E+ ADSEE

JobMarket Signalling
1. Exploratory Data Analysis

ADSEE

Co-funded by the
Erasmus+ Programme
of the European Union

Overview

It is important to explore your data before you make conclusions. **Data tends to be dirty** with the majority of time in a Research project spent on gathering and cleaning the data. Programming languages can do most of the busy work for you.

R has many useful built in functions and is very good and loading data from many formats or saving data to many formats.

Packages are a collection of functions centred on a specific theme. Using the right packages can save you a lot of time and busy work.

Packages used in the EDA notebook

Base R

- Sample function: For almost random sampling of your dataset

Packages

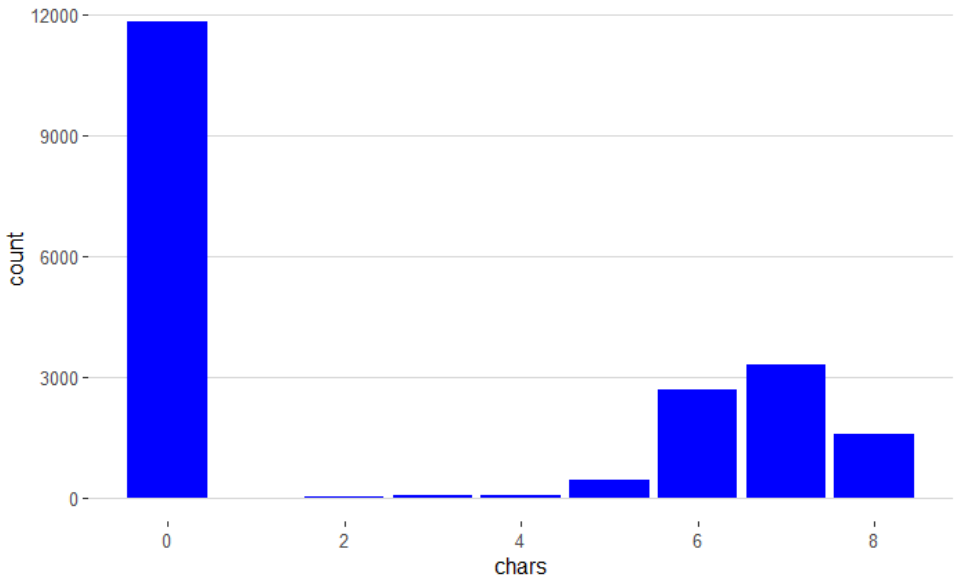
- GGPLOT2 and GGTHEMES for efficient graph generation
- Lubridate for the manipulation of data strings
- DataExplorer for automatic reporting of the essential features of your data

Data is Dirty

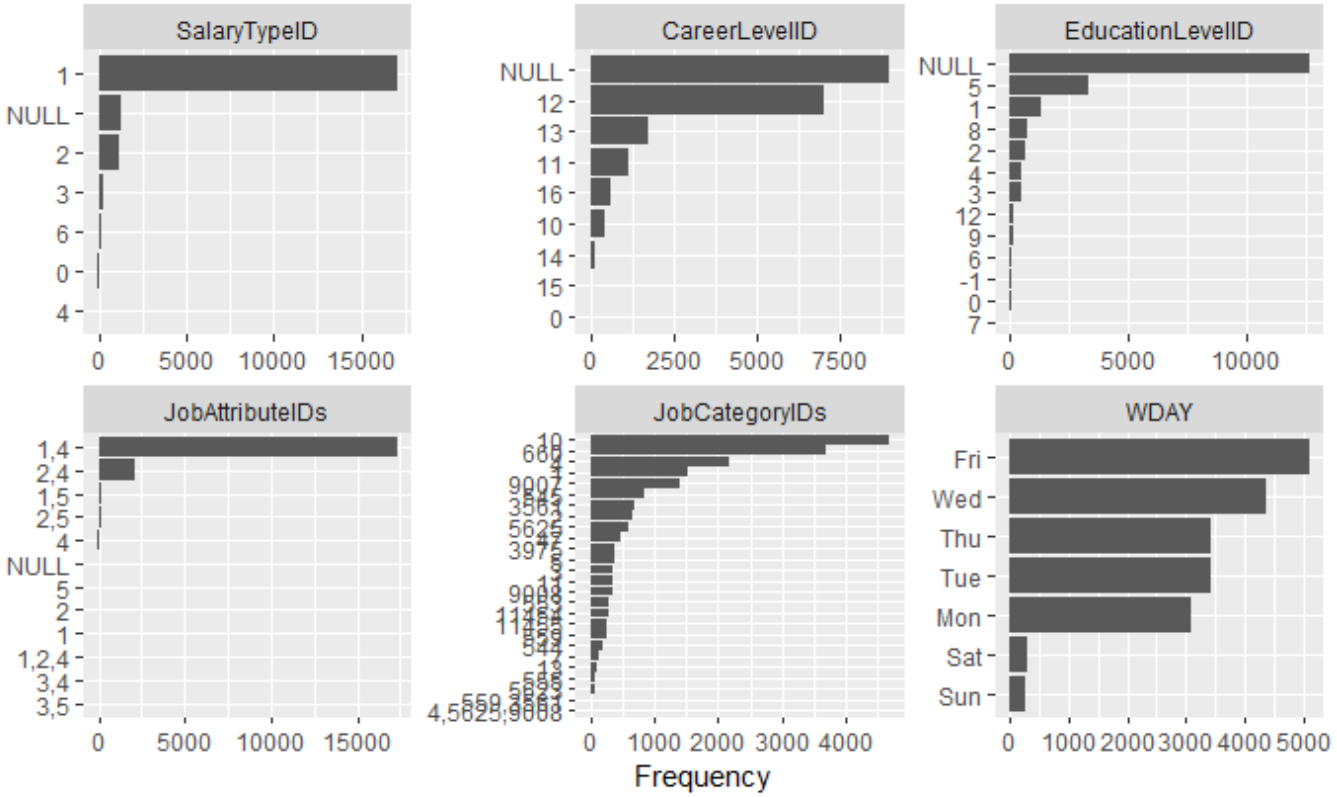
Human Input

	City <fctr>	Freq <int>
716	London	4110
714	london	99
717	LONDON	57
715	lONDON	10

Graph of **postcode data**
Missing or incorrect fields



Exploratory Data Analysis



Dataexplorer allows you to quickly explore your data.

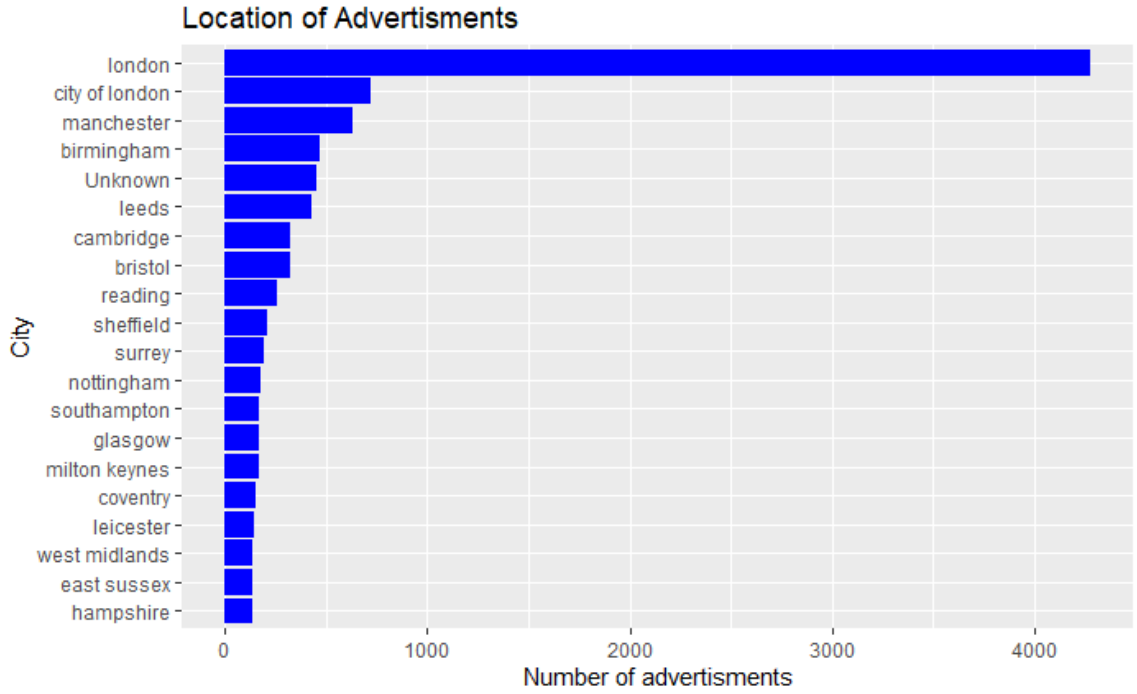
The more you look the more you will see.

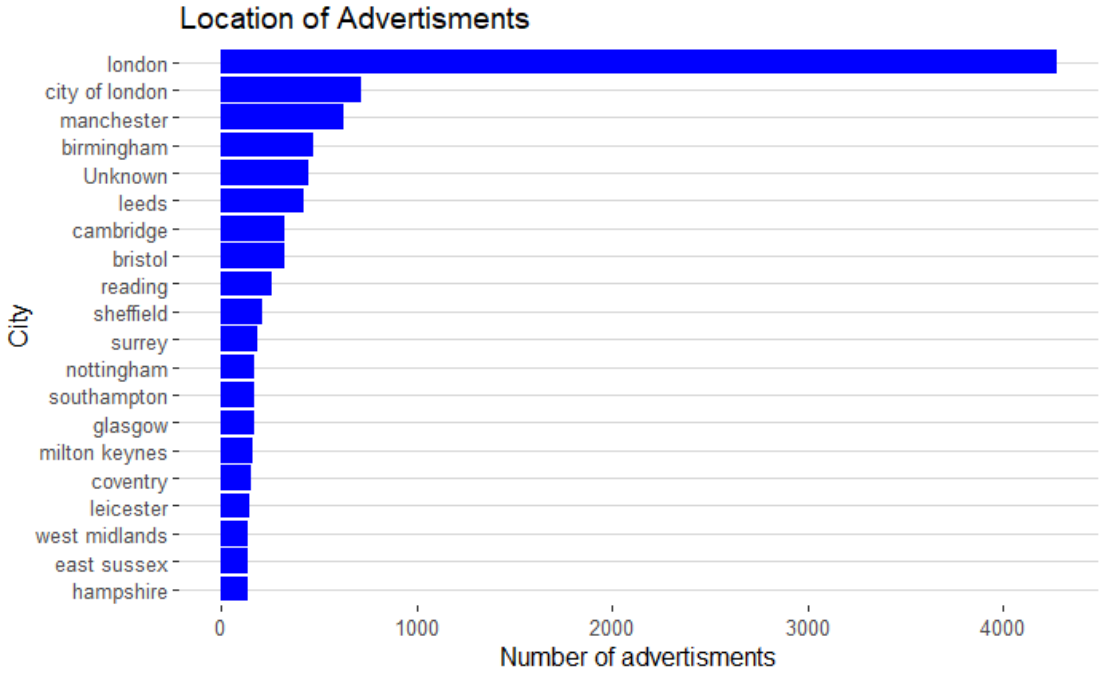
- The **EducationLevel** has a lot of missing data (NULL).
- Job adverts are seldom first published on Saturday or Sunday
- Certain Job Categories are better represented.

What can you see in this visualisation?

Plotting is straight forward

With a few lines of code via GGplot2 you can make publication ready graphics
GGthemes allows you to change the look and feel of a graph with one command





Resources

Blog: DataExplorer

- https://www.business-science.io/code-tools/2021/03/02/use-dataexplorer-for-EDA.html

Chapter of book

- https://r4ds.had.co.nz/exploratory-data-analysis.html

Cheatsheets

- Lubridate - https://rawgit.com/rstudio/cheatsheets/master/lubridate.pdf
- Import Data - https://github.com/rstudio/cheatsheets/raw/master/data-import.pdf
- Data Visualisation - https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf