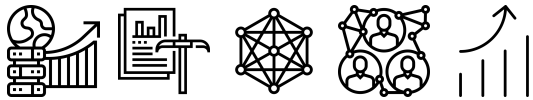


CSC 5741
Lecture 5: Exploratory Data Analysis

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia

Announcements—April 23, 2019

- Mini Project progress
  - Data extraction and collection. You want to make sure you collect A LOT of data. Alternatively, set up pipelines that you could easily reuse.
  - Any challenges you need help with?
    - OAI-PMH (NETD) group?
    - Zambian Advertisements?
    - YouTube group
    - WordPress posts group?

Implementation [8%]
30%: Data collection
30%: Code/scripts works correctly
20%: Novelty of key insights provided
10%: Relevance of implementation
10%: Demonstration

Presentation [4%]
20%: Contents of presentation
20%: Quality of presentation
20%: Visualisations
20%: Comprehensiveness of presentation
20%: Response to questions

Technical Report [8%]
10%: Abstract
10%: Aim/Problem Formulation and Background Work
10%: Implementation
10%: Dataset Description

https://groups.google.com/a/unza.zm/forum/?hl=en#!forum/csc5741

Lecture Series Outline

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

Lecture Series Outline

- Part I: Academic Talk
  - Andreya Kumwenda, Customer Data Mining and Analysis Specialist , MTN Zambia
  - Title: “Data, The Lifeblood and Differentiator In Telecommunications”
- Part III: Exploratory Data Analysis

Lecture Series Outline

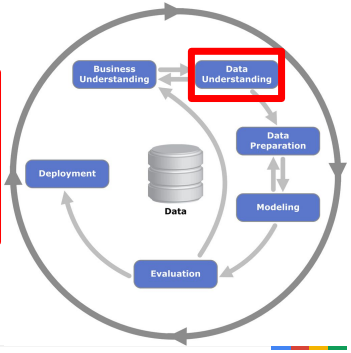
- Part I: Academic Talk
- Part III: Exploratory Data Analysis
  - Introduction
  - Exploratory Data Analysis
  - Jupyter Notebook Walkthrough

Introduction (1/2)

- 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

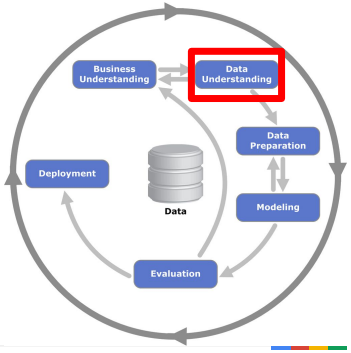
Introduction (2/2)

- 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



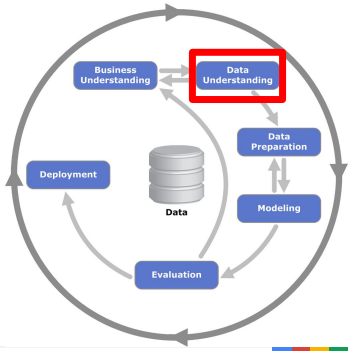
Exploratory Data Analysis (1/6)

- 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**



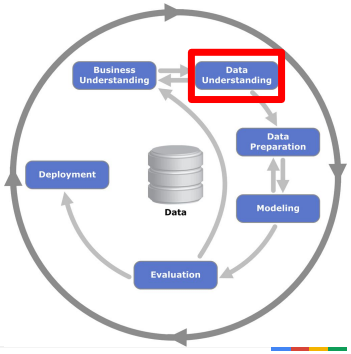
Exploratory Data Analysis (2/6)

- **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



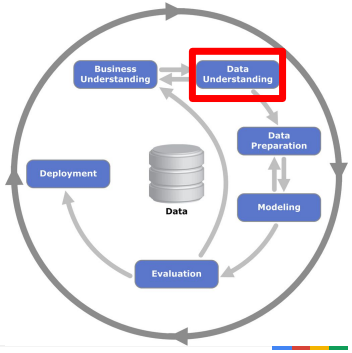
Exploratory Data Analysis (3/6)

- **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



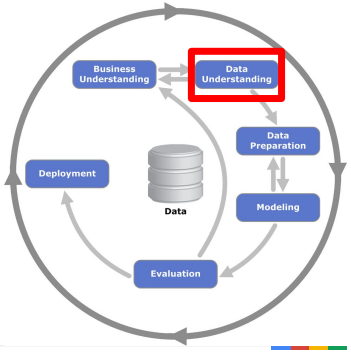
Exploratory Data Analysis (4/6)

- **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?



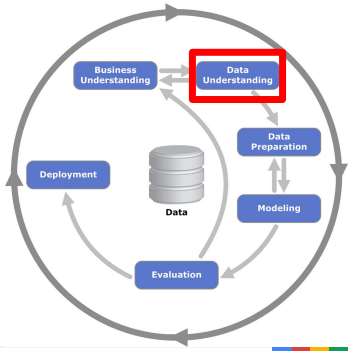
Exploratory Data Analysis (5/6)

- **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

- Part I: Academic Talk
- Part II: Paper Reading Discussion
- Part III: Exploratory Data Analysis
  - Introduction
  - Exploratory Data Analysis
  - Jupyter Notebook Walkthrough

Lecture Series Outline

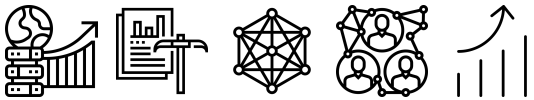
- Part I: Academic Talk
- Part III: Exploratory Data Analysis

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



CSC 5741
Lecture 5: Exploratory Data Analysis

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia