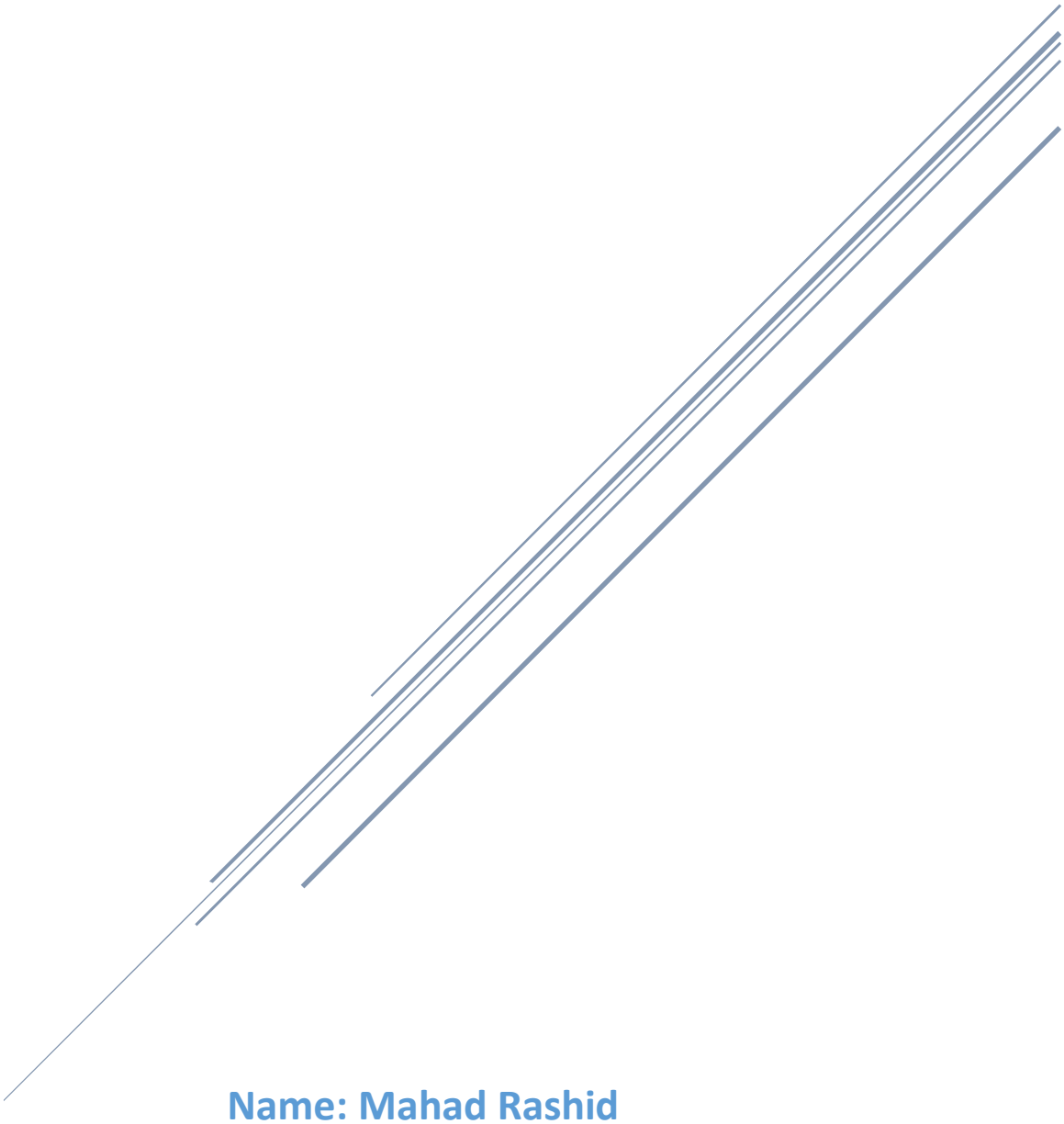
SIT742: MODERN DATA SCIENCE

2020 Assessment Task 01: Exploration for Data Scientists Survey Data



Name: Mahad Rashid
Student ID: 219367287

- Task 1.0.A

```
TitleFit                          4251
CurrentEmployerType               4275
MLToolNextYearSelect              4206
MLMethodNextYearSelect            4170
LanguageRecommendationSelect      4228
MajorSelect                       3952
FirstTrainingSelect               4324
JobSatisfaction                   4317
```

- Task 1.0.B

```
df_demog_ds=df_demog.loc[df_demog['CurrentJobTitleSelect'] == 'Data Scientist']
print('The number of users whose current job title is Data Scientist are: ')
print(len(df_demog_ds))
```

```
The number of users whose current job title is Data Scientist are:
1263
```

- Task 1.1.A

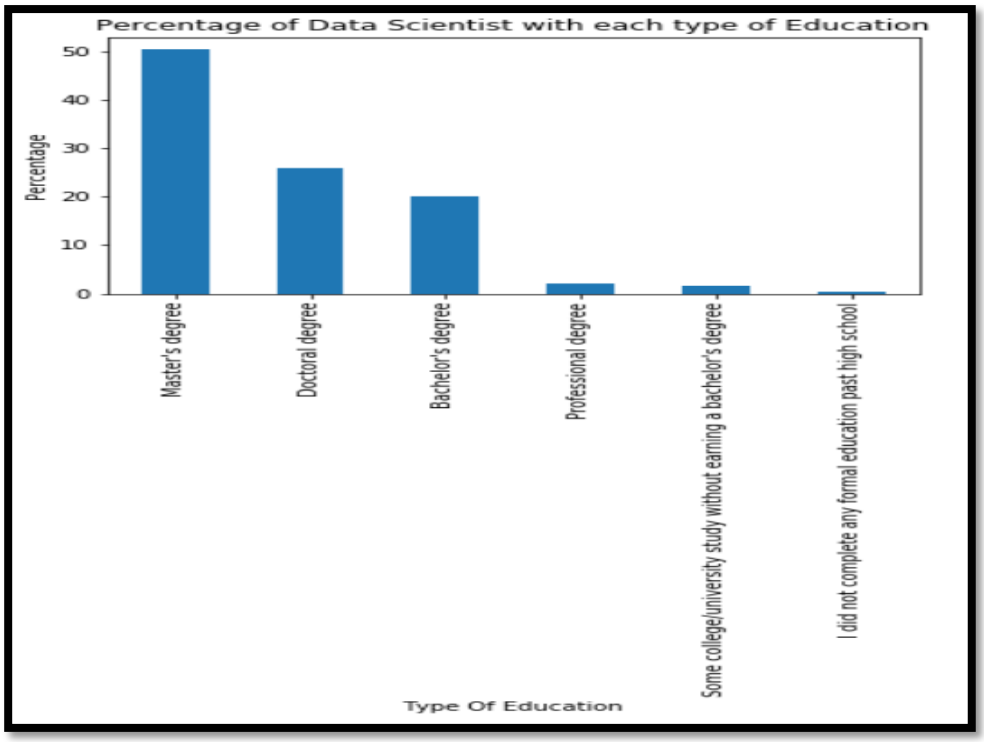
Total Number of data scientist in each formal education category:

```
FormalEducation
Bachelor's degree                                                252
Doctoral degree                                                  326
I did not complete any formal education past high school           6
Master's degree                                                  635
Professional degree                                               25
Some college/university study without earning a bachelor's degree 19
dtype: int64
```

Percentage of data scientist in each formal education category:

Master's degree	50.277118
Doctoral degree	25.811560
Bachelor's degree	19.952494
Professional degree	1.979414
Some college/university study without earning a bachelor's degree	1.504355
I did not complete any formal education past high school	0.475059

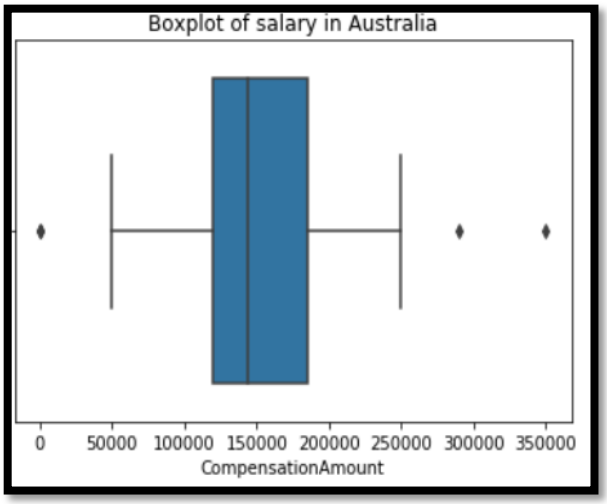
Name: FormalEducation, dtype: float64



- ## Task 1.2.A:

- The maximum salary for all the respondents is: 742720.4288
- The maximum salary for all the respondents is: 88829.6896

- ## Task 1.2.B



- The **Maximum salary** in AUD for Australian respondents: 350000.00
- The **Median salary** in AUD for Australian respondents: 143500.00

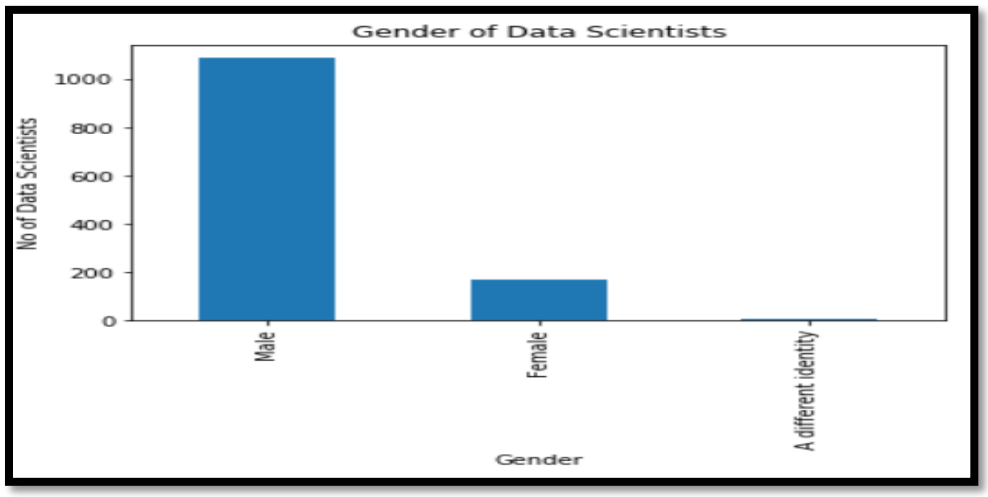
- ## Task 1.2.C

- The **Maximum salary** in AUD for Australian respondents: 250000.0
- The **Median salary** in AUD for Australian respondents is: 143500.0

Task 1.3.A

1. **What is the mean Age?**
   34
2. **What is the median Age**
   32
3. **How many Data Scientists are between 24 and 60?**
   1129
4. **How many respondents were under 18?**
   1

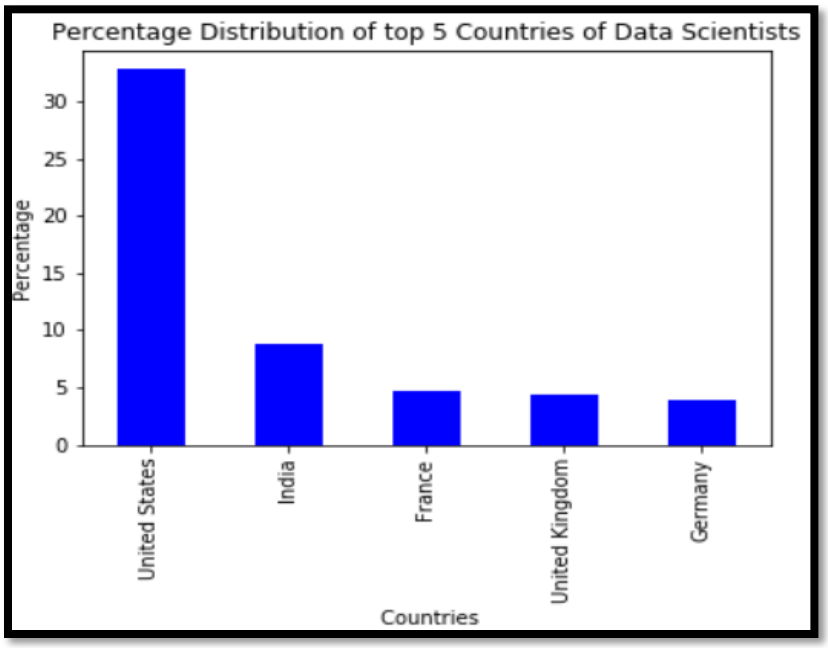
Task 1.3.B



- Task 1.3.C

Country	Number of Data Scientist
United States	414
India	111
France	60
United Kingdom	55
Germany	50

- Task 1.3.D



- Task 1.3.E

Mean and Median For United States:

- The mean age of Male Data Scientist is: 35.65 and the median is: 33.0
- The mean age of Female Data Scientist is: 33.44 and the median is: 31.0
- The mean age of Different Gender Data Scientist is: 31.0 and the median is: 31.0

Mean and Median For India:

- The mean age of Male Data Scientist in India is: 30.02 and the median is: 28.0
- The mean age of Female Data Scientist in India is: 29.0 and the median is: 27.0
- There are no Data Scientist having a different gender in India

Mean and Median For Australia:

- The mean age of Male Data Scientist in Australia is: 35.0 and the median is: 34.0
- The mean age of Female Data Scientist in Australia is: 32.6 and the median is: 31.0
- There are no Data Scientist having a different gender in Australia

Mean and Median For Pakistan:

- The mean age of Male Data Scientist in Pakistan is: 32.0 and the median is: 27.0
- There are no Data Scientist having a female gender in Pakistan
- There are no Data Scientist having a different gender in Pakistan

Task 2.1.A

```
#Reading the job posting file
df_text = pd.read_csv('JobPostings.csv')
stop_words = set(stopwords.words('english'))
tokenizer = RegexpTokenizer(r"\w+(?:[-']\w+)?") # defining the tokenizer
jobdescription=df_text['job_description']
raw_data=[]
for val in jobdescription:                    # for loop iterates through job description and appends values into raw_data
    raw_data.append(val)
raw_data=' '.join(raw_data)

# this function filters our raw_data by removing the stop words from the data
def tokenizerRawData(raw_data):
    tokens = tokenizer.tokenize(raw_data)
    tokens = [token.lower() for token in tokens]
    tokens_filtered = [token for token in tokens if token not in stop_words]
    return tokens_filtered

print("Tokenizer Created")
```

- Task 2.1.B

	words	frequency
0	ability	15686
1	across	7189
2	advanced	10627
3	algorithms	9070
4	analysis	20628
...	...	...
88	use	7574
89	using	12635
90	work	28160
91	working	13382
92	years	16235

93 rows × 2 columns

- Task 2.1.C

	words	frequency
15	data	124649
25	experience	59165
10	business	33571
90	work	28160
66	science	26875
34	learning	26867
6	analytics	21846
79	team	20729
4	analysis	20628
35	machine	20485

- Task 2.1.D

Determine that do the frequency of the most common 10 words vary between the one of the two major job posting boards Indeed and Monster. If yes what are the common words in each of them?

Discovery:

The most common words differ largely in jobs posted in both the job boards. Word “experience” is used almost 3 times more in job description of the jobs posted in Indeed than on jobs posted on Monster and the words “Analytics” is used almost 4 times more in job description of the jobs posted in Indeed than on jobs posted on Monster.

Most Common words present in job ads posted by Indeed:

	words	frequency
6895	data	72685
9324	experience	34512
4200	business	20983
26203	work	16912
21314	science	16292
14022	learning	15270
2388	analytics	13200
23630	team	13043
2367	analysis	12599
22112	skills	11588

Most Common words present in job ads posted by Monster:

	words	frequency
5399	data	18830
6897	experience	9814
9524	learning	4700
16581	work	4685
3876	business	4602
13798	science	4106
9124	job	4028
13814	scientist	3657
9852	machine	3402
2720	analytics	3080