Part 1: Data Wrangling

Gathering:

First, we searched for a dataset in Kaggle and chose ‘**Wuzzuf_Job_Posts_Sample.csv’** file in the ‘**WUZZUF Job Posts (2014-2016)’** dataset

Assessment:

Secondly, we searched for some quality and tidiness issues and found 9 quality issues and 2 tidiness issues.

Quality:

1- using the ‘info’ method we can see that ‘post_date’ is a string instead of date-time

2- by looking at the head and the tail of the data we can see that ‘Job Description’ contains HTML tags and HTML special characters

3- by looking at the head and the tail of the data we can see that ‘Job Requirements’ contains HTML tags and HTML special characters

4- by looking at the head and the tail of the data we can see that there are some Missing Values in the ‘Job Description’

5- by looking at the head and the tail of the data we can see that there are some Missing Values in the ‘Job Requirements’

6- by visual assessment we found that some salaries are 0

7- by looking at the head and the tail of the data we can see that there are some Missing Values in the ‘Job Category’ named "Select"

8- by looking at the head and the tail of the data we can see that there are some Missing Values in the ‘Job industry’ named "Select"

9- by looking at the head and the tail of the data we can see that ‘experience_years’ contains inconsistent data

Tidiness:

1- by looking at the head and the tail of the data we can see that ‘Job Categories’ are split over three columns

2- by looking at the head and the tail of the data we can see that ‘Job Industries’ are split over three columns

---

Cleaning:

1- we changed the data type of post_date column from string to date time using ‘pd.to_datetime’ method.

2,3- we replaced the html tags to an empty string using ‘.replace’ function and the following regex parameter ('(<[^>]*>|\\r|\\n|&.*;)', '',regex=True, case=False).

4,5- we used ‘.fillna’ method to fill the null values with a good description .

6- we used ‘.mean’ method to find the salaries mean and replaced the 0 by the mean using ‘.replace’ method.

7,8(quality) and 1,2(tidiness)- we will put the 3 columns in one list as a string and replace all "Select" words by empty string and replace all "/" by ‘,’ then we will add this list to a new data frame then convert the string to list using ‘.split’ method then remove the 3 columns using ‘.pop’ method then insert this new data frame column to the original data frame.

9- we will extract the first and second number and separate put ‘-‘ between them and if it has only 1 number won’t change the data