Unit 1
Introduction to Data Science

Dr. S. Thenmozhi

Contents

- What is Data Science? Where does data comes from?
- Working with Data at Scale
- Making Data tell its story
- Data Scientists
- Data Science in a Big Data world
- Facets of data
- The Data Science Process.
- Data
- Variables
- Organization of Data
- Describing Data by Tables and Graphs

What is data science?

- Producing Data-driven apps
- Using data is not data science, bringing insights is Data Science
- It should not be application with data
- It should be a data product
  - Eg: Google: Page ranking in web search, Spell checking, Voice search
  - LinkedIn and Facebook: patterns of relationships
  - Amazon: appropriate recommendations

Where does data come from?

- Moore's Law : Number of transistors per square inch on integrated circuits had doubled every year since their invention.
  Most experts expect Moore's law to hold for another two decades
- The extension of Moore's law is that computers, machines that run on computers, and computing power all become smaller and faster with time, as transistors on integrated circuits become more efficient.

Where does data come from?

- Web server logs
- Tweets
- Online transaction records
- Data from sensors
- Government data
- Mashup data
- Business firms
- Even your body….

How Data Science different from statistics?

- Statistics is a segmented approach – It is the grammar of data science – basic skill
- Data science is a Wholistic approach – not just making guesses. It is about testing hypotheses and making sure the conclusions drawn from the data are valid for the universe.

Working with data at scale

- Big data - ***Big data*** is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy.
- Data warehouse – history/repository of data
- Bigdata
- Data warehouse

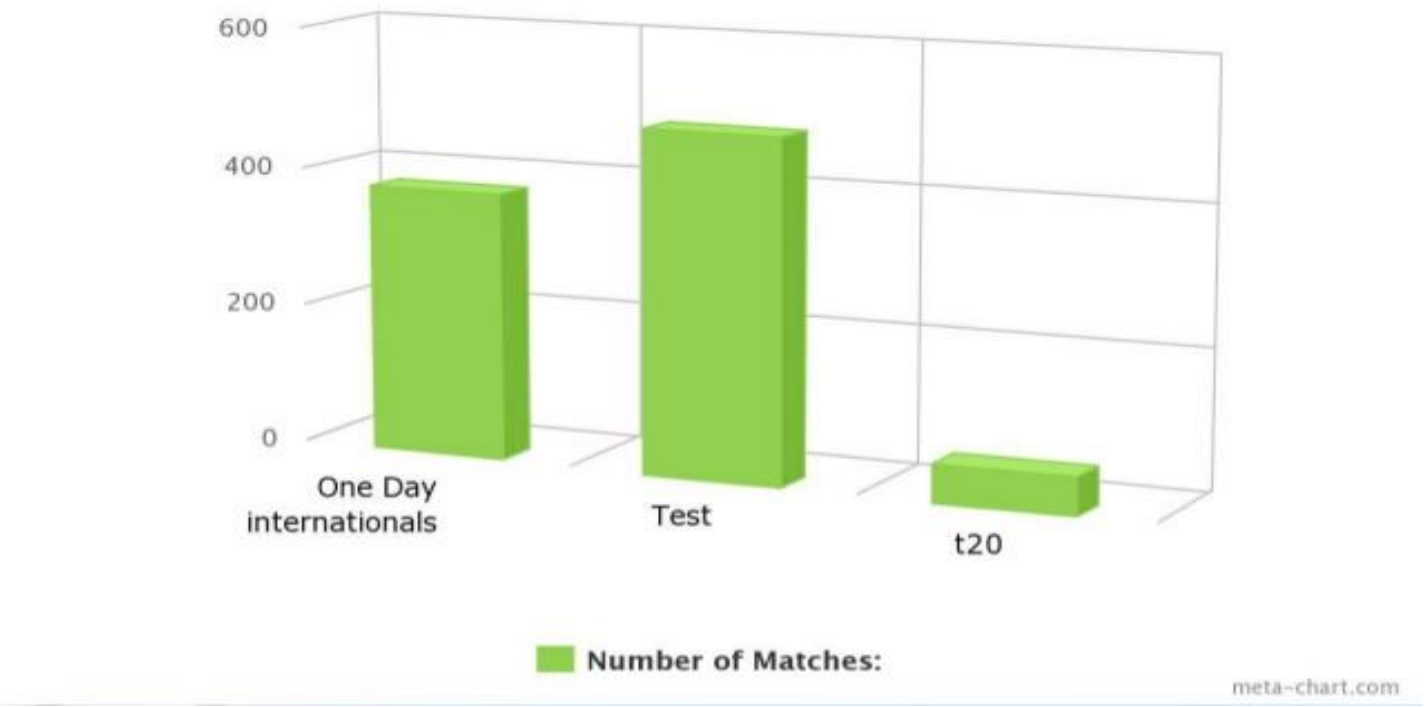
Schemas evolve as understanding of data changes

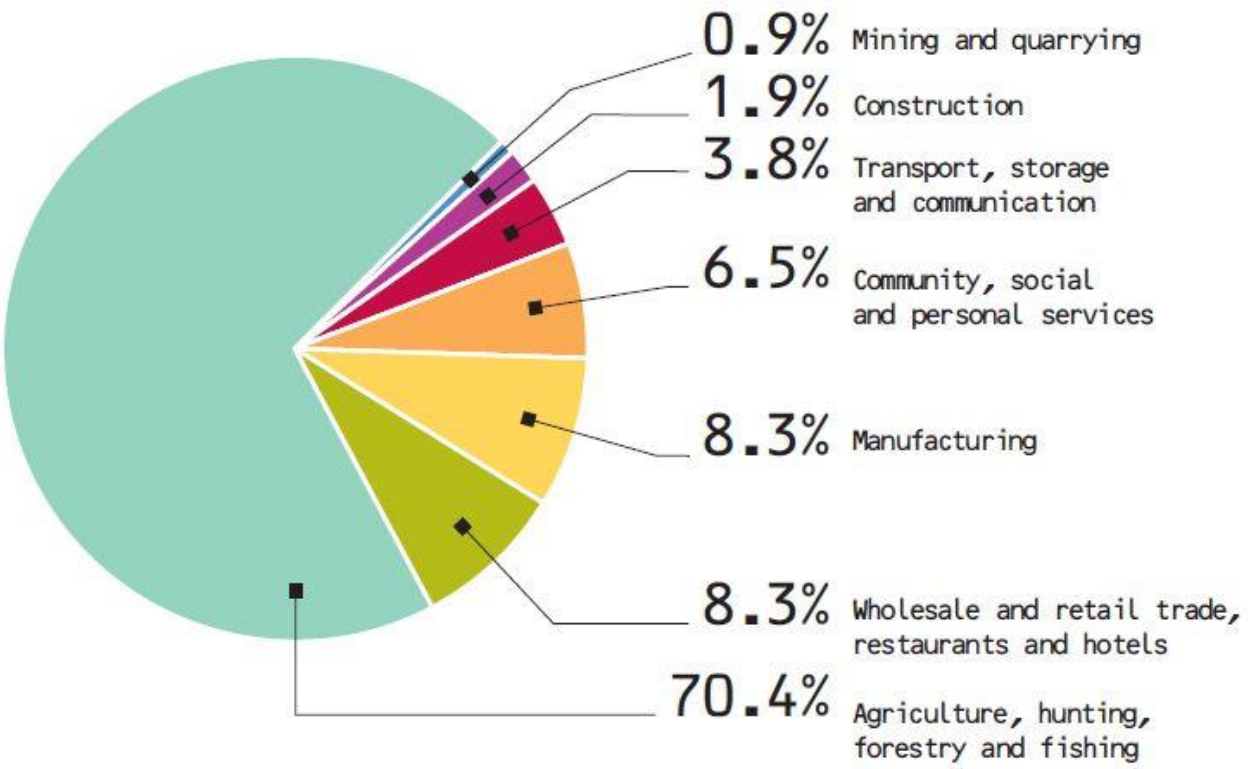
- NoSQL Databases
  - Cassandra - Facebook
  - Hbase, BigTable – Google (Billions of rows and millions of columns)
  - Dynamo – Amazon
- Bigger databases give rise to computational problems => MapReduce Approach (similar to divide and conquer strategy)

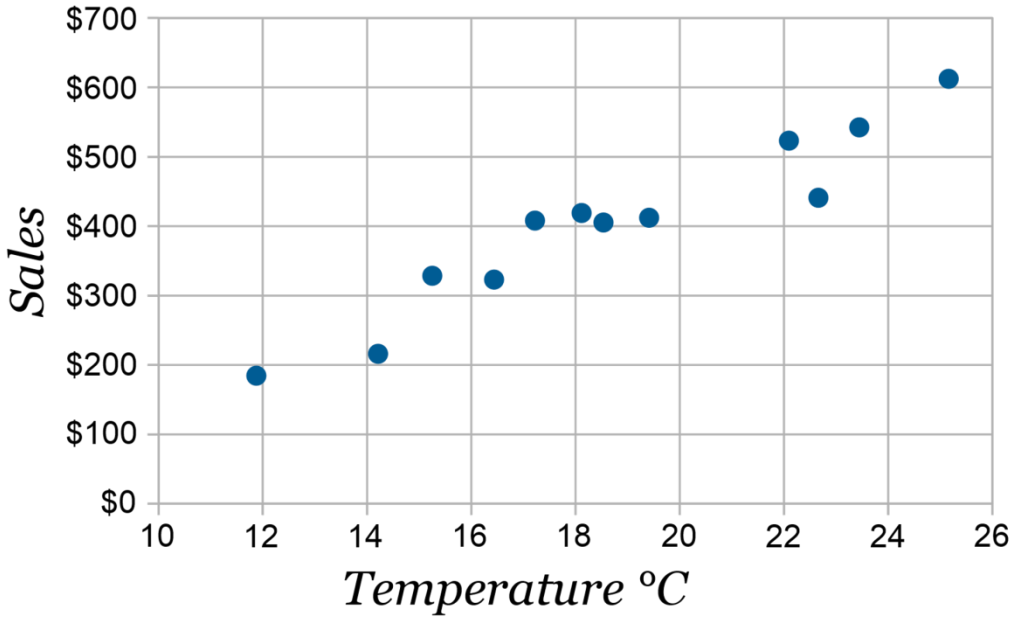
Making data tell its story

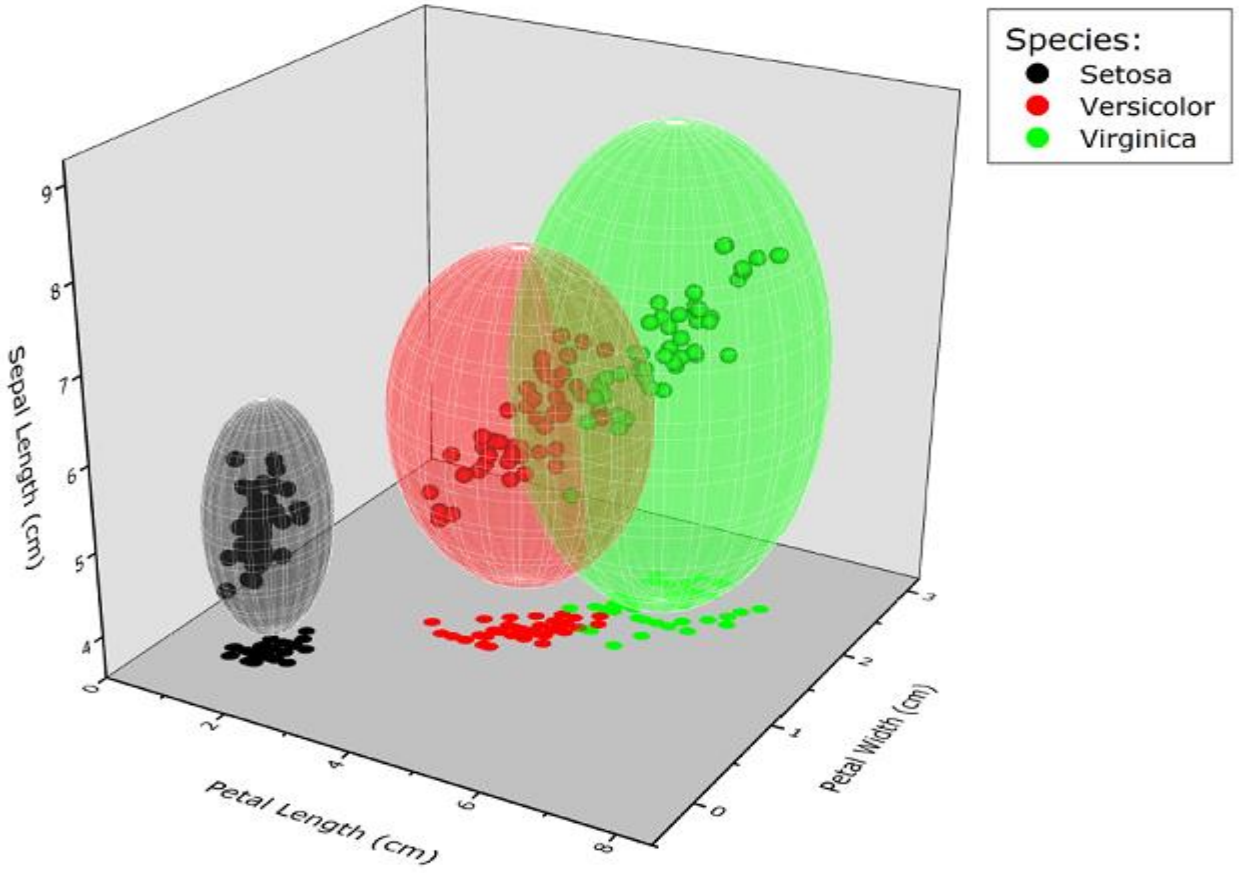
- A picture is worth of thousand words.
- Similarly the graph tells more about the data – visualizing data helps to understand data much better
- It is the first step of analysis

BAR GRAPH


600
400
200
0
One Day internationals
Test
t20
Number of Matches:
meta-chart.com


0.9% Mining and quarrying
1.9% Construction
3.8% Transport, storage and communication
6.5% Community, social and personal services
8.3% Manufacturing
8.3% Wholesale and retail trade, restaurants and hotels
70.4% Agriculture, hunting, forestry and fishing


$700
$600
$500
$400
$300
$200
$100
$0
Sales
10
12
14
16
18
20
22
24
26
Temperature °C


Species:
Setosa
Versicolor
Virginica
Sepal Length (cm)
Petal Length (cm)
Petal Width (cm)

Data Scientists

- Skills ranging from traditional computer science to mathematics
- Mathematician, Computer scientist, Physicists - any one who have strong mathematical background, computing skills can be a data scientist
- Think about big picture on big problem
- Data scientists combine entrepreneurship with patience, the willingness to build products incrementally, the ability to explore, the ability to iterate over a solution.
- It is interdisciplinary

- The ability to take data – to be able to understand it, process it, to extract value from it, to visualize it, to communicate it – that's going to be hugely important skill in the next decades.

Facets of Data

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

- Structured data - data that depends on a data model and resides in a fixed field within a record Eg: Excel files, Databases
- Unstructured data - data that isn't easy to fit into a data model because the content is context-specific or varying
- Natural language - special type of unstructured data; it's challenging to process because it requires knowledge of specific data science techniques and linguistics.

- Machine-generated data - information that's automatically created by a computer, process, application, or other machine without human intervention.
- Graph based - any data can be shown in a graph Eg: social media websites
- Audio, image, and video – data captured from multimedia
- Streaming data - The data flows into the system when an event happens instead of being loaded into a data store in a batch. Eg: What is trending??

1	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%

CDA@engineer.com
To xyz@program.com
Today 10:21
An investment banking client of mine has had the go ahead to build a new team of UI engineers to work on various areas of a cutting-edge single-dealer trading platform.
They will be recruiting at all levels and paying between 40k & 85k (+ all the usual benefits of the banking world). I understand you may not be looking. I also understand you may be a contractor. Of the last 3 hires they brought into the team, two were contractors of 10 years who I honestly thought would never turn to what they considered “the dark side.”

```
CSIPERF:TXCOMMIT;313236
2014-11-28 11:36:13, Info                  CSI    00000153 Creating NT transaction (seq
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000154 Created NT transaction (seq 69)
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info                  CSI    00000155@2014/11/28:10:36:13.471
Beginning NT transaction commit...
2014-11-28 11:36:13, Info                  CSI    00000156@2014/11/28:10:36:13.705 CSI perf
trace:
```

Data Science Process

- Setting the Research Goal – What, how, why?
- Retrieving Data – Finding and getting access from the data owner
- Data Preparation – raw data to be polished, transformed
- Data Exploration – look for patterns, correlations and deviations
- Data Modelling – gain the insights of the data
- Presentation and Automation – Present your results, automate if needed

- Setting the Research goal
  - Define research goal
  - Create project charter
- Retrieving data
  - Internal data
  - External data

- Data Preparation
  - **Data Cleansing**
    - Errors from data entry
    - Physically impossible values
    - Missing values
    - Outliers
    - Spaces, Typos..
    - Error Against codebook

- **Data Transformation**
  - Aggregating Data
  - Extrapolating Data
  - Derived Measures
  - Creating Dummies
  - Reducing number of Values
- **Combining Data**
  - Merging/Joining data sets
  - Set Operators
  - Creating Views

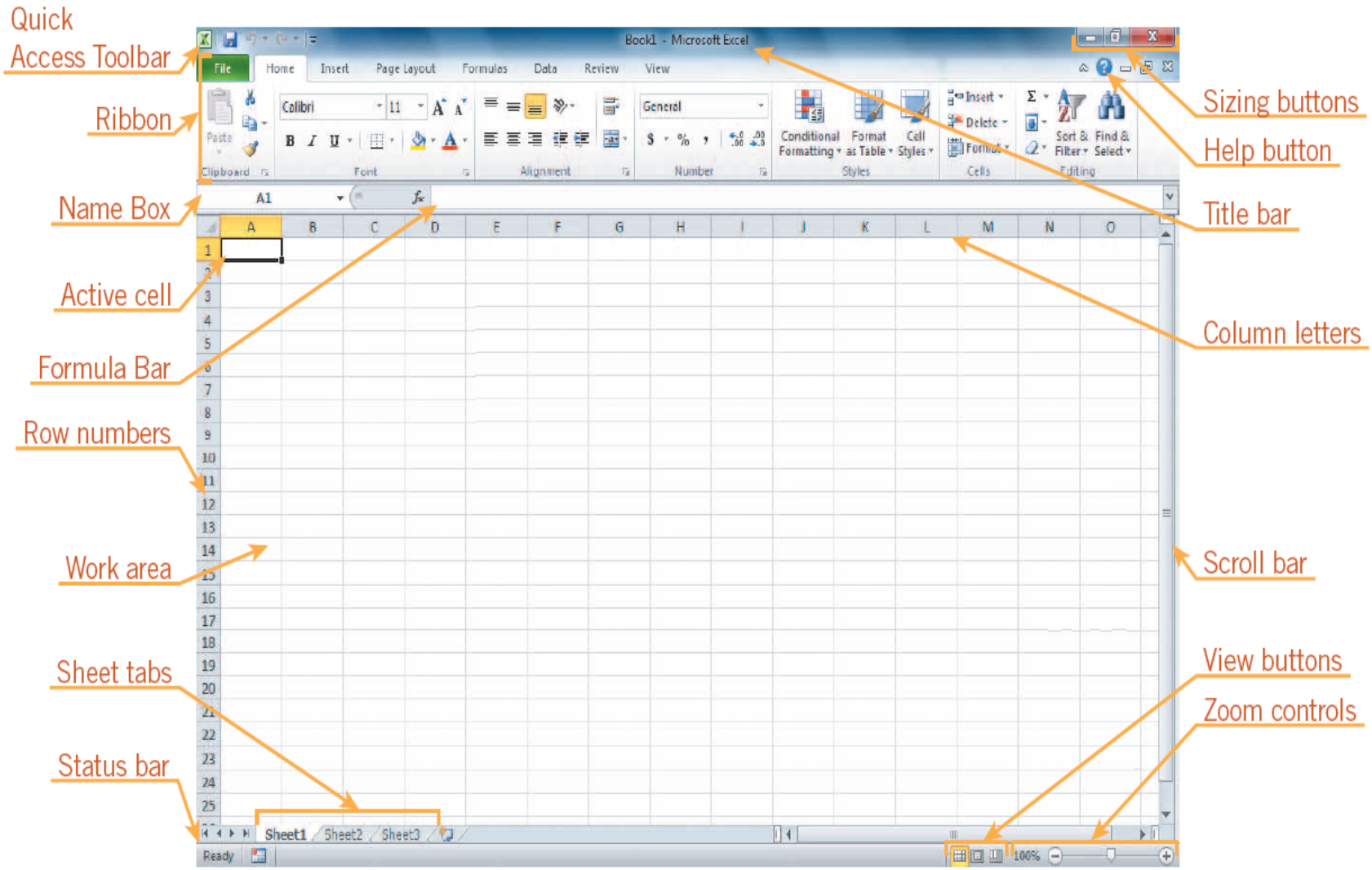
- Data Exploration
  - Simple Graphs
  - Combined Graphs
  - Link and brush
  - Non graphical techniques
- Data Modelling
  - Model and Variable selection
  - Model Execution
  - Model Diagnostic and model comparison

- Presentation and Automation
  - Presenting Data
  - Automating Data analysis

Data Handling using Spread sheets

- Formulas and Functions
- Filtering and Sorting
- Formulas for Locating and Pulling Values
- Inserting Charts
- Using Pivot Tables

Excel Program Window



Basic Operations

- Create a new workbook
- Open Workbook
- Save Workbook
- Close workbook
- Rename workbook
- Add a Worksheet
- Delete a worksheet
- Hide /Unhide a Worksheet
- Copy /Move a worksheet

Cell References

- Relative – when you copy and paste, the references change dynamically
- Absolute - do not change dynamically. (eg: A1) i.e, locking of row and column.

Condition function

- If()
  - Syntax: =IF(condition,True Statement, False Statement)
- And()
  - Syntax: =and(condition1,condition2,…) – results true or false
- Or()
  - Syntax: =and(condition1,condition2,…) – results true or false

Statistical functions

- sum(range)
- sumif(range, criteria, outputrange)
- counts(range)
- countif(range, criteria)
- average(range)
- averageif(range,criteria)
- min(range)
- max(range)
- large(range,which)
- small(range,which)

Data Cleaning functions

- **concatenate()**
  - useful to combine text from two or more cells into one cell.
  - Syntax: =Concatenate(Text1,Text2,…..Textn)
- **len() -** length of a cell
  - Syntax: =len(Text)
- **lower(), upper() and proper() -** change the text to lower, upper and sentence case
  - Syntax: =Upper(Text)/ Lower(Text) / Proper(Text)
- **trim():** used to clean text that has leading and trailing white space
  - Syntax: =Trim(Text)

- **Remove duplicate values**
  - Select data –> Go to Data ribbon –> Remove Duplicates
- **Text to Columns**
  - “Data” ribbon –> “Text to Columns”
  - Define delimiter

Sorting

- Sorting on one column
  - Select one cell in the column you want to sort.
  - On the Excel Ribbon, click the Data tab.
  - Click Sort A to Z (smallest to largest) or Sort Z to A (largest to smallest)
- Soring on Multiple columns
  - Select all the cells in the list.
  - On the Excel Ribbon, click the Data tab.
  - In the Sort & Filter group, click the Sort button.
  - Click the Add Level button, to add the first sorting level.
  - From the Sort by dropdown, select the first column you want to sort. In this example, Gender will be the first column sorted.

Filtering

- Go to the Data tab on Excel ribbon
- Select the Filter tool

Locating and Pulling values

- **Vlookup()**
  - It helps to search a value in a table and returns a corresponding value.
  - Syntax: =VLOOKUP(Key to lookup, Source_table, column of source table, are you ok with relative match?)

Charts

- A **chart**, or **graph**, is a visual representation of a set of data
- Select the data source with the range of data you want to chart
- In the Charts group on the Insert tab, click a chart type, and then click a chart subtype in the Chart gallery
- In the Location group on the Chart Tools Design tab, click the Move Chart button to place the chart in a chart sheet or embed it into a worksheet

Chart Type	Description
Column	Compares values from different categories. Values are indicated by the height of the columns.
Line	Compares values from different categories. Values are indicated by the height of the line. Often used to show trends and changes over time.
Pie	Compares relative values of different categories to the whole. Values are indicated by the areas of the pie slices.
Bar	Compares values from different categories. Values are indicated by the length of the bars.
Area	Compares values from different categories. Similar to the line chart except that areas under the lines contain a fill color.
XY (Scatter)	Shows the patterns or relationship between two or more sets of values. Often used in scientific studies and statistical analyses.
Stock	Displays stock market data, including the high, low, opening, and closing prices of a stock.
Surface	Compares three sets of values in a three-dimensional chart.
Doughnut	Compares relative values of different categories to the whole. Similar to the pie chart except that it can display multiple sets of data.
Bubble	Shows the patterns or relationship between two or more sets of values. Similar to the XY (Scatter) chart except the size of the data marker is determined by a third value.
Radar	Compares a collection of values from several different data sets.

Generating inference from Data

- **Pivot Table**
  - Pivot table is a summary table that lets you count, average, sum, and perform other calculations according to the reference feature you have selected
  - It converts a data table to inference table which helps us to take decisions

Steps to create pivot table

- Click somewhere in the list of data. Choose the *Insert* tab, and click *PivotTable*
- Now, you can see the PivotTable Field List panel, which contains the fields from your list; all you need to do is to arrange them in the boxes at the foot of the panel. Once you have done that, the diagram on the left becomes your PivotTable.

Pivot table fields

- Report filter. This section allows us to filter our table by one or more criteria. For example, we can only show data in our Pivot Table for the month of January.
- Column labels. This section allows us to summarize data across columns, placing data labels along the top of the screen.
- Row labels. This section allows us to summarize data across rows, placing data labels along the side of the screen.
- Values. This section allows us to specify what we're summarizing — for example, total sales or number of items ordered.

Reading Resources

What is Data Science ?

Introducing Data Science by Arno

Introducing Data Science - pdf

Excel - https://www.excel-easy.com/functions.html

https://www.excel-easy.com/data-analysis.html

End of Unit1

Dr. S. Thenmozhi