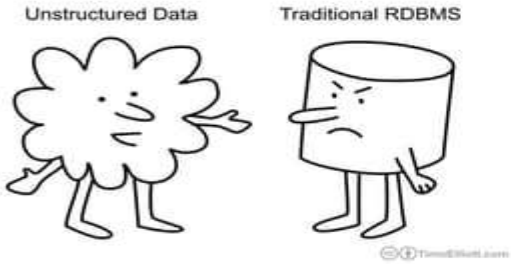
Introduction to Basic Data Analytics Tools

What is Data Analytics?

Data analytics is the science of analyzing raw data in order to make conclusions about that information.

Data Analytics Pipeline

Data Acquisition



"I'm sorry, I'm just not that into you..."

How To Collect Data!

- ★ REST API
- ★ From end users
- ★ Web scrape
- ★ Email and cloud storage
- ★ Client's server

REST : Representational State Transfer POST | PUT |Get

Tool for Data Acquisition:

Requests

A library for making HTTP requests in Python.

Key Features:

- Keep-alive & Connection Pooling
- Sessions with Cookie Persistence

```python
import requests

url = 'https://jsonplaceholder.typicode.com/posts/1'

response = requests.get(url=url)

print(response.text)

payload = {
    "userID": "1",
    "title": "thisTitle",
    "body": "thisBody"
}

url = 'https://jsonplaceholder.typicode.com/posts'

response = requests.post(url=url, data=payload)

print(response.status_code)
print(response.text)
```

Tool for Data Acquisition:

BeautifulSoup

A library for parsing HTML and XML documents.

Key Features:

- Multiple parser support (e.g., lxml, html5lib, and others)
- Creates parse tree which is easy to navigate

Flask, Flask-RESTPlus and Swagger UI

Flask is a micro web framework written in Python.

Flask-RESTPlus is an extension for Flask that adds support for quickly building REST APIs. It automatically documents the APIs which is visible in Swagger UI.

Data Pre-Processing and Storage

"What's a data lake for?
So you can drown in more data even faster!"

How To Clean Data!

- ★ Remove duplicate
- ★ Validate
- ★ Handle missing data
- ★ Fix errors
- ★ Filter outliers

How To Store Data!

- ★ RDBMS
- ★ ORM

Tool for Data Manipulation:

Pandas

A library for data manipulation and analysis.

Key Features:

- Loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.

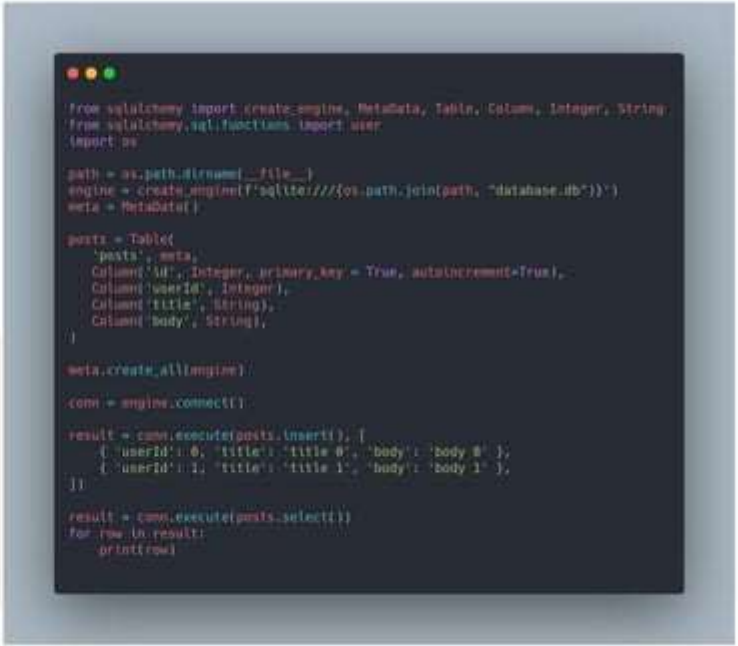
Tool for Database Operations:

SQLAlchemy

SQLAlchemy is a popular SQL toolkit and Object Relational Mapper.

Key Features:

- Function-based query construction.
- Multiple database support (e.g., SQLite, Postgresql, MySQL, Oracle, MS-SQL, Firebird, Sybase and others).



```
from sqlalchemy import create_engine, MetaData, Table, Column, Integer, String
from sqlalchemy.sql.functions import user
import os

path = os.path.dirname(__file__)
engine = create_engine(f'sqlite:///{os.path.join(path, "database.db")}')
meta = MetaData()

posts = Table(
    'posts', meta,
    Column('id', Integer, primary_key = True, autoincrement=True),
    Column('userId', Integer),
    Column('title', String),
    Column('body', String),
)

meta.create_all(engine)

conn = engine.connect()

result = conn.execute(posts.insert(), [
    { 'userId': 0, 'title': 'title 0', 'body': 'body 0' },
    { 'userId': 1, 'title': 'title 1', 'body': 'body 1' },
])

result = conn.execute(posts.select())
for row in result:
    print(row)
```

Data Analysis

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

How To Analyze Data!

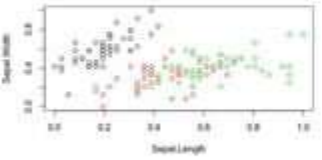
- ★ Five number summary (maximum, minimum, median, 1st quartile, 3rd quartile)
- ★ Average
- ★ Standard Deviation
- ★ Ratio
- ★ Interval
- ★ Trends
- ★ Aggregate and group by
- ★ Regression
- ★ Clustering

Tools for Data Analysis:

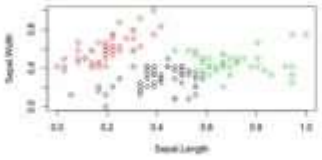
R and RStudio

R is a popular programming language for data analysis. RStudio is an IDE for R.

Original Classes



Clusters by k-means



```r
require("datasets")
data("iris")
head(iris)
summary(iris)

X <- iris[,c(1, 2)]
y <- iris[,"Species"]

normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

X$Sepal.Length <- normalize(X$Sepal.Length)
X$Sepal.Width <- normalize(X$Sepal.Width)
head(X)

result <- kmeans(X, 3)

result$size
result$centers

plot(X, col=result$cluster)
plot(X, col=y)
```

Data Presentation

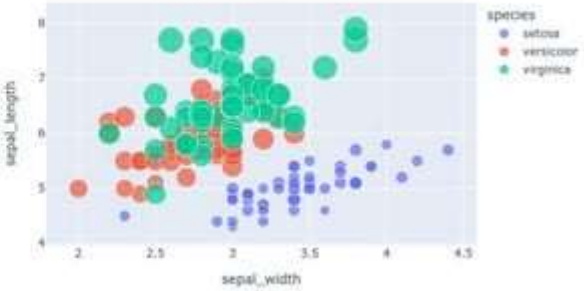
How To Visualize Data!

"Fake news!"

- ★ Charts
  - Line
  - Bar
  - Pie
  - Scatter
- ★ Graphs
- ★ Maps
  - Bubble
  - Polygon
- ★ Dashboards

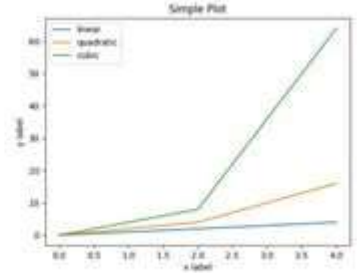
Plotly

An interactive graphing library.



Matplotlib

A plotting library for Python.



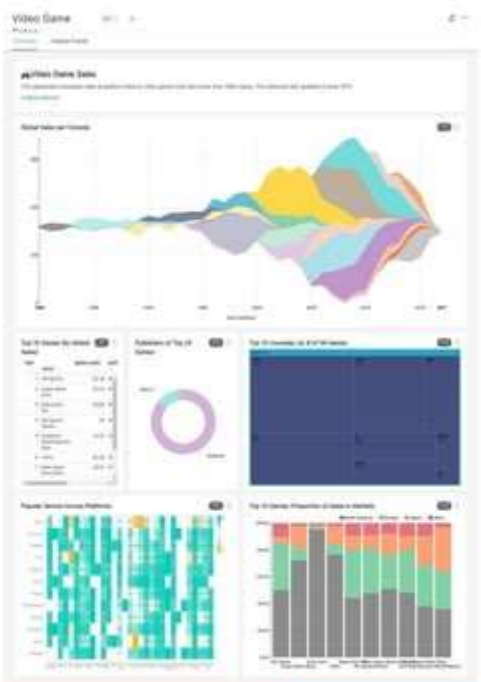
Tool for Data Visualization:

Apache Superset

A Data Visualization and Data Exploration Platform.

Key Features:

- It supports all the data sources that support SQL Alchemy and supports querying using SQL.
- Superset allows sharing dashboards.
- It comes with security features like Authentication, User Management and Roles.



Other Notable Tools

- ★ Excel
- ★ Tableau Public
- ★ Grafana
- ★ Microsoft Power BI
- ★ And many more . . .

Challenges

1. Poor quality data
2. Data privacy and security
3. Weak infrastructure
4. Data from multiple sources
5. Scaling data analysis

"Sir, some citizens have been complaining about our data privacy policies...

Do you want their names? Social security numbers? Most embarrassing secrets?"

Links

Flask: https://flask.palletsprojects.com/en/2.0.x/

Flask-RESTPlus: https://flask-restplus.readthedocs.io/en/stable/

SQLAlchemy: https://www.sqlalchemy.org/

R Programming Language: https://www.r-project.org/

k-means Clustering: https://en.wikipedia.org/wiki/K-means_clustering

Plotly: https://plotly.com/

Superset Docs: https://superset.apache.org/docs/intro

Presentation GitHub Link: https://github.com/saadrumon/basic-data-analytics-tools-presentation.git

Thank You

Any Questions?

"Information is the oil of the 21st century, and analytics is the combustion engine."

- Peter Sondergaard