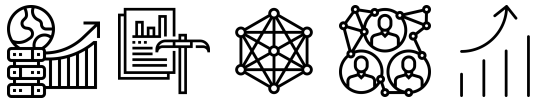


CSC 5741 (2020/21)
Data Mining and Warehousing
Lecture 5: Exploratory Data Analysis

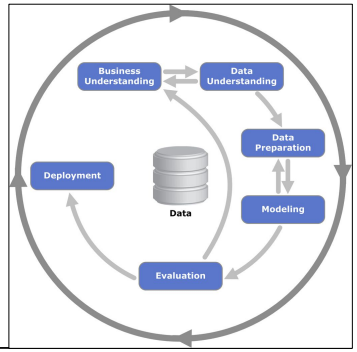
Lighton Phiri
Department of Library & Information Science
University of Zambia
http://lis.unza.zm/~lightonphiri

Lecture Series Outline

- Introduction
- Exploratory Data Analysis
- Descriptive Statistics
- Graphical Techniques
- Jupyter Notebook Walkthrough

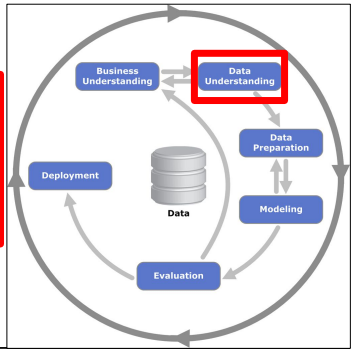
Introduction (1/3)

- The Cross-industry standard process for data mining (CRISP-DM) is a model commonly used to highlight approaches in data mining
  - CRISP-DM segments a data mining project into six phases with no strict order of execution
  - Surveys conducted suggest CRISP-DM is the most widely used methodology



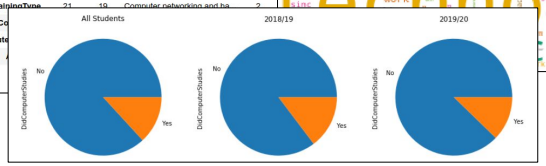
Introduction (2/3)

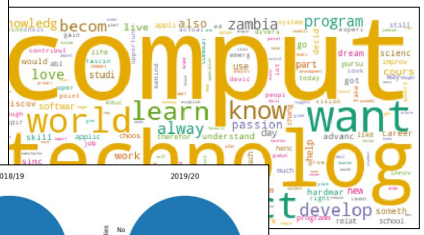
- Identify data sources
- Extract/collect required data
- Described and explore the data collected to gain some sense of what insights to derive
- Ascertain quality of data collected



Introduction (3/3)

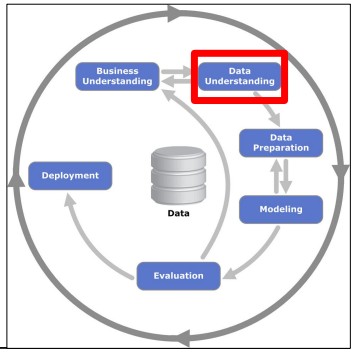
	count	unique	top	freq
Timestamp	91	91	2019/04/03 10:12:12 PM GMT+2	1
StudentName	89	55	Participant30	2
StudentID	91	91	57a47f39353c1f0dd1679e3263...	1
HomeTown	91	75	Lusaka	11
MinorProgramme	91	46	Mathematics	12
MinorProgrammeMotivation	91	90	passion	2
MajorProgrammeMotivation	91	89	love technolog	2
DidComputerStudies	91	2	No	79
HasComputerTraining	91	2	No	72
ComputerTrainingType	21	19	Computer networking and ha...	2
ExperienceWithCo				
HasCompute				





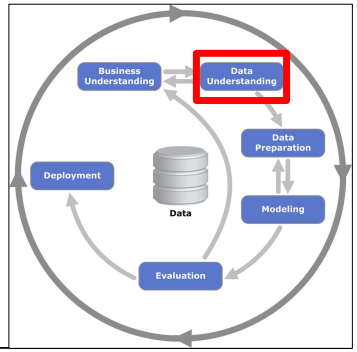
EDA Provides Insights From Datasets

- **The purpose of this EDA is to find insights from datasets and/or data sources**
  - Instrumental for setting the stage for data cleaning and transformation—output subsequently fed to machine learning algorithms.
  - Standard practice: **Data Understanding -> Data Preparation**



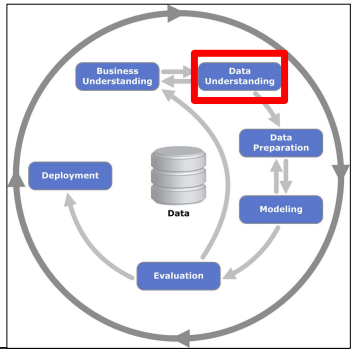
Goals and Objectives of EDA Process

- **Various techniques are employed during EDA in order to achieve the following broad objectives:**
  - Gain comprehensive insight of datasets
  - Identify important data characteristics
  - Identify outliers and anomalies
  - Determine correlations of various data characteristics



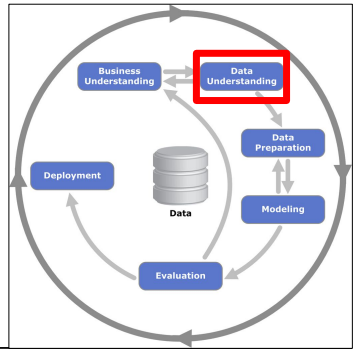
Outcomes of EDA Process

- **Outcome of EDA**
  - Important data attributes
  - Determine attribute characteristics—type of attribute, distribution and statistics (min, mode, median, mean)
  - Understand relationships between the different variables
    - DoB vs Age
  - List of outliers



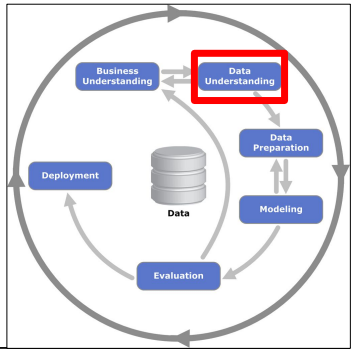
Exploratory Data Analysis (4/6)

- **Leading questions asked during EDA process**
  - What are the different types of data attributes (categorical, continuous, ordina)?
  - How is the data distributed (normal vs non-normal)?
  - Is there a correlation between data attributes and outcome?
  - What are the most important data attributes?



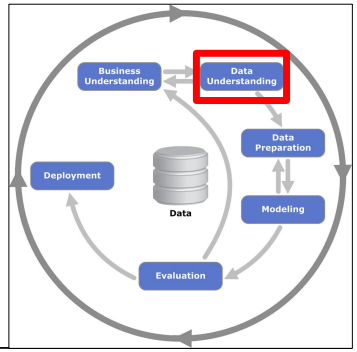
Exploratory Data Analysis (5/6)

- **Leading questions asked during EDA process**
  - What must be done to data attributes with missing values?
  - Do datasets have outliers?



Exploratory Data Analysis (6/6)

- **A graphical approach to EDA is generally effective, although summary tables could also be used**
  - Bar plots for categorical variables and aggregate data
  - Line plots for continuous variables
  - Histograms for continuous variables



Q & A Session

- **Comments, concerns and complaints?**

Lecture Series Outline

- Exploratory Data Analysis
- Descriptive Statistics
- Graphical Techniques
- Jupyter Notebook Walkthrough

Lecture Series Outline

- Part I: Academic Talk
- Part III: Exploratory Data Analysis

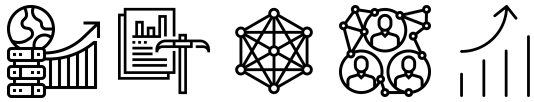
Bibliography

[1] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2017) Data Mining: Practical Machine Learning Tools and Techniques. Chapter 2 https://www.cs.waikato.ac.nz/ml/weka/book.html

[2] NIST/SEMATECH e-Handbook of Statistical Methods. Exploratory Data Analysis. Chapter 1 https://www.itl.nist.gov/div898/handbook/index.htm

[3] Seltman H. J.Experimental Design and Analysis. Exploratory Data Analysis. Chapter 4 https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf

csc5741@unza.zm
http://bit.ly/39HTdTK
http://bit.ly/2kK2ZkA



CSC 5741 (2020/21)
Data Mining and Warehousing
Lecture 5: Exploratory Data Analysis

Lighton Phiri
Department of Library & Information Science
University of Zambia
http://lis.unza.zm/~lightonphiri