**3803ICT Big Data Analysis 2020**

*Due: Friday, 22 May 2020  
Marks: 30%  
Partner: Ryan Taylor*

*­*

**Assignment 1**

**Luka Zoric  
s5027193  
luka.zoric@griffithuni.edu.au**

**Overview:**

In this assignment, we will apply data analytics, using the tools introduced in the labs. We are required to study the SEEK job market data and analyze. There are three parts to this assignment:

1. DATA PREPARATION AND PREPROCESSING
2. DATA ANALYSIS (EXPLORATORY, STATISTICAL AND PREDICTIVE)
3. EVALUATION

**Part 1 – Data Preparation and Preprocessing**

* **Dataset Description**

Given dataset is in a format of a CSV file. CSV stands for Comma Separated Values, simple format for representing a rectangular array of numeric and textual values. Each line contains the same number of fields, where each field that contains a special character such as comma or double quote must be escaped.

By looking at the given file, it appears to have 13 categories and all together 318477 entries. Only two of them appear to have correct data type, while rest are non-nullable objects. This suggests that we will have to do some data casting when cleaning up data. Additionally, we will take look into which columns are actually useful by looking into null values and evaluating their importance.

Table below lists all the columns and displays useful information gathered at the first look at the data frame:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column Name | Original dtype | Correct dtype | Entries | NoN-null Entries (%) | unique entries |
| Id | object | string/int64 | 318477 | 100% | 318477 |
| Title | object | string | 318477 | 100% | 168065 |
| Company | object | string | 306473 | 96% | 40628 |
| Date | object | DateTime | 318477 | 100% | 163 |
| Location | object | string | 197229 | 62% | 65 |
| Area | object | string | 122658 | 38% | 19 |
| Classification | object | string | 197229 | 62% | 30 |
| SubClassification | object | string | 197229 | 62% | 338 |
| Requirement | object | string | 318470 | 99% | 234287 |
| FullDescription | object | string | 302302 | 94% | 250901 |
| LowestSalary | int64 | int64 | 318477 | 100% | 11 |
| HighestSalary | int64 | int64 | 318477 | 100% | 11 |
| JobType | object | string | 302379 | 94% | 4 |

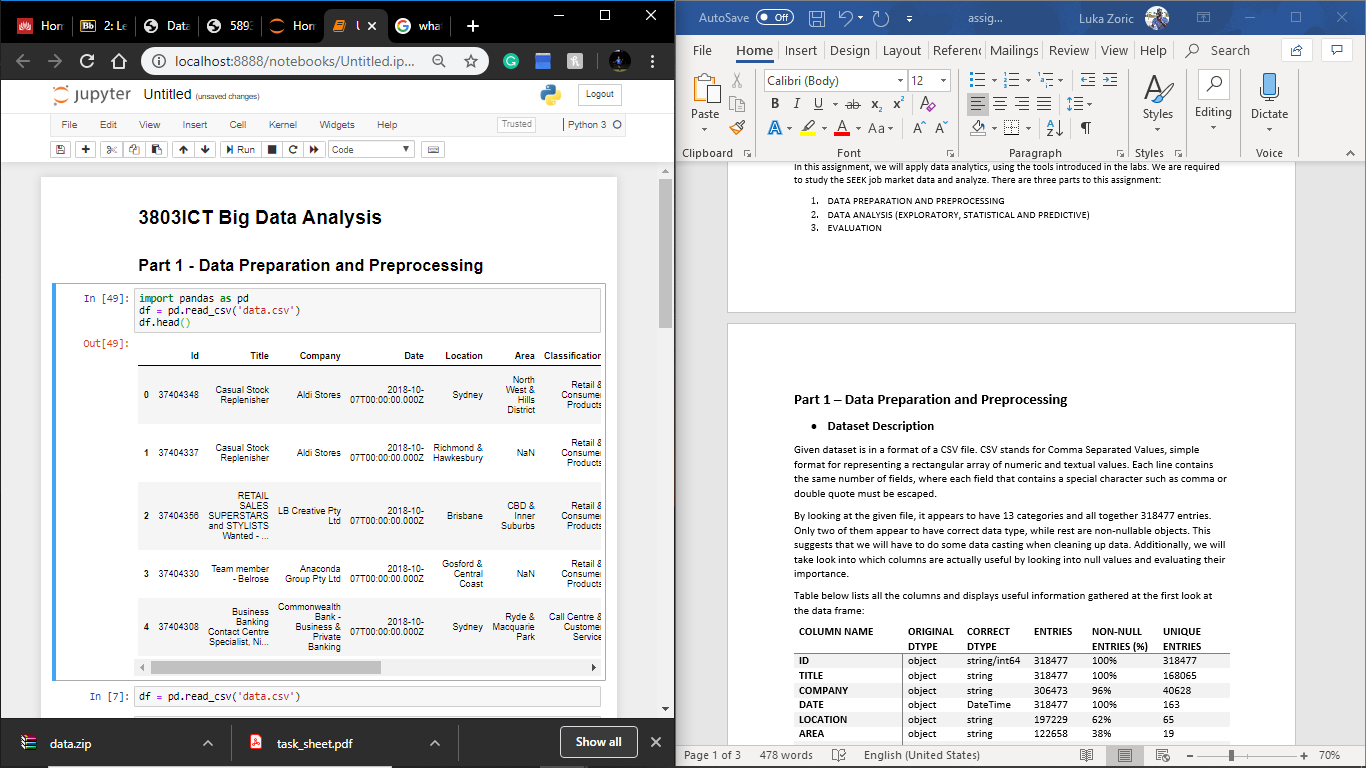
By looking at the table below, we see that most of the data has entries, however, column representing Area (location) has only 38% of non-null values. There is a possibility of filling up missing values by matching it with location column.

Additionally, this data describes 4 job types, from 40628 different companies across 65 locations. Great majority of the data is categorical. In order to process and analyse this data it is important to firstly perform data preparation and preprocessing. This will include normalizing the data and cleaning it up with method previously mentioned (for location) and many more. This will furthermore be discussed in the following part of the assignment.

* **Data Preparation and Preprocessing**

To load this data and perform analysis on it, we will use Python module Pandas. As our framework we will use Jupiter Notebook. CSV data is compatible with Pandas in a way that it is easily loaded using *read\_csv* function and stored as a *DataFrame*.

Pandas DataFrame is a two-dimensional size-mutable tabular data structure with labeled rows and columns, referred to as axes. Image below displays importing Pandas module, reading the file and displaying first 5 entries with *head()* function.



Firstly, we will convert all data types into what appears to be correct ones (mentioned in the first table).