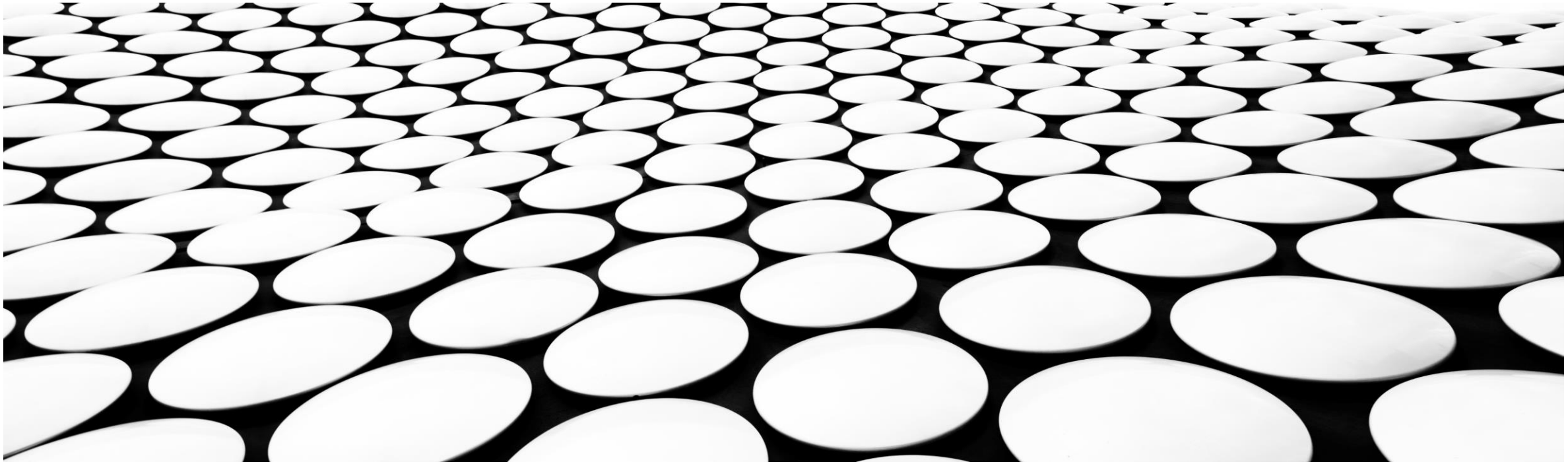
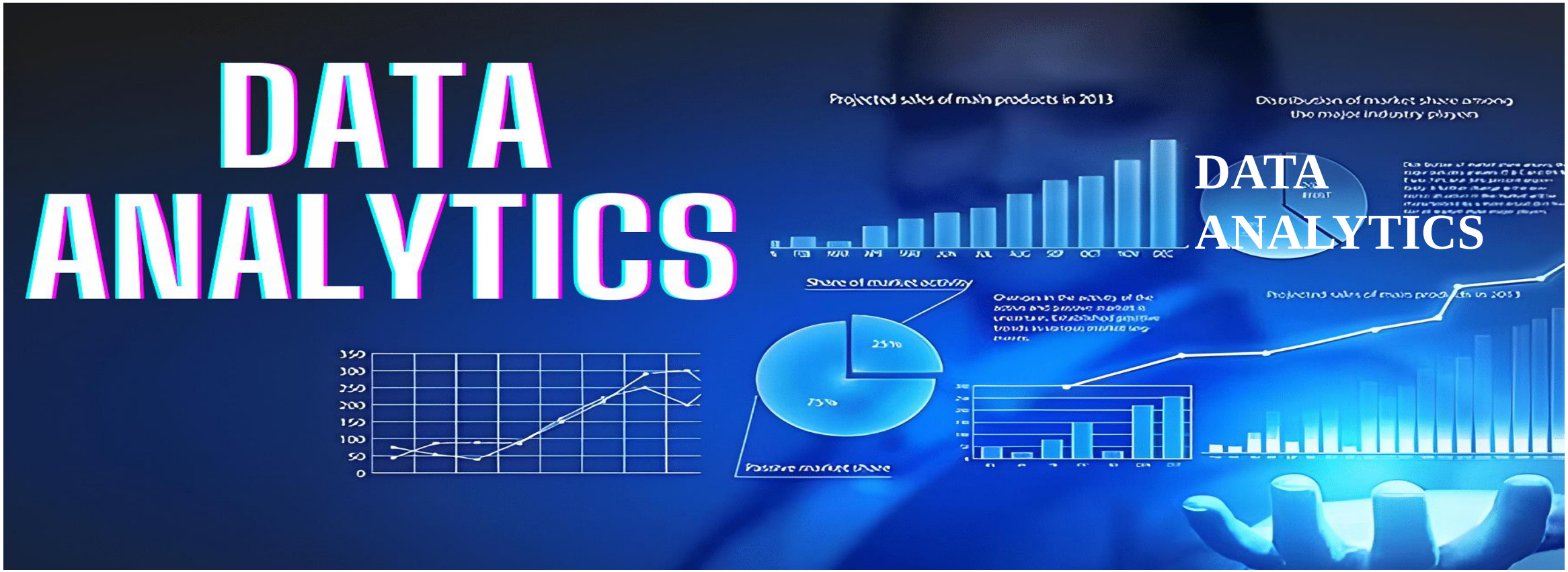
DATA ANALYTICS

SKILLS BUILD FOR COLLEGES



TITLE

- Definitions
- What is data analytics
- why data analytics is important to business?
- Data Analytics Tools
- Processes in data analytics
- Data collections
- ETL (Extract Transform LOAD)
- The main four types of data analytics
- Role of a data analyst
- Career opportunities


DATA
ANALYTICS
DATA
ANALYTICS

BASIC DEFINITIONS

- **Data :** Data is a set of values of qualitative or quantitative variables. It is information in raw or unorganized form. It may be a fact, figure, characters, symbols etc.
- **Information**: Meaningful or organized data is information.
- **Analytics :** Analytics is the discovery , interpretation, and communication of meaningful patterns or summery in data.
- **Data Analytics :**(DA) is the process of examining data sets in order to draw conclusion about the information it contains.
- Analytics is not a tool or technology, rather it is the way of thinking and acting on data.

WHAT IS DATA ANALYTICS?

Data analytics is the process of analyzing raw data in order to draw out meaningful, actionable insights, which are then used to inform and drive smart business decisions.

WHY DATA ANALYTICS IS IMPORTANT TO BUSINESS?

- Gain greater insight into target markets
- Enhance decision-making capabilities
- Create targeted strategies and marketing campaigns
- Improve operational inefficiencies and minimize risk
- Identify new product and service opportunities

DATA ANALYTICS TOOLS

- Python – This object-oriented open-source programming language is used for manipulating, visualizing, and modelling data.
- R – An open-source programming language used in numerical and statistical analysis.
- Tableau – This helps in creating several kinds of visualizations for presenting insights and trends in a better way.
- Power BI – This is a business intelligence tool that supports multiple data sources, helps in asking questions and getting immediate insights.
- SAS – This statistical analysis software helps in performing analytics, visualizing data, writing SQL queries, performing statistical analysis, and building ML models.

PROCESSES IN DATA ANALYTICS

The data analytics practice encompasses many separate processes, which can comprise a data pipeline:

- Collecting and ingesting the data
- Categorizing the data into structured/unstructured forms, which might also define next actions
- Managing the data, usually in databases, data lakes, and/or data warehouses
- Storing the data in hot, warm, or cold storage
- Performing ETL (extract, transform, load)
- Analyzing the data to extract patterns, trends, and insights
- Sharing the data to business users or consumers, often in a dashboard or via specific storage

PRIMARY DATA AND SECONDARY DATA

Primary data

Primary data collection involves the collection of original data directly from the source or through direct interaction with the respondents.

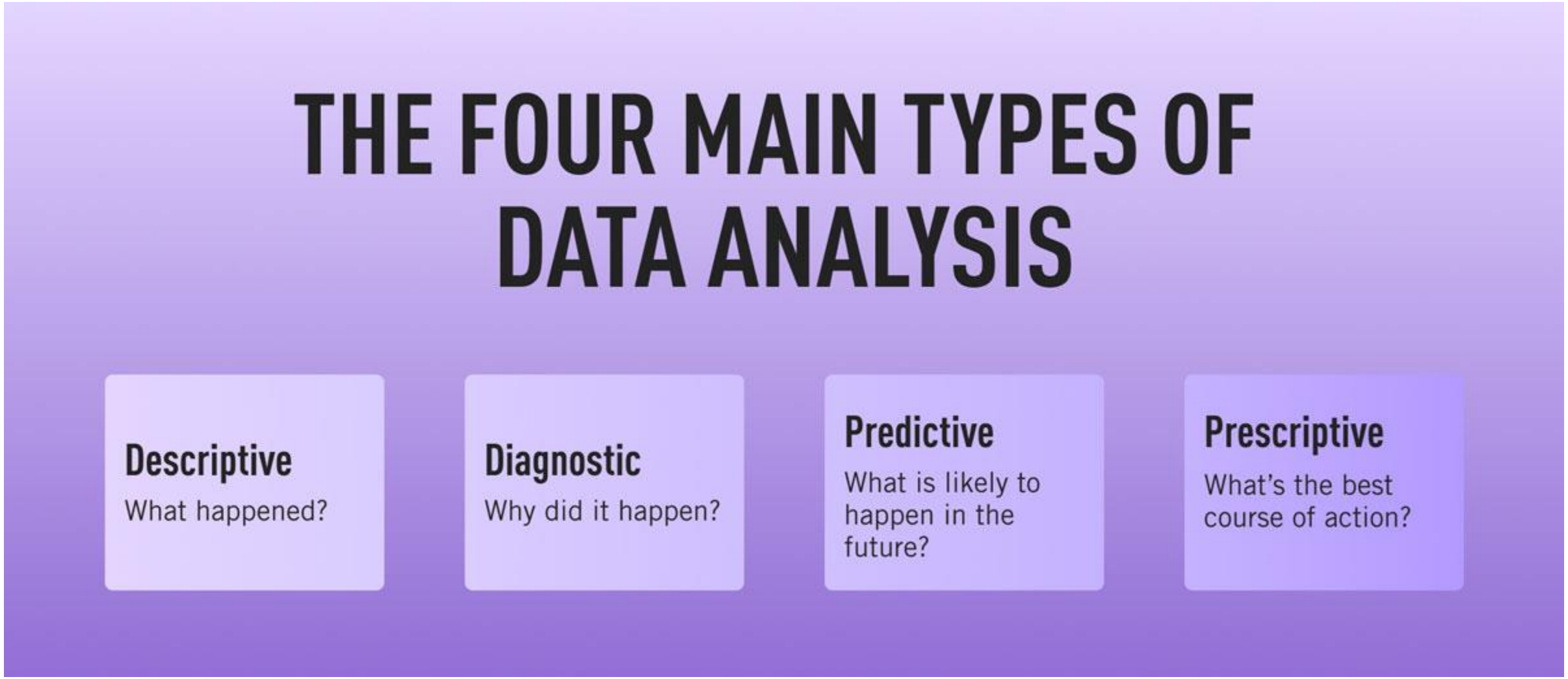
Secondary data

Secondary data collection involves using existing data collected by someone else for a purpose different from the original intent.

1. Primary Data Collection:
   - Surveys and Questionnaires
   - Interviews
   - Observations
   - Experiments
   - Focus Groups
2. Secondary Data Collection:
   - Published Sources
   - Online Databases
   - Government and Institutional Records
   - Publicly Available Data
   - Past Research Studies

ETL (EXTRACT TRANSFORM LOAD)

- **Extract: Retrieve data from various sources, such as databases, files, or APIs.**
- **Transform: Clean, filter, and manipulate data to ensure consistency and prepare it for analysis.**
- **Load: Store the transformed data into a target system or data warehouse for easy access and analysis.**


THE FOUR MAIN TYPES OF DATA ANALYSIS
Descriptive
What happened?
Diagnostic
Why did it happen?
Predictive
What is likely to happen in the future?
Prescriptive
What's the best course of action?

DIAGNOSTIC ANALYTICS

- Definition: Diagnostic analytics aims to determine the root causes and reasons behind certain events or trends observed in the data.
- Key Characteristics: Involves data exploration, drill-down analysis, and correlation identification. Diagnostic analytics answers the question of "why did it happen."
- Examples: Data mining techniques, regression analysis, cohort analysis.

DESCRIPTIVE ANALYTICS

- Definition: Descriptive analytics focuses on summarizing historical data to gain insights into past events and understand the current state.
- Key Characteristics: Involves data aggregation, visualization, and reporting. Descriptive analytics answers the questions of "what happened" and "what is happening."
- Examples: Bar charts, line graphs, dashboards displaying key performance indicators (KPIs).

PREDICTIVE ANALYTICS

- Definition: Predictive analytics leverages historical data to make predictions about future outcomes or events.
- Key Characteristics: Involves statistical modeling, machine learning algorithms, and pattern recognition. Predictive analytics answers the question of "what is likely to happen."
- Examples: Forecasting models, time series analysis, classification algorithms.

PRESCRIPTIVE ANALYTICS

- Definition: Prescriptive analytics recommends the best course of action based on predictive models, optimization techniques, and business rules.
- Key Characteristics: Involves simulation, optimization algorithms, and decision support systems. Prescriptive analytics answers the question of "what should be done."
- Examples: Optimization models, simulation tools, decision support systems.

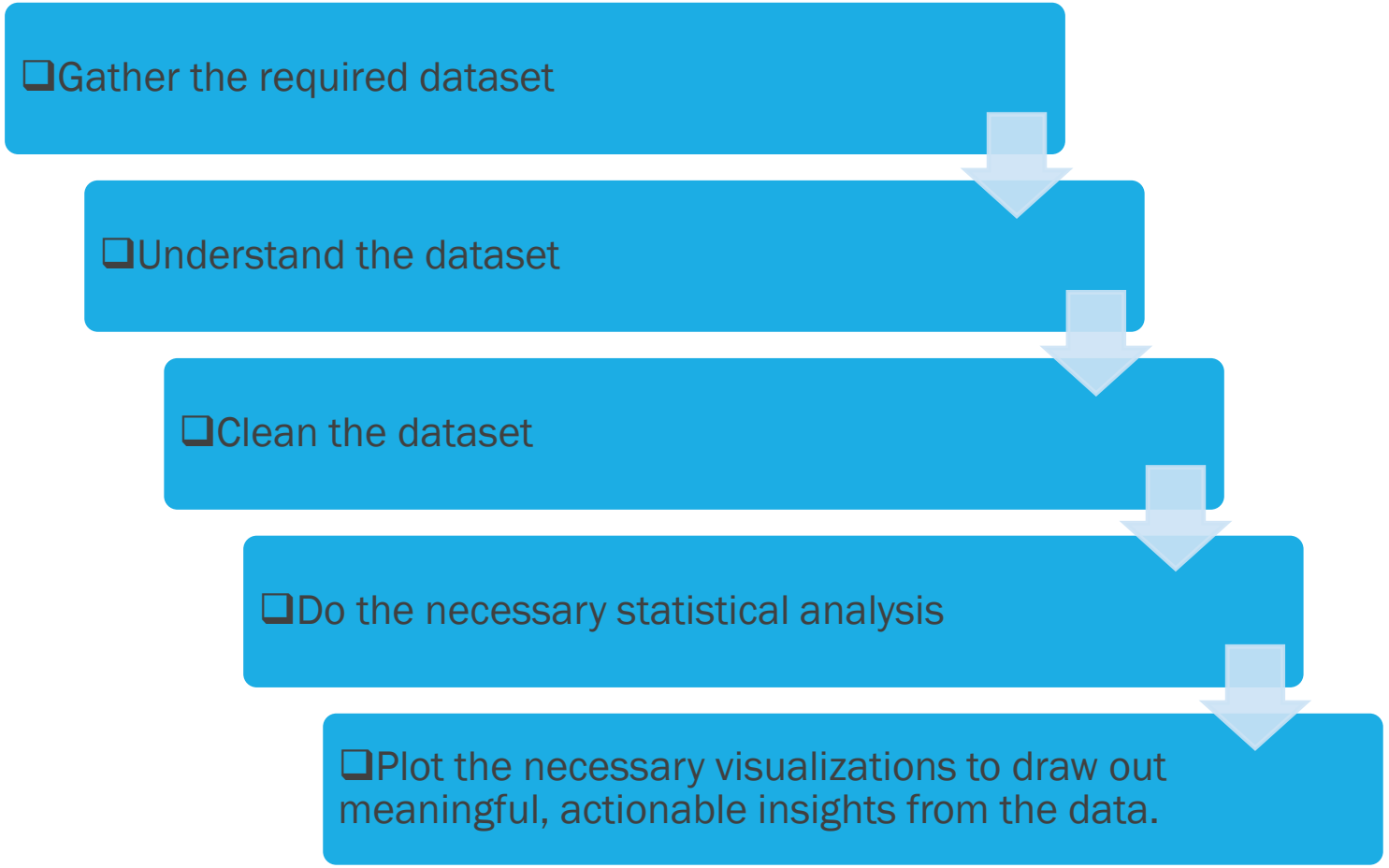
ROLE OF A DATA ANALYST

- A data analyst role is to answer specific questions or address particular challenges that have already been identified and are known to the business.
- To do this, they examine large datasets with the goal of identifying trends and patterns. They then “visualize” their findings in the form of charts, graphs, and dashboards.

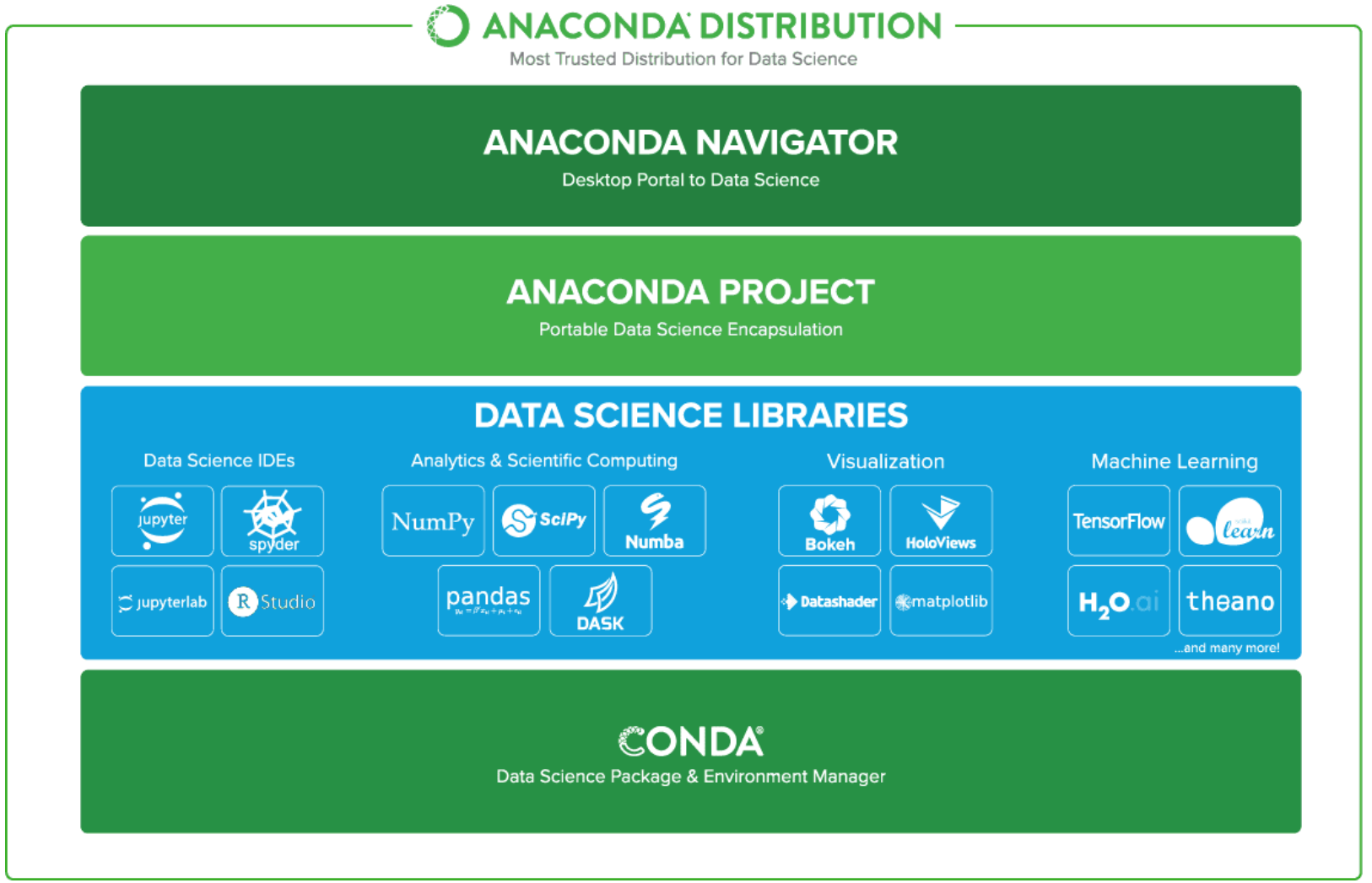
CAREER

- 1. Data Scientist
- 2. Business Intelligence Analyst
- 3. Data Engineer
- 4. Business Analyst
- 5. Marketing Analytics Manager
- 6. Financial Analyst
- 7. Quantitative Analyst
- 8. Risk Analyst
- 9. Data Governance Analyst
- 10. Data Visualization Engineer

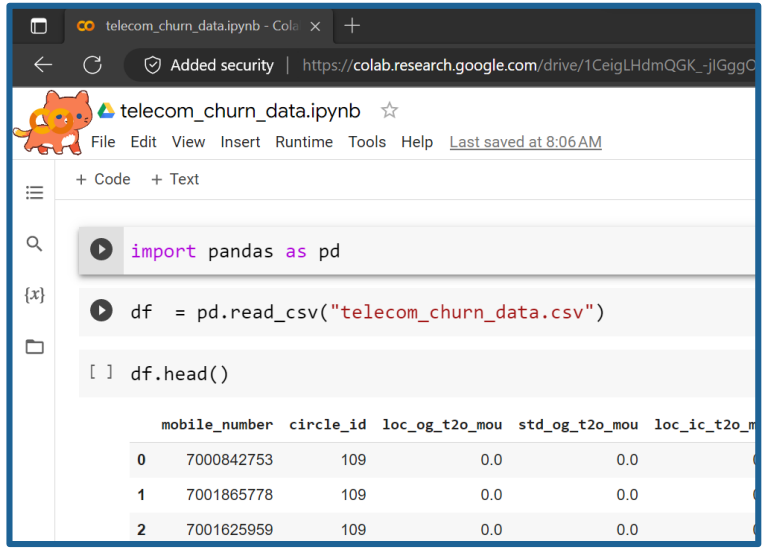
Steps involved in data analytics.



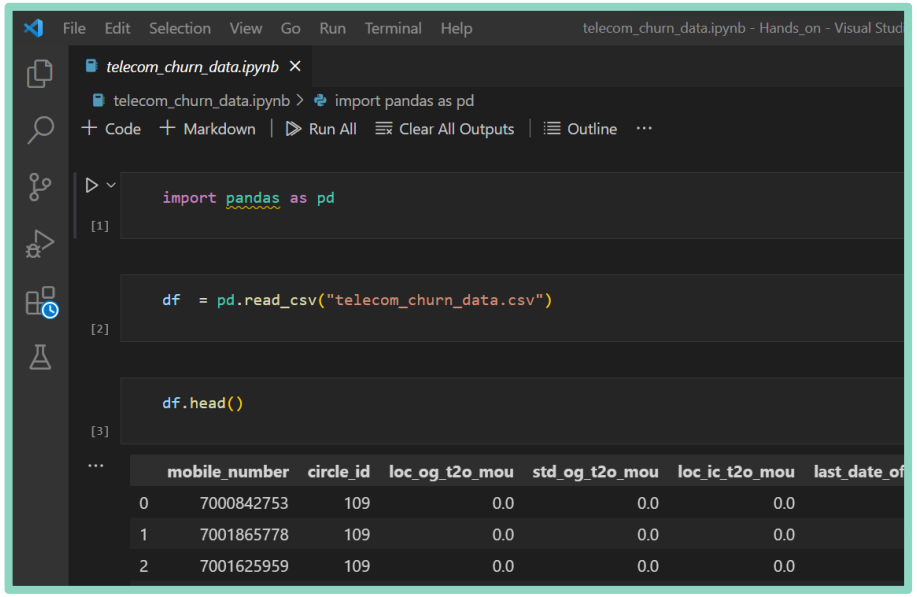
ABOUT ANACONDA NAVIGATOR



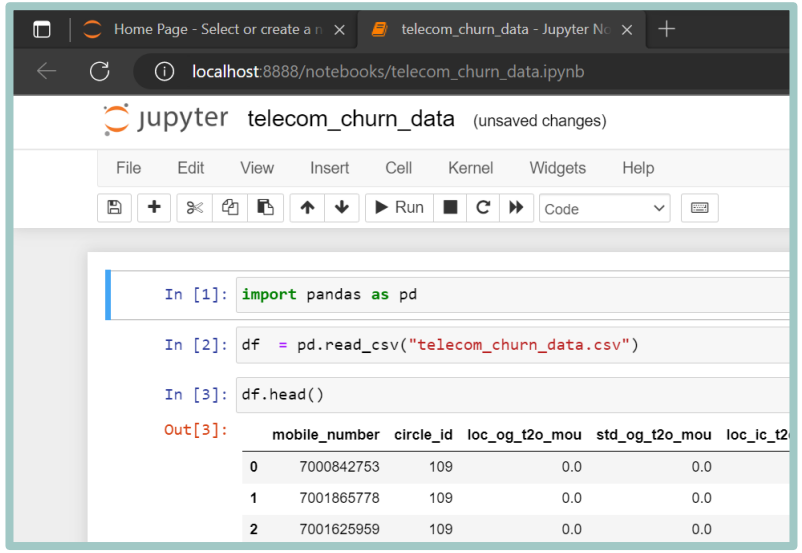
Platforms that we are going to use.



Google Colab

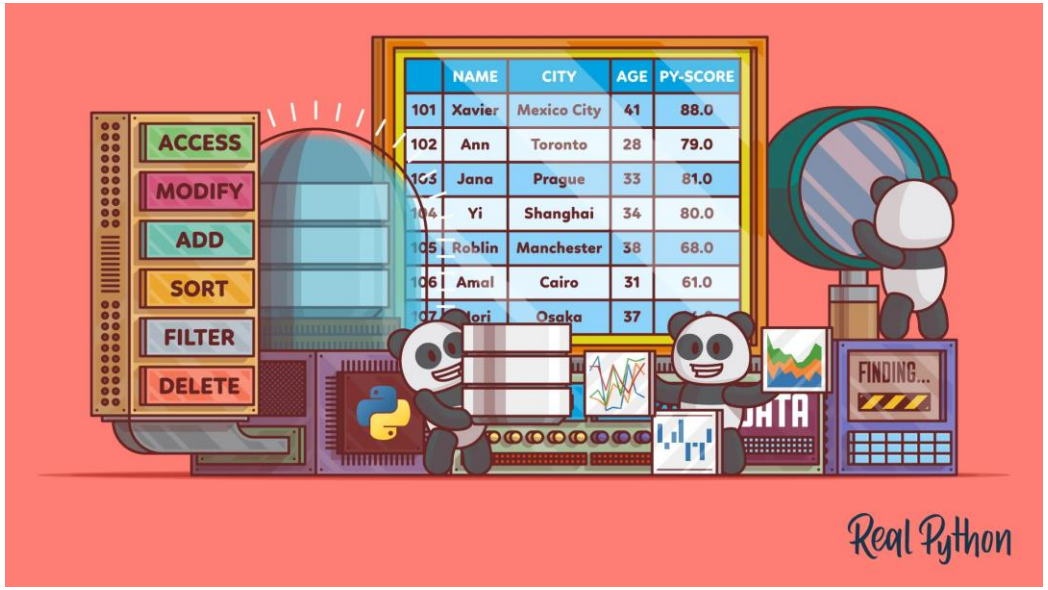


Visual studio code



Jupyter Notebook

INTRODUCTION TO PANDAS



- Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.
- Pandas is built on top of NumPy library.
- Pandas is well suited for many different of data

Image Source: https://realpython.com/pandas-dataframe/

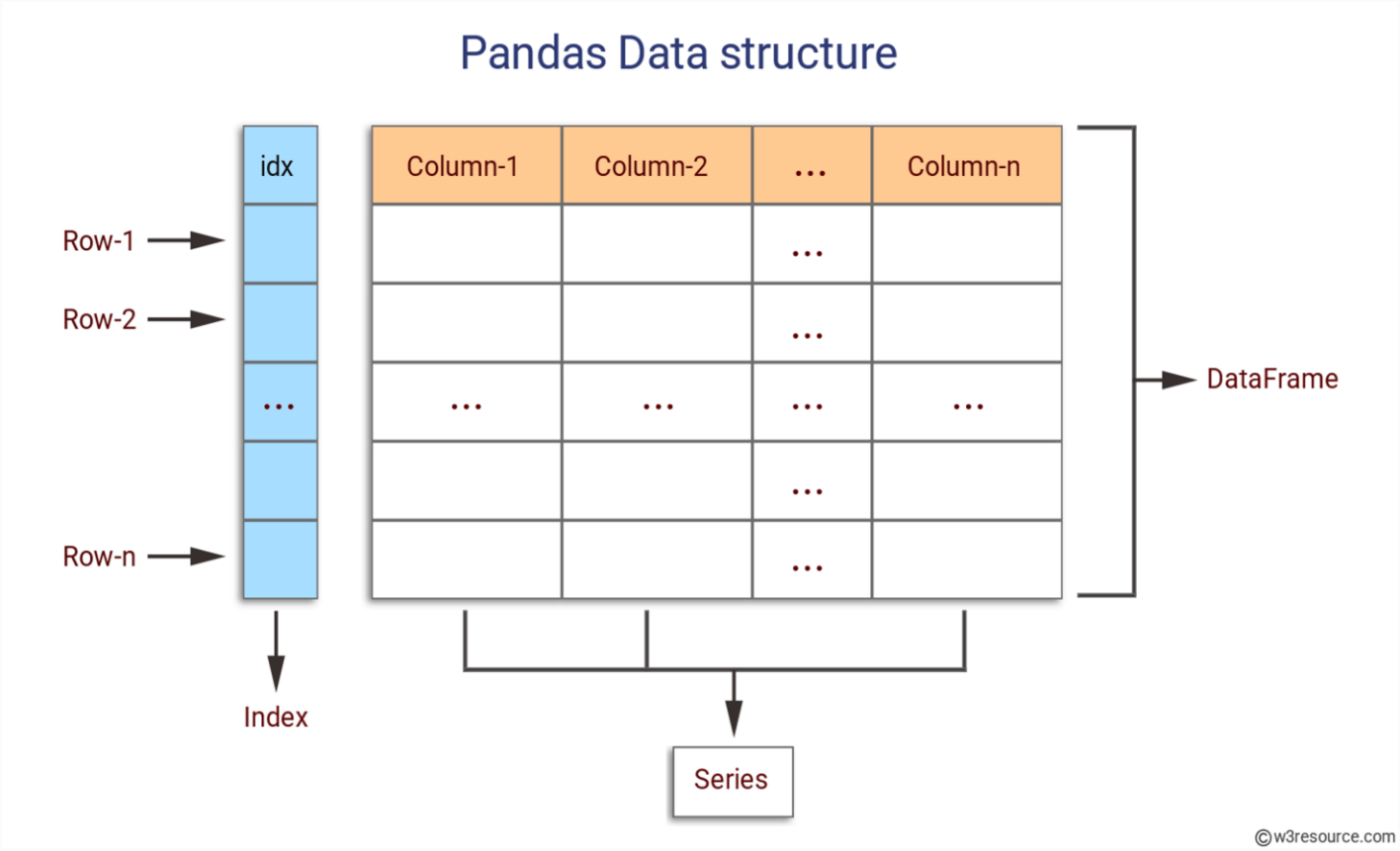
FEATURES OF PANDAS

Image Source: https://data-flair.training/blogs/python-pandas-features/

MOST USED FUNCTIONS IN PANDAS

read_csv()	head() /head(n)	describe()	memory_usage()	astype()
loc[:]	to_datetime()	value_counts()	drop_duplicates()	groupby()
	merge()	sort_values()	fillna()	

CORE COMPONENTS OF PANDAS : SERIES AND DATA FRAME



Series

	apples
0	3
1	2
2	0
3	1

+

Series

	oranges
0	0
1	3
2	7
3	2

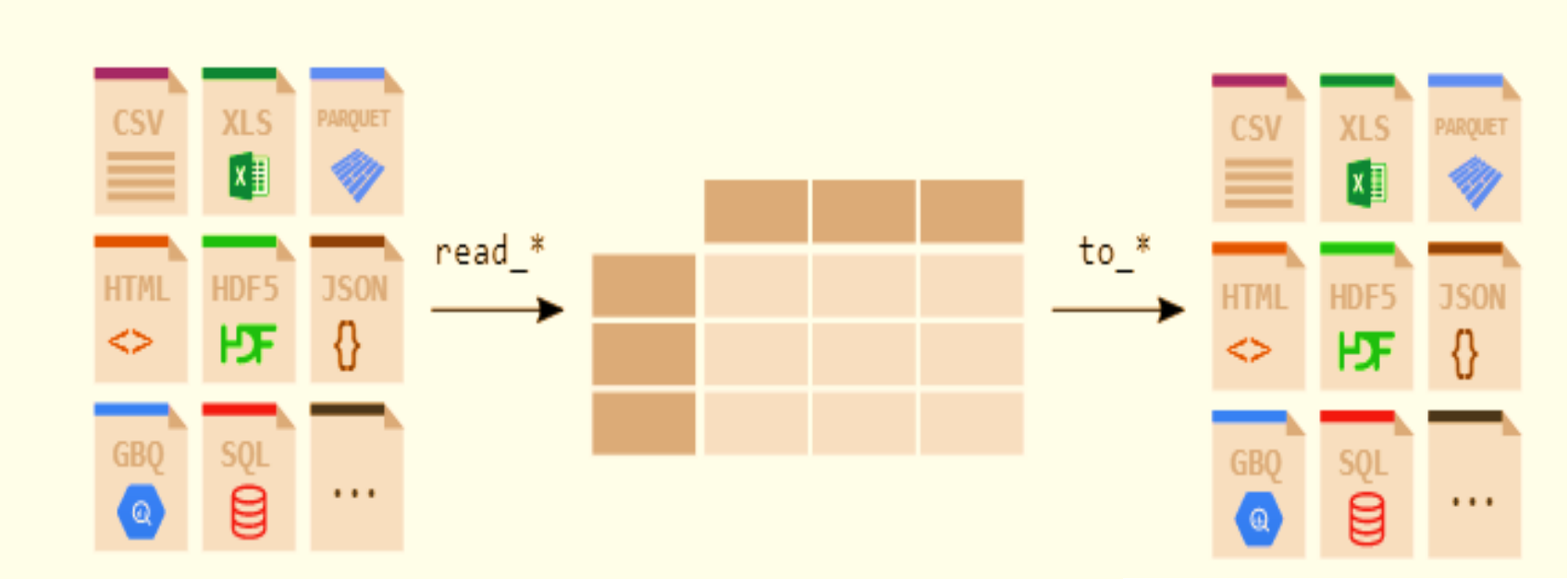
=

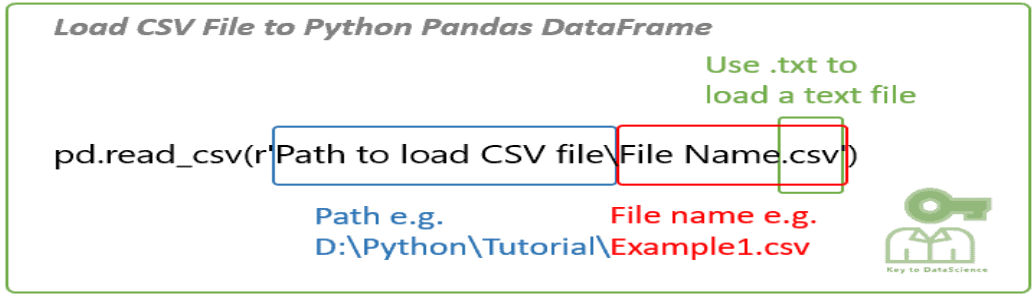
DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/

FILE HANDLING WITH PANDAS





Read csv Python Pandas

https://pandas.pydata.org/docs/getting_started/intro_tutorials/02_read_write.html

SAMPLE READING AND WRITING .CSV FILE

```
import pandas as pd
# Create a dataframe
raw_data = {'first_name':
['Sam','Ziva','Kia','Robin'],
        'degree':
['PhD','MBA','','MS'],
        'age': [25, 29, 19, 21]}
df = pd.DataFrame(raw_data)
df
#Save the dataframe
df.to_csv(r'Example1.csv')
```

```
import pandas as pd
# Read csv file
df = pd.read_csv(r'D:\Python\Tutorial\Example1.csv')
df
```

	first_name	degree	age
0	Sam	PhD	25
1	Ziva	MBA	29
2	Kia		19
3	Robin	MS	21

Output

EXPLORING A DATASET USING PANDAS

Download the dataset from: https://drive.google.com/file/d/1q7qK03njlzZRQ7PyYoprn12uVnE1gjH6/view

```
Import pandas as pd
data_1 = pd.read_csv(r'<datasetpath>')
data_1.head(6)
data_1.describe()
data_1.memory_usage(deep=True)
data_1['Gender'] = data_1.Gender.astype('category')
data_1.loc[0:4, ['Name', 'Age', 'State']]
data_1['DOB'] = pd.to_datetime(data_1['DOB'])
data_1['State'].value_counts()
data_1.drop_duplicates(inplace=True)
data_1.groupby(by='State').Salary.mean()
data_1.sort_values(by='Name', inplace=True)
data_1['City temp'].fillna(38.5, inplace=True)
```


Ref: https://www.analyticsvidhya.com/blog/2021/05/pandas-functions-13-most-important/

Convert list into series of elements

```
# convert element lists into series of elements, which have index from 0—5
import pandas as pd
my_data=[10,20,30,40,50]
pd.Series(data=my_data)
```

Convert dictionary into series of elements

```
import numpy as np
import pandas as pd
d={'a':10,'b':20,'c':30,'d':40}
#dictionary keys act as index and values with every key act as series values
pd.Series(d)
```

DATA MANIPULATION: DROP MISSING ELEMENTS

```
import pandas as pd
import numpy as np
d={'A':[1,2,np.NaN], 'B':[1,np.NaN,np.NaN],'C':[1,2,3]}
# np.NaN is the missing element in DataFrame
df=pd.DataFrame(d)   dictionary will get converted in to dataframe
df.dropna()   #pandas would drop any row with missing value
df.dropna(axis=1)   #drop column with NULL value
```

DATA MANIPULATION: FILLING SUITABLE VALUE

```
df.fillna(value='FILL VALUE')    #NaN is replaced by value=FILL VALUE

df['A'].fillna(value=df['A'].mean())

#Select column "A" and fill the missing value with mean value of the column A

df['A'].fillna(value=df['A'].std())

#Select column "A" and fill the missing value with standard deviation value of the column A
```

REPLACING A VALUE

```
import pandas as pd
df = pd.DataFrame({'one':[10,20,30,40,50,2000], 'two':[1000,0,30,40,50,60]})
print df.replace({1000:10,2000:60})
```

GROUPBY() FUNCTION

```
data = {'Company': [ 'CompA', 'CompA', 'CompB', 'CompB', 'CompC', 'CompC'],
        'Person': ['Rajesh', 'Pradeep', 'Amit', 'Rakesh', 'Suresh', 'Raj'],
        'Sales': [200, 120, 340, 124, 243, 350]}
df=pd.DataFrame(data)
df
comp=df.groupby("Company").mean()
comp
comp1=df.groupby("Company") #grouping done using label name "Company"
comp1.std()     #apply standard deviation on grouped data
```

FINDING MAXIMUM VALUE IN EACH LABEL

```
data = {'Company': [ 'CompA', 'CompA', 'CompB', 'CompB', 'CompC', 'CompC'],
        'Person': ['Rajesh', 'Pradeep', 'Amit', 'Rakesh', 'Suresh', 'Raj'],
        'Sales': [200, 120, 340, 124, 243, 350]}
df=pd.DataFrame(data)
df
df.groupby("Company").max()
```

FINDING UNIQUE VALUE & NUMBER OF OCCURRENCE FROM DATAFRAME

```
df = pd.DataFrame({'col1':[1,2,3,4],'col2':[444,555,666,444],'col3':['abc','def','ghi','xyz']})
# col1, col2 & col3 are column labels, each column have their own values
df['col2'].unique()           #fetches the unique values available in column
df['col2'].value_counts()  # count number of occurrence of every value
```

STATISTICAL FUNCTIONS

```
import numpy as np
import pandas as pd
s = pd.Series([1,2,3,4,5,4])
print s.pct_change()
df = pd.DataFrame(np.random.randn(5, 2))
print df.pct_change()
s1 = pd.Series(np.random.randn(10))
s2 = pd.Series(np.random.randn(10))
print s1.cov(s2) import numpy as np
frame = pd.DataFrame(np.random.randn(10, 5), columns=['a', 'b', 'c', 'd', 'e'])
print frame['a'].corr(frame['b'])
print frame.corr()
s = pd.Series(np.random.randn(5), index=list('abcde'))
s['d'] = s['b'] # so there's a tie
print s
print s.rank()
```


THANK
YOU