Master 2 Data Science et société numérique

MARCHÉ DU TRAVAIL DANS LE DOMAINE DATA SCIENCE SUR LINKEDIN EN FRANCE

PHAM Dong Pha

1. La problématique

Comme tous les autres étudiants qui sont en train de terminer leur Master 2, je souhaiterais chercher d'un emploi dans le domaine de Data Science après mon formation D2SN.

Lorsque vous êtes sur le marché du travail et que vous recherchez votre prochaine affectation, que feriez-vous? Parcourez les différents sites Web, consultez LinkedIn, Indeed, Glassdoor, actualisez votre boîte e-mail jour après jour pour vous assurer de ne manquer aucune opportunité?.

Une autre façon créative consiste à créer un outil de décapage des emplois. Vous pouvez collecter toutes les informations en cliquant sur un bouton.

Pour se préparer à l'avenir de poursuivre une carrière dans le domaine : Je me demande quelles sont les tendances du recrutement dans le domaine de Data Science?. Quelles sont les exigences pour les postes dans le domaine de Data Science?. Que je dois faire pour optimiser ma recherche d'emploi dans ce domaine?. Et bien d'autres questions que je poserai dans la section méthodologie ci-dessous. Donc, je me suis mis à chercher des réponses sur ce sujet.

2. Le plan

Pour ce projet, je souhaiterais scraper des données liés aux l'offres dans le domaine de Data Science, publiés en France sur LinkedIn en utilisant Python avec ses package Beautiful Soup et Selenium (*LinkedIn est le plus grand réseau social pour les professionnels et les demandeurs d'emploi (www.linkedin.com)*). Cela représente environ 900 offres. Une fois collectées par scraping via LinkedIn, les données d'offres d'emploi doivent être structurées pour l'analyse statistique. Après avoir de dataframe, j'abordais ce dataframe en effectuant une analyse exploratoire des données (Exploratory Data Analysis), une analyse de text, et WordCloud en utilisant Python.

Table des matières

PARTIE I : LES DONNÉES

1) Comment scrapper les données (l'informations du poste) sur LinkedLn?

C'est une question assez difficile, car ce n'est pas une tâche facile. Nous allons d'abord explorer le concept de "Le scrapping".

Le scraping, qu'est ce que c'est?.

Le scraping permet de retirer des informations dans une page web pour pouvoir les utiliser sur sa propre page. C'est une technique marketing qui vise la concurrence directe, parce qu'elle permet de récolter du contenu sur un site web et de le copier-coller tel qu'il est. Elle est souvent utile pendant une veille concurrentielle et très pratiquée par les sites e-commerce. On l'utilise dans le cadre d'une analyse marketing. Seulement, sur le plan déontologique, cela n'est pas conseillé. Toutefois, le scraping Linkedin fait partie des méthodes de "growth hacking", des méthodes qui permettent d'améliorer rapidement son marché, sa position face à la concurrence, de trouver des partenaires d'affaires, etc.

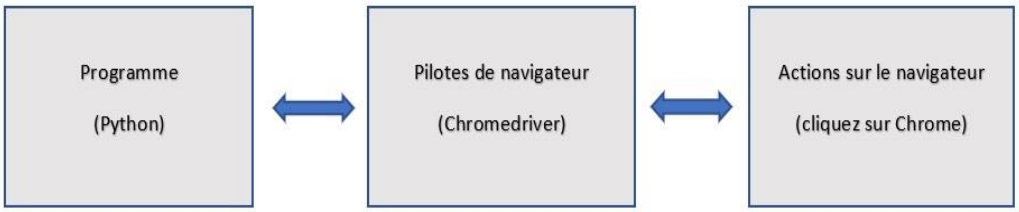
Sur ce projet, le scraping est un procédé automatique de récolte de données sur LinkedIn, ici, ils sont des informations des postes en France avec des mots-clés de recherche: "Data science" Ils sont des infos de: ID, Titre d'emploi, Nom de l'entreprise, Location, Nombre d'applications, Titre d'emploi, Date de publication, Description de l'emploi, Niveau d'expérience, Type d'emploi, Domaine de poste, Lien de candidature à un emploi, Lien de poste.

Comment faire du Scraping sur Linkedin?

Pour construire "le scraper" il faut commencer par chercher les emplacements des données dans HTML de la page web à copier. Balises, langage Python, code HTML… une inspection est nécessaire pour inspecter l'élément et trouver la balise. Pour les scrapping, voici ce qu'il faut faire:

+ Procédure au sélénium

La bibliothèque que j'utilise est «sélénium» car certaines actions sont demandées sur le site Web, telles que «cliquer», «déplacer la souris».



Ci-dessus, C'est une procédure simple pour comprendre l'utilisation du "Selenium".

Nous utilisons la bibliothèque Python "Selenium" pour envoyer une requête au navigateur.

Le programme envoie des demandes au pilote du navigateur qui agira comme un concentrateur pour transmettre les demandes au navigateur. Différents navigateurs ont des pilotes de navigateur différents. Ici, j'utilise Chromedriver. Assurez-vous de télécharger le chromedriver compatible avec la version de votre navigateur (*Pour trouver votre version Chrome, cliquez sur les trois points verticaux dans le coin supérieur droit de votre navigateur. Choisissez: Aide - À propos de Google Chrome*).

Le pilote du navigateur transmettra la demande au navigateur et effectuera les actions. Il renverra également les données du navigateur vers notre code par la suite.

Importer les packages sur Python :

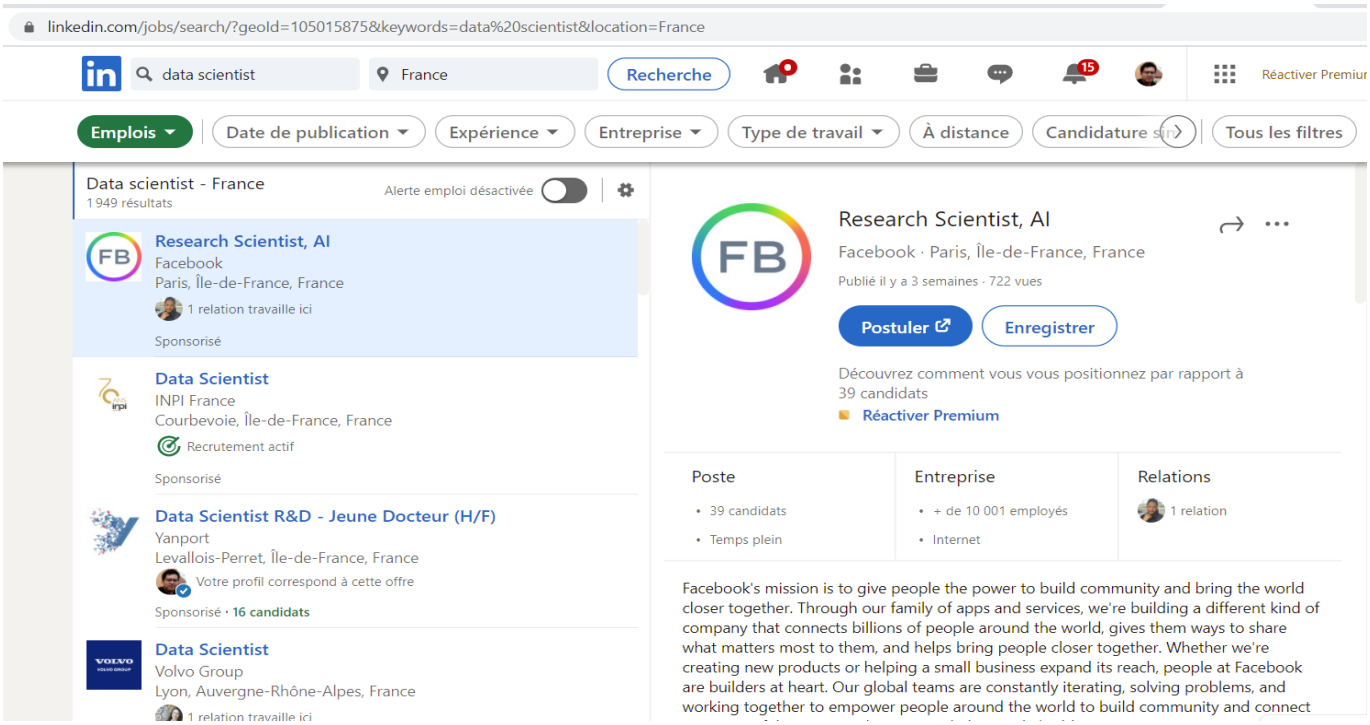
```
Entrée [25]: # importer les packages
import pandas as pd
import re

from bs4 import BeautifulSoup
from datetime import date, timedelta, datetime
from IPython.core.display import clear_output
from random import randint
from requests import get
from selenium import webdriver
from selenium.webdriver.common.action_chains import ActionChains
from time import sleep
from time import time
start_time = time()

from warnings import warn
```

- Vue rapide du site Web

Mon objectif pour ce projet est de décrocher des emplois en science des données situés à France avec le mot de recherche "data science", location : "France".

Voici mon URL:

https://www.linkedin.com/jobs/search/?geoId=105015875&keywords=data%20scientist&location=France&start=0

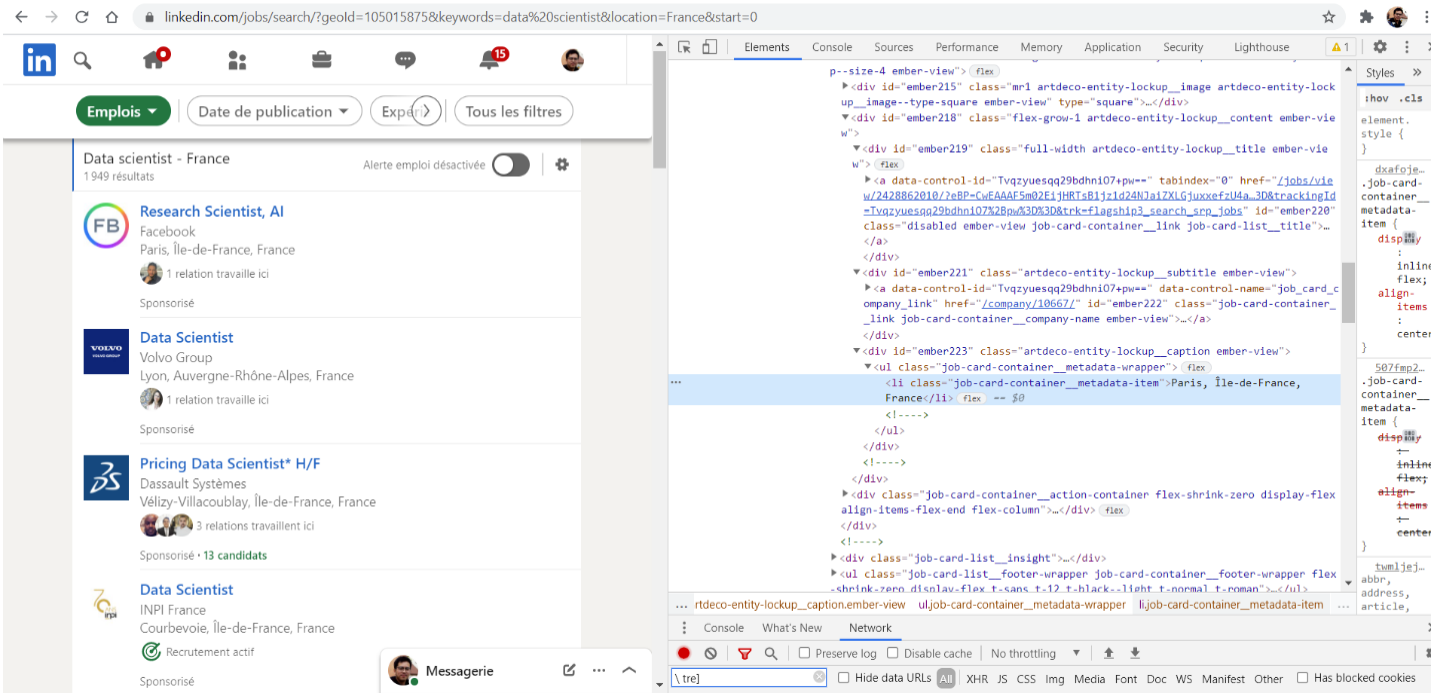


- Configurer chromedriver:

```
url = "https://www.linkedin.com/jobs/search/?geoId=105015875&keywords=data%20scientist&location=France&start=0"
no_of_jobs = 1000
```

```
driver = webdriver.Chrome('C:/webdrives/chromedriver.exe')
driver.get(url)
sleep(3)
action = ActionChains(driver)
```

Ensuite, nous devons savoir combien d'emplois nous avons trouvé grâce à cette recherche. Si vous n'êtes pas familiarisé avec le web scraping, vous pouvez cliquer avec le bouton droit de la souris, sélectionner "inspecter" ou appuyer sur F12. Cliquez sur «Éléments» et vous trouverez les codes HTML.



- Parcourir tous les emplois

La façon dont les offres d'emploi LinkedIn fonctionnent consiste à charger plus d'emplois si vous faites défiler la barre du navigateur. Cependant, lorsque vous faites glisser la barre plusieurs fois, elle ne se charge pas automatiquement, mais vous devez cliquer sur le bouton "Voir plus d'emplois".

```
# pour afficher plus d'emplois. Dépend du nombre d'emplois sélectionnés
i = 2
while i <= (no_of_jobs/25):
    driver.find_element_by_xpath('/html/body/main/div/section/button').click()
    i = i + 1
    sleep(5)
```

- Choisir l'élément

Le premier élément que nous aimerions obtenir est le nombre d'emplois en data science en France sur LinkedIn. Une fois que nous déplaçons la souris sur le 994 en surbrillance (Résultats trouvés), les codes correspondants sont mis en surbrillance.

Nous analyson de la page Web et nous utiliserons Beautiful Soup sur Python pour sélectionner 994 emplois.

- Trouvez tous les emplois

J'utilise la balise ‘jobs-search_results-list’ pour trouver les listes d’emplois. Ensuite, nous pouvons utiliser la boucle for pour examiner chaque travail et obtenir les détails.

```
# analyse de la page Web
pageSource = driver.page_source
lxml_soup = BeautifulSoup(pageSource, 'lxml')

# recherche de tous les emplois
job_container = lxml_soup.find('ul', class_ = 'jobs-search__results-list')
```

- Charger les détails de la tâche dans Dataframe

Une fois que nous utilisons la souris pour localiser les codes HTML de chaque élément, nous pourrions obtenir ci-dessous les détails de la description de poste:

Nom de l’entreprise, Location de poste, Identifiant de poste (ID), Titre de poste, Nombre d’application de poste, Date de publication de poste, Description de l'emploi, Niveau d'expérience, Type d'emploi (Full time, intership, contract, other), Industrie.

J’ai utilisé les codes dans mon script pour extraire les éléments en surbrillance, par exemple:

```python
# configuration de la liste pour les informations sur l'emploi
company_name = []
job_location = []
job_id = []
post_title = []
nombre_application =[]
post_date = []
job_desc = []
level = []
emp_type = []
industries = []

# for loop pour les élements
for job in job_container:

     # nom de l'entreprise
    company_names = job.find('li', class_ = 'job-card-container__link job-card-container__company-name ember-view').text
    company_name.append(company_names)

    # location de poste
    job_locations = job.find("li", class_ ="job-card-container__metadata-item").text
    job_location.append(job_locations)

    # identifiant de poste sur Linkedin
    job_ids = job.find('a', href=True)['href']
    job_ids = re.findall(r'(?!-)([0-9]*)(?=\?)',job_ids)[0]
    job_id.append(job_ids)

    # titre de poste
    job_titles = job.find('div', class_ ="disabled ember-view job-card-container__link job-card-list__title").text
    post_title.append(job_titles)

    # nombre_d'application de poste
    nombre_applications = job.find('li', class_ ='job-card-container__applicant-count job-card-container__footer-item job-card-c
    nombre_application.append(nombre_applications)
```

```python
    # date de publication de poste
    post_dates = job.select_one('time')['datetime']
    post_date.append(post_dates)

# for loop pour la description du poste et les critères
for x in range(1,len(job_id)+1):

    # cliquer sur différents conteneurs de travaux pour afficher des informations sur le travail
    job_xpath = '/html/body/main/div[3]/div[3]/div[3]/div[1]/div[1]/div[1]/div[1]/div[1]/div[1]/div[2]/a/img'.format(x)
    driver.find_element_by_xpath(job_xpath).click()
    sleep(3)

    # Description de poste
    jobdesc_xpath = '/html/body/main/div[4]/div[3]/div[1]/div[1]/div[1]/div[2]/article'
    job_descs = driver.find_element_by_xpath(jobdesc_xpath).text
    job_desc.append(job_descs)

    # conteneur de critères d'emploi sous la description
    job_criteria_container = lxml_soup.find('article', class_ = 'jobs-description__container jobs-description__container--condens
    all_job_criterias = job_criteria_container.find_all("div", class_='jobs-box__html-content jobs-description-content__text t-14

    # niveau d'expérience
    seniority_xpath = '/html/body/main/section/div[2]/section[2]/ul/li[1]'
    seniority = driver.find_element_by_xpath(seniority_xpath).text.splitlines(0)[1]
    level.append(seniority)

    # type d'emploi
    type_xpath = '/html/body/main/section/div[2]/section[2]/ul/li[2]'
    employment_type = driver.find_element_by_xpath(type_xpath).text.splitlines(0)[1]
    emp_type.append(employment_type)

    # industrie
    industry_xpath = '/html/body/main/section/div[2]/section[2]/ul/li[4]'
    industry_type = driver.find_element_by_xpath(industry_xpath).text.splitlines(0)[1]
    industries.append(industry_type)

    x = x+1
```

La dernière étape consiste à charger ces données dans un DataFrame pour une analyse plus approfondie.

Référence pour le partie scrapping des données :

1) Selenium with Python - Baiju Muthukadan : https://realpython.com/beautiful-soup-web-scraper-python/

2) Beautiful Soup: Build a Web Scraper With Python - Martin Breuss : https://selenium-python.readthedocs.io/

2) Web scrapping with Python - Kerry Parker : https://blog.lesjeudis.com/web-scraping-avec-python

PARTIE II: METHODOLOGIE

Comme mentionné dans la problématique, afin de bien préparer les compétences et optimiser la recherche d'emplois dans le domaine de la Data science, puis postuler aux entreprises, dans les villes de France, avec le type de travail que vous souhaitez. Ensuite, nous nous poserons des questions, par exemple :

- Quelles entreprises et villes de France ont le plus d'offres d'emploi dans le domaine de la Data science ?

- Quels titres de l'offre d'emploi sont les plus populairesdans le domaine de la Data science en France ?

- Si nous recherchons un poste dans le domaine de la science des données (par exemple Data scientist, Data Analyst, Consultant Big Data, Ingénieur Machine Learning, ...etc), quelles exigences devons-nous intégrer, quelles compétences, quelle programmation logiciel, quelles qualifications, ..etc pour avoir la possibilité de trouver facilement un emploi à ce poste?

- Quel est le nombre moyen d'années d'expérience requis dans ce domaine de la science des données ?

- Quels outils et logiciels de programmation les employeurs en France vous demandent-ils d'intégrer dans le domaine de la Data science ? Quels logiciels et outils sont les plus demandés ?

- Quel industrie est le plus populaire pour les offres d'emploi ?

- Quelles sont les catégories d'offres d'emploi dans le domaine de la science des données ? Quels genres sont les plus populaires ?

- Quel type de travail le plus populaire pour les offres d'emploi ? Temps plein? Stage? Temp partiel,..etc?

- Quels logiciels et outils sont les plus demandés ?

Pour répondre à ces questions:

D'abord, j'aborde le dataset en effectuant une l'analyse exploratoire des données (EDA) sur l'ensemble de données général et une analyse de texte et utilisant les méthodes machine learning sur Python prédire le résultat

D'abord, nous clarifions le concept et le travail de méthode EDA. Donc, qu'est-ce que l'analyse exploratoire des données (EDA)?

L'analyse exploratoire des données (EDA), également connue sous le nom d'exploration de données, est une étape du processus d'analyse des données, où une certain nombre de techniques sont utilisées pour mieux comprendre l'ensemble de données utilisé.

- Extraire des variables importantes et laisser des variables inutiles

- Identifier les valeurs aberrantes, les valeurs manquantes ou les erreurs humaines

- Comprendre la (les) relation(s), ou l'absence de, entre les variables

- En fin de compte, maximiser vos connaissances sur un ensemble de données et minimiser les erreurs potentielles plus tard dans le processus

En réalisant une EDA, vous pouvez transformer un ensemble de données presque utilisable en un ensemble de données entièrement utilisable. Je ne dis pas que EDA peut par magie nettoyer n'importe quel ensemble de données - ce n'est pas vrai. Cependant, de nombreuses techniques EDA peuvent remédier à certains problèmes courants présents dans chaque ensemble de données. EDA permet de garantir que vous choisissez les techniques statistiques correctes pour analyser et prévoir les données.

EDA fait deux choses principales:

1. Cela aide à nettoyer un ensemble de données.

2. Cela vous permet de mieux comprendre les variables et les relations entre elles.

Composantes de l'EDA

Pour moi, il y a les principaux composants de l'exploration des données:

- Comprendre vos variables

- Nettoyage de votre jeu de données

- Analyse des relations entre les variables

Ici, j'utilise des techniques EDA et technique de l'analyse des donnees textuelle, Nuage de points (WordCloud). Ils montrent les propriétés clés d'un ensemble de données dans un format pratique. Il est souvent plus facile de comprendre les propriétés d'une variable et les relations entre les variables en regardant des graphiques plutôt que de regarder les données brutes.

Voici les tâches que j'ai effectuées en utilisant Python.

I) Préparation des données

Cette étape impliquera la conversion des chaînes de l'ensemble de données dans le type de données approprié pour l'analyse et la suppression de certains mots pour plus de facilité. D'abord, nous importons les bibliothèques sur Python et nous importons le dataset:

	Entreprise	Location	ID	Titre	Application	Date	Description	Niveau_experience	job_category	Type_poste	Industrie
0	Davidson consulting	Valbonne, Provence-Alpes-Côte d'Azur, France	2542453760	Consultant BI (H/F)	Be among the first 25 applicants	2021-05-10	Je découvre les filiales \n\nRejoindre Davids...	Associate	Information Technology	Full-time	Computer Software | Internet | Staffing and Re...
1	GrAI Matter Labs	Paris, Île-de-France, France	2526303569	Application Engineer - Machine Learning	Be among the first 25 applicants	2021-04-05	GrAI Matter Labs utilizes brain-inspired, neur...	Entry level	Engineering | Information Technology	Full-time	Information Technology and Services | Computer...
2	Capgemini Engineering	Greater Toulouse Metropolitan Area	2486573814	Lead Data Scientist	Be among the first 25 applicants	2021-04-27	Notre offre\n\nTESSELLA est le World Class Cen...	Not Applicable	Engineering | Information Technology | Research	Full-time	Information Technology and Services
3	Faurecia	Paris, Île-de-France, France	2522211414	Data Scientist	35 applicants	2021-05-06	Data Scientist (H/F) \n[CDI] \n\nNew trends a...	Not Applicable	Engineering | Information Technology	Full-time	Automotive
4	UNLCK	Paris, Île-de-France, France	2553254896	Lead Data Scientist - Scoring - H/F	Be among the first 25 applicants	2021-05-17	Le poste \n\nPython / Sql / Tableau / Agile ...	Associate	Engineering | Information Technology	Full-time	Information Technology and Services | Computer...

Cet dataframe contient 993 offres d'emploi dans le domain Data Science avec 11 colonnes suivantes:

Entreprise : Nom de l'entreprise de recrutement

Location : Localisation d'offre d'emploi (en France)

ID : Identifiant d'offre d'emploi

Titre : Titre d'offre d'emploi

Application : Nombre de candidats à l'offre d'emploi (mis à jour au moment du scrapping des données)

Date : Date de publication de l'offre d'emploi

Description : Description de l'offre d'emploi

Niveau_experience :

Job_categorie :

Type_poste : Type de l'offre d'emploi, par exemple: Full-time (Temps plein), Internship (Stage), Contract (Contrat), Part-time (Temps partiel), Other (Autre), Temporary (Temporaire).

Industrie : Industrie d'offre d'emploi

Afin de nettoyer les données, j'ai vérifié le nombre de NaN dans chaque colonne.

```
Entreprise            0
Location              0
ID                    0
Titre                 0
Application           0
Date                  0
Description           0
Niveau_experience    15
job_category         15
Type_poste            0
Industrie            15
```

Nous voyons qu'il y a une petite quantité des offres d'emploi manquent ses informations sur le niveau d'experience, la categorie et l'industrie avec un rapport de 15/993. Cela n'aura pas beaucoup d'impact sur la structure des données si je supprime les 15 offres d'emploi manquantes ses informations.

Donc, après avoir supprimé le "Nan", nous avons un dataset avec 978 offres d'emploi.

II) Analyse exploratoire des données (EDA)

Pour répondre aux questions es questions posées au dessus du sujet. Nous analyserons et explorerons chaque colonne de ce dataset pour trouver les réponses.

A) Entreprise :

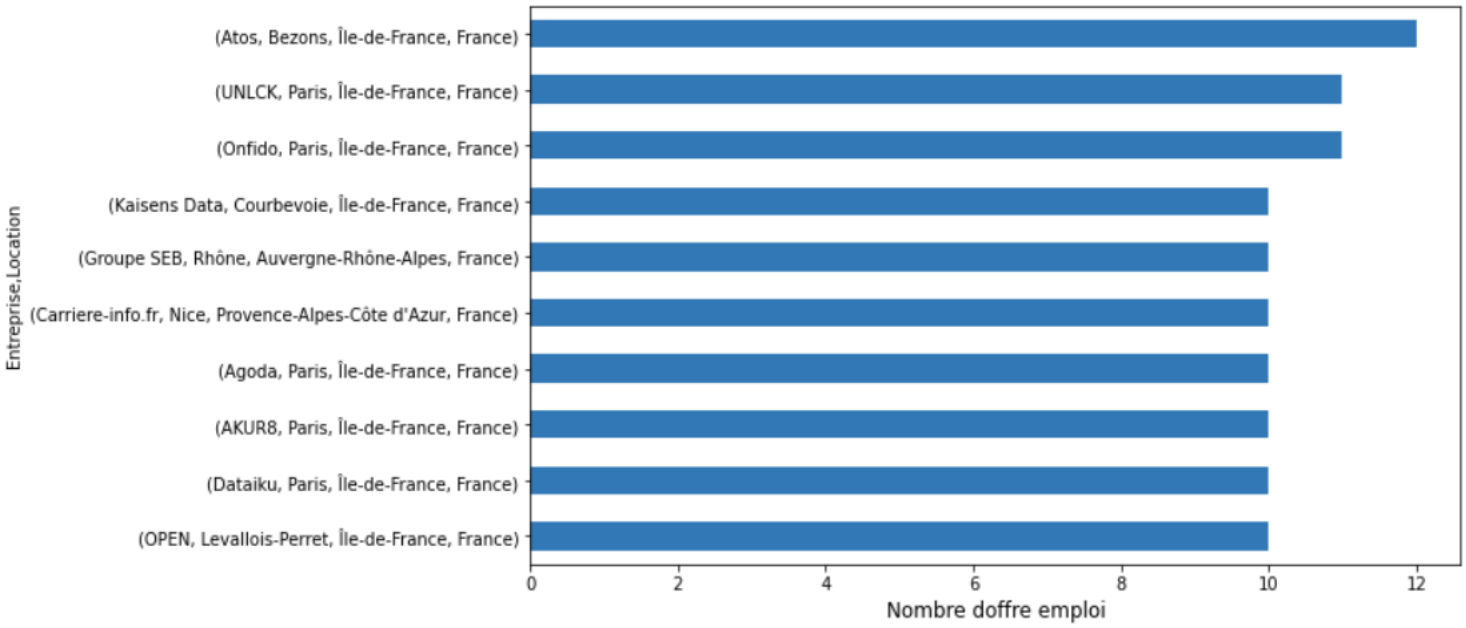
Nous voyons le nombre d'entreprises embauchant dans **993 offres d'emploi**. Bien sûr, chaque entreprise affichera un ou plusieurs offres d'emploi à différents postes.

```
# Nombre d'entreprise
df['Entreprise'].nunique()
```

```
312
```

Donc, il y a **312 entreprises différentes** qui affichent des offres d'emploi dans le domaine Data science sur LinkedIn en France.

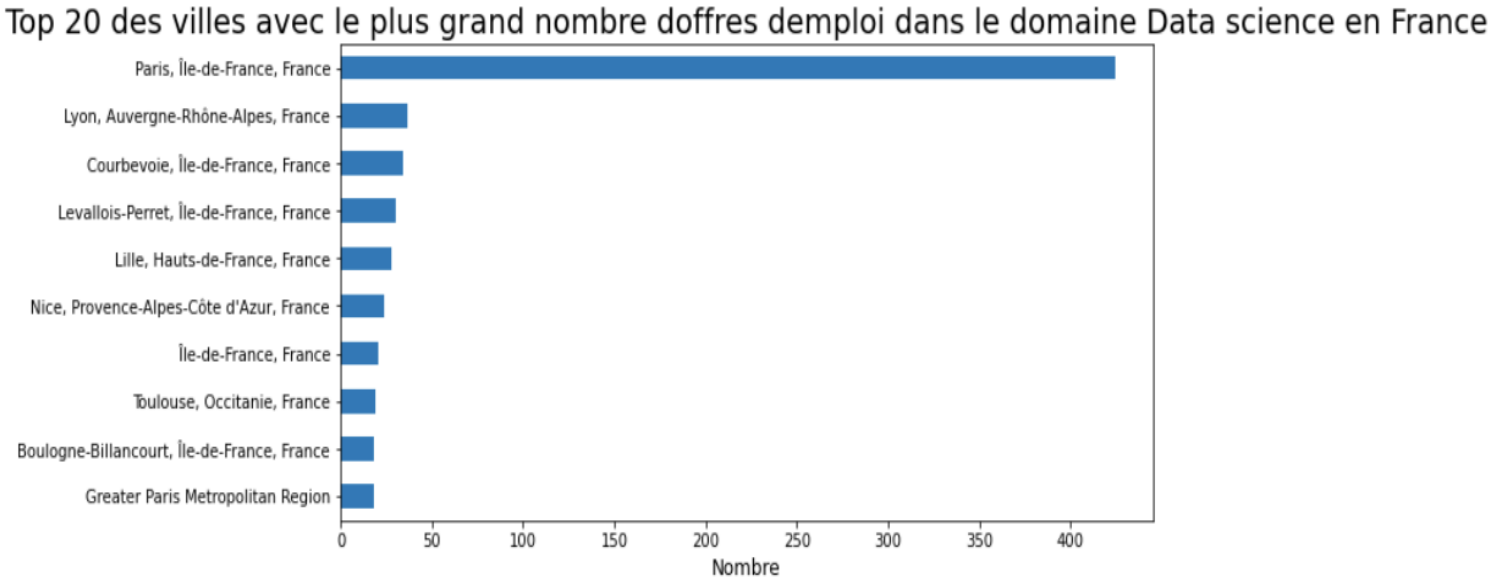
D'abord, nous exportons un graphique à barres pour montrer **top 10 des entreprises**, y compris leur location avec les plus nombre d'offre de l'emploi dans le domaine Data science en France.

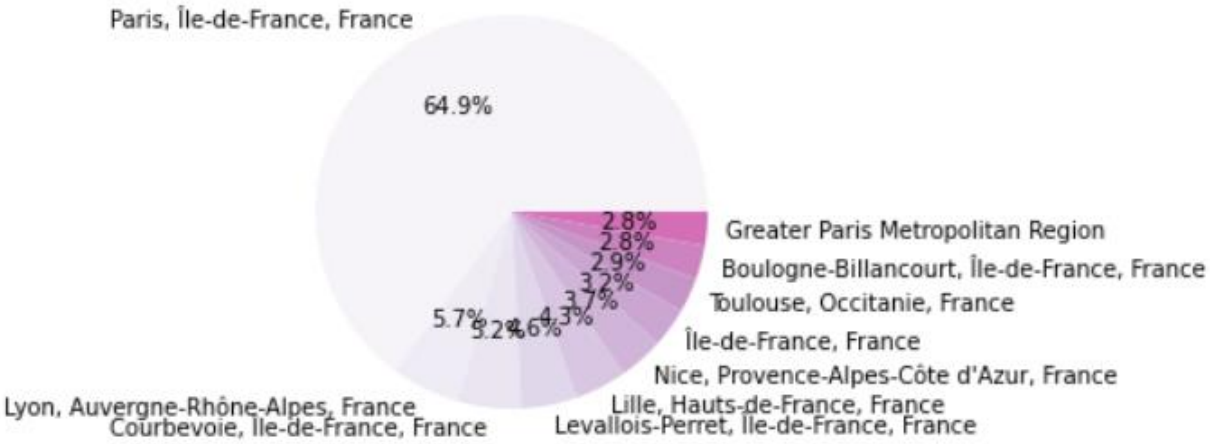


Nous voyons que l'entreprise **Atos à Bezons** est l'entreprise avec le plus grand nombre d'offres d'emploi dans le domaine Data science en France avec 12 offres. Les deuxième rang sont des sociétés **UNLCK, Onfido en Ile-de-France** avec 11 offres. Après, nous verrons quel pourcentage de régions/villes ont des offres d'emploi en France.

B) Location

Nous exportons un graphique à barres pour montrer **top 20 des villes** avec le plus grand nombre d'offres d'emploi dans le domaine Data science en France.

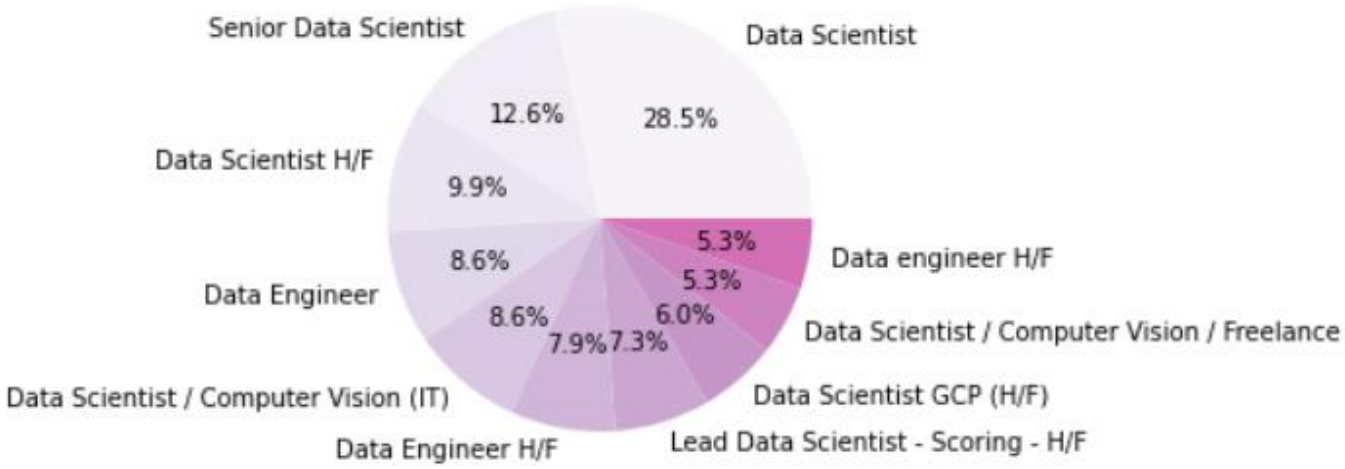




Nous voyons que **Paris** compte près de 65% du Top 10 des villes avec le plus grand nombre d'offres d'emploi en France, le suivant est **Lyon** avec 5,7% et Coubevoir 5,2%. Cela se prouve facilement car Paris est la capitale de la France, qui compte le plus grand nombre d'entreprises en France, donc la demande d'emplois dans le domaine de Data science est ici très élevée.

C) Titre

Nous allons répondre à la question laquelle des postes de Data Science est la plus populaire, en exportant un graphique du Top 10 des emplois les plus populaires dans le domaine Data science.



Titre	
Data Scientist	43
Senior Data Scientist	19
Data Scientist H/F	15
Data Scientist / Computer Vision (IT)	13
Data Engineer	13
Data Engineer H/F	12
Lead Data Scientist - Scoring - H/F	11
Data Scientist GCP (H/F)	9
Machine Learning Engineer	8
Data engineer H/F	8

Nous allons clarifier un peu la notion de métiers clés dans le domaine de la Data en général et de Data science en particulier, ils comprennent : Data scientist, Data Analyst, Data Engineer, Machine Learning Engineer, Big Data Consultant**.**

Sur les graphique ci-dessus, nous voyons que dans le domaine de la Data science en France, le poste de **Data Scientist** sera le plus embauche. Il est proportionnel au total : 28,5% + 12,6% + 9,9% + 8,6% + 7,3% + 6% + 5,35% = 72,9%. Pour le reste, **Data Engineer** et **Machine Learning Engineer** ne représente que près de 20% du nombre d'offres d'emploi dans le top 10.

Pour expliquer cela, Data scientist : connu pour avoir été élu “métier le plus sexy du 21ème siècle” par la Harvard Business Review, le métier de data scientist a également été récompensé comme emploi le plus prometteur en 2020 par la resource LinkedIn. Ce spécialiste de la donnée est toujours très recherché par les entreprises en 2021.

D) Description

Dans cette partie, nous répondrons à la question: Quelles sont les exigences pour d’offre emploi dans le domaine Data science?. (*Logiciel programmation, Tool, Diplôme, Nombre moyen d'années d'expérience, d'autres compétences,…etc*).

Comme défini dans la section précédente, Nous combinons les titres des emplois du domaine Data sceince en 4 titres principaux comme suit: **Data Scientist, Data Analyst, Big Data Engineer (Big Data Consultant), et Machine Learning Engineer**. Cette étape, nous explorerons les exigences pour ces 4 titres, entre autres.

Les étapes de ce travail sont les suivantes:

D’abord, nous supprimerons les mots vides (stopword) dans les descriptions de l’ofrre d’emploi (Définition : *Les stopwords sont les mots en anglais qui n'ajoutent pas beaucoup de sens à une phrase. Ils peuvent être ignorés en toute sécurité sans sacrifier le sens de la phrase. Par exemple, les mots comme the, he, have, je, il, le, la, les, avoir... etc. De tels mots sont déjà saisis ceci dans le corpus nommé corpus. Nous le téléchargeons d'abord dans notre environnement python*).

Nous avons aussi des packages à appliquer à la langue française. Dans ce projet, nous utilisons à la fois l'anglais et le français car les offres d'emploi sont en anglais ou en français

Cette étape, je vais convertir et remplacer des chaînes de l'ensemble de données dans le type de données approprié pour l'analyse et supprimer de certains mots pour plus de faciliter l'analyser plus tard.

Après le traitement sur Python, nous obtenons un dataframe:

	Titre	Description
0	Consultant BI (H/F)	Je découvre les filiales \n\nRejoindre Davids...
1	Application Engineer - Machine Learning	GrAI Matter Labs utilizes brain-inspired, neur...
2	Lead Data Scientist	Notre offre\n\nTESSELLA est le World Class Cen...
3	Data Scientist	Data Scientist (H/F) \n[CDI] \n\nNew trends a...
4	Lead Data Scientist - Scoring - H/F	Le poste \n\nPython / Sql / Tableau / Agile ...

Ensuite, nous sélectionnons les 4 titres de poste mentionnés ci-dessus pour analyser. Compte tenu de la nature du traitement du langage naturel (NLP) des ensembles de données, les méthodes d'exploration similaires seront appliquées à chaque ensemble de données pour dévoiler leurs caractéristiques distinctes. Cela inclura l'utilisation de visualisations telles que :

I) Nuages de mots (Wordcloud)

Les nuages de mots peuvent identifier des tendances et des modèles qui seraient autrement peu clairs ou difficiles à voir sous forme de tableau. Les mots-clés fréquemment utilisés ressortent mieux dans un nuage de mots. Les mots courants qui pourraient être négligés sous forme de tableau sont mis en évidence dans un texte plus gros, ce qui les fait ressortir lorsqu'ils sont affichés dans un nuage de mots.

II) N-Gram (Unigramme, Bigramme et Trigramme)

Un n-gramme est une séquence contiguë de n éléments d'un échantillon donné de texte ou de parole. Différentes définitions des n-grammes permettront d'identifier les mots/phrases les plus répandus dans les données d'apprentissage et ainsi aider à distinguer ce qui comprend des questions non sincères et sincères.

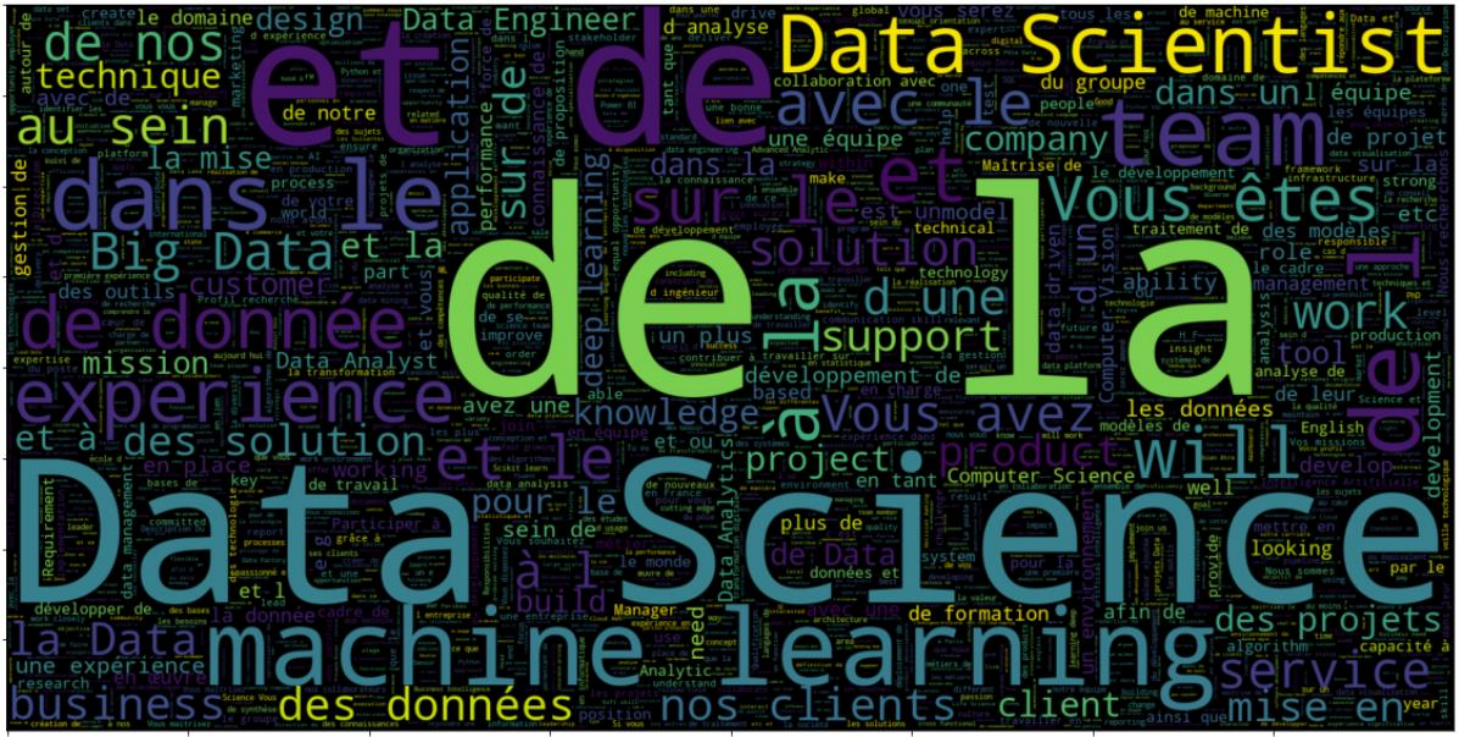
Il convient de noter qu'avant d'afficher des mots ou des phrases individuels, le texte sera d'abord marqué (basé sur un entier souhaité), puis placé dans un cadre de données qui sera utilisé pour construire des tracés côte à côte. La tokenisation est, en général, une première étape du processus de NLP, une étape qui divise des chaînes de texte plus longues en morceaux plus petits, ou jetons. De plus gros morceaux de texte peuvent être segmentés en phrases, les phrases peuvent être segmentées en mots, ...etc.

Après avoir sélectionné les descriptions des titres, nous convertissons les éléments de la liste en minuscules, supprimons les liens html de la liste, supprimons les caractères spéciaux restants et convertissons en dataframe.

- Nous exécutons la fonction sur les 4 titres de l'emploi: Data Scientist, Data Analyst, Big Data Consultant, et Machine Learning Engineer et nous supprimons les valeurs manquants pour une visualisation plus claire.

I) Wordcloud de description

Nous voyons que, dans le Wordcloud ci-dessous, les descriptions de poste de Data scientist, Machine Learning, Big Data, Data, Client, Computer Science, Data Engineer, Projet, Tool, Travail en équipe, Knowleage (Connaissance), …etc sont courantes dans le Wordcloud. La prochaine étape consiste à explorer spécifiquement à travers chaque poste dans le domaine Data science.

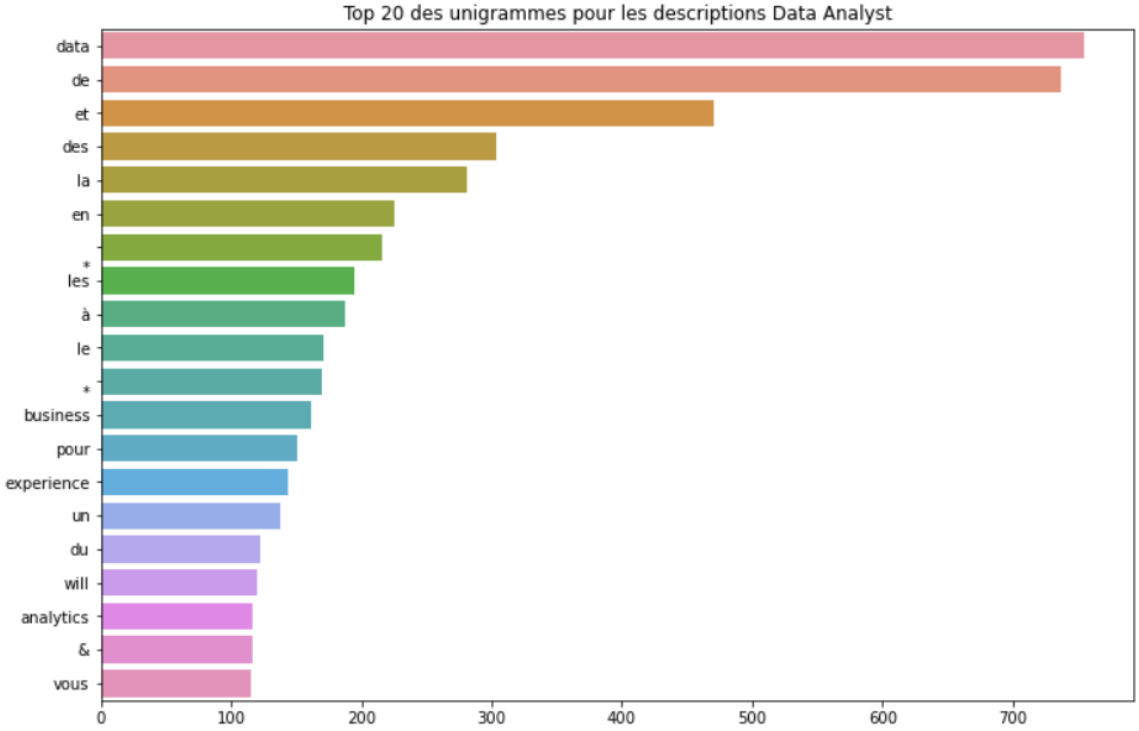


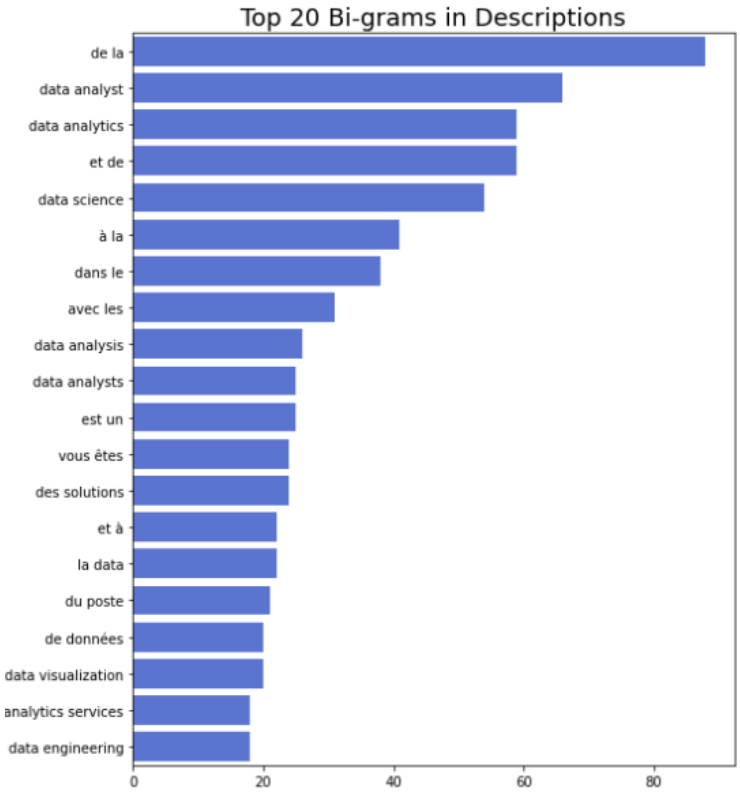
3) N-gram Analysis

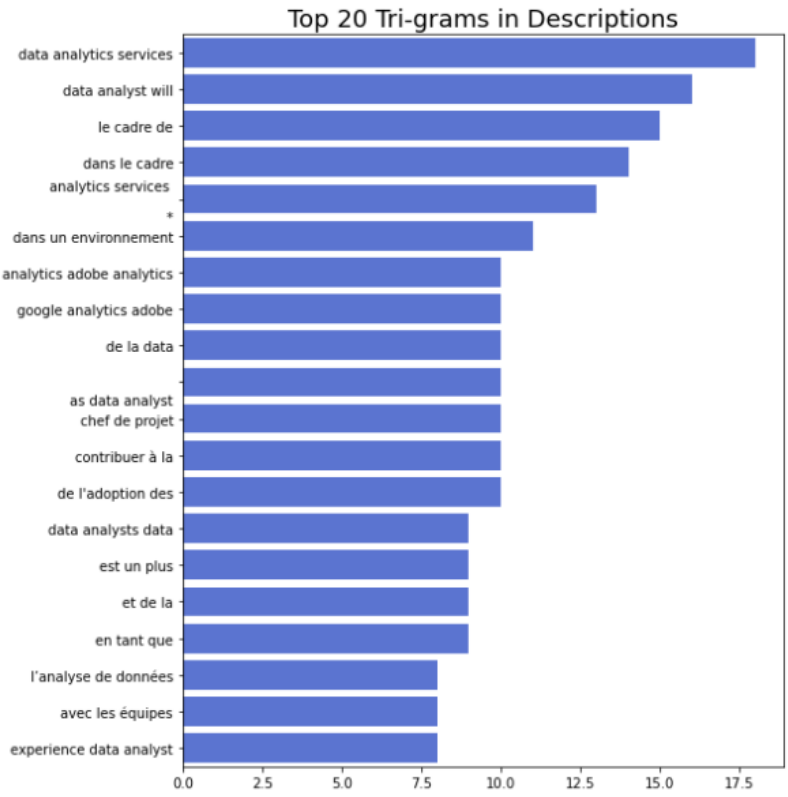
Dans cette partie, nous analyserons les N-grammes de 4 titres Data Scientist, Data Analyst, Big Data Consultant, et Machine Learning Engineer pour trouver les exigences mentionées ci-dessus telles que: les logiciels de programmation, les outils spéciaux dans le domaine Data science, les diplômes, les années d'expérience, ...etc et leurs genres les plus populaires

A) Analyse de N-gramme par Data Analyst

Top 20 des unigrammes pour les descriptions Data Analyst



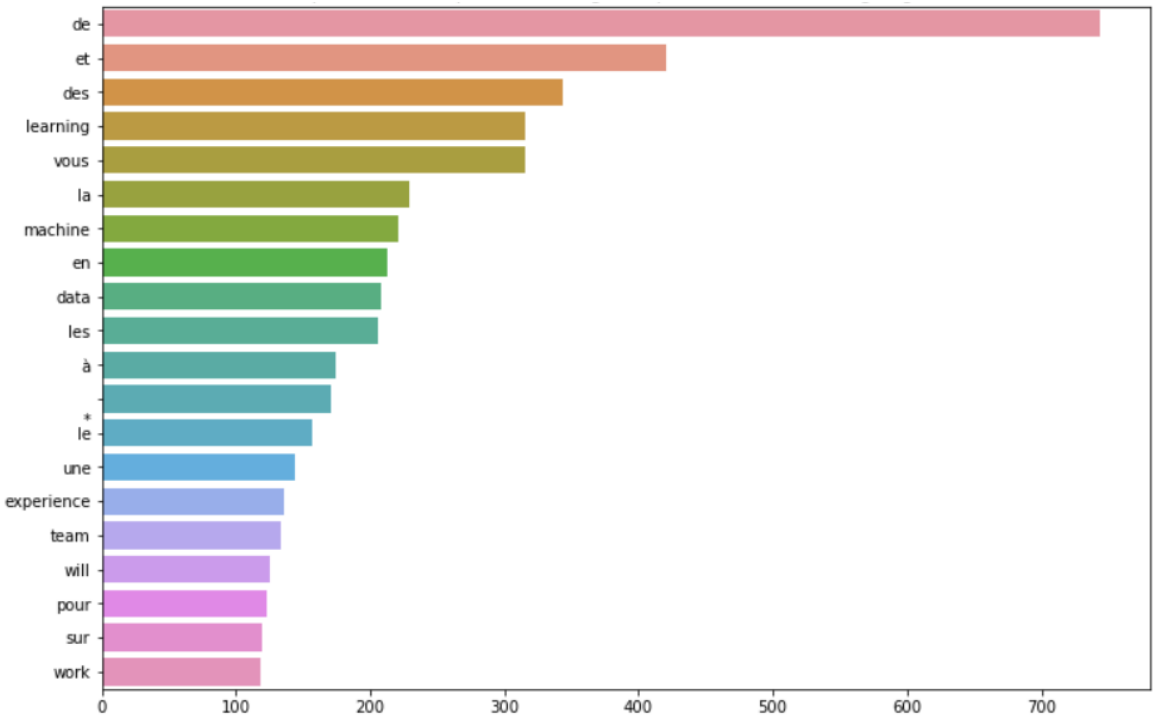


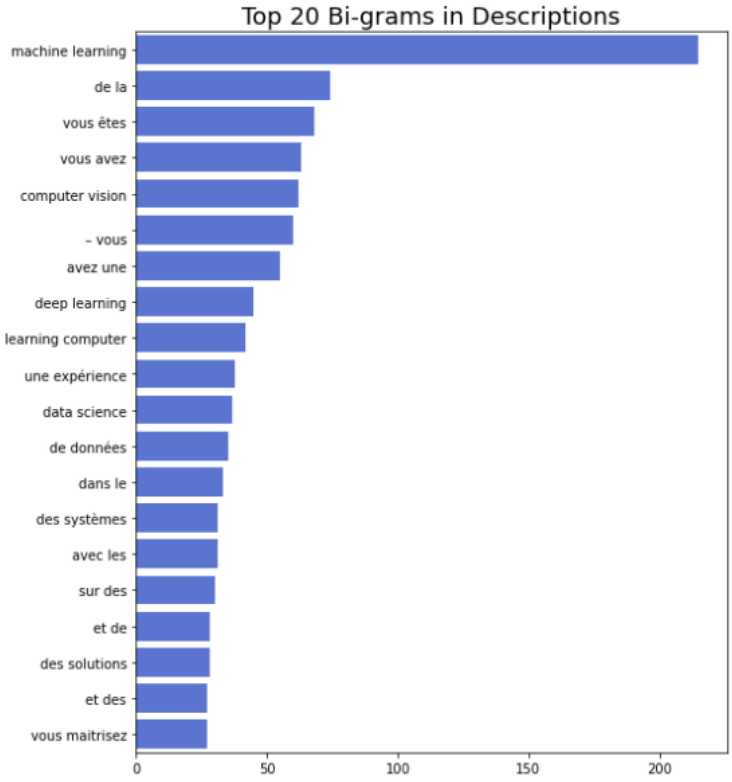


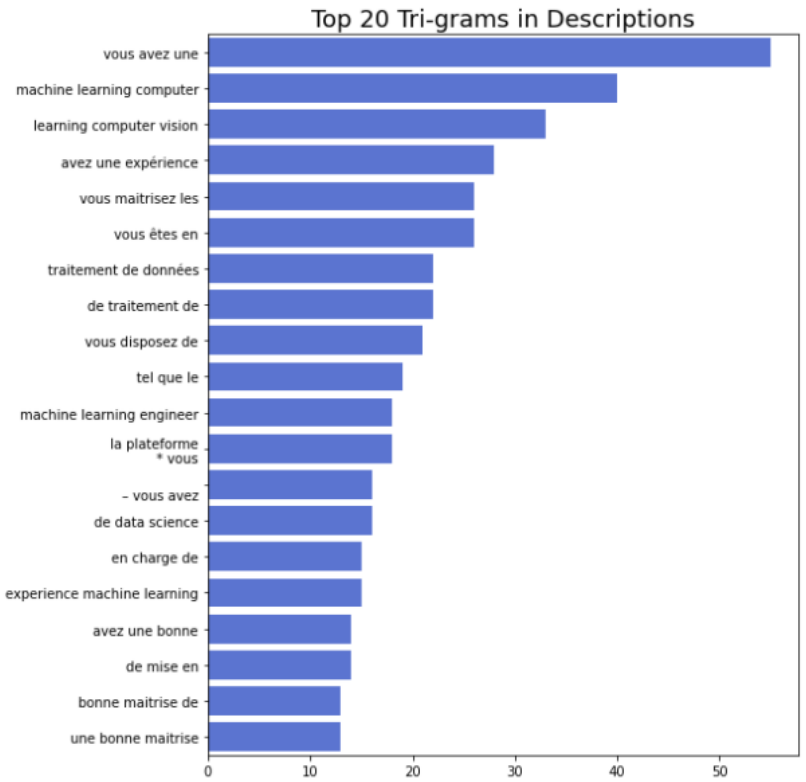
L'analyse n-gram pour le titre “Data Analyst” indique en outre que le rôle peut être très général, avec des responsabilités allant de l'analyse des données, des affaires, du service d'analyse analytique, de Google Analytics, d'Adobe, chef de projet, travail d'équipe. Nous pouvons déduire de ce qui précède que les analystes de données (sur la base de l'échantillon de données) ont généralement travailler dans un environnement commercial, et il y a une grande exigence d'expérience dans l'analyse de données, la visualisation de données, et une compétence en google analytics et en adobe et des compétences en travail d'équipe sont requises. En plus de la position de Data Analyst, ils embauchent également des chefs de projet dans ce domaine.

B) Analyse de N-gramme de Machine leaning Engineer

Top 20 des unigrammes pour les descriptions Data Scientist



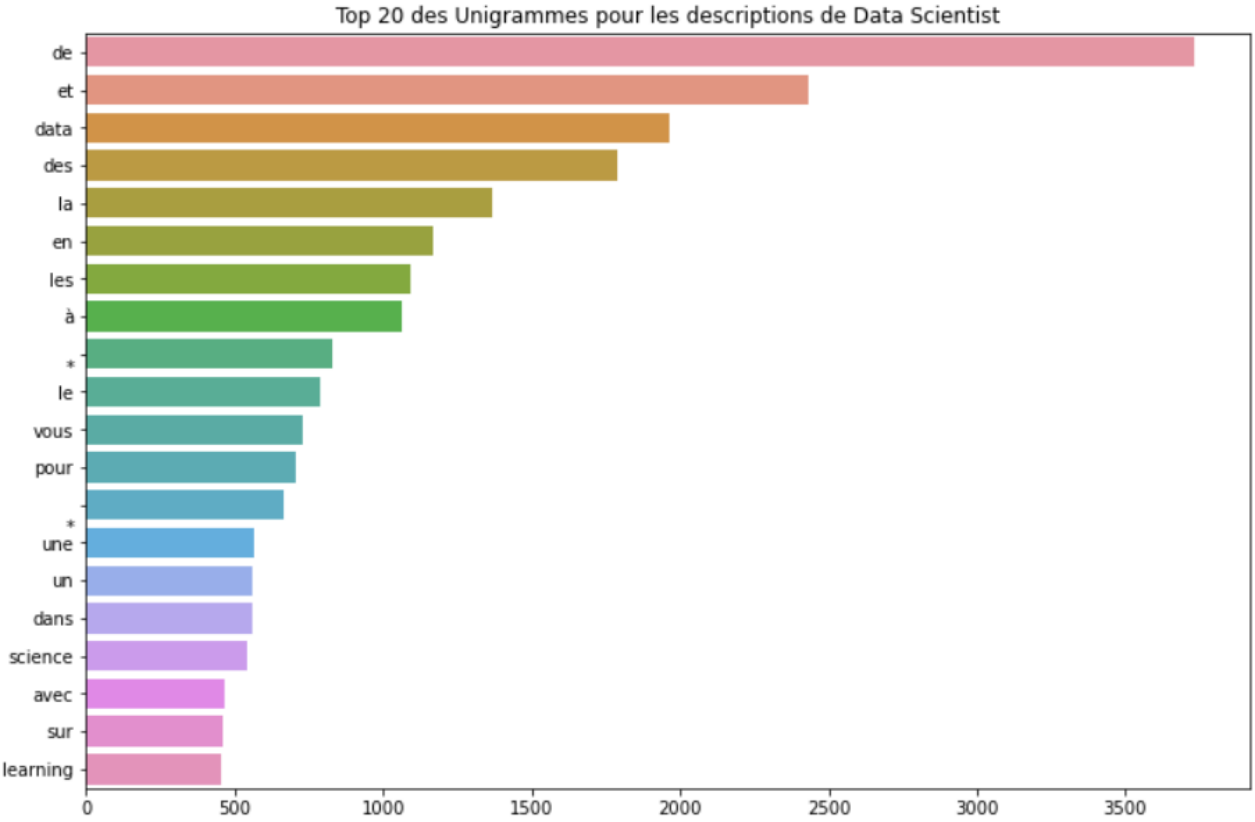


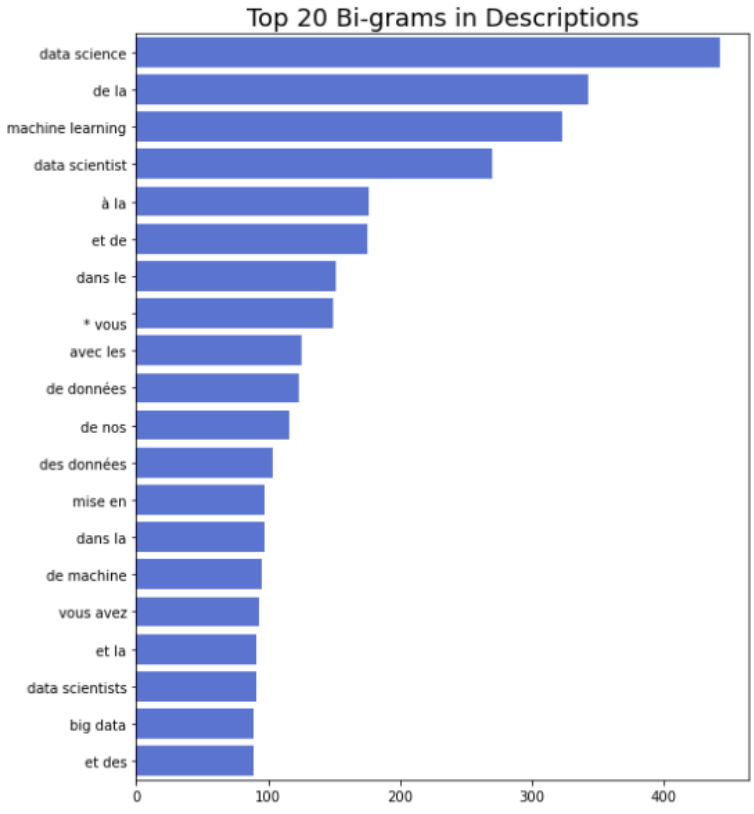


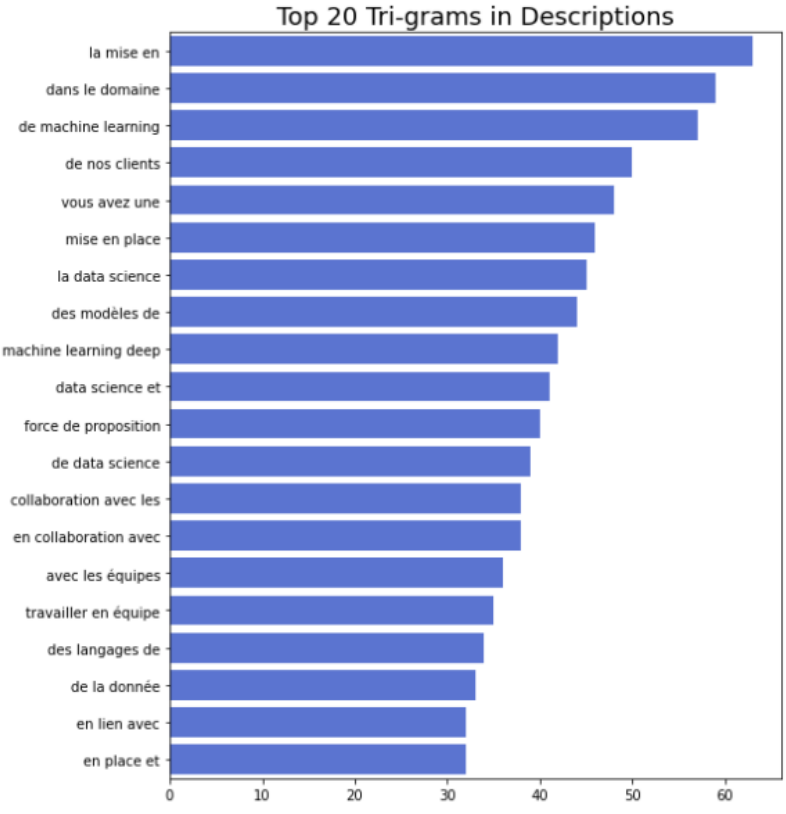
L'analyse N-gramme pour les postes de Machine Learning Engineer révèle qu'il s'agit bien d'un rôle lié à l'ingénierie, avec un diplôme en computer science étant généralement requis. De plus, il s'avère également qu'il s'agit d'un rôle plus spécifique que son homologue “Data Scientist” où des termes tels que l'apprentissage en profondeur (deep learning) et les systèmes. En outre, ils nécessitent également des compétences en traitement de données et maîtrise des logiciel.

C) Analyse de N-gramme de Data Scientist

Top 20 des unigrammes pour les descriptions Data Scientist



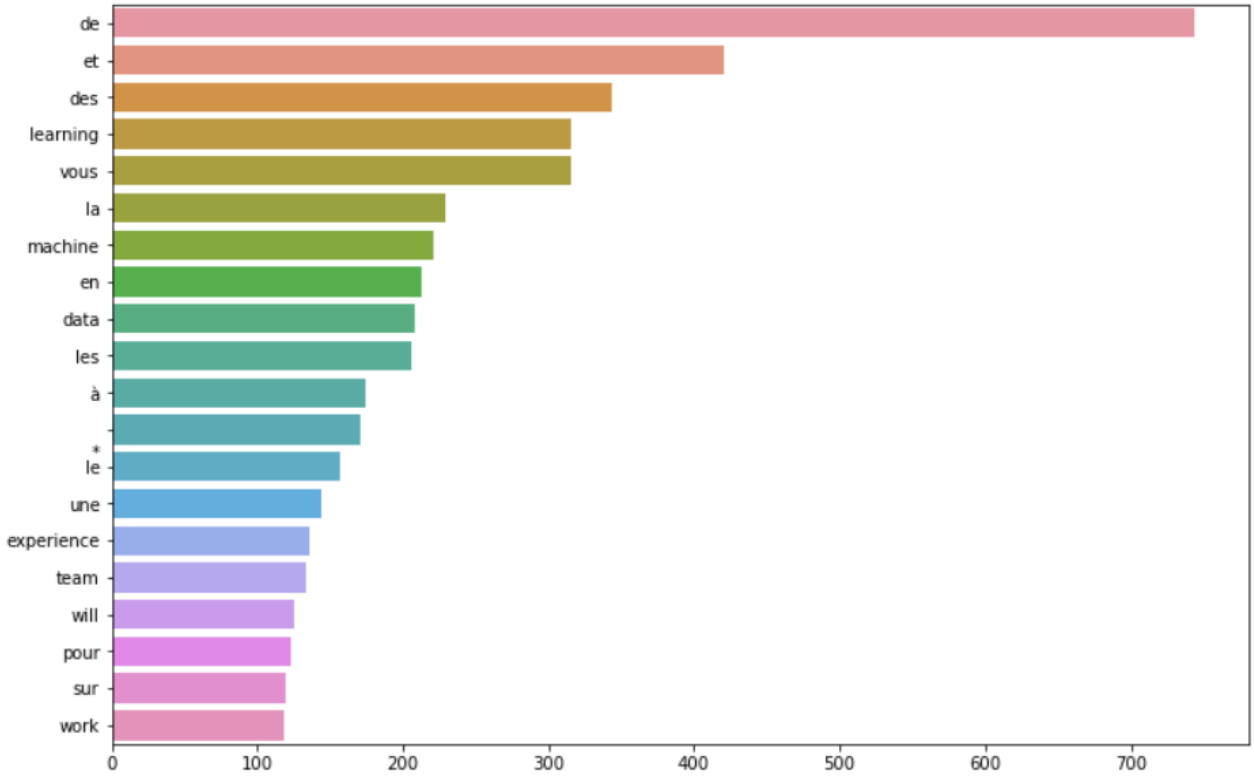


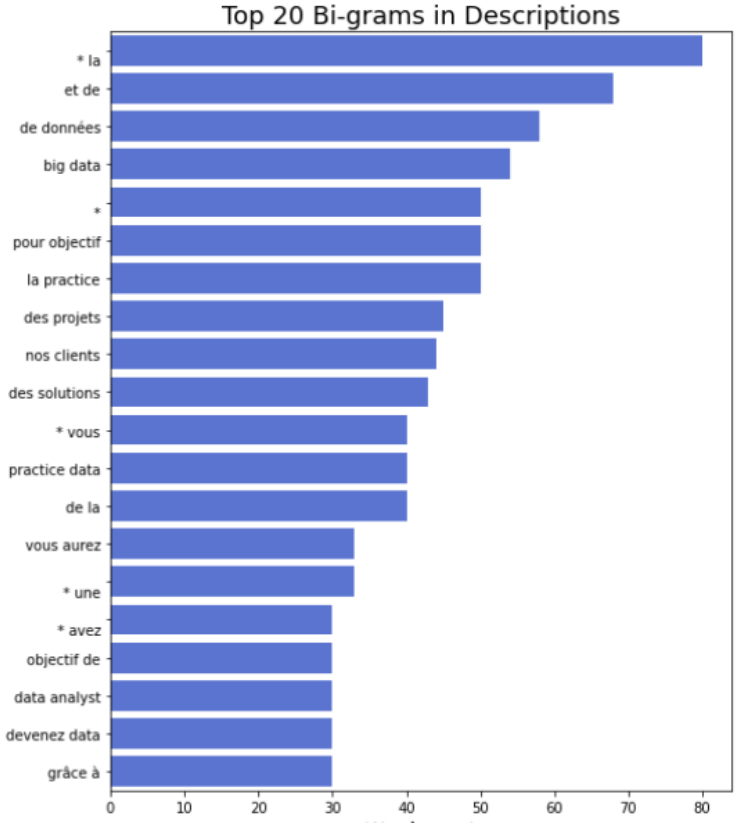


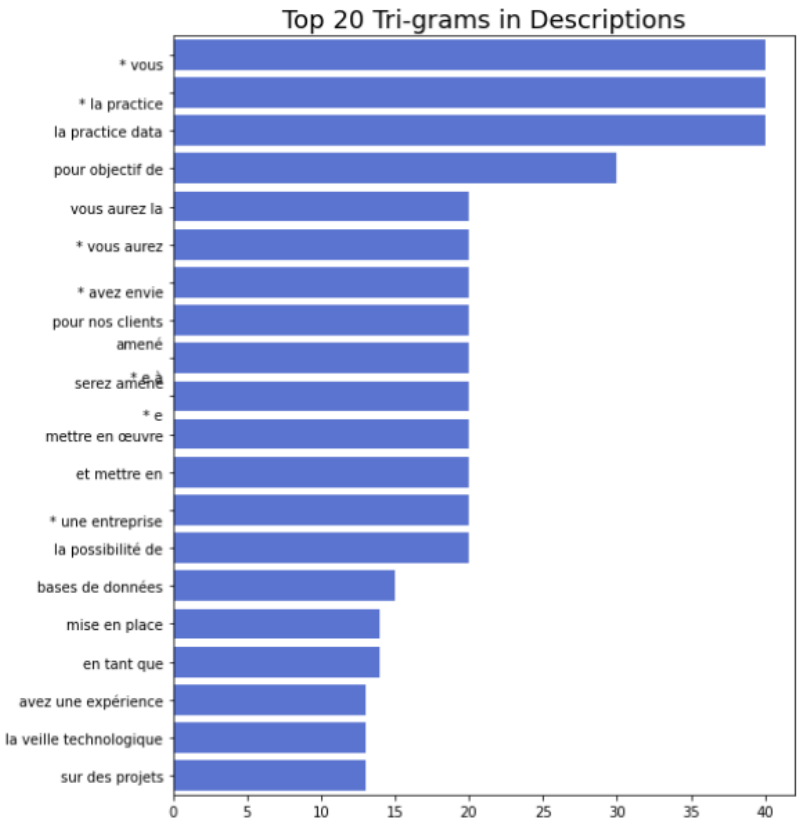
L'analyse N-gramme pour le poste de "Data Scientist" semble en outre que le rôle peut se concentrer sur le data science s et les modèles de machine learning, le deep learning, avec les compétences requises en langage de programmation, force de proposition, travail d'équipe, pour analyser des mégadonnées (Big Dat)a et travailler avec leurs clients.

D) Analyse de N-gramme de Big Data Consultant

Top 20 des unigrammes pour les descriptions Big Data Consultant







L'analyse N-gramme pour les rôles Big Data Consultant semble confirmer qu'il s'agit de rôles d'ingénierie qui reposent sur l'utilisation de technologies de Big Data et Machine Learning. Ils semblent nécessiter une formation liée à l'informatique, technologie et semblent assumer certaines des responsabilités des "Data Scientists", vise à exécuter et conseiller les projets de leurs clients en travaillant avec des données, l'analyse de données, ...etc.

En résumé, cette étape, cette analyse indique que pour les positions ci-dessous, les éléments suivants semblent être retenus:

Postes de Data Analyst - Formation de niveau supérieur pour un poste de niveau d'entrée (par rapport aux autres postes analysés dans ce noyau) qui nécessite des connaissances en science des données, big data, analytique et apprentissage automatique.

Postes de Data Scientist - Axé sur la satisfaction des besoins de l'entreprise et la direction des équipes pour répondre à ces derniers. Il est généralement nécessaire d'utiliser des techniques statistiques et des modèles d'apprentissage automatique pour analyser de grands ensembles de données. Semblable au poste "Data Analyst", le "Data Scientist" doit utiliser des compétences dans les domaines de l'exploration de données, du big data, de l'analyse et de l'apprentissage automatique.

Postes de Machine Learning Engineer - Rôle axé sur l'ingénierie, un diplôme en informatique étant généralement requis. Il semble être un rôle plus spécifique que son homologue "Data Scientist" où des termes tels que l'apprentissage en profondeur, le développement de logiciels, le traitement du langage et les systèmes sont utilisés.

Postes Big Data Consultant - Rôles d'ingénierie qui reposent sur l'utilisation des technologies Big Data et de l'analyse. Ils semblent nécessiter une formation liée à l'informatique et semblent assumer certaines des responsabilités des "Data Scientists".

E) Les mots-clés les plus populaires dans les offres d'emploi en Data Science

Dans cette étape, nous trouverons la réponse à la question, quelles sont les exigences les plus demandées dans les offres d'emploi, pour les 4 types de titres spécifiés? Par exemple : logiciel de programmation, sur les outils spéciaux, quelles qualifications sont les plus demandées dans les offres d'emploi, ..etc

Ici, j'ai écrit des instructions def pour trouver des mots-clés

- pour les langages de programmation dans la description, par exemple : Python, C++, MATLAB, .NET, C#, JavaScript, Java, Scala, SQL, ...etc

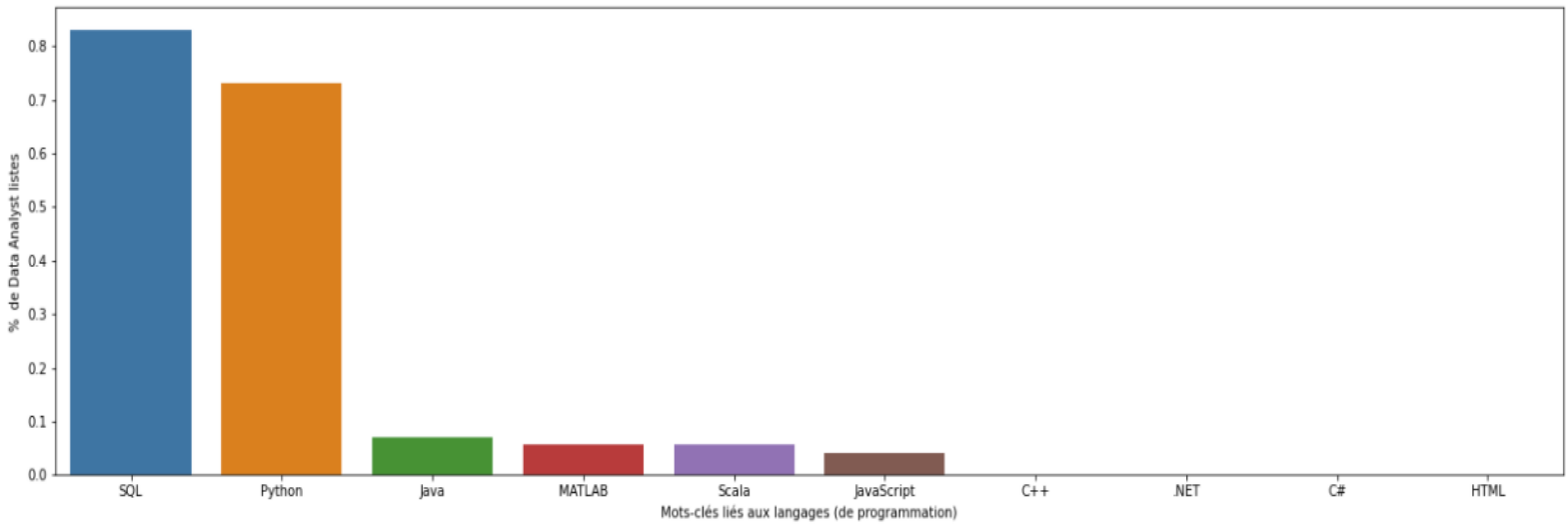
- pour des mots-clés de des logiciels spécifiques dans le domaine Data Science tels que : Hadoop, Spark, Impala, Cassandra, Kafka, HDFS, HBase, Hive, Kubernetes, Kubeflow, Airflow, BigQuery, ou

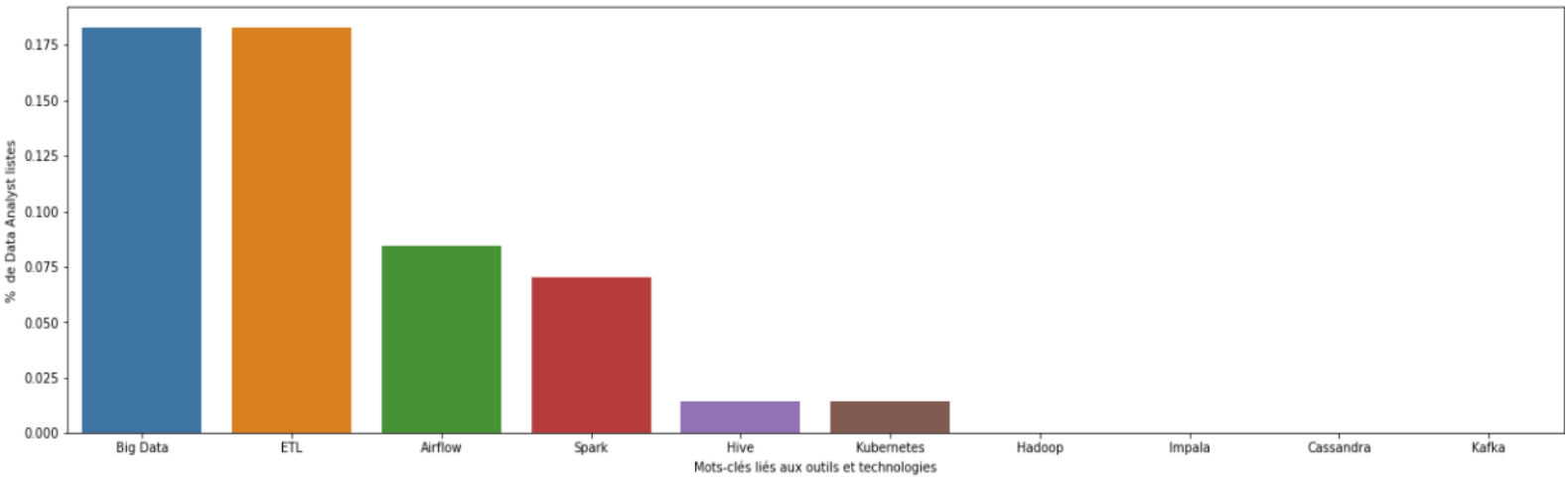
- pour les clouds et les stockages tels que : AWS, GCP, Azure, Google Cloud, S3 , Redshift, EC2, Lambda, Route S3, Dynamo DB,...etc et enfin

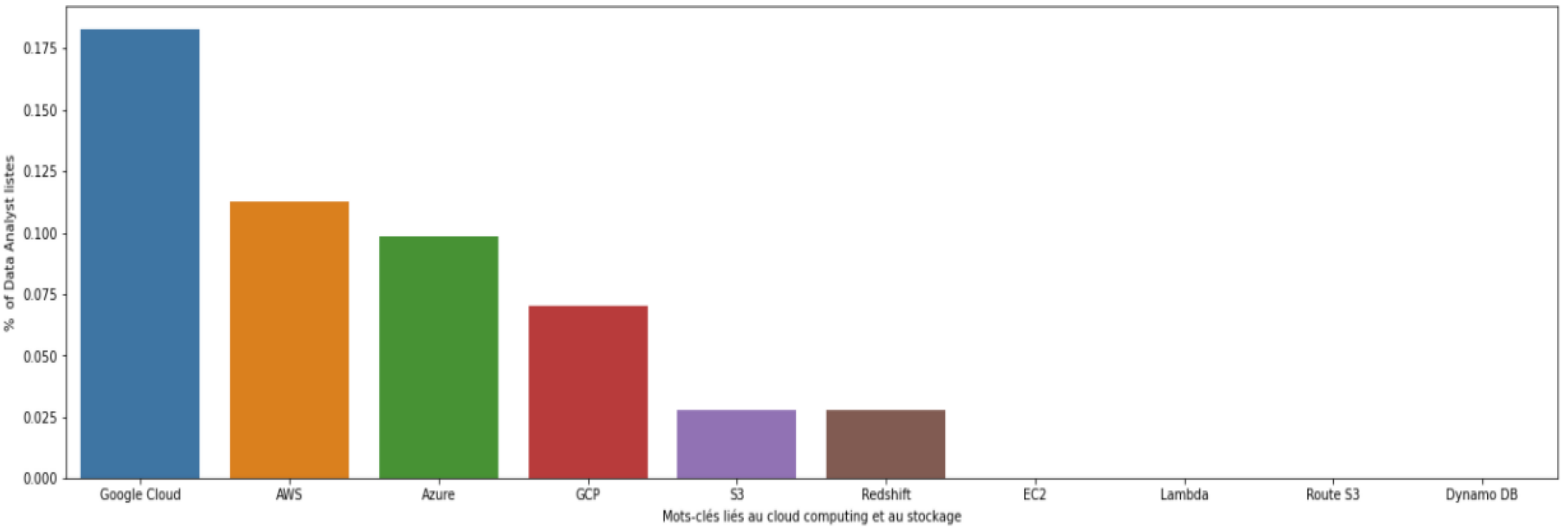
- Pour les diplômes requises et les types de travail tels que : BSc, MSc, PhD, Stage, Junior, Senior, "Master", Doctorate, Bachelor, Post-Doc, Temps plein, Stage, Remote,…etc.

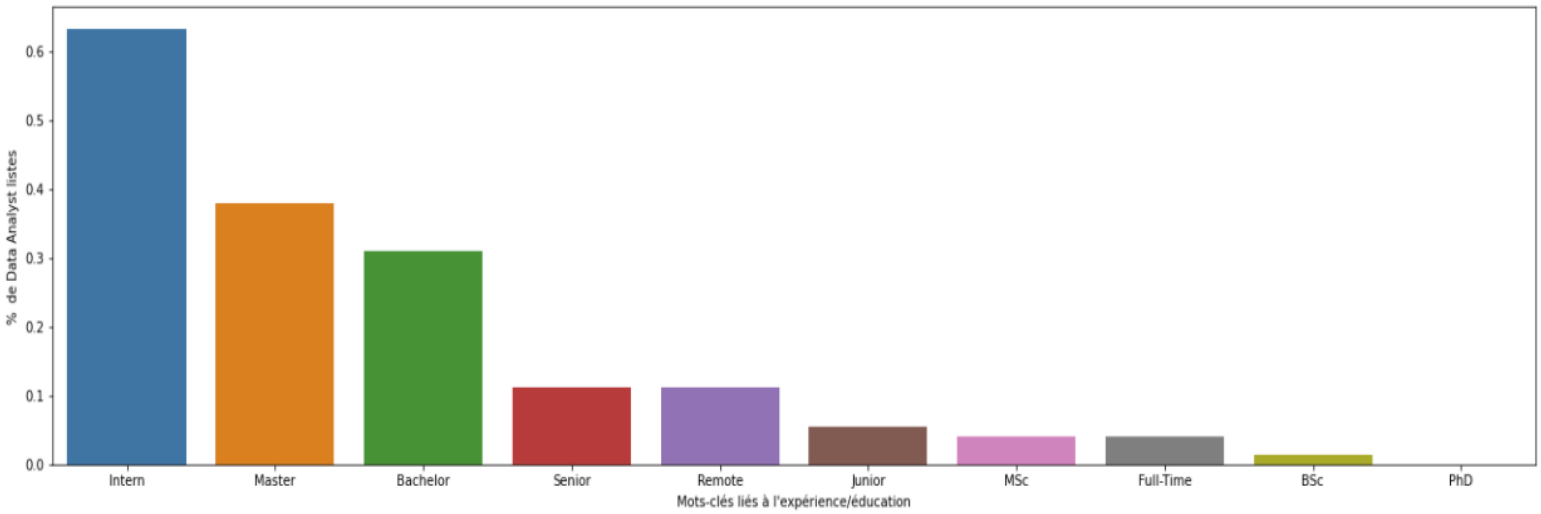
Nous allons ensuite sortir des graphiques à barres pour visualiser ces informations.

***) Data Analyst**


% de Data Analyst listes
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
SQL
Python
Java
MATLAB
Scala
JavaScript
C++
.NET
C#
HTML
Mots-clés liés aux langages (de programmation)

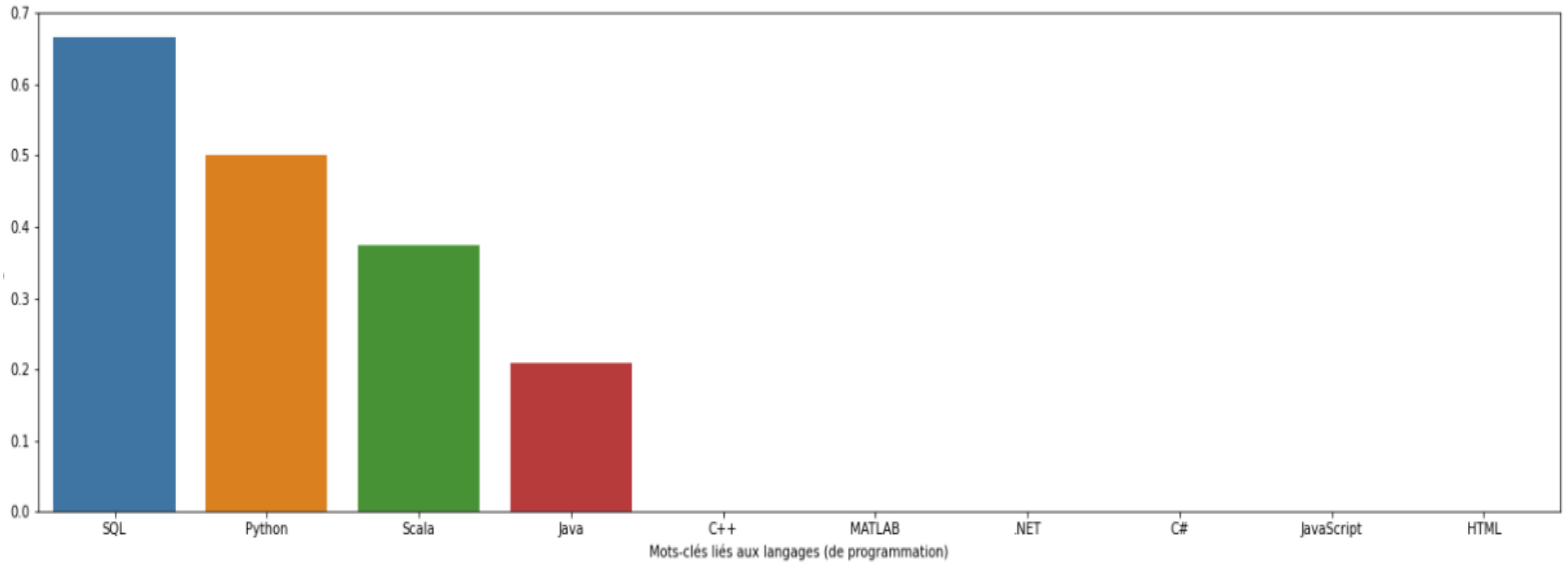

% de Data Analyst listes
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Big Data
ETL
Airflow
Spark
Hive
Kubernetes
Hadoop
Impala
Cassandra
Kafka
Mots-clés liés aux outils et technologies

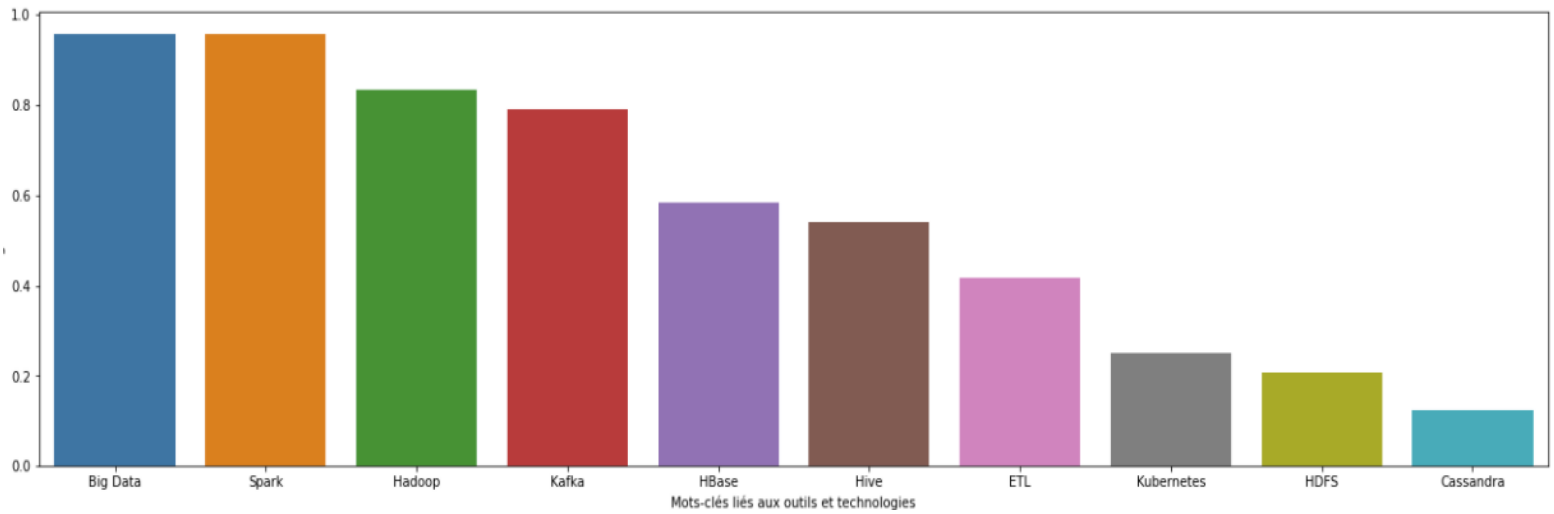

% of Data Analyst listes
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Google Cloud
AWS
Azure
GCP
S3
Redshift
EC2
Lambda
Route S3
Dynamo DB
Mots-clés liés au cloud computing et au stockage

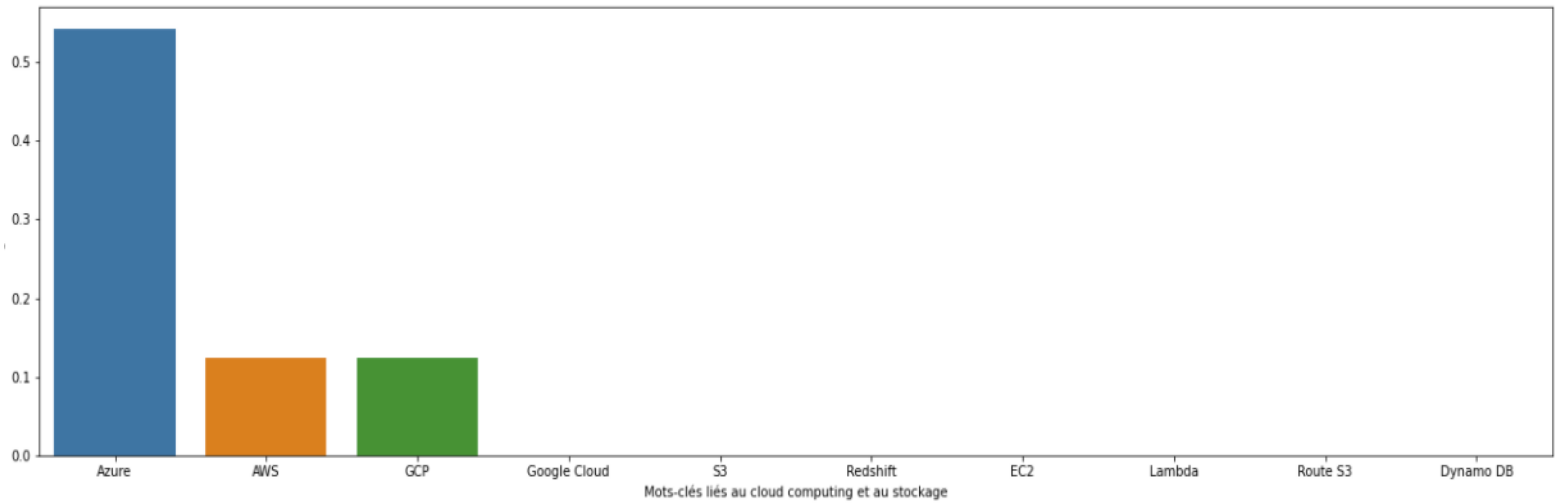


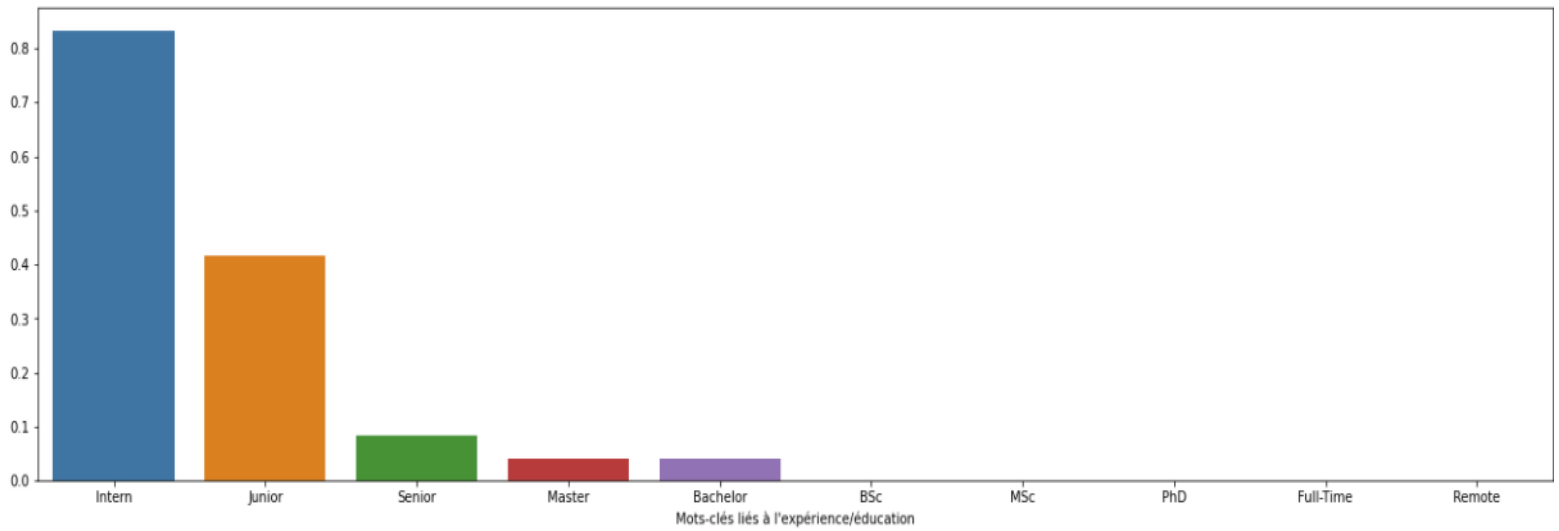
Les exigences de base d'un profil "Data Anlyst" sont : Logiciel (SQL, Python), l'outils et les technologies (Big Data, ETL, Airflow, Spark), les connaissance du cloud et du stockage (Google Cloud, AWS, Azure, GCP), diplôme (Master, Bachelor), type de travail (Stage, Senior ou travail à distance).

***) Big Data Consultant**



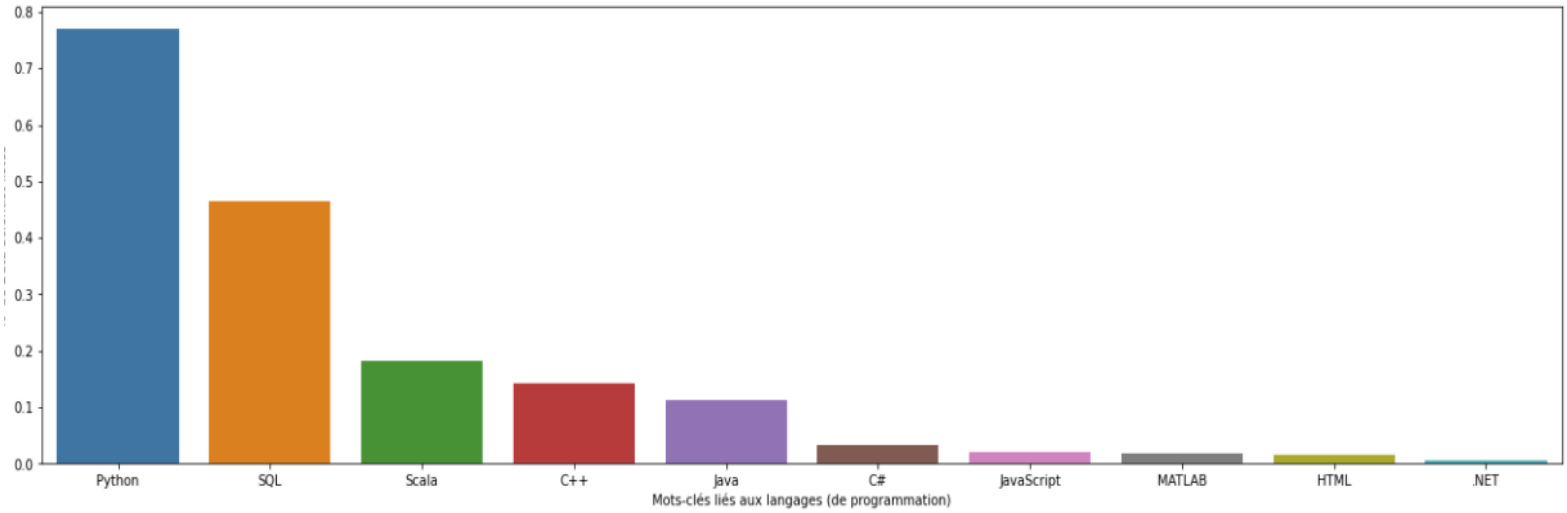


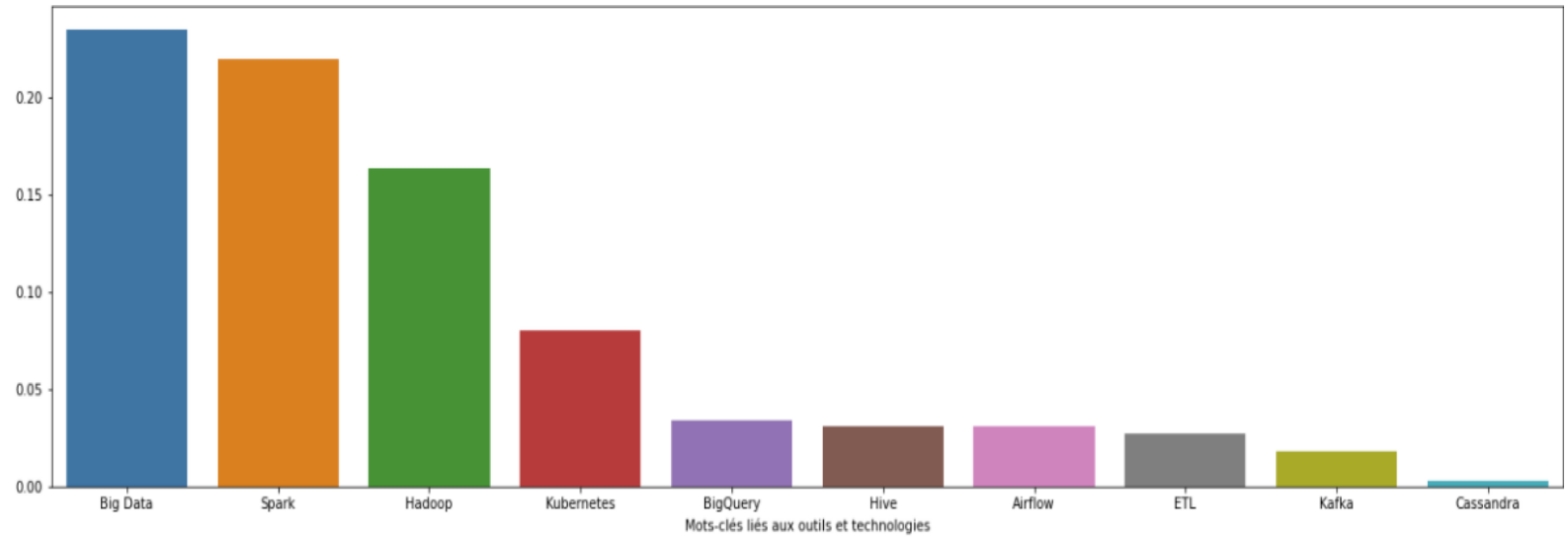


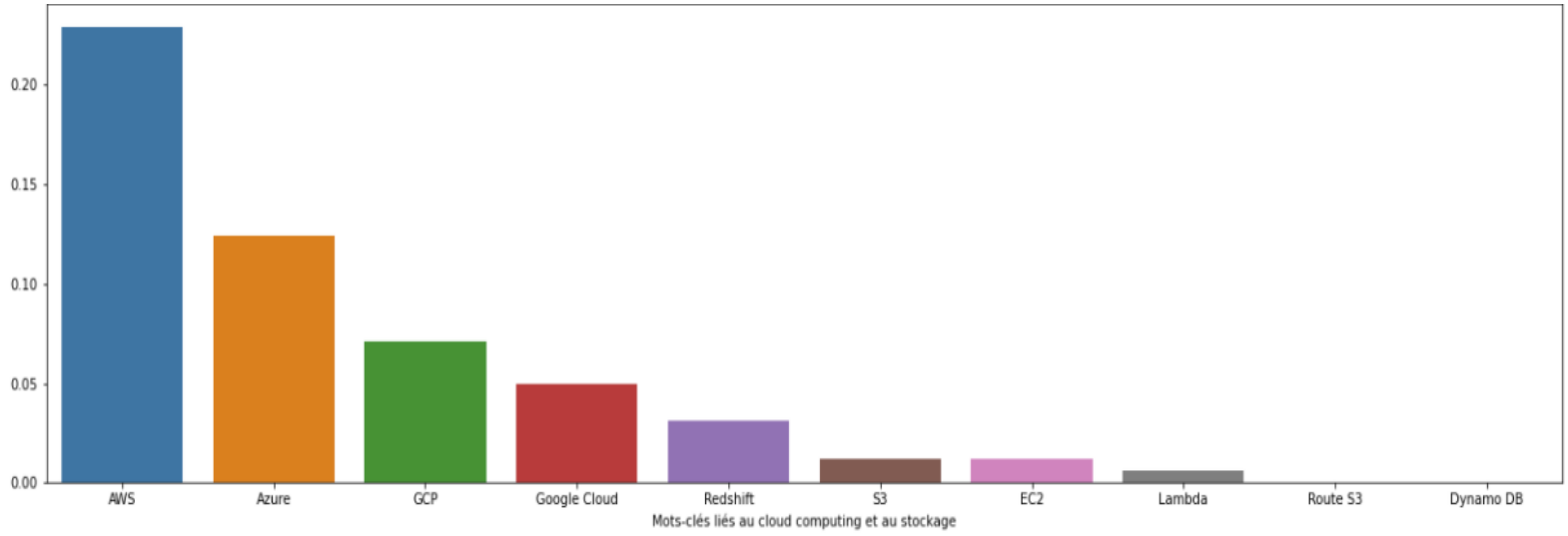


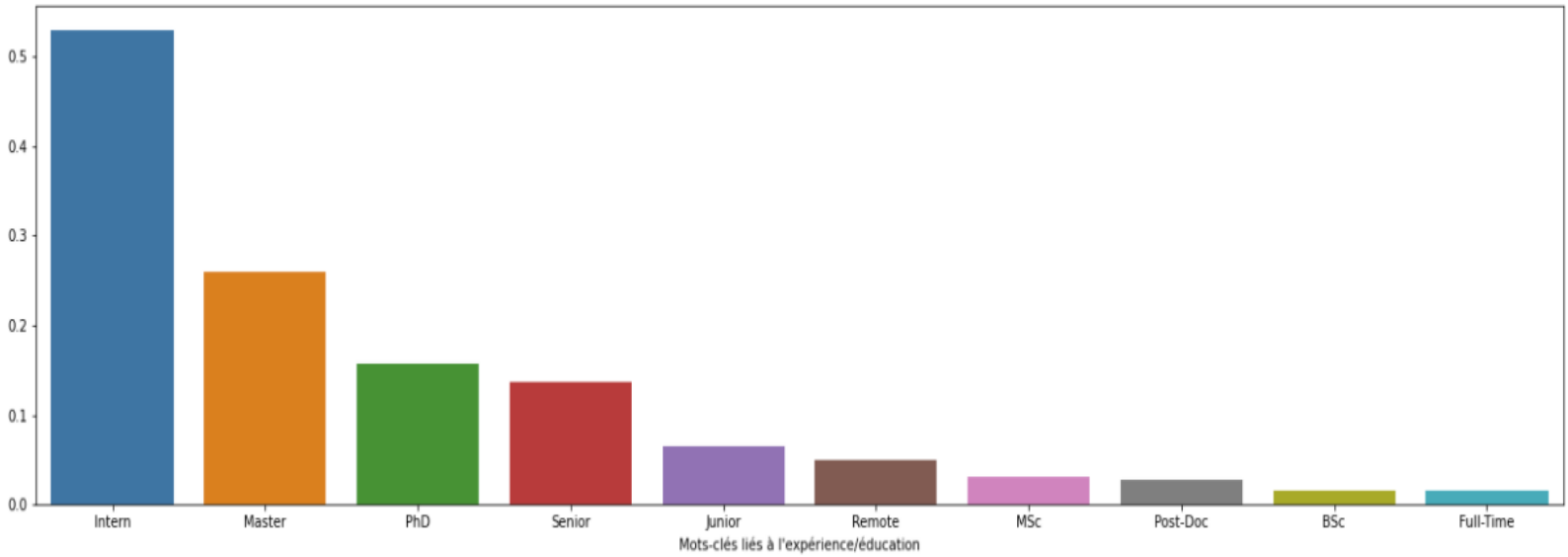
Les exigences de base d'un profil "Big Data Consultant" sont : Logiciel (SQL, Python, Scala, Java), l'outils et les technologies (Big Data, Spark, Handoop, Kafka, Hbase, Hive, ETL,), les connaissance du cloud et du stockage (Azure, AWS, GCP), diplôme (Master, Bachelor), type de travail (Stage, Junior, Senior).

***) Data scientist**



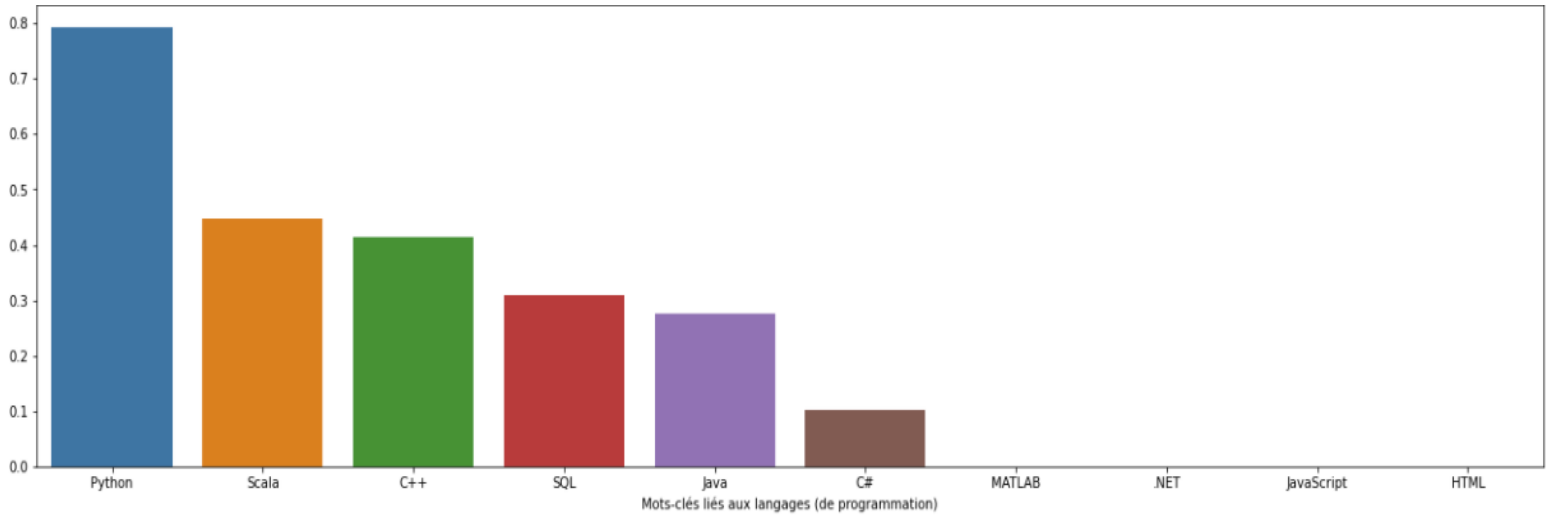


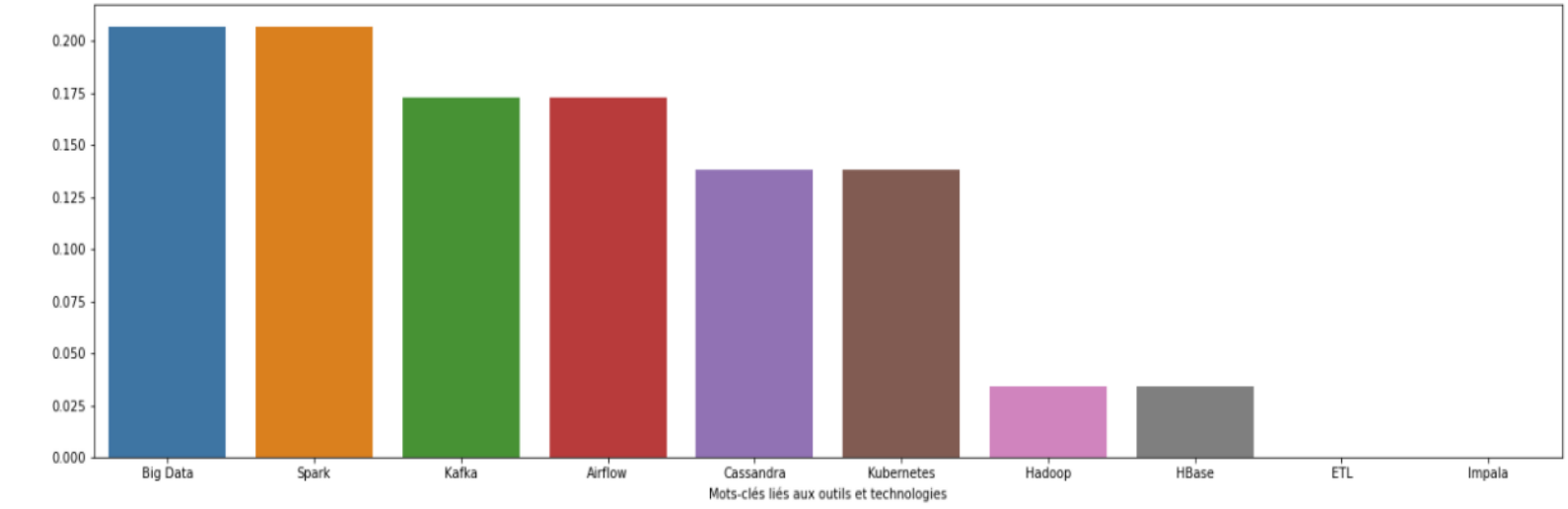


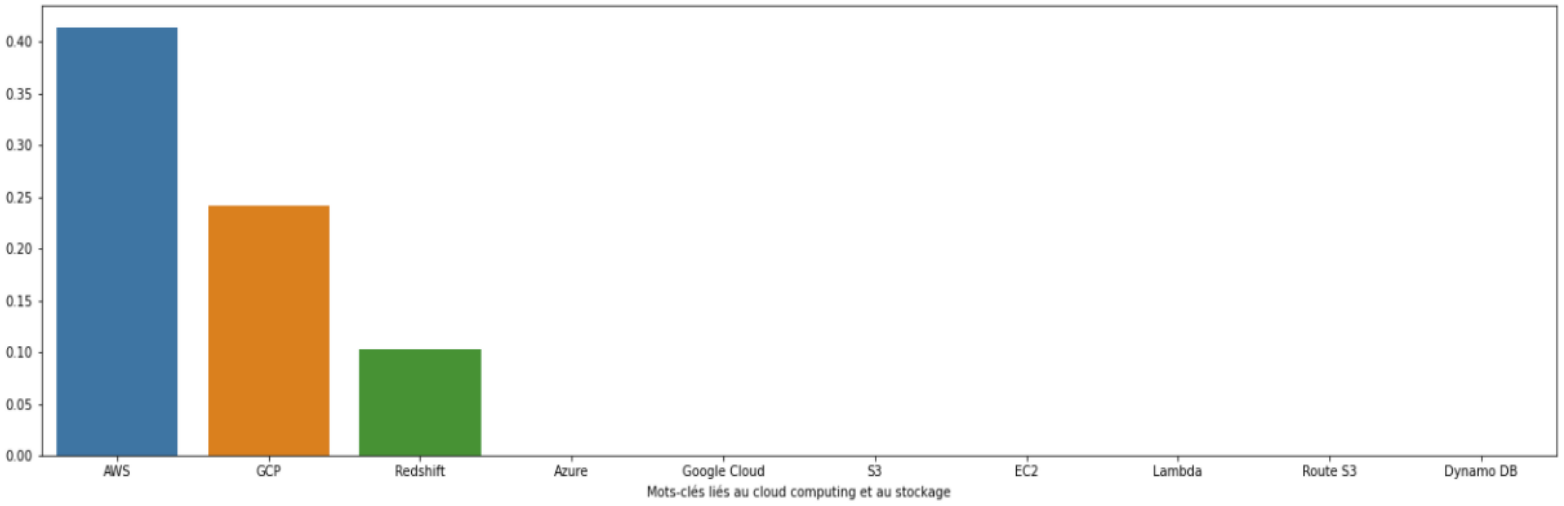


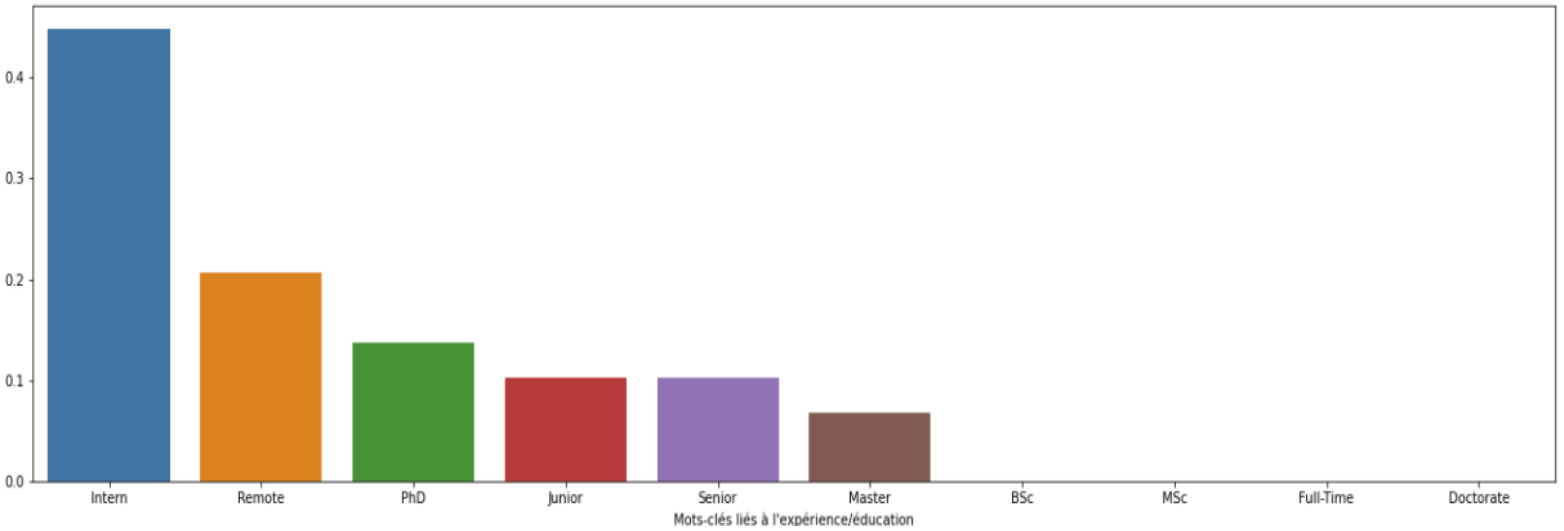
Les exigences de base d'un profil "Data scientist" sont : Logiciel (Python, SQL, Scala, C++, Java), l'outils et les technologies (Big Data, Spark, Handoop, Kubemetes, Bigquery, Hive, Airflow, ETL), les connaissance du cloud et du stockage (AWS, Azure, GCP, Google Cloud, RedShit), diplôme (Master, PhD, Msc, Post-doc), type de travail (Stage, Senior, Junior ou travail à distance).

*) Machine Learning Engineer



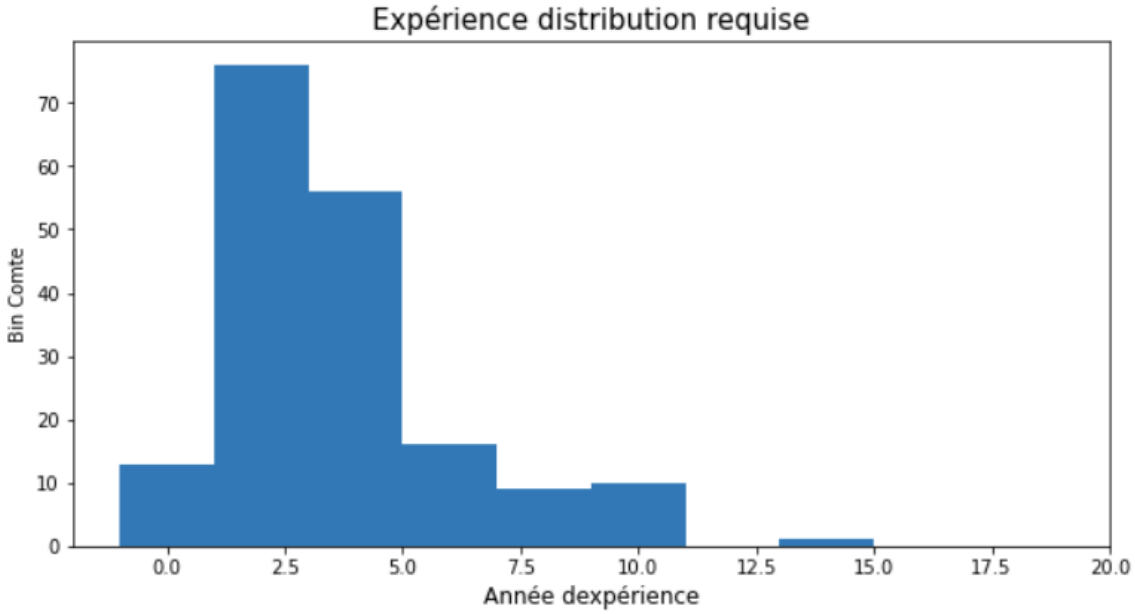






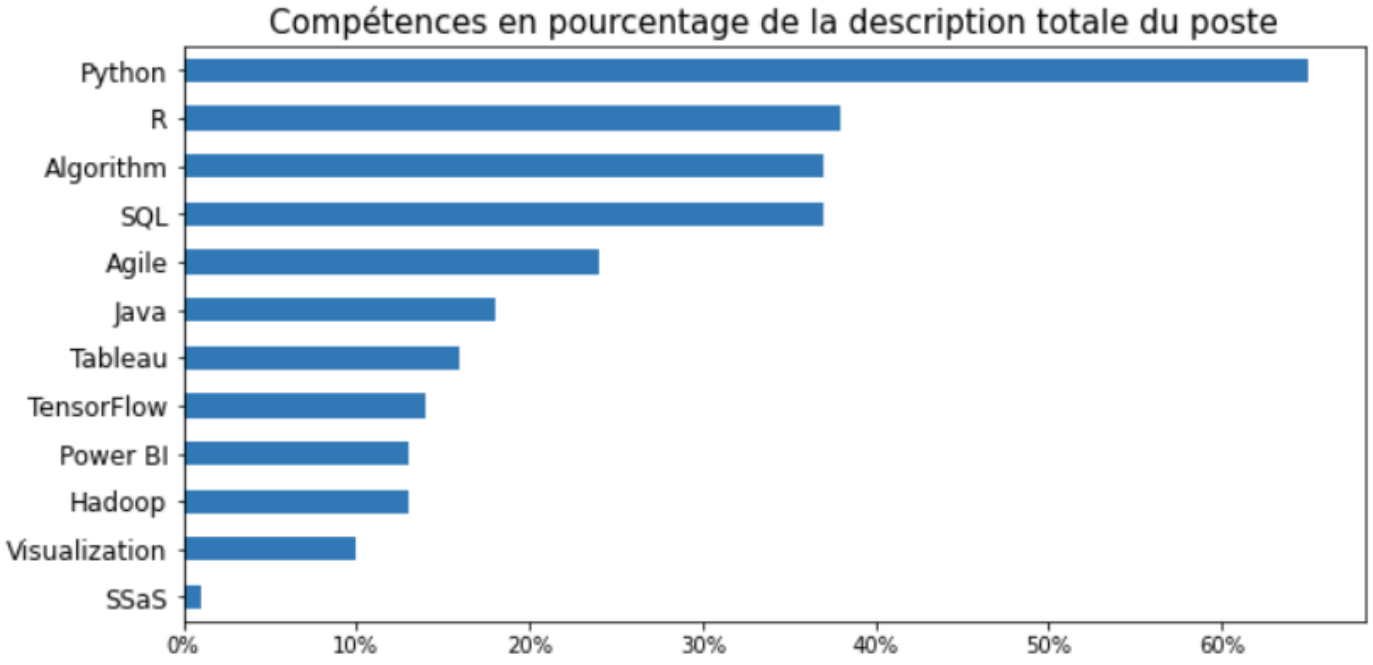
Les exigences de base d'un profil "Data scientist" sont : Logiciel (Python, Scala, C++, SQL, Java, C#), l'outils et les technologies (Big Data, Spark, Kafka, Airflow Cassandra, Kubernetes, Handoop, Hbase), les connaissance du cloud et du stockage (AWS, GCP, RedShit), diplôme (PhD, Master) type de travail (Stage, Senior).

F) Nombre moyen d'années d'expérience requis



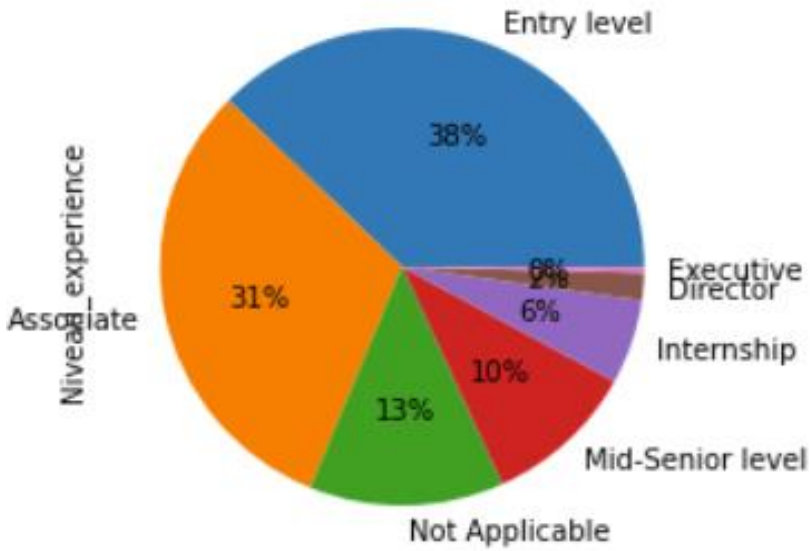
Nombre moyen d'années d'expérience requis est 4.5 ans dans les description des offres d'emploi.

G) Compétences en pourcentage de la description totale du poste



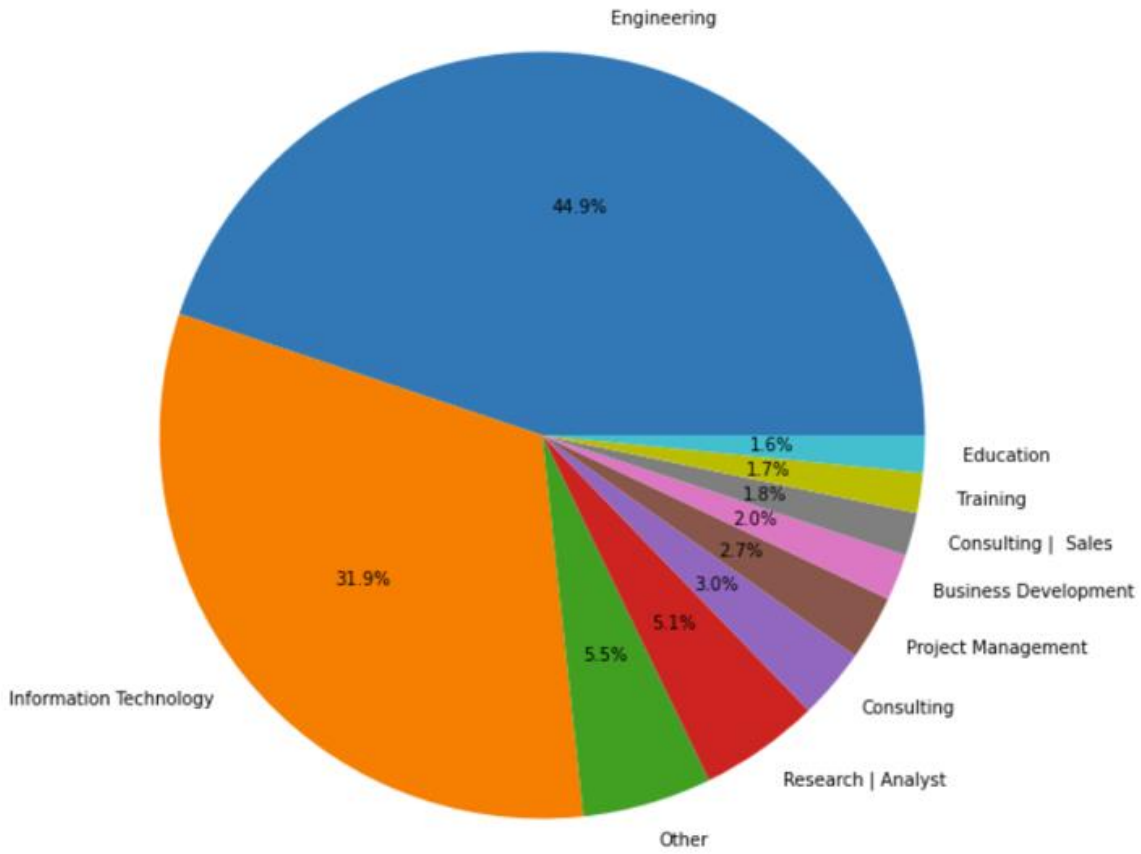
Les connaissances Python, R, SQL, Agile, Java, Tableau et les connaissance d'Algothime sont les compétences les plus demandées dans les descriptions d'offre d'emploi dans le domaine Data science en France.

H) Niveau d'expérience de l'offre d'emploi



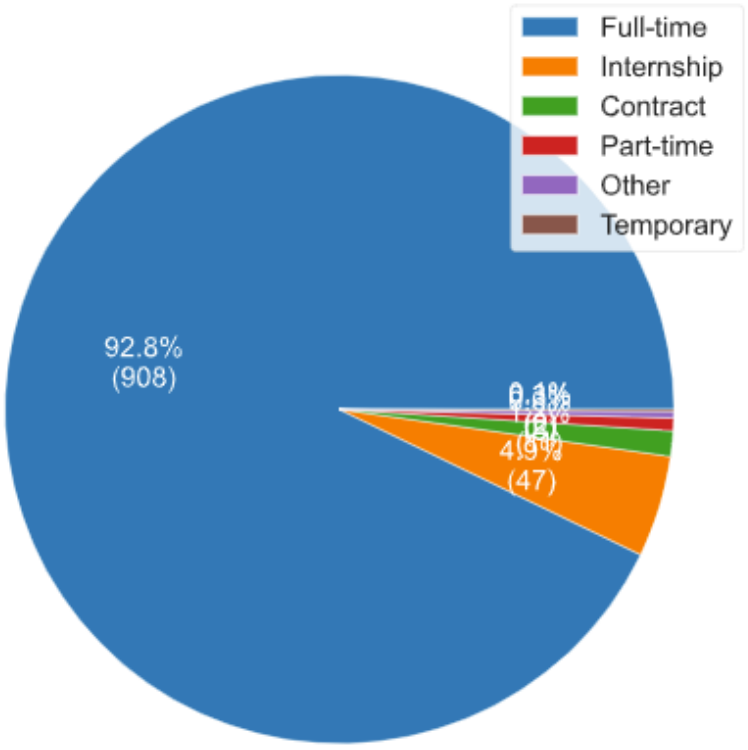
Nous analysons la colonne "Niveau d'expérience" et produisons un diagramme circulaire comme ci-dessus. Nous voyons que le niveau d'entrée représente la plus grande proportion: 38%, suivie par le niveau de l'associé représentant 31%. Le niveau exécutif n'a presque pas d'offres d'emploi (0%). Par ailleurs, le niveau du directeur et niveau de stage ne représentent qu'un faible pourcentage avec 2% et 6%, le reste étant de niveau de mi-senior et le niveau sans-applicables.

I) Catégorie



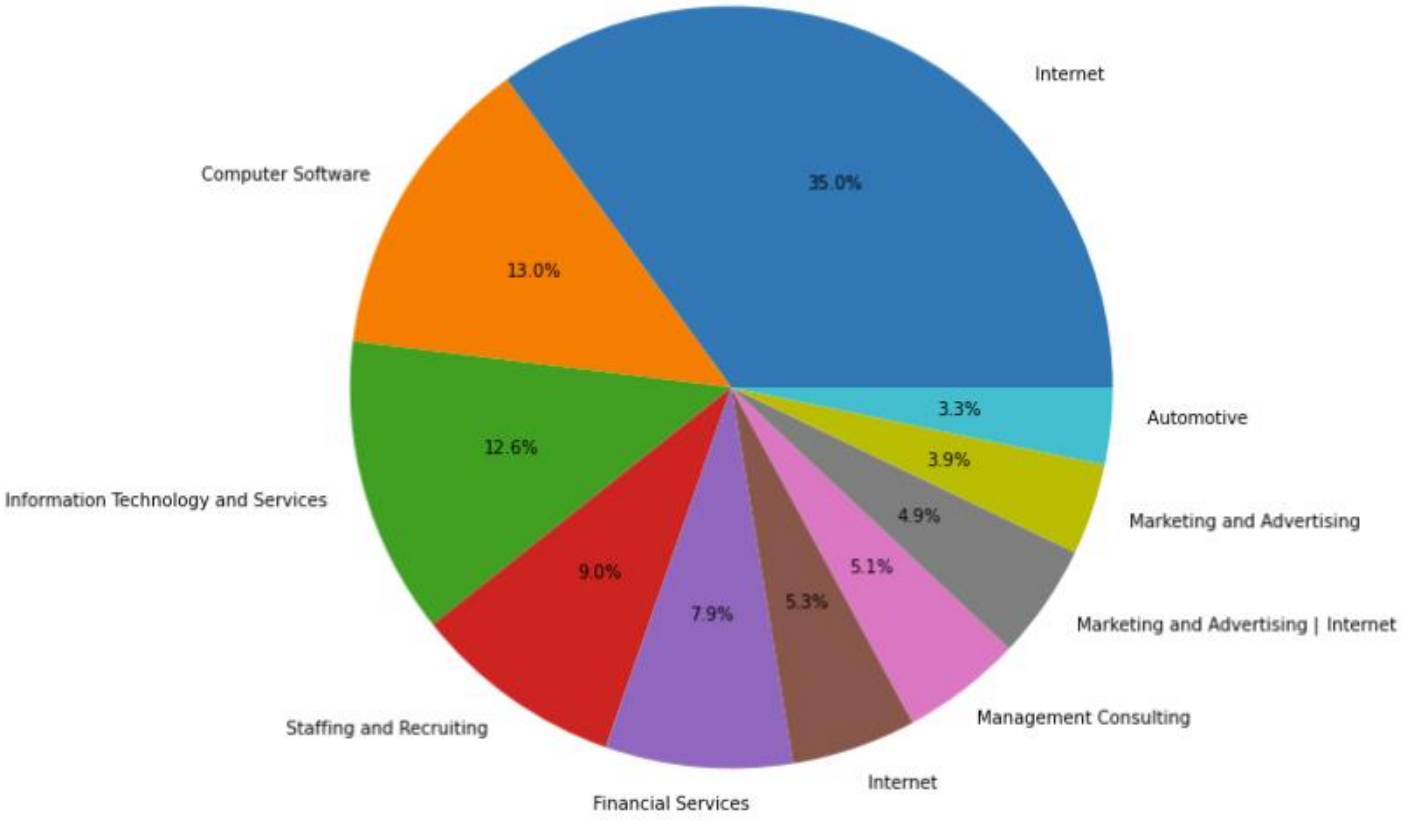
Les catégories Ingénieure sont les plus présente dans les offres d'emploi dans le domaine Data science en France avec un taux d'environ 45%, la deuxième catégorie Technologie de l'information étant la catégorie information avec un taux d'environ 32%,

K) Type de travail



Nous voyons que le type d'emploi de temps plein est majoritaire environ 93%, ensuite est le type de stage environ 5% de nombre des offres d'emploi dans le domaine Data science en France

L) Industrie



Nous voyons que dans l'industrie de l'emploi, l'industrie Internet le plus grand taux de recrutement avec 35% du nombre d'offres d'emploi dans le domaine de la Data Science en France, suivi est de domaine de Computer Software et Technologies de l'information et Services avec le taux de 13 % et 12,6 %. Des domaines comme: Autonomie, Marketing and Advertising, Management Consulting ne représentent qu'un faible pourcentag.

PARTIE III: RÉSULTAT ET CONCLUSION

Ce projet m'a fourni de précieuses informations sur ce que comment extraire les données des réseaux sociaux d'offres d'emploi comme LinkedIn, ou cela peut également s'appliquer à d'autres sites d'offres d'emploi comme Indeen, Monster, Lesjeudis,...etc et comment préparer le prétraitement des données, ainsi que comment appliquer la méthode EDA, la méthode analyse textuelle, N-gram, et les méthodes Machine learni et et quelques autres méthodes utiles (vous pouvez les voir dans mon script Python)

D'autre part, dans le processus de mise en œuvre, j'ai eu beaucoup de difficultés à gérer des données qui ne sont pas parfaites, il y a beaucoup de "poubelles" dans cet ensemble de données et le problème le plus difficile est de les filtrer et ensuite, comment choisir et créer les bons graphiques pour donner les résultats les plus claire pour les analyser et les prédire aux problèmes et aux questions posées.

Il y a quelques points que je souhaite améliorer, en particulier dans la partie analyse de texte. J'espère en savoir plus sur la meilleure façon d'utiliser les WordCloud afin d'avoir une meilleure précision sur mes résultats.

Nous obtenons des réponses aux questions posées:

-En effet, il y a 312 entreprises différentes qui affichent des offres d'emploi dans le domaine Data science sur LinkedIn en France.

-L'entreprise Atos à Bezons est l'entreprise avec le plus grand nombre d'offres d'emploi dans le domaine Data science en France avec 12 offres. Les deuxième rang sont des sociétés UNLCK, Onfido en Ile-de-France avec 11 offres.

-Paris compte près de 65% du Top 10 des villes avec le plus grand nombre d'offres d'emploi en France, le suivant est Lyon avec 5,7% et Coubevoir 5,2%. Cela se prouve facilement car Paris est la capitale de la France, qui compte le plus grand nombre d'entreprises en France.

-Dans le domaine de la Data science en France, le titre de Data Scientist sera le plus embauche.

Nous avons spécifié 4 titres de base dans le domaine de la science des données : Data Scientist, Data Analyst, Big Data Consultant, et Machine Learning Engineer, plus précisément, après de l'danalyse N-gram, nous avons les résultats comme suit:

- Postes de Data Analyst - Formation de niveau supérieur pour un poste de niveau d'entrée (par rapport aux autres postes analysés dans ce noyau) qui nécessite des connaissances en science des données, big data, analytique et apprentissage automatique.

- Postes de Data Scientist - Axé sur la satisfaction des besoins de l'entreprise et la direction des équipes pour répondre à ces derniers. Il est généralement nécessaire d'utiliser des techniques statistiques et des modèles d'apprentissage automatique pour analyser de grands ensembles de données. Semblable au poste "Data Analyst", le "Data Scientist" doit utiliser des compétences dans les domaines de l'exploration de données, du big data, de l'analyse et de l'apprentissage automatique.

- Postes de Machine Learning Engineer - Rôle axé sur l'ingénierie, un diplôme en informatique étant généralement requis. Il semble être un rôle plus spécifique que son homologue "Data Scientist" où des termes tels que l'apprentissage en profondeur, le développement de logiciels, le traitement du langage et les systèmes sont utilisés.

- Postes Big Data Consultant - Rôles d'ingénierie qui reposent sur l'utilisation des technologies Big Data et de l'analyse. Ils semblent nécessiter une formation liée à l'informatique et semblent assumer certaines des responsabilités des "Data Scientists".

- Les exigences de base d'un profil "Data Anlyst" sont : Logiciel (SQL, Python), l'outils et les technologies (Big Data, ETL, Airflow, Spark), les connaissance du cloud et du stockage (Google Cloud, AWS, Azure, GCP), diplôme (Master, Bachelor), type de travail (Stage, Senior ou travail à distance).

- Les exigences de base d'un profil "Big Data Consultant" sont : Logiciel (SQL, Python, Scala, Java), l'outils et les technologies (Big Data, Spark, Handoop, Kafka, Hbase, Hive, ETL), les connaissance du cloud et du stockage (Azure, AWS, GCP), diplôme (Master, Bachelor), type de travail (Stage, Junior, Senior).

- Les exigences de base d'un profil "Data scientist" sont : Logiciel (Python, SQL, Scala, C++, Java), l'outils et les technologies (Big Data, Spark, Handoop, Kubemetes, Bigquery, Hive, Airflow, ETL), les connaissance du cloud et du stockage (AWS, Azure, GCP, Google Cloud, RedShit), diplôme (Master, PhD, Msc, Post-doc), type de travail (Stage, Senior, Junior ou travail à distance).

- Les exigences de base d'un profil "Data scientist" sont : Logiciel (Python, Scala, C++, SQL, Java, C#), l'outils et les technologies (Big Data, Spark, Kafka, Airflow Cassandra, Kubernetes, Handoop, Hbase), les connaissance du cloud et du stockage (AWS, GCP, RedShit), diplôme (PhD, Master) type de travail (Stage, Senior).

- Les connaissances Python, R, SQL, Agile, Java, Tableau et les connaissance d'Algothime sont les compétences les plus demandées dans les descriptions d'offre d'emploi dans le domaine Data science en France.

- Nombre moyen d'années d'expérience requis est 4.5 ans dans les description des offres d'emploi.

- Les connaissances Python, R, SQL, Agile, Java, Tableau et les connaissance d'Algothime sont les compétences les plus demandées dans les descriptions d'offre d'emploi dans le domaine Data science en France.

- Le niveau d'entrée représente la plus grande proportion: 38%, suivie par le niveau de l'associé représentant 31%. Par ailleurs, le niveau du directeur et niveau de stage ne représentent qu'un faible pourcentage avec 2% et 6%, le reste étant de niveau de mi-senior et le niveau sans-applicables.

- Pour la catégorie, Ingénieure est le plus présente dans les offres d'emploi dans le domaine Data science en France avec un taux d'environ 45%, la deuxième catégorie est Technologie de l'information étant la catégorie information avec un taux d'environ 32%.

- Pour le type de l'emploi, le type temps plein est majoritaire environ 93%, ensuite est le type de stage environ 5% de nombre des offres d'emploi dans le domaine Data science en France.

- Pour l'industrie de l'emploi, Internet le plus grand taux de recrutement avec 35% du nombre d'offres d'emploi dans le domaine de la Data Science en France, suivi est de domaine de Computer Software et Technologies de l'information et Services avec le taux de 13 % et 12,6 %. Des domaines comme: Autonomie, Marketing and Advertising, Management Consulting ne représentent qu'un faible pourcentag.

Nous allons voir un profil de base

Entant un étudiant bientôt diplômé dans le domaine Data science et cherchant à percer dans le domaine de Data science et à poursuivre une carrière en tant qu'analyste de données, voici les principaux points à retenir:

-Pour améliorer mes chances, je devrais chercher des emplois à Paris.

Étant donné que j'ai déjà quelques connaissances en Python, SQL, je devrais continuer à améliorer ces compétences en Scala, C++, Java, l'outils et les technologies (Big Data, Spark, Handoop, Kubemetes, Bigquery, Hive, Airflow, ETL), les connaissance du cloud et du stockage (AWS, Azure, GCP, Google Cloud, RedShit),

- Pour trouver un moyen de mettre en valeur mes connaissances.

Je devrais vraiment envisager de faire un Doctorat car beaucoup de travail préfèrent un candidat avec un l'expérience de travail moyenne requise est de 4.5 ans, je dois donc trouver une poste en analyste débutant et faire plus de projets pour satisfaire cette exigence.

Réferences :

- Méthode d'analyse statistique exploratoire - Magali Coupaud, Jérémy Castéra, Michel Larini, Alice Delserieys (https://hal-amu.archives-ouvertes.fr/hal-01794680/document)

- ANALYSE DES OFFRES D'EMPLOI EN LIGNE : COMMENT CODER LE MÉTIER ? - Paul ANDREY, Maxime BERGEAT (http://www.jms.insee.fr/2018/S25_3_RESUME_ANDREY_JMS2018.pdf)

- Analyse exploratoire(Exploratory Data Analysis) - Sylvain Sardy (http://www.unige.ch/math/mgene/cours/slides2.pdf)