Introduction to data Analytics

By Dr Shaik A Qadeer

Professor MJCET

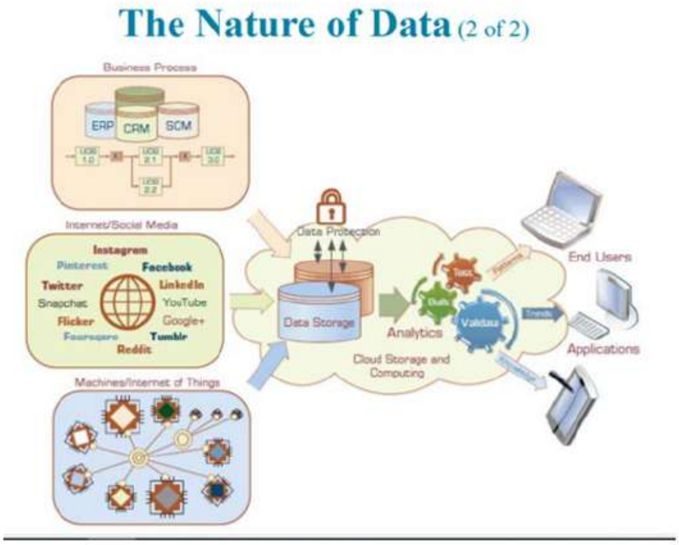
Content

- Intro to data analytics
- Data analytics life cycle
- Discovery
- Data preparation
- Model planning
- Model building implementation
- Communicate Result(Documentation)
- Operationalize(Quality Assurance)

Introduction to data analytics: Data and their nature

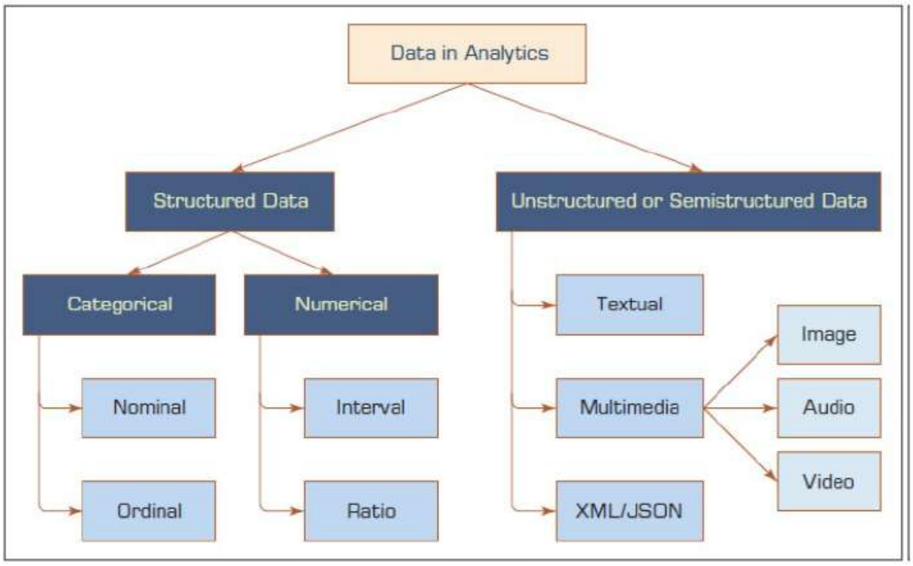
- Data: a collection of facts(usually obtained as the result of **experiences**, **observations**, or **experiments**)
- Data may consist of **numbers, words, images**
- Data is the lowest level of abstraction in analytics
- Data is the source for information and knowledge
- Data quality and data integrity → critical to analytics

Introduction to data analytics: Data and their nature..



The Nature of Data (2 of 2)

Introduction to data analytics: Data Taxonomy

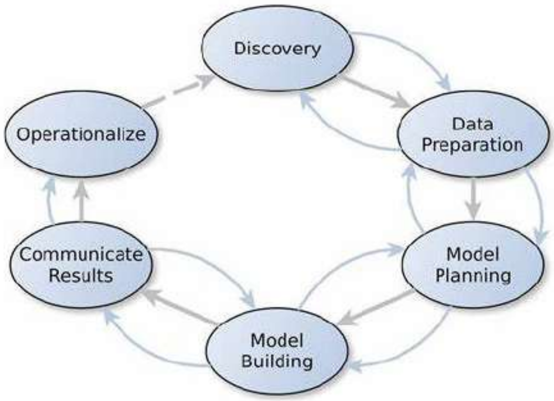


Introduction to data analytics

- **Data analytics** is the process of collecting, transforming, and organizing data to make informed decisions. It involves analyzing raw data to draw conclusions and predictions, ultimately driving better decision-making.
- Data analytics=Data discovery+Data analysis(extracts meaning from data)+data science (using data to theorise and forecast) + data engineering (building data systems).

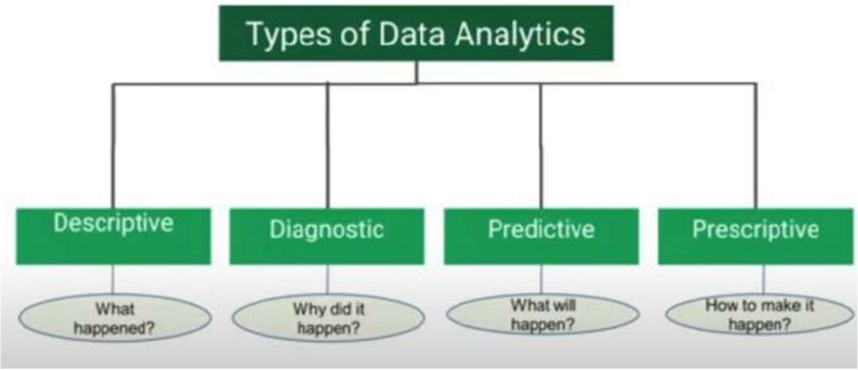
Data analytics life cycle

- Figure shows the life cycle diagram: Data Discovery to model deployment



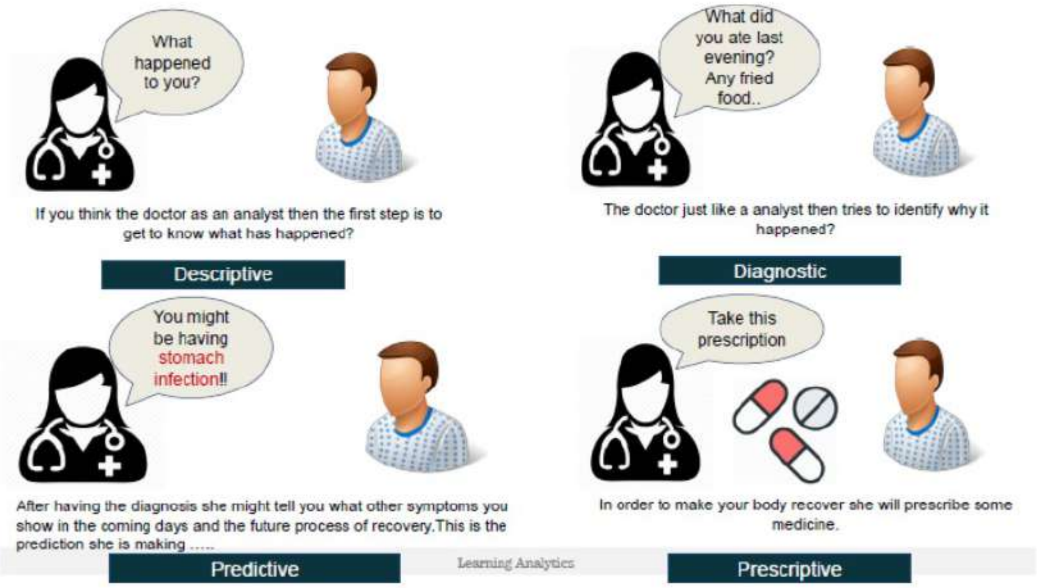
Data analytics life cycle

It can mapped to this 4 types of data analytics:
Descriptive->Data discovery,
Diagnostic->data preparation,
Predictive->model planning +
model development
Prescriptive ->Communicate result
operationalize



Types of Data Analytics

Data analytics life cycle: A case study



What happened to you?
What did you ate last evening? Any fried food..
You might be having stomach infection!!
Take this prescription

If you think the doctor as an analyst then the first step is to get to know what has happened?

Descriptive

The doctor just like a analyst then tries to identify why it happened?

Diagnostic

After having the diagnosis she might tell you what other symptoms you show in the coming days and the future process of recovery.This is the prediction she is making .....

Predictive

In order to make your body recover she will prescribe some medicine.

Prescriptive

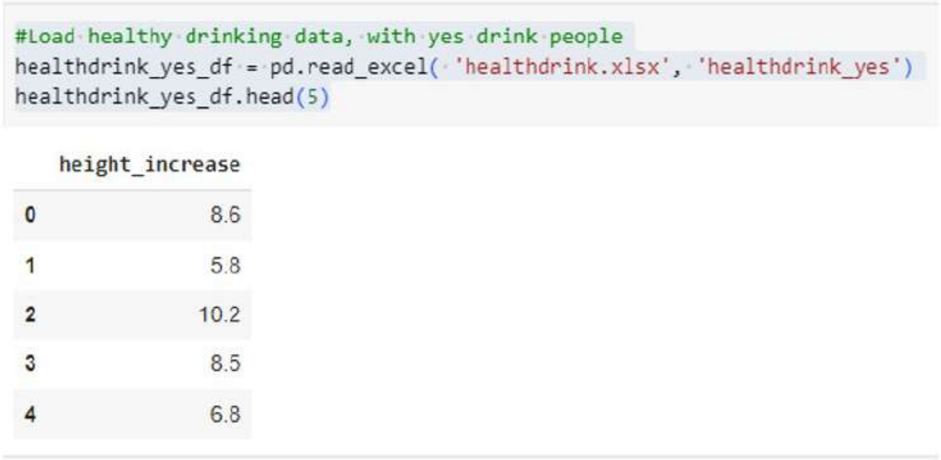
Learning Analytics

Data discovery(Descriptive analytics is used)

- Purpose: Understand Business Objectives and Data Requirements
- Tasks:
  - Identify Goals and Objectives
  - Explore Available Data Sources
  - Conduct Exploratory Data Analysis (EDA)
- Tools: Data Visualization, Descriptive Statistics(central tendency and measure of dispersion)
- Example: Exploring healthy drink data to Identify Trends and Patterns

Data discovery: Example: Exploring healthy drink data to Identify Trends and Patterns

- Loading data from excel file into python data frame with yes option

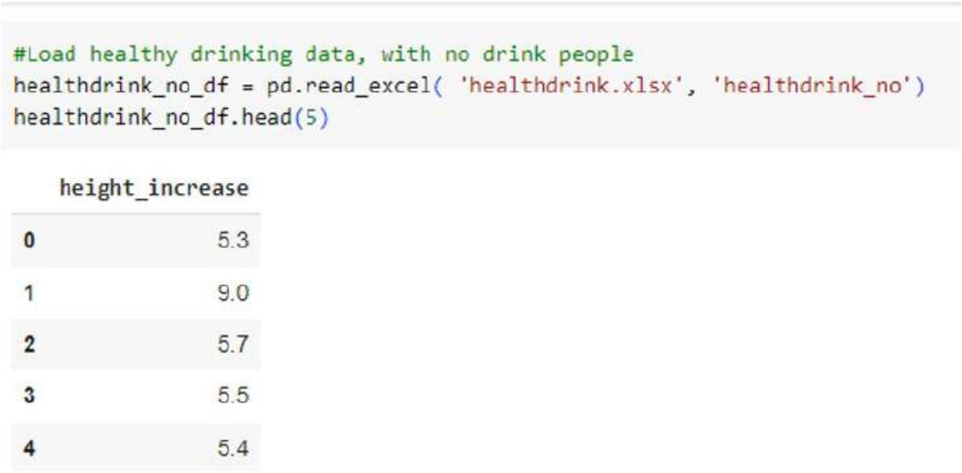


```
#Load healthy drinking data, with yes drink people
healthdrink_yes_df = pd.read_excel( 'healthdrink.xlsx', 'healthdrink_yes')
healthdrink_yes_df.head(5)
```

```
   height_increase
0              8.6
1              5.8
2             10.2
3              8.5
4              6.8
```

Data discovery: Example: Exploring healthy drink data to Identify Trends and Patterns

- Loading data from excel file into python data frame with no option

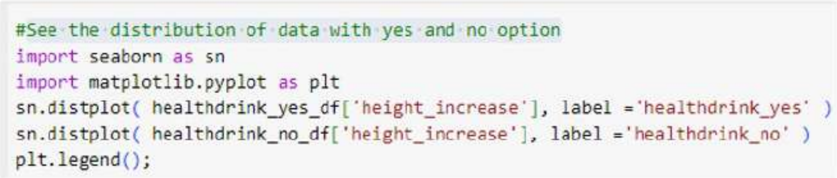


```
#Load healthy drinking data, with no drink people
healthdrink_no_df = pd.read_excel( 'healthdrink.xlsx', 'healthdrink_no')
healthdrink_no_df.head(5)
```

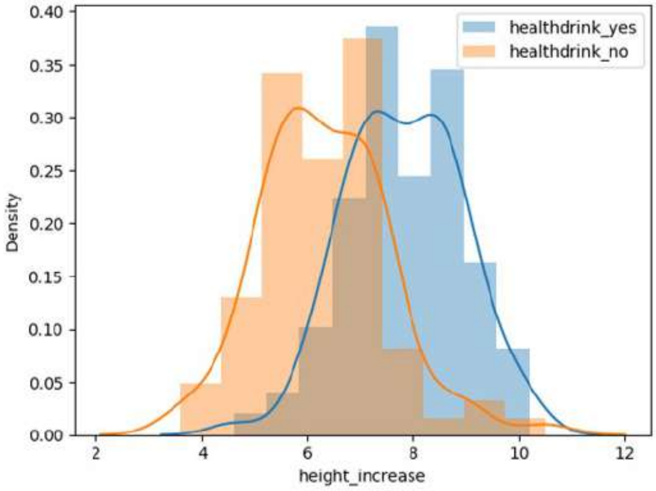
	height_increase
0	5.3
1	9.0
2	5.7
3	5.5
4	5.4

Data discovery: Example: Exploring healthy drink data to Identify Trends and Patterns

- The normal or Gaussian distribution of data with yes and no option



```
#See the distribution of data with yes and no option
import seaborn as sn
import matplotlib.pyplot as plt
sn.distplot( healthdrink_yes_df['height_increase'], label ='healthdrink_yes' )
sn.distplot( healthdrink_no_df['height_increase'], label ='healthdrink_no' )
plt.legend();
```

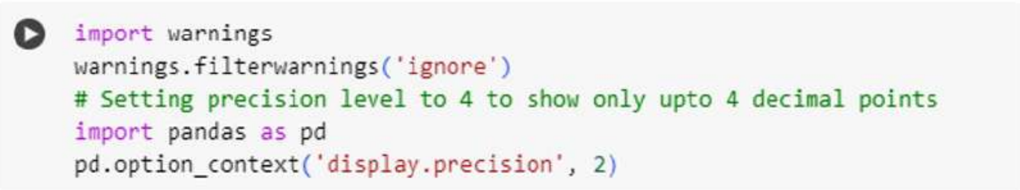


Data Preparation(Diagnostics analytic is used for this)

- Purpose: Clean and Prepare Data for Analysis
- Tasks:
    - ETL(Extract transform and load)
    - Data Cleaning and Preprocessing
    - Feature Engineering
- Importance of Data Quality and Consistency
- Tools: correlation, covariance, Inferential statistics
- Example: Cleaning and Transforming Raw Stock Data for investment analysis

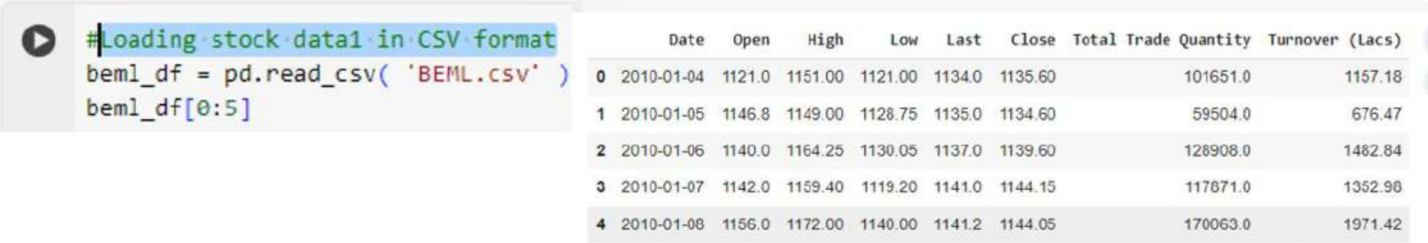
Data Preparation:
Example: ETL and data cleaning

- Importing libraries



```
import warnings
warnings.filterwarnings('ignore')
# Setting precision level to 4 to show only upto 4 decimal points
import pandas as pd
pd.option_context('display.precision', 2)
```

- Loading stock data1 in CSV format

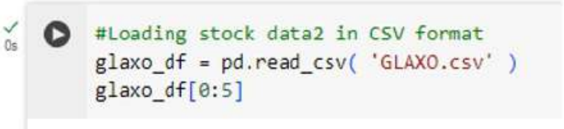


```
#Loading stock data1 in CSV format
beml_df = pd.read_csv( 'BEML.csv' )
beml_df[0:5]
```

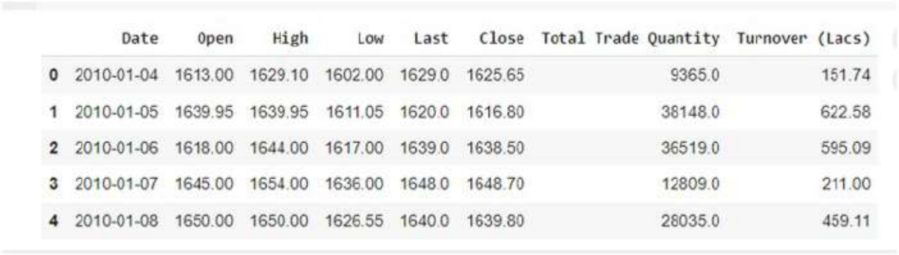
	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
0	2010-01-04	1121.0	1151.00	1121.00	1134.0	1135.60	101651.0	1157.18
1	2010-01-05	1146.8	1149.00	1128.75	1135.0	1134.60	59504.0	676.47
2	2010-01-06	1140.0	1164.25	1130.05	1137.0	1139.60	128908.0	1482.84
3	2010-01-07	1142.0	1159.40	1119.20	1141.0	1144.15	117871.0	1352.98
4	2010-01-08	1156.0	1172.00	1140.00	1141.2	1144.05	170063.0	1971.42

Data Preparation:
Example: ETL and data cleaning

- Loading stock data2 in CSV format



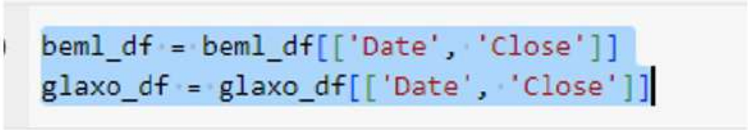
```
#Loading stock data2 in CSV format
glaxo_df = pd.read_csv( 'GLAXO.csv' )
glaxo_df[0:5]
```



	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
0	2010-01-04	1613.00	1629.10	1602.00	1629.0	1625.65	9365.0	151.74
1	2010-01-05	1639.95	1639.95	1611.05	1620.0	1616.80	38148.0	622.58
2	2010-01-06	1618.00	1644.00	1617.00	1639.0	1638.50	36519.0	595.09
3	2010-01-07	1645.00	1654.00	1636.00	1648.0	1648.70	12809.0	211.00
4	2010-01-08	1650.00	1650.00	1626.55	1640.0	1639.80	28035.0	459.11

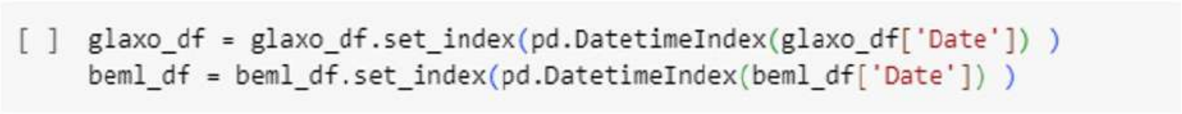
Data Preparation:
Example: ETL and data cleaning

- Selecting one two feature vector



```
beml_df = beml_df[['Date', 'Close']]
glaxo_df = glaxo_df[['Date', 'Close']]
```

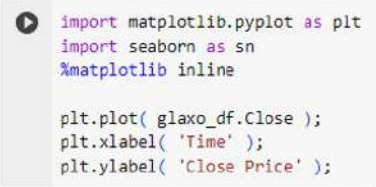
- Converting time to index(which is needed in ETL operation)



```
[ ]  glaxo_df = glaxo_df.set_index(pd.DatetimeIndex(glaxo_df['Date']) )
     beml_df = beml_df.set_index(pd.DatetimeIndex(beml_df['Date']) )
```

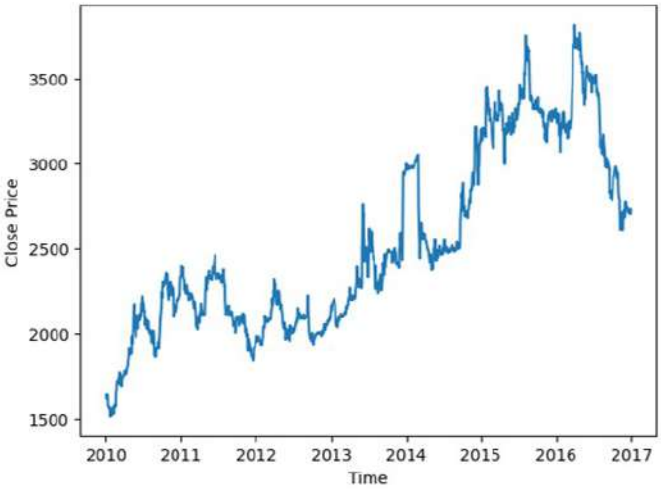
Data Preparation:
Example: ETL and data cleaning

- Time plot to see the relationship between time and close



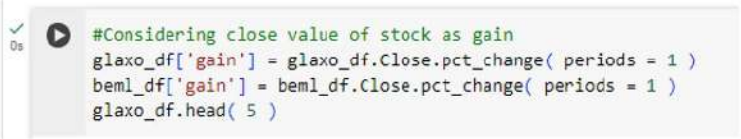
```
import matplotlib.pyplot as plt
import seaborn as sn
%matplotlib inline

plt.plot( glaxo_df.Close );
plt.xlabel( 'Time' );
plt.ylabel( 'Close Price' );
```



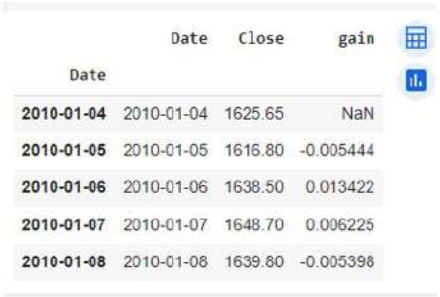
Data Preparation:
Example: data cleaning

- Considering close value of stock as gain



```
#Considering close value of stock as gain
glaxo_df['gain'] = glaxo_df.Close.pct_change( periods = 1 )
beml_df['gain'] = beml_df.Close.pct_change( periods = 1 )
glaxo_df.head( 5 )
```

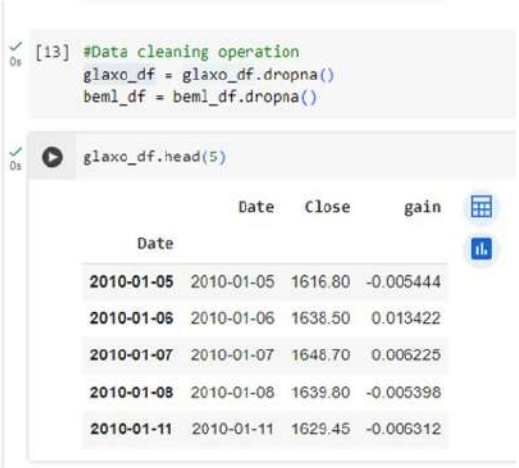
- See in the data there is some missing value
- It can heal with data cleaning operation



	Date	Close	gain
Date			
2010-01-04	2010-01-04	1625.65	NaN
2010-01-05	2010-01-05	1616.80	-0.005444
2010-01-06	2010-01-06	1638.50	0.013422
2010-01-07	2010-01-07	1648.70	0.006225
2010-01-08	2010-01-08	1639.80	-0.005398

Data Preparation:
Example: data cleaning

- Data cleaning



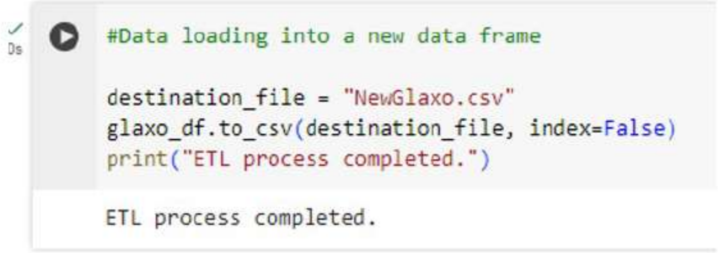
```
[13] #Data cleaning operation
     glaxo_df = glaxo_df.dropna()
     beml_df = beml_df.dropna()
```

```
glaxo_df.head(5)
```

	Date	Close	gain
Date			
2010-01-05	2010-01-05	1616.80	-0.005444
2010-01-06	2010-01-06	1638.50	0.013422
2010-01-07	2010-01-07	1648.70	0.006225
2010-01-08	2010-01-08	1639.80	-0.005398
2010-01-11	2010-01-11	1629.45	-0.006312

Data Preparation:
Example: data load operation

- Loading data into a new file



```
#Data loading into a new data frame

destination_file = "NewGlaxo.csv"
glaxo_df.to_csv(destination_file, index=False)
print("ETL process completed.")
```

```
ETL process completed.
```

Model planning

- Purpose: Define Analytical Approach and Methods
- Tasks:
  - Define Problem Statement and Objectives
  - Select Relevant Variables and Features
  - Choose Suitable Algorithms and Techniques
- Considerations: Model Complexity, Interpretability, Scalability
- Example: Planning a Machine Learning Model for Customer Churn Prediction

Model building

- Purpose: Develop and Train Predictive Models
- Tasks:
  - Split Data into Training and Testing Sets
  - Build and Train Models
  - Fine-Tune Model Parameters
- Importance of Validation and Evaluation Metrics
- Example: Building a Neural Network for Image Recognition

Communicate Results (Documentation)

- Purpose: Document and Communicate Findings
- Tasks:
  - Create Reports, Dashboards, and Visualizations
  - Document Insights and Recommendations
- Importance of Clear and Effective Communication to Stakeholders
- Example: Presenting Data Analysis Results to Company Executives

Operationalize (Quality Assurance)

- Purpose: Implement Models into Production Environment
- Tasks:
  - Deploy Models into Production Systems
  - Monitor Model Performance
  - Address Ethical and Regulatory Considerations
- Importance of Continuous Quality Assurance and Improvement
- Example: Deploying a Fraud Detection Model in Banking Systems