

# ***Job offers in France***

*Data acquisition, extraction, & storage Project*

---

Marwa Nair  
Fatma Zohra Rezkellah  
Abderrahmane Kasmi

# ***Presentation plan***

---

**01.** INTRODUCTION

**02.** DATA ACQUISITION

**03.** DATA TRANSFORMATION

**04.** DATA STORAGE & QUALITY ASSESSMENT

**05.** CONCLUSION

# ***Introduction***

---

## *GOAL OF THE PROJECT*

- **Objective:** Create a clean and structured dataset of job offers in France.
- **Use cases:**
  - Develop a machine learning model for job recommendations, job salary prediction,...
  - Build a job search engine.
  - Design a graphical interface for easy job offer navigation.



# ***Introduction***

---

## ***CHALLENGES***

### ***Website Scraping Restrictions***

*Many platforms like LinkedIn and Indeed block scraping efforts.*

### ***Paid APIs***

*Most job APIs require subscriptions (e.g., Glassdoor, LinkedIn).*

# *Data Acquisition*

---

## *DATA SOURCES*



<https://developer.adzuna.com>



<https://www.francetravail.fr>

# Data Acquisition



## USING ADZUNA API

**Libraries:** *Requests, JSON, Pandas.*

### Process:

1. API Access with authentication.
2. Send requests and parse JSON responses.
3. Extract job details.
4. Convert data to a DataFrame and save as a CSV.

```
→ "latitude": 48.57042,  
→ "salary_is_predicted": "0",  
→ "company": {  
    "__CLASS__": "Adzuna::API::Response::Company",  
    "display_name": "SCHMIDT"  
},  
→ "contract_type": "contract",  
    "__CLASS__": "Adzuna::API::Response::Job",  
→ "longitude": 0.2171,  
→ "description": "Imaginez\u00e0 ce matin vous franchissez les portes de  
→ "created": "2024-11-07T16:47:15Z",  
    "adref": "eyJhbGciOiJIUzI1NiJ9.eyJzIjoibm1lSF1EQ243eEdBaHpkVW5SdEdIUSIs  
    "id": "4930018941",  
→ "redirect_url": "https://www.adzuna.fr/details/4930018941?utm_medium=ap  
→ "category": {  
    "label": "Emplois Vente",  
    "__CLASS__": "Adzuna::API::Response::Category",  
    "tag": "sales-jobs"  
},  
→ "title": "Franchis\u00e9 Schmidt h/f",  
    "location": {  
        "area": [  
            "France",  
            "Normandie",  
            "Orne"
```

# Data Acquisition

FROM FRANCE TRAVAIL

**Libraries:** *Selenium, BeautifulSoup, Pandas.*

## Process:

1. Use Selenium with Chrome WebDriver in headless mode.
2. Scrape jobs details using BeautifulSoup.
3. Handle Pagination.
4. Save extracted data to a CSV file using Pandas.

Offre n° 185KPKZ

## Coordinateur santé et bien-être (H/F)

75 - PARIS 07 - [Localiser avec Mappy](#)



Actualisé le 12 décembre 2024

Le coordinateur santé et bien-être prend en charge les sports et les activités récréatives sur le campus, tout en assurant la liaison avec les institutions et les partenaires extérieurs au campus, tels que les agents de la mairie et les associations sportives locales.

### RESPONSABILITES

#### Activités de compétition

- Assister dans l'organisation et la coordination du programme d'activités compétitives de l'AUP.
- Assurer la liaison avec les entraîneurs des équipes et les étudiants pour les activités de compétition.
- Assurer la liaison avec la Fédération française du sport universitaire (FFSU), la LIFSU (Ligue Île de France Sport Université) et d'autres organisations sportives.

-  CDD - 6 Mois  
Contrat travail
-  35H Travail en journée
-  Salaire brut : Annuel de 24000.0 Euros à 25000.0 Euros sur 12.0 mois  
Chèque repas
-  Déplacements : Jamais

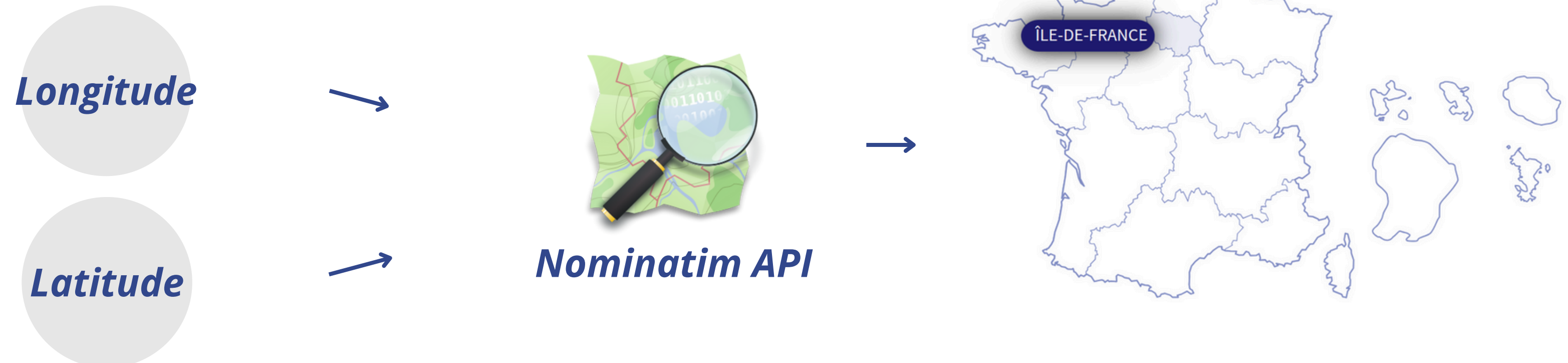
```
▼ <span itemtype="http://schema.org/PostalAddress" itemscope  
  itemprop="address">  
    <span content="75007" itemprop="postalCode"></span>  
    <span content="Paris" itemprop="addressLocality"></span>  
    <span content="Île-de-France" itemprop="addressRegion"></span>  
    <span content="FRANCE" itemprop="addressCountry"></span>  
    <span itemprop="name">75 - PARIS 07</span>  
</span>
```

# Data transformation

- Price extraction (FranceTravail).



- Region extraction (Adzuna).





# ***Data transformation***

---

- Price Normalization (Adzuna).
- Description Translation  $\mathcal{T}_A$
- Job category alignment

## ***France Travail***

- Achats / Comptabilité / Gestion
- Arts / Artisanat d'art
- Banque / Assurance
- Bâtiment / Travaux Publics

*22 Category*



## ***Adzuna***

- Accounting & Finance Jobs
- IT Jobs
- Sales Jobs
- Engineering Jobs
- Other Jobs

*30 Category*

# Quality Assessment

# Overview of the Dataset

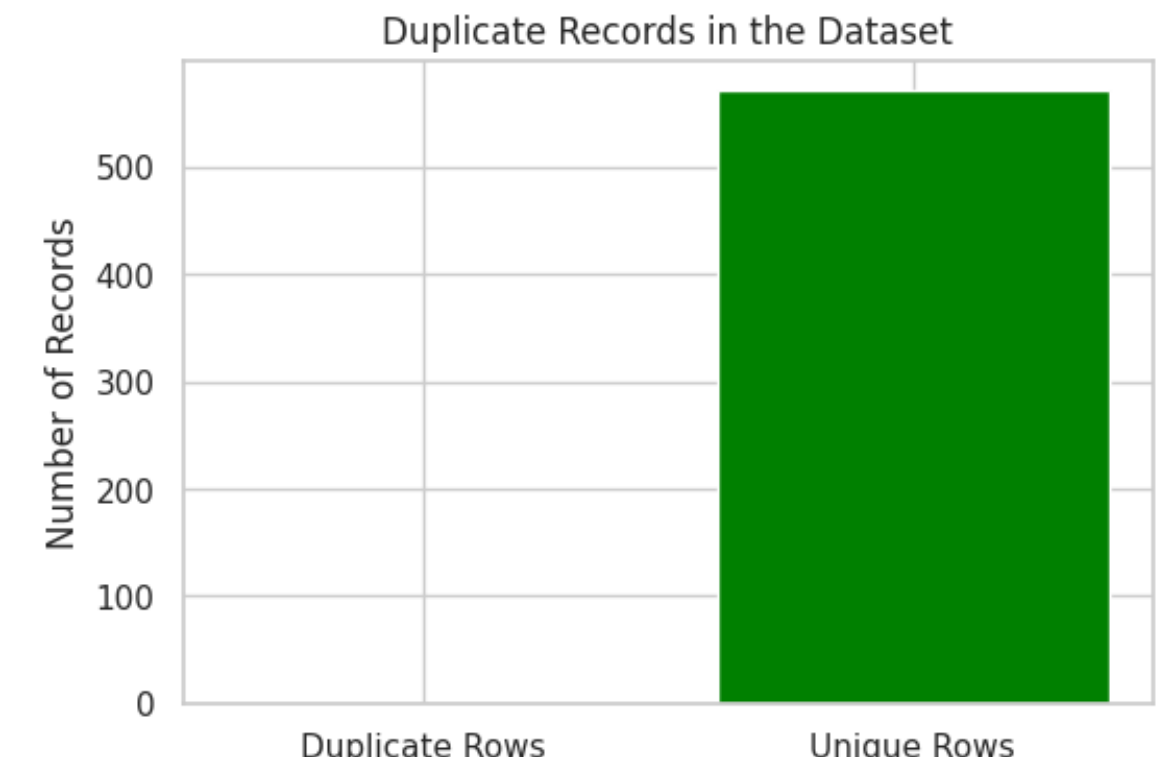
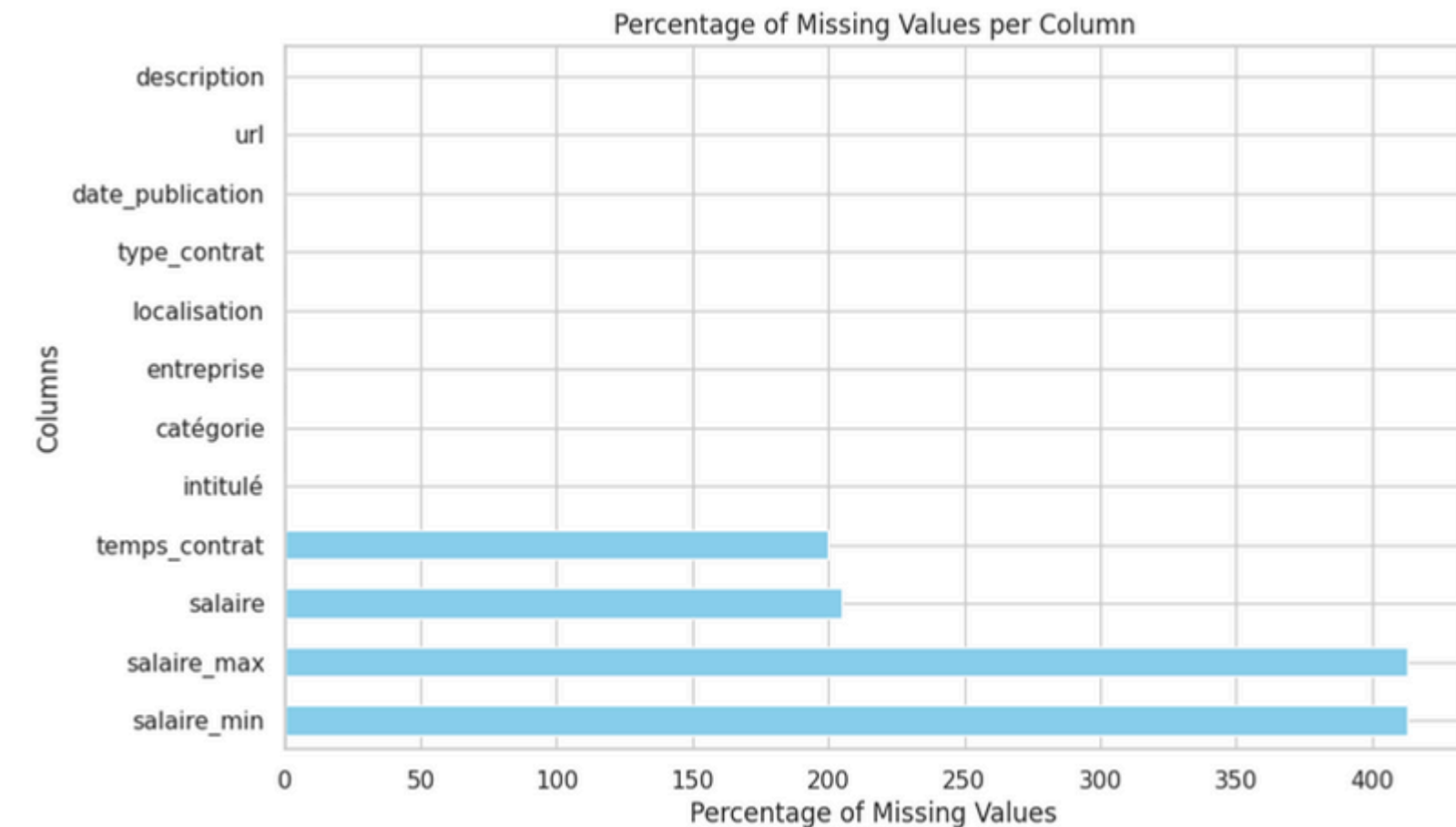
- (571 rows, 12 columns)
- Key columns such as: intitulé, entreprise, localisation, type\_contrat, salaire, etc.

|   | intitulé                          | catégorie | entreprise | localisation       | type_contrat | temps_contrat | date_publication | url                                                                                                             | salaire_min | salaire_max | salaire | description                                       |
|---|-----------------------------------|-----------|------------|--------------------|--------------|---------------|------------------|-----------------------------------------------------------------------------------------------------------------|-------------|-------------|---------|---------------------------------------------------|
| 0 | Agent Commercial Immobilier (H/F) | Unknown   | SAFTI      | Nouvelle-Aquitaine | CDD          | NaN           | 2024-05-14       | <a href="https://www.adruna.fr/details/46923484067">https://www.adruna.fr/details/46923484067</a><br>utm_m...   | 2500.0      | 8333.333333 | NaN     | Nous recrutons et formons des agents indépendants |
| 1 | Agent Commercial Immobilier (H/F) | Unknown   | SAFTI      | Nouvelle-Aquitaine | CDD          | NaN           | 2024-05-14       | <a href="https://www.adruna.fr/details/469234895277">https://www.adruna.fr/details/469234895277</a><br>utm_m... | 2500.0      | 8333.333333 | NaN     | Nous recrutons et formons des agents indépendants |

# Quality Assessment

## Quality Check Methodology

- *Missing values*
- *Duplicates*
- *Outliers: Extreme values in the salary column (0)*
- *Invalid URLs (0)*



# *Quality Assessment*

---

- *Handling missing values for the 'Salary feature' where salary\_min and salary\_max exist: Replaced with the average of the two.*
- *The absence of certain information (e.g., contract duration) is often meaningful and should not be hidden --> Replaced by 'Not Specified'*
  - *Keeps the dataset closer to real-world job market conditions.*
  - *Prevents the introduction of misleading or false information (Replacing the missing salary with the mean?)*

# Data Storage

---

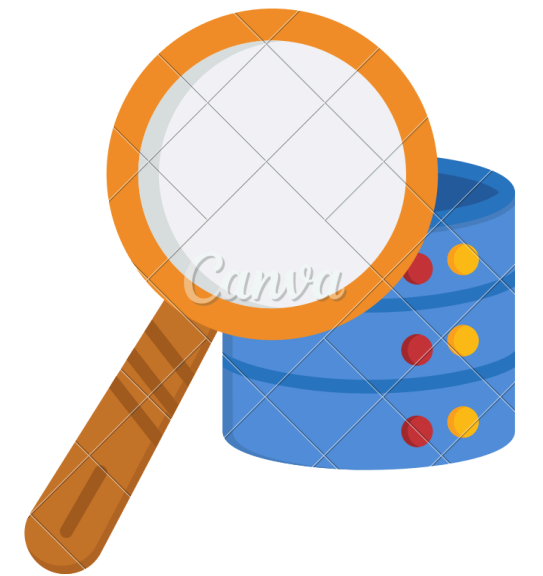
- *The dataset is relatively small, containing a single table, which makes the **CSV** file format a simple and highly accessible solution.*
- *No need for additional infrastructure like relational databases or NoSQL systems for simple datasets!*
- *This simplicity supports quick analysis, ensuring that stakeholders or collaborators can easily access the data.*



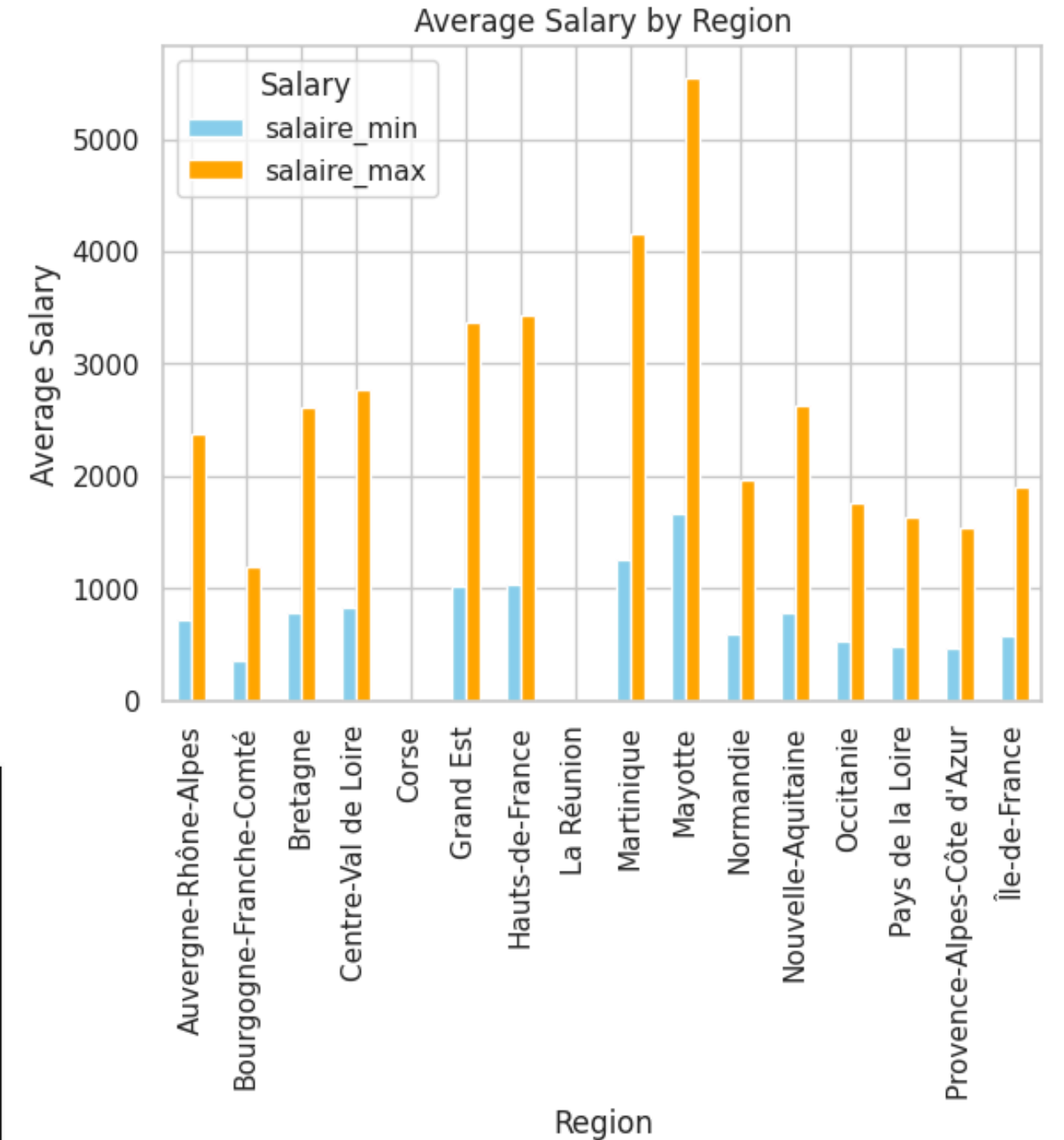
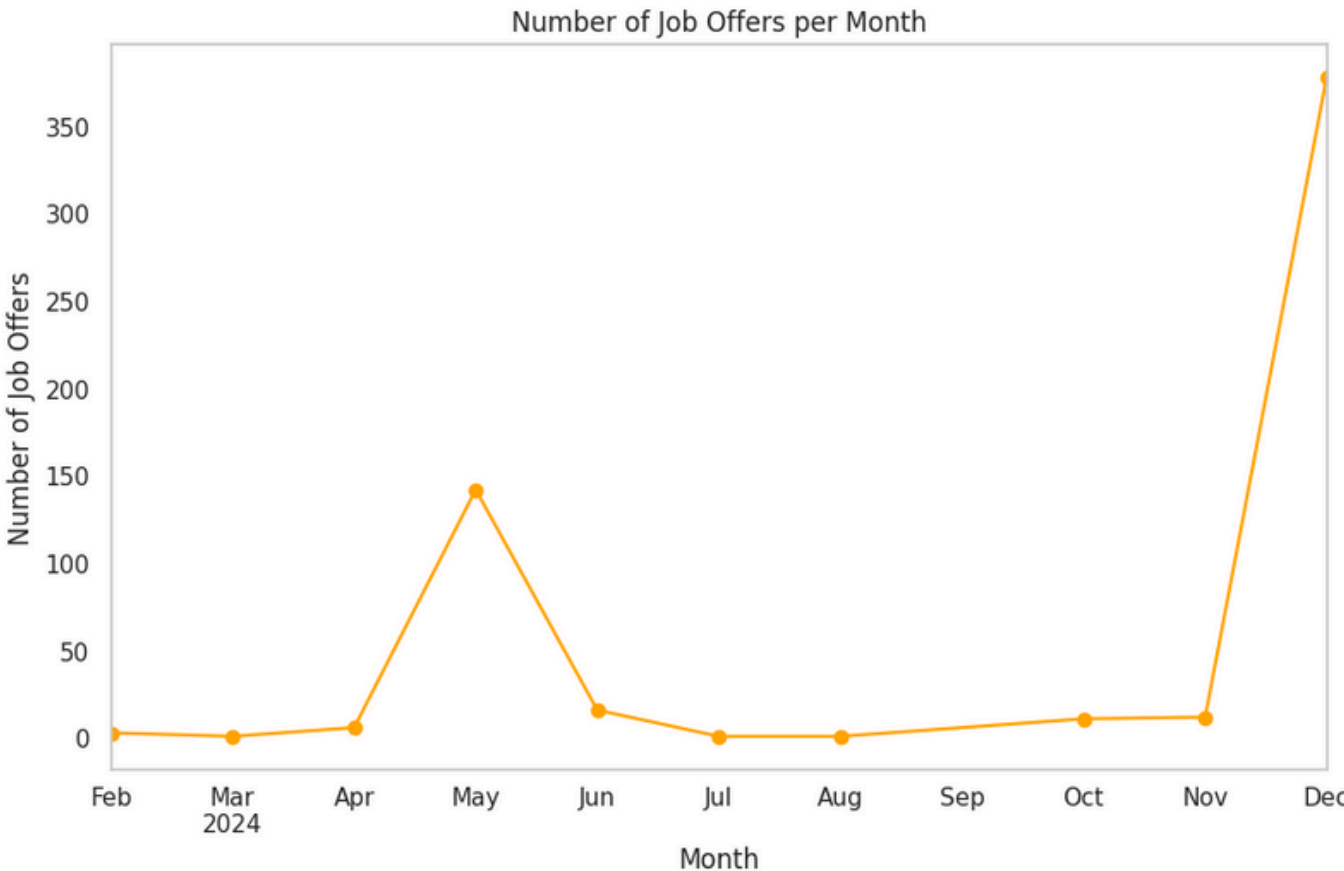
# Queries execution

---

- A set of representative queries were executed using **pandas** to extract valuable insights from the dataset.
- These queries are relevant to the job offers analysis and provide an understanding of the dataset's structure and trends.
- Type of queries:
  - Job Offers per Region, Year
  - Top hiring companies
  - Contract type distribution
  - Most common job titles

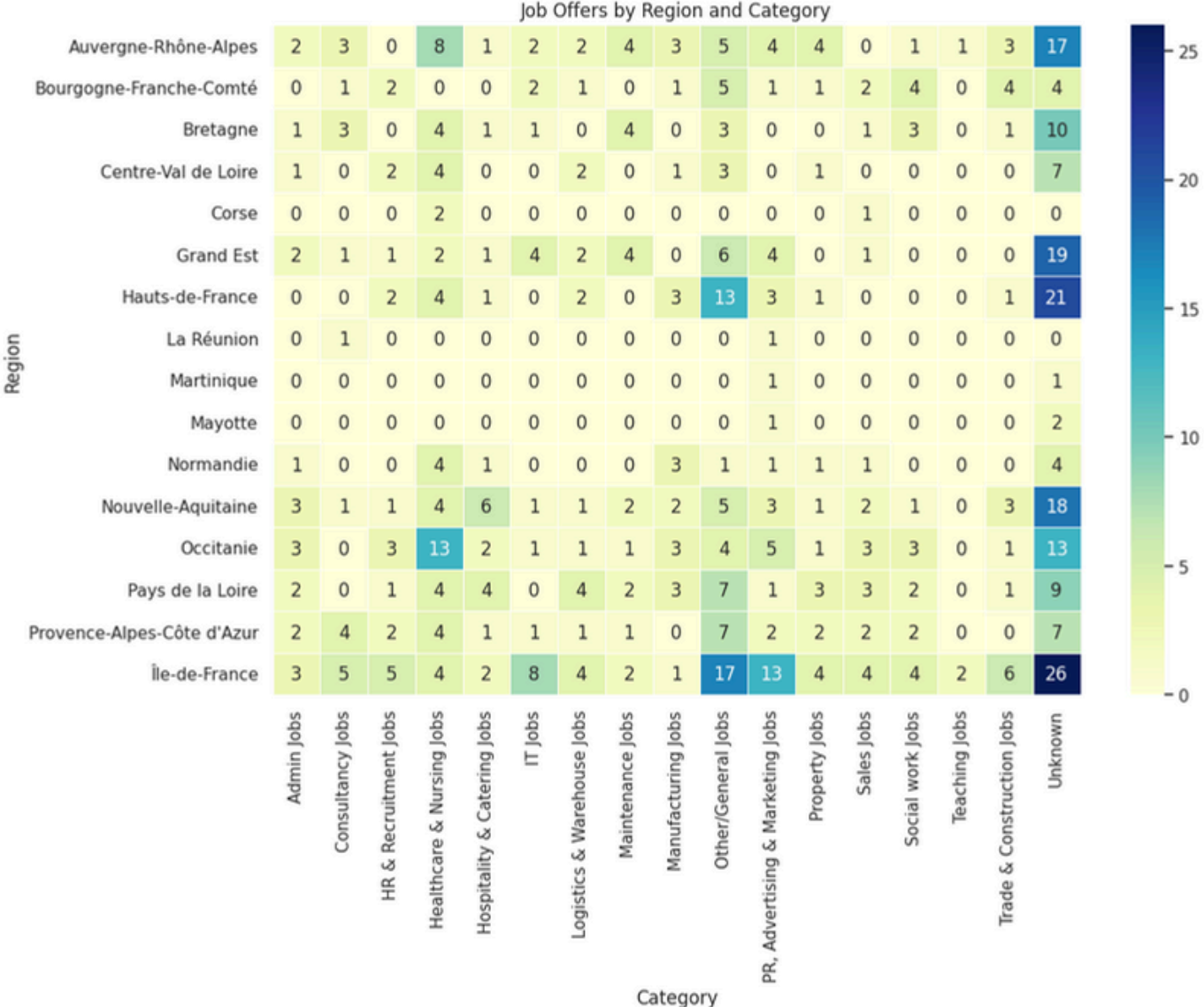
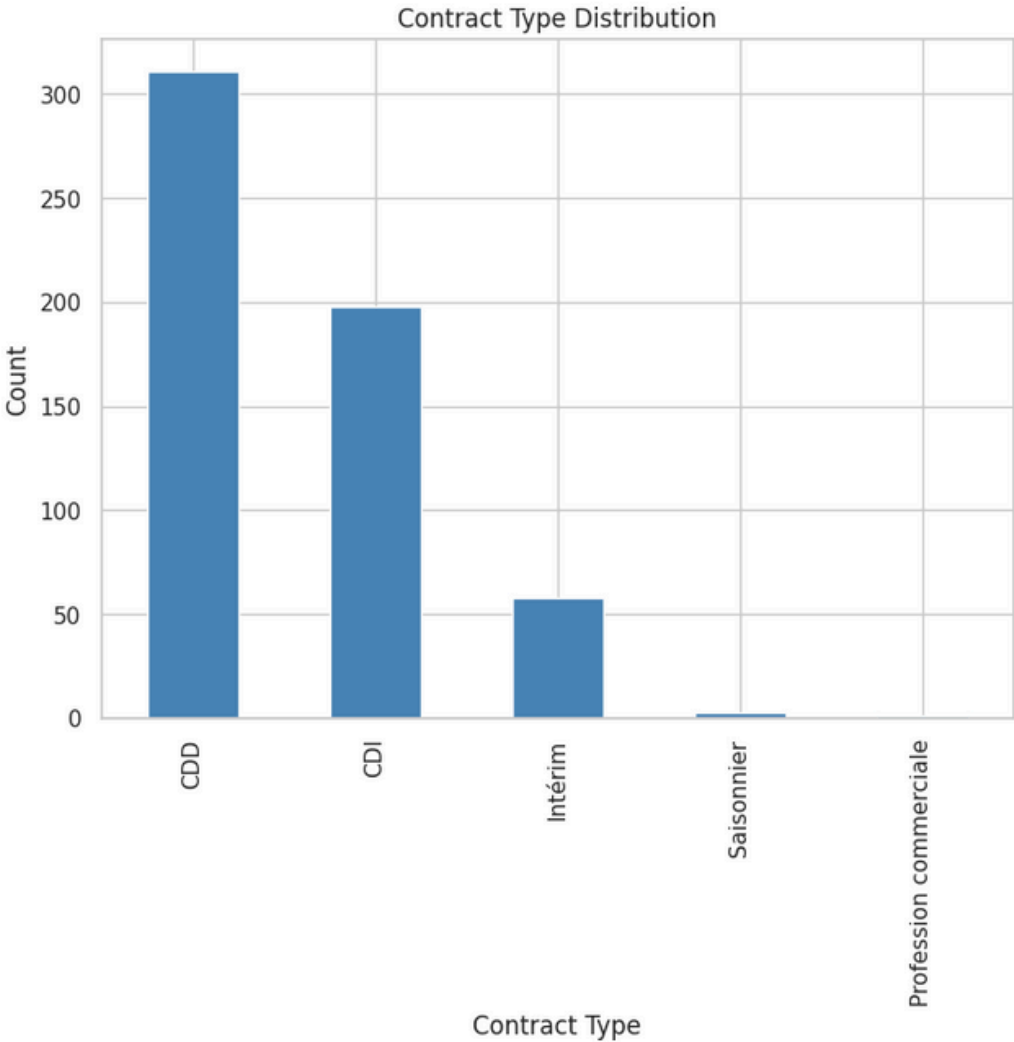
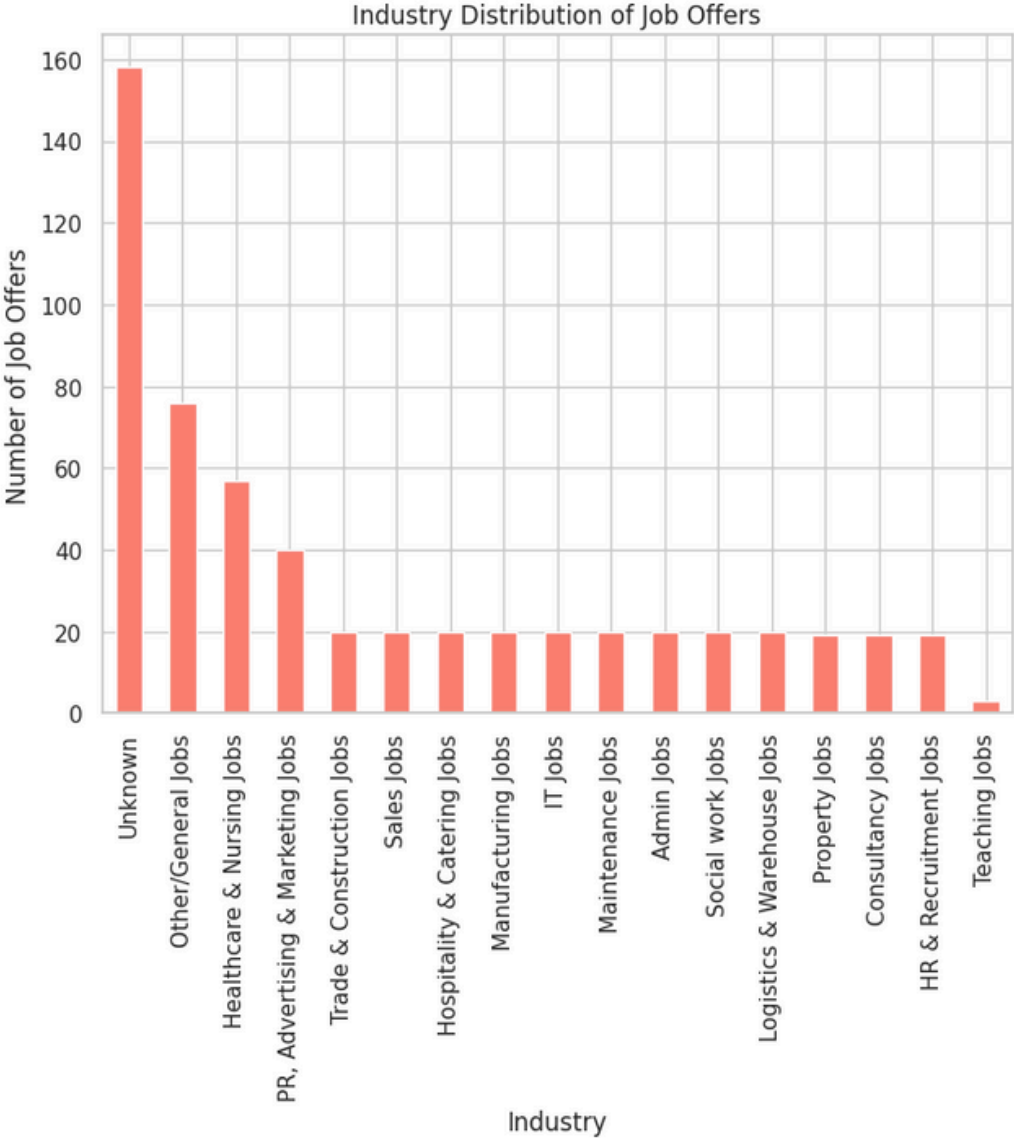


# Queries Results





# Queries Results





# Conclusion

---

- *A dataset that can be used for machine learning application, study of phenomena related to job offers or for building graphical interfaces ...*
- **Possible improvement:**
  - *Improve the mechanism of price extraction (fine-tuning).*
  - *Improve the translation mechanism.*
  - *Extract more details about the location (not only the region).*

***Thank you!***