Qi Haodi

Haodi.qi@gmail.com

Abstract

The analysis consists of three main parts:  
Exploratory Data Analysis on Salary and Firm Data (Section 2-3)  
Correlation Matrix and Decision Tree Regression Model (Section 4-6)  
Distribution of Salary, Male-Female Ratio and Population across Departments (section 7-9)

dATASET ANALYSIS:  
FRENCH EMPLOYMENT BY TOWN

Contents

[Section 2: Exploratory Data Analysis on Salary Data 2](#_Toc4626564)

[Section 3: Exploratory Data Analysis on Firm Data 3](#_Toc4626565)

[Section 4: Correlation Matrix and Model: Gender, Position vs Salary 4](#_Toc4626566)

[Section 5: Correlation Matrix and Model: Gender, Age vs Salary 4](#_Toc4626567)

[Section 6: Random Forest Regressor 5](#_Toc4626568)

[Section 7: Salary distribution in different departments 5](#_Toc4626569)

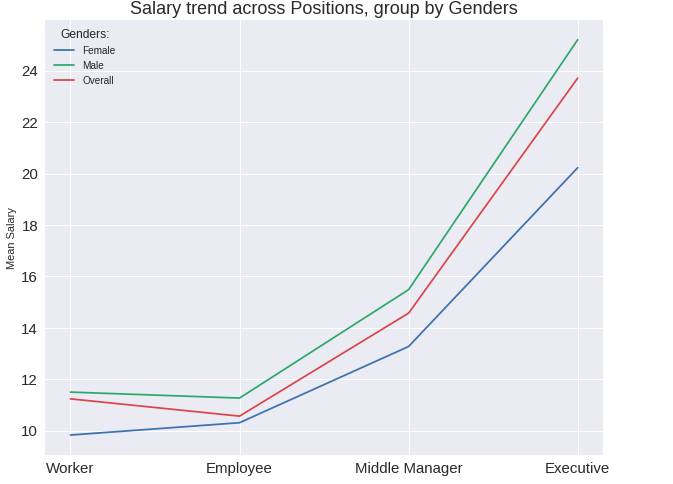
[Section 8: Male-Female distribution in different departments 6](#_Toc4626570)

[Section 9: Population across ages in different departments 7](#_Toc4626571)

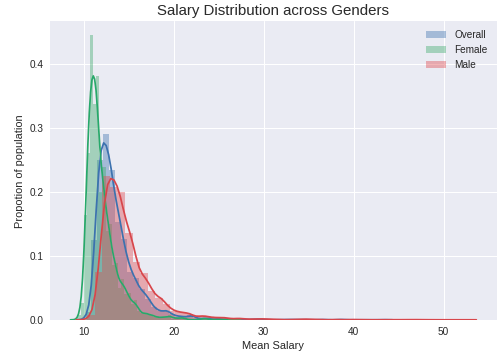
## Section 2: Exploratory Data Analysis on Salary Data

The salary data contains mean salary across, genders, ages and position levels. They are plotted in the two graphs below for trend identification. The third graph is salary distribution in the population.

The following insights are generated from the graphs:

* Males are earning higher than women of the same age or position.
* The wage differential increases with age.
* The wage differential increases at higher level of management positions, meaning that as females moves up the ladder, the increase in pay is much lower than that of males.
* The overall mean salary of workers does not follow the trend and is higher than that of employee.
* Most females are earning at the wage rate at which very few males are earning.

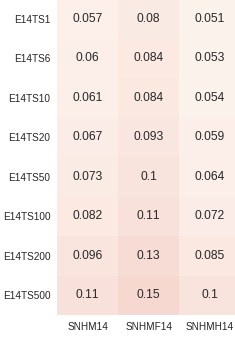
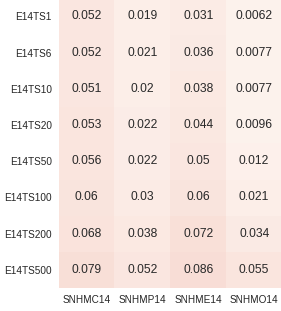
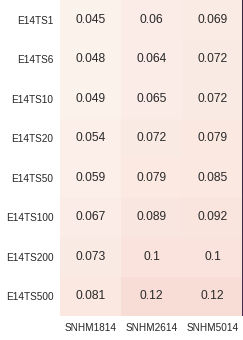
As such, a clear gender inequality can be identified in terms of wage. The inequality worsens as one grows older or reaches a higher management position.



## Section 3: Exploratory Data Analysis on Firm Data

Hypothesis: Towns with larger firms should be more developed and thus offer more job opportunities with a higher salary.

Correlation Matrix between Salary and 1) Genders; 2) Positions 3) Ages



Based on the three graphs:

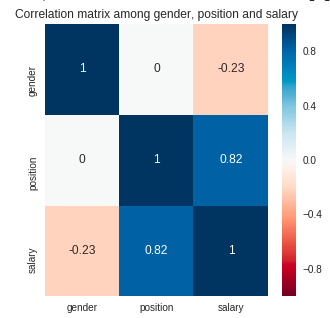
* Larger firms will indeed lead to higher mean salaries, regardless of gender, age and positions.
* It is interesting to see that firm sizes are more positively correlated with female mean salary than male, suggesting that females in less developed towns may not get employed or receive very low wage.
* It is interesting to see that firm sizes are more positively correlated with employee mean salary than middle manager.
* The correlations with mean salary of age 26 to 50 and that of above 50 are the same. It may suggest that firms now do not focus on seniority rather than performances. However, the reason why they are still higher than that of 18 – 25 may be that the older one is, the more working experience he has and thus the better job performance he achieves.

As such, there may be wage inequalities across towns due to the presence of firms of varying sizes. Despite the positive correlation, the values are very small, suggesting minimal impact of firm sizes on salary. Therefore, the hypothesis is proven wrong.

## Section 4: Correlation Matrix and Model: Gender, Position vs Salary

Correlation matrix is used to further examine the strengths of relationships among gender, position and salary.

Based on the matrix, position is strongly positively correlated with salary while gender is negatively correlated with salary. The positive correlation between position and salary suggests the higher the position one has, the more wages he will earn, as values increase with the level of positions. The negative correlation between position and gender suggests males are earning a higher salary than females, as females are denoted with 2 and males are denoted as 1.

Given mainly categorical attributes, decision tree regressor is used to predict salaries.

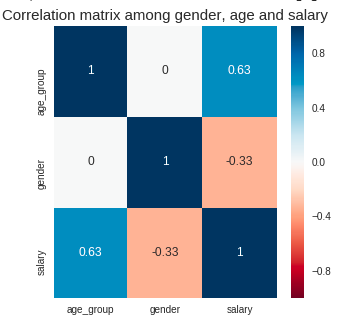
The R square value is 0.881 and the mean standard error (MSE) is 3.44

The R square value shows the model is useful.

## Section 5: Correlation Matrix and Model: Gender, Age vs Salary

The strengths of relationships among gender, age and salary are also examined using correlation matrix

Based on the matrix, age is strongly positively correlated with salary. The positive correlation between age and salary suggests the higher the one’s age is, the more wages he will earn, as values of age group increase with age.

Similar to the previous section, decision tree regressor is used.

The R square value is 0.478 and the mean standard error (MSE) is 7.53

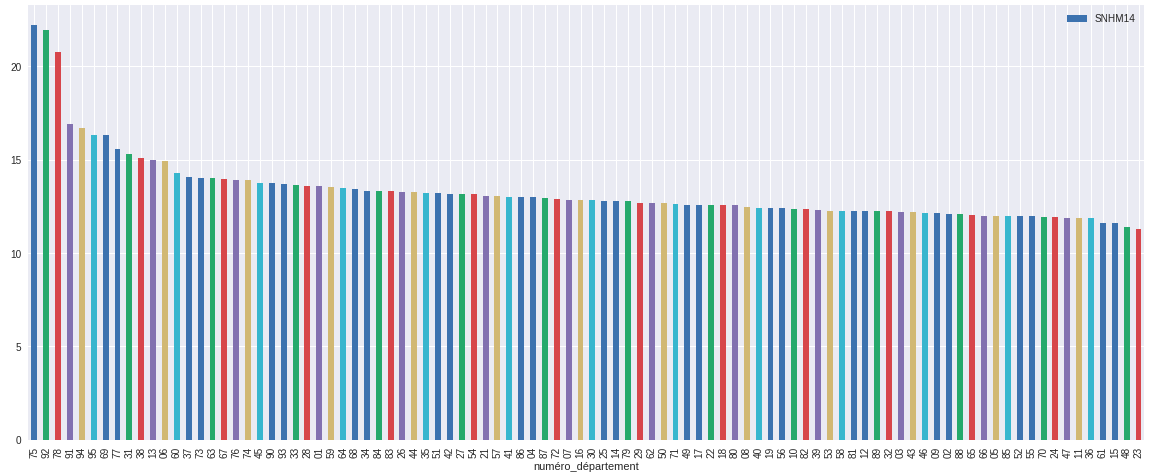
The R square value shows the model is not useful. Gender and age are not good predictor of salary.

It is worth noted that the correlation between gender and salary becomes more negative, suggesting possible interaction effect between gender and age, and gender and position. However, due to time constraint, this is not examined in the analysis.

## Section 6: Random Forest Regressor

To build a more accurate model, random forest regressor is used. The R square value is 0.885, which is slightly better than the first model (section 5). Considering the complexity of two models, the first model is good enough.

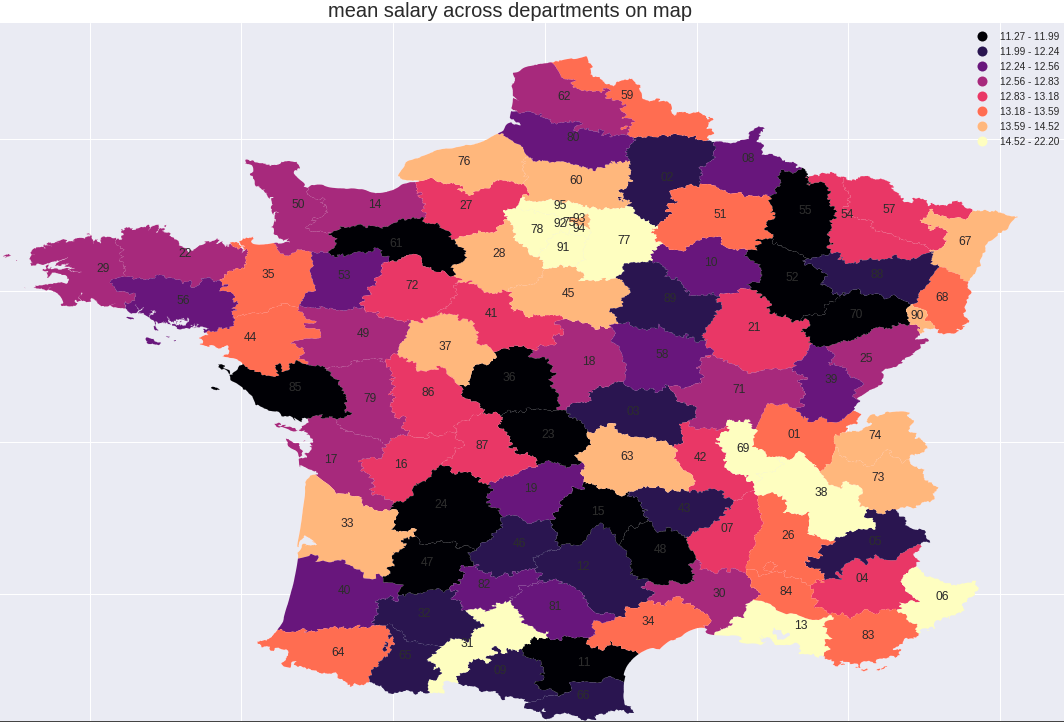
## Section 7: Salary distribution in different departments

To have to bigger picture, starting from this section, the analysis will be more on the department level.

The overall mean salary across departments plot above shows two distinct groups: 1) the top three departments with significantly higher salary than the rest; 2) the rest of departments with varying salaries but the salaries are not far from one another.

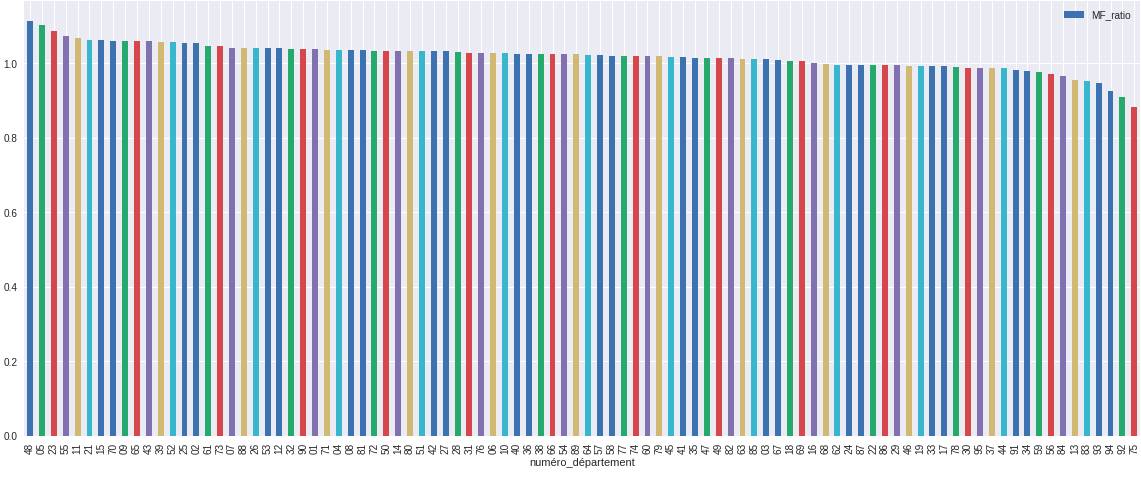
The top three departments are Paris, Yvelines and Hauts-de-Seine, which are found to be at the top departments in GDP per capita.

The distribution is also plotted on the map. The north and the south east regions are more developed than other regions, shown by a higher mean salary.



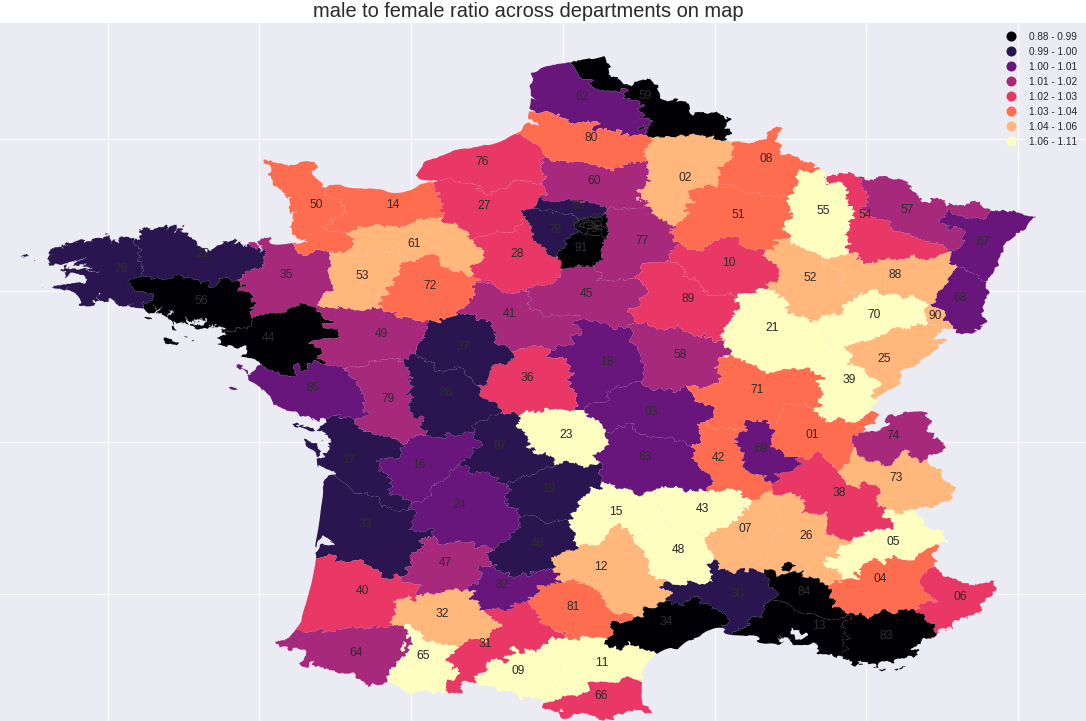
## Section 8: Male-Female distribution in different departments

Since males are earning a higher salary, towns with more males than females should have a higher overall mean salary.



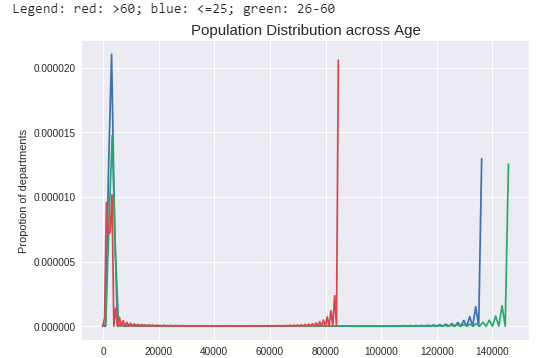
Based on the graph of Male-Female Ratio (mf-ratio) across departments is plotted above shows that more than two third of towns have more males than females, with a mf-ratio more than 1. Overall the balance in terms of gender of the population is rather balanced.

The map below shows that the departments at the east side tends to have a higher mf-ratio while those at the west side tend to have lower.



The correlation matrix between mf-ratio and salaries across positions and ages are mostly positive but the values are too small to render them useful.

## Section 9: Population across ages in different departments

The population distribution across towns are plotted below. 

The graph clearly suggests two categories of town in terms of population size. It is interesting to note that for the group of towns with smaller population (left side of graph), more departments tend to have more population of <=25 years old, followed by 26-60 and then > 60 years old. However, the other group tends to have more people with age of 26-60, followed by <=25 and > 60.

The map below shows that except for the region around Paris, the departments at the border tend to have a higher population.

