HumanMatters_Final-Copy1

July 12, 2019

1 Data Science for Good - Jobs in LA

1.0.1 Goal

In 2020, 1/3 of the 50000 employees of the City of Los Angeles will retire. The goal of this project is to uncover biases in job postings provided by the city of L.A to help optimize recruitment and decrease unconscious discriminations.

1.0.2 Entry Data

The entry data is composed of a set of 683 job postings as text files. Each file is composed of a title, the job descrription, the requirements, the selection methods, the deadline to apply and other parts that we are going to explore.

1.0.3 Action plan

We'll be performing the following actions : ##### 1. Exploratory Data Analysis ##### 2. Uncover gender bias > Requirements length : studies show the length of requirements can discourage women from applying

3. Explore other biases by correlation analysis

Number of steps in the recruiting process Deadline for applying : is it too short ? Do the candidates have time to get aware of the job and prepare to apply ?

4. Listing suspicious Job postings

5. Text analysis

- Word cloud
- Named Entity Recognition

6. Modeling

1.0.4 1. Exploratory Data Analysis

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\techv\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\techv\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

1.a Gather all job postings into one dataframe to manipulate the data

```
'NoneType' object has no attribute 'group'
umatched sequence
'NoneType' object has no attribute 'group'
umatched sequence
'NoneType' object has no attribute 'group'
umatched sequence
'NoneType' object has no attribute 'group'
umatched sequence
'NoneType' object has no attribute 'group'
umatched sequence
'NoneType' object has no attribute 'group'
umatched sequence
'NoneType' object has no attribute 'group'
umatched sequence
'NoneType' object has no attribute 'group'
umatched sequence
```

Some attributes were not parsed but not too much apparently. Let's go further.

1.b Descriptive statistics

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 675 entries, 0 to 674
Data columns (total 27 columns):
File Name              675 non-null object
Position               675 non-null object
salary_start           675 non-null object
salary_end             575 non-null object
opendate               675 non-null datetime64[ns]
requirements           675 non-null object
duties                 675 non-null object
deadline               625 non-null object
deadline_date          625 non-null datetime64[ns]
validity_duration      625 non-null object
selection              675 non-null object
nb_lines               675 non-null object
```

```
nb_chars               675 non-null object
Essay                  675 non-null float64
Exercices              675 non-null float64
Interview              675 non-null float64
MultiChoice            675 non-null float64
OralPres               675 non-null float64
PhysicalTest           675 non-null float64
WTest                  675 non-null float64
nb_requirements        675 non-null float64
nb_selection_steps     675 non-null float64
raw_job_text           675 non-null object
EXPERIENCE_LENGTH      576 non-null object
FULL_TIME_PART_TIME    576 non-null object
EDUCATION_YEARS        122 non-null object
SCHOOL_TYPE            122 non-null object
dtypes: datetime64[ns](2), float64(9), object(16)
memory usage: 142.5+ KB
```

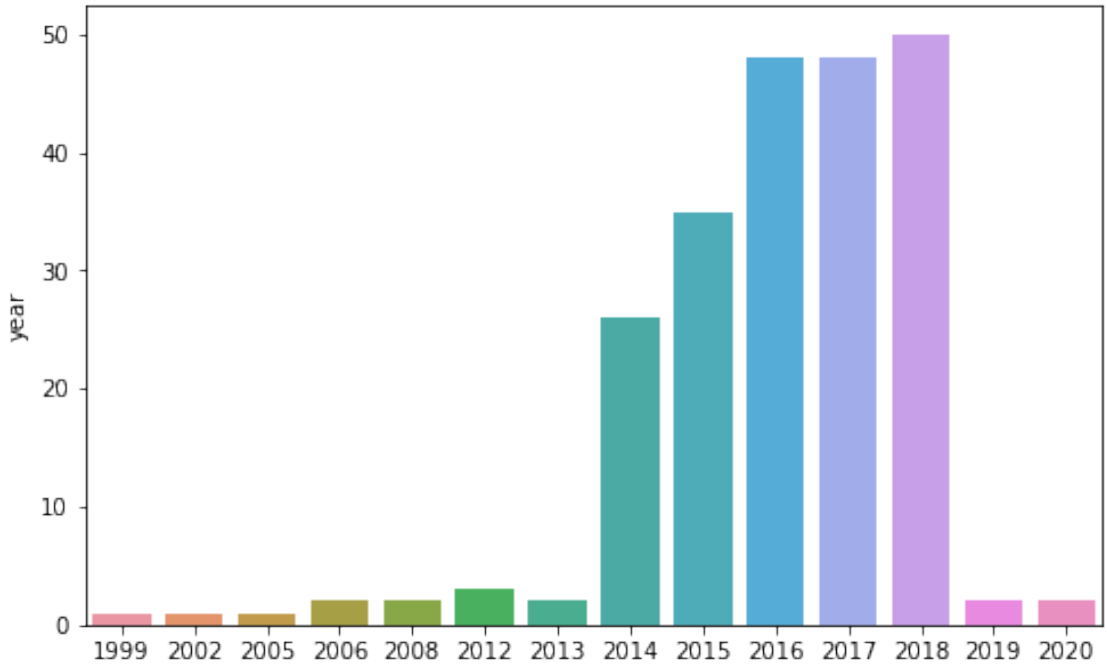
- On the 683 files we've been processing, 675 are now in our dataframe, so only a few presented a problem during parsing. We have most of them though (98%) so we can move on.
- We can notice we don't have all values for the following fields :
- salary_end
- deadline
- validity_duration
- EXPERIENCE_LENGTH
- FULL_TIME_PART_TIME
- EDUCATION_YEARS
- SCHOOL_TYPE

The two last fields especially are not often filled.
Let's look at how it looks in the dataframe :

```
Out[964]:                                        File Name  \
          0                  311 DIRECTOR  9206 041814.txt
          1                     ACCOUNTANT 1513 062218.txt
          2                ACCOUNTING CLERK 1223 071318.txt
          3  ACCOUNTING RECORDS SUPERVISOR 1119 072718.txt
          4         ADMINISTRATIVE ANALYST 1590 060118.txt

                              Position salary_start salary_end   opendate  \
          0                  311 director      125,175   $155,514 2014-04-18
          1                    accountant       49,903    $72,996 2018-06-22
          2               accounting clerk       49,005    $71,618 2018-07-13
          3  accounting records supervisor       55,332    $80,930 2018-07-27
          4         administrative analyst       60,489    $88,468 2018-06-01
```

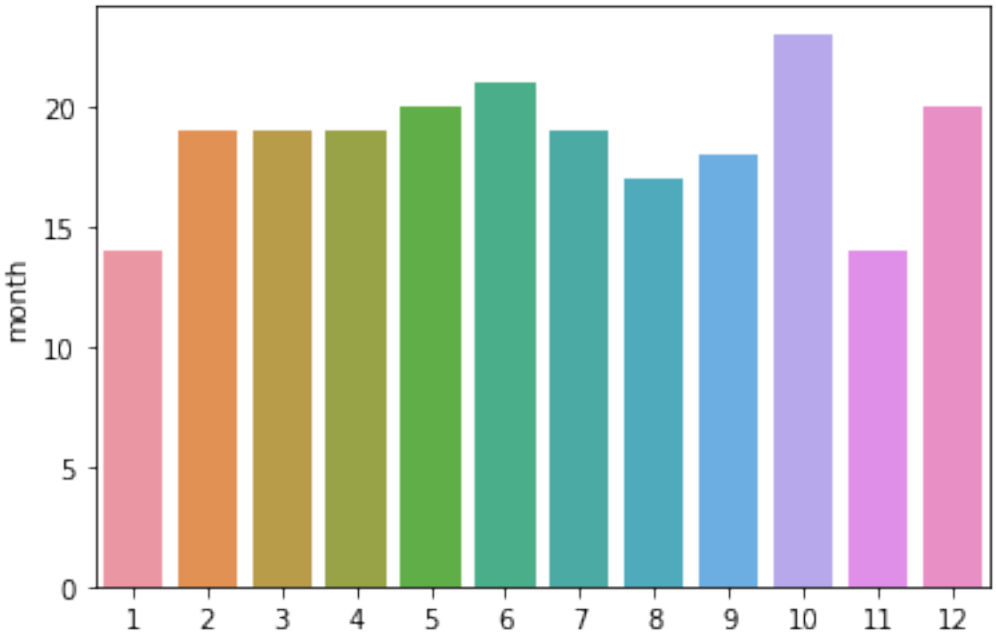
```
                                        requirements  \
0  1. One year of full-time paid experience as a ...
1  Graduation from an accredited four-year colleg...
2  Two years of full-time paid office clerical ex...
3  Two years of full-time paid experience as an A...
4  1. One year of full-time paid professional exp...

                                              duties         deadline  \
0  A 311 Director is responsible for the successf...      MAY 1, 2014
1  An Accountant does professional accounting wor...  AUGUST 25, 2018
2  An Accounting Clerk performs difficult and res...              NaN
3  An Accounting Records Supervisor assigns, revi...   AUGUST 9, 2018
4  An Administrative Analyst performs professiona...    JUNE 14, 2018

  deadline_date validity_duration  ... OralPres PhysicalTest WTest  \
0    2014-05-01                13  ...      0.0          0.0   0.0
1    2018-08-25                64  ...      0.0          0.0   1.0
2           NaT               NaN  ...      0.0          0.0   1.0
3    2018-08-09                13  ...      0.0          0.0   1.0
4    2018-06-14                13  ...      0.0          0.0   1.0

   nb_requirements  nb_selection_steps  \
0              3.0                 1.0
1              1.0                 2.0
2              1.0                 1.0
3              1.0                 2.0
4              3.0                 3.0

                                       raw_job_text  EXPERIENCE_LENGTH  \
0  311 DIRECTOR Class Code:       9206 Open Date:...                One
1  ACCOUNTANT  Class Code:       1513 Open Date: ...                NaN
2  ACCOUNTING CLERK  Class Code:       1223 Open ...                Two
3  ACCOUNTING RECORDS SUPERVISOR  Class Code:    ...                Two
4  ADMINISTRATIVE ANALYST  Class Code:       1590...                One

   FULL_TIME_PART_TIME  EDUCATION_YEARS            SCHOOL_TYPE
0            FULL_TIME              NaN                    NaN
1                  NaN             four  College or University
2            FULL_TIME              NaN                    NaN
3            FULL_TIME              NaN                    NaN
4            FULL_TIME             four  College or University

[5 rows x 27 columns]
```

1.b.1 Opendate's distribution - Job postings by year Opendate is the field indicating the date the job was posted.



- Before 2014, very few employment opportunities were offered to the citizen. As we approach 2020, we can see that the number of bulletins is increasing, there's even already job postings for 2020. There is a strong issue in managing the turnover since 2014. The number of job opportunities offred has almost doubled between 2014 and 2016, and then the number of published bulletins remain high.
- This makes our job even more challenging !

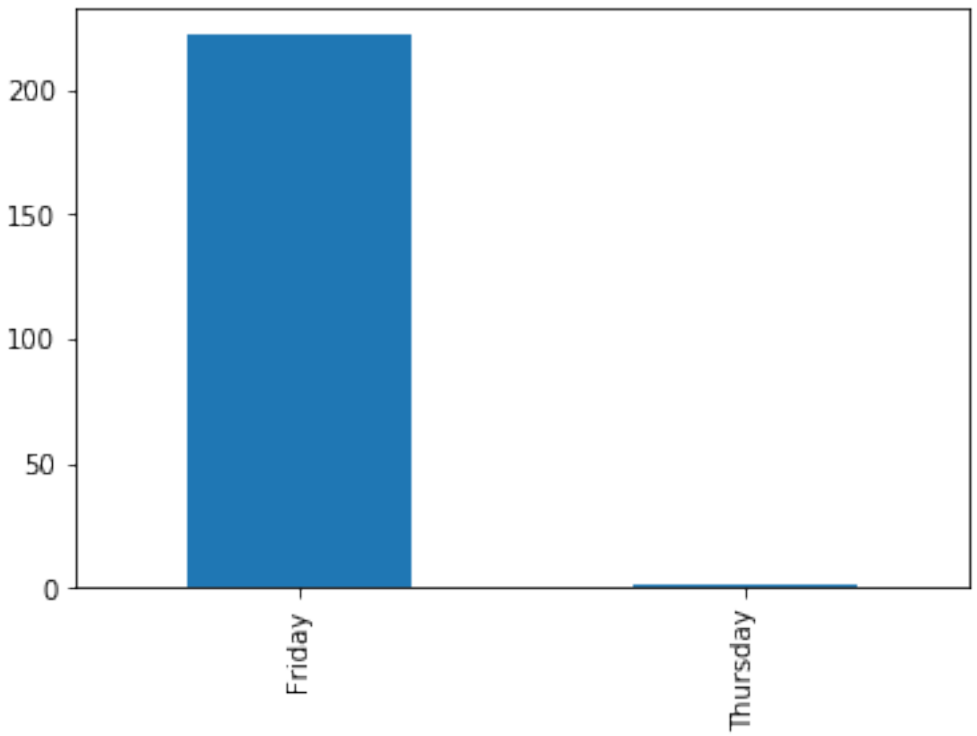
1.b.2 Job postings by month over the years



- The job openings seem about the same over the years throughout the months. January and November seem the months when there are less job openings; October on the other hand seems to be the month when most of job openings occur.
- January and Novemeber are the months with less postings.
- October concentrates more postings than other months, maybe this can be explained by the fact that it is a "back to business" period, the city assesses what is needed in september after school holidays and posts in October.
- _____ Budget decisions may be taken in November as well which leads to concentrate lots of postings in October. ____

1.b.3 Job postings by weekdays over the years When are the job posted during the week ?

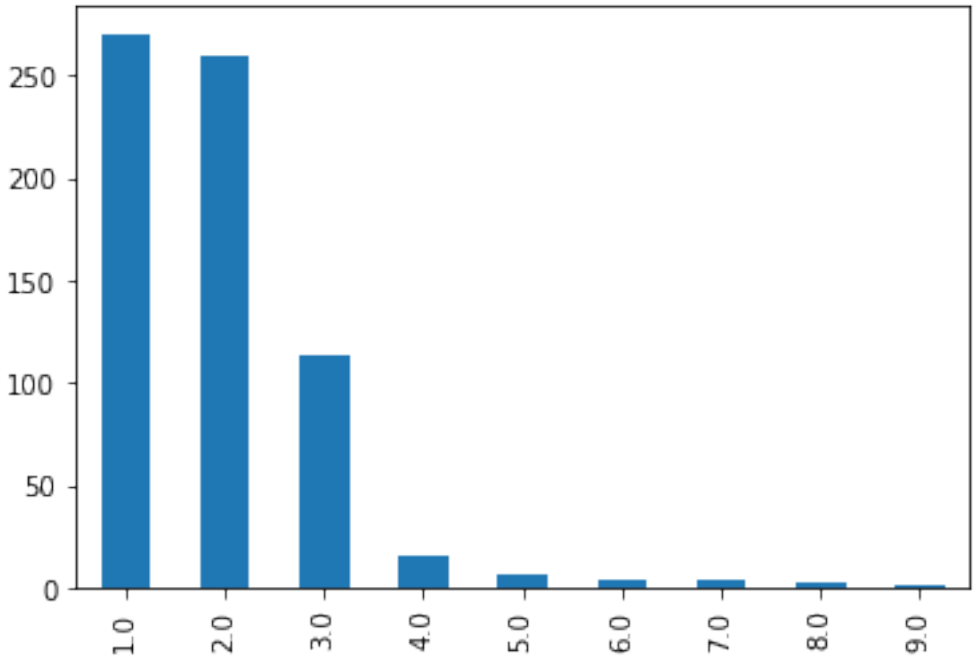
```
Out[967]: <matplotlib.axes._subplots.AxesSubplot at 0x22f97224780>
```



Apparently, almost every job openning is posted on a friday! Why is that, is it the best option ? It leaves candidates time to review them on weekends ?

1.b.4 Number of requirements specified

```
Out[968]: count    675.000000
          mean       1.924444
          std        1.080682
          min        1.000000
          25%        1.000000
          50%        2.000000
          75%        2.000000
          max        9.000000
          Name: nb_requirements, dtype: float64

Out[969]: <matplotlib.axes._subplots.AxesSubplot at 0x22f961c98d0>
```



The number of requirements can have a big impact and the more is indicated, the more female candidate can be discouraged, it may represent an unconscious bias. This parameter is to be looked up, because when confronted to a lot of requirements, A candidate can feel uncomfortable. The number of requirements should be moderate to allow more candidates to apply.

Here:

- The large majority of the bulletins indicate less that 4 requirements.
- However there are few bulletins that include more than 4 requirements and up to 9 !

What we can infer :

- Including more than 3 requirements can add to much complexity in the reading of the job posting and can be due to the intend of having a dedicated candidate, which may constitute a bias.

1.b.5 Number of steps to go through during the recruting process Let's check the different steps, what are they, how many are required and in which proportion

```
Out[970]: [Interview]                 162
          [Essay, Interview]          130
          []                           99
          [Test]                       93
          [Questionnaire]              30
          [Test, Interview]            29
          [Essay]                      25
```

```
[Test, Essay, Interview]         22
[Review]                         10
[Questionnaire, Interview]       10
[Test, Test]                      9
[Test, Questionnaire]             8
[Experience]                      6
[Evaluation]                      5
[Choice, Essay, Interview]        5
[Exercise, Interview]             4
[Written, Interview]              4
[Essay, Test, Interview]          3
[Choice, Interview]               3
[Written]                         3
[Test, Essay]                     2
[Test, Test, Test]                2
[Essay, Test]                     2
[Essay, Exercise, Interview]      2
[Test, Exercises, Interview]      1
[Choice, Test]                    1
[Written, Essay, Interview]       1
[Test, Defense]                   1
[Choice]                          1
[Interview, Essay]                1
[Abilities, Interview]            1
Name: selection, dtype: int64
```

- Several selection steps can be asked for one job (maximum 3).

Let's get a list of distinct possible selection steps

```
Out[972]: ['Experience',
           'Essay',
           'Test',
           'Written',
           'Exercises',
           'Abilities',
           'Exercise',
           'Defense',
           'Evaluation',
           'Choice',
           'Interview',
           'Questionnaire',
           'Review']
```

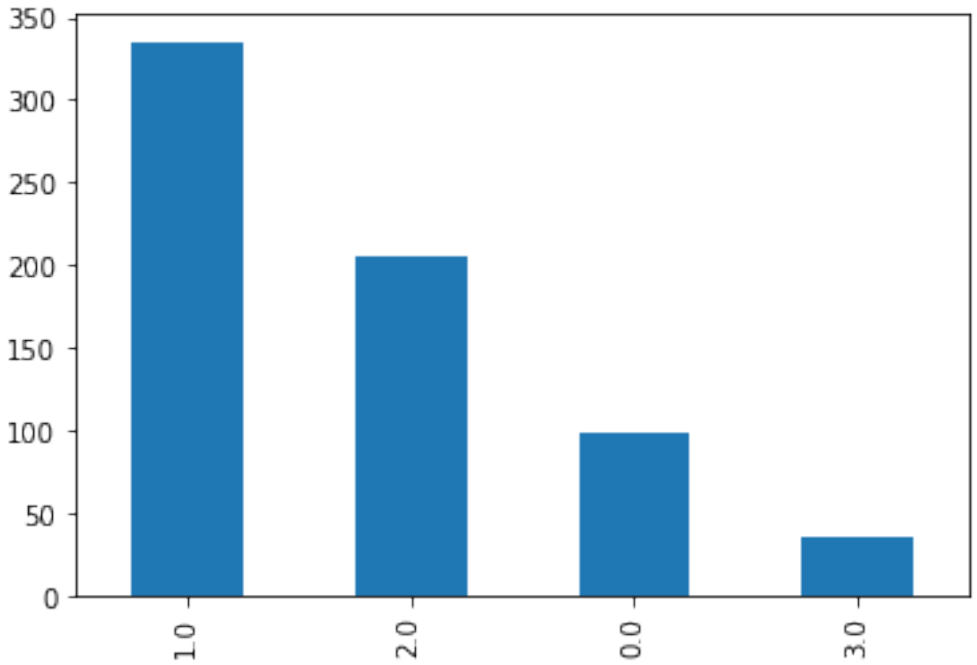
- 13 types of evaluation are possible but some of them seem weird (Abilities, Review and Defense), we'll check them later

```
Out[973]: count             675
          unique             31
```

```
top       [Interview]
freq              162
Name: selection, dtype: object
```

Out[974]:
```
count    675.000000
mean       1.263704
std        0.771392
min        0.000000
25%        1.000000
50%        1.000000
75%        2.000000
max        3.000000
Name: nb_selection_steps, dtype: float64
```

Out[975]: <matplotlib.axes._subplots.AxesSubplot at 0x22f972fa898>



Having a complex selection process may dissuade potential candidates, like disabled ones or women because of its duration and the availability required for attending each appointment. Enabling a complex selection process can be legitimate when the city wants to hire a high responsibility profile.

Here:

- There are up to 3 steps for the selection.
- This procedure helps the collectivity ensure they are hiring the appropriate candidate.
- 80% of the job opportunities include 1 or 2 steps, the most common being interview and tests.

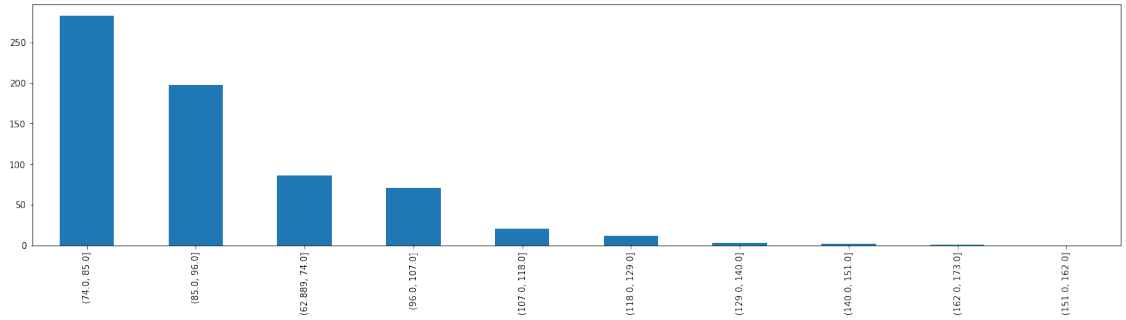
- 15% of them do not require any selection step.
- The remaning 5% of bulletins suggest a selection performed in 3 steps. Are they related to a specific kind of job ?

Next steps: By intituion, we would say that a 3-step selection process should be reserved to high responsibility position, where a hiring mistake can have strong impacts on the organzation. We then need to look for a correlation between the number of selection steps and the responsibility level.

1.b.6 Number of lines in the job description

```
Out[976]: count     675
          unique     63
          top        83
          freq       31
          Name: nb_lines, dtype: int64
```

```
Out[977]: <matplotlib.axes._subplots.AxesSubplot at 0x22f9735e4a8>
```



Analysis of the number of lines in the job description Having a long description may be interesting for high responsibility positions in order to provide sufficient context elements on the job offer and the performance of the work. However a long bulletin can dissuade potential candidates to apply because the text would be too long.

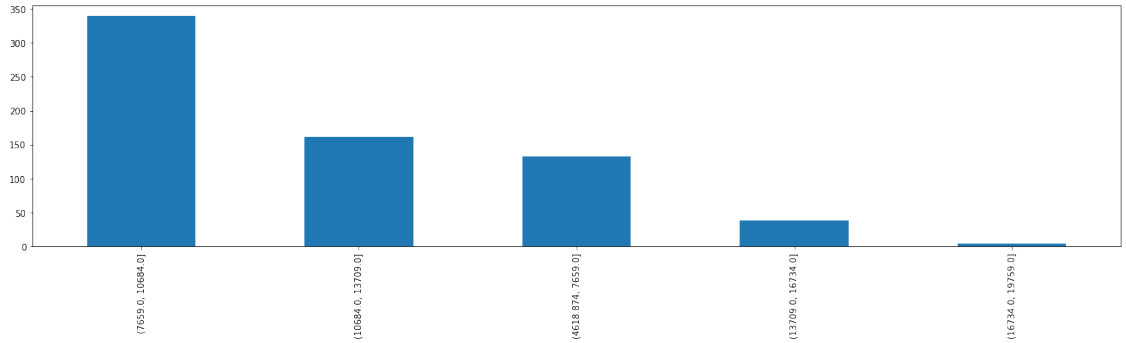
Here:

- Most of the job postings include less than 100 lines.
- The vast majority is between 74 and 96 lines

Next steps: We will list the positions according to a scale of responibility and check if a long job description is legitimato or not. if it is not the case, maybe is it due to an unconsious biais. We will use a scale from 1 to 5

1.b.7 Number of caracters in the job description

```
Out[978]: count      675
          unique     649
          top       8215
          freq         3
          Name: nb_chars, dtype: int64

Out[979]: (7659.0, 10684.0]     339
          (10684.0, 13709.0]    161
          (4618.874, 7659.0]    133
          (13709.0, 16734.0]     38
          (16734.0, 19759.0]      4
          Name: nb_chars, dtype: int64

Out[980]: <matplotlib.axes._subplots.AxesSubplot at 0x22f973e1470>
```



Analysis Having a long description may be interesting for high responsibility positions in order to provide sufficient context elements on the job offer anr the performance of the work. However a charged (in terms of chars) can dissuade potential candidates to apply because the text would be too complex.

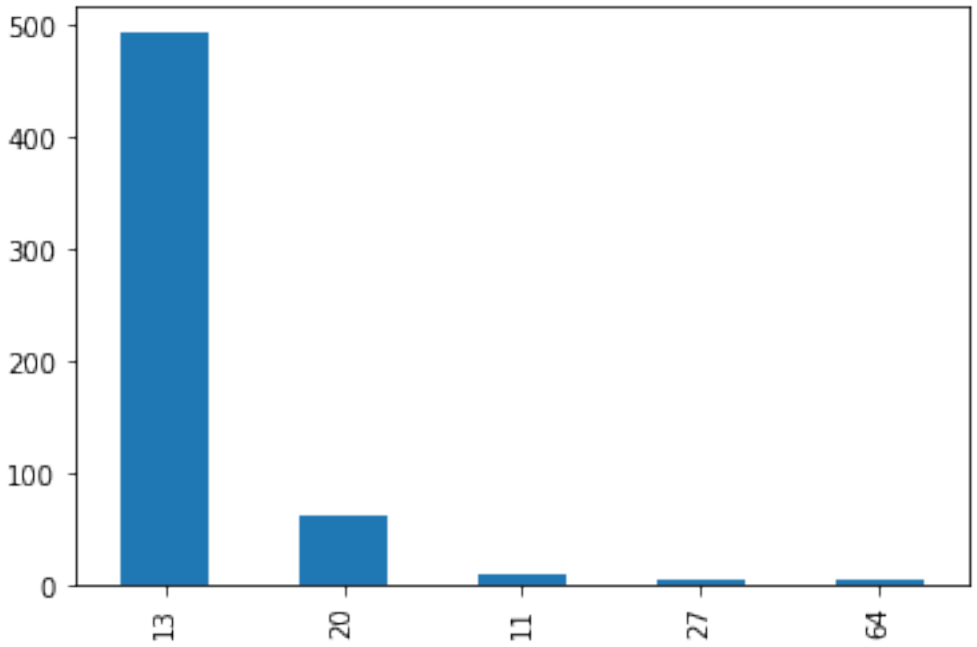
Here:

- Most of the job postings (527 of them i.e 78%) include 4.600 to 13.700 caracters.
- the remaining 148 bulletins (about 22%) may be too 'verbose'.

Next steps: We will list the positions according to a scale of responibility and check if a verbose description is legitimate or not. if it is not the case, maybe is it due to an unconsious biais. We will use a scale from 1 to 5

1.b.8 Deadline - Time to apply validity_duration field has been computed to give us the time between the date the job was posted and the deadline to apply.

```
Out[981]: count      625
          unique      38
          top         13
          freq       494
          Name: validity_duration, dtype: int64

Out[982]: <matplotlib.axes._subplots.AxesSubplot at 0x22f974466d8>
```



Analysis of the time remaining to apply Validity duration is an important parameter. Offering little time to apply reduces the number of candidates . We can expect a job opening to leave enough time to people to apply. Maybe more for disabled people who may have access to the job openings less easily. We can also expect the deadline to be extended for rare profiles like those intended for high responsibility positions.

If a low validity duration is given for a high responsibility position or for a position open to all, it can represent a barrier to the City to meet interesting external candidates. We should here offer a prescriptive action to the City.

Here

- Most of the job opportunities (about 65%) are to be applied within 13 days, equivalent to 2 weeks since the release of the bulletin.
- The next most common validity duration is 20 days equivalent to 3 weeks.
- Up to 10 bulletins offer a validity duration of 10 days, which is rather short. This is the shortest validity duration.

Next steps: We have to explore those bulletins with 11-day validity duration, check if they are opent to all and check if the position leads to high responibilities. We will foculs on 11-day validity bulletin because this is the shortest duration, and a 13-day validity duration is to common.

1.1 Descriptive Analysis Summary

File parsing performance

- Over the 683 files we managed to keep 675 of them after parsing (98%).

Offered employment

- Before 2014, very few employment opportunities were offered to the citizen. As we approch 2021, we can see that the number of bulletins is increasing. There is a strong issue in managing the turnover since 2014. The niumber of job opportunities offred is almost doubled between 2014 and 2016, and then the number of published bulletins remain high.
- This makes our job even more challenging !

Job posting all over the year

- It seems that about the same amount of jobs have been posted every month throughout the years.

Number of requirements

- In large majority, the bulletins indicate less that 4 requirements.
- However thet are few bulletins that include more requirements even up to 9 !
- Including lots of requirements may have a negative impact on female applications and therefore be part of an unconscious bias. This parameter is to be looked into.

Number of selection steps

- There are up to 3 steps for the selection.
- This procedure helps the collectivity ensure they are hiring the appropriate candidate.
- 80% of the job opportunities include 1 or 2 steps, the most common being interview, essay and test.
- 15% of them do not require a complex selection process.
- The remaning 5% of bulletins suggest a selection performed in 3 steps. Are they related to a specific kind of job ?
- Having a 3-step selection process may dissuade potential candidates, like disabled ones or women because of its duration and the availability required for attending each appointment.

Validity duration

- Most (about 65%) of the job opportunities are to be applied within 13 days equivalent to 2 weeks since publishing of the bulletin.
- The next validity duration is 20 days equivalent to 3 weeks.
- Up to 10 bulletin offer a validity duration of 10 days, which is rather short. This is the shortest validity duration.

- Validity duration is an important parameter, offering little time to apply may reduce the number of candidates. We also need to identify if some job positions are only opened to current employees, which can explain why validity duration is short.

Next steps Correlate Responsibility level with : - Validity duration, - Nb of requirements in the job description - Nb steps in the selection process - Nb lines in the job description - Nb chars in the job description

1.1.1 1.d Feature engineering - Enriching the dataframe with computed fields

In this section we will enhance our dataframe with additional computed fields: - nb_line_scale : number of lines on a scale from 0 to 5 - nb_char_scale : number of chars on a scale from 0 to 5 - full_time_part_time_code : indicates if job is part time (1) or full time (2) - exp_years : number of years of experience needed - high_education : 1 if requiring University or College, 0 else - Open_To_All : indicates if the position is open to all including actual city employees - Resp_level : scale of responsibility from 0 to 5

Resp_level based on the job title : - Director = 5, - Manager, Principal, Chief, Captain = 4, - Engineer, Specialist, Representative, Advocate, Inspector, Supervisor = 3 - Officer = 2 - Other = 0

Out[983]:
```
count                675
unique               665
top       annual salary
freq                   3
Name: Position, dtype: object
```

Out[986]:

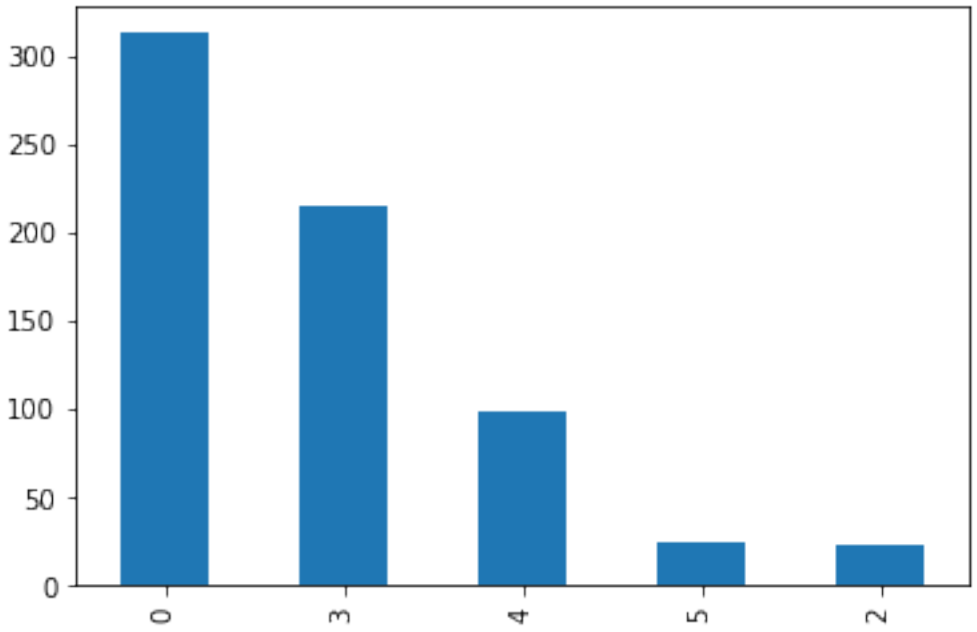
	File Name	Position	salary_start	salary_end	opendate	requirements	duties	deadline
0	311 DIRECTOR 9206 041814.txt	311 director	125,175	$155,514	2014-04-18	1. One year of full-time paid experience as a ...	A 311 Director is responsible for the successf...	MAY 1, 2014
1	ACCOUNTANT 1513 062218.txt	accountant	49,903	$72,996	2018-06-22	Graduation from an accredited four-year colleg...		
2	ACCOUNTING CLERK 1223 071318.txt	accounting clerk	49,005	$71,618	2018-07-13	Two years of full-time paid office clerical ex...		
3	ACCOUNTING RECORDS SUPERVISOR 1119 072718.txt	accounting records supervisor	55,332	$80,930	2018-07-27	Two years of full-time paid experience as an A...		
4	ADMINISTRATIVE ANALYST 1590 060118.txt	administrative analyst	60,489	$88,468	2018-06-01	1. One year of full-time paid professional exp...		

```
1  An Accountant does professional accounting wor...  AUGUST 25, 2018
2  An Accounting Clerk performs difficult and res...              NaN
3  An Accounting Records Supervisor assigns, revi...   AUGUST 9, 2018
4  An Administrative Analyst performs professiona...    JUNE 14, 2018

  deadline_date validity_duration  ... EDUCATION_YEARS            SCHOOL_TYPE  \
0    2014-05-01                13  ...             NaN                    NaN
1    2018-08-25                64  ...            four  College or University
2           NaT               NaN  ...             NaN                    NaN
3    2018-08-09                13  ...             NaN                    NaN
4    2018-06-14                13  ...            four  College or University

  nb_line_scale  nb_char_scale  Resp_level  Open_To_All  Open_To_Mention  \
0             1              1           5            1                0
1             2              1           0            1                1
2             2              1           0            1                1
3             2              1           3            0                1
4             2              3           0            0                1

   exp_years  high_education  full_time_part_time_code
0        1.0               0                         2
1        NaN               1                         0
2        2.0               0                         2
3        2.0               0                         2
4        1.0               1                         2

[5 rows x 35 columns]
```

Out[987]:
```
count    675.000000
mean       1.795556
std        1.748160
min        0.000000
25%        0.000000
50%        3.000000
75%        3.000000
max        5.000000
Name: Resp_level, dtype: float64
```

Out[988]:
```
0    313
3    215
4     99
5     25
2     23
Name: Resp_level, dtype: int64
```

Out[989]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22f974465c0>
```

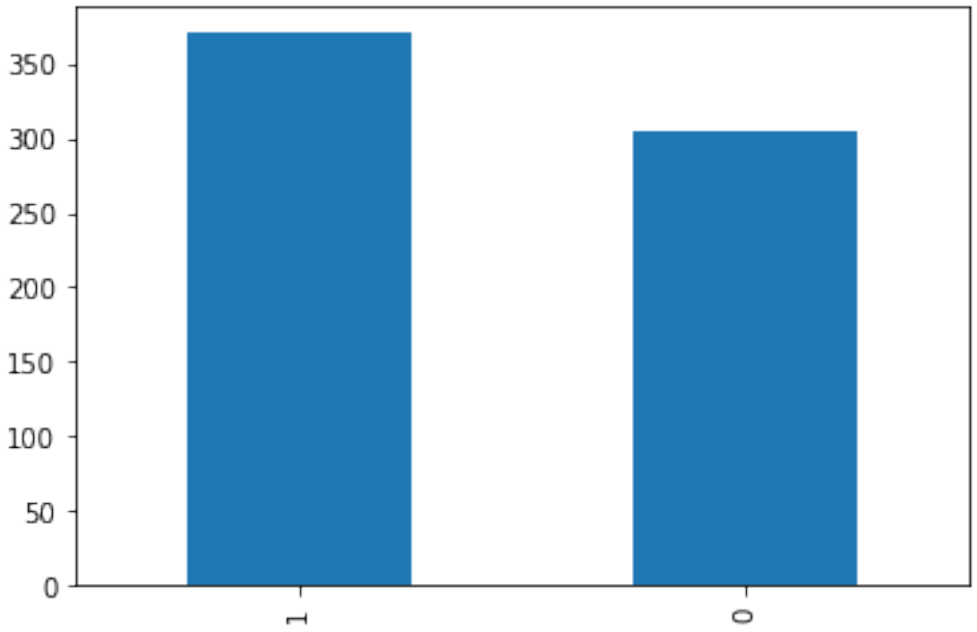


1.d.1 Responsibility Level

- Most (about 50%) of the job opportunities are very low responsibility levels (levels 0 to 2)
- About 46% of the job positions present a medium or high responibility level (level 3 and 4).
- About 4% of the job positions deal with very high responsibility (level 5)

Our feeling is that the should not have displayed biais especially for medium to very high responsibility levels. If there is a bias, for instance en gender bias, it should be considered as a critical one.

```
Out[990]: count    675.000000
          mean       0.548148
          std        0.498045
          min        0.000000
          25%        0.000000
          50%        1.000000
          75%        1.000000
          max        1.000000
          Name: Open_To_All, dtype: float64

Out[991]: 1    370
          0    305
          Name: Open_To_All, dtype: int64

Out[992]: <matplotlib.axes._subplots.AxesSubplot at 0x22f9751fc50>
```



1.d.2 Open to All

- About 370 Job bulletins (about 55%) are open to all kind of candidates including already city employees.
  - About 305 Job bulletins (about 50%) clearly specify the are open to all.
  - About 35 Job bulletins (about 5%) do not specify whether the job is opent to external cadidates.
- About 305 Job bulletins (about 45%) are only open to current city employees

This is huge, only 50% of the job postings are open to every candidate, this may reduce chances to have new candidates.

```
Out[993]: 2    576
          0     99
          Name: full_time_part_time_code, dtype: int64
```

1.d.3 Part time or full time ?

- The very large most of the position are specifically indicated as open in Full time (about 85%)
- The remaining bulletins (99 of them) DO NOT specify if they are open to PART_TIME or NOT. Wee assume that they are fulltime.

```
Out[994]: 0    553
          1    122
          Name: high_education, dtype: int64
```

1.d.4 High eductation or Not ?

- Only 122 job position require a college or university education (about 18%)
- The remaining bulletins DO NOT specify anything. We will assume that they do don't require it.

1.d.5 Salaries analysis

Out[995]:

	File Name	Position	salary_start	salary_end	opendate	requirements	duties	deadline	deadline_date	validity_duration	...	EDUCATION_YEARS	SCHOOL_TYPE	nb_line_scale	nb_char_scale	Resp_level	Open_To_All	Open_To_Mention
0	311 DIRECTOR 9206 041814.txt	311 director	125175	$155,514	2014-04-18	1. One year of full-time paid experience as a ...	A 311 Director is responsible for the successf...	MAY 1, 2014	2014-05-01	13	...	NaN	NaN	1	1	5	1	0
1	ACCOUNTANT 1513 062218.txt	accountant	49903	$72,996	2018-06-22	Graduation from an accredited four-year colleg...	An Accountant does professional accounting wor...	AUGUST 25, 2018	2018-08-25	64	...	four	College or University	2	1	0	1	1
2	ACCOUNTING CLERK 1223 071318.txt	accounting clerk	49005	$71,618	2018-07-13	Two years of full-time paid office clerical ex...	An Accounting Clerk performs difficult and res...	NaN	NaT	NaN	...	NaN	NaN	2	1	0	1	1
3	ACCOUNTING RECORDS SUPERVISOR 1119 072718.txt	accounting records supervisor	55332	$80,930	2018-07-27	Two years of full-time paid experience as an A...	An Accounting Records Supervisor assigns, revi...	AUGUST 9, 2018	2018-08-09	13	...	NaN	NaN	2	1	3	0	1
4	ADMINISTRATIVE ANALYST 1590 060118.txt	administrative analyst	60489	$88,468	2018-06-01	1. One year of full-time paid professional exp...	An Administrative Analyst performs professiona...	JUNE 14, 2018	2018-06-14	13	...	four	College or University					

```
4            2              3          0           0                1

   exp_years  high_education  full_time_part_time_code
0        1.0               0                         2
1        NaN               1                         0
2        2.0               0                         2
3        2.0               0                         2
4        1.0               1                         2

[5 rows x 35 columns]
```

```
Out[996]: (72938.0, 116143.0]     304
          (29516.974, 72938.0]    299
          (116143.0, 159348.0]     66
          (159348.0, 202553.0]      5
          (202553.0, 245758.0]      1
          Name: salary_start, dtype: int64
```

- Let's encode the salary ranges to add that variable to the model

```
Out[999]:                                    File Name  \
          0                  311 DIRECTOR  9206 041814.txt
          1                     ACCOUNTANT 1513 062218.txt
          2                ACCOUNTING CLERK 1223 071318.txt
          3  ACCOUNTING RECORDS SUPERVISOR 1119 072718.txt
          4          ADMINISTRATIVE ANALYST 1590 060118.txt

                              Position  salary_start salary_end   opendate  \
          0                  311 director        125175   $155,514 2014-04-18
          1                    accountant         49903    $72,996 2018-06-22
          2               accounting clerk         49005    $71,618 2018-07-13
          3  accounting records supervisor         55332    $80,930 2018-07-27
          4          administrative analyst         60489    $88,468 2018-06-01

                                              requirements  \
          0  1. One year of full-time paid experience as a ...
          1  Graduation from an accredited four-year colleg...
          2  Two years of full-time paid office clerical ex...
          3  Two years of full-time paid experience as an A...
          4  1. One year of full-time paid professional exp...

                                                    duties         deadline  \
          0  A 311 Director is responsible for the successf...      MAY 1, 2014
          1  An Accountant does professional accounting wor...  AUGUST 25, 2018
          2  An Accounting Clerk performs difficult and res...              NaN
          3  An Accounting Records Supervisor assigns, revi...   AUGUST 9, 2018
          4  An Administrative Analyst performs professiona...    JUNE 14, 2018
```

```
  deadline_date validity_duration  ...            SCHOOL_TYPE nb_line_scale  \
0    2014-05-01                13  ...                    NaN             1
1    2018-08-25                64  ...  College or University             2
2           NaT               NaN  ...                    NaN             2
3    2018-08-09                13  ...                    NaN             2
4    2018-06-14                13  ...  College or University             2

  nb_char_scale  Resp_level  Open_To_All  Open_To_Mention  exp_years  \
0             1           5            1                0        1.0
1             1           0            1                1        NaN
2             1           0            1                1        2.0
3             1           3            0                1        2.0
4             3           0            0                1        1.0

   high_education  full_time_part_time_code  salary_start_code
0               0                         2                  3
1               1                         0                  1
2               0                         2                  1
3               0                         2                  1
4               1                         2                  1

[5 rows x 36 columns]
```

Out[1000]:
```
0    155514
1     72996
2     71618
3     80930
4     88468
Name: salary_end_strip, dtype: object
```

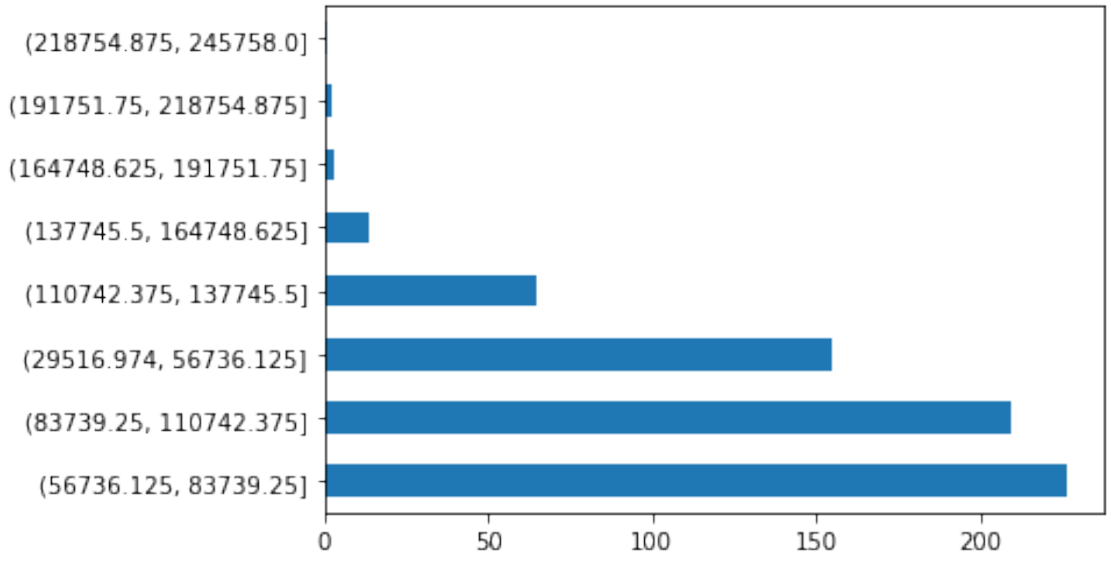
Out[1001]:
```
0    30339.0
1    23093.0
2    22613.0
3    25598.0
4    27979.0
Name: salary_variation, dtype: float64
```

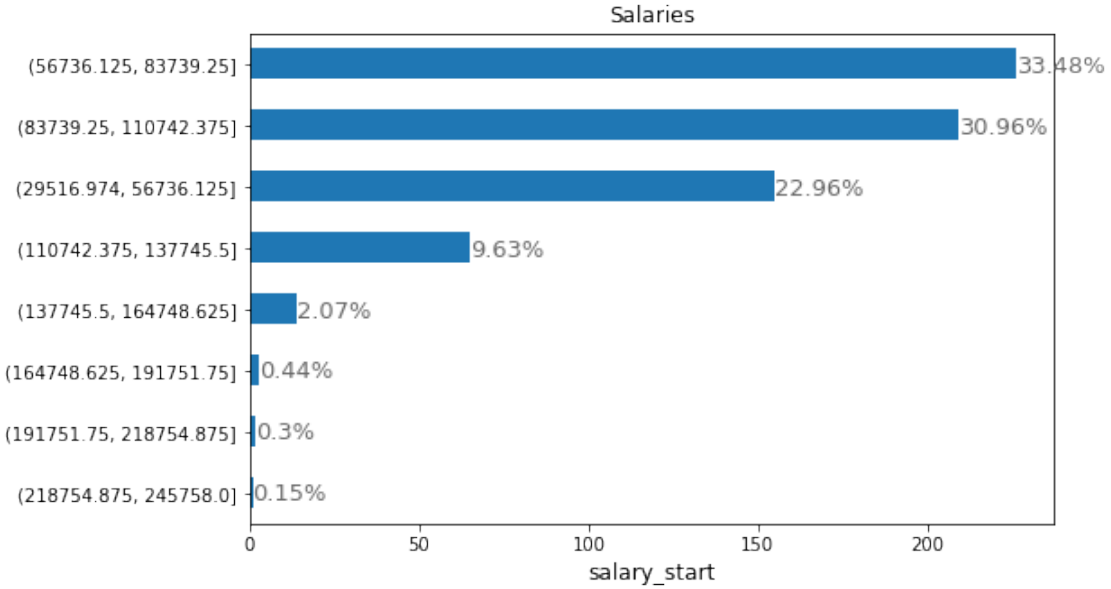
Out[1002]:
```
(17063.4, 31620.8]     265
(31620.8, 46178.2]     137
(2433.212, 17063.4]    120
(46178.2, 60735.6]      52
(60735.6, 75293.0]       1
Name: salary_variation, dtype: int64
```

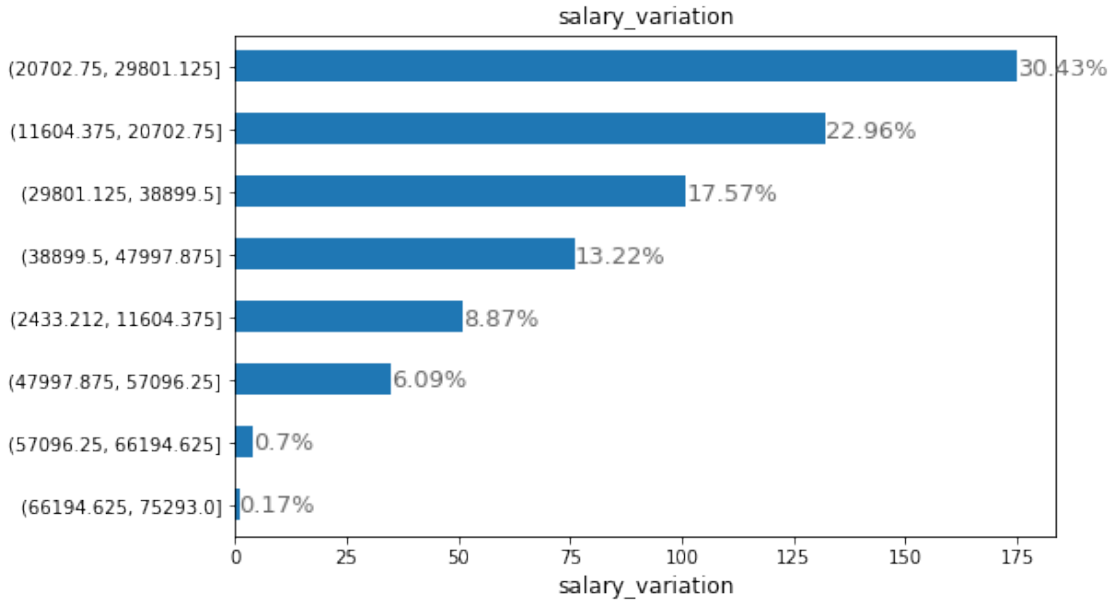
Out[1003]:
```
                                    File Name  \
0                 311 DIRECTOR  9206 041814.txt
1                    ACCOUNTANT 1513 062218.txt
2               ACCOUNTING CLERK 1223 071318.txt
3  ACCOUNTING RECORDS SUPERVISOR 1119 072718.txt
```

```
4         ADMINISTRATIVE ANALYST 1590 060118.txt

                      Position  salary_start salary_end   opendate  \
0                  311 director        125175   $155,514 2014-04-18
1                    accountant         49903    $72,996 2018-06-22
2               accounting clerk         49005    $71,618 2018-07-13
3  accounting records supervisor         55332    $80,930 2018-07-27
4         administrative analyst         60489    $88,468 2018-06-01

                                      requirements  \
0  1. One year of full-time paid experience as a ...
1  Graduation from an accredited four-year colleg...
2  Two years of full-time paid office clerical ex...
3  Two years of full-time paid experience as an A...
4  1. One year of full-time paid professional exp...

                                           duties         deadline  \
0  A 311 Director is responsible for the successf...      MAY 1, 2014
1  An Accountant does professional accounting wor...  AUGUST 25, 2018
2  An Accounting Clerk performs difficult and res...              NaN
3  An Accounting Records Supervisor assigns, revi...   AUGUST 9, 2018
4  An Administrative Analyst performs professiona...    JUNE 14, 2018

  deadline_date validity_duration  ... Resp_level Open_To_All Open_To_Mention  \
0    2014-05-01                13  ...          5           1               0
1    2018-08-25                64  ...          0           1               1
2           NaT               NaN  ...          0           1               1
3    2018-08-09                13  ...          3           0               1
4    2018-06-14                13  ...          0           0               1

   exp_years  high_education  full_time_part_time_code  salary_start_code  \
0        1.0               0                         2                  3
1        NaN               1                         0                  1
2        2.0               0                         2                  1
3        2.0               0                         2                  1
4        1.0               1                         2                  1

   salary_end_strip  salary_variation  salary_variation_code
0            155514           30339.0                      3
1             72996           23093.0                      2
2             71618           22613.0                      2
3             80930           25598.0                      3
4             88468           27979.0                      3

[5 rows x 39 columns]
```

Out[1004]: <matplotlib.axes._subplots.AxesSubplot at 0x22f97589780>


(218754.875, 245758.0]
(191751.75, 218754.875]
(164748.625, 191751.75]
(137745.5, 164748.625]
(110742.375, 137745.5]
(29516.974, 56736.125]
(83739.25, 110742.375]
(56736.125, 83739.25]
0
50
100
150
200


Salaries
(56736.125, 83739.25]
(83739.25, 110742.375]
(29516.974, 56736.125]
(110742.375, 137745.5]
(137745.5, 164748.625]
(164748.625, 191751.75]
(191751.75, 218754.875]
(218754.875, 245758.0]
33.48%
30.96%
22.96%
9.63%
2.07%
0.44%
0.3%
0.15%
0
50
100
150
200
salary_start




```
Out[1008]: count      575.000000
           mean     27491.669565
           std      12717.306214
           min       2506.000000
           25%      18311.000000
           50%      25891.000000
           75%      36634.500000
           max      75293.000000
           Name: salary_variation, dtype: float64
```

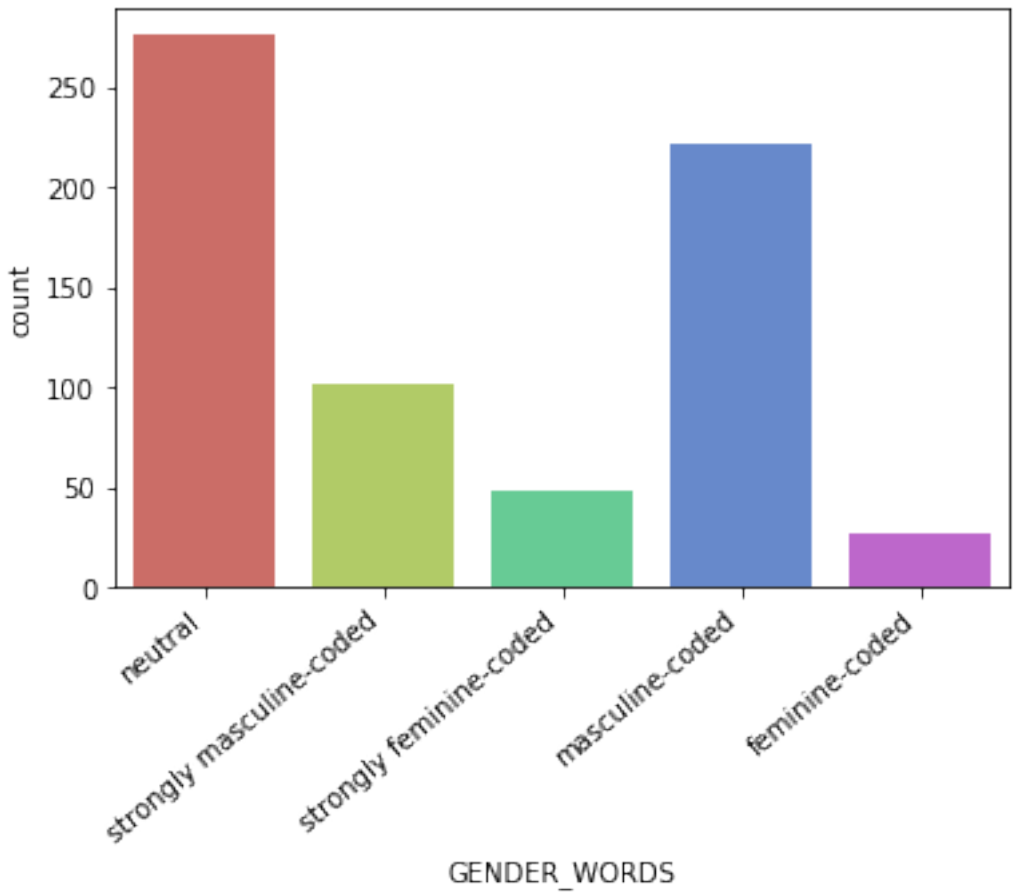
1.1.2 Analysis

- The majority of job postings offer a salary between 56K$ and 110K$
- Most jobs have a pretty big amplitude btw the entry salary and the end salary offered

1.2 2. Gender bias analysis

Let's assign a "tendency" to each job posting based on the following paper : https://www.hw.ac.uk/services/docs/gendered-wording-in-job-ads.pdf

Are there any indication of a gender bias in the duties part ?




```
Out[1011]: neutral                    276
           masculine-coded            222
           strongly masculine-coded   102
           strongly feminine-coded     48
           feminine-coded              27
           Name: GENDER_WORDS, dtype: int64
```

2.a Gender tendency analysis There is an insight here !

- 41% of the bulletins are masculine coded including
- 33% are masculine coded
- 8% are strongly masculine coded
- Only 11 % are feminine or strongly feminine coded

So the job postings are three times more inclined towards masculine words than feminine.

Next steps:

- We need to enrich the dataset with the tendancy separetedly masculine or feminine

Out[1014]:

	File Name	Position	salary_start	salary_end	opendate
0	311 DIRECTOR 9206 041814.txt	311 director	125175	$155,514	2014-04-18
1	ACCOUNTANT 1513 062218.txt	accountant	49903	$72,996	2018-06-22
2	ACCOUNTING CLERK 1223 071318.txt	accounting clerk	49005	$71,618	2018-07-13
3	ACCOUNTING RECORDS SUPERVISOR 1119 072718.txt	accounting records supervisor	55332	$80,930	2018-07-27
4	ADMINISTRATIVE ANALYST 1590 060118.txt	administrative analyst	60489	$88,468	2018-06-01

	requirements	duties	deadline
0	1. One year of full-time paid experience as a ...	A 311 Director is responsible for the successf...	MAY 1, 2014
1	Graduation from an accredited four-year colleg...	An Accountant does professional accounting wor...	AUGUST 25, 2018
2	Two years of full-time paid office clerical ex...	An Accounting Clerk performs difficult and res...	NaN
3	Two years of full-time paid experience as an A...	An Accounting Records Supervisor assigns, revi...	AUGUST 9, 2018
4	1. One year of full-time paid professional exp...	An Administrative Analyst performs professiona...	JUNE 14, 2018

	deadline_date	validity_duration	...	full_time_part_time_code
0	2014-05-01	13	...	2
1	2018-08-25	64	...	0
2	NaT	NaN	...	2
3	2018-08-09	13	...	2
4	2018-06-14	13	...	2

	salary_start_code	salary_end_strip	salary_variation	salary_variation_code
0	3	155514	30339.0	3
1	1	72996	23093.0	2
2	1	71618	22613.0	2
3	1	80930	25598.0	3
4	1	88468	27979.0	3

	GENDER_WORDS	Too_Feminine	Too_Masculine	gender_bias
0	neutral	0	0	0
1	strongly masculine-coded	0	1	2
2	neutral	0	0	0
3	neutral	0	0	0
4	strongly masculine-coded	0	1	2

```
   gender_score
0             0
1             4
2             0
3             0
4             4

[5 rows x 44 columns]
```

The enriching with the gender tendency is over

1.3 4. What are the bulletins that require immediate action to reduce unconscious biases ?

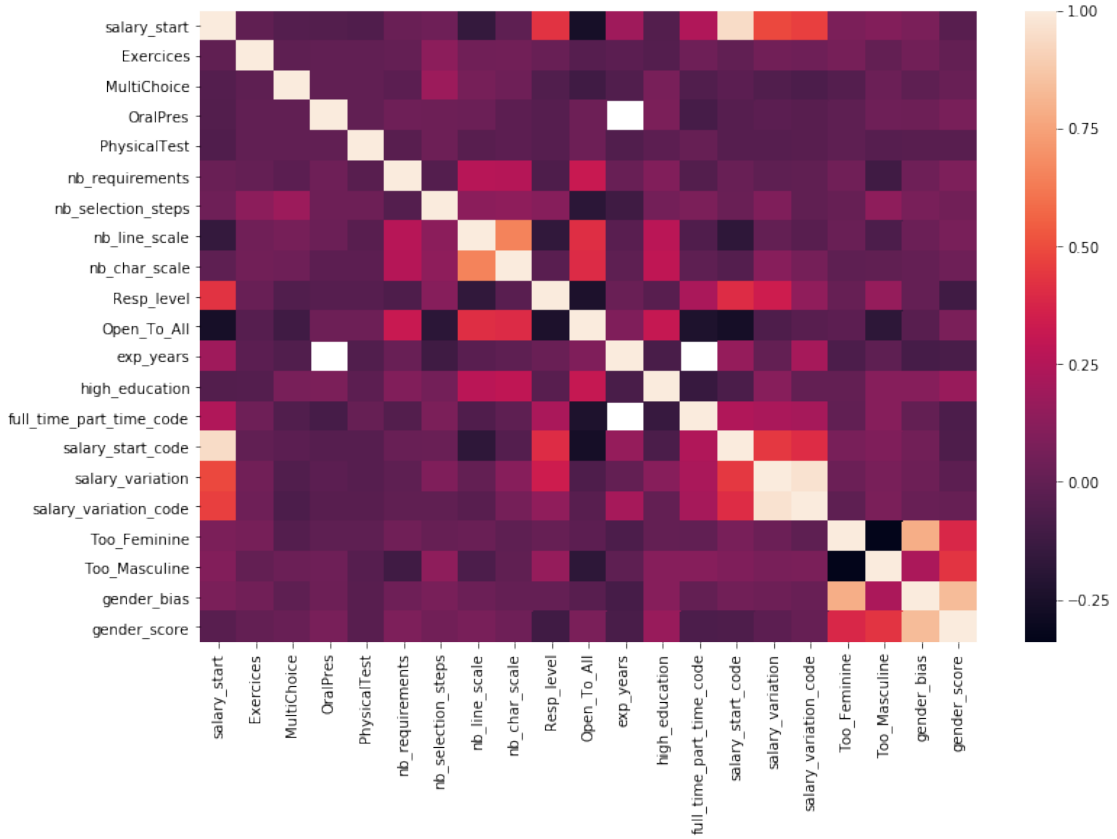
For this, we will look for bulletins with high masculine or high feminine coded language and check their responsibility level as well as the complexity of the selection, and the validity duration

4.a Interesting correlations In this section, we will check if there are 'unlegitimate' correlation between - responisibility level, - vadility duration - nb_lines_scale (in the text description) - Nb_chars_scale - nb_requirement - nb_selection steps, - Open_to_All - exp_years - full_time_part_time_code - high_education - toomasculine - toofeminine

Out[1015]:

```
                                      File Name  \
0                 311 DIRECTOR  9206 041814.txt
1                    ACCOUNTANT 1513 062218.txt
2               ACCOUNTING CLERK 1223 071318.txt
3  ACCOUNTING RECORDS SUPERVISOR 1119 072718.txt
4         ADMINISTRATIVE ANALYST 1590 060118.txt

                         Position  salary_start salary_end validity_duration  \
0                  311 director         125175   $155,514                13
1                    accountant          49903    $72,996                64
2               accounting clerk          49005    $71,618               NaN
3  accounting records supervisor          55332    $80,930                13
4         administrative analyst          60489    $88,468                13

   nb_lines nb_chars  Exercices  MultiChoice  OralPres  ...  \
0        72     9781        0.0          0.0       0.0  ...
1        90    10260        0.0          0.0       0.0  ...
2        88    10291        0.0          0.0       0.0  ...
3        87     7821        0.0          0.0       0.0  ...
4        93    11451        0.0          0.0       0.0  ...

   full_time_part_time_code  salary_start_code  salary_end_strip  \
0                         2                  3            155514
1                         0                  1             72996
2                         2                  1             71618
```

```
3                        2                 1              80930
4                        2                 1              88468

   salary_variation  salary_variation_code             GENDER_WORDS  \
0           30339.0                      3                  neutral
1           23093.0                      2  strongly masculine-coded
2           22613.0                      2                  neutral
3           25598.0                      3                  neutral
4           27979.0                      3  strongly masculine-coded

   Too_Feminine  Too_Masculine  gender_bias  gender_score
0             0              0            0             0
1             0              1            2             4
2             0              0            0             0
3             0              0            0             0
4             0              1            2             4

[5 rows x 29 columns]
```



4.b Interesting correlations summary

Open_To_All VS nb_requirements

No straight correlation ==> Difficult to suspect a biais on this situation

high_education VS Open_To_All

No straight correlation ==> Difficult to suspect a biais on this situation

Resp_level VS tooMasculine and tooFeminine :

There is very little (0.25) trend that responsibility level can correlate with masculine coded. The opposite trend is obtained for feminine coded (-0.25)

Resp_Level VS nb_line_scale

At the opposite to what we could expect there is not a strait correlation between the nuber of lines in the job description and the responsibility level.

Resp_Level VS nb_char_scale

At the opposite to what we could expect there is not a straight correlation between the nuber of chars in the job description and the responsibility level.

Resp_Level VS nb_char_scale

At the opposite to what we could expect there is not a straight correlation between the nuber of chars in the job description and the responsibility level.

Resp_Level VS nb_selection_steps

There is very little (0.25) trend that responsibility level can correlate with the number of selection steps which makes sense.

Resp_Level VS nb_requirements

No straight correlation ==> Difficult to suspect a biais on this situation.

Resp_Level VS : other observations

There is also 1 little correlation betweel Resp_Level and Fulltime job. Indeed, Responibility jobs require full time. We can also be surprised by the fact that there is no straigh correlation between Resp_Level and exp_years, neither high_education. The backgroung does not seem to be important for offering rresponsibility jobs.

Too_Masculine VS ???

There is a little trend on the correlation of TooMasculine with the number of requirements, the complexity of the selection, the requirement of a high education background. The sexist trend is very subtle.

Too_Feminine VS ???

there is no straight correlation possible between a feminine coded bulletin and other bulletin characteristics.

Salaries

The entry salary is - positively correlated to the responsability level (seems consistent) - negatively correlated to the "open to all" criteria

4.c Suspicious bulletins We need to identify the most suspicious bulletins inside our dataframe of 675. By "suspicious", we mean buletins that would be biased, or that would need to follow basic recommandations about the validity duration, or other parameters.

For this, we are going to score bulletin. The score is a combiations of penalties based on the main indcators of a biais which are - Resp_level - gender_score - nb_line_scale - nb_char_scale - nb_selection_steps - nb_requirements - Open_To_All - validity_duration

The higher the socre is, the higher is the necessity to look up the bulletin. > When it comes to medium to high responnsibility position, biased bulletins are sanctioned even harder.

```
Out[1019]: 0      434
           15     105
           20      63
           60      29
           80      13
           25      13
           108      4
           4        4
           100      3
           8        2
           2        2
           24       1
           16       1
           10       1
           Name: score, dtype: int64
```

```
Out[1020]:                                                File Name  \
           145  COMMUNICATIONS INFORMATION REPRESENTATIVE 1461...
           617  TRANSPORTATION ENGINEERING ASSOCIATE 7280 0724...
           163    CUSTOMER SERVICE REPRESENTATIVE 1230 020918.txt
           195  ELECTRICAL ENGINEERING ASSOCIATE 7525 093016 R...
           55   ASSISTANT DIRECTOR INFORMATION SYSTEMS 9377 03...
           172  DIRECTOR OF AIRPORT OPERATIONS 9304 062317 (3)...
           181  DIRECTOR OF PRINTING SERVICES 1488 101218 REV ...
           113             CHIEF FORENSIC CHEMIST 2237 092818.txt
           120     CHIEF OF OPERATIONS 7258 042018 REV 051318.txt
           119      CHIEF OF DRAFTING OPERATIONS 7271  042018.txt

                                                 Position  salary_start  score  \
```

```
145          communications information representative         41697    108
617  transportation engineering associate                     68549    108
163                    customer service representative         57148    108
195                  electrical engineering associate          66231    108
55              assistant director information systems        145199    100
172                     director of airport operations        141858    100
181                     director of printing services         117596    100
113                             chief forensic chemist        106362     80
120                                chief of operations         90473     80
119                       chief of drafting operations        135302     80

    salary_end validity_duration nb_lines nb_chars  Exercices  MultiChoice  \
145    $59,340                11      111    13006        1.0          0.0
617    $85,149                11       88    12544        0.0          0.0
163    $71,012                11       89     9186        0.0          0.0
195    $94,252                11       94    11977        0.0          0.0
55    $180,382                13       77     8157        0.0          0.0
172   $201,867                13       80     9286        0.0          0.0
181   $171,946                20       74     8078        0.0          0.0
113   $155,493                13       77     6419        0.0          0.0
120   $132,274                20      103    14062        0.0          0.0
119   $168,084                13       80     8166        0.0          0.0

     ...  full_time_part_time_code  salary_start_code  salary_end_strip  \
145  ...                        2                  1             59340
617  ...                        0                  1             85149
163  ...                        2                  1             71012
195  ...                        0                  1             94252
55   ...                        2                  4            180382
172  ...                        2                  4            201867
181  ...                        2                  3            171946
113  ...                        2                  3            155493
120  ...                        2                  2            132274
119  ...                        2                  4            168084

     salary_variation  salary_variation_code             GENDER_WORDS  \
145           17643.0                      2   strongly feminine-coded
617           16600.0                      2  strongly masculine-coded
163           13864.0                      1   strongly feminine-coded
195           28021.0                      3  strongly masculine-coded
55            35183.0                      3   strongly feminine-coded
172           60009.0                      5  strongly masculine-coded
181           54350.0                      5  strongly masculine-coded
113           49131.0                      5  strongly masculine-coded
120           41801.0                      4  strongly masculine-coded
119           32782.0                      3   strongly feminine-coded

     Too_Feminine  Too_Masculine  gender_bias  gender_score
```

```
145             1               0               4               4
617             0               1               2               4
163             1               0               4               4
195             0               1               2               4
55              1               0               4               4
172             0               1               2               4
181             0               1               2               4
113             0               1               2               4
120             0               1               2               4
119             1               0               4               4

[10 rows x 30 columns]
```

4.c Top 10 Suspicious bulletins scoring analysis In this top 10: > 40% are of the bulletins are strongly feminine coded > 60% are strongly masculine coded. > All of them are medium to high responibility positions which make sense, as our schme add more penalites to thos profiles. > The highest scores are given to 11-days validity duration bulletins open to all, which can be considered as too short to allow external candidates to apply

Revolting habits: the score also unviels that the city is very traditionalist and tends to follow stereotypes likes > Jobs for women consists in secretary, communication or sales. > Jobs for men consists in more technical jobs such as engineers.

2 5. Text Analysis

```
---------------------------------------------------------------------------

ModuleNotFoundError                       Traceback (most recent call last)

<ipython-input-1022-9f2b669dad53> in <module>
----> 1 import spacy
      2 from pprint import pprint
      3 from collections import Counter
      4
      5 nlp = spacy.load("en_core_web_sm")


ModuleNotFoundError: No module named 'spacy'
```

3 6. Modeling

We managed to get a dataset with the number of applicants to each position and their characteristics (gender and race). Let's check our work on the gender scoring.

```
Out[783]:   Fiscal Year          Job Number                Job Description  \
          0   2013-2014  9206 OP 2014/04/18              311 DIRECTOR 9206
```

```
1   2013-2014   1223 P 2013/08/09           ACCOUNTING CLERK 1223
2   2013-2014  7260 OP 2014/02/14            AIRPORT MANAGER 7260
3   2013-2014   3227 P 2013/11/15  AIRPORT POLICE LIEUTENANT 2013
4   2013-2014   2400 O 2014/05/02                   AQUARIST 2400

   Apps Received  Female  Male  Unknown_Gender  Black  Hispanic  Asian  \
0             54      20    31               3     25        18      1
1            648     488   152               8    151       204    123
2             51      13    37               1      8        12      9
3             48       9    38               1     21        14      3
4             40      15    24               1      3         7      7

   Caucasian  American Indian/ Alaskan Native  Filipino  Unknown_Ethnicity  \
0          6                                0         0                  4
1         62                                3        79                 26
2         20                                0         0                  2
3          7                                0         1                  2
4         19                                1         1                  2

  JobNumber
0      9206
1      1223
2      7260
3      3227
4      2400
```

Out[847]:

```
                                        File Name  \
0  CUSTOMER SERVICE REPRESENTATIVE 1230 020918.txt
1    CHIEF OF DRAFTING OPERATIONS 7271  042018.txt

                          Position  salary_start  score salary_end  \
0  customer service representative         57148    108    $71,012
1     chief of drafting operations        135302     80   $168,084

  validity_duration nb_lines nb_chars  Exercices  MultiChoice  ...  Female  \
0                11       89     9186        0.0          0.0  ...   19892
1                13       80     8166        0.0          0.0  ...       2

   Male  Unknown_Gender  Black  Hispanic  Asian  Caucasian  \
0  7968             370  12618     10214   1094       1958
1    11               0      1         7      2          1

   American Indian/ Alaskan Native  Filipino  Unknown_Ethnicity
0                              131       740               1475
1                                0         1                  1

[2 rows x 45 columns]
```

5.2 Simple encoding of target result based on what we know Let's code a target label based on the number of male/female applicants. First, simple : if more female applicants, let's code it as attract_female_applicants = 1 else 0

Out[849]:

	File Name	Position	salary_start	score	salary_end	validity_duration	nb_lines	nb_chars	Exercices	MultiChoice	...	salary_variation_code	JobNumber	Fiscal Year	Job Number	Job Description	Apps Received	Female	Male	ratio	attract_female_applicants
0	CUSTOMER SERVICE REPRESENTATIVE 1230 020918.txt	customer service representative	57148	108	$71,012	11	89	9186	0.0	0.0	...	1	1230	2013-2014	1230 O 2013/12/27	CUSTOMER SERVICE REPRESENTATIVE 1230	28230	19892	7968	2.496486	1
1	CHIEF OF DRAFTING OPERATIONS 7271 042018.txt	chief of drafting operations	135302	80	$168,084	13	80	8166	0.0	0.0	...	3	7271	2013-2014	7271 P 2013/11/08	CHIEF OF DRAFTING OPERATIONS 7271	13	2	11	0.181818	0
2	CHIEF MANAGEMENT ANALYST 9182 020918.txt	chief management analyst	123667	80	$179,944	13	76	7385	0.0	0.0	...	5	9182	2013-2014	9182 P 2014/06/20	CHIEF MANAGEMENT ANALYST 9182	143	78	54	1.444444	1
3	CONSTRUCTION INSPECTOR 7291 042117.txt	construction inspector	80283	60	$97,092	NaN	133	15445	0.0	0.0	...	2	7291	2014-2015	7291 P 2014/07/04-ARCHIVE	CONSTRUCTION INSPECTOR 7291 - ARCHIVE	471	17	443	0.038375	0
4	FIRE PROTECTION ENGINEERING ASSOCIATE 7978 041...	fire protection engineering associate	66231	60	$96,841	13	89	9384	0.0	0.0	...	3	7978	2013-2014	7978 O 2014/6/6	FIRE PROTECTION ENGINEERING ASSOCIATE	107	16	89	0.179775	0

[5 rows x 39 columns]

Out[850]: Index(['File Name', 'Position', 'salary_start', 'score', 'salary_end',

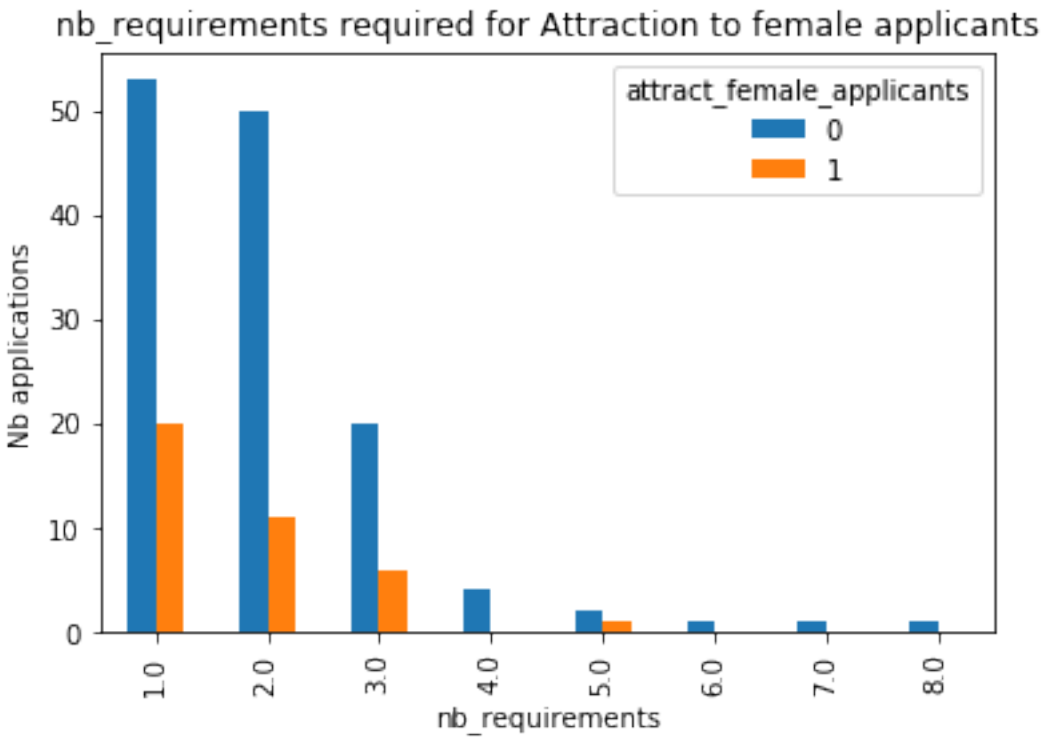
```
       'validity_duration', 'nb_lines', 'nb_chars', 'Exercices', 'MultiChoice',
       'OralPres', 'PhysicalTest', 'nb_requirements', 'nb_selection_steps',
       'nb_line_scale', 'nb_char_scale', 'Resp_level', 'Open_To_All',
       'exp_years', 'high_education', 'full_time_part_time_code',
       'GENDER_WORDS', 'Too_Feminine', 'Too_Masculine', 'gender_bias',
       'gender_score', 'salary_start_code', 'salary_end_strip',
       'salary_variation', 'salary_variation_code', 'JobNumber', 'Fiscal Year',
       'Job Number', 'Job Description', 'Apps Received', 'Female', 'Male',
       'ratio', 'attract_female_applicants'],
      dtype='object')
```

3.0.1 Let's analyse some relationships

```
Out[851]: Text(0, 0.5, 'Nb applications')
```

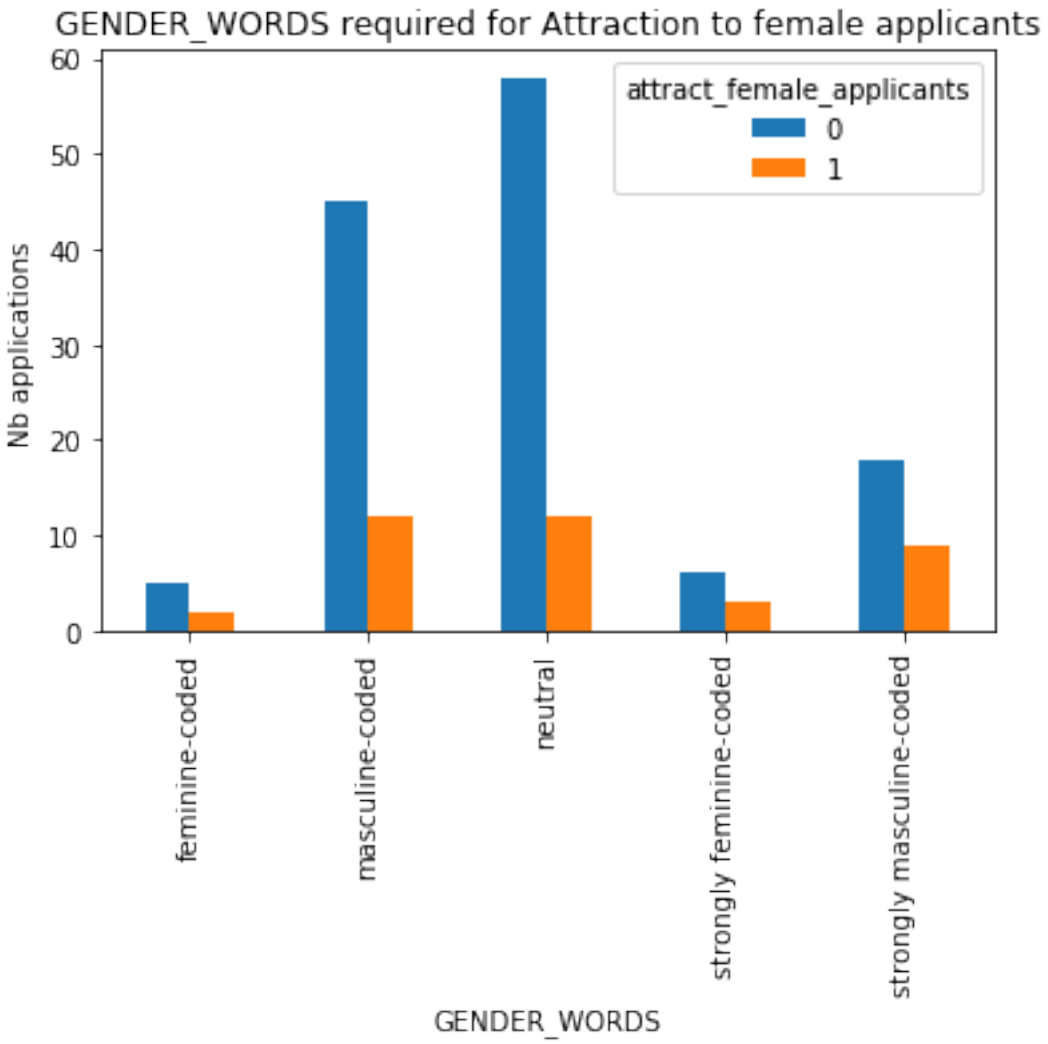
- globally, jobs attract men and women until the 4th level of salary where no job that offered the salary levels 4 and 5 (the highest) attracted more women than men.
- In fact, they all attracted more men than women to apply.

```
Out[852]: Text(0, 0.5, 'Nb applications')
```



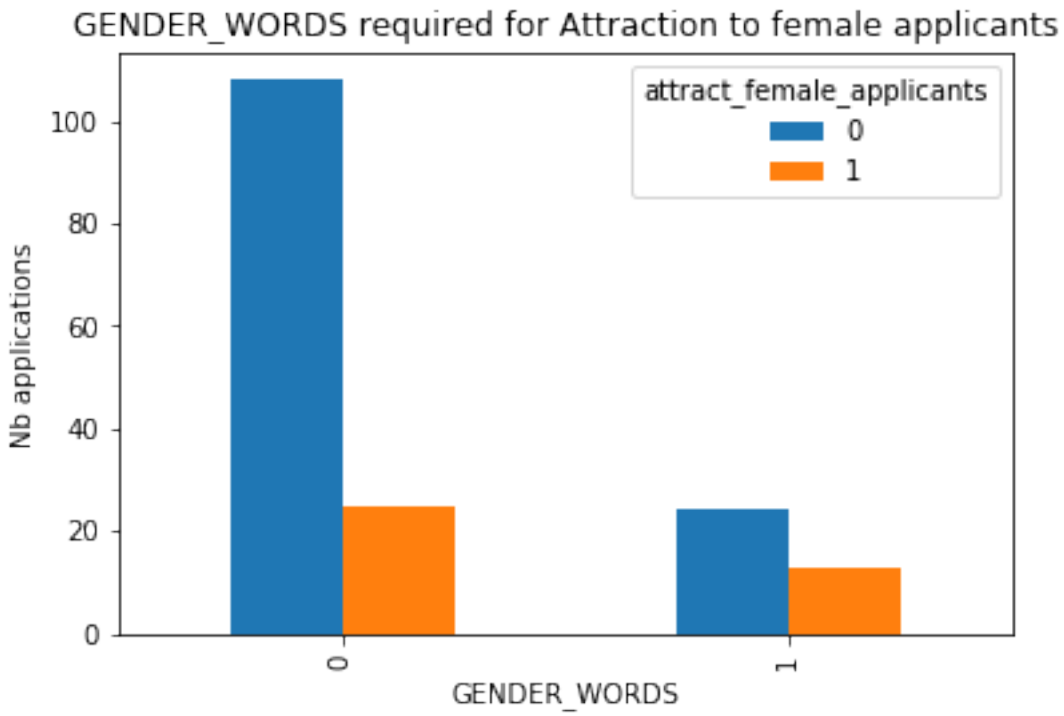
- The more the number of requiresments, the less applications are done mostly by women

```
Out[853]: Text(0, 0.5, 'Nb applications')
```



- We would have expected a clearer relationship between the trend of candidate (mostly feminine or masculine) and the words used.
- Here we can see indeed that the difference is very little when the job posting is strongly feminine-coded, meaning women apply more on those jobs in average
- but we observe kind of the same for strongly masculine-coded words..

```
Out[854]: Text(0, 0.5, 'Nb applications')
```



3.0.2 5.3 Training and testing

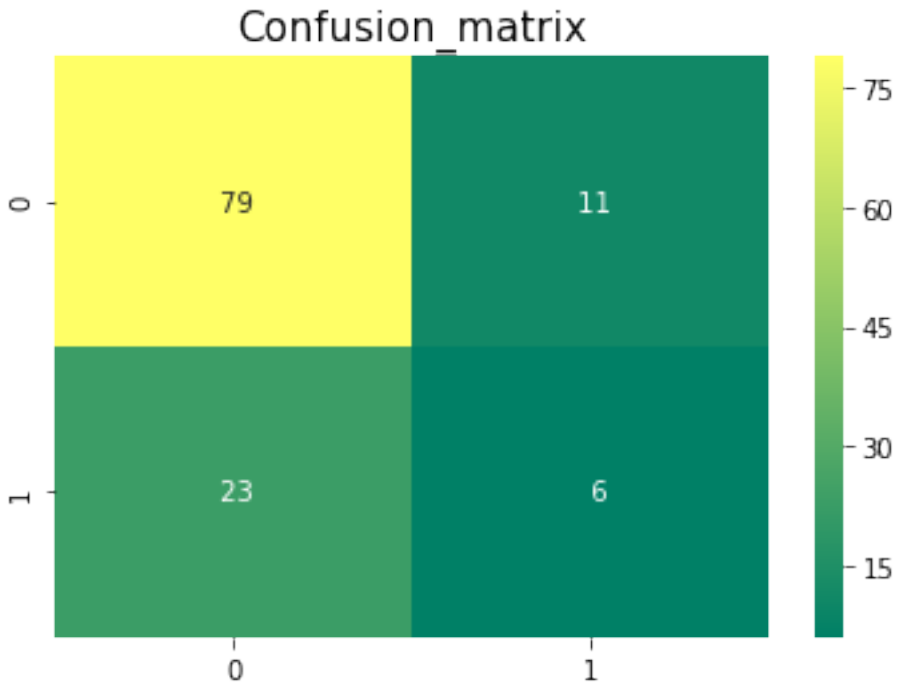
Let's get our train and test datasets from the labeled one

```
Out[856]: Index(['validity_duration', 'nb_lines', 'nb_chars', 'Exercices', 'MultiChoice',
                 'OralPres', 'PhysicalTest', 'nb_requirements', 'nb_selection_steps',
                 'nb_line_scale', 'nb_char_scale', 'Resp_level', 'Open_To_All',
                 'exp_years', 'high_education', 'full_time_part_time_code',
                 'gender_score', 'salary_start_code', 'salary_end_strip',
                 'salary_variation', 'salary_variation_code',
                 'attract_female_applicants'],
                dtype='object')
```

5.3.1 First model : Decision Tree

```
--------------The Accuracy of the model----------------------------
The accuracy of the Decision Tree Classifier is 60.0
The cross validated score for Decision Tree Classifier is: 71.36
```

```
Out[859]: Text(0.5, 1.05, 'Confusion_matrix')
```



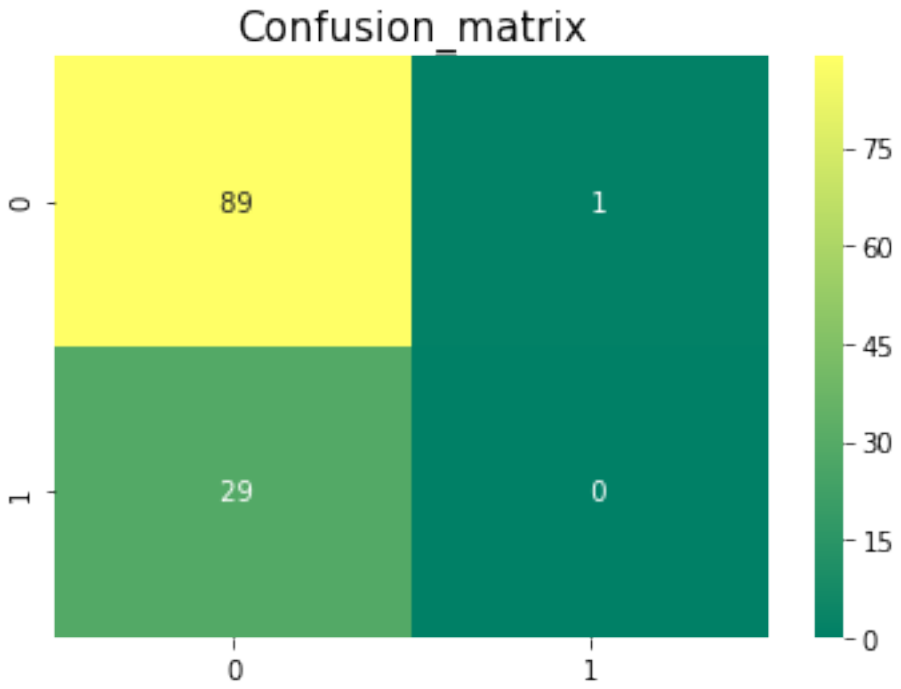
The confusion matrix tells us :

```
* 79 true positive were predicted as positive
* 6 true negative were predicted as negative
* 11 true positive were predicted as negative
* 23 true negative were predicted as positive
```

We get a score of 71.36% and an accuracy of 60, not very good...
Let's check other models :

3.0.3 5.3.2 Logistic Regression

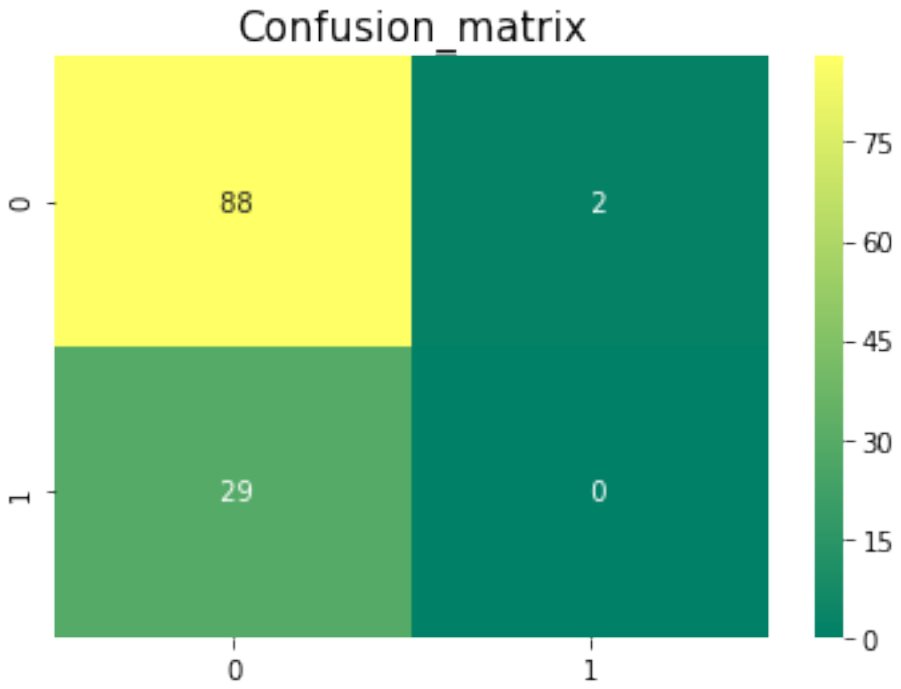
```
--------------The Accuracy of the model----------------------------
The accuracy of the Logistic Regression is 76.67
The cross validated score for Logistic Regression is: 74.85


Out[860]: Text(0.5, 1.05, 'Confusion_matrix')
```



5.3.3 Random Forests

```
--------------The Accuracy of the model----------------------------
The accuracy of the Random Forest Classifier is 80.0
The cross validated score for Random Forest Classifier is: 73.94
```

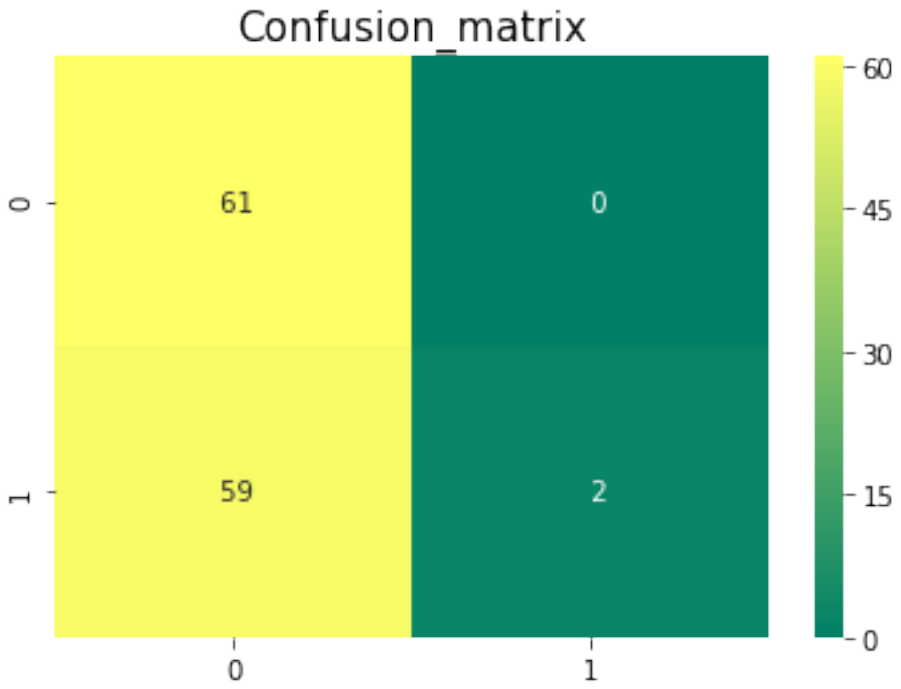
```
Out[861]: Text(0.5, 1.05, 'Confusion_matrix')
```



	precision	recall	f1-score	support
0	0.83	0.96	0.89	25
1	0.00	0.00	0.00	5
micro avg	0.80	0.80	0.80	30
macro avg	0.41	0.48	0.44	30
weighted avg	0.69	0.80	0.74	30

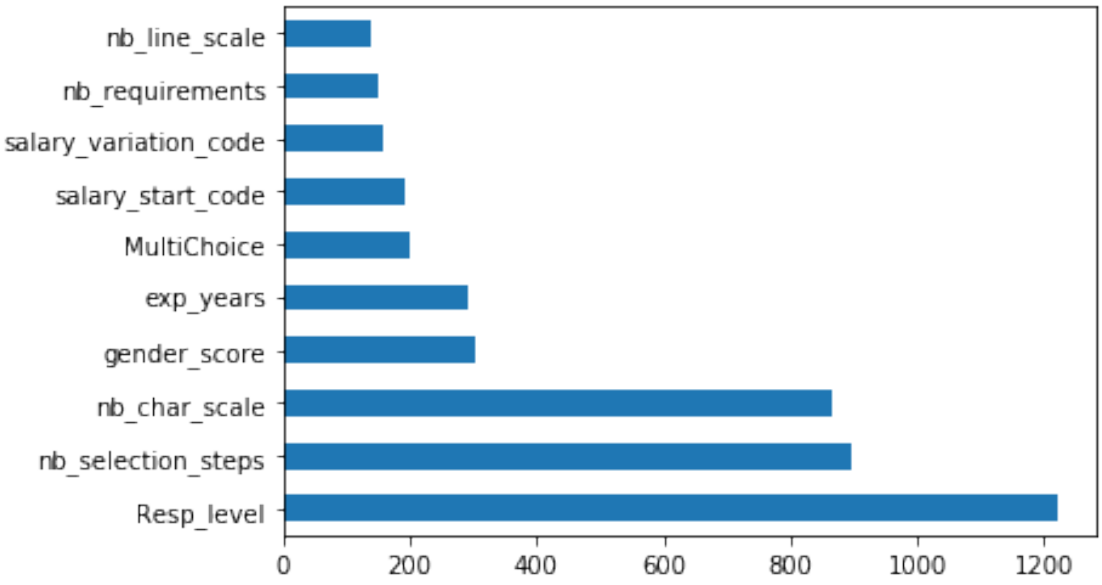
```
--------------The Accuracy of the model----------------------------
The accuracy of the Support Vector Machines Classifier is 40.54
The cross validated score for Support Vector Machines Classifier is: 51.67
```

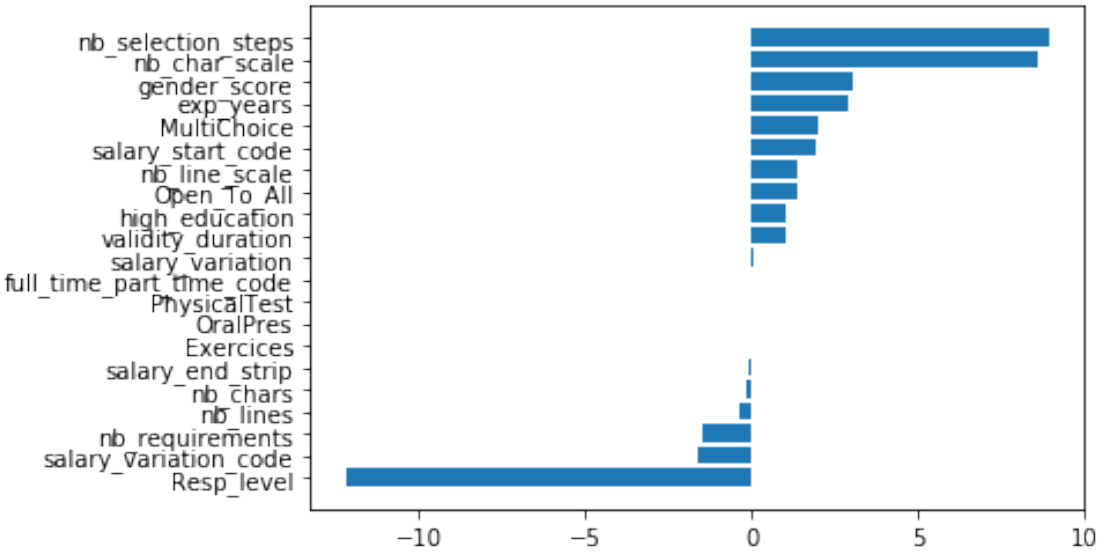
Out[925]: Text(0.5, 1.05, 'Confusion_matrix')



We get a good score with a good accuracy, let's try to check the most important features

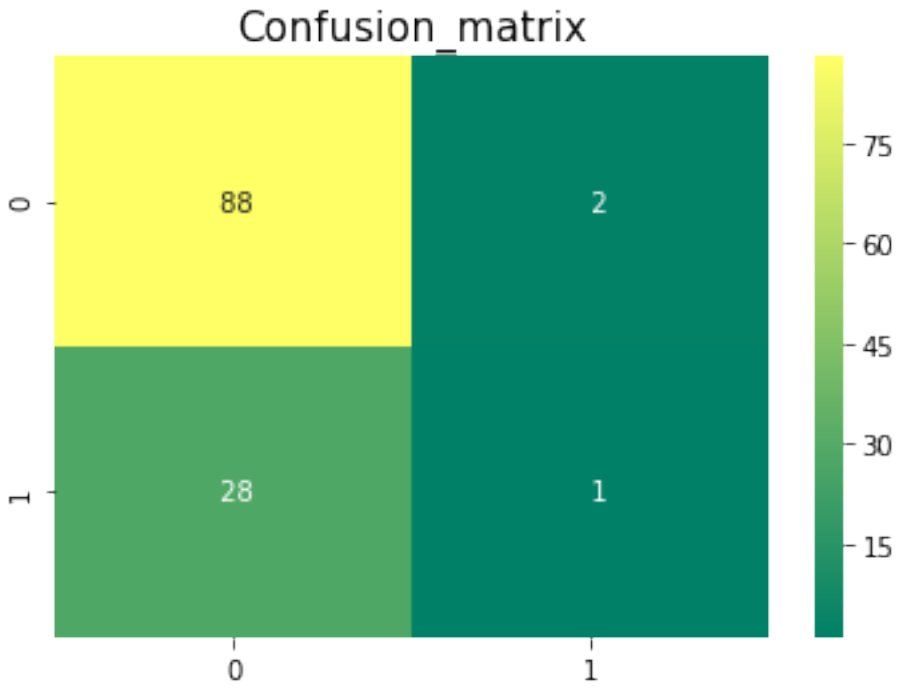
Out[864]: <matplotlib.axes._subplots.AxesSubplot at 0x22f92f12da0>





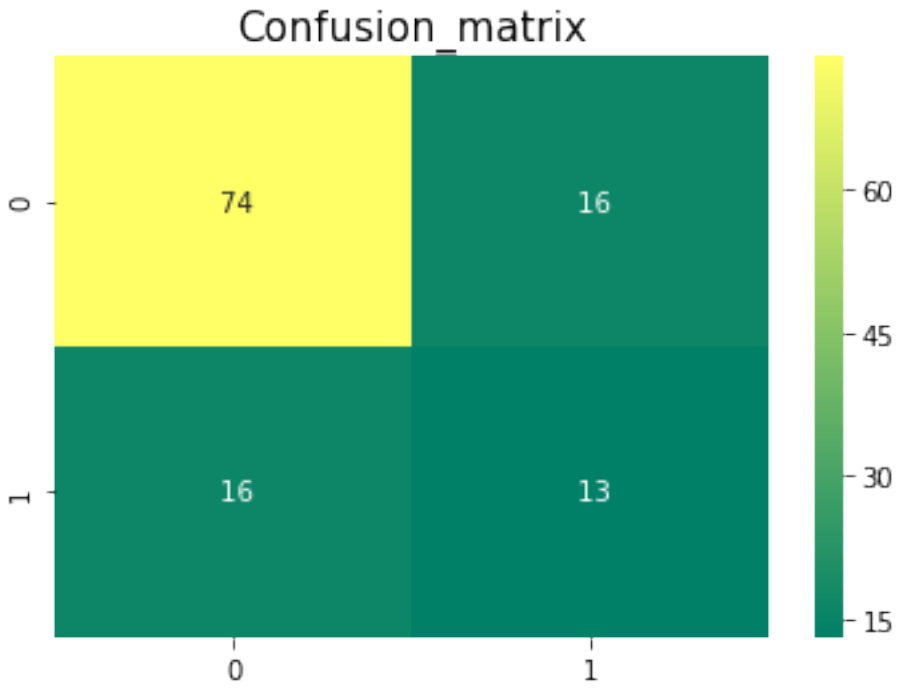
- The number of selection steps, the "gender score" (wether it's biased in favor of men or not), the years of experience and entry salary seem to be impactful in favor of female applications.
- On the other end, the responsability level seems to have a huge negative impact on female applications.

```
--------------The Accuracy of the model----------------------------
The accuracy of the K Nearst Neighbors Classifier is 80.0
The cross validated score for K Nearest Neighbors Classifier is: 74.77


Out[866]: Text(0.5, 1.05, 'Confusion_matrix')
```



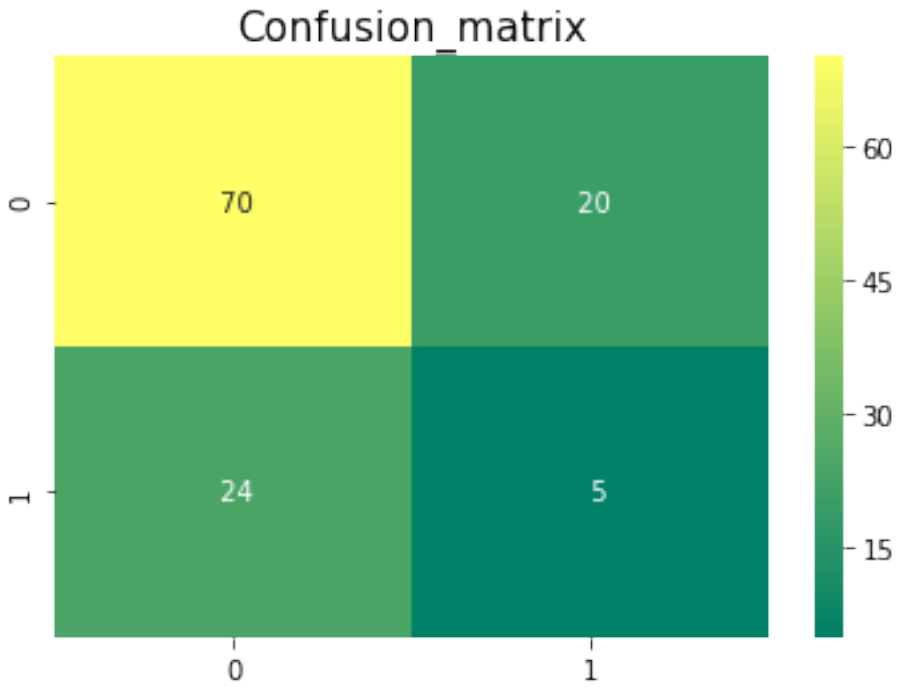

```
--------------The Accuracy of the model----------------------------
The accuracy of the Gaussian Naive Bayes Classifier is 70.0
The cross validated score for Gaussian Naive Bayes classifier is: 73.26
```

```
Out[867]: Text(0.5, 1.05, 'Confusion_matrix')
```



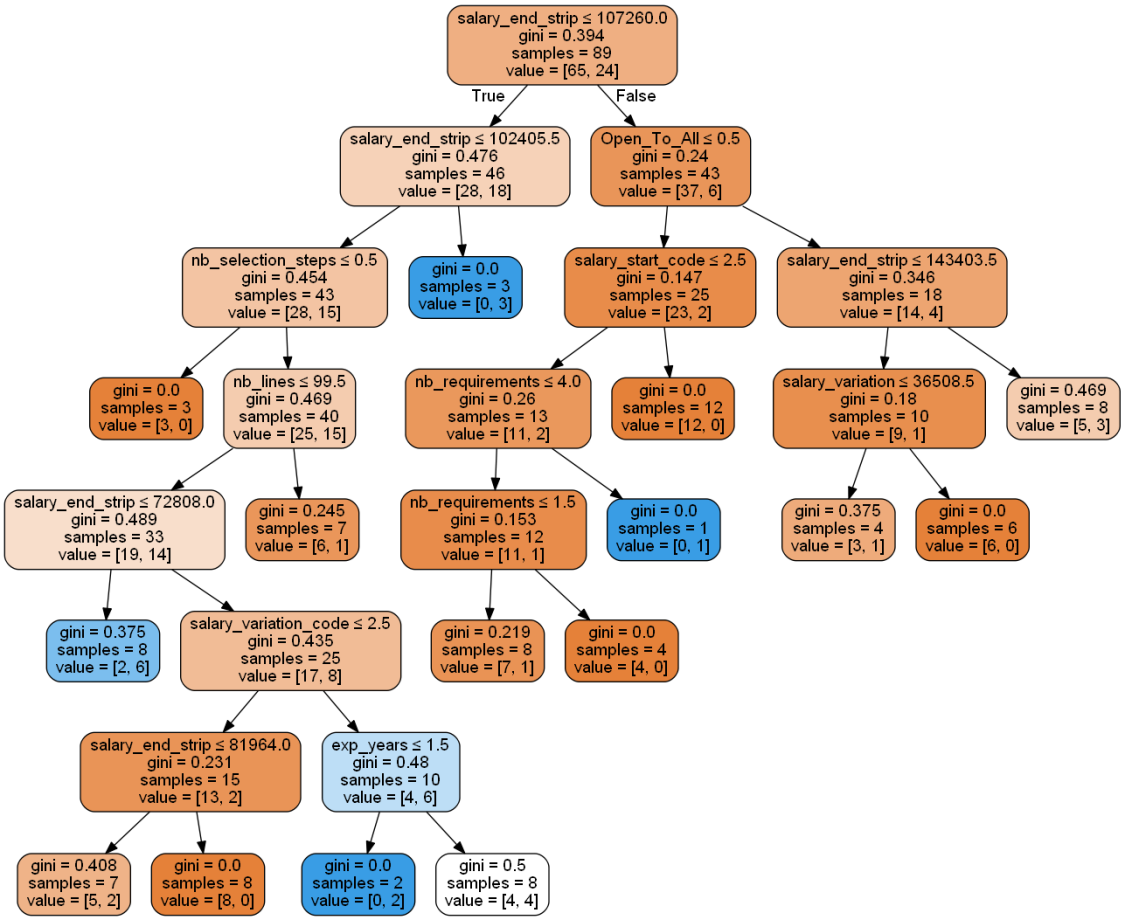

```
--------------The Accuracy of the model----------------------------
The accuracy of the DecisionTree Classifier is 66.67
The cross validated score for Decision Tree classifier is: 70.53
```

Out[868]: Text(0.5, 1.05, 'Confusion_matrix')



Let's check the decision tree in detail :

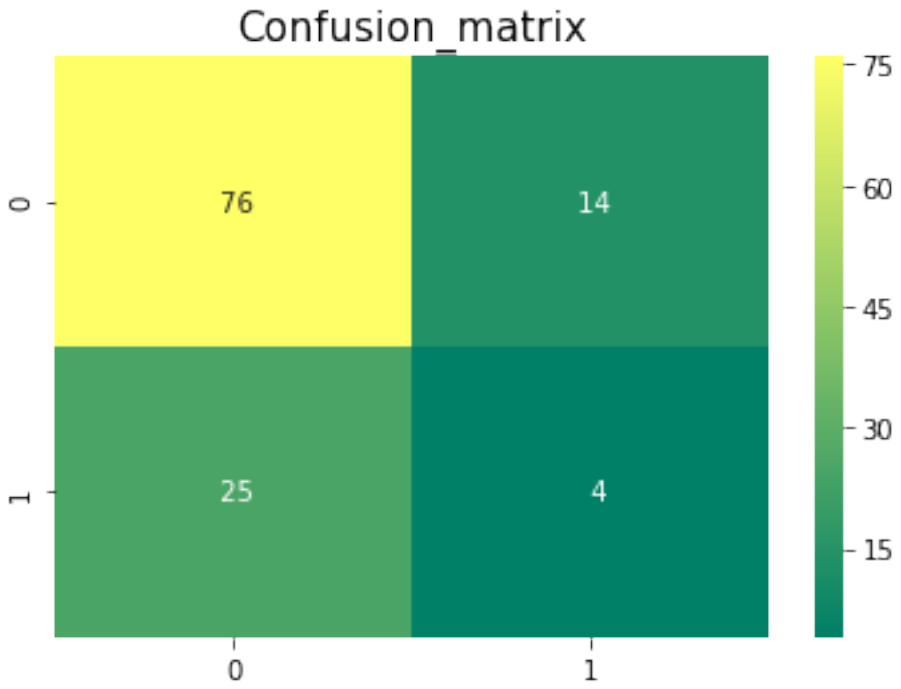
Out[869]:



- The maxmum salary in the job posting seems to be very important in that model to determine wether we'll have more male applicants over female ones.

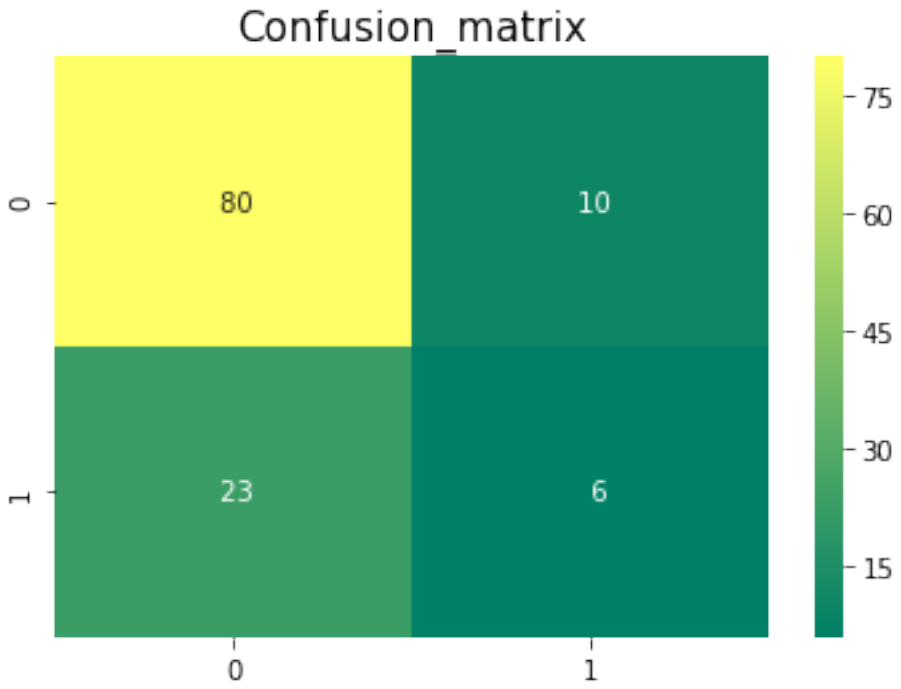
```
--------------The Accuracy of the model----------------------------
The accuracy of the AdaBoostClassifier is 76.67
The cross validated score for AdaBoostClassifier is: 67.12
```

```
Out[870]: Text(0.5, 1.05, 'Confusion_matrix')
```



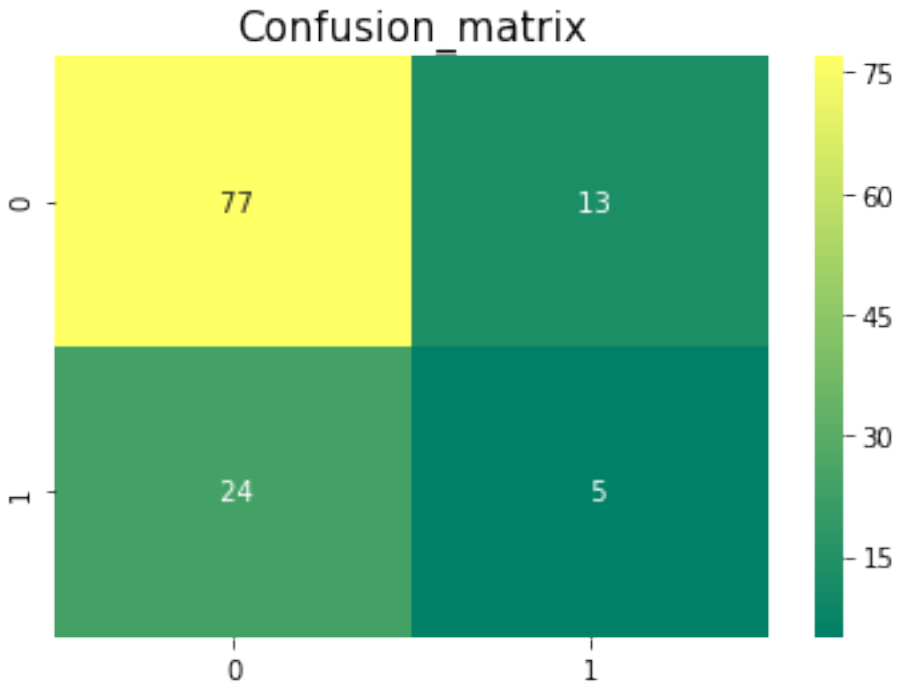

```
--------------The Accuracy of the model----------------------------
The accuracy of the LinearDiscriminantAnalysis is 60.0
The cross validated score for AdaBoostClassifier is: 72.27
```

```
Out[871]: Text(0.5, 1.05, 'Confusion_matrix')
```




```
--------------The Accuracy of the model----------------------------
The accuracy of the Gradient Boosting Classifier is 70.0
The cross validated score forGradient Boosting Classifier is: 69.02
```

Out[872]: Text(0.5, 1.05, 'Confusion_matrix')



3.0.4 Let's compute a summary of all our models to compare

Out[873]:

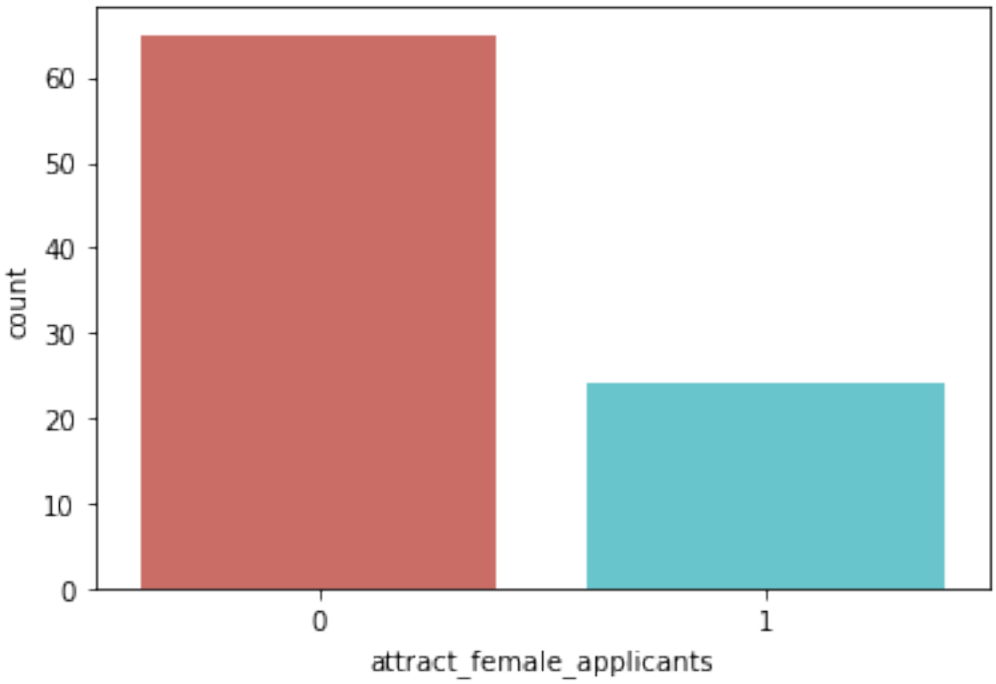
	Model	Score
0	Support Vector Machines	0.756818
2	Logistic Regression	0.748485
1	KNN	0.747727
3	Random Forest	0.739394
4	Naive Bayes	0.732576
7	Linear Discriminant Analysis	0.722727
8	Decision Tree	0.705303
6	Gradient Decent	0.690152
5	AdaBoostClassifier	0.671212

- The Support Vector Machines seems to be the best model, we should try it on our dataframe !

However, there is a test we didn't perform to check how balanced our classes are... In our classes are imbalanced, our model can be wrong..

```
Out[874]: 69     0
          164    1
          154    0
          80     0
          6      1
          Name: attract_female_applicants, dtype: int64
```

```
Out[876]: 0    65
          1    24
          Name: attract_female_applicants, dtype: int64
```




```
percentage of NOT 'attract_female_applicants' is 73.03370786516854
percentage of 'attract_female_applicants' 26.96629213483146
```

- Indeed, our classes are imbalanced. We'll could the SMOTE algo to up-sample and improve our model. https://arxiv.org/pdf/1106.1813.pdf

(We're just going to make first step but no time to investigate more...)

```
Out[883]: Index(['validity_duration', 'nb_lines', 'nb_chars', 'Exercices', 'MultiChoice',
                 'OralPres', 'PhysicalTest', 'nb_requirements', 'nb_selection_steps',
                 'nb_line_scale', 'nb_char_scale', 'Resp_level', 'Open_To_All',
                 'exp_years', 'high_education', 'full_time_part_time_code',
                 'gender_score', 'salary_start_code', 'salary_end_strip',
                 'salary_variation', 'salary_variation_code',
                 'attract_female_applicants'],
                dtype='object')
```

```
length of oversampled data is  122
Number of no attract_female_applicants in oversampled data 61
Number of subscription 61
Proportion of no attract_female_applicants data in oversampled data is  0.5
Proportion of attract_female_applicants data in oversampled data is  0.5
```

Let's try SVM on our unlabeled dataframe !

Out[449]:

	validity_duration	nb_lines	nb_chars	nb_requirements	nb_selection_steps	nb_line_scale	nb_char_scale	Resp_level	Open_To_All	exp_years	high_education	full_time_part_time_code	predicted
145	11	111	13006	3.0	2.0	3	4	3	1	1.0	0	2	1
163	11	89	9186	1.0	2.0	2	1	3	1	2.0	0	2	0
55	13	77	8157	1.0	2.0	1	1	5	0	2.0	0	2	1
172	13	80	9286	3.0	1.0	1	1	5	1	4.0	0	2	1
181	20	74	8078	2.0	1.0	1	1	5	0	3.0	0	2	1

Out[450]:

	File Name	Position	salary_start	score	salary_end	validity_duration_x	nb_lines_x	nb_chars_x	nb_requirements_x	nb_selection_steps_x	...	nb_requirements_y	nb_selection_steps_y	nb_line_scale_y	nb_char_scale_y	Resp_level_y	Open_To_All_y	exp_years_y	high_education_y	full_time_part_time_code_y	predicted
145	COMMUNICATIONS INFORMATION REPRESENTATIVE 1461...	communications information representative	41,697	108	$59,340	11	111	13006	3.0	2.0	...	3.0	2.0	3	4	3	1	1.0	0	2	1
163	CUSTOMER SERVICE REPRESENTATIVE 1230 020918.txt	customer service representative	57,148	108	$71,012	11	89	9186	1.0	2.0	...	1.0	2.0	2	1	3	1	2.0	0	2	0

[2 rows x 36 columns]

```
Out[451]:      score  predicted
          145    108          1
          163    108          0
```