

Understanding Socio-Economic Dynamics in France

Insights into population, establishments,
salaries & gender inequalities



TABLE OF CONTENTS

- 01 INTRODUCTION**
Goals & Project planning
- 02 DATA GATHERING**
Flat files | API | Web Scrap | BigQuery
- 03 DATA PREPARATION**
Data Cleaning & Wrangling | SQL
- 04 DATA EXPLORING**
EDA & Visualization
- 05 DATA DISPLAYING**
French Population, Establishments, and Salaries Insights API
- 06 MACHINE LEARNING**
Gender pay gap prediction
- 07 CHALLENGES | NEXT STEPS**

01

INTRODUCTION

On November 6, 2023, at 11:25 a.m., women in France began working for free compared to men, highlighting the ongoing gender pay gap issue

In 2020, this disparity was evident on November 4, at 04:16 p.m.

INTRODUCTION

- Analyzing socio-economic factors in France, including population demographics, establishment distributions, and job categories
- Leveraging datasets from the French National Institute of Statistics and Economic Studies (INSEE) from 2020
- Focusing on gender inequalities in the labor market, particularly in salary structures and representation in leadership roles
- Aim to uncover patterns and trends through advanced data analysis to inform policy-making and promote socio-economic progress



Goals & Project Planning



To illuminate socio-economic dynamics in France and advocate for a more inclusive future...

Explore demographic trends and disparities in France

Investigate the representation of women in managerial positions

Analyze salary distributions across different demographics and job categories

Identify correlations between population demographics and establishment data

GOALS



TRELLO FOR PROJECT PLANNING

The screenshot shows a Trello board titled "Ironhack final project". It has three main columns: "À faire", "En cours", and "Terminé".

- À faire:** Contains a pinned card for "Rédiger rapport RNCP" due on "23 avr." and a card for "1.3. BigQuery".
- En cours:** Contains cards for "2.1 Clean data", "2.2. Execute exploratory data analysis (EDA)", "2.3. Visualisation", "3. Choose the database type (compare several types and explain why)", "4. SQL : Create a database (database, tables)", and "6. Create an entity-relationship diagram (at least 4 entities)".
- Terminé:** Contains cards for "1. Data collection" (status 8/8), "1.1. Data from API" (status 2/2), "1.2. Data from webscraping" (status 5/5), "5. SQL : Add data to the database", and "7. SQL : Create 5 scripts showing the insights" (status 5/5).

The screenshot shows a detailed view of a Trello card for "1. Data collection".

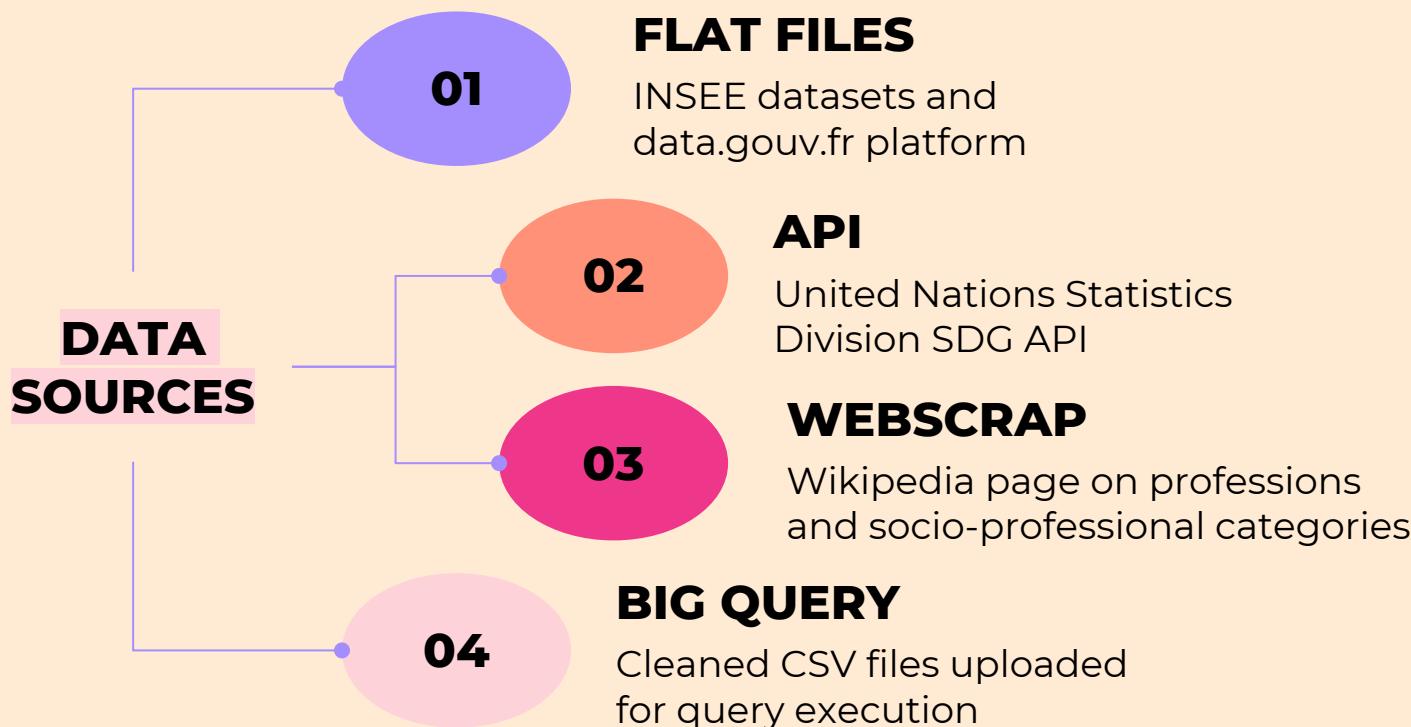
- Description:** Includes a pinned note about employment, salaries, population per town.
- Pièces jointes:** A file named "French employment, salaries, population per town" is attached.
- Checklist:** A progress bar shows 100% completion for tasks: "télécharger-les-csv", "les mettre-sur-un-notebook", and "Webscraping".
- Sidebar:** Shows suggestions like "Rejoindre", "Ajouter à la carte", and a "Power-Ups" section with options like "Ajouter des Pow..." and "Automatisation".

02

DATA GATHERING

- Flat files | API | Web Scrap | BigQuery

DATA GATHERING



DATA GATHERING

SOURCES

Reliable sources: National Institute of Statistics and Economic Studies (INSEE) & data.gouv.fr platform

FLAT FILES

DATA YEAR

2020

RATIONALE

Ensures consistency, reliability, and completeness in analysis, facilitating meaningful comparisons across different variables and indicators.

DATA GATHERING : FLAT FILES



Number of Active Establishments

- Description: Provides insights into the distribution of establishments across five major sectors.
- Source: INSEE
- File: etablissement.csv | Rows: 34,980 | Columns: 61



Average Net Hourly Wage

- Description: Details average net hourly wage by socio-professional category, gender, and age.
- Source: INSEE
- File: salaire_par_commune.csv | Rows: 5,421 | Columns: 25

DATA GATHERING : FLAT FILES

Population Demographics



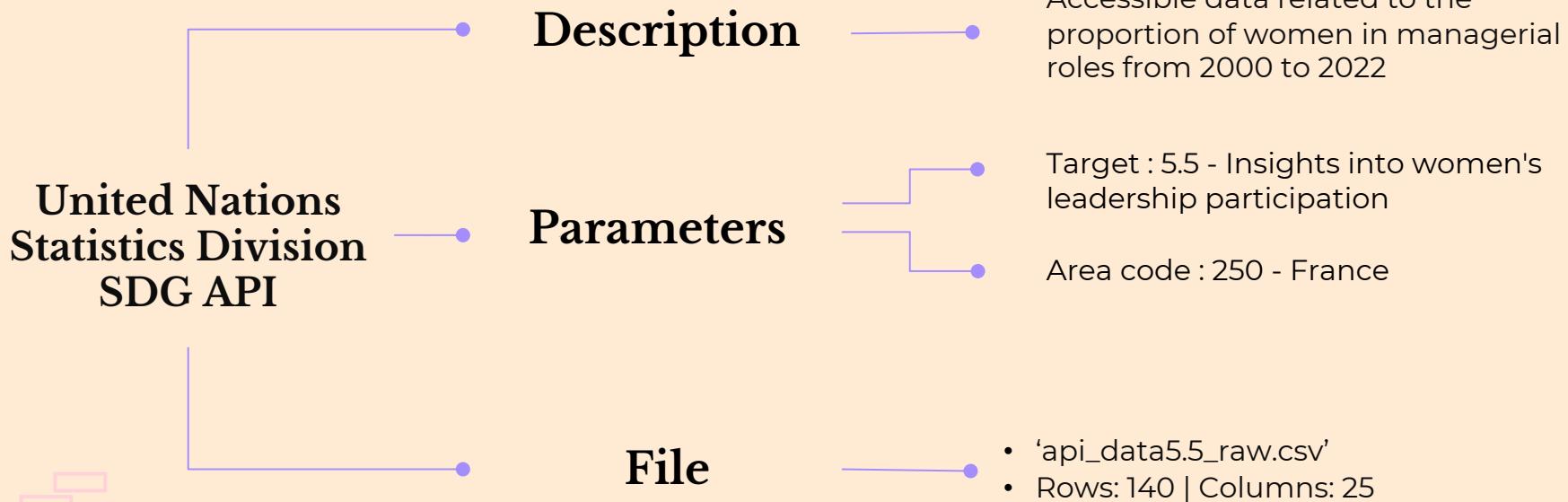
- Description: Presents demographic information on population over 15 years old by gender, age, and socio-professional category
- Source: INSEE
- File: population.csv | Rows: 1,738,965 | Columns: 7

French Municipalities - Base of Postal Codes

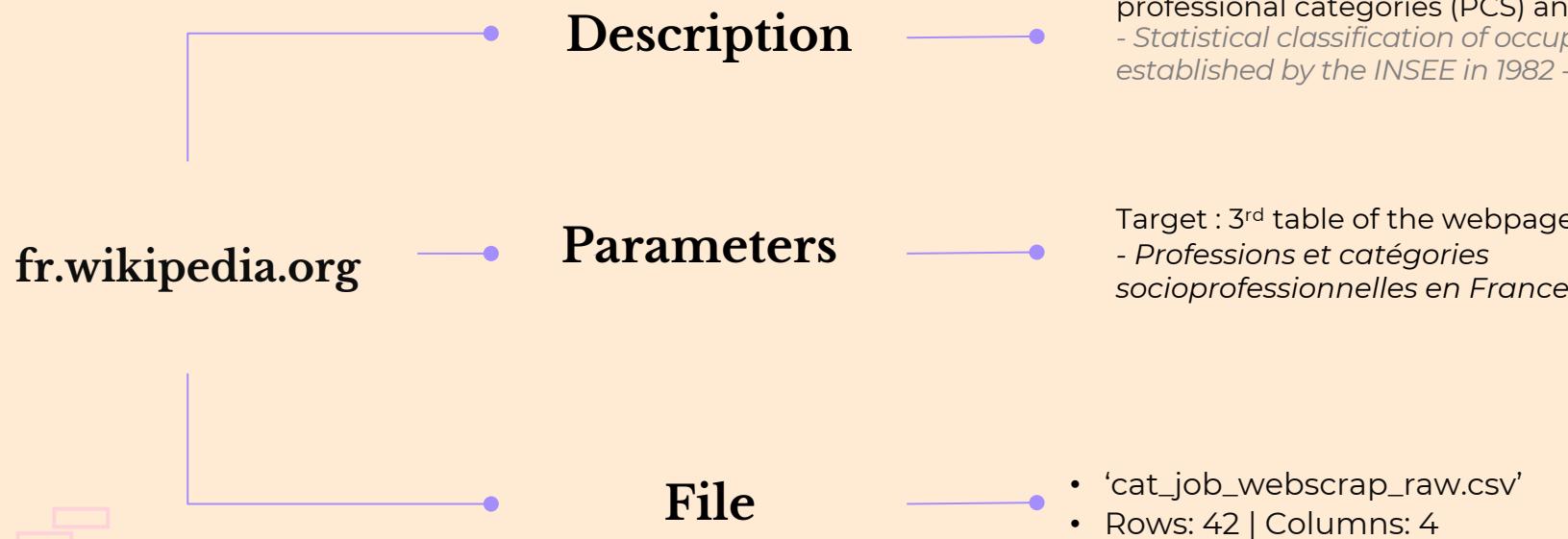


- Description: Includes names of regions, departments, and municipalities along with their respective INSEE codes
- Source: La Poste - data.gouv.fr
- File: communes-departement-region.csv | Rows: 5,421 | Columns: 15

DATA GATHERING : API



DATA GATHERING : WEBSRAPING



03

DATA PREPARATION

Data Cleaning & Wrangling |
SQL : Database creation | ERD | Queries |

DATA CLEANING & WRANGLING



- Handling null values, duplicates
- Value formating
- Renaming, dropping & reorganizing columns
- Reduce the number of categories (age groups & firm sizes)
- Data homogenization & normalization

AFTER CLEANING

population

- Shape : 826,266 x 6
- Before : 1,738,965 x 7



establishment

- Shape : 34,980 x 11
- Before : 34,980 x 61



geography

- Shape : 36,370 x 10
- Before : 5,421 x 15



Cleaned datasets

salary

- Shape : 10,842 x 10
- Before : 5,421 x 25



woman_managerial_position

- Shape : 19 x 4
- Before : 140 x 25



french_cat_job

Shape : 42 x 4



AFTER CLEANING

population*

- Shape : 826,266 x 6
- Before : 1,738,965 x 7



establishment

- Shape : 34,980 x 11
- Before : 34,980 x 61



geography

- Shape : 36,370 x 10
- Before : 5,421 x 15



Cleaned datasets

salary**

- Shape : 10,842 x 10
- Before : 5,421 x 25



woman_managerial_position

- Shape : 19 x 4
- Before : 140 x 25



french_cat_job

Shape : 42 x 4



*population_all = population before age grouping

**salary_all = salary before normalization (male/female)



ERD

SQL DATABASE CREATION : 'final_project'

SQL QUERIES*

used for ...



Pearson Corr. for EDA

between total population
& establishments



Views for API endpoints

with joint tables & new
aggregated columns



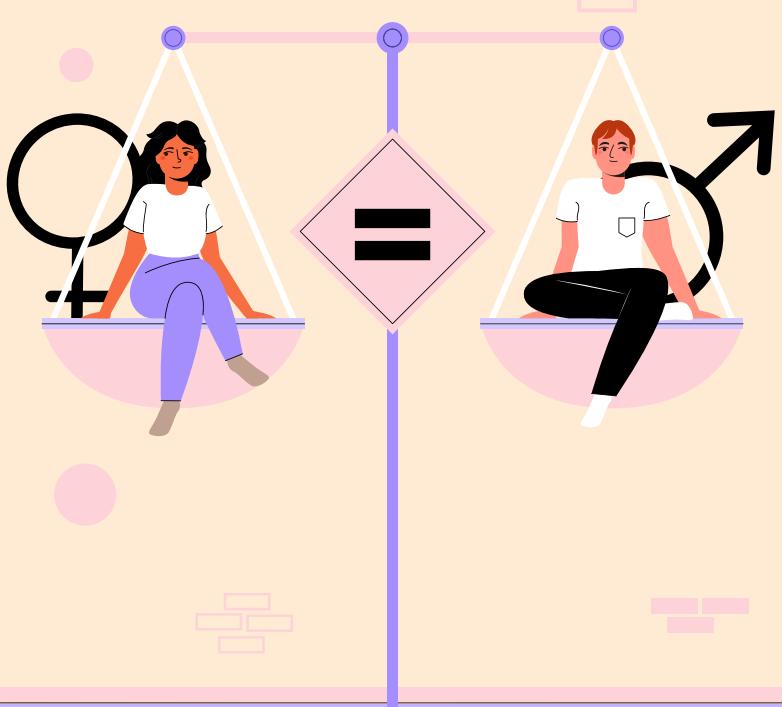
Paygap for Tableau

exported as .csv from MySQL
& uploaded for visualization



TOP 10 for EDA

- highest & lowest gender pay gap by department

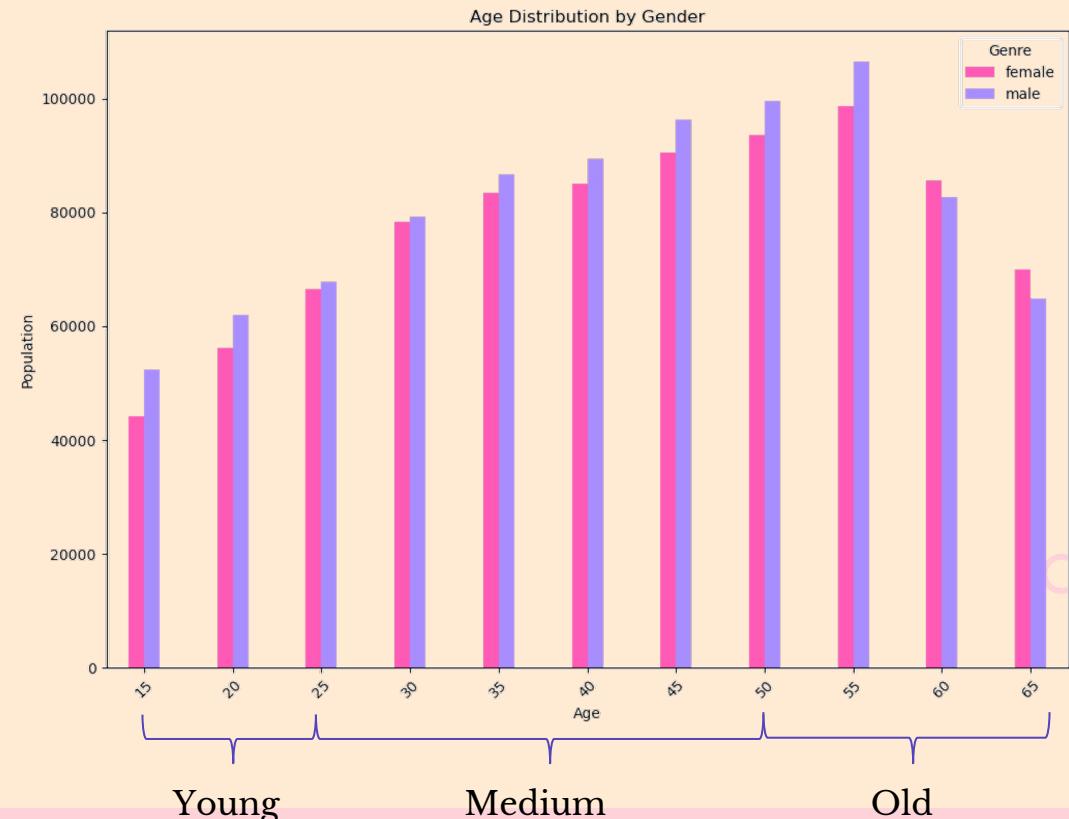
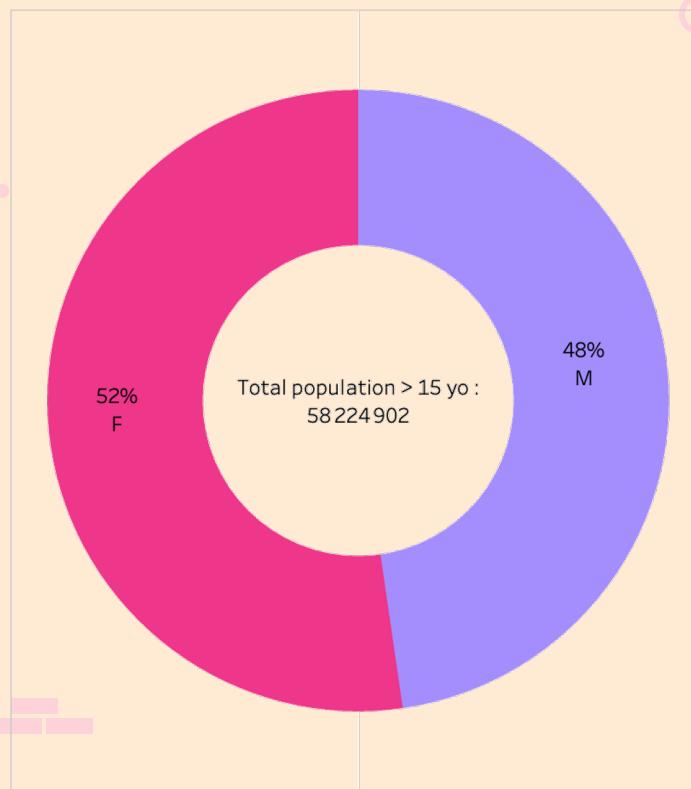


04

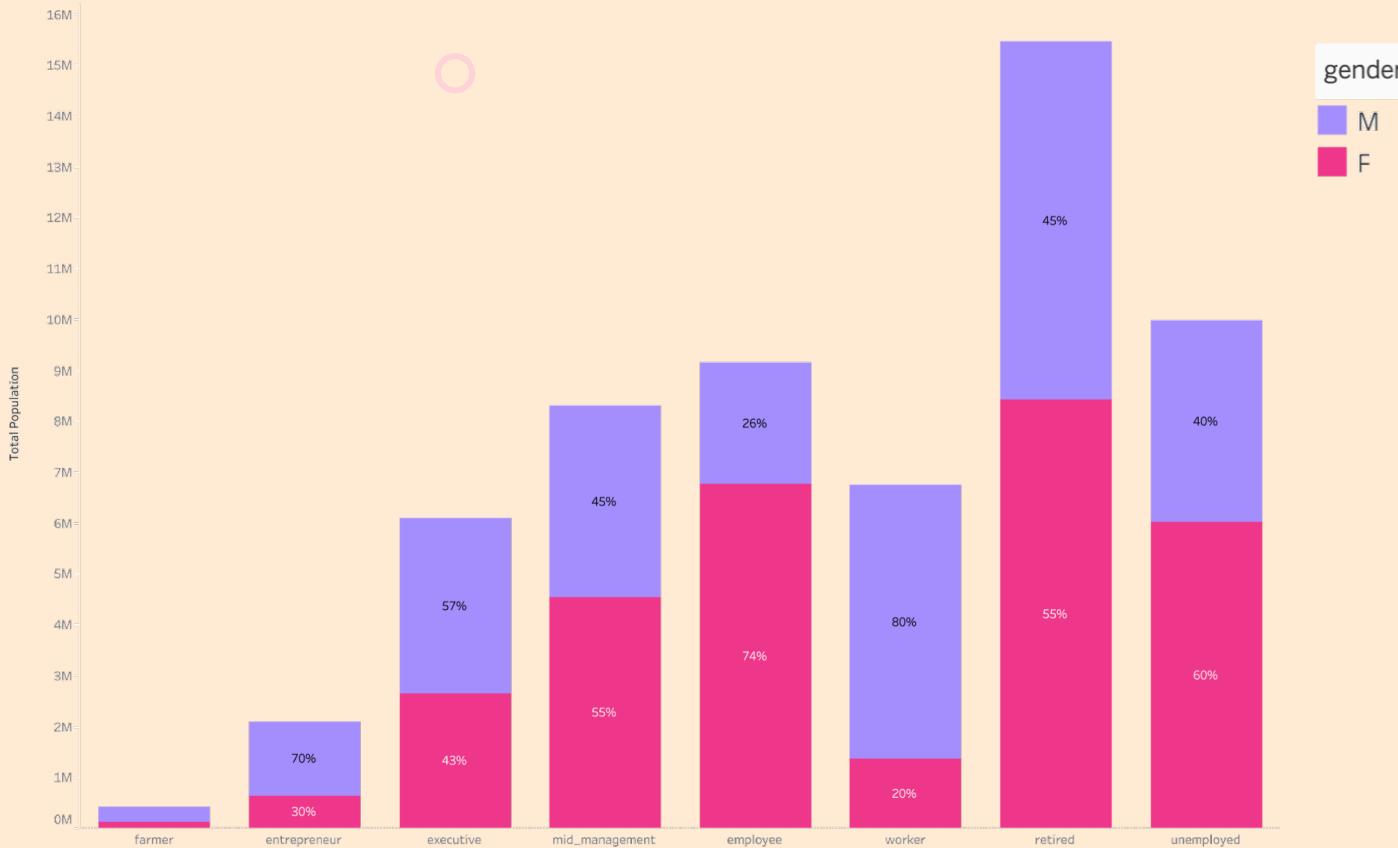
DATA EXPLORING

Exploratory Data Analysis
& Visualization

POPULATION INFOGRAPHICS BY GENDER



GENDER REPRESENTATION BY JOB CATEGORY



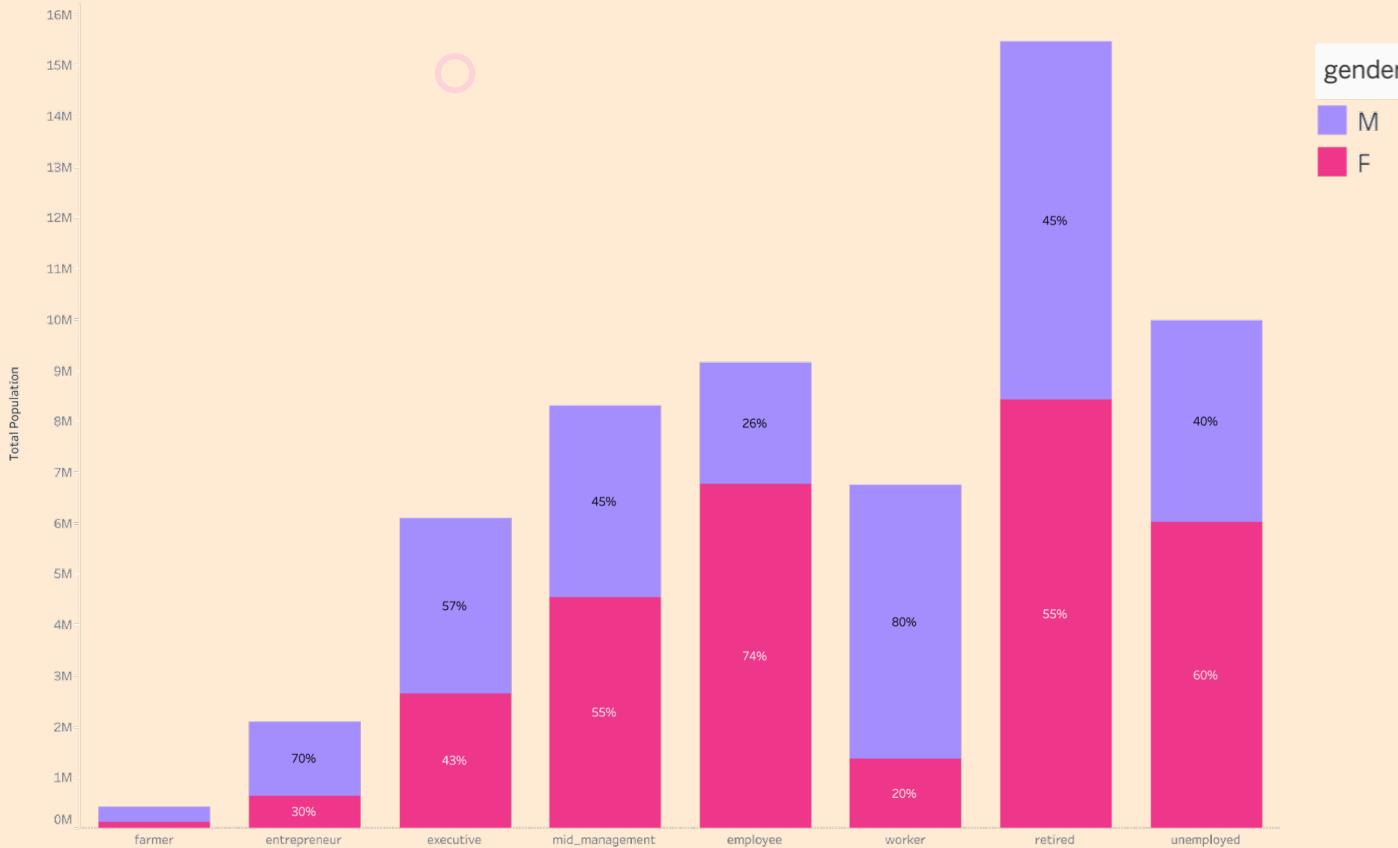
GENDER REPRESENTATION BY JOB CATEGORY

Proportion of woman in a managerial position

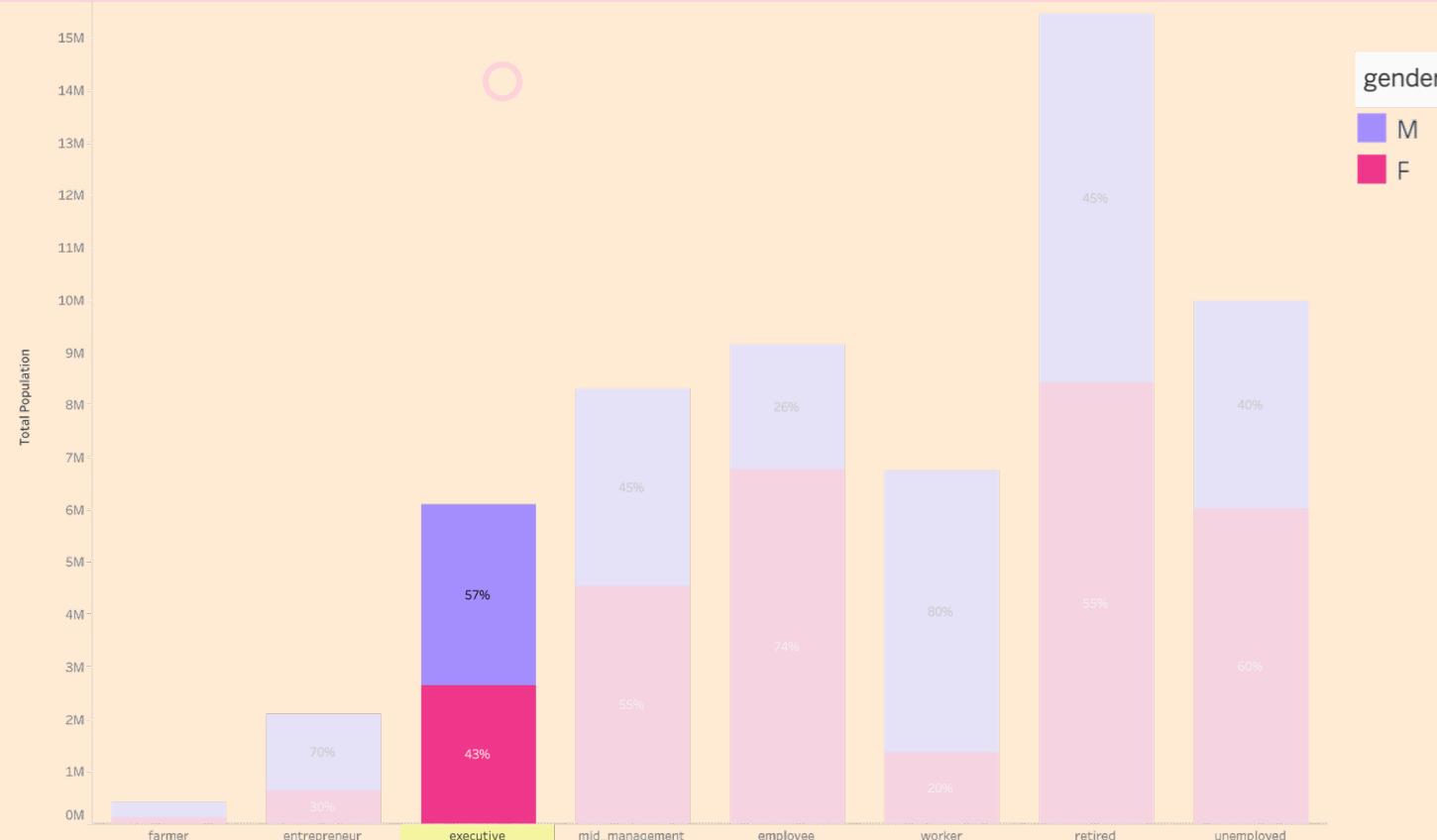
| Year of Time Period | % Women In Managerial Positions | % Women In Senior Middle Management |
|---------------------|---------------------------------|-------------------------------------|
| 2004 | 35,89 | 33,46 |
| 2005 | 37,60 | 36,37 |
| 2006 | 37,88 | 37,01 |
| 2007 | 37,83 | 36,46 |
| 2008 | 38,50 | 36,79 |
| 2009 | 38,09 | 37,37 |
| 2010 | 38,54 | 38,59 |
| 2011 | 39,34 | 39,13 |
| 2012 | 39,32 | 39,20 |
| 2013 | 36,00 | 35,58 |
| 2014 | 32,74 | 32,03 |
| 2015 | 31,66 | 30,93 |
| 2016 | 32,91 | 31,07 |
| 2017 | 33,44 | 32,64 |
| 2018 | 34,44 | 34,38 |
| 2019 | 34,67 | 34,30 |
| 2020 | 35,53 | 34,95 |
| 2021 | 37,79 | 36,77 |
| 2022 | 39,89 | 39,39 |

The joint calculation of these two measures provides information on whether **women are more represented in junior management** than in senior and middle management, thus pointing to an eventual **ceiling for women** to access higher-level management positions

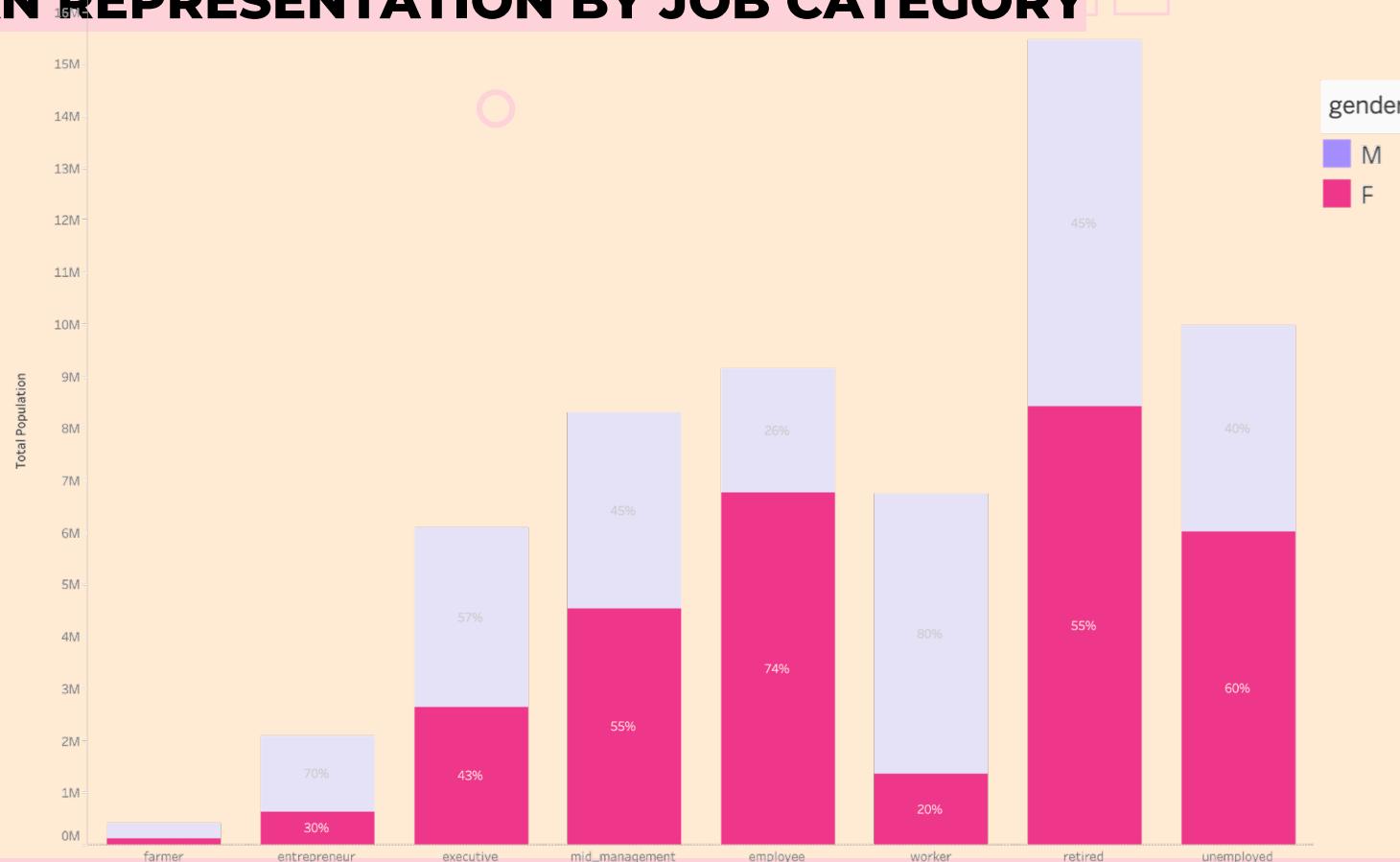
GENDER REPRESENTATION BY JOB CATEGORY



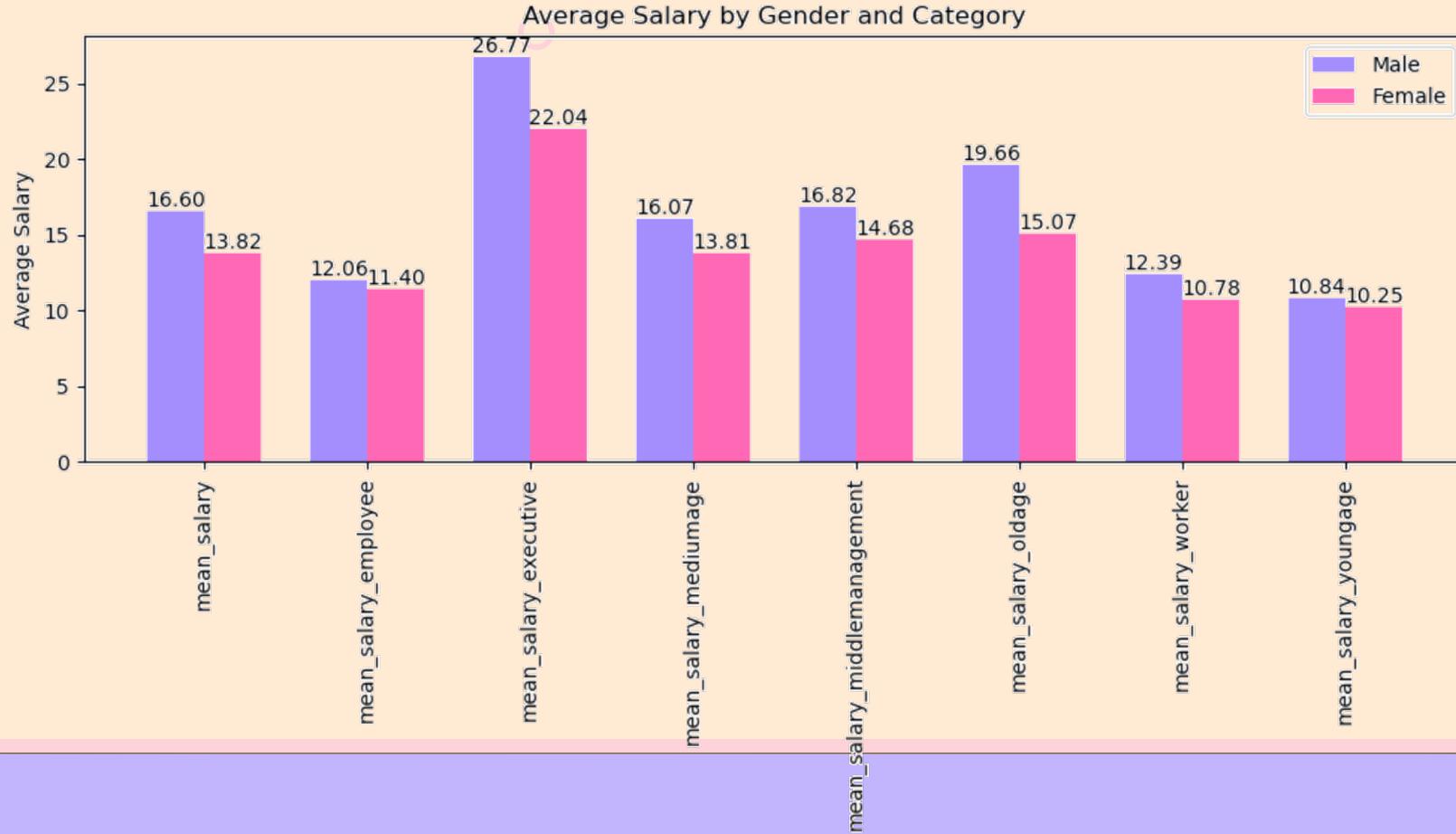
GENDER REPRESENTATION IN SENIOR MANAGERIAL POSITION



WOMAN REPRESENTATION BY JOB CATEGORY



GENDER PAY GAP BY JOB CATEGORY



Population & establishment correlation

Pearson correlation =

population_establishment_correlation

0.59

Population / Region

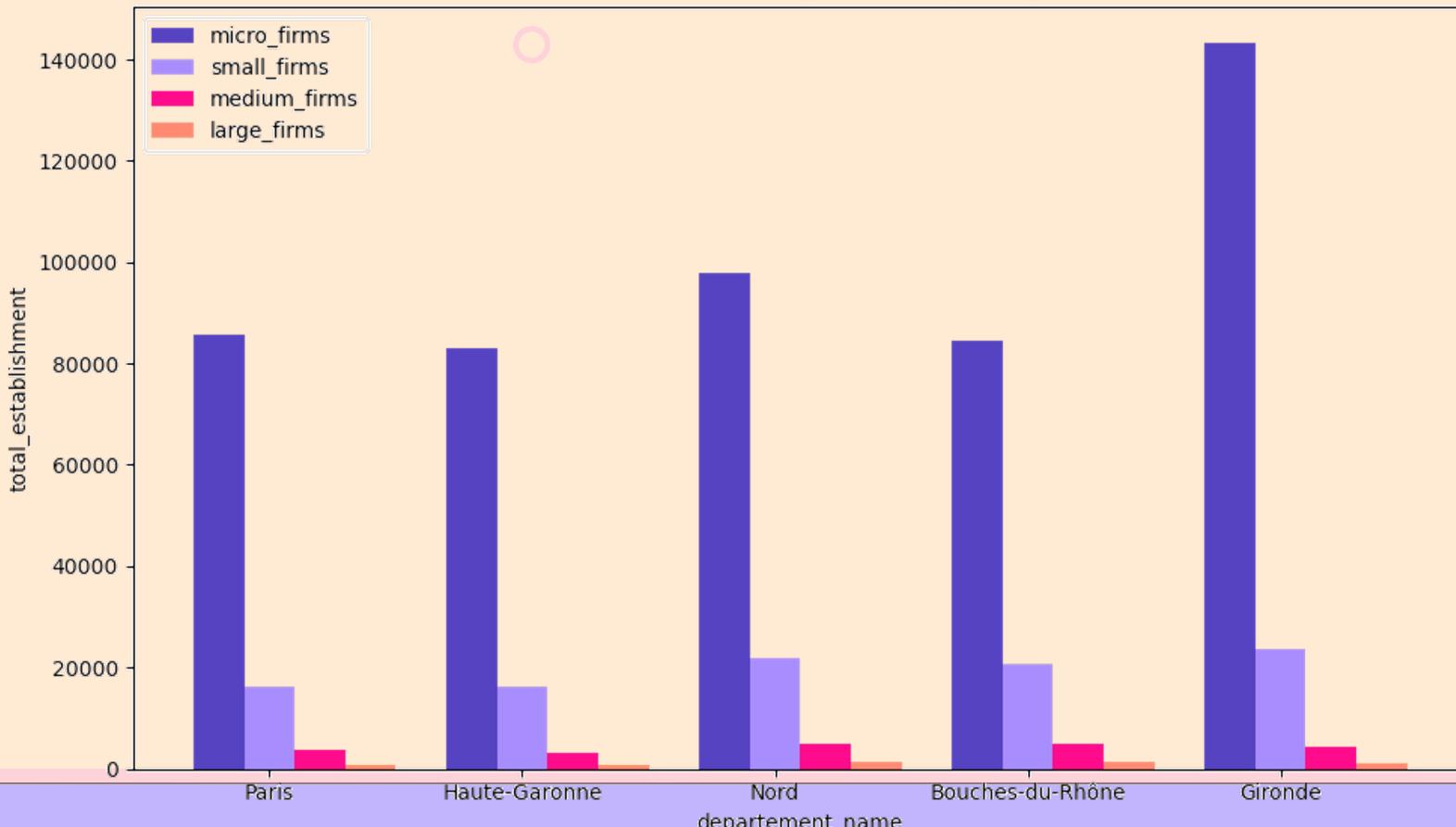


Establishment / Region



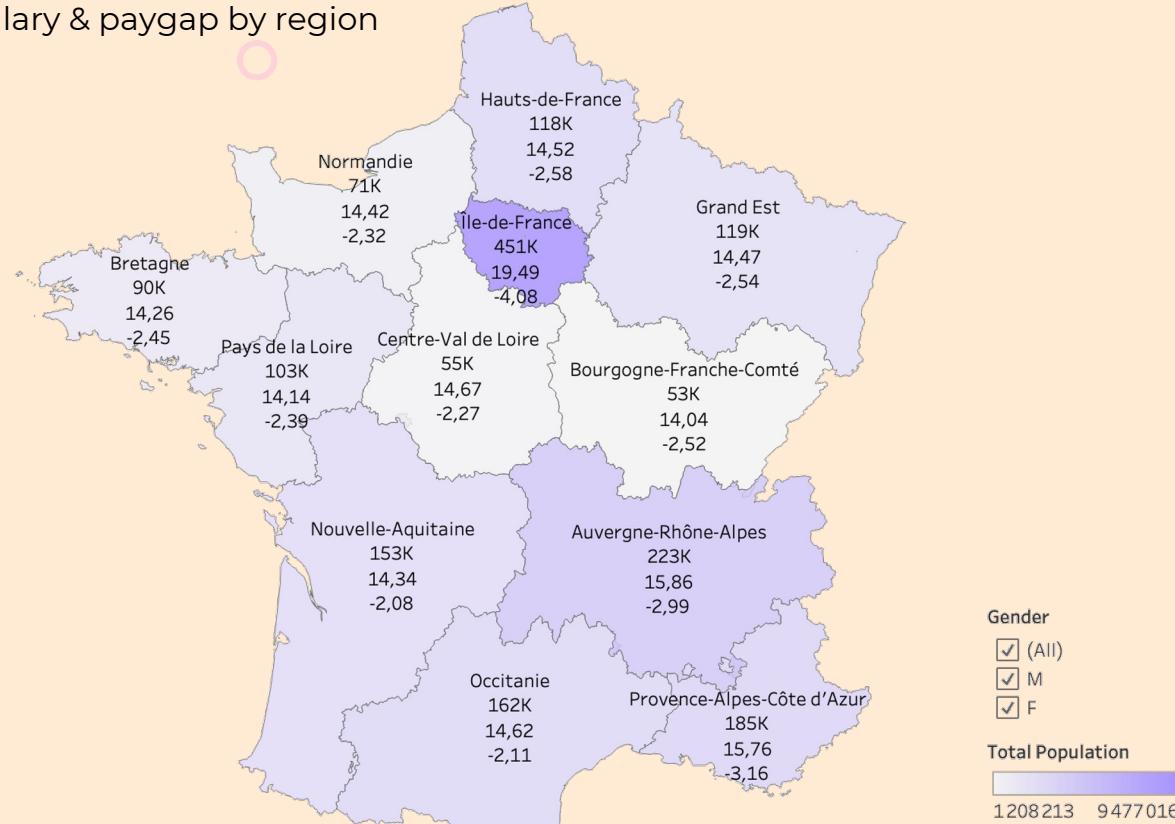
Population & establishment correlation

establishment distribution by size for top 5 department



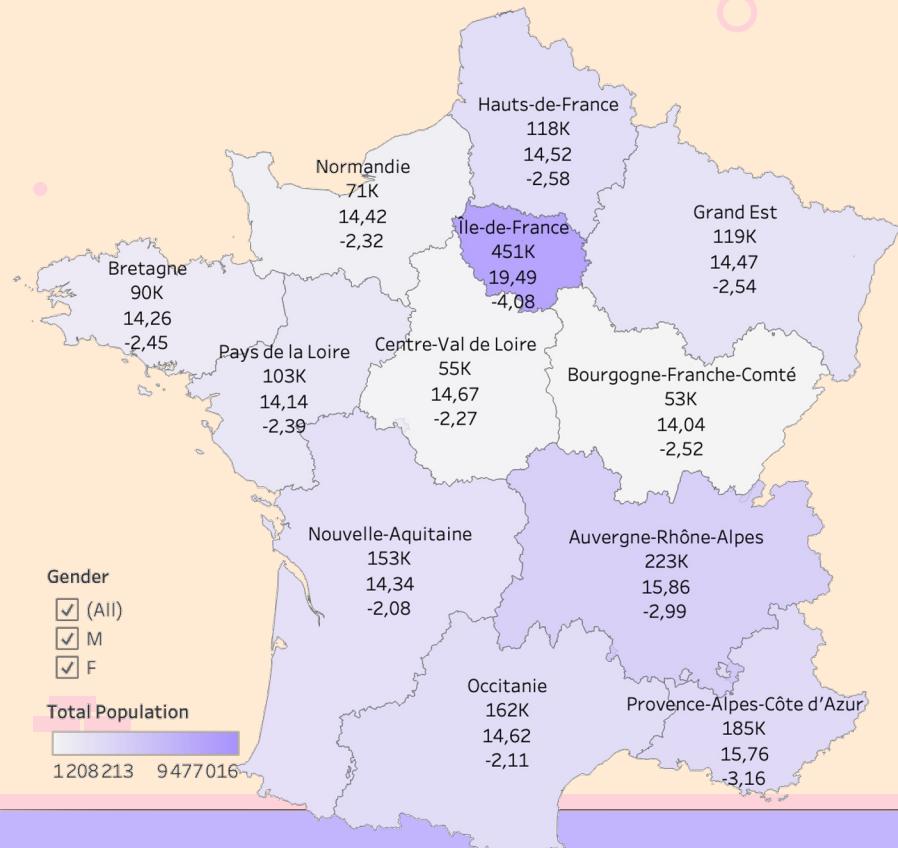
Regions most populated have a higher salary, but also a higher gender pay gap

Establishments, mean salary & paygap by region

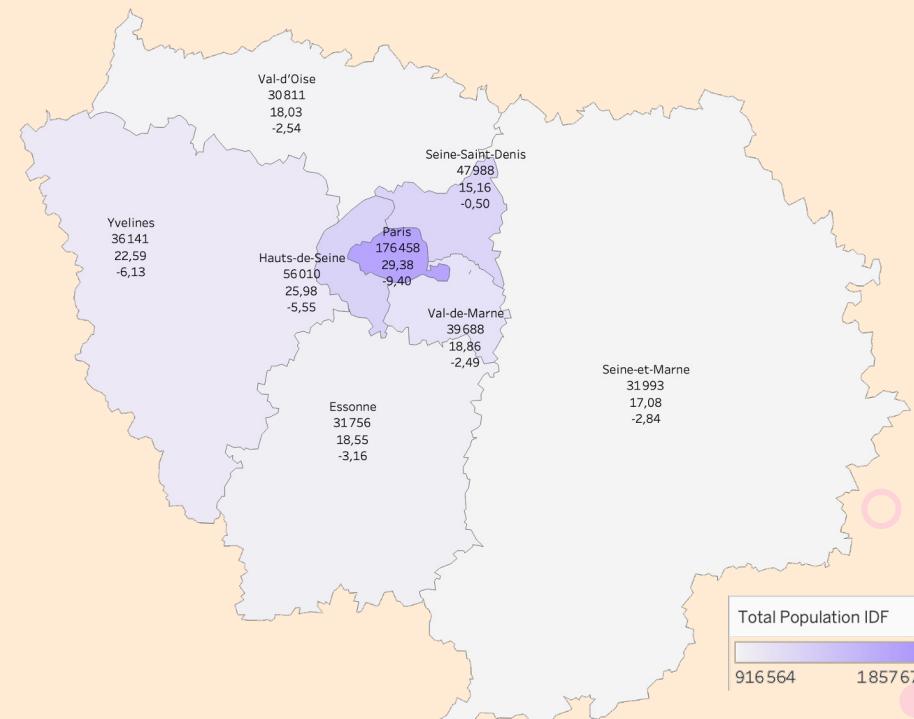


Regions most populated have a higher salary, but also a higher gender pay gap

Establishments, mean salary & paygap by region



Establishments, mean salary & paygap in IDF



The pay gap is negative all over France



Top 10 – highest
pay gap by
department

| code_departement | departement_name | avg_salary_male | avg_salary_female | avg_salary_all | gender_pay_gap |
|------------------|-----------------------|-----------------|-------------------|----------------|----------------|
| 75 | Paris | 34.67 | 25.27 | 29.97 | -9.4 |
| 78 | Yvelines | 25.65 | 19.52 | 22.59 | -6.13 |
| 92 | Hauts-de-Seine | 28.78 | 23.22 | 26 | -5.55 |
| 69 | Rhône | 20.43 | 16.16 | 18.3 | -4.27 |
| 90 | Territoire de Belfort | 17.83 | 13.73 | 15.78 | -4.11 |
| 31 | Haute-Garonne | 18.75 | 14.99 | 16.87 | -3.76 |
| 13 | Bouches-du-Rhône | 18.53 | 14.76 | 16.64 | -3.76 |
| 38 | Isère | 18.47 | 14.82 | 16.65 | -3.65 |
| 73 | Savoie | 17.26 | 13.96 | 15.61 | -3.3 |
| 68 | Haut-Rhin | 16.5 | 13.28 | 14.89 | -3.22 |

Top 10 – lowest
pay gap by
department

| code_departement | departement_name | avg_salary_male | avg_salary_female | avg_salary_all | gender_pay_gap |
|------------------|-------------------|-----------------|-------------------|----------------|----------------|
| 93 | Seine-Saint-Denis | 15.38 | 14.88 | 15.13 | -0.5 |
| 23 | Creuse | 12.82 | 12.14 | 12.48 | -0.68 |
| 48 | Lozère | 13.15 | 12.16 | 12.65 | -0.99 |
| 974 | La Réunion | 13.89 | 12.77 | 13.33 | -1.12 |
| 971 | Guadeloupe | 14.69 | 13.56 | 14.12 | -1.12 |
| 973 | Guyane | 15.09 | 13.96 | 14.52 | -1.13 |
| 79 | Deux-Sèvres | 14.64 | 13.19 | 13.91 | -1.45 |
| 972 | Martinique | 15.09 | 13.64 | 14.36 | -1.45 |
| 36 | Indre | 13.82 | 12.31 | 13.06 | -1.51 |
| 24 | Dordogne | 13.98 | 12.4 | 13.19 | -1.58 |

05

DATA DISPLAYING

French Population, Establishments,
and Salaries Insights API

SWAGGER

This Swagger documentation outlines the structure of the French Population, Establishments, and Salaries API

The screenshot shows the Swagger UI interface for the "French Population, Establishment and Salary API". The title bar includes the Swagger logo, the URL "/static/tp_myAPI_swagger.yml", and a green "Explore" button. Below the title, the API name "French Population, Establishment and Salary API" is displayed with a version of "1.0.0" and an "OAS3" badge. A brief description states: "An API to retrieve population, establishment and salary data by department".
The "Servers" section shows a dropdown menu set to "http://localhost:5000".
The "default" section contains two GET requests:

- GET /population_salary/{code_departement} Retrieve population and salary data by department code
- GET /population_salary Retrieve population, establishment and salary data for all departments

The "Schemas" section lists three schema definitions:

- PopulationSalary >
- PopulationSalaryResponse >
- Establishment >

/population_salary

```
Non sécurisé 192.168.1.18:8080/population_salary
Terminer la mise à jour

Raw Parsed

[{"id": 1, "average_salary_all": 13.53, "average_salary_female": 12.45, "average_salary_male": 14.61, "code_departement": "10", "departement_name": "Aube", "gender_paygap": -2.16, "population_establishment_correlation": 0.58, "total_establishments": "687238", "total_population": "515469"}, {"id": 2, "average_salary_all": 13.43, "average_salary_female": 12.25, "average_salary_male": 14.01, "code_departement": "11", "departement_name": "Aude", "gender_paygap": -1.75, "population_establishment_correlation": 0.53, "total_establishments": "607828", "total_population": "630998"}, {"id": 3, "average_salary_all": 13.42, "average_salary_female": 12.42, "average_salary_male": 14.42, "code_departement": "12", "departement_name": "Aveyron", "gender_paygap": -1.89, "population_establishment_correlation": 0.49, "total_establishments": "656544", "total_population": "507778"}, {"id": 4, "average_salary_all": 16.44, "average_salary_female": 14.76, "average_salary_male": 18.53, "code_departement": "13", "departement_name": "Bouches-du-Rhône", "gender_paygap": -3.76, "population_establishment_correlation": 0.51, "total_establishments": "10108796", "total_population": "4983278"}, {"id": 5, "average_salary_all": 14.34, "average_salary_female": 13.15, "average_salary_male": 15.54, "code_departement": "14", "departement_name": "Calvados", "gender_paygap": -2.38, "population_establishment_correlation": 0.55, "total_establishments": "605228", "total_population": "391110"}]
```

/population_salary/<code_department>

```
← → ⌂ Non sécurisé 192.168.1.18:8080/population_salary/75
{
  "establishment": [
    {
      "code_department": "75",
      "department_name": "Paris",
      "population_establishment_correlation": 0.88,
      "total_agriculture_sector": "5841",
      "total_agriculture_over50yo": "7867273",
      "total_construction_sector": "347689",
      "total_establishments": "8186115",
      "total_industry_sector": "2077657",
      "total_micro_firms": "179133",
      "total_medium_firms": "179133",
      "total_micro_firms": "1913983",
      "total_population": "1997602",
      "total_public_sector": "5576805",
      "total_small_firms": "1004958"
    }
  ],
  "population_salary": [
    {
      "average_salary": 34.58,
      "average_salary_25to49yo": 31.4,
      "average_salary_50plus": 13.24,
      "average_salary_executive": 44.77,
      "average_salary_middlemanagement": 20.68,
      "average_salary_over50yo": 46.87,
      "average_salary_under25yo": 14.81,
      "average_salary_over50yo": 11.95,
      "code_department": "75",
      "department_name": "Paris",
      "geog": "Paris",
      "total_population": "823881",
      "total_population_25to49yo": "4246698",
      "total_population_employee": "823595",
      "total_population_executive": "311982",
      "total_population_over50yo": "2118875",
      "total_population_over50yo": "355987",
      "total_population_under25yo": "142682",
      "total_population_worker": "59674"
    },
    {
      "average_salary": 25.27,
      "average_salary_25to49yo": 25.48,
      "average_salary_50plus": 12.47,
      "average_salary_executive": 33.27,
      "average_salary_middlemanagement": 18.84,
      "average_salary_over50yo": 28.9,
      "average_salary_under25yo": 13.48,
      "average_salary_over50yo": 11.44,
      "code_department": "75",
      "department_name": "Paris",
      "geog": "Paris",
      "total_population": "1073300",
      "total_population": "1073300",
      "total_population_25to49yo": "4520813",
      "total_population_employee": "1507880",
      "total_population_executive": "207080",
      "total_population_over50yo": "2162726",
      "total_population_over50yo": "451885",
      "total_population_under25yo": "1507880"
    }
  ]
}
```

06

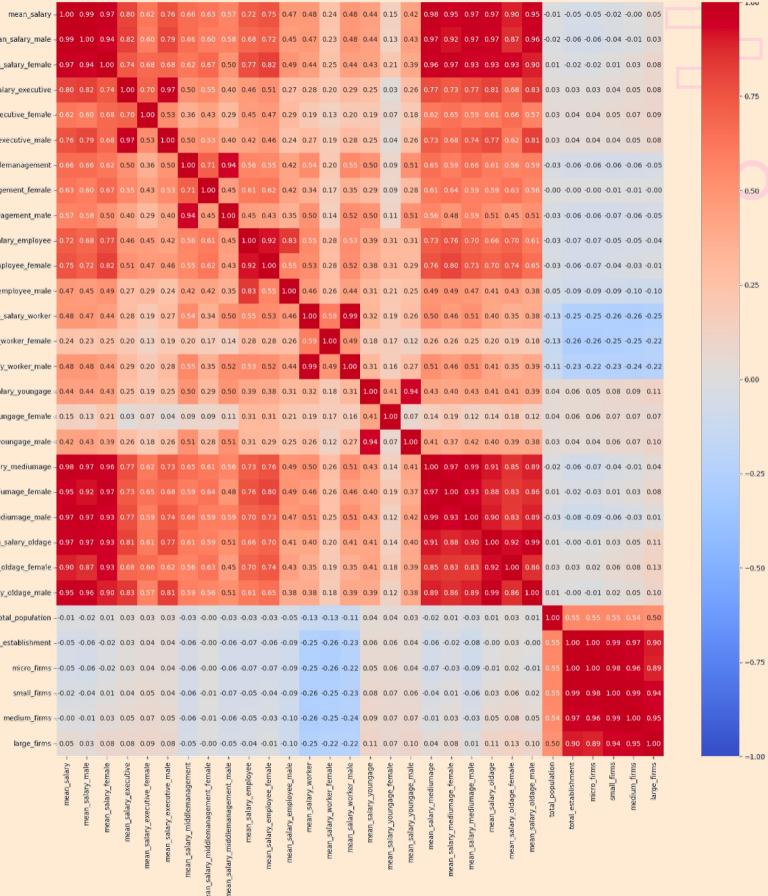
MACHINE LEARNING

Gender pay gap prediction within municipalities

Correlation Matrix

Heatmap visualization of feature correlations to note potential relationships between features

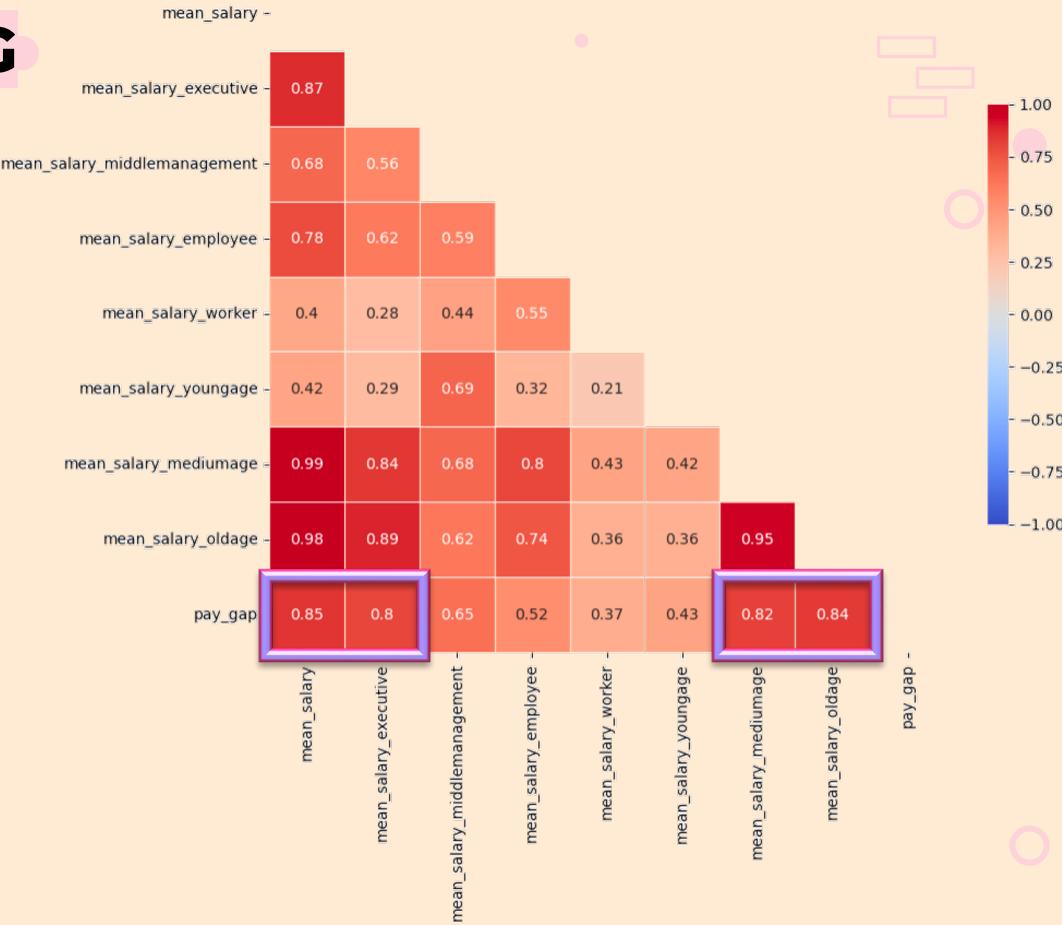
- Merged DataFrames **establishment**, **geography**, **salary_all** and **population** based on the CODGEO column
 - Selected columns for creating the heatmap, including various mean salary categories, total population, and establishment sizes
 - Calculated the correlation matrix between the selected variables



FEATURE ENGINEERING

Features selection

Pay gap has a high correlation with the mean salary, the executive salary, and the middle and old age salaries.



MACHINE LEARNING OVERVIEW

Objective :

- Predict the **gender pay gap** within various communes

Target Variable :

- **Pay gap** defined as the **numerical** difference in mean salaries between genders

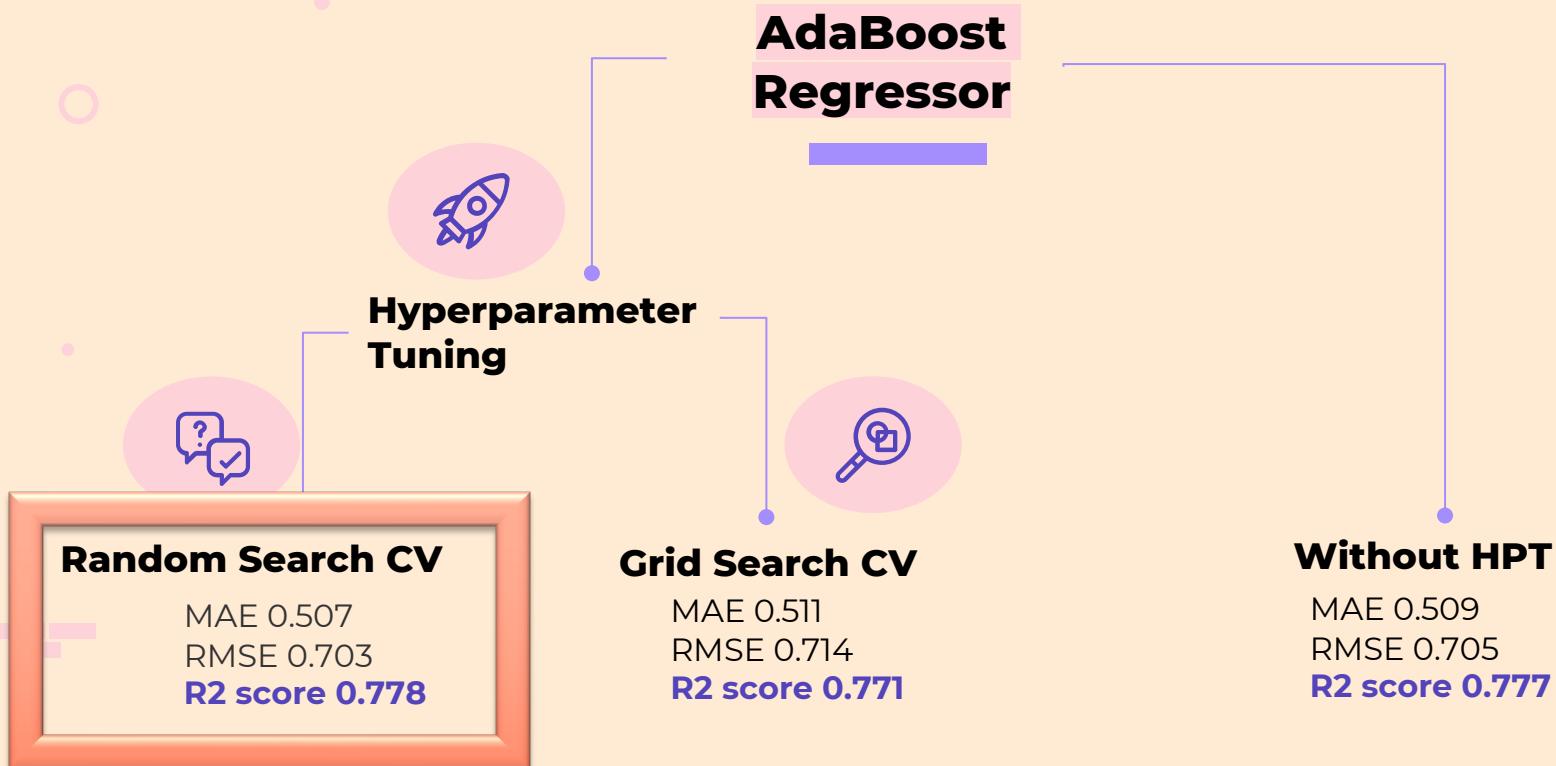
Regression Model Choice :

- Effective for predicting **numerical outcomes** with linear or combined linear relationships between input variables (features) and the output variable (target)
- Our model will help analyze how different salary components, like mean salaries across various job categories and age groups, influence the gender pay gap

Model Selection :

- Random forest regressor
- Bagging regressor with and without pasting
- AdaBoost regressor

EVALUATION METRICS



EVALUATION METRICS

Without HPT



Random Forest

MAE 0.516
RMSE 0.711
R2 score 0.773



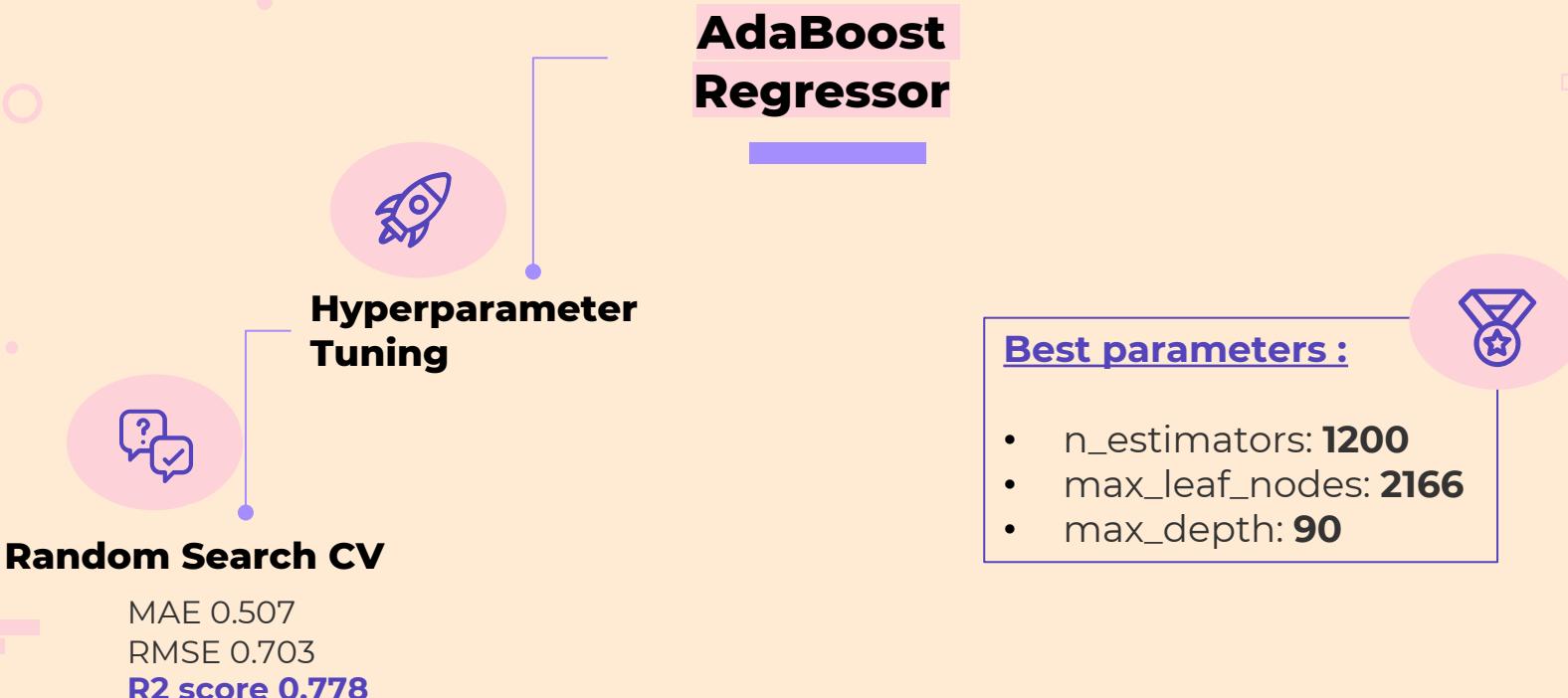
Bagging

MAE 0.530
RMSE 0.724
R2 score 0.765

Bagging w/ pasting

MAE 0.528
RMSE 0.725
R2 score 0.764

EVALUATION METRICS



07

CHALLENGES | NEXT STEPS



THANKS!

**DO YOU HAVE ANY
QUESTIONS?**

pragassam.stephanie@gmail.com