**DATA SCIENCE** – a field that combines math, statistics, programming and domain knowledge

**DATA SCIENCE LIFE CYCLE**

* **Collect Data**
* **Clean Data**
* **Analyze Data**
* **Visualize**
* **Apply Insights**

**DATA MINING** – focuses on discovering patterns in a large dataset.

**DATA ANALYTICS** – focuses on applying statistical technique to interpret data

1. TYPES:
   1. **Descriptive**
   2. **Diagnostic**
   3. **Predictive**
   4. **Prescriptive**

**TYPES OF DATA**

* **NOMINAL DATA –** used to label variables without any order or quanti value
* **ORDINAL DATA –** have a natural ordering, the arrangement is in order
* **DISCRETE DATA –** integer or whole numbers
* **CONTINUOUS DATA –** decimal/fraction

**DATA SOURCE** – refers to the origin of a specific set of information

**DATA SETS** – collection of raw data gathered during research process in the form of numerical data

**DATA FRAME** – a two-dimensional, tabular data structure commonly used in data analysis and manipulation

**MATRIX** – a two-dimensional data structure with rows and columns

**DASBOARDS** - a data visualization tool that can be customized for viewing the overall data at a glance.

* **STRATEGIC** – focuses on long term goals and strategies at the highest level of metrics.
* **OPERATIONAL** – short-term performance tracking and intermediate goals.
* **ANALYTICAL** – consists of the datasets and the mathematics used in these sets.

**EXPLORATORY DATA ANALYSIS(EDA)**

* A crucial step in data analysis to summarize the main characteristics of a datasets
* Helps uncover patterns, detect anomalies, and check assumptions using statistics and visualization
* The initial investigation of data.

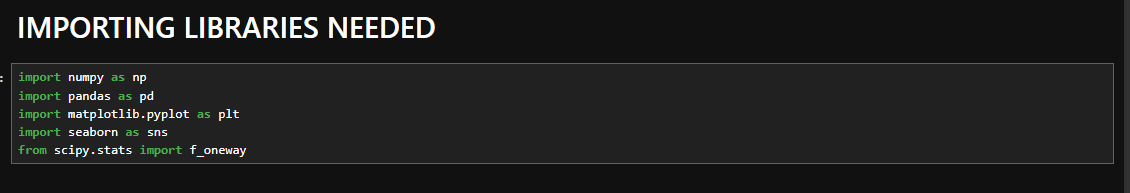
**DATA AGGREGATOIN** – the process of collecting, organizing, and summarizing data to create a simplified view

**CORRELATION** – a relationship between two variables (they move together)

1. **Types**
   1. **Positive Correlation** – both value increases/decreases together
   2. **Negative Correlation** – one increases while the other decreases
   3. **No Correlation** – one changes but the other one remain unchanged

**CAUSATION** – one event directly causes another

**DATA VISUALIZATION** – graphical representation of data



.describe() – return the description of the data (count, mean, std, min, max, 25%, 50%, 70%)

.info() – shows the index range and data type, shows the data’s size and structure

.dtypes – returns the data type

.astype() – changes the dtype of a column

.isnull() – returns the number of cells containing null values

.mode() – returns the frequent values (qualilative)

.mean() returns the average

.fillna() – fills the null values with the specified value

Pd.datetime – appropriate dtype for dates

.value\_counts – counts occurences of unique values

.count() – returns the size excempting rows containing null values

.size() – returns the size without excemption