HumanMatters_DS4Good_pdf

July 15, 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 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
```

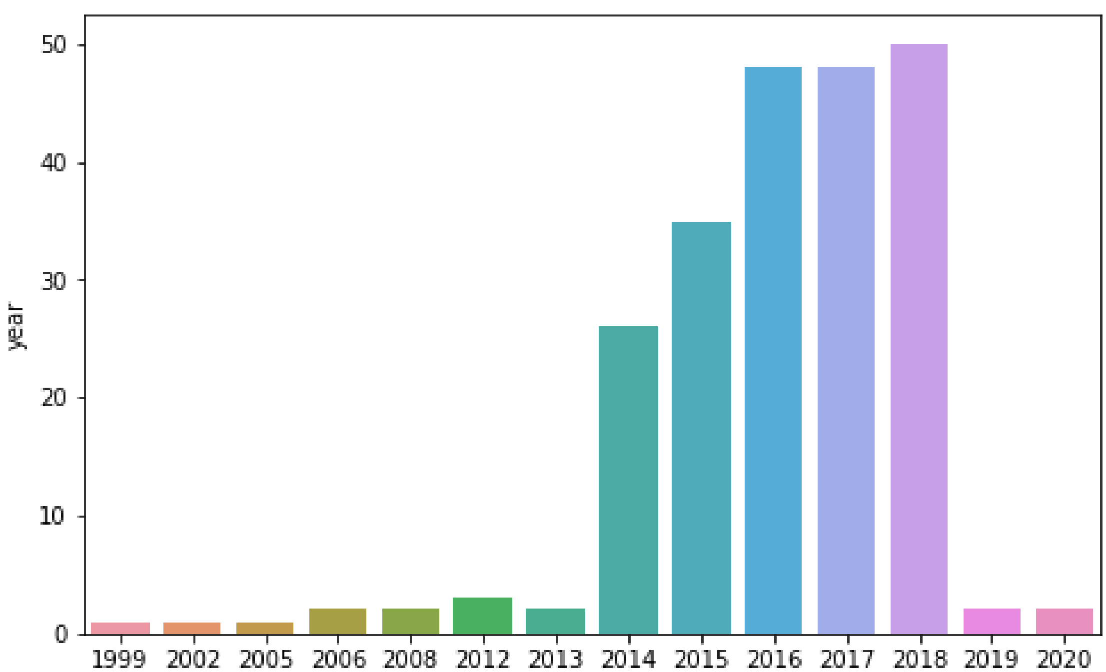
- On the 683 files we've been processing, 675 are now in our dataframe, so only a few presented a problem during parsing. We have most of them though (98%) so we can move on.
- We can notice we don't have all values for the following fields :
- salary_end
- deadline
- validity_duration
- EXPERIENCE_LENGTH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

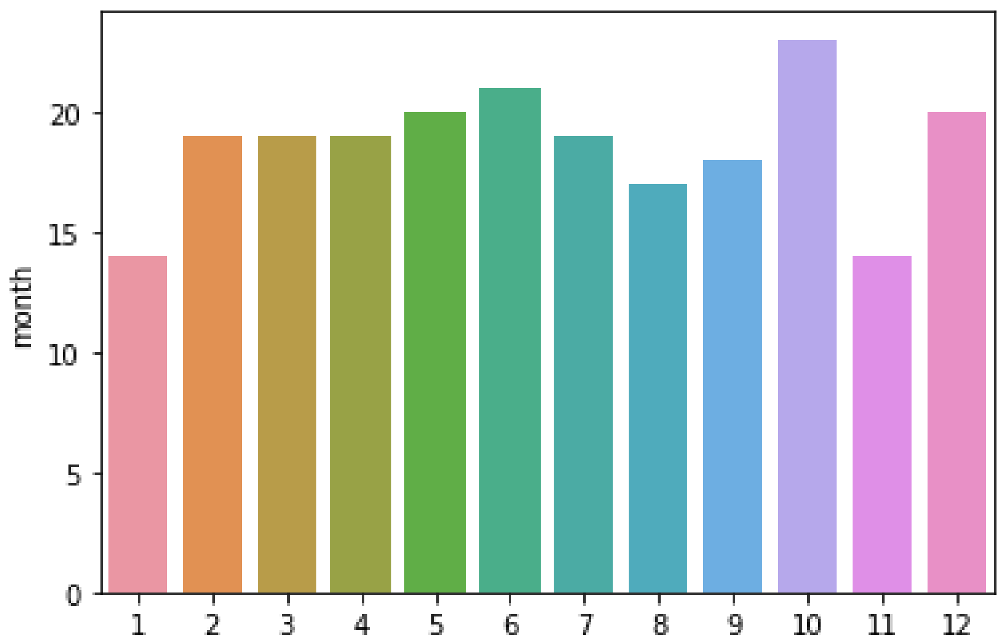
```
4    2018-06-14                13  ...      0.0          0.0   1.0

   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 !

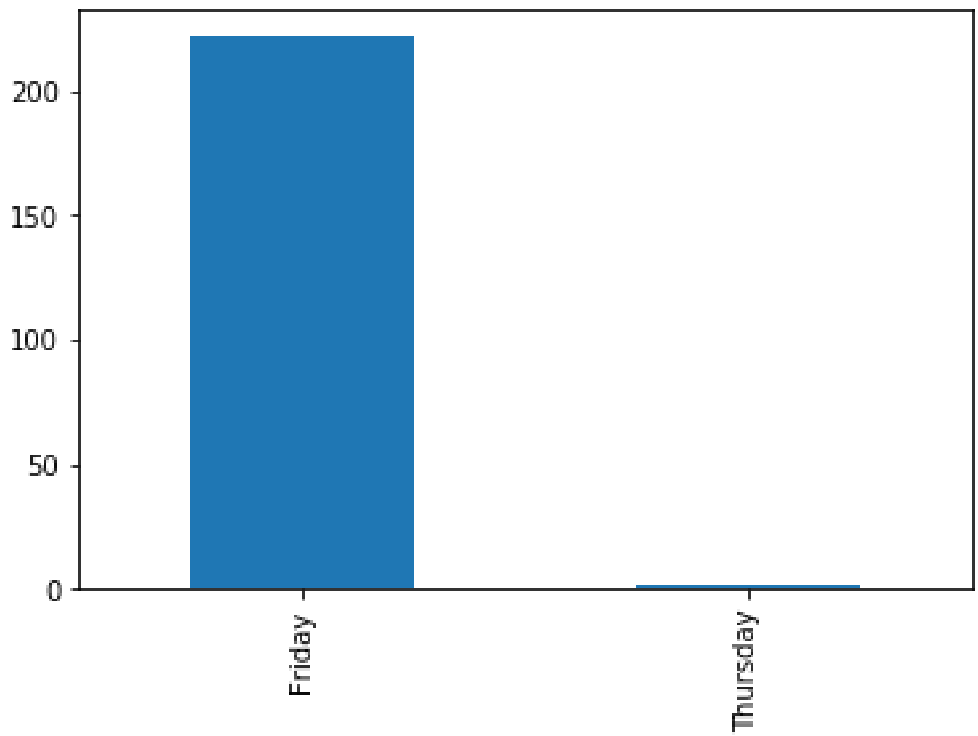
1.b.2 Job postings by month over the years



- The job openings seem about the same over the years throughout the months. January and November seem the months when there are less job opennings; October on the other hand seems to be the month when most of job opennings 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 ?

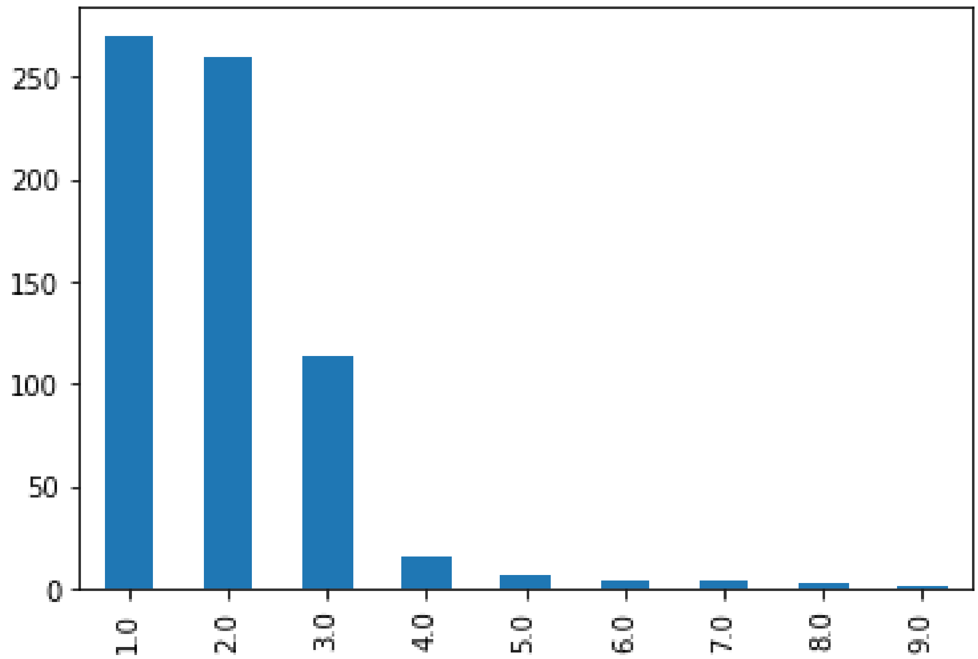
```
Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x16b0a5e6f60>
```



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[11]: <matplotlib.axes._subplots.AxesSubplot at 0x16b0a652da0>



The number of requirements can have a big impact on the reading of the bulletin. The more is displayed, the more female candidate can be discouraged, which might lead to 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 displayed 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 excessive 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[12]: [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

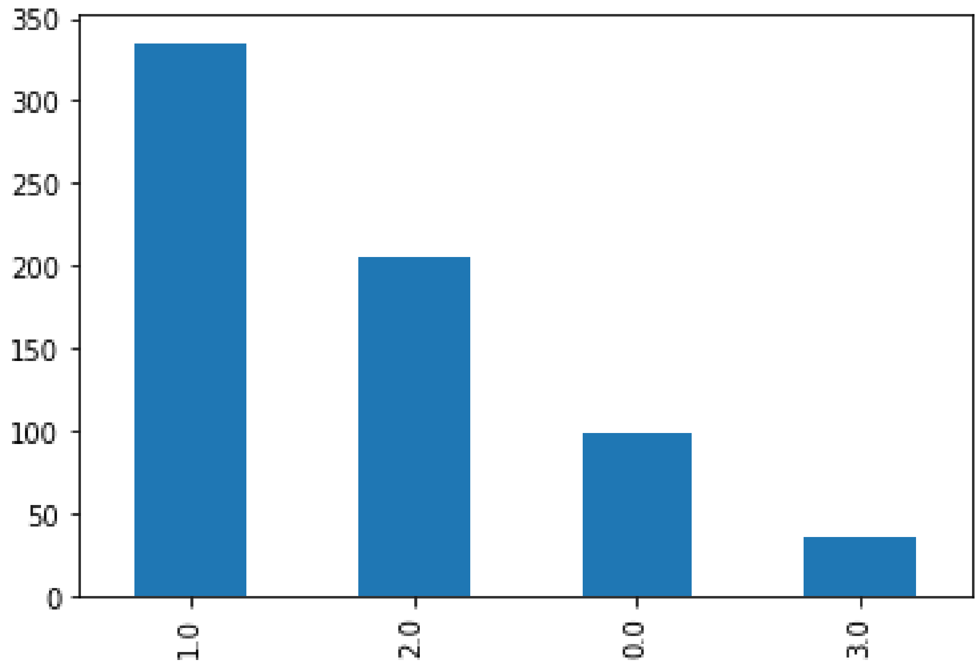
```
Out[14]: ['Test',
          'Questionnaire',
          'Experience',
          'Exercise',
          'Defense',
          'Review',
          'Evaluation',
          'Abilities',
          'Exercises',
          'Interview',
          'Choice',
          'Essay',
          'Written']
```

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

```
Out[15]: count             675
         unique             31
```

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

Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x16b0a8360f0>



The number of steps in the selection process is an flat indicator of the complexity of the selection process. 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.

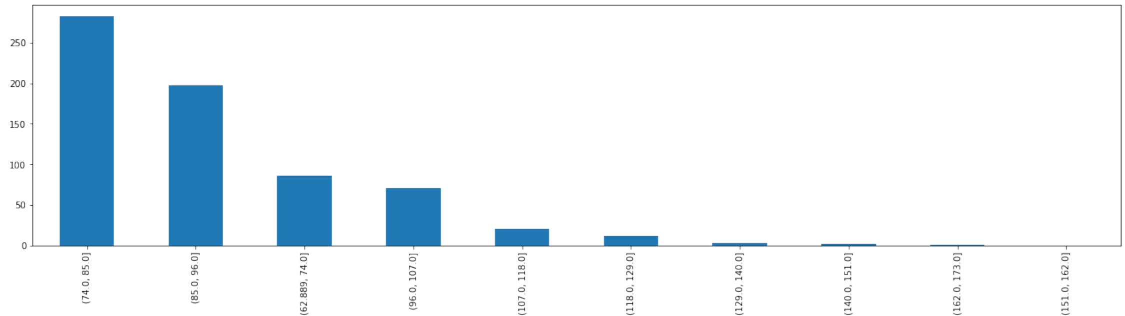
Here:

- 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 intuition, 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[19]: <matplotlib.axes._subplots.AxesSubplot at 0x16b0a886f98>



Analysis of the number of lines in the job description The number of lines is a first indicator of the complexity in the reading of 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.

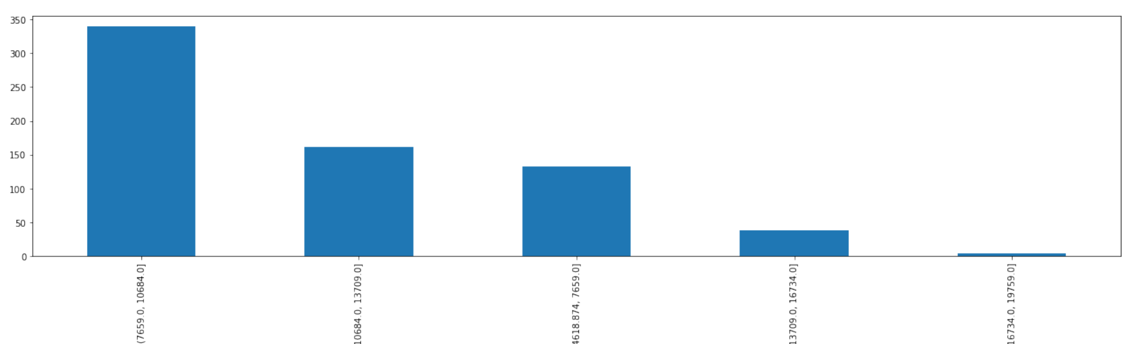
Here:

- 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 legitimate or not. If it is not the case, maybe is it due to an unconsious bias. We will use a scale from 1 to 5

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

Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x16b0a91ab00>



Analysis The number of chars is a second indicator of the complexity in the reading of 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 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.

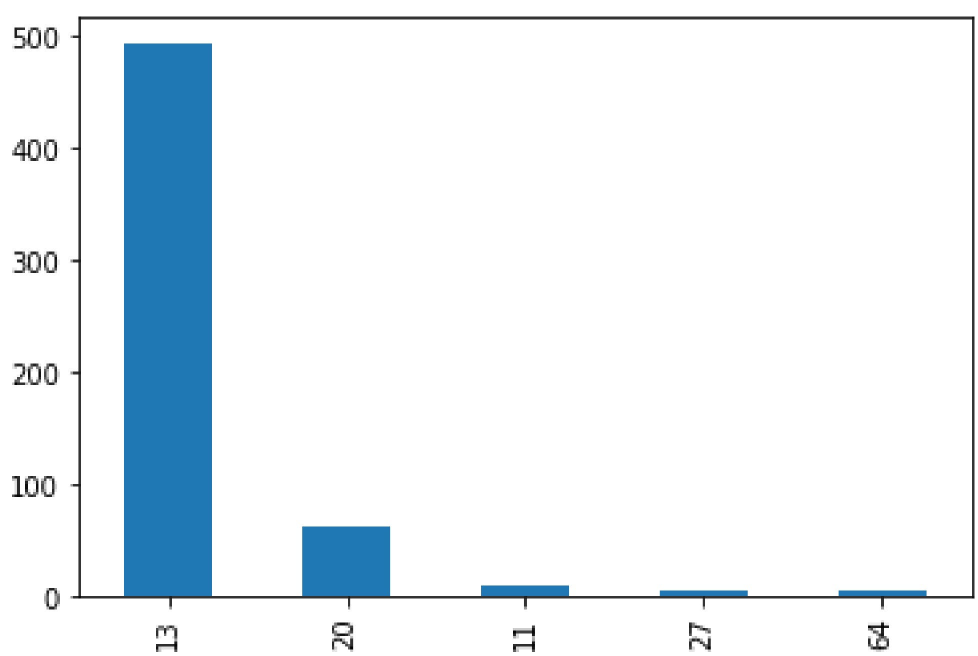
Here:

- 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 bias. 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[24]: <matplotlib.axes._subplots.AxesSubplot at 0x16b0a97ac88>



Analysis of the time remaining to apply Validity duration is an important parameter that tell us about restrincting applicants. Offering little time to apply reduces the number of candidates. We can expect a job opening to leave enough time to people to apply. For instance, disabled people may difficulties in access to the job opennings and may not have enough time to apply 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 open 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 number 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

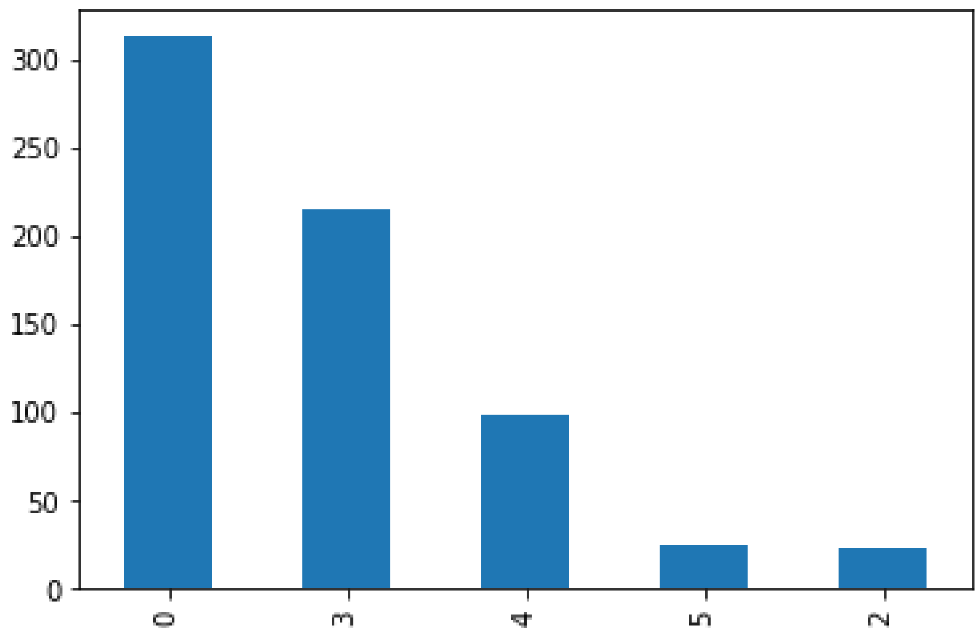
Resp_level based on the job title : - Director = 5, - Manager, Principal, Chief, Captain = 4, - Engineer, Specialist, Representative, Advocate, Inspector, Supervisor = 3 - Officer = 2 - Other = 0

Out[25]:
```
count                          675
unique                         665
top       campus interviews only
freq                             3
Name: Position, dtype: object
```

Out[28]:
```
        Position  nb_line_scale  nb_char_scale  Resp_level  Open_To_All  \
0   311 director              1              1           5            1
1     accountant              2              1           0            1

   Open_To_Mention  exp_years  exp_years  high_education  \
```

```
0                 0        1.0        1.0                  0
1                 1        NaN        NaN                  1

   full_time_part_time_code
0                          2
1                          0
```

Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x16b0aa0a710>

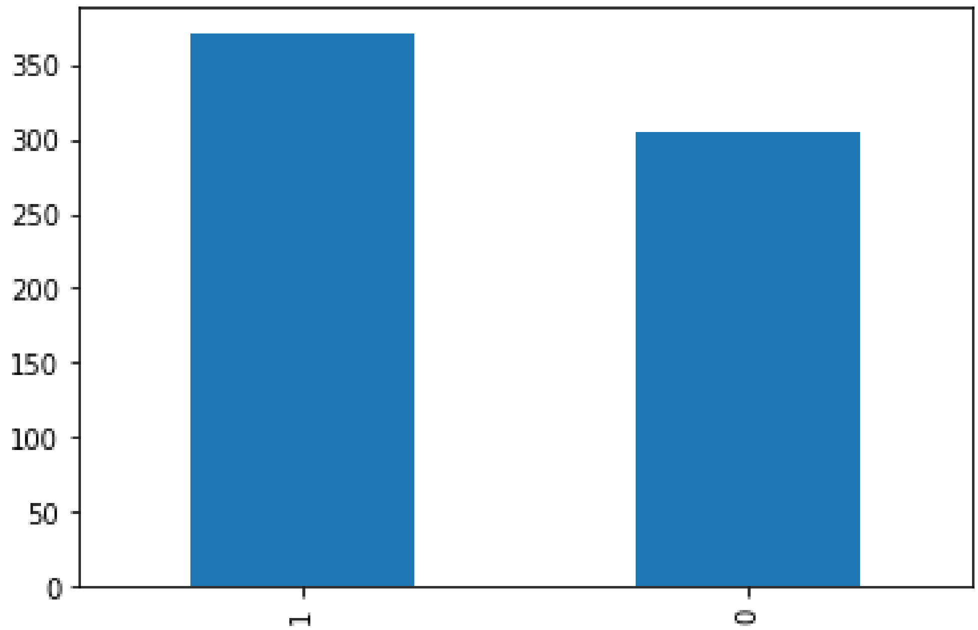


1.d.1 Responsibility Level

- 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 Ethic makes that the shoud not be displated bias when it comes to hire someone 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[34]: <matplotlib.axes._subplots.AxesSubplot at 0x16b0aa87dd8>



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 open 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 external candidate, this may reduce chances to have new candidates.

```
Out[35]: 2    576
         0     99
         Name: full_time_part_time_code, dtype: int64
```

1.d.3 Part time or full time ?

- The very large majority of the positions 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[36]: 0    553
         1    122
         Name: high_education, dtype: int64
```

1.d.4 High eductation or Not ?

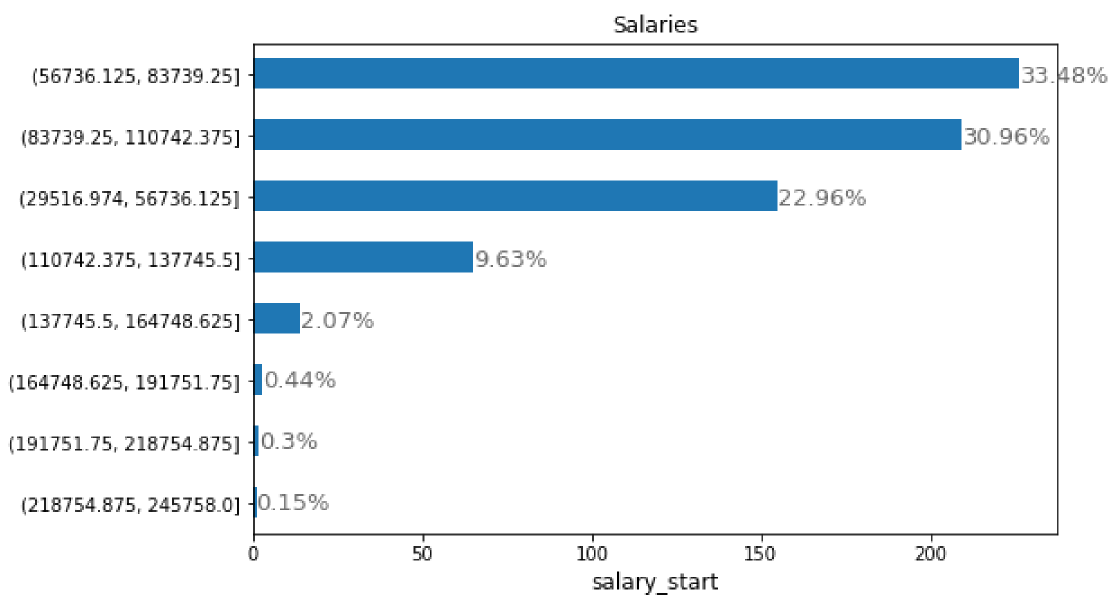
- Only 122 job positions require a college or university education (about 18%)
- The remaining bulletins DO NOT specify anything. We will assume that they don't require it.

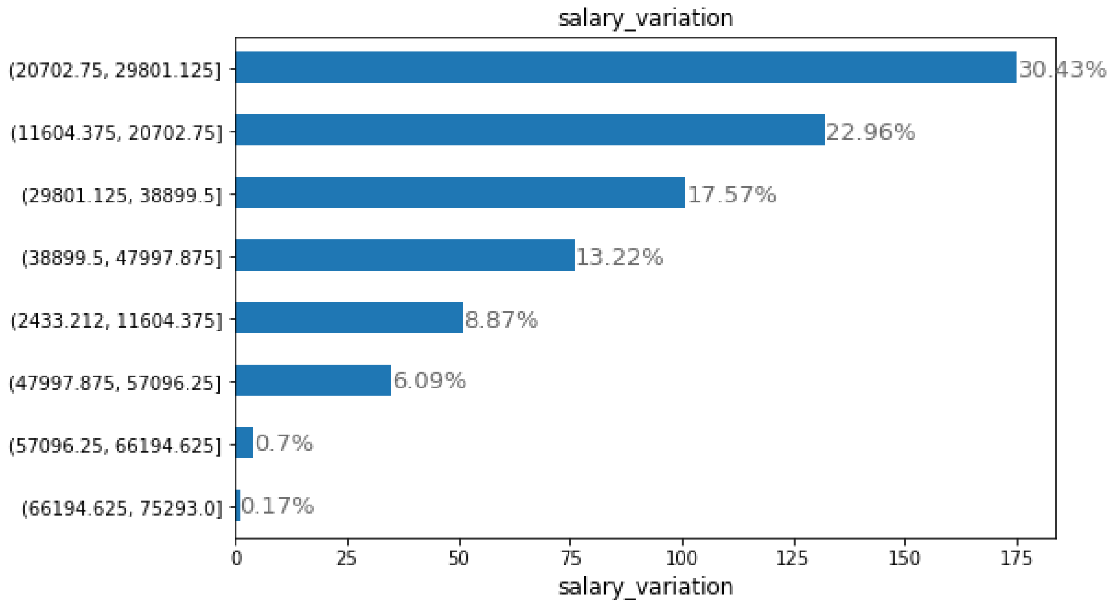
1.d.5 Salaries analysis

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

Out[41]:

	Position	salary_start	salary_start_code
0	311 director	125175	3
1	accountant	49903	1
2	accounting clerk	49005	1
3	accounting records supervisor	55332	1
4	administrative analyst	60489	1





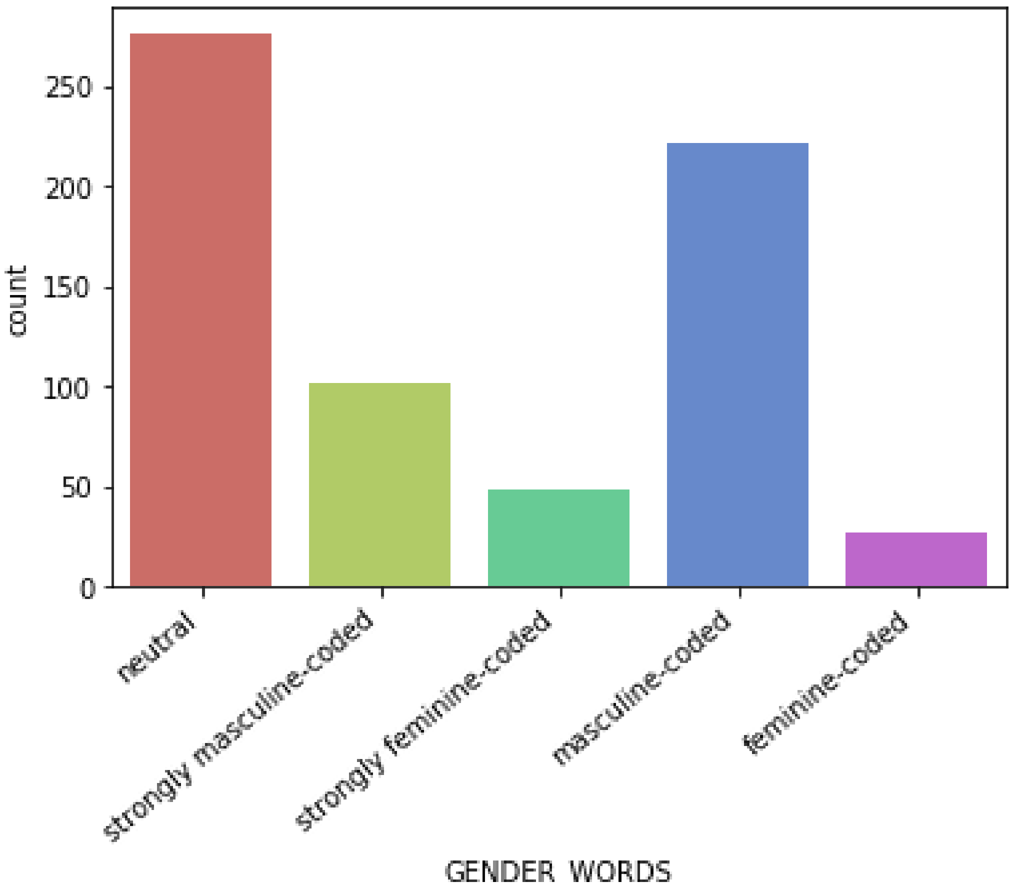
Analysis

- The majority of job postings offer a salary between 56k$ and 110k$
- Most jobs have a pretty big amplitude between 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[51]: 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[54]:

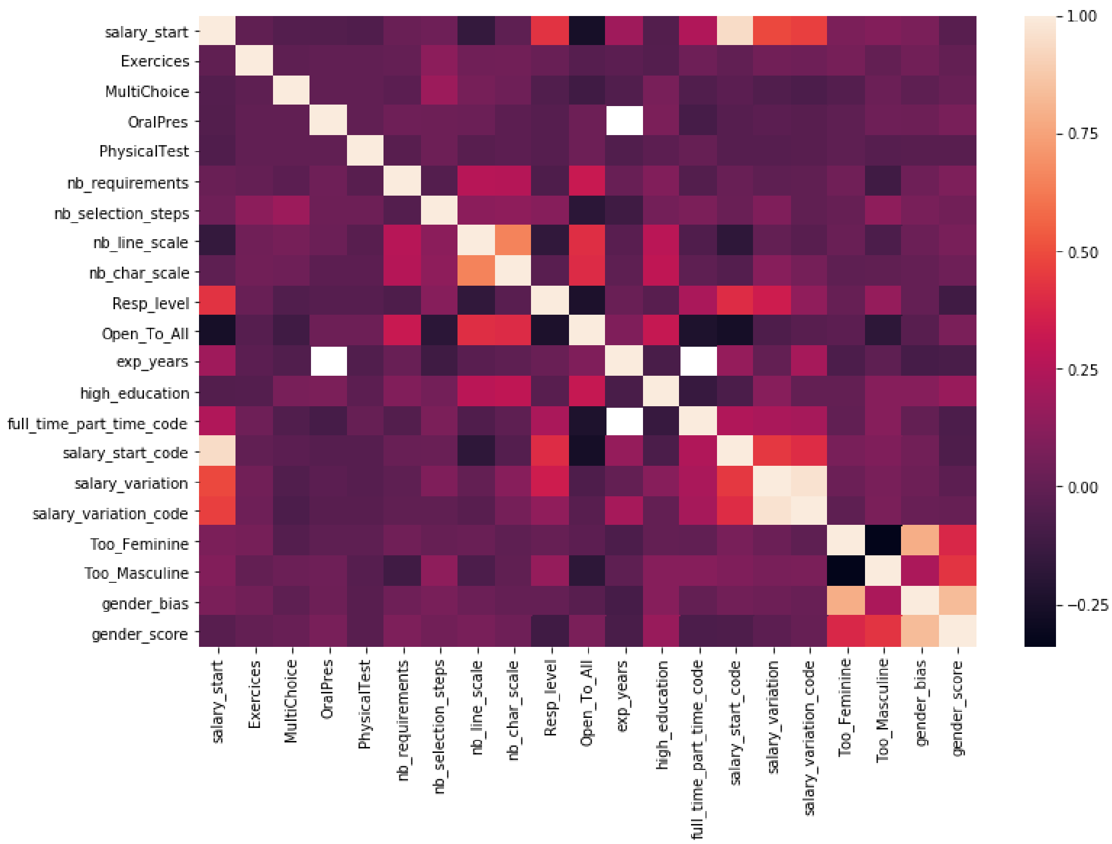
```
                       Position              GENDER_WORDS  Too_Feminine  \
0                  311 director                   neutral             0
1                    accountant  strongly masculine-coded             0
2              accounting clerk                   neutral             0
3  accounting records supervisor                   neutral             0
4        administrative analyst  strongly masculine-coded             0

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

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



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 bulletins 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 every 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[59]: 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[60]:                                          Position  score  Resp_level  \
         145        communications information representative    108           3
         617  transportation engineering associate              108           3
         163                    customer service representative    108           3
         195                   electrical engineering associate    108           3
         55               assistant director information systems    100           5
         172                       director of airport operations    100           5
         181                       director of printing services     100           5
         113                               chief forensic chemist     80           4
         120                                  chief of operations     80           4
         119                         chief of drafting operations     80           4

                         GENDER_WORDS  Too_Feminine  Too_Masculine  gender_bias  \
         145   strongly feminine-coded             1              0            4
         617  strongly masculine-coded             0              1            2
         163   strongly feminine-coded             1              0            4
         195  strongly masculine-coded             0              1            2
         55    strongly feminine-coded             1              0            4
         172  strongly masculine-coded             0              1            2
         181  strongly masculine-coded             0              1            2
         113  strongly masculine-coded             0              1            2
         120  strongly masculine-coded             0              1            2
         119   strongly feminine-coded             1              0            4
```

```
     gender_score
145             4
617             4
163             4
195             4
55              4
172             4
181             4
113             4
120             4
119             4
```

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 (when writing the job description) like: > - Jobs for women consist in secretary, communication or sales. > - Jobs for men consist in more technical jobs such as engineers.

2 5. Text Analysis

2.0.1 5.1 Named Entity Recognition

Let's see if we can find something interesting by getting NER out of the offers.

```
Out[236]:                        Position  \
          0                  311 director
          1                    accountant
          2              accounting clerk
          3  accounting records supervisor
          4        administrative analyst

                                                            NER
          0  {'CARDINAL': 35, 'MONEY': 3, 'NORP': 2, 'DATE'...
          1  {'ORG': 28, 'MONEY': 3, 'PRODUCT': 4, 'CARDINA...
          2  {'ORG': 24, 'MONEY': 2, 'DATE': 16, 'NORP': 5,...
          3  {'CARDINAL': 14, 'ORG': 25, 'MONEY': 4, 'PRODU...
          4  {'ORG': 38, 'DATE': 20, 'MONEY': 4, 'PRODUCT':...
```

We now have NER for each job bulletin. Let's go further.

```
<IPython.core.display.HTML object>
```

2.0.2 5.2 Word Cloud

Let's see what we can gather from the most used words in the different parts.

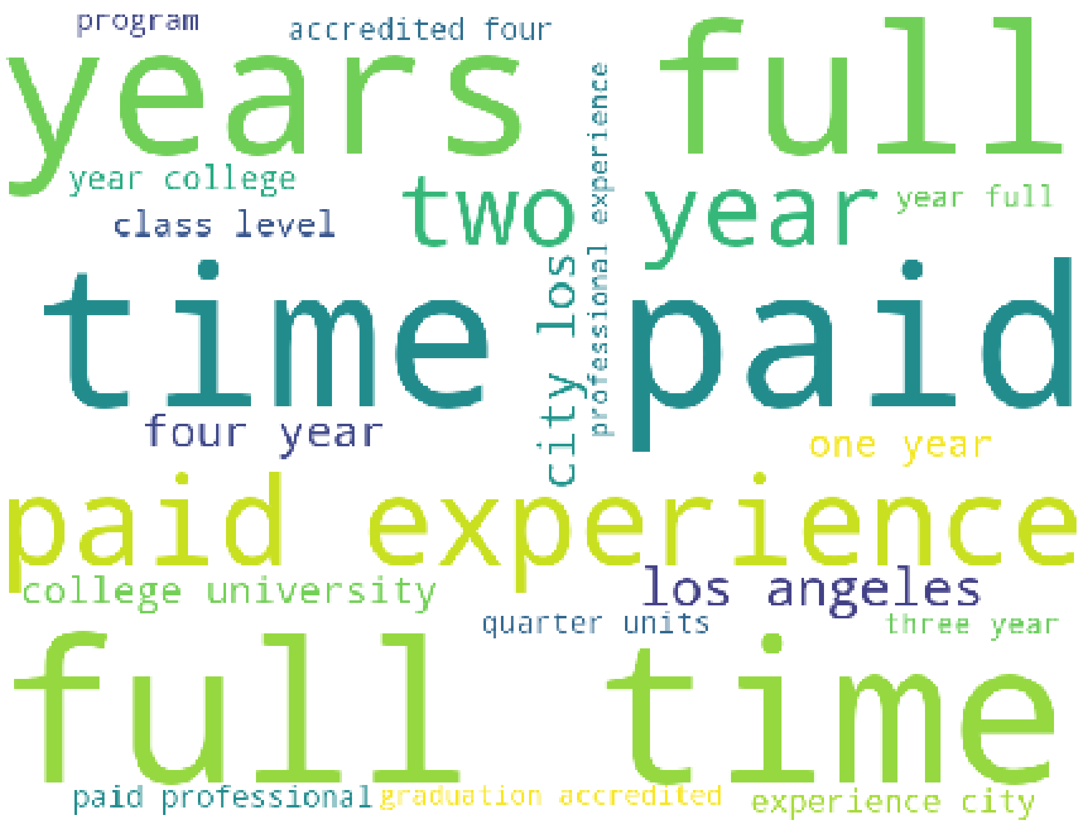
Let's perform a first test without word cloud on the first offer to see the most common words

```
Out[219]: [('City', 7), ('311', 5), ('1', 4)]
```

```
Out[220]: [('the City of Los Angeles', 343),
           ('Two years of full-time', 199),
           ('Two years', 180),
           ('four-year', 145),
           ('Four years', 136),
           ('1', 116),
           ('One year', 111),
           ('one year', 95),
           ('and2', 91),
           ('Three years', 85)]
```

Let's try our first word cloud on the duties, is there something poping out ?

What about requirements ?



From this word cloud, we can see that previous experience is often required in job offers.

3 6. Modeling

We managed to get a dataset with the number of applicants for some of the positions and the candidates 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]:

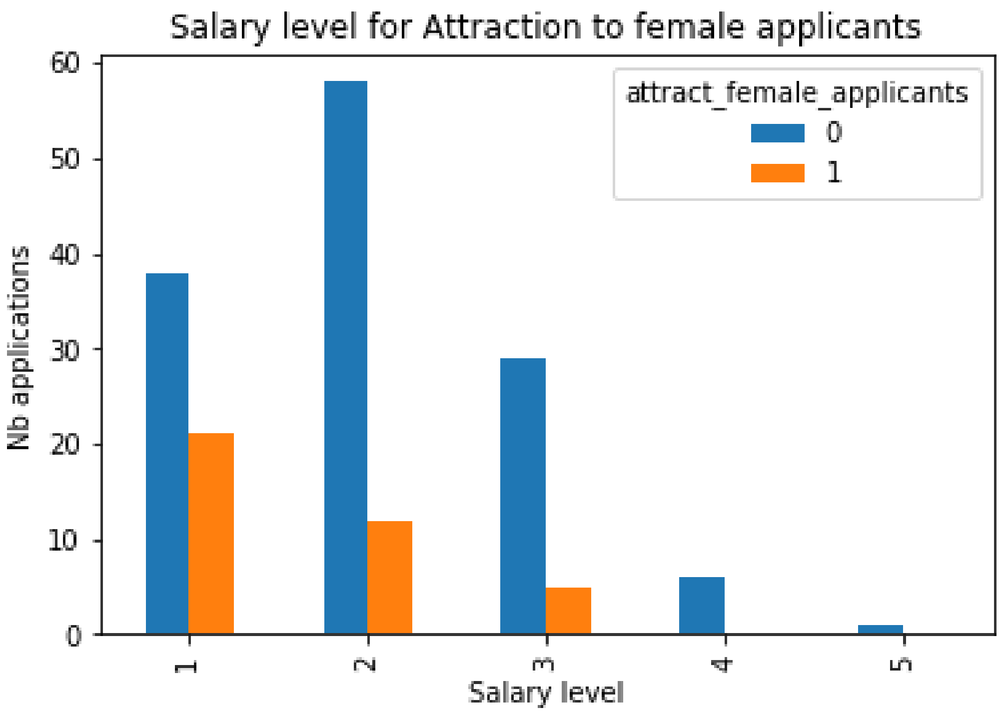
```
                                         File Name  \
0    CUSTOMER SERVICE REPRESENTATIVE 1230 020918.txt
1      CHIEF OF DRAFTING OPERATIONS 7271  042018.txt
2          CHIEF MANAGEMENT ANALYST 9182 020918.txt
3            CONSTRUCTION INSPECTOR 7291 042117.txt
4  FIRE PROTECTION ENGINEERING ASSOCIATE 7978 041...

                                          Position  salary_start  score salary_end  \
```

```
0        customer service representative         57148    108    $71,012
1           chief of drafting operations        135302     80   $168,084
2               chief management analyst        123667     80   $179,944
3                 construction inspector         80283     60    $97,092
4  fire protection engineering associate         66231     60    $96,841

  validity_duration nb_lines nb_chars  Exercices  MultiChoice  ...  \
0                11       89     9186        0.0          0.0  ...
1                13       80     8166        0.0          0.0  ...
2                13       76     7385        0.0          0.0  ...
3               NaN      133    15445        0.0          0.0  ...
4                13       89     9384        0.0          0.0  ...

   salary_variation_code  JobNumber  Fiscal Year                Job Number  \
0                      1       1230    2013-2014         1230 O 2013/12/27
1                      3       7271    2013-2014         7271 P 2013/11/08
2                      5       9182    2013-2014         9182 P 2014/06/20
3                      2       7291    2014-2015  7291 P 2014/07/04-ARCHIVE
4                      3       7978    2013-2014           7978 O 2014/6/6

                          Job Description  Apps Received  Female  Male  \
0   CUSTOMER SERVICE REPRESENTATIVE 1230          28230   19892  7968
1      CHIEF OF DRAFTING OPERATIONS 7271             13       2    11
2          CHIEF MANAGEMENT ANALYST 9182            143      78    54
3  CONSTRUCTION INSPECTOR 7291 - ARCHIVE            471      17   443
4  FIRE PROTECTION ENGINEERING ASSOCIATE            107      16    89

      ratio  attract_female_applicants
0  2.496486                          1
1  0.181818                          0
2  1.444444                          1
3  0.038375                          0
4  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')
```

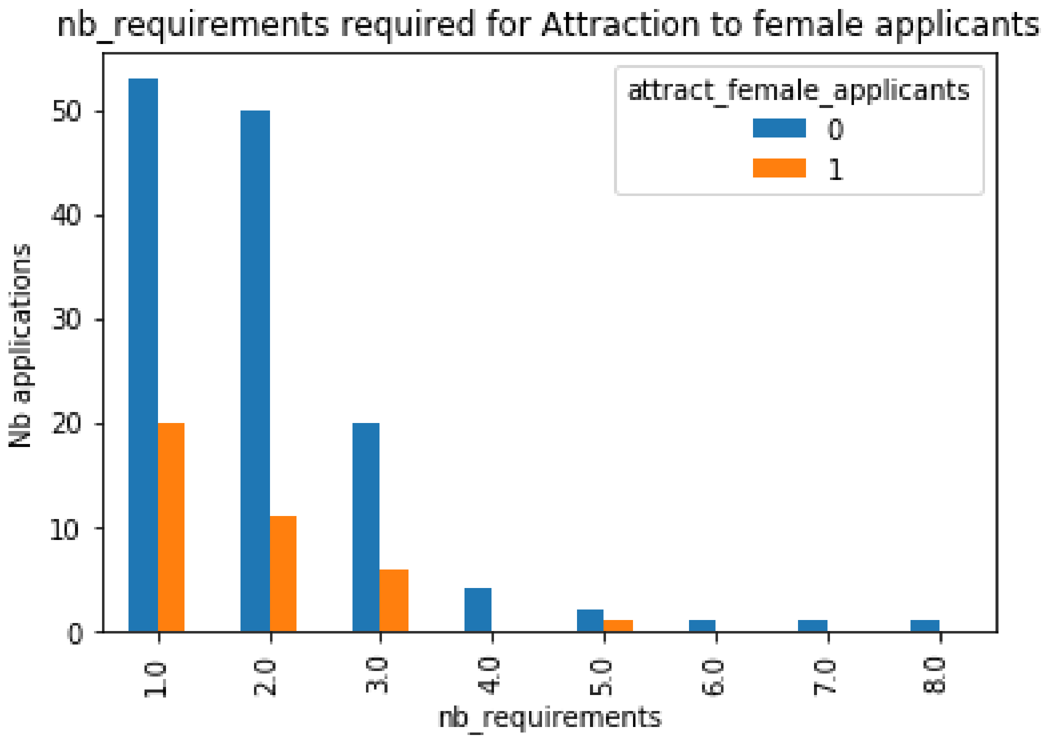
3.0.1 Let's analyse some relationships

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



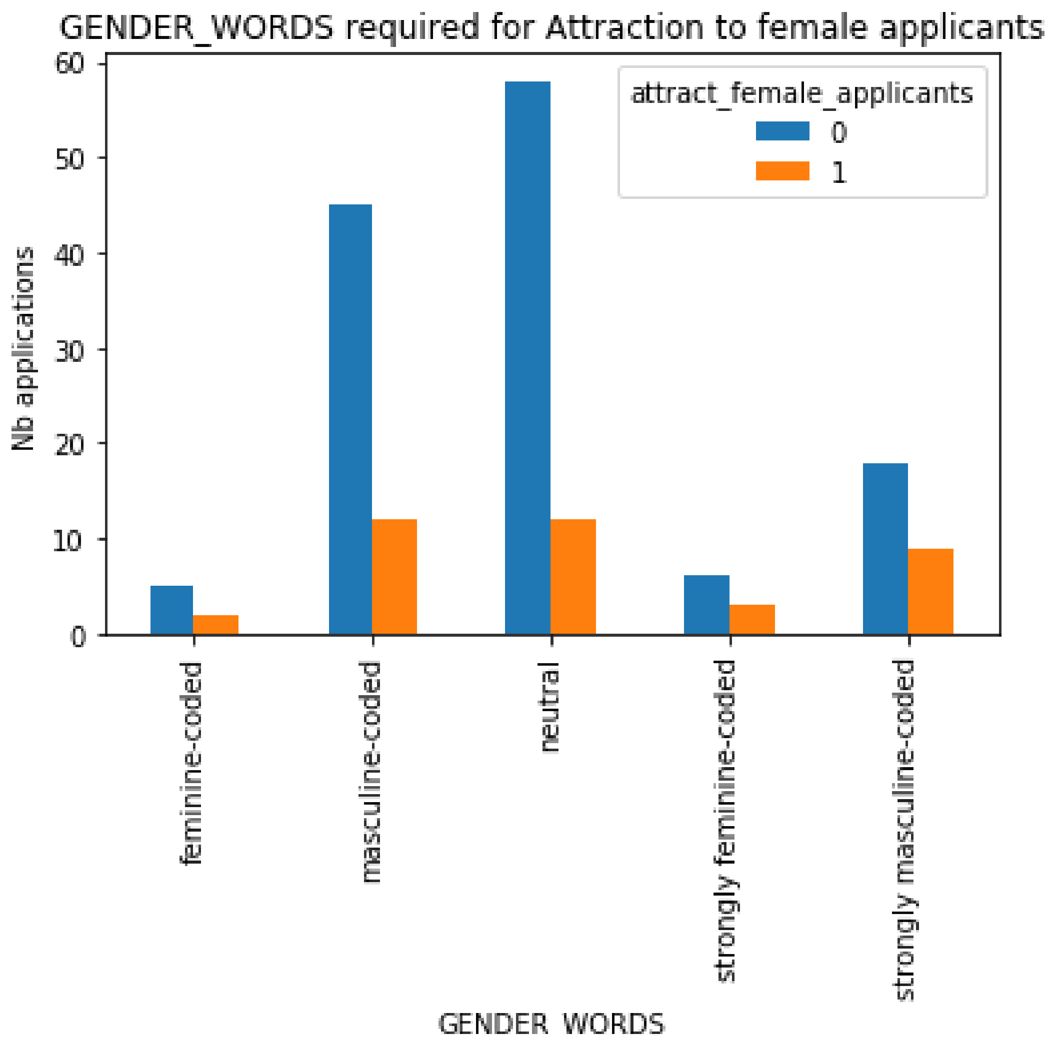
- globally, jobs attract men and women until the 4th level of salary where no job that offered the salary levels 4 and 5 (the highest) attracted more women than men.
- In fact, they all attracted more men than women to apply.

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



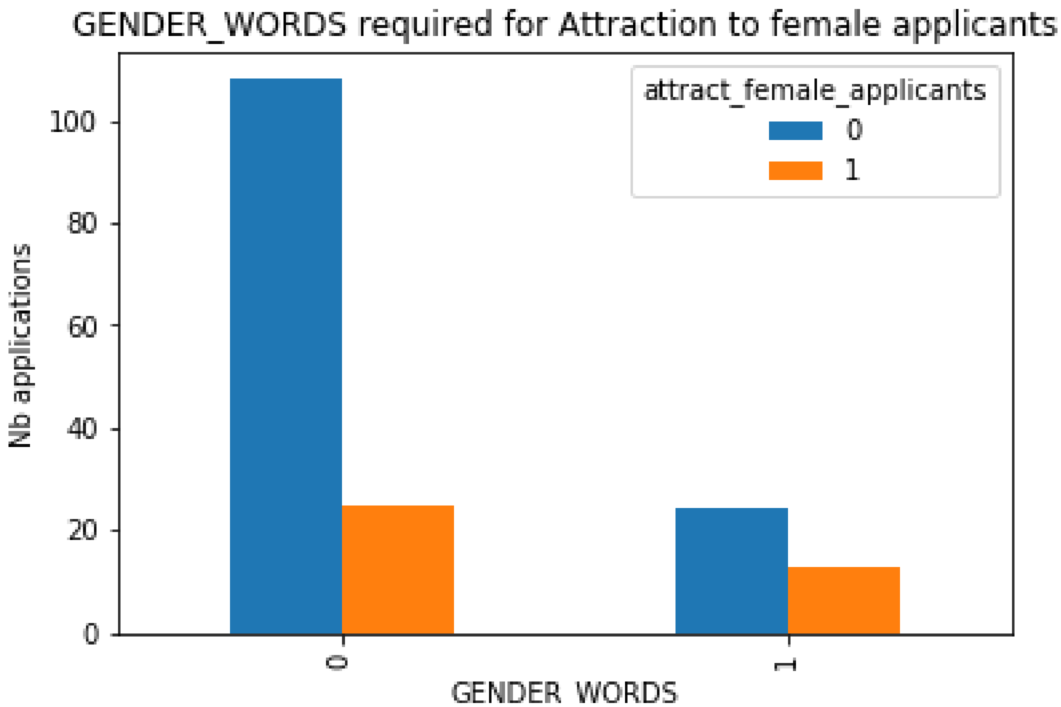
- The more the number of requiresments, the less applications are done mostly by women

```
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

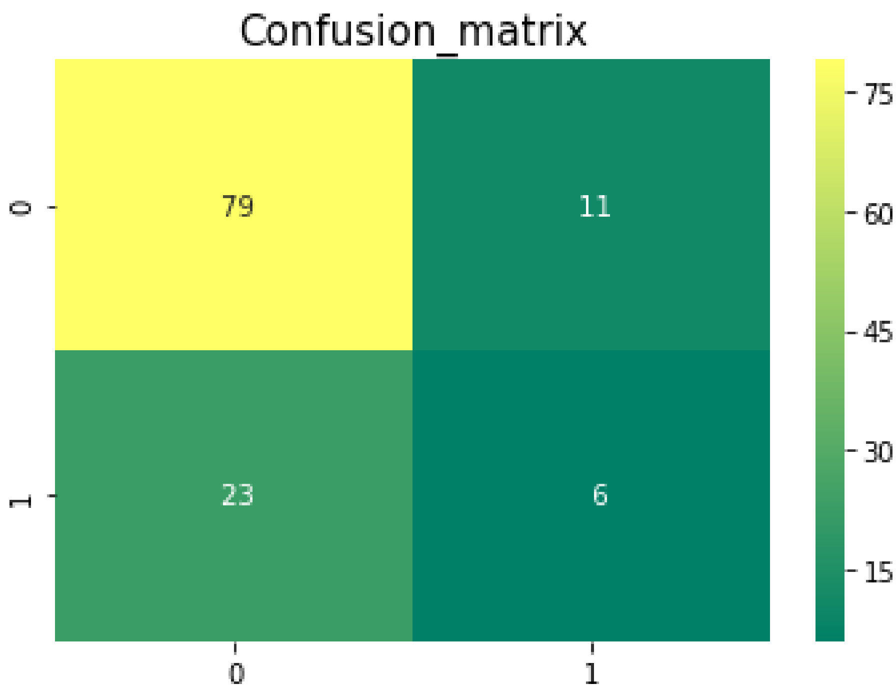
Let's get our train and test datasets from the labeled one

```
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 :

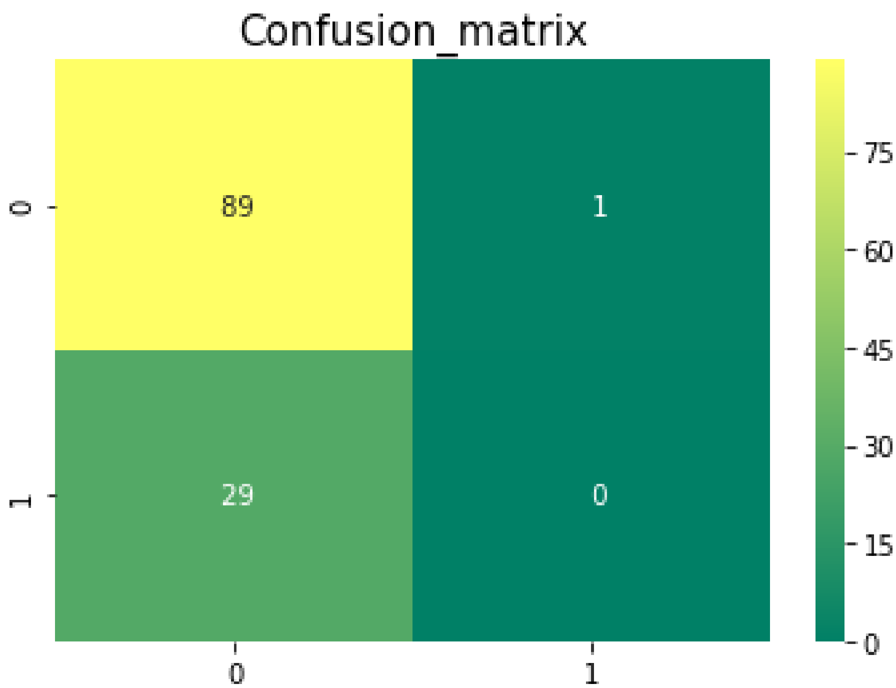
```
* 79 true positive were predicted as positive
* 6 true negative were predicted as negative
* 11 true positive were predicted as negative
* 23 true negative were predicted as positive
```

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
```

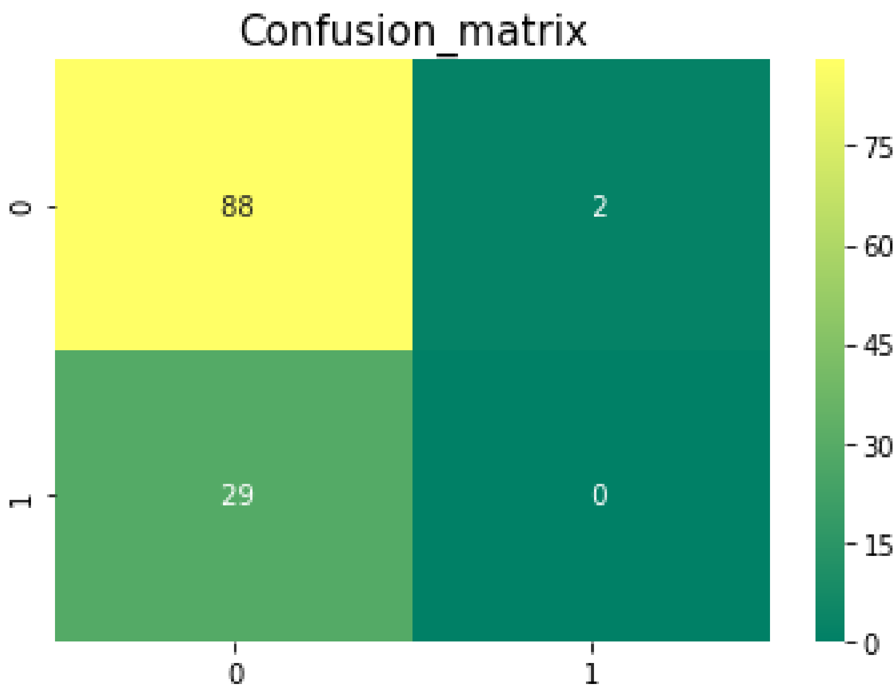
```
Out[860]: Text(0.5, 1.05, 'Confusion_matrix')
```



5.3.3 Random Forests

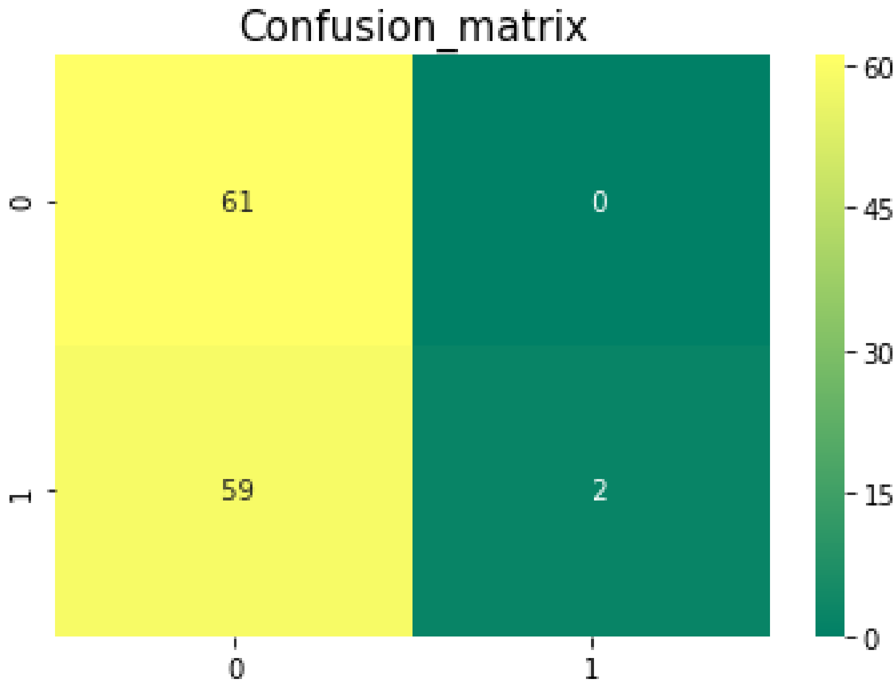
```
--------------The Accuracy of the model----------------------------
The accuracy of the Random Forest Classifier is 80.0
The cross validated score for Random Forest Classifier is: 73.94
```

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



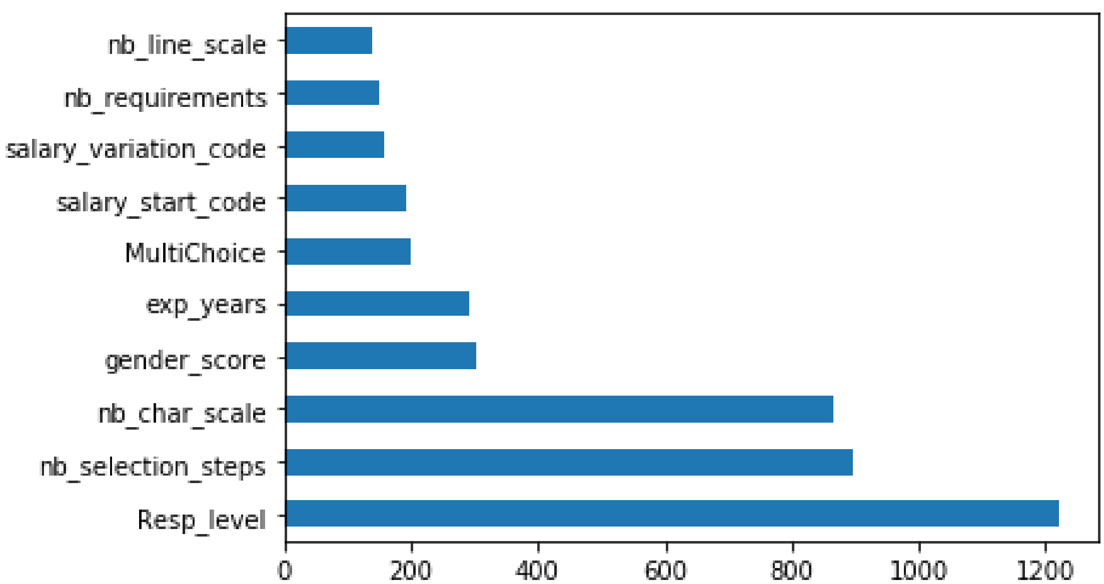

```
              precision    recall  f1-score   support

           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
```

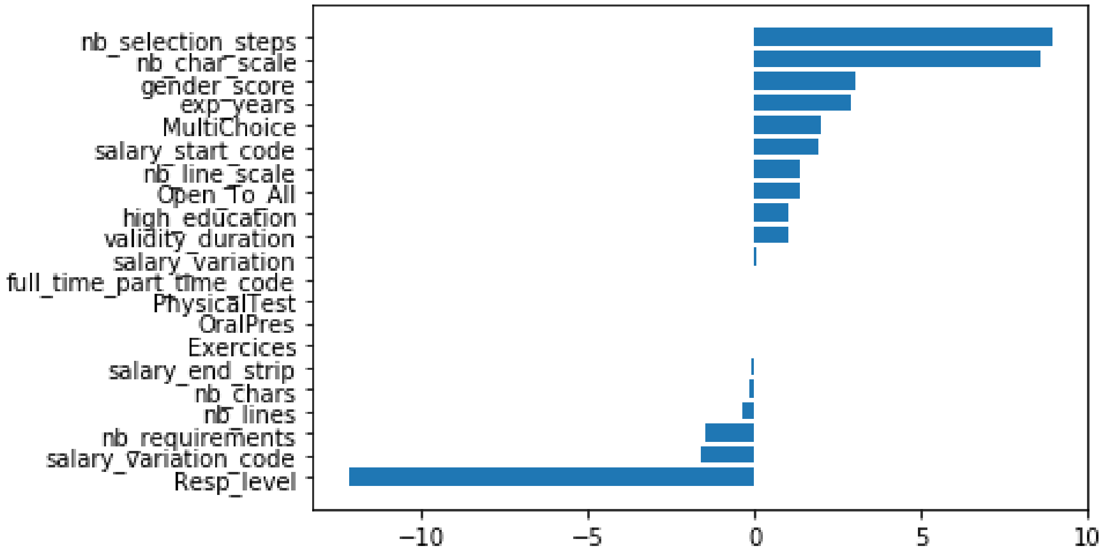
Out[925]: Text(0.5, 1.05, 'Confusion_matrix')



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>
```

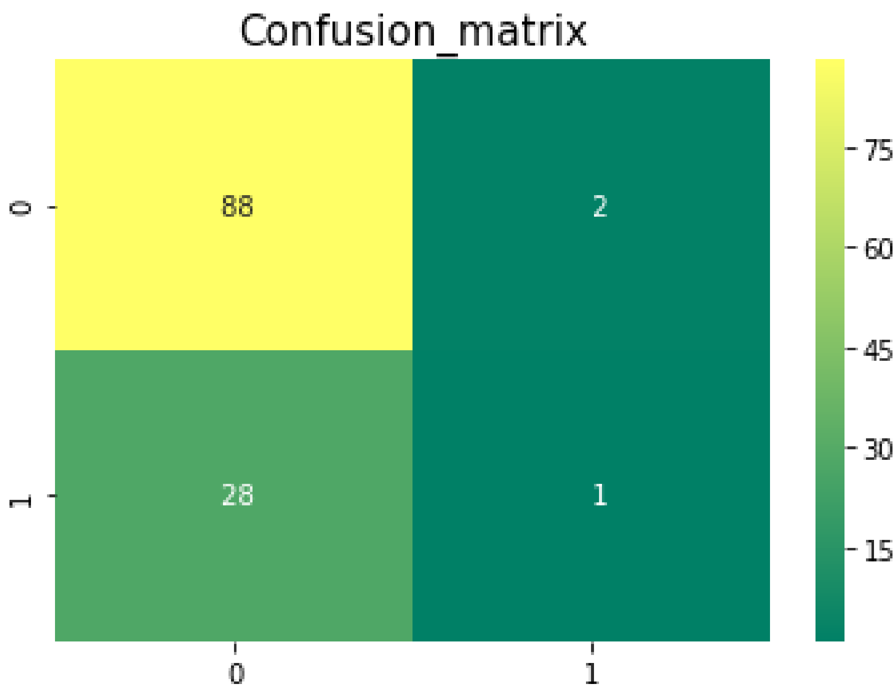




- The number of selection steps, the "gender score" (wether it's biased in favor of men or not), the years of experience and entry salary seem to be impactful in favor of female applications.
- On the other end, the responsability level seems to have a huge negative impact on female applications.

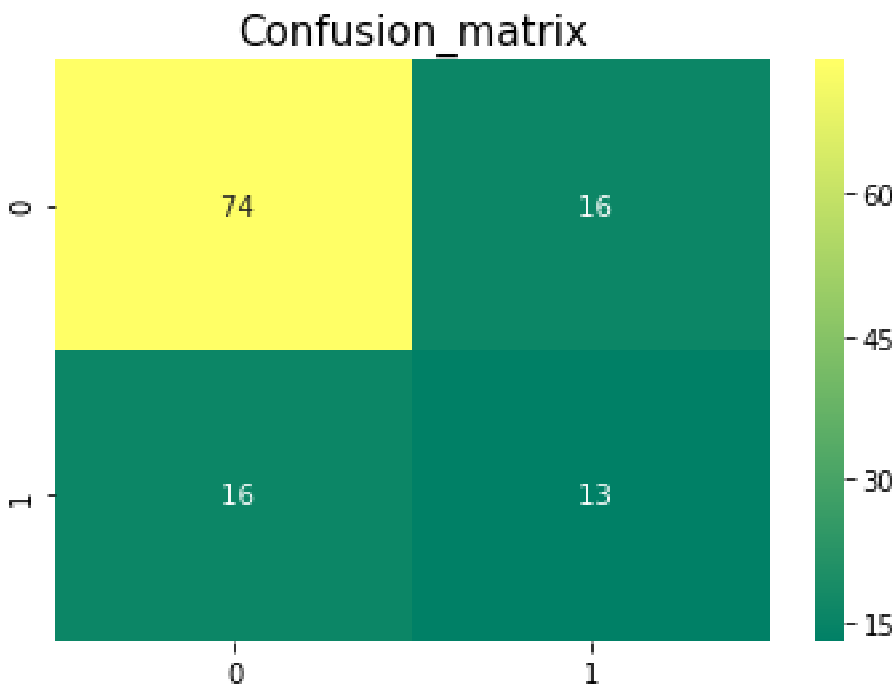
```
--------------The Accuracy of the model----------------------------
The accuracy of the K Nearst Neighbors Classifier is 80.0
The cross validated score for K Nearest Neighbors Classifier is: 74.77
```

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



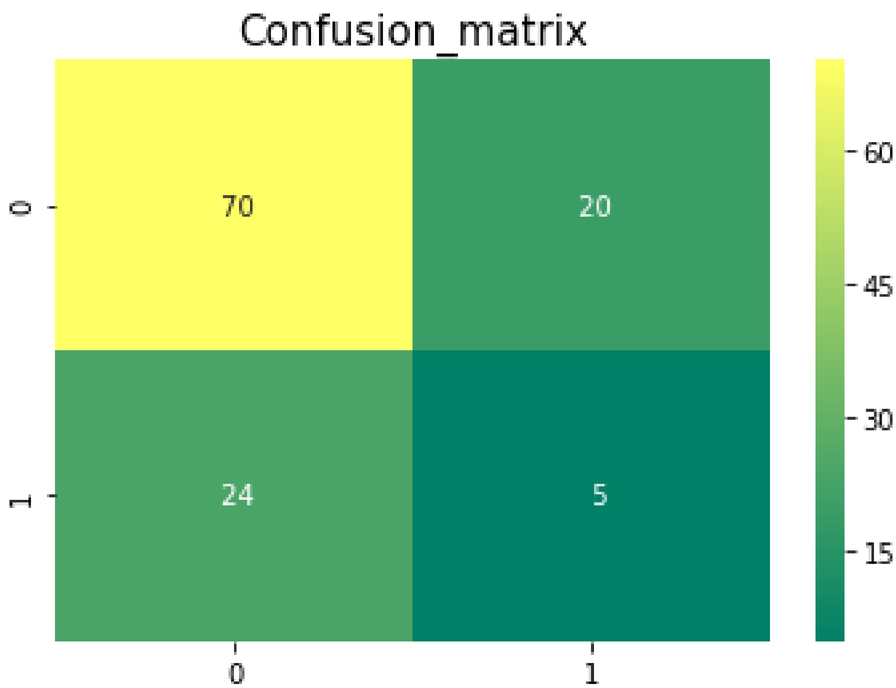

```
--------------The Accuracy of the model----------------------------
The accuracy of the Gaussian Naive Bayes Classifier is 70.0
The cross validated score for Gaussian Naive Bayes classifier is: 73.26
```

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



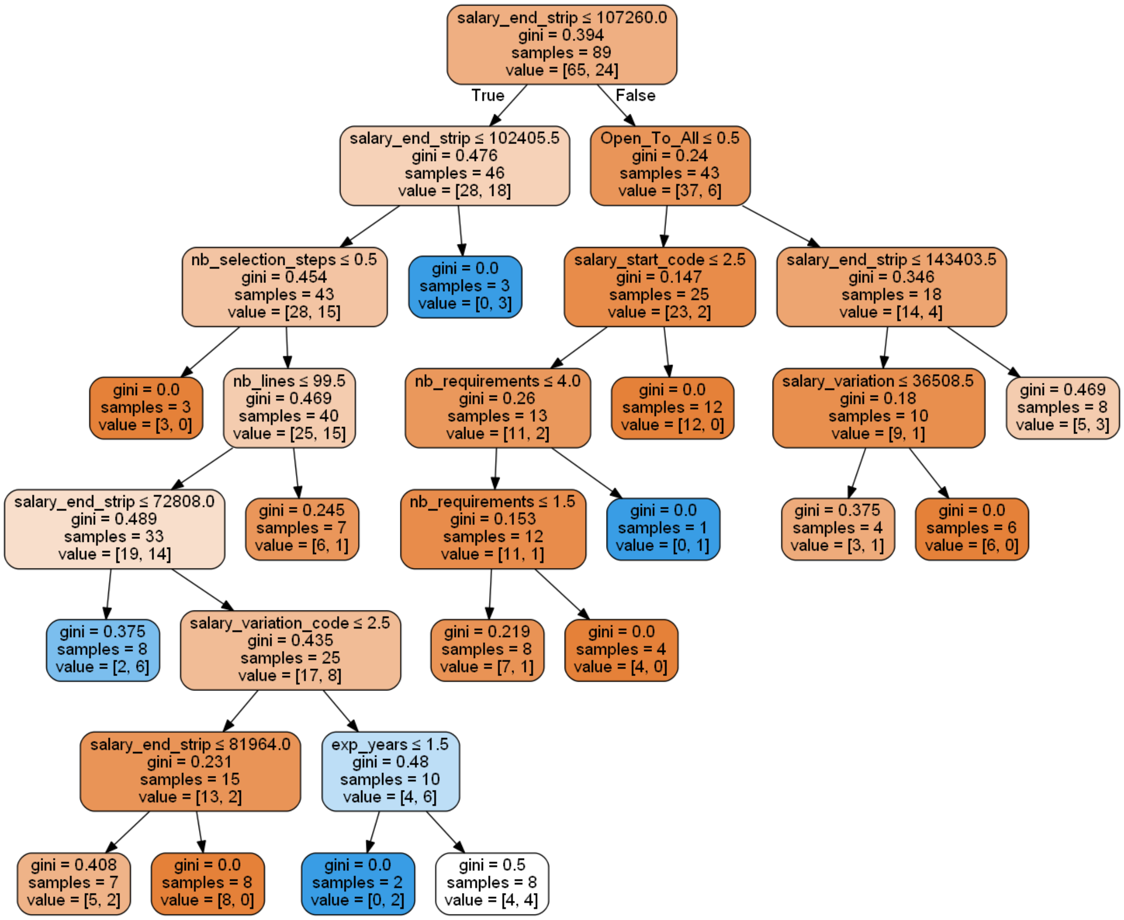

```
--------------The Accuracy of the model----------------------------
The accuracy of the DecisionTree Classifier is 66.67
The cross validated score for Decision Tree classifier is: 70.53
```

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



Let's check the decision tree in detail :

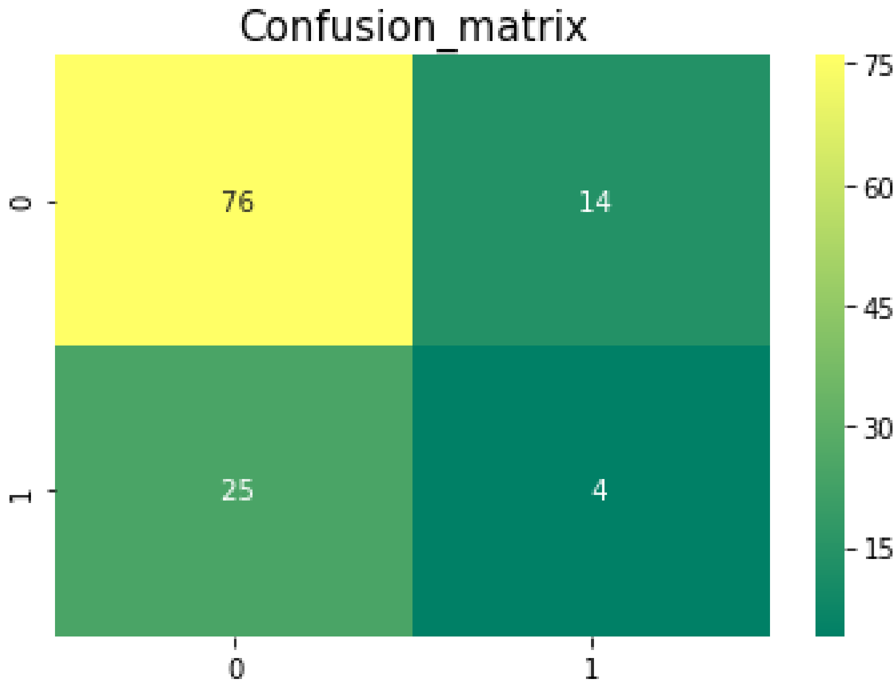
Out[869]:



- 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.

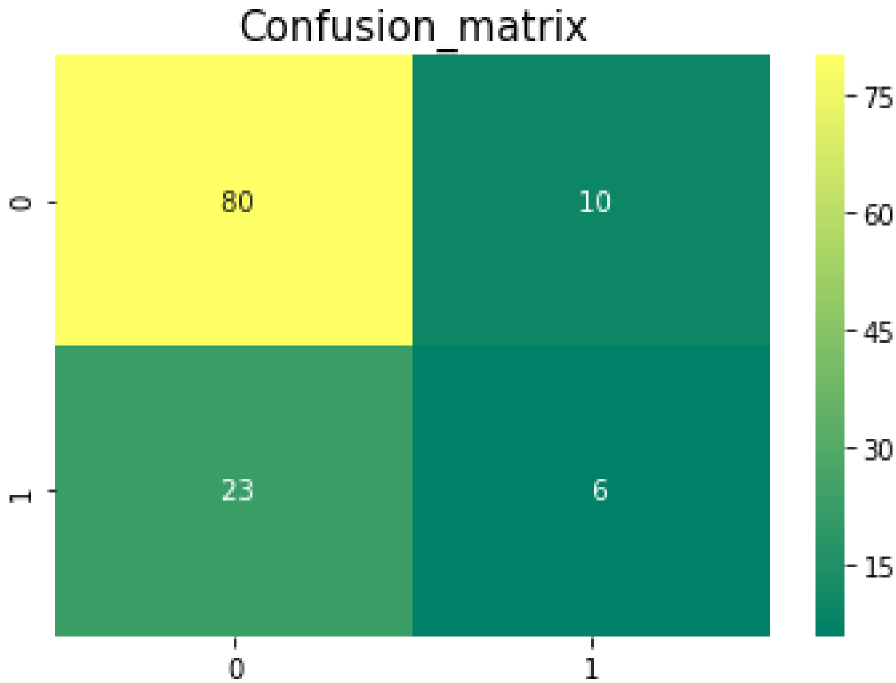
```
--------------The Accuracy of the model----------------------------
The accuracy of the AdaBoostClassifier is 76.67
The cross validated score for AdaBoostClassifier is: 67.12
```

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



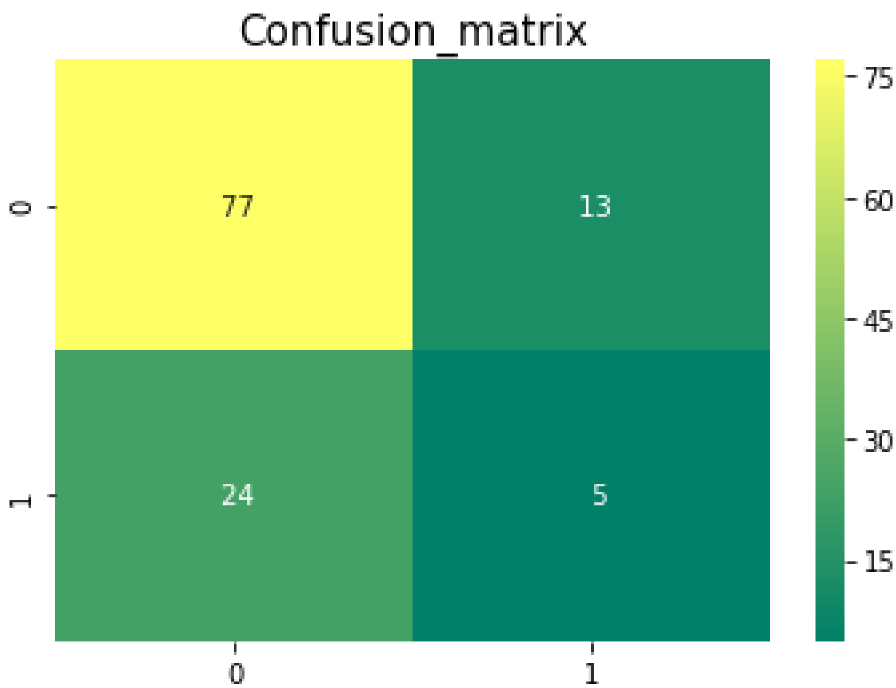

```
--------------The Accuracy of the model----------------------------
The accuracy of the LinearDiscriminantAnalysis is 60.0
The cross validated score for AdaBoostClassifier is: 72.27
```

```
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

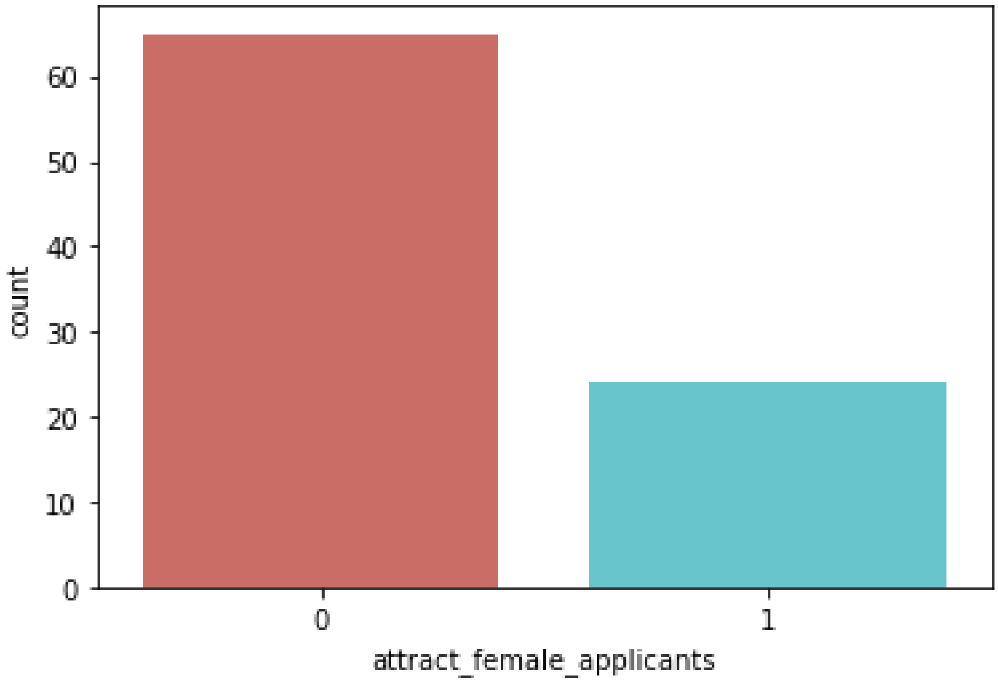
- The Support Vector Machines seems to be the best model, we should try it on our dataframe !

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  \
        145                  11       111    13006              3.0                 2.0
        163                  11        89     9186              1.0                 2.0
        55                   13        77     8157              1.0                 2.0
        172                  13        80     9286              3.0                 1.0
        181                  20        74     8078              2.0                 1.0

             nb_line_scale  nb_char_scale  Resp_level  Open_To_All  exp_years  \
        145              3              4           3            1        1.0
        163              2              1           3            1        2.0
        55               1              1           5            0        2.0
        172              1              1           5            1        4.0
        181              1              1           5            0        3.0

             high_education  full_time_part_time_code  predicted
        145               0                         2          1
        163               0                         2          0
        55                0                         2          1
        172               0                         2          1
        181               0                         2          1
```

```
Out[450]:                                             File Name  \
        145  COMMUNICATIONS INFORMATION REPRESENTATIVE 1461...
        163    CUSTOMER SERVICE REPRESENTATIVE 1230 020918.txt

                                      Position salary_start  score salary_end  \
        145  communications information representative       41,697    108    $59,340
        163            customer service representative       57,148    108    $71,012

            validity_duration_x nb_lines_x nb_chars_x  nb_requirements_x  \
        145                  11        111      13006                3.0
        163                  11         89       9186                1.0

             nb_selection_steps_x  ...  nb_requirements_y  nb_selection_steps_y  \
        145                   2.0  ...                3.0                   2.0
        163                   2.0  ...                1.0                   2.0

             nb_line_scale_y  nb_char_scale_y  Resp_level_y  Open_To_All_y  \
        145                3                4             3              1
        163                2                1             3              1

             exp_years_y high_education_y  full_time_part_time_code_y  predicted
        145          1.0                0                           2          1
        163          2.0                0                           2          0

        [2 rows x 36 columns]
```

```
Out[232]: Index(['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_line_scale_x', 'nb_char_scale_x',
       'Resp_level_x', 'Open_To_All_x', 'exp_years_x', 'high_education_x',
       'full_time_part_time_code_x', 'GENDER_WORDS', 'Too_Feminine',
       'Too_Masculine', 'gender_bias', 'gender_score', 'JobNumber',
       'validity_duration_y', 'nb_lines_y', 'nb_chars_y', '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'],
      dtype='object')
```

Out[233]:

	score	gender_bias	GENDER_WORDS	gender_score	predicted
145	108	4	strongly feminine-coded	4	1
163	108	4	strongly feminine-coded	4	0
55	100	4	strongly feminine-coded	4	1
172	100	2	strongly masculine-coded	4	1
181	100	2	strongly masculine-coded	4	1

The predicted column indicates if female will be more likely to apply than men. Here our prediction is not really aligned with the previous treatment we had done.