UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
MSc in Science Data Science Program (2023)

STA5092Z - EXPLORATORY DATA ANALYSIS – Lecture Outline

As part of the MSc specializing in Data Science, this course aims to introduce the essential techniques for performing exploratory data analysis. These techniques are typically applied before formal modeling commences and allow the researcher to discover patterns, spot anomalies, test hypotheses and check assumptions with the help of summary statistics and graphical representations. Different types of data will be described and the appropriate exploratory data analysis techniques for each data type will be introduced. The course will distinguish between univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical techniques. Special attention will focus on the visualization of large data dets using appropriate software. Some of the topics to be covered include:

1) Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots).
2) Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
3) Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.
4) Plotting geocoded data and creating dashboards
5) Dimensionality reduction and clustering of similar observations

Lecture Format:

Lectures will take place on Mondays, Wednesdays and Fridays 4-6pm face to face starting on the 13th of February – ending on the 10th of March 2023.

Lecture Venues:

Mondays and Wednesdays in **Hahn 3**
Fridays in **Hahn 4**

Lecturers:

Dr Sebnem Er sebnem.er@uct.ac.za
Mr Stefan Britz stefan.britz@uct.ac.za

Resources

There are some really good free online textbooks by well known and respected teachers in this area – most of the material we need can be based on these three sources:

1. Exploratory Data Analysis with R (Roger Peng = RP): https://bookdown.org/rdpeng/exdata/
2. STA545: Data wrangling, exploration, and analysis with R (Jenny Bryan = JB): https://stat545.com/index.html
3. R for data science (Hadley Wickham = HW): https://r4ds.had.co.nz/

Lecture philosophy

While some of the topics below might seem a bit dry, the idea is for the course to be very applied and hands-on, and based around worked examples and case studies (this is the way the books above are written), with a focus on the code used to generate an analysis.

The course is entirely in R, and the goal of the course is as much to introduce students to R and develop R skills as to cover the theory of EDA – which is after all relatively simple. To quote one of the books above, the course: "is motivated by the need to provide more balance in applied statistical training. Data analysts spend a considerable amount of time on project organization, data cleaning and preparation, and communication. These activities can have a profound effect on the quality and credibility of an analysis. Yet these skills are rarely taught, despite how important and necessary they are."

Assessment

The final mark is calculated using the three assignment marks. Two final marks might be calculated as in option 1 and 2 and the maximum of the two might be recorded as the final mark.

Option 1: 15% Assignment 1, 15% Assignment 2, 70% Assignment 3,
Option 2: 25% Assignment 1, 25% Assignment 2, 50% Assignment 3

Final Mark = Max(option 1, option2)

Lecture outline

Lecture	Material	Possible resources
1	R installation, basics, workflows	JB2
2	Visualizing raw data with ggplot	HW3, RP6
3	Managing data frames with dplyr	JB5-7, HW5, RP3
4	(filter, select, arrange, mutate, summarize) (visualizing univariate summaries)	Some from DSFI
5	EDA checklist	RP4, HW7
6	(right questions, correlation, missing values, outliers)	
7	Reshaping, tidying, joining dataframes	JB8, JB14-16, HW12-13
8	(tidyr, more dplyr)	
9	Principles of good graphics	JB24-27, RP5, RP14-15
10		
11	Exploring time series data	JB10-13, HW14-16
12		
13	Exploring spatial data	

14	(mapview, leaflet, sf)	
15	Functional programming with R	
16	(writing functions, purrr)	
17	Visualising data using animations	RP11-12
18	(gganimate)	
19	Version control with Git and GitHub	JB3, some from DSFI
20		
21	Dashboards	JB42
22	(flexdashboard, shiny)	
23	Dimensionality reduction and Clustering	RP13
24		