UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

School of Computer Science and Applied Mathematics

COMS4048A: Data Analysis and Exploration

Course Outline

Semester 1, 2017

1 Lecturer

Name:	Dr. Terence van Zyl
Email:	terence.vanzyl@wits.ac.za
Office:	MSB UG 21
Consultation Hours:	Tuesdays 14:30-17:30

2 The Course

2.1 Welcome

Welcome to the course. Make sure you read this outline. This outline will be the primary authority on the content and assessment of the course. Please also not that the outline is likely to change throughout the semester and as such the latest version found on Sakai[1] should be consulted.

2.2 Topics Covered

This course introduces the concepts and methodology of Data Analysis for Data Management. In particular the Exploratory Data Analysis (EDA) which falls under the general topic of Data Analysis. EDA is a procedure which assists the discovering of new information from a given dataset. This course will focus on using descriptive statistics to clean and analyse dataset and thus leads to understanding of defining hypotheses for real life problem solving.

- Python programming to perform analysis
- Problem Definition: defining the problems that need to be solved in the context of a given dataset
- Data requirements and collection
- Data PreProcessing
  - Data cleansing
    - Deduplication
    - Missing values
    - Noise elimination
    - Editing
  - Transformation
    - Scaling, Standardization and Normalization

[1] https://cle.wits.ac.za/home/index

        - Binarization
        - Discretization
        - Aggregation and Generalisation
    - Feature Engineering
        - Feature extraction (Encoding/Vectorization, Edge Detection, Fourier Transform, ...)
        - Feature Learning
        - Dimensionality Reduction
    - Feature selection

- Exploratory Data Analysis (EDA)
    - Descriptive Statistics
    - Exploratory Visualisations
    - Exploratory Models

- Multivariate Statistics
    - Multivariate

- Modelling and Analysis
    - Inference and Prediction
    - Forecasting and Extrapolation
    - Modelling Distributions
    - Multivariate Regression

2.3 Expectations

You should have knowledge of

- Descriptive Statistics; Permutations & Combinations; Probability; Discrete & Continuous Random Variable; Sampling & Distributions & Tests of Hypothesis About a Mean; Correlation & Regression

It is useful if you have knowledge or are in a position to acquire knowledge of some of the following:

- Programming in Python.

- You are able to get Anaconda up and running.

3 Outcomes

Having successfully completed this course, the student will be able to:

- Target and define problems related to a given dataset.

- Suggest hypotheses from a given dataset.

- Perform data preparation and assess deficiencies in the collected data.
- Find hidden relationships in a dataset.
- Assess assumptions made in statistical models used in analysis of a dataset.
- Select appropriate statistical techniques for analysis of a dataset.
- Program the above procedures.
- Plan implementations based on results from data analysis.

4 Teaching Methods

A blended learning approach that makes use of the following teaching/learning methodology opportunities and experiences is used:

- lectures;
- tutorials;
- subject-related learning materials; and
- consultations with the course lecturer.

The following scheduled learning opportunities will take place. The content of all tutorial labs may be used for assessment opportunities.

Format	**When**		**Recurrence**	**Location**
Lectures	Monday	14:15-16:30	14 weeks	MSL 110
Labs	NA	NA	NA	NA

4.1 Attendance

Attendance of all lectures is required and or recommended.

4.2 Course Schedule

Learners are requested to review the work schedule below, paying particular attention to the dates of the assessment opportunities. Please note that the schedule pertaining to the material covered is tentative and subject to change.

Week	Date	Format	Subject(s) Covered
01	06/02	Lecture/Lab	Welcome and Introduction
02	13/02	Lecture/Lab	Stating and Refining the Question (ADS Ch1-3) Introduction to Numpy (PDSH Ch2) Assignment 1
03	20/02	Lecture/Lab	Exploratory Data Analysis (ADS Ch4, TS2E Ch1) Data Manipulation with Pandas (PDSH Ch3) Assignment 2
04	27/02	Lecture/Lab	Data Wrangling (PDSH Ch3, P4DA 7) Data Manipulation with Pandas (PDSH Ch3) Assignment 3
05	06/03	Lecture/Lab	Using Models to Explore Your Data (ADS Ch5) Modelling Distributions (TS2E Ch2, 5) PMF, CDF, PDF (TS2E Ch3-4,6) *Self Reading* Visualisation with Pandas (Pandas Docs[2]) Assignment 4
06	13/03	Lecture/Lab	Inference: A Primer (ADS Ch6) Formal Modeling (ADS Ch7) Estimation (TS2E Ch 8) *Self Reading* Inference vs. Prediction: Implications for Modeling Strategy (ADS Ch8) *Self Reading* Getting Started (FPP Ch1) *Self Reading* Data Manipulation with Pandas (PDSH Ch3.Working with Time Series)
07	20/03	Lecture/Lab	Relationships between variables (TS2E Ch7) *Self Reading* Linear Least Square & Regression(TS2E Ch10-11) Assignment 5
08	03/04	Lecture/Lab	The forecaster's toolbox (FPP Ch2) Judgemental forecasts (FPP Ch3)
09	10/04	Test	**Short Practical Project**
10	17/04	Lecture/Lab	Simple regression (FPP Ch4)
11	24/04	Lecture/Lab	Multiple regression (FPP Ch5)
12	01/05	Lecture/Lab	Hypothesis testing (TS2E Ch 8) *Self Reading* Time Series Analysis (TS2E Ch12)
13	08/05	Lecture/Lab	Interpreting Your Results and Communication (ADS Ch9-10) Survival analysis (TS2E CHh13)
14	15/05	Lecture	**Exam Prep**

5 Assessments

An integrated approach to assessment whereby assessment forms an integral part of teaching and learning is followed, this integrated assessment takes the form of:

Formative Assessments The learner is assessed throughout the semester in the form of class assignments, lab assignments, a project and a semester test.

Summative Assessments The learner is required to complete a written/programmed examination that is representative of all the work covered at the end of the semester.

Class Test and Assignments	**28%**
Short Practical Project	**10%**
Exam	**62%**

Exam Date TBA.

5.1 Rubrics

5.1.1 Short Practical Project

	0%	**-50%**	**-75%**	**-100%**	**Weight**
					20%
					5%
					20%
					20%
					35%
				Total	100%

5.1.2 Assignments

The class test and assignment

	0-20%	**-40%**	**-50%**	**-60%**	**-70%**	**-75%**	**-100%**	**Weight**
Data Wrangling								25%
EDA								25%
Modeling								25%
								25%
			Total					100%

5.2 Academic Integrity

Copying, communicating, or using disallowed materials during an exam is cheating, of course. Students caught cheating on a test or final exam will be reported to the university disciplinary committee. Academic integrity is a more complicated issue for programming assignments, but one we take very seriously. Students naturally want to work together, and it is clear they learn a great deal by doing so. Getting help is often the best way to interpret error messages and find bugs, even for experienced programmers. In response to this, the following rules will be in force for programming assignments:

- Students are allowed to work together in designing algorithms, in interpreting error messages, and in discussing strategies for finding bugs, but NOT in writing code.

- Students may not share code, may not copy code, and may not discuss code in detail (line-by-line or loop-by-loop) while it is being written or afterwards.

- Similarly, students may not receive detailed help on their code from individuals outside the course. This restriction includes tutors, students from prior terms, Internet resources, etc.

- Students may not show their code to other students as a means of helping them. Sometimes good students who feel sorry for struggling students are tempted to provide them with “just a peek” at their code. Such “peeks” often turn into extensive copying, despite prior claims of good intentions.

- Students may not leave their code (either electronic versions or printed copies) in publicly accessible areas. Students may not share computers in any way when there is an assignment pending.

We may use various code comparison tools including automated tools to help spot assignments that have been submitted in violation of these rules. The tools take all assignments from all sections and all prior terms and compares them, highlighting regions of the code that are similar. We (the instructor and the teaching assistants) check flagged pairs of assignments very carefully ourselves, and make our own judgement about which students violated the rules of academic integrity on programming assignments. When we believe an incident of academic dishonesty has occurred, we contact the students involved. All students caught cheating on a programming assignment (both the copier and the provider) will receive an automatic 0 for that assignment. No excuses, no discussions, no exceptions! The university rules regarding academic dishonesty and plagiarism can be found at

http://www.wits.ac.za/academic/science/stats/courses/5721/plagiarism_policy.html

Due to problems with plagiarism, **all assignments will be checked.** If plagiarism is found, action will be taken, and university disciplinary processes will be activated.

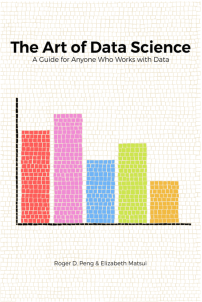
6 Course Material

Students are encouraged to make use of the available resources for the module. Sources include: lecture notes, other information on the e-Learning platforms(s) and the prescribed textbook.

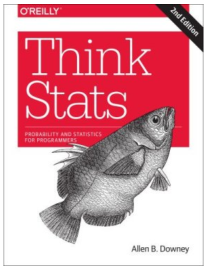
You are encouraged to use Google, YouTube, OpenCourseWare, StackOverflow and any other online resources. OpenCourseWare specifically is very good.

6.1 Textbook

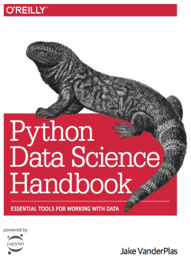
The textbooks for this course is available online for free:


The Art of Data Science
A Guide for Anyone Who Works with Data
Roger D. Peng & Elizabeth Matsui

Title:	The Art of Data Science
Edition:	First Edition
Author:	Roger D. Peng and Elizabeth Matsui
Publisher:	Lean Publishing
ISBN:	
Web Site:	https://leanpub.com/artofdatascience



Title:	Think Stats
Edition:	Second Edition
Author:	Allen B. Downey
Publisher:	O'Reilly
ISBN:	
Web Site:	http://greenteapress.com/wp/think-stats-2e/


O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by jupyter
Jake VanderPlas

Title:	Excerpt from (Python Data Science Handbook)
Edition:	Second Edition
Author:	Jake VanderPlas
Publisher:	O'Reilly
ISBN:	
Web Site:	https://github.com/jakevdp/PythonDataScienceHandboo

Title:	Forecasting: principles and practice
Edition:	Online Edition
Author:	Rob J Hyndman and George Athanasopoulos
Publisher:	Copyright © 2017, OTexts.
ISBN:	
Web Site:	https://www.otexts.org/fpp

6.2 Reading List

- Leek, Jeffery T., and Roger D. Peng. ”What is the question?.” Science 347.6228 (2015): 1314-1315.
- https://github.com/jvns/pandas-cookbook

6.3 Viewing and Listening List

6.4 e-Learning Resources

Links to online resources will be provided via Sakai[3] when relevant, but you are encouraged to find your own as well.

Appendix A: Notes

Notes on Project

1. For all projects, the project descriptions as well as reference code will be distributed using Sakai.
2. You must implement the projects in the language requested.
3. You may do the projects individually.

[3] https://cle.wits.ac.za/

4. Projects usually last a week. Submissions up to 24 hours late will be accepted with a 20% penalty. No further submissions will be accepted. Do not leave them to the last minute; congestion on Sakai 5 minutes before the deadline is not an acceptable excuse.

5. Projects will be submitted using Sakai. It is your responsibility to figure out how to submit projects. If you have any technical problems, you may contact Mr. Shunmuga Pillay (shunmuga.pillay@wits.ac.za).

6. Each project description will come with a detailed breakdown of the grading scheme and you will receive credit for different tasks of a project; however, no partial credit will be given for individual tasks. Also, there may sometimes be an optional part for extra credit. So, you can decide how much of each project you can/want to do. All projects will be graded out of 100.

7. The projects are designed to help you learn and apply the material taught in lectures and laboratory sessions; but you will only learn and understand by doing the projects.

8. If you need help with any of the projects, then please ask for the appropriate help (teaching assistant or lecturer) and help will gladly be given. We will not solve the problems for you. You must show that you have attempted the problem from a number of different aspects before we will intervene.

9. You will be required to make use of the basic features of a distributed version control system; specifically we will use GIT.

Notes on Labs

1. The purpose of the laboratories is for you to get hands-on experience with the theory that has been discussed in classes.

2. You will work on these small programming problems during the laboratory sessions and you may ask the teaching assistants for help.

3. You are not expected to complete all programming problems during the laboratory sessions. You may complete the work sheets in your own time if necessary.

4. You should not expect sample solutions of work sheets and projects to be handed out. If you cannot solve a programming problem on your own, then please ask for help and help will gladly be given.

Notes on Tests/Exam

1. All tests/exams will be closed-book.

2. Each test will have a separate lab test component that will be tested during normal lab times.

3. All tests/exams will also test whether you have done the worksheets and projects. So, be prepared for questions about the projects and worksheet problems.

4. If you have queries regarding the marking of your test script, you must write a short paragraph to submit with your script for remarking, that describes why you believe that a specific question deserves more marks. If I've made a mistake adding this is not necessary, just bring the script to me.

Appendix B: Assessment

1. Why would you use a Tab delimited file over a CSV?
   - ____
   - ____
2. Python is a dynamically typed language. What does this mean?
   - ____
   - ____
3. Why would you model the heights of humans using a Gaussian distribution?
   - ____
   - ____
4. What does the so called p-value tell you?
   - ____
   - ____
5. What is the difference between regression, classification and forecasting?
   - ____
   - ____
6. Name a few common aggregation functions you can perform on a data column.
   - ____
   - ____
7. If I had to take a slice from a row major 2D array of size 11x15 say A[3:10] how many items would I land up with?
   - ____
   - ____