ANALYSIS AND PREDICTION OF FRENCH EMPLOYMENT SECTOR

Optimization for data science

Abhigna Doddiganahalli chandrashekar
Krishna Teja Kancherla

Objective

In this project we will analyze publicly available job data of France with an aim to uncover social trends such as urbanization, Regional variation of industry/company concentration, Regional variation in payscale, gender based payscale difference.

Finally, we predict regional “mean_net_salary” using the predictors “mean net salary for women” and “mean net salary for men”

We also try to analyze ratio between Population of a region and job concentration inorder to isolate causes of urbanization.

Data

Four files are in the dataset :

base_etablissement_par_tranche_effectif :
Information on the number of firms in every french town, categorized by size , come from INSEE.

*CODGEO : geographique code for the town (can be joined with *code_insee* column from "name_geographic_information.csv')*
LIBGEO : name of the town (in french)
REG : region number
DEP : depatment number
E14TST : total number of firms in the town
E14TS0ND : number of unknown or null size firms in the town
E14TS1 : number of firms with 1 to 5 employees in the town
E14TS6 : number of firms with 6 to 9 employees in the town
E14TS10 : number of firms with 10 to 19 employees in the town
E14TS20 : number of firms with 20 to 49 employees in the town
E14TS50 : number of firms with 50 to 99 employees in the town
E14TS100 : number of firms with 100 to 199 employees in the town
E14TS200 : number of firms with 200 to 499 employees in the town
E14TS500 : number of firms with more than 500 employees in the town

name_geographic_information :
gives geographic data on french town (mainly latitude and longitude, but also region / department codes and names)

EU_circo : name of the European Union Circonscription
code_région : code of the region attached to the town
nom_région : name of the region attached to the town
chef.lieu_région : name the administrative center around the town
numéro_département : code of the department attached to the town
nom_département : name of the department attached to the town
préfecture : name of the local administrative division around the town

numéro_circonscription : number of the circumpscription
nom_commune : name of the town
codes_postaux : post-codes relative to the town
code_insee : unique code for the town
latitude : GPS latitude
longitude : GPS longitude
éloignement : i couldn't manage to figure out what was the meaning of this number

net_salary_per_town_per_category :
salaries around french town per job categories, age and sex

CODGEO : unique code of the town
LIBGEO : name of the town
SNHM14 : mean net salary
SNHMC14 : mean net salary per hour for executive
SNHMP14 : mean net salary per hour for middle manager
SNHME14 : mean net salary per hour for employee
SNHMO14 : mean net salary per hour for worker
SNHMF14 : mean net salary for women
SNHMFC14 : mean net salary per hour for feminin executive
SNHMFP14 : mean net salary per hour for feminin middle manager
SNHMFE14 : mean net salary per hour for feminin employee
SNHMFO14 : mean net salary per hour for feminin worker
SNHMH14 : mean net salary for man
SNHMHC14 : mean net salary per hour for masculin executive
SNHMHP14 : mean net salary per hour for masculin middle manager
SNHMHE14 : mean net salary per hour for masculin employee
SNHMHO14 : mean net salary per hour for masculin worker
SNHM1814 : mean net salary per hour for 18-25 years old
SNHM2614 : mean net salary per hour for 26-50 years old
SNHM5014 : mean net salary per hour for >50 years old
SNHMF1814 : mean net salary per hour for women between 18-25 years old
SNHMF2614 : mean net salary per hour for women between 26-50 years old
SNHMF5014 : mean net salary per hour for women >50 years old
SNHMH1814 : mean net salary per hour for men between 18-25 years old
SNHMH2614 : mean net salary per hour for men between 26-50 years old
SNHMH5014 : mean net salary per hour for men >50 years old

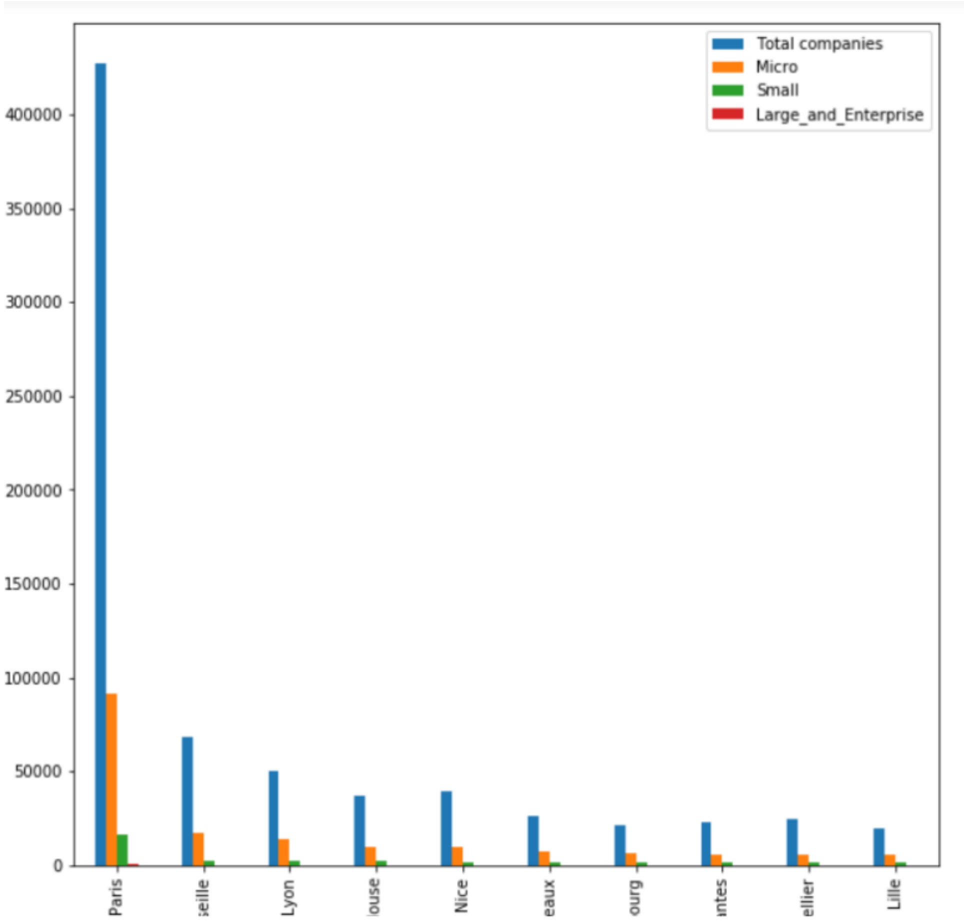
population :
demographic information in France per town, age, sex and living mode

NIVGEO : geographic level (arrondissement, communes...)
CODGEO : unique code for the town
LIBGEO : name of the town (might contain some utf-8 errors, this information has better quality name_geographic_information)
MOCO : cohabitation mode : [list and meaning available in Data description]
AGE80_17 : age category (slice of 5 years) | ex : 0 -> people between 0 and 4 years old
SEXE : sex, 1 for men | 2 for women
NB : Number of people in the category

These datasets can be merged by : CODGEO = code_insee
For the sake of simplicity we are using limited number of features from these datasets

Data analysis

Hypothesis 1 : Cities have more industry/Job concentration leading to urbanization
Finding 10 major economic cities in France

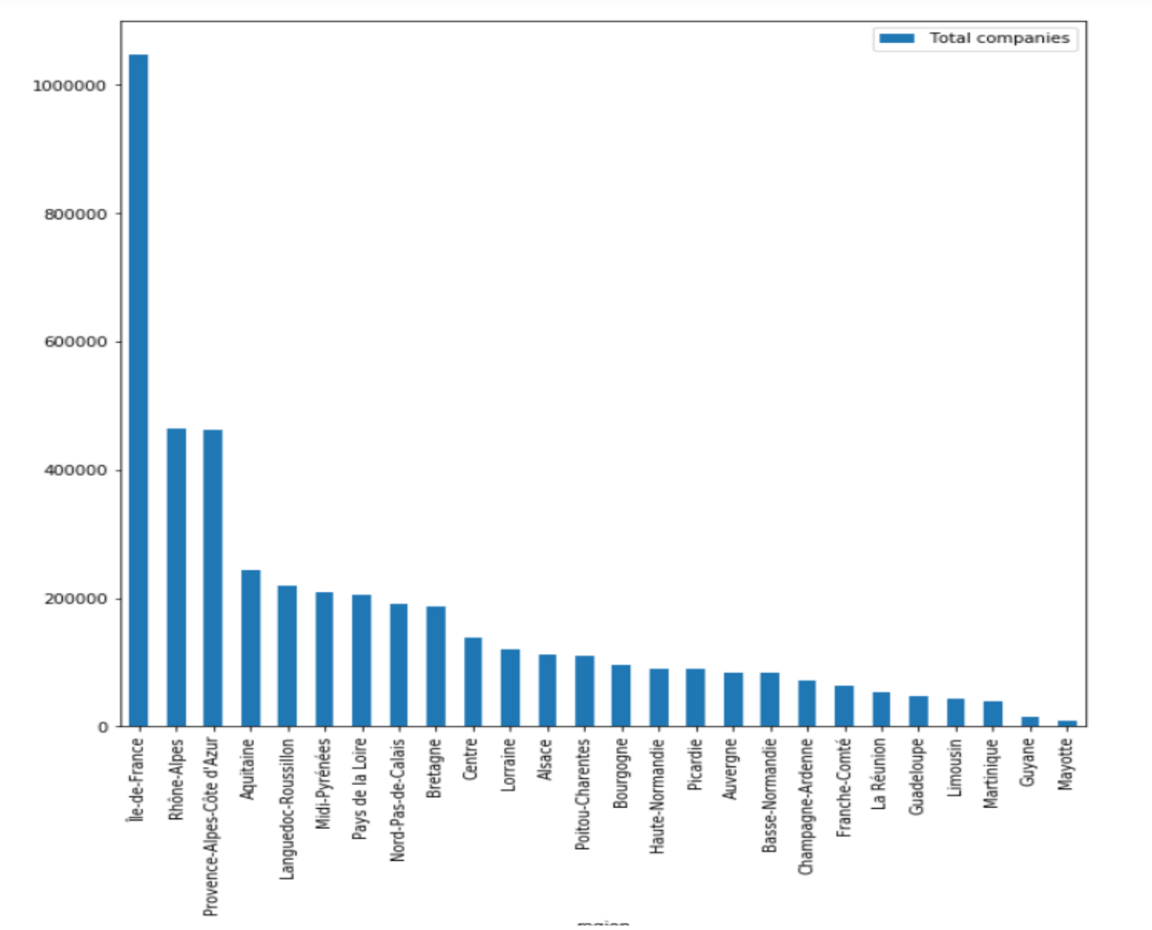


This plot shows the huge gap in concentration of industries in France. Paris has employment concentration out of proportions and also we can see that the top 10 places are all major cities in France, which is expected.

This shows a trend towards urbanization.

Hypothesis 2: industries/jobs are also concentrated in suburbs for major cities thus influencing region wise concentration

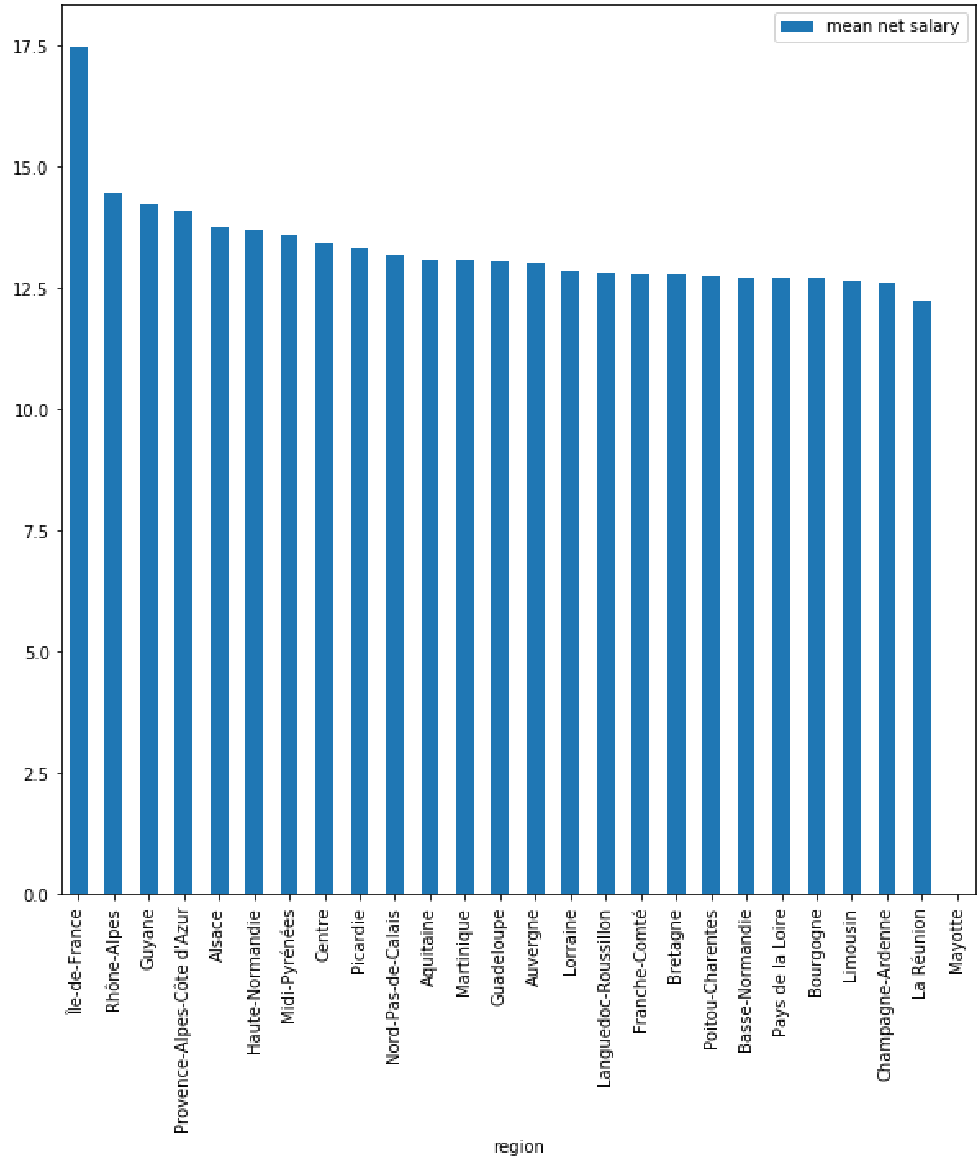
Plotting regions of France sorted based on job concentration



This plot confirms our hypothesis. Ile-de-France tops the list.

Hypothesis 3: More jobs = More salary

Showing PayScale variations across regions in France

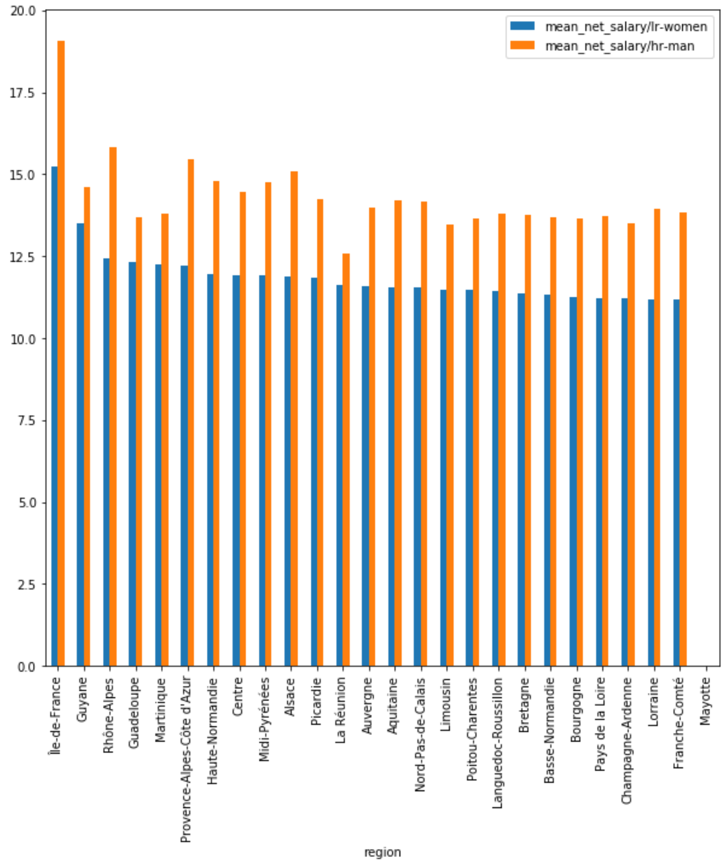


Comparing this with our graph for regional concentration of jobs, we can see that our assumption is not true.
Although for regions Ile-de-France, Rhone-alpes, and provence-alpes-cote d'azur our hypothesis is valid, it is false for regions like Aquitane,Languedoc Rousillon etc.

This graph can also be used as an indicator for highest per capita income based on regions

Hypothesis 4: Men earn more than women

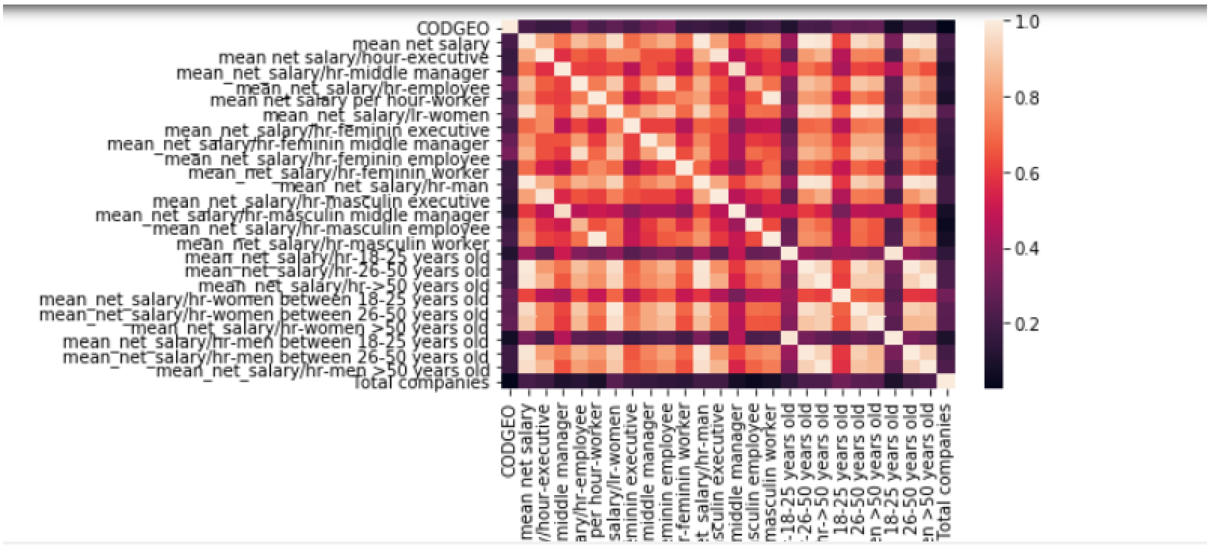
Plotting average salary for men and average salary for women



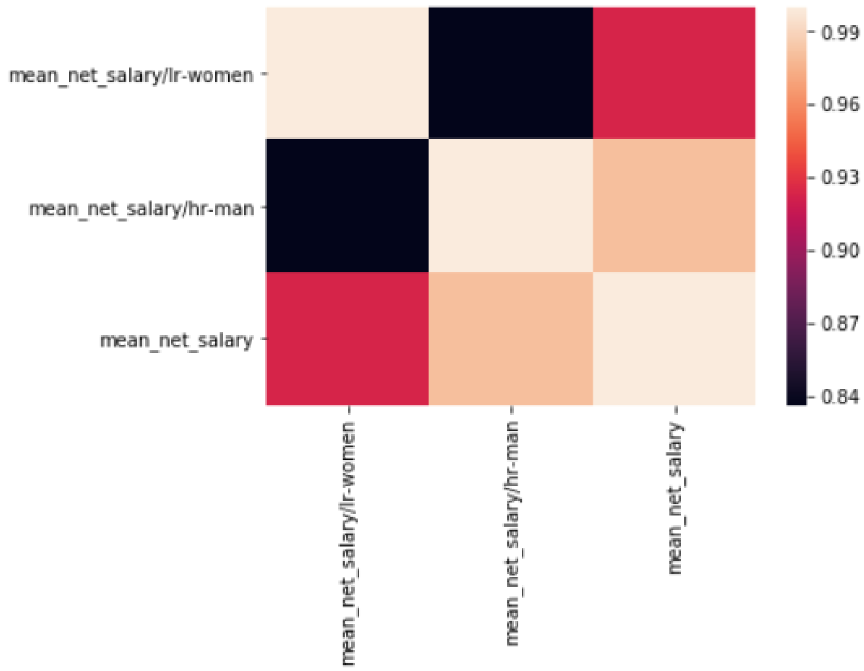
This plot proves our assumption. every region in France has lesser average salary for women.

Prediction

The goal here is to predict average salary of a region based on average salary of women and average salary of men. although this seems intuitive, we would like to check if this intuitive hypothesis “total of men and women salary/ total no. of men and women = mean net salary” is true. These features are also chosen as they have a good correlation with our required variable “mean net salary” and are most common values in general terms

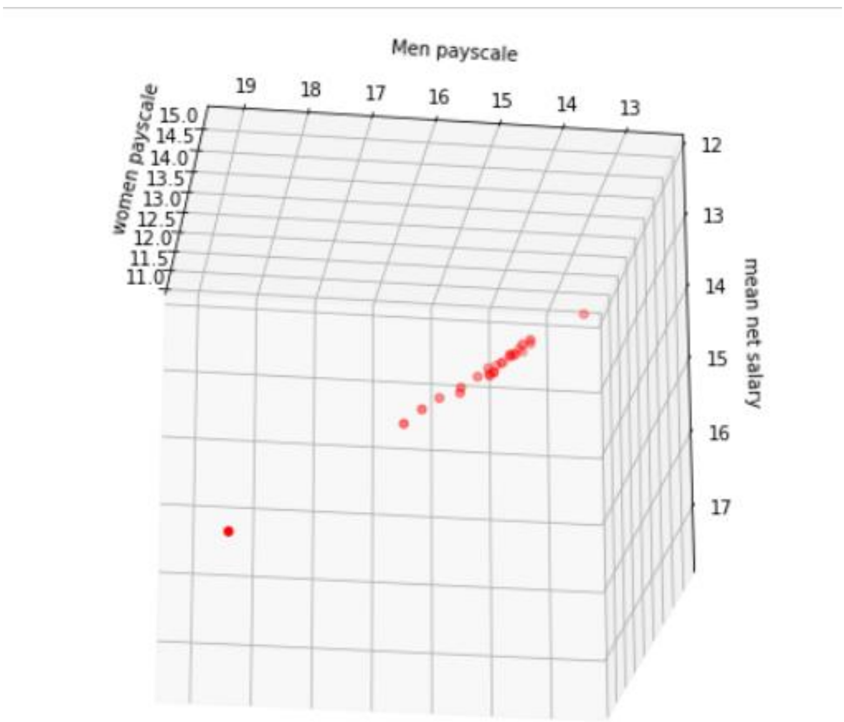


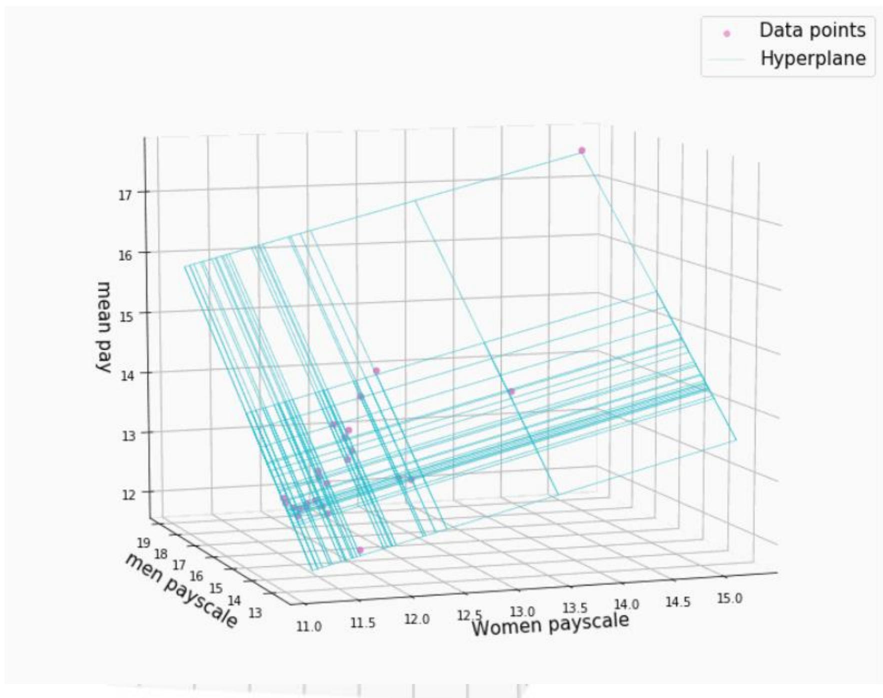
	mean_net_salary/lr-women	mean_net_salary/hr-man	mean_net_salary
mean_net_salary/lr-women	1.000000	0.836610	0.923672
mean_net_salary/hr-man	0.836610	1.000000	0.981414
mean_net_salary	0.923672	0.981414	1.000000



Choice of Algorithm

We scatter plot our data points to see





There seems to be a linear pattern, so we will use Linear regression to make our predictions

Choice of ratio for training and test data with results

Initial data split (training/test)	Prediction (training data, test data)
60/40	99.81, 80.5
70/30	99.81, 89.95
80/20	99.82, 87.82
85/15	99.82, 90.64
90/10	99.83, 79.29

We can see the ratio 90/10 overfits as the test data accuracy decreases drastically, the best ratio for our data is 85/15.

Is it overfitting?

The mean square error is 0 on both our training and test data. Although this is ideal , raises question of overfitting. But at the ratio of 85/15 , the model performs well even on test data, so this might be fine as our features well correlated and we are also dealing with non-noisy data

Conclusion

We get the below results

intercept : 0.048850092627617414

Coefficients:

[0.40621753 0.59019565]

Mean squared error for training data: 0.00

Mean squared error for test data: 0.00

Variance score: 0.91

In order to further tune the model,

Data for test and train can be randomly divided, instead of head and tail

Use of different machine learning algorithm like logistic regression.

Draw more interesting features like age classification