Practices of Business Intelligence

(Descriptive Analytics I: Nature of Data, Statistical Modeling, and Visualization)

Min-Yuh Day
Assistant Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2018-10-03
<table>
<thead>
<tr>
<th>週次 (Week)</th>
<th>日期 (Date)</th>
<th>內容 (Subject/Topics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2018/09/12</td>
<td>商業智慧實務課程介紹</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Course Orientation for Practices of Business Intelligence)</td>
</tr>
<tr>
<td>2</td>
<td>2018/09/19</td>
<td>商業智慧、分析與資料科學</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Business Intelligence, Analytics, and Data Science)</td>
</tr>
<tr>
<td>3</td>
<td>2018/09/26</td>
<td>人工智慧、大數據與雲端運算</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(ABC: AI, Big Data, and Cloud Computing)</td>
</tr>
<tr>
<td>4</td>
<td>2018/10/03</td>
<td>描述性分析I：數據的性質、統計模型與可視化</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Descriptive Analytics I: Nature of Data, Statistical Modeling, and Visualization)</td>
</tr>
<tr>
<td>5</td>
<td>2018/10/10</td>
<td>國慶紀念日 (放假一天) (National Day) (Day off)</td>
</tr>
<tr>
<td>6</td>
<td>2018/10/17</td>
<td>描述性分析II：商業智慧與資料倉儲</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Descriptive Analytics II: Business Intelligence and Data Warehousing)</td>
</tr>
<tr>
<td>週次 (Week)</td>
<td>日期 (Date)</td>
<td>內容 (Subject/Topics)</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>7 2018/10/24</td>
<td>2018/10/24</td>
<td>預測性分析 I：資料探勘流程、方法與演算法 (Predictive Analytics I: Data Mining Process, Methods, and Algorithms)</td>
</tr>
<tr>
<td>8 2018/10/31</td>
<td>2018/10/31</td>
<td>預測性分析 II：文本、網路與社群媒體分析 (Predictive Analytics II: Text, Web, and Social Media Analytics)</td>
</tr>
<tr>
<td>10 2018/11/14</td>
<td>2018/11/14</td>
<td>期中考試 (Midterm Exam)</td>
</tr>
<tr>
<td>12 2018/11/28</td>
<td>2018/11/28</td>
<td>社會網絡分析 (Social Network Analysis)</td>
</tr>
</tbody>
</table>
週次 (Week) 日期 (Date) 內容 (Subject/Topics)
13 2018/12/05 機器學習與深度學習
(Machine Learning and Deep Learning)
14 2018/12/12 自然語言處理
(Natural Language Processing)
15 2018/12/19 AI交談機器人與對話式商務
(AI Chatbots and Conversational Commerce)
16 2018/12/26 商業分析的未來趨勢、隱私與管理考量
(Future Trends, Privacy and Managerial Considerations in Analytics)
17 2019/01/02 期末報告 (Final Project Presentation)
18 2019/01/09 期末考試 (Final Exam)
Business Intelligence (BI)

1. Introduction to BI and Data Science
2. Descriptive Analytics
3. Predictive Analytics
4. Prescriptive Analytics
5. Big Data Analytics
6. Future Trends
Descriptive Analytics I: Nature of Data, Statistical Modeling, and Visualization
Outline

• Descriptive Analytics I
• Nature of Data
• Statistical Modeling
• Visualization
Chapter 1 • An Overview of Business Intelligence, Analytics, and Data Science

Computer technology, management science techniques, and statistics to solve real problems. Of course, many other organizations have proposed their own interpretations and motivations for analytics. For example, SAS Institute Inc. proposed eight levels of analytics that begin with standardized reports from a computer system. These reports essentially provide a sense of what is happening with an organization. Additional technologies have enabled us to create more customized reports that can be generated on an ad hoc basis. The next extension of reporting takes us to OLAP-type queries that allow a user to dig deeper and determine specific sources of concern or opportunities. Technologies available today can also automatically issue alerts for a decision maker when performance warrants such alerts. At a consumer level we see such alerts for weather or other issues. But similar alerts can also be generated in specific settings when sales fall above or below a certain level within a certain time period or when the inventory for a specific product is running low. All of these applications are made possible through analysis and queries on data being collected by an organization. The next level of analysis might entail statistical analysis to better understand patterns. These can then be taken a step further to develop forecasts or models for predicting how customers might respond to a specific marketing campaign or ongoing service/product offerings. When an organization has a good view of what is happening and what is likely to happen, it can also employ other techniques to make the best decisions under the circumstances. These eight levels of analytics are described in more detail in a white paper by SAS (sas.com/news/sascom/analytics_levels.pdf).

This idea of looking at all the data to understand what is happening, what will happen, and how to make the best of it has also been encapsulated by INFORMS in proposing three levels of analytics. These three levels are identified (informs.org/Community/Analytics) as descriptive, predictive, and prescriptive. Figure 1.11 presents a graphical view of these three levels of analytics. It suggests that these three are somewhat independent steps and one type of analytics applications leads to another. It also suggests that there is actually some overlap across these three types of analytics. In either case, the interconnected nature of different types of analytics applications is evident. We next introduce these three levels of analytics.

### Three Types of Analytics

<table>
<thead>
<tr>
<th>Questions</th>
<th>Descriptive</th>
<th>Predictive</th>
<th>Prescriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>What happened?</td>
<td>What is happening?</td>
<td>What will happen?</td>
<td>What should I do?</td>
</tr>
<tr>
<td>What is happening?</td>
<td>Why will it happen?</td>
<td>Why should I do it?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enablers</th>
<th>Descriptive</th>
<th>Predictive</th>
<th>Prescriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business reporting</td>
<td>Data mining</td>
<td>Optimization</td>
<td></td>
</tr>
<tr>
<td>Dashboards</td>
<td>Text mining</td>
<td>Simulation</td>
<td></td>
</tr>
<tr>
<td>Scorecards</td>
<td>Web/media mining</td>
<td>Decision modeling</td>
<td></td>
</tr>
<tr>
<td>Data warehousing</td>
<td>Forecasting</td>
<td>Expert systems</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Descriptive</th>
<th>Predictive</th>
<th>Prescriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well-defined business problems and opportunities</td>
<td>Accurate projections of future events and outcomes</td>
<td>Best possible business decisions and actions</td>
<td></td>
</tr>
</tbody>
</table>

---

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Although its value proposition is undeniable, to live up its promise, the data has to comply with some basic usability and quality metrics. Not all data is useful for all tasks, obviously. That is, data has to match with (have the coverage of the specifics for) the task for which it is intended to be used. Even for a specific task, the relevant data on hand needs to comply with the quality and quantity requirements. Essentially, data has to be analytics ready. So what does it mean to make data analytics ready? In addition to its relevancy to the problem at hand and the quality/quantity requirements, it also has to have a certain data structure in place with key fields/variables with properly normalized values. Furthermore, there must be an organization-wide agreed-on definition for common variables and subject matters (sometimes also called master data management), such as how you define a customer (what characteristics of customers are used to produce a holistic enough representation to analytics) and where in the business process the customer-related information is captured, validated, stored, and updated.

Sometimes the representation of the data may depend on the type of analytics being employed. Predictive algorithms generally require a flat file with a target variable, so making data analytics ready for prediction means that data sets must be transformed into a flat-file format and made ready for ingestion into those predictive algorithms. It is also imperative to match the data to the needs and wants of a specific predictive algorithm and/or a software tool—for instance, neural network algorithms require all input variables to be numerically represented (even the nominal variables need to be converted).
A Simple Taxonomy of Data

Data in Analytics

Structured Data

Categorical
- Nominal
- Ordinal

Numerical
- Interval
- Ratio

Unstructured or Semistructured Data

Textual
- Image
- Audio
- Video

XML/JSON

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Data Preprocessing Steps

Data Consolidation
- Collect data
- Select data
- Integrate data

Data Cleaning
- Impute values
- Reduce noise
- Eliminate duplicates

Data Transformation
- Normalize data
- Discretize data
- Create attributes

Data Reduction
- Reduce dimension
- Reduce volume
- Balance data

Well-Formed Data

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
An Analytics Approach to Predicting Student Attrition

A Graphical Depiction of the Class Imbalance Problem

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Relationship between Statistics and Descriptive Analytics

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Understanding the Specifics about Box-and-Whiskers Plots

- **Outliers**: Larger than 1.5 times the upper quartile
- **Max**: Largest value, excluding larger outliers
- **Upper Quartile**: 25% of data is larger than this value
- **Median**: 50% of data is larger than this value—middle of data set
- **Mean**: Simple average of the data set
- **Lower Quartile**: 25% of data is smaller than this value
- **Min**: Smallest value, excluding smaller outliers
- **Outliers**: Smaller than 1.5 times the lower quartile

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Relationship between Dispersion and Shape Properties.

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
A Scatter Plot and a Linear Regression Line


1. Tabulated Data

2. Data Assessment
   - Scatter plot
   - Correlations

3. Model Fitting
   - Transform data
   - Estimate parameters

4. Model Assessment
   - Test assumptions
   - Assess model fit

5. Deployment
   - One-time use
   - Recurrent use

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
The Logistic Function

\[ f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \]

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Predicting NCAA Bowl Game Outcomes

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
A Sample Time Series of Data on Quarterly Sales Volumes

Quarterly Product Sales (in Millions)

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
The Role of Information Reporting in Managerial Decision Making

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
A Taxonomy of Charts and Graphs

What would you like to show in your chart or graph?

Composition

Distribution

Relationship

Comparison

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
A Gapminder Chart That Shows the Wealth and Health of Nations

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
Magic Quadrant for Business Intelligence and Analytics Platforms

Source: https://www.tableau.com/reports/gartner
A Storyline Visualization in Tableau Software

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
An Overview of SAS Visual Analytics Architecture

DATA BUILDER
- Join data from multiple sources
- Create calculated and derived columns
- Load data

ADMINISTRATOR
- Monitor SAS® LASR™ Analytic server
- Load/unload data
- Manage security

EXPLORER
- Perform ad hoc analysis and data discovery
- Apply advanced analytics

DESIGNER
- Create dashboard style reports for web or mobile

MOBILE BI
- Native iOS and Android applications that deliver interactive reports

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson
A Screenshot from SAS Visual Analytics

A Sample Executive Dashboard

igraph

igraph is a collection of network analysis tools with the emphasis on **efficiency**, **portability** and ease of use. igraph is **open source** and free.

igraph can be programmed in **R**, **Python** and **C/C++**.

[igraph R package](https://igraph.org/R)  [python-igraph](https://igraph.org/python)  [igraph C library](https://igraph.org/C)

**Recent news**

**R/igraph 1.0.0**

June 24, 2015

**Release Notes**

This is a new major release, with a lot of UI changes. We tried to make it easier to use, with short and easy to remember, consistent function names. Unfortunately

[http://igraph.org/redirect.html](http://igraph.org/redirect.html)
Gephi

The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform »

Support us! We are non-profit. Help us to innovate and empower the community by donating only 8€:

Donate

Applications

- **Exploratory Data Analysis**: intuition-oriented analysis by networks manipulations in real time.
- **Link Analysis**: revealing the underlying structures of associations between objects.
- **Social Network Analysis**: easy creation of social

Like Photoshop™ for graphs.

— the Community

PAPERS

LATEST NEWS

- Gephi updates with 0.9.1 version

https://gephi.org/
Discovering, Analyzing, Visualizing and Presenting Data with Python in Google Colab
Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. See our FAQ for more info.

Getting Started

- Overview of Colaboratory
- Loading and saving data: Local files, Drive, Sheets, Google Cloud Storage
- Importing libraries and installing dependencies
- Using Google Cloud BigQuery
- Forms, Charts, Markdown, & Widgets
- TensorFlow with GPU
- Machine Learning Crash Course: Intro to Pandas & First Steps with TensorFlow

Highlighted Features

Seedbank
Looking for Colab notebooks to learn from? Check out Seedbank, a place to discover interactive machine learning examples.

TensorFlow execution

Colaboratory allows you to execute TensorFlow code in your browser with a single click. The example below adds two matrices.

\[
\begin{bmatrix}
1 & 1 & 1 \\
\end{bmatrix}
+ \begin{bmatrix}
1 & 2 & 3 \\
\end{bmatrix}
= \begin{bmatrix}
2 & 3 & 4 \\
\end{bmatrix}
\]
The Quant Finance PyData Stack

Quantopian

PyThalesians

Zipline

DX Analytics

PyAlgoTrade

QuantLib

StatsModels Statistics in Python

scikit-image image processing in python

matplotlib

pandas

$y_{it} = \beta x_{it} + \mu_i + \epsilon_{it}$

SciPy

NumPy

SymPy

IPython

Python

Jupyter

Source: http://nbviewer.jupyter.org/format/slides/github/quantopian/pyfolio/blob/master/pyfolio/examples/overview_slides.ipynb#5
Python
Pandas

\[ y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it} \]

http://pandas.pydata.org/
Iris Classification

### iris.data


<table>
<thead>
<tr>
<th>Sepal Length</th>
<th>Sepal Width</th>
<th>Petal Length</th>
<th>Petal Width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.6</td>
<td>3.1</td>
<td>1.5</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.0</td>
<td>3.6</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.4</td>
<td>3.9</td>
<td>1.7</td>
<td>0.4</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.6</td>
<td>3.4</td>
<td>1.4</td>
<td>0.3</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.0</td>
<td>3.4</td>
<td>1.5</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.4</td>
<td>2.9</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.9</td>
<td>3.1</td>
<td>1.5</td>
<td>0.1</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.4</td>
<td>3.7</td>
<td>1.5</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.8</td>
<td>3.4</td>
<td>1.6</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.8</td>
<td>3.0</td>
<td>1.4</td>
<td>0.1</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.3</td>
<td>3.0</td>
<td>1.1</td>
<td>0.1</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.8</td>
<td>4.0</td>
<td>1.2</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.7</td>
<td>4.4</td>
<td>1.5</td>
<td>0.4</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.4</td>
<td>3.9</td>
<td>1.3</td>
<td>0.4</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.3</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.7</td>
<td>3.8</td>
<td>1.7</td>
<td>0.3</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.1</td>
<td>3.8</td>
<td>1.5</td>
<td>0.3</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.4</td>
<td>3.4</td>
<td>1.7</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.1</td>
<td>3.7</td>
<td>1.5</td>
<td>0.4</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.6</td>
<td>3.6</td>
<td>1.0</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.1</td>
<td>3.3</td>
<td>1.7</td>
<td>0.5</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.8</td>
<td>3.4</td>
<td>1.9</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.0</td>
<td>3.0</td>
<td>1.6</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.0</td>
<td>3.4</td>
<td>1.6</td>
<td>0.4</td>
<td>Iris-setosa</td>
</tr>
</tbody>
</table>

**Species:**
- **setosa**
- **virginica**
- **versicolor**
Iris Data Visualization

Source: https://seaborn.pydata.org/generated/seaborn.pairplot.html
Connect Google Colab in Google Drive
Google Colab
Google Colab

Colaboratory
offered by https://colab.research.google.com
A data analysis tool that combines code, output, and descriptive text into one collaborative document.
Connect Colaboratory to Google Drive

Colaboratory was connected to Google Drive.

Make Colaboratory the default app for files it can open

OK
Google Colab
Google Colab
Google Colab
Run Jupyter Notebook
Python3 GPU
Google Colab
Google Colab Python Hello World

print('Hello World')
import seaborn as sns
iris = sns.load_dataset("iris")
g = sns.pairplot(iris, hue="species")
import seaborn as sns
sns.set(style="ticks", color_codes=True)
iris = sns.load_dataset("iris")
g = sns.pairplot(iris, hue="species")

Source: https://seaborn.pydata.org/generated/seaborn.pairplot.html
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from pandas.plotting import scatter_matrix

# Load dataset
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
df = pd.read_csv(url, names=names)

print(df.head(10))
print(df.tail(10))
print(df.describe())
print(df.info())
print(df.shape)
print(df.groupby('class').size())

plt.rcParams["figure.figsize"] = (10,8)
df.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
plt.show()

df.hist()
plt.show()

scatter_matrix(df)
plt.show()

sns.pairplot(df, hue="class", size=2)
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from pandas.plotting import scatter_matrix

# Import Libraries
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from pandas.plotting import scatter_matrix
print('imported')

imported

names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']

df = pd.read_csv(url, names=names)
print(df.head(10))

# Load dataset
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
df = pd.read_csv(url, names=names)
print(df.head(10))

<table>
<thead>
<tr>
<th>sepal-length</th>
<th>sepal-width</th>
<th>petal-length</th>
<th>petal-width</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.6</td>
<td>3.1</td>
<td>1.5</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.0</td>
<td>3.6</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.4</td>
<td>3.9</td>
<td>1.7</td>
<td>0.4</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.6</td>
<td>3.4</td>
<td>1.4</td>
<td>0.3</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>5.0</td>
<td>3.4</td>
<td>1.5</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.4</td>
<td>2.9</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4.9</td>
<td>3.1</td>
<td>1.5</td>
<td>0.1</td>
<td>Iris-setosa</td>
</tr>
</tbody>
</table>
```python
print(df.tail(10).)
```

<table>
<thead>
<tr>
<th>sepal-length</th>
<th>sepal-width</th>
<th>petal-length</th>
<th>petal-width</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>6.7</td>
<td>3.1</td>
<td>5.6</td>
<td>2.4 Iris-virginica</td>
</tr>
<tr>
<td>141</td>
<td>6.9</td>
<td>3.1</td>
<td>5.1</td>
<td>2.3 Iris-virginica</td>
</tr>
<tr>
<td>142</td>
<td>5.8</td>
<td>2.7</td>
<td>5.1</td>
<td>1.9 Iris-virginica</td>
</tr>
<tr>
<td>143</td>
<td>6.8</td>
<td>3.2</td>
<td>5.9</td>
<td>2.3 Iris-virginica</td>
</tr>
<tr>
<td>144</td>
<td>6.7</td>
<td>3.3</td>
<td>5.7</td>
<td>2.5 Iris-virginica</td>
</tr>
<tr>
<td>145</td>
<td>6.7</td>
<td>3.0</td>
<td>5.2</td>
<td>2.3 Iris-virginica</td>
</tr>
<tr>
<td>146</td>
<td>6.3</td>
<td>2.5</td>
<td>5.0</td>
<td>1.9 Iris-virginica</td>
</tr>
<tr>
<td>147</td>
<td>6.5</td>
<td>3.0</td>
<td>5.2</td>
<td>2.0 Iris-virginica</td>
</tr>
<tr>
<td>148</td>
<td>6.2</td>
<td>3.4</td>
<td>5.4</td>
<td>2.3 Iris-virginica</td>
</tr>
<tr>
<td>149</td>
<td>5.9</td>
<td>3.0</td>
<td>5.1</td>
<td>1.8 Iris-virginica</td>
</tr>
</tbody>
</table>
```python
print(df.describe())
```

<table>
<thead>
<tr>
<th></th>
<th>sepal-length</th>
<th>sepal-width</th>
<th>petal-length</th>
<th>petal-width</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>150.000000</td>
<td>150.000000</td>
<td>150.000000</td>
<td>150.000000</td>
</tr>
<tr>
<td>mean</td>
<td>5.843333</td>
<td>3.054000</td>
<td>3.758667</td>
<td>1.198667</td>
</tr>
<tr>
<td>std</td>
<td>0.828066</td>
<td>0.433594</td>
<td>1.764420</td>
<td>0.763161</td>
</tr>
<tr>
<td>min</td>
<td>4.300000</td>
<td>2.000000</td>
<td>1.000000</td>
<td>0.100000</td>
</tr>
<tr>
<td>25%</td>
<td>5.100000</td>
<td>2.800000</td>
<td>1.600000</td>
<td>0.300000</td>
</tr>
<tr>
<td>50%</td>
<td>5.800000</td>
<td>3.000000</td>
<td>4.350000</td>
<td>1.300000</td>
</tr>
<tr>
<td>75%</td>
<td>6.400000</td>
<td>3.300000</td>
<td>5.100000</td>
<td>1.800000</td>
</tr>
<tr>
<td>max</td>
<td>7.900000</td>
<td>4.400000</td>
<td>6.900000</td>
<td>2.500000</td>
</tr>
</tbody>
</table>
```python
print(df.info())
print(df.shape)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
sepal-length  150 non-null float64
sepal-width   150 non-null float64
petal-length  150 non-null float64
petal-width   150 non-null float64
class         150 non-null object
dtypes: float64(4), object(1)
memory usage: 5.9+ KB
None
```

```python
print(df.shape)
```

```
(150, 5)
```
df.groupby('class').size()

```
print(df.groupby('class').size())
```

class
Iris-setosa       50
Iris-versicolor   50
Iris-virginica    50
dtype: int64
```python
plt.rcParams["figure.figsize"] = (10,8)
df.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
plt.show()
```
df.hist()
plt.show()
```
scatter_matrix(df)
plt.show()
```
```python
sns.pairplot(df, hue="class", size=2)
```

![Sns Pairplot Diagram](image)
Summary

• Descriptive Analytics I
• Nature of Data
• Statistical Modeling
• Visualization
References

• Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.

• EMC Education Services (2015), Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley