Table of Contents

[***Lecturer handout*** 2](#_Toc90993605)

[Overview 2](#_Toc90993606)

[Motivation 2](#_Toc90993607)

[Working with the code 2](#_Toc90993608)

[Objectives 3](#_Toc90993609)

[Session 1 3](#_Toc90993610)

[Exploratory Data Analysis 3](#_Toc90993611)

[Schedule 3](#_Toc90993612)

[Session 2 4](#_Toc90993613)

[Key Words in Context 4](#_Toc90993614)

[Schedule 4](#_Toc90993615)

[Session 3 5](#_Toc90993616)

[Data Pipelines 5](#_Toc90993617)

[Quick Visualizations 5](#_Toc90993618)

[Schedule 5](#_Toc90993619)

[Extra – Data Enrichment 5](#_Toc90993620)

# Lecturer handout

Here are the notes on how the author plans to teach the ADSEE course on Job Market signalling.

This plan is configured for all 3 themes split over 3 periods of 4 hours.

# Overview

This is a hands-on course where non-technical students get to interact with statistical programming based on the R language. The student is immersed in reading the code and theory on a particular topic. In this case how to derive information (signals) from the Job market based on analyzing a realistic dataset, 20 thousand job advertisements.

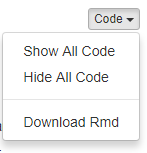
## Motivation

This course is a stepping stone for non-technical researchers who wish to involve themselves with data science approaches. The emphasis is on learning to read code (not necessarily write code due to the time constraints). At the end of the course the student should understand that data science is relevant to their studies, that data is dirty and it takes effort to clean, that there are simple programmatic recipes that achieve much with little coding, that the R language is a viable tool.

## Working with the code

The code is in RMarkdown Notebook format and can be run by the students on their computers via R studio. The expected output and code are stored in a web browser readable format. Therefore, a student that has failed to install R studio may still work along via a web browser. The teacher can share their screen while video conferencing and run the code via R studio. In a classroom setting via a beamer or video conferencing the shared screen.

Each student is provided with an **.nb.html** file which they can read via a standard web browser. The code is also included in the html file. At the top of the web page, they will need to download it via the code button.



The code is approximately explained in a slide deck.

Code breaks, most likely to do with the libraries being improved over time. For each code notebook there is a provided list of version numbers.

A theory slide deck is available per theme.

## Objectives

By the of this course, you will be able to:

1. Describe a reproducible data science work flow to process text from job vacancies
2. Enact a common business process lifecycle that leads to describing the operationalization of the results
3. Review a textual data source of 20,000 job advertisements (vacancies)
4. Explain at least five examples of how text in vacancies signal specific mechanisms.
5. Install a development environment for R
6. Run R code
7. Become familiar with numerous functions within R
8. Run recipes for common text mining methods so that you can visualise differences in occupation and Education level
9. Explore, visualise using common methods within the R language.

# Session 1

Session has extra content due to the setup requirements.

## Exploratory Data Analysis

* It is important to analysis the problem that you want to solve for the business. However, there is a potential contradiction, a Data Science workflow is iterative in its nature.
* In this session we will review Job Market Signalling and discuss gender issues.
* The code associated with this session deals with exploring the data and its weaknesses as early in the cycle as possible. The motivation is to discover issues early, where it is cheapest to solve.

### Schedule

|  |  |  |
| --- | --- | --- |
| Time | Event | Resources (shared via Chat) |
| 14:00  14:50 | Introduction / Theory | 1.EDA.Part1.pdf |
| 14:50  15:00 | Q/A or theory continued |  |
| 15:00  15:10 | Break |  |
| 15:10  15:30 | Check setup Rstudio – Self help  Breakout rooms | Setup.pdf |
| 15:30  16:30 | Code Review | 1.EDA.nb.html  1.EDA.Demo.rmd  1.EDA.Part2.pdf |
| 16:30  16:45 | Break |  |
| 16:45  17:00 | Q/A or Code review | All resources |
| 17:00  17:45 | Exercises  Breakout room | Work through exercises in slides. |
| 17:45 | Summarize & answer any questions |  |

# Session 2

## Key Words in Context

* In this session we look at a number of basic approaches to analyzing Job Market Data centered around text mining. During this course we will exercise a number of the approaches.
* The coding section is about looking at the text, seeing specific words in their native context. In this case Job vacancies. This is a well-tested method which enables problem domain experts to create dictionaries of words that represent different signals.
* In the results decomposition session, the code section will detail one method to measure the reliability of the expert opinion.

### Schedule

|  |  |  |
| --- | --- | --- |
| Time | Event | Resources |
| 13:45  14:00 | Teacher available to answer questions | Setup.pdf |
| 14:00  14:45 | Theory | 1.EDA-Part1.pdf |
| 14:45  15:00 | Break |  |
| 15:00  15:15 | Check setup Orange 3 – Self help  Breakout rooms |  |
| 15:15  16:00 | Code Review | 2.Keywords.In.Context.nb.html  Kwic.demo.Rmd  1.EDA-Part2.pdf |
| 16:00  16:15 | Break |  |
| 16:15  16:45 | Demonstration Orange 3 | Orange3 & dataflow files  Kwic.ows  Visualization.of.Data.Subsets.ows |
| 16:45  17:00 | Break |  |
| 17:00  17:45 | Exercises  Breakout rooms | Exercises in presentations |
| 17:45 | Summarize |  |

# Session 3

## Data Pipelines

* In this section we review an opportunistic source of data, job vacancies and an example of the dimensions associated with the data.
* The coding section is describes building a reproducible data pipeline for the Job descriptions.

## Quick Visualizations

* The primary source of data is textual, in this section we again look at mainstream techniques.
* The coding section builds on your skills to quickly visualize relevant features in your data. Visualization support intuitive understanding of the techniques.

### Schedule

|  |  |  |
| --- | --- | --- |
| Time | Event | Resources |
| 13:45  14:00 | Teacher available to answer questions |  |
| 14:00  14:45 | Data Pipeline | 3.Data.Selection-Part1.pdf  Data.Pipeline.Demo.Rmd |
| 14:45  15:00 | Break | Exercises in presentations |
| 15:00  15:30 | Exercises  Breakout rooms |  |
| 15:30  16:15 | Quick visualizations | 3.Quick.Visualizations.nb-Part2.pdf  Quick.Visualizations.Demo.Rmd |
| 16:15  16:30 | Break |  |
| 16:30  17:30 | Exercises  Breakout rooms | Exercises in presentations |
| 17:30  18:00 | Q/A & Roundup |  |

# Extra – Data Enrichment

It is difficult to judge the level of the audience. Extra content is provided around the subject of enriching data as support.

* Depending on the question you wanted answer you can combine the vacancy dataset with other types of mostly freely available data.
* There are a number of ways of gathering useful data through the R programming language which we explore