 is the first quartile (or lower quartile).

Quartiles are values that divide a dataset into four equal parts. As such, there are three quartiles, namely, *Q*1, *Q*2, and *Q*3. *Q*2 is also the median of the data. An example of how quartiles can be plotted in a box and whiskers plot is shown in **Figure 3-20**.

**Figure 3-20**

To find quartiles from a dataset, the dataset must be arranged in ascending order.

**How To Find Outliers Using The IQR**

A data point is considered an outlier if it is less than *Q*1 – 1.5(IQR) or more than *Q*3 + 1.5(IQR).

Consider the following data containing the heights of 10 individuals in centimeters. One of the measurements is an outlier.

155, 167, 300, 168, 188, 170, 180, 177, 165, 175

The methods outlined above help to determine whether the data contains an outlier.

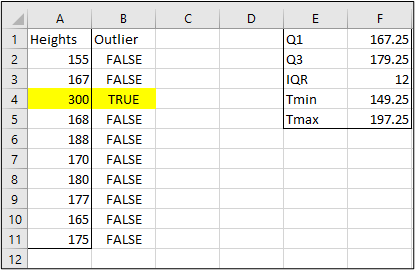
**Detecting outliers using R**

|  |
| --- |
| Code implementation |
| Histogram | 1. Store the data in a variable called heights. 2. Use the function hist() to create a histogram. 3. heights <-c (155, 167, 300, 168, 188, 170, 180, 177, 165, 175) 4. hist(heights, breaks = 6)   Note: By setting breaks = 6, you are specifying that you want the histogram to have 6 bins. The number of bins determines the granularity or level of detail in the histogram.  **Figure 3-21**Histogram of height, output from R that shows most data clustered in the first three bins, with a single value in the 6th bin. |
| IQR | 1. Use the function IQR() to obtain the interquartile range. 2. Use the summary() function to obtain  and . 3. Obtain threshold values for the outliers, i.e., Tmin and Tmax denoting minimum and maximum threshold, respectively (Tmin =  – 1.5(IQR) and Tmax =  + 1.5(IQR)). 4. Outliers should be all points greater than Tmax and all points less than Tmin. 5. heights<-c(155, 167, 300, 168, 188, 170, 180, 177, 165, 175)   IQR(heights)  This code produces the following output:  [1] 12  Run the following code to determine  and :  summary(heights)  This code produces the following output:  Min. 1st Qu. Median Mean 3rd Qu. Max.  155.0 167.2 172.5 184.5 179.2 300.0  Run the following code to determine the minimum and maximum threshold:  Tmin = 167.2 - (1.5\*12)  Tmax = 179.2 + (1.5\*12)  heights[which(heights < Tmin | heights > Tmax)]  This code produces the following output:  [1] 300  From the R output, you can see that the outlier is 300. |

**Finding outliers in Excel**

You can use formulas or a histogram to detect outliers in Excel. Let's look at both of these techniques.

1. Enter the data in cells A2:A11 in a new Excel worksheet. Compute Q1 by entering the formula =QUARTILE(A2:A11,1) in cell F1.
2. Compute Q3 by entering the formula =QUARTILE(A2:A11,3) in cell F2.
3. Compute the IQR by entering the formula =F2-F1 in cell F3.
4. Compute Tmin by entering the formula =F1-(1.5\*F3) in cell F4.
5. Compute Tmax by entering the formula =F2+(1.5\*F3) in cell F5.
6. To determine whether a data point is greater than Tmax or less than Tmin, enter the formula =OR(A2>$F$5, A2<$F$4) in cell B2. Copy this formula to the remaining cells in column B by double-clicking on the fill handle of the cell. A TRUE value should appear for all outliers in the data, as shown in **Figure 3-22**.

**Figure 3-22**

OR

Create a histogram in Excel:

1. Select the data: Select the range of cells that contain your data.
2. Insert a histogram: Go to the **Insert**tab in the Excel ribbon and click on the **Recommended Charts** or **Insert Statistic Chart** button (the exact location may vary depending on your Excel version).
3. Choose the histogram: In the **Recommended Charts** or**Insert Chart** window, select the **Histogram**option. It is typically found under the **Column** or **Bar**chart category.
4. Customize the histogram: After inserting the histogram, you can customize its appearance and settings. Right-click on the chart and select **Format Chart Area** or use the various formatting options available in the ribbon to modify the chart's appearance, labels, axes, etc.
5. Adjust the bin size: By default, Excel automatically determines the bin size for the histogram. However, you can adjust the bin size to fit your needs. Right-click on the horizontal axis of the histogram and select **Format Axis** to open the axis formatting options. In the **Axis Options** panel, you can specify the bin width or number of bins under the **Bounds**or **Axis Options** section.

**Data drilling**

**Data drilling** is a method of analyzing data by pulling out statistically interesting subsets or subcategories. It involves an in-depth investigation of the underlying data to allow an analyst to understand it better and enhance the decision-making process. One key component of data drilling is **granularity**. Granularity is how separately and distinctly you view each data point. Sometimes looking closer at data (more granular), you can notice features that aren’t apparent when looking at the dataset as a whole. Sometimes looking at the dataset from a broader perspective (less granular) allows you to see overall trends or patterns. For example, if a sales report shows a decline in sales, an analyst can drill down into the report to find the products or departments that contributed to this decline.

**Types Of Data Drilling**

There are two main types of data drilling:

* Drill down
* Drill up

**Data Drill Down**

Data drill down starts by looking at the calculated summary statistics for details on the underlying data. It helps the analyst shift from an overall view of the data to a more detailed and specific view.

**Data Drill Up**

Data drill up is the opposite of data drill down. It enables the data analyst to shift from a detailed view of the data to a more overall view of the situation.

**Data mining**

**Video: Describe data mining**

**Data mining** is the process of extracting information from large datasets. It involves analyzing large datasets to find odd entries, patterns, and correlations that can help solve business problems. For example, businesses can use data mining techniques to develop customer profiles from customer data. These profiles help businesses find their best customers and tailor marketing strategies to attract others with similar behaviors.

Although data mining is often used interchangeably with machine learning, the two terms are different. **Machine learning** involves using data and algorithms to develop methods that learn and change in response to reinforcement or new data, similar to the way humans learn.

The following are some examples of data mining techniques:

**Anomaly Detection**

This technique is also referred to as outlier analysis. It helps identify suspicious and rare events that differ significantly from most of the data.

An **anomaly** is an event or item that does not follow the expected pattern. An example of an anomaly is a spike in unsuccessful login attempts in an online banking system.

**Clustering**

This technique finds clusters or groupings of data points that are similar to one another in a dataset. It aims to make groups that are similar on the inside of the cluster, while making sure the clusters are as different as possible from one another. For example, business owners can use clustering to identify distinct groups of customers and develop marketing strategies specific to each group.

**Trends and interpretation introduction**

The primary goal of data analysis is to get fair and honest information from data. These insights should be supported by the data and have practical value. In other words, analysts should be able to use the results of their analysis to make recommendations that can cause real change in an organization. Exploratory data analysis helps uncover patterns in data, and provides the context needed for these patterns. Exploratory data analysis identifies the most important variables in your dataset. Analysts can then perform more in-depth tests to obtain detailed predictions and insights from the data.

This skill covers how to:

* Perform a simple linear regression
* Interpret the results of a simple linear regression
* Use regression analysis for prediction

**Simple linear regression**

Scatter plots and correlations are methods used to explore the relationship between two numerical variables. If a scatter plot and a correlation coefficient reveal that the relationship between two variables is linear and strong, this relationship can be explored further using techniques such as linear regression.

**Simple Linear Regression**

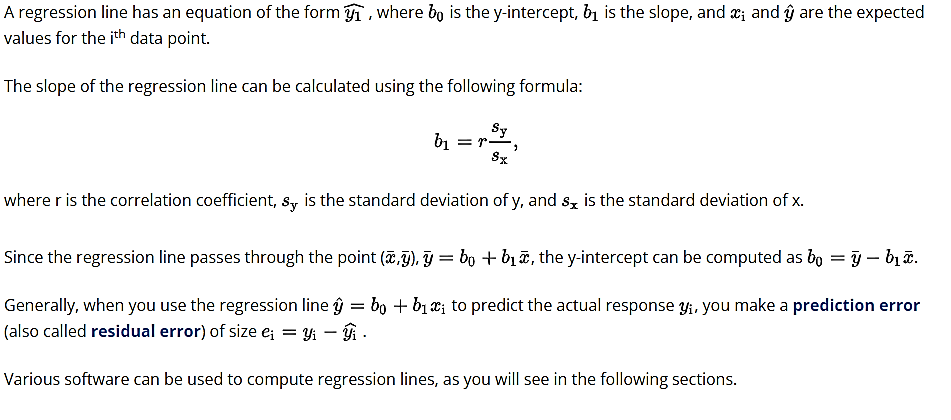
**Simple linear regression** is a statistical method used to study the relationship between two continuous variables.

The variable that is changing in response to another factor is called the **response variable**.

The variable used to predict the response is called the **explanatory** or **predictor variable**.

For example, if a dataset is comparing the height of basketball players and their shooting percentage, the predictor variable would be the player height and the response variable would be their shooting percentage.

Typically, data points do not exactly fit a straight line. One method for finding the line that best fits the data as a whole is called **least squares regression**. This method is similar to the variance calculations from earlier, in that it compares how far each data point is from a line and finds the line that makes these distances the smallest possible overall. A **regression line** is a line that best fits the data.



**Simple linear regression using a real dataset in R**

**Video: Fit simple linear models**

Earlier, the marketing dataset from the datarium package was used in R to explore relationships in data. A strong linear relationship was found to exist between the variables YouTube and sales. In this section, the relationship between these two variables will be explored more using simple linear regression.

Load the marketing dataset into a variable in R using the following code:

require(datarium)

md <- marketing

Use the function lm() to perform linear regression with the variables sales and YouTube. In this case, sales is the response variable, and YouTube is the explanatory variable.

model <- lm(sales~youtube, data=md)

summary(model)

This code produces the following output:

Call:

lm(formula = sales ~ youtube, data = md)

Residuals:

Min 1Q Median 3Q Max

-10.0632 -2.3454 -0.2295 2.4805 8.6548

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.439112 0.549412 15.36 <2e-16 \*\*\*

YouTube 0.047537 0.002691 17.67 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.91 on 198 degrees of freedom

Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099

F-statistic: 312.1 on 1 and 198 DF, p-value: < 2.2e-16

**Interpretation Of R Output**

* + The estimated regression line equation is sales = 8.439 + 0.048 · youtube. The value of the y-intercept is 8.439 (to 3 decimal places) and the slope is 0.048 (to 3 decimal places). Note: the coefficient labeled as "Estimate" under the "Coefficients" section represents the slope of the YouTube estimated regression line.
  + Alternatively, compute the slope of the regression line using the formula  (to 3 decimal places), where r is the correlation coefficient,  is the standard deviation of the dependent variable (sales), and  is the standard deviation of the independent variable (YouTube). Calculate this value in R using the following code:
  + cor(md$youtube,md$sales)\*(sd(md$sales)/sd(md$youtube))

This code produces the following output:

[1] 0.04753664

* + Alternatively, compute the y-intercept of the regression line using the formula ▁▁ (to 1 decimal place). Calculate this value in R using the following code:
  + mean(md$sales) - (0.048 \* mean(md$youtube))

This code produces the following output:

[1] 8.357352

Note: The slope and y-intercept calculated here are slightly different from the results of the R linear regression due to rounding the slope to two significant figures.

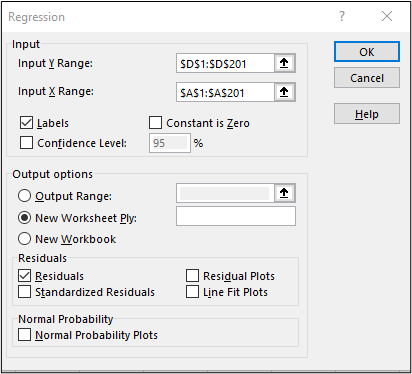
* + When the YouTube advertising budget is 0, sales are expected to amount to 8.439 = 8,439 dollars (recall that sales and YouTube units are in thousands of dollars)
  + A 1 unit increase in the YouTube budget should result in a 0.048 unit increase in sales. As sales and YouTube units are given in thousands of dollars, it means that a 1000 dollar increase in the YouTube budget should result in a 48-dollar increase in sales.

**Simple linear regression using a real dataset in Excel**

Import the *marketing.csv* dataset in the folder DATA to a new Excel worksheet.

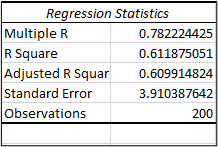
Perform a regression analysis using the **Data Analysis ToolPak** in Excel.  
To install the ToolPak, click **File > Options > Add-ins > Analysis ToolPak**.  
In the **Manage** drop-down list, select **Excel Add-ins** and click **Go**.  
In the **Add-ins** window that pops up, select **Analysis ToolPak** and click **OK**.  
The **Data Analysis** button appears on the **Data** tab.

Next, click the **Data Analysis** button on the **Data** tab.  
Select **Regression** and click **OK**.  
In the **Regression** dialog box that pops up, configure the input Y range (sales) and input X range (YouTube).  
Check **Labels** if your X and Y ranges contain the headers YouTube and sales, respectively (the range for Y should be $D$1:$D$201 and that for X should be $A$1:$A$201).  
Under the **Output option**, select **New Worksheet Ply**. Check **Residuals** to obtain the difference between the predicted and actual values. Click **OK**.  
**Figure 3-23** displays the **Regression** dialog box populated with required values.

**Figure 3-23**

**Interpreting The Regression Analysis Output**

The table in **Figure 3-24** provides statistical measures that indicate how well the model fits the data.

**Figure 3-24**

**R-square** is a statistical measure that explains how much of the variance in the response variable (sales) is explained by the explanatory variable (YouTube). Often, the larger the value of R-square, the better the regression model fits your observations. The R-squared value of **0.6119** indicates that the YouTube predictor accounts for approximately 61% of the variance in sales. The **Multiple R** is the correlation coefficient that we computed earlier.

The **standard error of the regression** is a summary of how far each of the observed values falls from the regression line. In this example, the distance is **3.91**. A low distance value is better. Such a value would indicate that the distances between the data points and the fitted values are small.

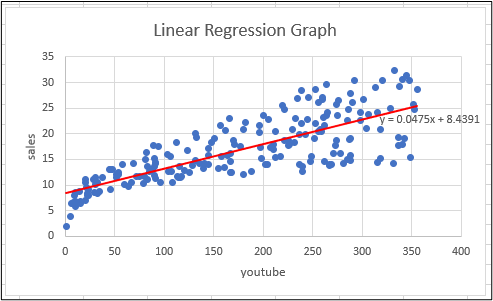
An analyst can show the relationship between sales and YouTube using a linear regression chart.

To create this chart, first, create a scatter plot of sales against YouTube using the method from earlier in the lesson.

Now, draw the least squares regression line. Right click on any point and select **Add Trendline** from the context menu.

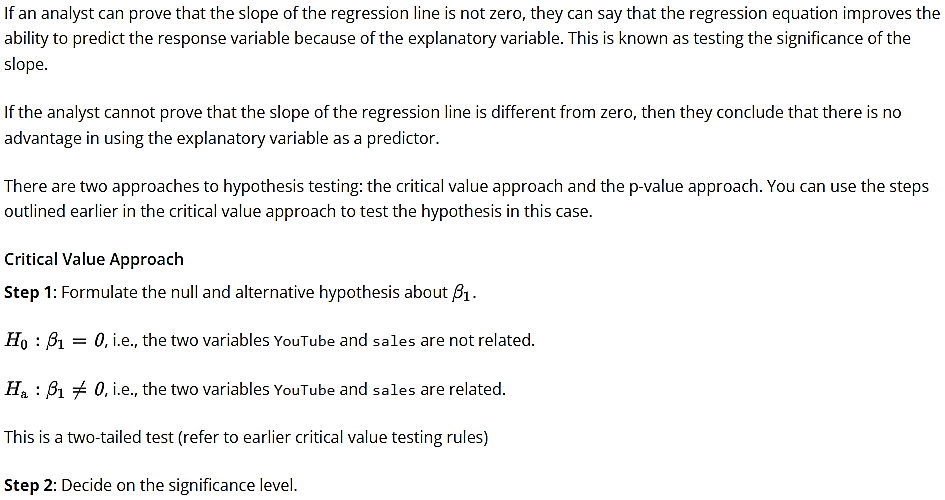
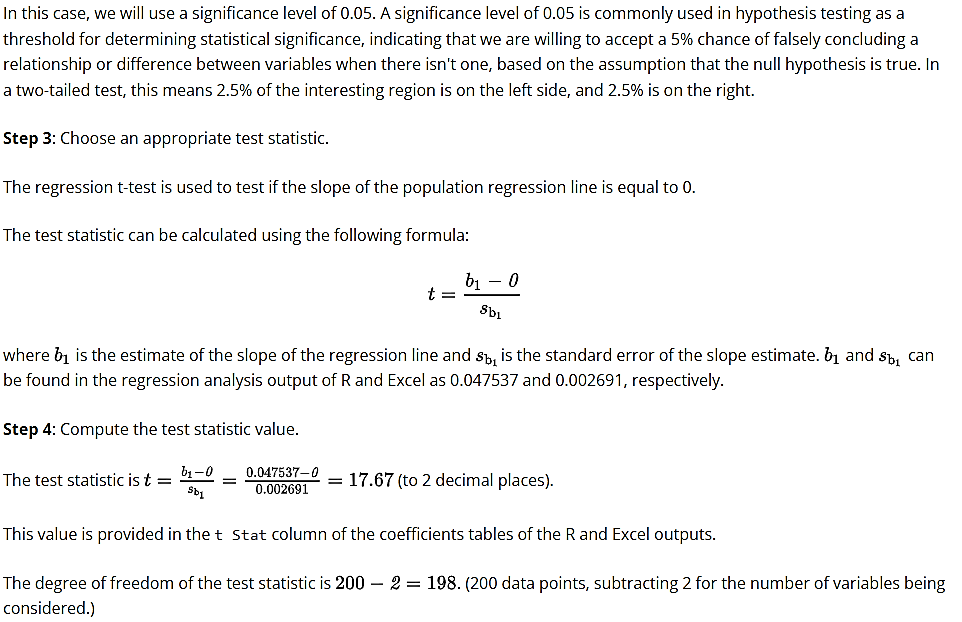
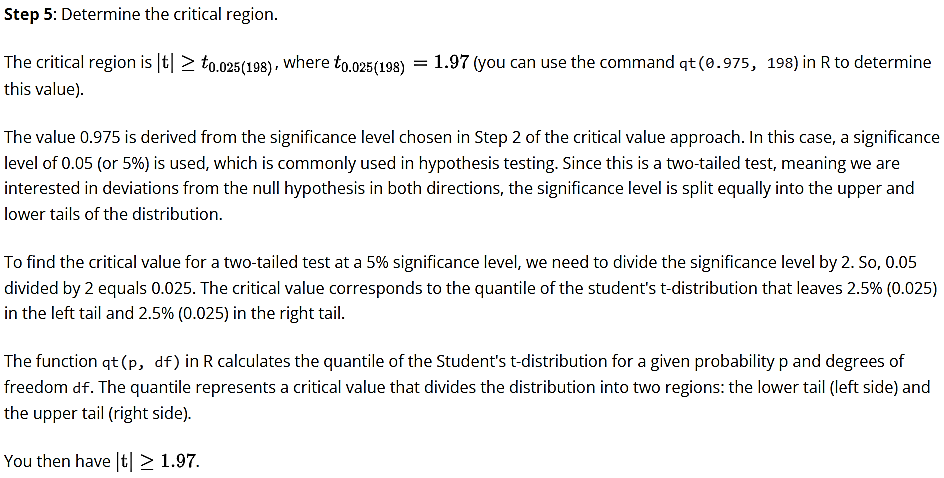
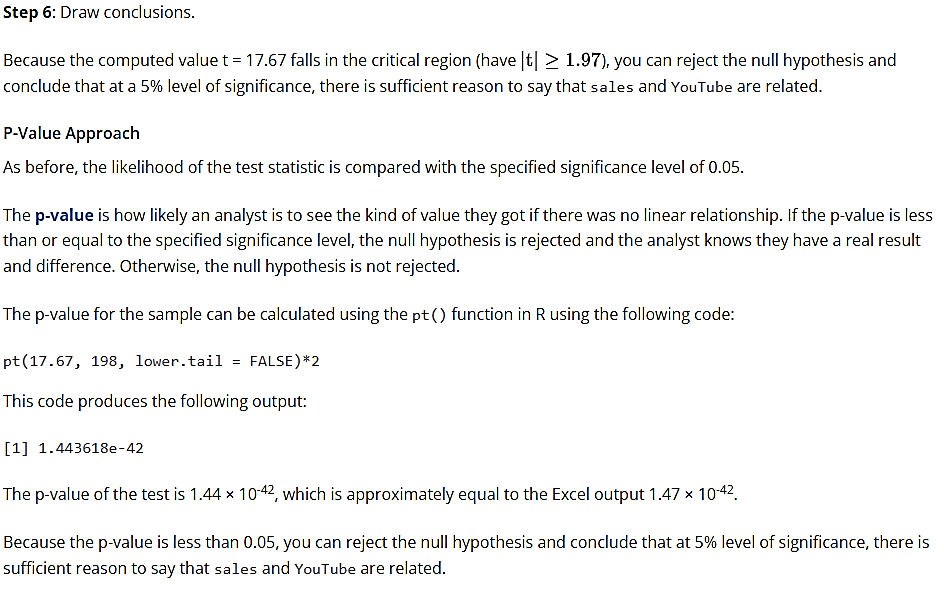
On the right pane that appears, select **Linear** under **Trendline Options** and check **Display Equation on Chart**.

Use the **Fill & Line** tab to customize the line, e.g., change the line to a solid line, the color of the line to red, and the dash type to unbroken line. **Figure 3-27** shows the resulting linear regression chart.

**Figure 3-27**

You will see that the regression line in **Figure 3-27** has the same coefficients as those obtained from the R and Excel regression outputs.

**Testing the significance of the slope**

**Using regression analysis for prediction**

Simple linear regression helps you to determine the relationship between a response and an explanatory variable. A simple linear regression model can also be used to predict the values of new observations.

In this section, you will use the marketing dataset from the datarium package in R. In the previous section, you developed a simple linear model using the variables YouTube and sales. In this section, you will use this model to answer the following question:

What do you predict the sales will be when YouTube is (1) 200 and (2) 340?

**Implementation In R And Excel**

|  |  |
| --- | --- |
|  | Code implementation |
| R | 1. Load the marketing data in a variable called md. 2. Perform simple linear regression using the variables sales and YouTube. 3. Use the function predict() to predict sales when YouTube is (i) 200 and (ii) 340: 4. require(datarium) 5. md <- marketing 6. model <- lm(sales~youtube, data=md) 7. predict(model, data.frame(youtube = c(200, 340)))   This code produces the following output:  1 2  17.94644 24.60157 |
| Excel | 1. Import *marketing.csv* to a new worksheet. The sales data are in the range D2:D201, and YouTube data are in the range A2:A201. 2. Create a column called newYT (cell H1), denoting new YouTube values. 3. Enter the values 200 and 340 in the column newYT (cells H2 and H3). 4. Create a column called predicted (cell I1). 5. Use the FORECAST function to predict the sales for the two new YouTube values. Enter the formulas:   =FORECAST(H2,D2:D201,A2:A201) in cell I2 and  =FORECAST(H3,D2:D202,A2:A202) in cell I3.  The result is shown in **Figure 3-28**. **Figure 3-28**A table showing the prediction results in Excel. |

From the analysis above, you can see that sales are predicted to be 17.964 and 24.602 when YouTube spend is 200 and 340, respectively. All values are in thousands of dollars.

**Role of artificial intelligence introduction**

Artificial intelligence (AI) studies how to make machines perform tasks commonly associated with human beings. It looks at how the human brain works and how human beings learn and make decisions when solving problems. AI then uses these results to develop systems that are adaptive and can learn progressively, much like humans.

Today, the volume, velocity, and variety of data generated in the world are massive. In other words, the data is too big, moves too fast, and data sources are numerous. These characteristics of big data require artificially intelligent systems that can help human beings in handling data.

This skill covers how to:

* Define artificial intelligence, algorithm, machine learning, and deep learning
* Discuss how machine learning algorithms help in data analysis
* Discuss how artificial intelligence algorithms work in data analysis

**Artificial intelligence and machine learning**

**Video: Machine learning and artificial intelligence**

**Artificial intelligence** (AI) is a broad field of science concerned with building machines that can perform tasks requiring human intelligence. It aims to design machines with human-like intelligence.

An **algorithm** is a step-by-step procedure used for solving a problem. Algorithms are crucial components of AI systems. These systems use algorithms to perform calculations, process data, analyze data, detect anomalies in data, etc.

**Machine learning** is a branch of AI that aims to develop systems that can learn from the data they receive. Machine learning algorithms use sample data to build models that can perform laid-out tasks. For example, models can help organizations to predict possible outcomes based on historical data. A simple example of a machine learning algorithm is a simple linear regression model.

**Uses Of Machine Learning In Data Analysis**

* Collecting and analyzing various data types
* Exploring and cleaning datasets
* Building and training models for predictive purposes

**Deep learning** is a branch of AI that uses algorithms called artificial neural networks to learn from data. Artificial neural networks are designed to think and learn like humans. Examples of systems based on deep learning are self-driving cars, virtual assistants, and facial recognition.

Machine learning and deep learning are two well-known subsets of AI.

**Use Of Artificial Intelligence In Data Analysis**

AI is typically used to analyze big data to find patterns that can be used to derive insights to improve work processes.

**Big data** can be defined as data that is too much in scope for desktop software or calculators to process and analyze. The three features of big data are volume, velocity, and variety.

Big data and AI complement each other. AI requires considerable data to learn, and big data analytics requires AI technologies for efficient data analysis.

**Uses Of AI In Data Analytics Process:**

* Building new data analysis methods
* Processing large volumes of data quickly
* Solving common data problems, e.g., detecting outliers and missing values, de-duplicating data, or reducing dimensions of data
* Performing various types of data analyses, from simple descriptive and diagnostic analyses to complex predictive and prescriptive analyses

**Summary**

* Data analysis is the process of collecting, cleaning, transforming, and processing data to yield information that can be useful in decision-making.
* The main methods for data analysis are descriptive analysis, diagnostic analysis, predictive analysis, prescriptive analysis, and hypothesis testing.
* Descriptive analysis answers the question “What happened?”.
* Diagnostic analysis answers the question “Why did it happen?”.
* Predictive analysis uses current and historical data to answer the question “What might happen in the future?”.
* Prescriptive analysis answers the question “What should be done?”.
* Hypothesis testing is a method of data analysis that uses data from a sample to draw conclusions about the overall population.
* Data aggregation is the process of collecting data from multiple sources and working it into a summarized form.
* Data interpretation is how an analyst attaches meaning to processed and analyzed data.
* Exploratory data analysis summarizes datasets by their main characteristics.
* Correlation is a statistical measure that explains how much two variables are linearly related.
* An outlier is a data point that is far from other points.
* Cross tabulation is a method used to analyze the relationship between two or more categorical variables.
* Data drilling is a method of analyzing data by providing different perspectives of the data in reports or spreadsheets.
* Data mining is the process of extracting information from large datasets.
* Simple linear regression is a statistical method used to study the relationship between two numerical variables.
* Artificial intelligence works to make machines perform tasks commonly associated with human thinking and decision-making abilities.
* Machine learning focuses on developing systems that learn from the data they receive.