

Career-Crafters.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share RAM Disk

```
[ ] import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
import numpy as np
import re
```

Project Introduction

Overview

In today's competitive job market, indiscriminately applying to every available job posting is not only futile but also inefficient. A more strategic approach involves identifying job opportunities that align with an individual's specific skill set, thereby increasing the chances of the resume being shortlisted by Application Tracking Systems (ATS). This project aims to address this challenge by developing a recommender system that matches job seekers with the most relevant job postings based on their resumes.

Project Goals

- Objective:** To build a recommender system that informs job seekers of the best job opportunities aligned with their skills and experiences.
- Target:** Enhance job application effectiveness by ensuring resumes are matched with suitable job postings, potentially increasing interview chances.

Current Status

- Progress:** We are on track with our initial goals and the project scope remains unchanged.
- Insights:** The problem statement's open-ended nature has been acknowledged, leading to a structured plan for the upcoming phases of the project. Discussed at the end in up-coming work.

Notebook Content Overview

Datasets

The foundational dataset(s) for this project were sourced from Kaggle, featuring job listings scraped from LinkedIn across four major categories:

1. Data Analyst
2. Software Engineer
3. Data Scientist
4. Data Engineer

These categories encapsulate a substantial segment of the current Computer Science-related job market.

Methodology

Data Preparation

- Data Wrangling and Exploratory Data Analysis (EDA): Initial data cleaning and exploration to understand dataset characteristics.
- Job Location Processing: Deriving a 'city' column from 'job_location' to facilitate geocode retrieval.
- Geocode Retrieval: Implementing a caching mechanism for efficient batch retrieval of geocodes to avoid service provider blocks.

Text Processing

Comprehensive text processing on `job_description` and `job_skills`, including:

- Tokenization
- Case normalization
- Stopword removal
- Punctuation handling
- Lemmatization
- Part-of-Speech (POS) tagging
- Vectorization

Modeling and Recommendations

- Utilization of `CountVectorizer` and `Tf-IDF Vectorizer` for text vectorization.
- Matching job vectors to resume text to calculate similarity scores and identify top-n job recommendations.
- Application of K-Nearest Neighbors (KNN) for refined job recommendations.
- Visualization of recommended jobs in comparison to a baseline resume using t-SNE, highlighting the proximity of job opportunities in a reduced dimensional space.

Link: [Notebook in Repo](https://github.com/saurayjoshi/Career-Crafters/blob/main/Career_Crafters.ipynb)

Note: - If any visualization is not properly rendered in the exported pdf, please find all the visualizations at the end attached.

```
[ ] df_analyst = pd.read_csv("./master/data_analyst.csv")
df_software = pd.read_csv("./master/software_engineering.csv")
df_scientist = pd.read_csv("./master/data_scientist.csv")
df_data = pd.read_csv("./master/data_engineer.csv")

[ ] merged_dataframe = pd.concat([df_analyst, df_software, df_scientist, df_data], ignore_index=True)
merged_dataframe.to_csv("./merged_jobs.csv", index=False)

[ ] df_master = pd.read_csv("./master/merged_jobs.csv")
```

Based on initial data exploration, following can be observed:

- Presence of Null-values in `job_summary` and `job_skills`.
- Presence of junk in `job-description` such as emoticons, symbols and invalid characters.
- Over 32K rows of job postings.

```
[ ] df_master.shape
(32641, 11)
```

```
[ ] df_master.head()
```

	job_title	company	job_location	job_link	first_seen	search_city	search_country	job_level	job_type	job_summary	job_skills
0	Data Analyst-SQL, Tableau	Zortech Solutions	Mountain View, CA	https://www.linkedin.com/jobs/data-analyst-jobs	2023-12-20	Bloomington	United States	Associate	Onsite	NaN	NaN
1	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/marketing-rese...	2023-12-20	Bloomington	United States	Mid senior	Onsite	Company Description\nAre you a high-perform...	Data analysis, Market research, Survey develop...
2	Business Systems Analyst '1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-sys...	2023-12-20	Bloomington	United States	Mid senior	Onsite	Overview\nThe Business Systems	Business Analysis, Technical Writing, Database...

3	Senior VAT and Indirect Tax Analyst	Epic	Bloomington, IN	https://www.linkedin.com/jobs/view/senior-vat...sy...	2023-12-20	Bloomington	United States	Mid senior	Onsite	We're looking for an experienced tax professio...	Accounting, Finance, VAT/GST tax regimes, US a...	Analyst 1 perfo...	Writing, Software...	
4	Senior HRIS Analyst (Timekeeping and Payroll)	Nordson Corporation	Greater Bloomington Area	https://www.linkedin.com/jobs/view/senior-hris...y...	2023-12-20	Bloomington	United States	Mid senior	Remote	Collaboration drives Nordson's success as a ma...	Workday HCM, UKG Dimensions, Ceridian Dayforce...			

```
[ ] df_master.describe()
```

	job_title	company	job_location	job_link	first_seen	search_city	search_country	job_level	job_type	job_summary	job_skills		
count	32641	32641	32641		32641	32641	32641	32641	32641		32230		31368
unique	13465	9150	3259		29331	3	960	7	2		24147		31244
top	Senior Software Engineer	Jobs for Humanity	United States	https://ca.linkedin.com/jobs/view/customer-ser...	2023-12-20	Greater London	United States	Mid senior	Onsite	Who is Recruiting from Scratch :InRecruiting f...	Databricks, SQL, Python, PySpark, Azure Data F...		
freq	794	805	598		3	17236	341	25283	26244	15707		136	7

Since job_skills and job_summary are most important features for modelling, removing those with no information. Also fixing the job_countries text irregularities.

```
[ ] df_master.isna().sum()
```

```
job_title      0
company        0
job_location   0
job_link       0
first_seen     0
search_city    0
search_country 0
job_level      0
job_type       0
job_summary    411
job_skills     1273
dtype: int64
```

```
[ ] df_master.dropna(inplace=True)
```

```
[ ] df_master.isna().sum()
```

```
job_title      0
company        0
job_location   0
job_link       0
first_seen     0
search_city    0
search_country 0
job_level      0
job_type       0
job_summary    0
job_skills     0
dtype: int64
```

```
[ ] df_master.shape
```

```
(31368, 11)
```

```
[ ] df_master["search_country"].value_counts()
```

```
United States    24434
United Kingdom   2732
Canada          1779
Australia        810
Canada          767
United Kingdom   617
Australia        229
Name: search_country, dtype: int64
```

```
[ ] df_master['search_country'] = df_master['search_country'].str.strip()
```

```
[ ] df_master["search_country"].value_counts()
```

```
United States    24434
United Kingdom   3349
Canada          2546
Australia        1039
Name: search_country, dtype: int64
```

```
[ ] df_master["job_title"].value_counts()
```

```
Senior Software Engineer           793
Software Engineer                 587
Data Analyst                      572
Senior Data Engineer              531
Data Engineer                     472
Senior Business Systems Analyst, Finance Systems  ...
Senior FMV Analyst                1
BCBA - Board Certified Behavior Analyst 1
Treasury Analyst Senior           1
Installation Supervisor Data Cables / International Travel 1
Name: job_title, Length: 13155, dtype: int64
```

```
[ ] unique_list = []
```

```
for col in df_master.columns:
    item = (col, df_master[col].nunique(), df_master[col].dtypes, df_master[col].unique())
    unique_list.append(item)

unique_counts = pd.DataFrame(
    unique_list,
    columns=["Column", "Number_of_unique_values", "Type", "Unique_category"]
).sort_values(by="Number_of_unique_values")
```

```
display(unique_counts)
```

The first_seen column serves no value for us. Removing it.

```
[ ] df_master.drop('first_seen', axis=1, inplace=True)
```

```
[ ] df_copy = df_master.copy()
```

```
[ ] df_master = df_copy
```

This section does the following:

1. Tries to extract city from the job_location.
2. For cities marked as unknown as result of (1), information is tried to extract via a custom script defined in cells below.

```
[ ] from geotext import Geotext
[ ] def extract_city_with_geotext(location):
    places = Geotext(location)
    cities = list(places.cities)
    return cities[0] if cities else 'Unknown'

df_master['city'] = df_master['job_location'].apply(extract_city_with_geotext)

[ ] df_master.head()
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	city
1	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/market-rese...	Bloomington	United States	Mid senior	Onsite	Company DescriptionIn are you a high-performer ...	Data analysis, Market research, Survey develop...	Bloomington
2	Business Systems Analyst '1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-sy...	Bloomington	United States	Mid senior	Onsite	Overview in The Business Systems Analyst 1 perfo...	Business Analysis, Technical Writing, Software...	Bloomington
3	Senior VAT and Indirect Tax Analyst	Epic	Bloomington, IN	https://www.linkedin.com/jobs/view/senior-vat...	Bloomington	United States	Mid senior	Onsite	We're looking for an experienced tax professio...	Accounting, Finance, VAT/GST tax regimes, US a...	Bloomington
4	Senior HRIS Analyst (Timekeeping and Payroll)	Nordson Corporation	Greater Bloomington Area	https://www.linkedin.com/jobs/view/senior-hris...	Bloomington	United States	Mid senior	Remote	Collaboration drives Nordson's success as a ma...	Workday HCM, UKG Dimensions, Ceridian Dayforce...	Unknown
5	Business Intelligence Reporting Analyst 2	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-in...	Bloomington	United States	Mid senior	Hybrid	Overview in The Business Intelligence Analyst 2 ...	SAP Business Objects, SQL, Qlik, Data Modeling...	Bloomington

```
[ ] unknown_city_rows = df_master[df_master['city'] == 'Unknown']
unknown_city_rows.count()
```

```
[ ] job_title      5847
[ ] company       5847
[ ] job_location   5847
[ ] job_link        5847
[ ] search_city     5847
[ ] search_country   5847
[ ] job_level       5847
[ ] job_type         5847
[ ] job_summary      5847
[ ] job_skills        5847
[ ] city              5847
[ ] dtype: int64
```

```
[ ] known_countries = ['United States', 'Australia', 'Canada']
known_regions = ['England', 'Scotland', 'Wales', 'Northern Ireland']

known_states_us = [
    'AL', 'AK', 'AZ', 'AR', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA',
    'HI', 'ID', 'IL', 'IN', 'IA', 'KS', 'KY', 'LA', 'ME', 'MD',
    'MA', 'MI', 'MN', 'MS', 'MO', 'MT', 'NE', 'NV', 'NH', 'NJ',
    'NM', 'NY', 'NC', 'ND', 'OH', 'OK', 'OR', 'PA', 'RI', 'SC',
    'SD', 'TN', 'TX', 'UT', 'VT', 'VA', 'WA', 'WV', 'WI', 'WY'
]
```

```
[ ] def refine_city_extraction(row):
    if row['city'] == 'Unknown':
        location = row['job_location']
        parts = [part.strip() for part in location.split(',')]

        if len(parts) == 2 and parts[1] in known_states_us:
            return parts[0]
        elif len(parts) == 3:
            return parts[0]
        else:
            return parts[0]

    return row['city']

df_master['city'] = df_master.apply(refine_city_extraction, axis=1)

print(df_master[['job.location', 'city']])
```

	job_location	city
1	Bloomington, IN	Bloomington
2	Bloomington, IN	Bloomington
3	Bloomington, IN	Bloomington
4	Greater Bloomington Area	Unknown
5	Bloomington, IN	Bloomington
...
32635	Northamptonshire, England, United Kingdom	Northamptonshire
32636	Brackley, England, United Kingdom	Brackley
32637	Northampton, England, United Kingdom	Northampton
32638	Milton Keynes, England, United Kingdom	Milton Keynes
32639	Manchester Area, United Kingdom	Unknown

```
[31368 rows x 2 columns]
```

- Even after processing the cities, the left-over rows are dropped.
- Number of unknown cities went from 5847 to 2217

```
[ ] unknown_city_count = df_master[df_master['city'] == 'Unknown']['job.location'].count()
print(f"Number of rows with 'Unknown' city: {unknown_city_count}")

df_master = df_master[df_master['city'] != 'Unknown']
df_master.reset_index(drop=True, inplace=True)

df_master.head()
```

Number of rows with 'Unknown' city: 2217

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	city
0	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/market-rese...	Bloomington	United States	Mid senior	Onsite	Company DescriptionIn are you a high-performer ...	Data analysis, Market research, Survey develop...	Bloomington
1	Business Systems Analyst '1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-sy...	Bloomington	United States	Mid senior	Onsite	Overview in The Business Systems Analyst 1 perfo...	Business Analysis, Technical Writing, Software...	Bloomington
2	Senior VAT and Indirect Tax Analyst	Epic	Bloomington, IN	https://www.linkedin.com/jobs/view/senior-vat...	Bloomington	United States	Mid senior	Onsite	We're looking for an experienced tax professio...	Accounting, Finance, VAT/GST tax regimes, US a...	Bloomington
3	Business Intelligence Reporting Analyst 2	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-in...	Bloomington	United States	Mid senior	Hybrid	Overview in The Business Intelligence Analyst 2 ...	SAP Business Objects, SQL, Qlik, Data Modeling...	Bloomington
4	Lead Senior Business Analyst	Nashville Toyota North	Laughlin AFB, TX	https://www.linkedin.com/jobs/view/lead-senior...	Del Rio	United States	Mid senior	Hybrid	R2C, Inc. is a rapidly growing, SDVOSB headqua...	Information technology, Microsoft Office, Data...	Laughlin AFB

- Next we try to get the geocode of the cities.
- Seeing the number of unique cities as 2569, it was better to cache the geocodes, to avoid repeated request for same city from API.

```
[ ] df_master['city'].nunique()
2569
```

```
[ ] from geopy.geocoders import Nominatim
from geopy.exc import GeocoderTimedOut, GeocoderUnavailable
from time import sleep
```

```
[ ] l = geolocator.geocode("Monrovia, CA")
```

```
[ ] len(geocode_cache)
1701

[ ] geolocator = Nominatim(user_agent="sauravjoshi_33@rediffmail.com")

# Cache for storing city geocodes to avoid repeated requests
geocode_cache = {}

def get_lat_lon(city):
    if city in geocode_cache:
        return geocode_cache[city]
    else:
        max_retries = 5
        retries = 0
        while retries < max_retries:
            try:
                location = geolocator.geocode(city)
                if location:
                    geocode_cache[city] = (location.latitude, location.longitude)
                    print(f"Found: {geocode_cache[city]}")
                    return geocode_cache[city]
                else:
                    print(f"Not found: {city}")
                    geocode_cache[city] = (None, None)
                    return (None, None)
            except (GeocoderTimedOut, GeocoderUnavailable):
                retries += 1
                sleep_time = 5
                print(f"Timeout or unavailable, retrying... ({retries}/{max_retries})")
                sleep(sleep_time)
        return (None, None)

def process_batch(cities):
    return cities.apply(get_lat_lon)

# Batch processing
batch_size = 1000
for start in range(0, len(df_master), batch_size):
    end = start + batch_size
    df_master.loc[start:end, 'latitude_longitude'] = process_batch(df_master['job_location'][start:end])
    print(f"Processed batch {start//batch_size + 1}/{(len(df_master) - 1)//batch_size + 1}")
    sleep(1)

print(df_master.head())

[ ] none_count = df_master['latitude_longitude'].apply(lambda x: x == (None, None)).sum()
print(f"Number of (None, None) entries: {none_count}")
print(df_master.shape)

Number of (None, None) entries: 397
(29147, 14)

[ ] df_master = df_processed_coordinates.copy()

[ ] df_processed_coordinates = df_master.copy()
df_master = df_master[df_master['latitude_longitude'] != (None, None)]
df_master.reset_index(drop=True, inplace=True)
print(df_master.shape)

(28750, 14)

[ ] df_master.to_csv("./sample_data/coordinates_processed_merged.csv", index=False)

[ ] df_master = pd.read_csv("./drive/MyDrive/418/coordinates_processed_merged.csv")

[ ] df_master.head(2)



|   | job_title                          | company                       | job_location    | job_link                                          | search_city | search_country | job_level  | job_type | job_summary                                       | job_skills                                        | city        | latitude_longitude        | latitude | longitude  |
|---|------------------------------------|-------------------------------|-----------------|---------------------------------------------------|-------------|----------------|------------|----------|---------------------------------------------------|---------------------------------------------------|-------------|---------------------------|----------|------------|
| 0 | Market Research & Insights Analyst | Indiana University Foundation | Bloomington, IN | https://www.linkedin.com/jobs/view/market-rese... | Bloomington | United States  | Mid senior | Onsite   | Company Description\nAre you a high-performer ... | Data analysis, Market research, Survey develop... | Bloomington | (39.1670396, -86.5342881) | 39.16704 | -86.534288 |
| 1 | Business Systems Analyst           | Cook Medical                  | Bloomington, IN | https://www.linkedin.com/jobs/view/business-sy... | Bloomington | United States  | Mid senior | Onsite   | Overview\nThe Business Systems Analyst 1 perfo... | Business Analysis, Technical Writing, Software... | Bloomington | (39.1670396, -86.5342881) | 39.16704 | -86.534288 |



Now as we have latitude_longitude, we try to extract some information out of the jobs posted.

[ ] import folium
from folium.plugins import HeatMap

[ ] # df_master = df_processed_coordinates.copy()

[ ] df_processed_coordinates = df_master.copy()

[ ] df_master.head()



|   | job_title                               | company                       | job_location     | job_link                                          | search_city | search_country | job_level  | job_type | job_summary                                       | job_skills                                        | city         | latitude_longitude        | latitude  | longitude   |
|---|-----------------------------------------|-------------------------------|------------------|---------------------------------------------------|-------------|----------------|------------|----------|---------------------------------------------------|---------------------------------------------------|--------------|---------------------------|-----------|-------------|
| 0 | Market Research & Insights Analyst      | Indiana University Foundation | Bloomington, IN  | https://www.linkedin.com/jobs/view/market-rese... | Bloomington | United States  | Mid senior | Onsite   | Company Description\nAre you a high-performer ... | Data analysis, Market research, Survey develop... | Bloomington  | (39.1670396, -86.5342881) | 39.16704  | -86.534288  |
| 1 | Business Systems Analyst                | Cook Medical                  | Bloomington, IN  | https://www.linkedin.com/jobs/view/business-sy... | Bloomington | United States  | Mid senior | Onsite   | Overview\nThe Business Systems Analyst 1 perfo... | Business Analysis, Technical Writing, Software... | Bloomington  | (39.1670396, -86.5342881) | 39.16704  | -86.534288  |
| 2 | Senior VAT and Indirect Tax Analyst     | Epic                          | Bloomington, IN  | https://www.linkedin.com/jobs/view/senior-vat-... | Bloomington | United States  | Mid senior | Onsite   | We're looking for an experienced tax professio... | Accounting, Finance, VAT/GST tax regimes, US a... | Bloomington  | (39.1670396, -86.5342881) | 39.167040 | -86.534288  |
| 3 | Business Intelligence Reporting Analyst | Cook Medical                  | Bloomington, IN  | https://www.linkedin.com/jobs/view/business-in... | Bloomington | United States  | Mid senior | Hybrid   | Overview\nThe Business Intelligence Analyst 2 ... | SAP Business Objects, SQL, ODBC, Data Modeling... | Bloomington  | (39.1670396, -86.5342881) | 39.167040 | -86.534288  |
| 4 | Lead Senior Business Analyst            | Nashville Toyota North        | Laughlin AFB, TX | https://www.linkedin.com/jobs/view/lead-senior... | Del Rio     | United States  | Mid senior | Hybrid   | R2C, Inc. is a rapidly growing, SDVOB headqua...  | Information Technology, Microsoft Office, Data... | Laughlin AFB | (29.3614054, -100.778572) | 29.361405 | -100.778572 |


```

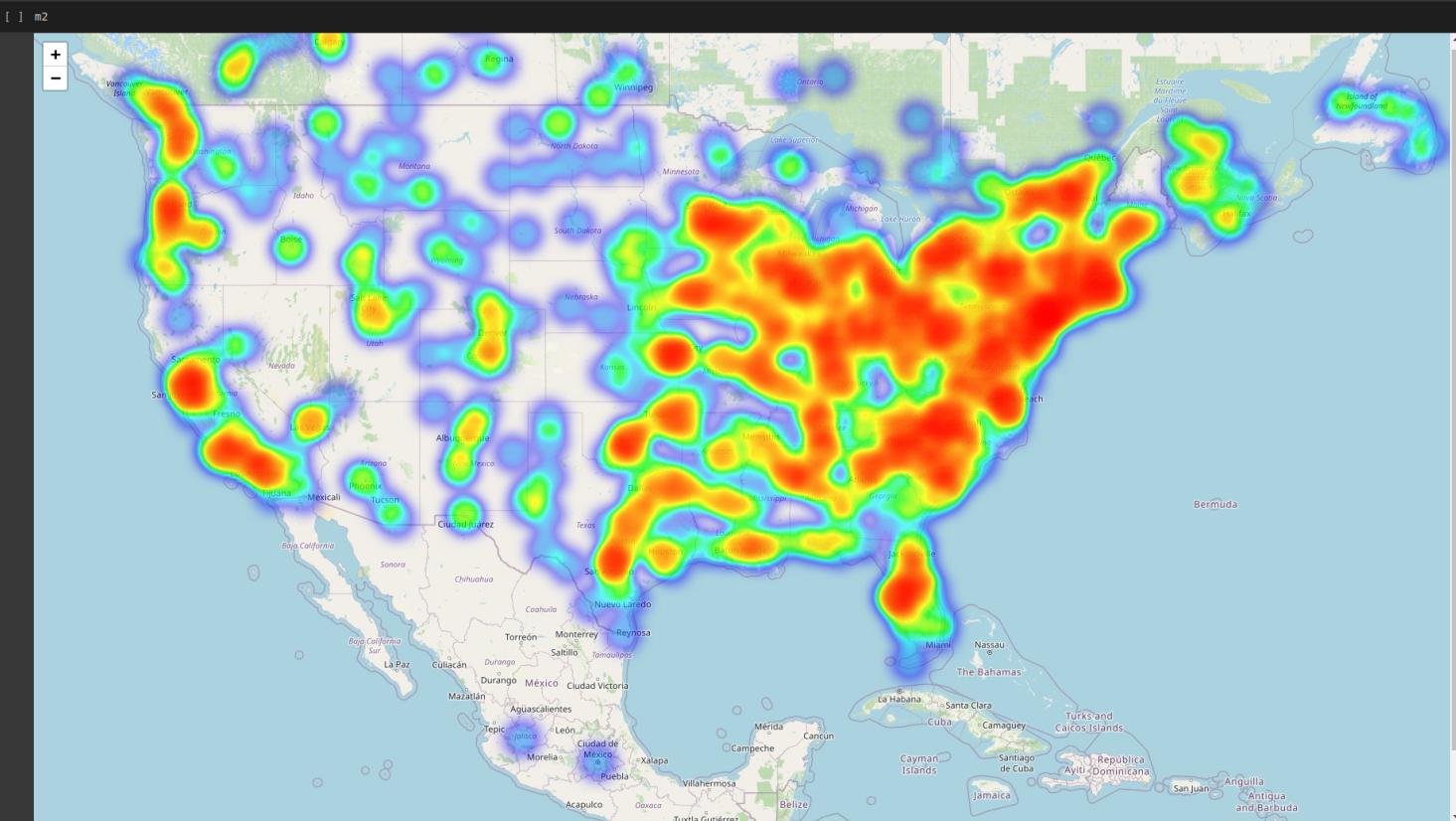
Visualization 1: Saurav Joshi

1. This heatmap shows us the job locations which are present across different countries.
2. The jobs seems to be more concentrated over West & Mid-West of America than Easts, Almost entirely in Australia towards the East-coast,

```
mostly in London, Manchester and Birmingham in UK and interestingly almost every job in Canada, comes with a location at a closer proximity with US such as at Vancouver and Montreal.
```

```
3. This visualization in-turn gives rise to know what jobs are prevalent in which regions.
```

```
[ ] coords = df_master[['latitude', 'longitude']].values.tolist()
map_center = [df_master['latitude'].mean(), df_master['longitude'].mean()]
m2 = folium.Map(location=map_center, zoom_start=2)
HeatMap(coords).add_to(m2)
<folium.plugins.heat_map.HeatMap at 0x7e781129a3e0>
```



```
[ ] # HeatMap(coords).add_to(m)
map_file_path = './drive/MyDrive/418/heatmap.html'
m2.save(map_file_path)
map_file_path
```

```
'./drive/MyDrive/418/heatmap.html'
```

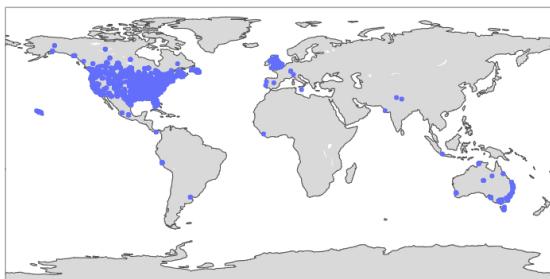
```
[ ] import pandas as pd
import plotly.express as px

fig = px.scatter_geo(df_master,
                     lat='latitude',
                     lon='longitude',
                     title='Geographical Plot of Cities')

fig.update_layout(showlegend=True, geo=dict(
    landcolor='rgb(217, 217, 217)')
)

fig.show()
```

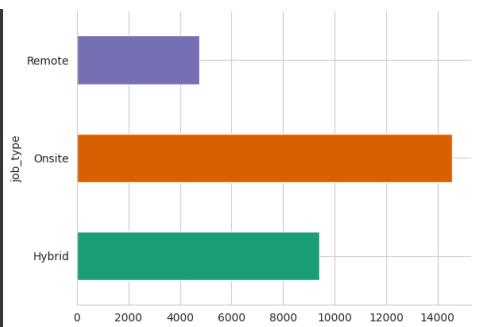
Geographical Plot of Cities



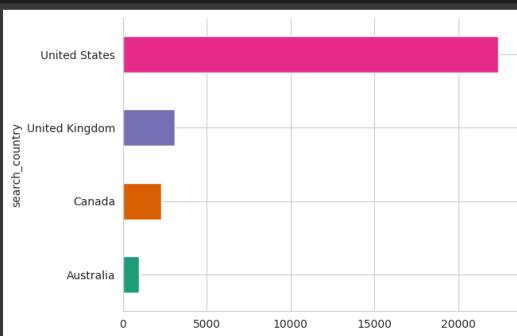
```
[ ]
```

```
Some more supportive visualizations for knowing the split of data on job_types and the countries they were searched from.
```

```
[ ] from matplotlib import pyplot as plt
import seaborn as sns
df_master.groupby('job_type').size().plot(kind='barh', color=sns.palettes.mpl_palette('Dark2'))
plt.gca().spines[['top', 'right']].set_visible(False)
```



```
[ ] from matplotlib import pyplot as plt
import seaborn as sns
df_master.groupby('search_country').size().plot(kind='barh', color=sns.palettes.mpl_palette('Dark2'))
plt.gca().spines[['top', 'right']].set_visible(False)
```



```
[ ] df_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28750 entries, 0 to 28749
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype  
 --- 
 0   job_title        28750 non-null   object 
 1   company          28750 non-null   object 
 2   job_location     28750 non-null   object 
 3   job_link         28750 non-null   object 
 4   search_city      28750 non-null   object 
 5   search_country   28750 non-null   object 
 6   job_level        28750 non-null   object 
 7   job_type         28750 non-null   object 
 8   job_summary      28750 non-null   object 
 9   job_skills        28750 non-null   object 
 10  city             28750 non-null   object 
 11  latitude_longitude 28750 non-null   object 
 12  latitude          28750 non-null   float64 
 13  longitude         28750 non-null   float64 
dtypes: float64(2), object(12)
memory usage: 3.14 MB
```

To get more information on job titles the irregularities present in the values of job_titles need to be handled.

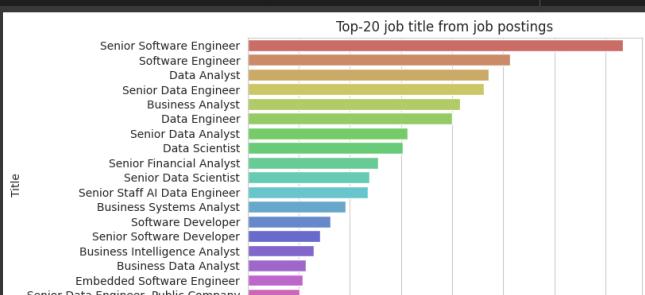
```
[ ] df_master['job_title'].value_counts().head(25)
# It can be seen that there are repetitions present like Senior Software Engineer and Sr. Software Engineer
# Also Senior Data Analyst vs Sr. Data Analyst
# Such titles need to be handled and changed into a single category.
# Not only these, some job title come along with name of company or some additional metadata which creates them as a new category.
```

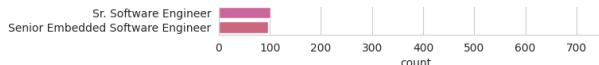
Senior Software Engineer	735
Software Engineer	513
Data Analyst	471
Senior Data Engineer	463
Business Analyst	415
Data Engineer	400
Senior Data Analyst	313
Data Scientist	304
Senior Financial Analyst	256
Senior Data Scientist	238
Senior Staff AI Data Engineer	235
Business Systems Analyst	192
Software Developer	162
Senior Software Developer	142
Business Intelligence Analyst	139
Business Data Analyst	113
Embedded Software Engineer	108
Senior Data Engineer, Public Company	101
Sr. Software Engineer	101
Senior Embedded Software Engineer	97
Senior VAT and Indirect Tax Analyst	97
Data Engineer, Scala(U.S. remote)	91
Sr. Data Analyst	88
Data Research Analyst, gt.school (Remote) - \$60,000/year USD	83
3SF Intelligence Analyst	82

Name: job_title, dtype: int64

```
[ ] job_title = df_master["job_title"].value_counts().reset_index()[:20]
job_title.columns = ["Title", "count"]

sns.barplot(x="count", y="Title", data=job_title, color="blue", palette="hls").set(title="Top-20 job title from job postings");
```





Resolving Cases such as:

1. Resolve Sr. Data Engineer to Senior Data Engineer
2. Sr. Data Analyst to Senior Data Analyst
3. Or Data Engineer - Scala(US, remote) to Data Engineer
4. Senior Data Engineer, Public Company to Senior Data Engineer
5. Labs - Data Scientist - Senior Associate to Data Scientist
6. Data Analyst (Bangkok Based, relocation provided) to Data Analyst

```
[ ] import re

def normalize_job_title(title):
    title = title.replace('Sr.', 'Senior')
    title = re.sub(r'\bSDE\b', 'Software Development Engineer', title, flags=re.IGNORECASE)
    title = re.sub(r'\bSW Engineer\b', 'Software Engineer', title, flags=re.IGNORECASE)

    title = re.sub(r'(\s*\.\s*)', '', title)
    title = re.sub(r'(\s*\.\s*)', '', title)
    title = re.sub(r'(\.\*)', '', title)

    title = title.strip()
    title = re.sub(r'^[A-Za-z0-9\s\-\_]+$', '', title)

    return title

df_master['normalized_job_title'] = df_master['job_title'].apply(normalize_job_title)

print(df_master[['job_title', 'normalized_job_title']]
```

```
          job title \
0      Market Research & Insights Analyst
1      Business Systems Analyst `1
2      Senior VAT and Indirect Tax Analyst
3      Business Intelligence Reporting Analyst 2
4      Lead Senior Business Analyst
...
28745   Customer Service Representative/Data Analyst/D...
28746       HR Systems and Data Analyst
28747           Senior Oracle Data Analyst
28748               Data Governance Analyst
28749       Energy Data and Forecast Analyst

          normalized_job_title
0      Market Research Insights Analyst
1      Business Systems Analyst 1
2      Senior VAT and Indirect Tax Analyst
3      Business Intelligence Reporting Analyst 2
4      Lead Senior Business Analyst
...
28745   Customer Service RepresentativeData AnalystDat...
28746       HR Systems and Data Analyst
28747           Senior Oracle Data Analyst
28748               Data Governance Analyst
28749       Energy Data and Forecast Analyst
```

[28750 rows x 2 columns]

```
[ ] df_master['normalized_job_title'].value_counts().head(25)
```

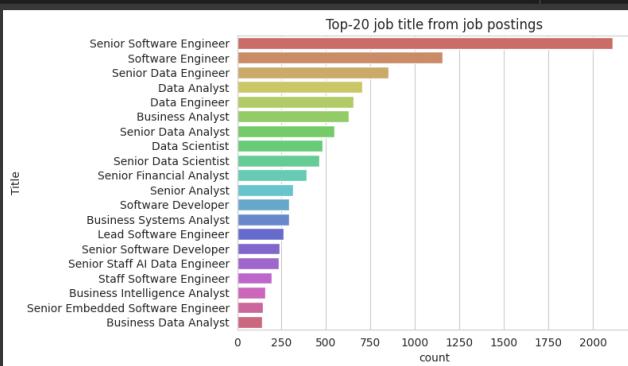
Normalized Job Title	Count
Senior Software Engineer	2111
Software Engineer	1155
Senior Data Engineer	852
Data Analyst	704
Data Engineer	653
Business Analyst	627
Senior Data Analyst	549
Data Scientist	482
Senior Data Scientist	461
Senior Financial Analyst	389
Senior Analyst	313
Software Developer	294
Business Systems Analyst	294
Lead Software Engineer	262
Senior Software Developer	238
Senior Staff AI Data Engineer	236
Staff Software Engineer	193
Business Intelligence Analyst	158
Senior Embedded Software Engineer	143
Business Data Analyst	142
Senior Business Analyst	128
Data Center Engineer	128
Embedded Software Engineer	125
Principal Software Engineer	116
Lead Data Engineer	112

Name: normalized_job_title, dtype: int64

```
[ ] job_title = df_master["normalized_job_title"].value_counts().reset_index()[:20]
```

```
job_title.columns = ["Title", "count"]
```

```
sns.barplot(x="count", y="Title", data=job_title, color="blue", palette="hls").set(title="Top-20 job title from job postings");
```



```
[ ] df_master.head()
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	city	latitude_longitude	longitude	normalized_job_title
0	Market Research & Insights	Indiana University	Bloomington, IN	https://www.linkedin.com/jobs/view/market-researcher-at-indiana-university-in-bloomington-in-1670396	Bloomington	United States	Mid	Onsite	Company DescriptionInArea	Data analysis, Market research	(39.1670396, -86.534288)	39.1670396, -86.534288	Market Research Insights	

	Insights Analyst	Foundation	Rese...	Senior	you a high performer ...	Research, Survey develop...	(-80.5542681)	Analyst							
1	Business Systems Analyst 1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-sy...	Bloomington	United States	Mid senior	Onsite	Overview in The Business Systems Analyst 1 perfo...	Business Analysis, Technical Writing, Software...	Bloomington	(-39.1670396, -86.5342881)	39.167040	-86.534288	Business Systems Analyst 1
2	Senior VAT and indirect Tax Analyst	Epic	Bloomington, IN	https://www.linkedin.com/jobs/view/senior-vat-...	Bloomington	United States	Mid senior	Onsite	We're looking for an experienced tax professio...	Accounting, Finance, VAT/GST tax regimes, US a...	Bloomington	(-39.1670396, -86.5342881)	39.167040	-86.534288	Senior VAT and Indirect Tax Analyst
3	Business Intelligence Reporting Analyst 2	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-in...	Bloomington	United States	Mid senior	Hybrid	Overview in The Business Intelligence Analyst 2 ...	SAP Business Objects, SQL, Qlik, Data Modeling...	Bloomington	(-39.1670396, -86.5342881)	39.167040	-86.534288	Business Intelligence Reporting Analyst 2
4	Lead Senior Business Analyst	Nashville Toyota North	Lahguhn AFB, TX	https://www.linkedin.com/jobs/view/lead-senior...	Del Rio	United States	Mid senior	Hybrid	R2C, Inc. is a rapidly growing, SDVOSB headqua...	Information technology, Microsoft Office, Data...	Laughlin AFB	(-29.3614054, -100.778572)	29.361405	-100.778572	Lead Senior Business Analyst

```
[ ] df_2 = df_master.copy()
# df_master = df_2.copy()
```

Visualization 2: Saurav Joshi

Now since we have normalized job title, we can now investigate dominant jobs based on cities.

- The concentration of points illustrates the relative number of jobs in each region, giving a visual representation of job availability and market demand.
- The clusters of job points may correlate with economic centers and tech hubs, suggesting where certain industries are thriving.
- This can allow candidates to target specific cities which are having more job postings as per their required job title.

```
▶ dominant_jobs = df_master.groupby('latitude_longitude')['normalized_job_title'].agg(
    lambda x: x.value_counts().idxmax()).reset_index(name='dominant_job')
```

```
[ ] dominant_jobs.head()
```

	latitude_longitude	dominant_job
0	(-11.9547401, -77.0612044)	eCommerce Analyst
1	(-11.9862556, -77.0096138)	Senior Data Engineer
2	(-12.1106354, -77.0471789)	Business Intelligence Analyst
3	(-12.46044, 130.8410469)	GIS Senior Analyst
4	(-12.463508000000001, 130.84355545363502)	Business Analyst for a Period of 6 Months

```
[ ] df_master = pd.merge(df_master, dominant_jobs, on='latitude_longitude', how='left')
fig = px.scatter_mapbox(df_master,
                        lat="latitude",
                        lon="longitude",
                        hover_name="dominant_job",
                        zoom=3,
                        height=600,
                        mapbox_style="carto-positron")
```

```
fig.show()
```



```
[ ] # Next thing to do is Text-preprocessing. Primarily for Job description and Normalized job title column.
```

```
# - tokenization
# - case handling, stop word removal
# - punctuation handling
# - lemmatization
# - POS tagging
# - Vectorization
```

```
import nltk
nltk.download("stopwords")
nltk.download('averaged_perceptron_tagger')
nltk.download('punkt')
nltk.download('wordnet')
from nltk import word_tokenize, pos_tag, pos_tag_sents
from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from string import punctuation
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
[nltk data] Downloading package stopwords to /root/nltk_data...
[nltk data]  Unzipping corpora/stopwords.zip.
[nltk data] Downloading package averaged_perceptron_tagger to
[nltk data]   /root/nltk_data...
[nltk data]  Unzipping taggers/averaged_perceptron_tagger.zip.
[nltk data] Downloading package punkt to /root/nltk_data...
[nltk data]  Unzipping tokenizers/punkt.zip.
[nltk data] Downloading package wordnet to /root/nltk_data...
```

```

[ ] def text_preprocessing(data_master):
    data_master["job_skills_summary"] = data_master["job_skills_summary"].apply(lambda x: re.sub(r"([A-Z][^a-z]*)", r" \1", x))
    data_master["job_skills_summary"] = data_master["job_skills_summary"].str.lower()

    return data_master

def get_pos(data_master):
    texts = data_master["job_skills_summary"].tolist()
    tagged_texts = pos_tag_sents(map(word_tokenize, texts))
    data_master["POS"] = tagged_texts

    return data_master

def get_informative_token(data_master):
    pos_noninformative = [".", "CC", "CD", "DT", "IN", "LS", "MD", "POS", "PRP",
                          "PRP$", "TO", "UH", "WDT", "WP", "WP$", "WRB"]

    data_master["POS_clean"] = data_master["POS"].apply(lambda x: [pair for pair in x if pair[0] != "nbsp" and pair[1] not in pos_noninformative])

    return data_master

def get_only_token(data_master):
    data_master["clean_token"] = data_master["POS_clean"].apply(lambda x: [word[0] for word in x])

    return data_master

def get_count_of_tokens(data_master):
    data_master["token_number"] = data_master["clean_token"].apply(lambda x: len(x))

    return data_master

```

```
[ ] duplicates = df_master[df_master.duplicated(subset='job_summary', keep=False)]
duplicates sorted = duplicates.sort_values(by='job_summary')
duplicates_sorted.head(4)
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	city	latitude_longitude	latitude	longitude	normalized_job_title
23615	Remote - Senior Application Engineer (Data Sim...	TechFetch.com - On Demand Tech Workforce hirin...	Canonsburg, PA	https://www.linkedin.com/jobs/view/remote-senior-application-engineer-data-simulation-at-t...	Weirton	United States	Mid senior	Remote	"All our jobs are US based and candidates must...	Data Simulation, PLM, Aras Innovator, Siemens ...	Canonsburg	(40.2588431, -80.186732)	40.258843	-80.186732	Remote
25121	Remote - Senior Application Engineer (Data Sim...	TechFetch.com - On Demand Tech Workforce hirin...	Canonsburg, PA	https://www.linkedin.com/jobs/view/remote-senior-application-engineer-data-simulation-at-t...	Steubenville	United States	Mid senior	Remote	"All our jobs are US based and candidates must...	Data Simulation, PLM, Aras Innovator, Siemens ...	Canonsburg	(40.2588431, -80.186732)	40.258843	-80.186732	Remote
28708	Big Data Developer	ASK Consulting	Ridgefield Park, NJ	https://www.linkedin.com/jobs/view/big-data-developer-at-ask-consulting-1perfo...	Greenwich	United States	Mid senior	Onsite	"All candidates must be directly contracted by...	SQL, Python, Spark, Hadoop, Kafka, Hive, Impal...	Ridgefield Park	(40.8570442, -74.0215285)	40.857044	-74.021529	Big Data Develop...
20421	Big Data Developer	ASK Consulting	Ridgefield Park, NJ	https://www.linkedin.com/jobs/view/big-data-developer-at-ask-consulting-1perfo...	Nyack	United States	Mid senior	Onsite	"All candidates must be directly contracted by...	SQL, Hive, Python, Spark, Hadoop, Kafka, Impal...	Ridgefield Park	(40.8570442, -74.0215285)	40.857044	-74.021529	Big Data Develop...

```
[ ] df_master = df_master.drop_duplicates(subset=['job_summary'], keep='first')

[ ] df_master["job_summary"].duplicated().sum()

0

[ ] df_master['job_skills_summary'] = df_master['job_skills'] + " " + df_master['job_summary']

[ ] df_master.head(2)
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	city	latitude_longitude	latitude	longitude	normalized_job_title
0	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/market-research-and-insights-analyst-at-indiana-university-foundation-1perfo...	Bloomington	United States	Mid senior	Onsite	Company Description: Are you a high-performer ...	Data analysis, Market research, Survey develop...	Bloomington	(39.1670396, -86.5342881)	39.16704	-86.534288	Market Research Insights Analyst
1	Business Systems Analyst '1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-systems-analyst-at-cook-medical-1perfo...	Bloomington	United States	Mid senior	Onsite	Overview: In The Business Systems Analyst 1 perfo...	Business Analysis, Technical Writing, Software...	Bloomington	(39.1670396, -86.5342881)	39.16704	-86.534288	Business Systems Analyst 1

```
[ ] df_master = text_preprocessing(df_master)

[ ] df_master["job_skills_summary"].describe()

count                           21926
unique                          21926
top      data analysis, market research, survey deve...
freq                           1
Name: job_skills_summary, dtype: object

[ ] df_master = get_pos(df_master)
df_master = get_informative_token(df_master)
df_master = get_only_token(df_master)
df_master = get_count_of_tokens(df_master)
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	city	latitude_longitude	latitude	longitude	normalized_job_title
0	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/market-research-and-insights-analyst-at-indiana-university-foundation-1perfo...	Bloomington	United States	Mid senior	Onsite	Company Description: Are you a high-performer ...	Data analysis, Market research, Survey develop...	Bloomington	(39.1670396, -86.5342881)	39.167040	-86.534288	Market Research Insights Analyst
1	Business Systems Analyst '1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-systems-analyst-at-cook-medical-1perfo...	Bloomington	United States	Mid senior	Onsite	Overview: In The Business Systems Analyst 1 perfo...	Business Analysis, Technical Writing, Software...	Bloomington	(39.1670396, -86.5342881)	39.167040	-86.534288	Business Systems Analyst 1

2	Senior VAT and Indirect Tax Analyst	Epic	Bloomington, IN	https://www.linkedin.com/jobs/view/senior-vat...	Bloomington	United States	Mid senior	Onsite	We're looking for an experienced tax professio...	Accounting, Finance, VAT/GST tax regimes, US a...	Bloomington	(39.1670396, -86.5342881)	39.167040	-86.534288	Senior VAT and Indirect Tax Analyst
3	Business Intelligence Reporting Analyst 2	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-in...	Bloomington	United States	Mid senior	Hybrid	OverviewInThe Business Intelligence Analyst 2 ...	SAP Business Objects, SQL, Qlik, Data Modeling...	Bloomington	(39.1670396, -86.5342881)	39.167040	-86.534288	Business Intelligence Reporting Analyst 2
4	Lead Senior Business Analyst	Nashville Toyota North	Laughlin AFB, TX	https://www.linkedin.com/jobs/view/lead-senior...	Del Rio	United States	Mid senior	Hybrid	R2C, Inc. is a rapidly growing, SDVOSB headqua...	Information technology, Microsoft Office, Data...	Laughlin AFB	(29.3614054, -100.778572)	29.361405	-100.778572	Lead Senior Business Analyst
5	Business Analyst	American National	Springfield, MO	https://www.linkedin.com/jobs/view/business-an...	Spokane	United States	Associate	Onsite	Are you a Business Analyst with a proven abili...	Business Analysis, Problem Solving, Root Cause...	Springfield	(37.2081729, -93.2922715)	37.208173	-93.292271	Business Analyst
6	GIS Analyst	Associated Electric Cooperative Inc.	Springfield, MO	https://www.linkedin.com/jobs/view/gis-analyst...	Spokane	United States	Associate	Hybrid	Are you passionate about the power of geograph...	GIS, ArcGIS, ArcGIS Pro, ArcGIS Hub, ArcGIS Po...	Springfield	(37.2081729, -93.2922715)	37.208173	-93.292271	GIS Analyst
7	Sr Business Systems Analyst	O'Reilly Auto Parts	Springfield, MO	https://www.linkedin.com/jobs/view/sr-business...	Spokane	United States	Mid senior	Onsite	O'Reilly Auto Parts has a proven track record ...	Tuition Educational Assistance Programs	Springfield	(37.2081729, -93.2922715)	37.208173	-93.292271	Sr Business Systems Analyst
8	Senior Network Analyst - DDI	Jobs for Humanity	Springfield, MO	https://www.linkedin.com/jobs/view/senior-netw...	Spokane	United States	Mid senior	Onsite	Company DescriptionInJobs for Humanity is part...	Information Technology, Network Solutions, DDI...	Springfield	(37.2081729, -93.2922715)	37.208173	-93.292271	Senior Network Analyst - DDI
9	Sr Business Systems Analyst	O'Reilly Auto Parts	Springfield, MO	https://www.linkedin.com/jobs/view/sr-business...	Spokane	United States	Mid senior	Onsite	Compensation Pay RangeIn\$103,000.00 - \$130,000...	Business Analysis, Software Development, SDLC...	Springfield	(37.2081729, -93.2922715)	37.208173	-93.292271	Sr Business Systems Analyst

- After doing pre-processing of text, here we collate the clean tokens.

- We derive POS tags for words in job description, then finally after cleaning POS tags we have the clean tokens.

- Some stopwords that are present in job description are skipped.

```
[ ] df_master["clean_token"] = [
    [token for token in tokens if len(token) > 1 and token not in ["company", "description", "title", "job", "skills"]]
    for tokens in df_master["clean_token"]
]

df_master = get_count_of_tokens(df_master)

df_master.head()
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	city	latitude_longitude	latitude	longitude	normalized_job_title
0	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/market-rese...	Bloomington	United States	Mid senior	Onsite	Company DescriptionInAre you a high-performer ...	Data analysis, Market research, Survey develop...	Bloomington	(39.1670396, -86.5342881)	39.167040	-86.534288	Market Research Insights Analyst
1	Business Systems Analyst '1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-sy...	Bloomington	United States	Mid senior	Onsite	OverviewInThe Business Systems Analyst 1 perfo...	Business Analysis, Technical Writing, Software...	Bloomington	(39.1670396, -86.5342881)	39.167040	-86.534288	Business Systems Analyst '1
2	Senior VAT and Indirect Tax Analyst	Epic	Bloomington, IN	https://www.linkedin.com/jobs/view/senior-vat...	Bloomington	United States	Mid senior	Onsite	We're looking for an experienced tax professio...	Accounting, Finance, VAT/GST tax regimes, US a...	Bloomington	(39.1670396, -86.5342881)	39.167040	-86.534288	Senior VAT and Indirect Tax Analyst
3	Business Intelligence Reporting Analyst 2	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-in...	Bloomington	United States	Mid senior	Hybrid	OverviewInThe Business Intelligence Analyst 2 ...	SAP Business Objects, SQL, Qlik, Data Modeling...	Bloomington	(39.1670396, -86.5342881)	39.167040	-86.534288	Business Intelligence Reporting Analyst 2
4	Lead Senior Business Analyst	Nashville Toyota North	Laughlin AFB, TX	https://www.linkedin.com/jobs/view/lead-senior...	Del Rio	United States	Mid senior	Hybrid	R2C, Inc. is a rapidly growing, SDVOSB headqua...	Information technology, Microsoft Office, Data...	Laughlin AFB	(29.3614054, -100.778572)	29.361405	-100.778572	Lead Senior Business Analyst

```
[ ] df_master["clean_job_desc"] = [* .join(x) for x in df_master["clean_token"]]
df_master.head(5)
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	...	latitude_longitude	latitude	longitude	normalized_job_title	job_s
0	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/market-rese...	Bloomington	United States	Mid senior	Onsite	Company DescriptionInAre you a high-performer ...	Data analysis, Market research, Survey develop...	...	(39.1670396, -86.5342881)	39.167040	-86.534288	Market Research Insights Analyst	data research
1	Business Systems Analyst '1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-sy...	Bloomington	United States	Mid senior	Onsite	OverviewInThe Business Systems Analyst 1 perfo...	Business Analysis, Technical Writing, Software...	...	(39.1670396, -86.5342881)	39.167040	-86.534288	Business Systems Analyst '1	business systems
2	Senior VAT and Indirect Tax Analyst	Epic	Bloomington, IN	https://www.linkedin.com/jobs/view/senior-vat...	Bloomington	United States	Mid senior	Onsite	We're looking for an experienced tax professio...	Accounting, Finance, VAT/GST tax regimes, US a...	...	(39.1670396, -86.5342881)	39.167040	-86.534288	Senior VAT and Indirect Tax Analyst	accounting vat/gst
3	Business Intelligence Reporting Analyst 2	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-in...	Bloomington	United States	Mid senior	Hybrid	OverviewInThe Business Intelligence Analyst 2 ...	SAP Business Objects, SQL, Qlik, Data Modeling...	...	(39.1670396, -86.5342881)	39.167040	-86.534288	Business Intelligence Reporting Analyst 2	sap sql
4	Lead Senior Business Analyst	Nashville Toyota North	Laughlin AFB, TX	https://www.linkedin.com/jobs/view/lead-senior...	Del Rio	United States	Mid senior	Hybrid	R2C, Inc. is a rapidly growing, SDVOSB headqua...	Information technology, Microsoft Office, Data...	...	(29.3614054, -100.778572)	29.361405	-100.778572	Lead Senior Business Analyst	information technology

5 rows × 21 columns

- As last part of text-processing, lemmatization is done.
- This gives us our final job description lemmatized feature.

- Some extra stop words are added that are content-specific.

```
[ ] wnl = WordNetLemmatizer()
patterns = "[a-zA-Z \n.]"

stopwords_eng = stopwords.words("english")
stopwords_eng.extend(["race", "ethnicity", "religion", "color", "sex", "age", "national", "origin", "genetic", "information", "sexual", "orientation", "disability", "gender", "identity", "week", "per", "please", "offer", "part time", "example", "compensation", "monday", "tuesday", "wednesday", "thursday", "friday", "saturday", "sunday", "pm", "am"])

def lemmatize_sentence(text):
    text = re.sub(patterns, " ", text)
    tokens = []

    for token in text.split():
        if token and token not in stopwords_eng:
            token = token.strip()
            token = wnl.lemmatize(token)

        tokens.append(token)

    return " ".join(tokens)

[ ] print("Before lemmatization:\n", df_master["clean_job_desc"].iloc[555])
print("\nAfter lemmatization:\n", lemmatize_sentence(df_master["clean_job_desc"].iloc[555]))
```

Before lemmatization:
data analysis data visualization data reconciliation ms excel data discrepancy identification cms fmis tableau dashboard power point access sql odbc service now business system analyst position hybrid location bost
After lemmatization:
data analysis data visualization data reconciliation m excel data discrepancy identification cm fmis tableau dashboard power point access sql odbc service business system analyst position hybrid location boston dur

```
[ ] df_master["job_desc_lem"] = df_master["clean_job_desc"].apply(lemmatize_sentence)
df_master.tail(5)
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	...	latitude	longitude	normalized_job_title	job_skills_summary
28735	Lead Data Engineer / Gaming space / onsite	Motion Recruitment Partners LLC	Santa Monica, CA	https://www.linkedin.com/jobs/view/lead-data-e...	Thousand Oaks	United States	Mid senior	Onsite	Dice is the leading career destination for tec...	Data Engineering, Data Analytics, Machine Lear...	...	34.019470	-118.491227	Lead Data Engineer Gaming space onsite	data engineering, analytics, machi
28738	Data Engineer	Whitbread	Dunstable, England, United Kingdom	https://uk.linkedin.com/jobs/view/data-engineer...	Milton Keynes	United Kingdom	Mid senior	Onsite	Data Engineer, Full-Time, Permanent, Dunstable...	Data Engineering, Data Pipelines, CI/CD, MLops...	...	51.886132	-0.520934	Data Engineer	data engineering, pipelines, ci/c
28742	Installation Supervisor (Data Cables / Internat...	Ernest Gordon Recruitment	Northampton, England, United Kingdom	https://uk.linkedin.com/jobs/view/installation-s...	Milton Keynes	United Kingdom	Mid senior	Onsite	40,000 - 45,000 + Company Van + (47k - 52k) O...	Installation, Data Cables, Telecommunications...	...	52.237885	-0.896364	Installation Supervisor	installation, data cables, telecommunic
28743	Rewards Data / HR Analyst	Latcom Ltd	Milton Keynes, England, United Kingdom	https://uk.linkedin.com/jobs/view/rewards-data-a...	Milton Keynes	United Kingdom	Mid senior	Onsite	Rewards Data / HR Analyst, required to work in...	Microsoft Excel, Microsoft Word, Microsoft Pow...	...	52.040650	-0.759409	Rewards Data HR Analyst	microsoft excel, microsoft word, microsoft pow
28744	Installation Supervisor Data Cables / Internat...	Ernest Gordon Recruitment	Northampton, England, United Kingdom	https://uk.linkedin.com/jobs/view/installation-s...	Milton Keynes	United Kingdom	Mid senior	Onsite	Installation Supervisor (Data Cables / Internat...	Installation Supervisor, Data Cables, Telecomm...	...	52.237885	-0.896364	Installation Supervisor Data Cables Internat...	installation supervisor, data cables, internat...

5 rows × 22 columns

```
[ ] df_master["token_number_after_lem"] = [len(word.split()) for word in df_master["job_desc_lem"]]
df_master.head()
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_summary	job_skills	...	longitude	normalized_job_title	job_skills_summary	POS	POS
0	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/market-rese...	Bloomington	United States	Mid senior	Onsite	Company Are you a high-performer ...	Data analysis, Market research, Survey develop...	...	-86.534288	Market Research Insights Analyst	data analysis, market research, survey deve...	[[data, NNS), ((data, NN), (analysis, NN), (market, NN))]	[[data, NN), ((data, NN), (analysis, NN), (market, NN))]
1	Business Systems Analyst '1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-sy...	Bloomington	United States	Mid senior	Onsite	Overview in The Business Systems Analyst 1 perfo...	Business Analysis, Technical Writing, Software...	...	-86.534288	Business Systems Analyst 1	business analysis, technical writing, soft...	[[business, NN), ((analysis, NN), (tech, NN))]	[[business, NN), ((analysis, NN), (tech, NN))]
2	Senior VAT and Indirect Tax Analyst	Epic	Bloomington, IN	https://www.linkedin.com/jobs/view/senior-vat-...	Bloomington	United States	Mid senior	Onsite	We're looking for an experienced tax professio...	Accounting, Finance, VAT/GST tax regimes, US a...	...	-86.534288	Senior VAT and Indirect Tax Analyst	accounting, finance, vat/gst tax regimes, ...	[[accounting, NN), ((finance, NN), (tax, NN))]	[[accounting, NN), ((finance, NN), (tax, NN))]
3	Business Intelligence Reporting Analyst 2	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-in...	Bloomington	United States	Mid senior	Hybrid	Overview in The Business Intelligence Analyst 2 ...	SAP Business Objects, SQL, Qlik, Data Modeling...	...	-86.534288	Business Intelligence Reporting Analyst 2	sap business objects, sql, qlik, data mod...	[[sap, NN), ((business, NN), (objects, NN), (...))]	[[sap, NN), ((business, NN), (objects, NN), (...))]
4	Lead Senior Business Analyst	Nashville Toyota North	Laughlin AFB, TX	https://www.linkedin.com/jobs/view/lead-senior...	Del Rio	United States	Mid senior	Hybrid	R2C, Inc. is a rapidly growing, SDVOSB headqua...	Information technology, Microsoft Office, Data...	...	-100.778572	Lead Senior Business Analyst	information technology, microsoft office, ...	[[information, NN), ((technology, NN), (tec NN))]	[[information, NN), ((technology, NN), (tec NN))]

5 rows × 23 columns

```
[ ] df_master.to_csv('/content/drive/MyDrive/418/text_processed.csv', index=False)
```

This is our final processed data which will now be used for modelling.

```
[ ] df_master = pd.read_csv('/content/drive/MyDrive/418/text_processed.csv')
```

```
[ ] df_master.columns
```

```
Index(['job_title', 'company', 'job_location', 'job_link', 'search_city', 'search_country', 'job_level', 'job_type', 'job_summary', 'job_skills', 'city', 'latitude_longitude', 'longitude', 'normalized_job_title', 'job_skills_summary', 'POS', 'POS_clean', 'clean_token', 'token_number', 'clean_job_desc', 'job_desc_lem', 'token_number_after_lem', dtype='object')
```

Visualization 3: Shahbaz Syed

- The chart presents the demand for various job skills across key global cities, categorizing roles into Hybrid, Onsite, and Remote.
- It reveals a notable preference for remote jobs in cities like New York and San Francisco, while onsite work remains prevalent in Chicago and London.

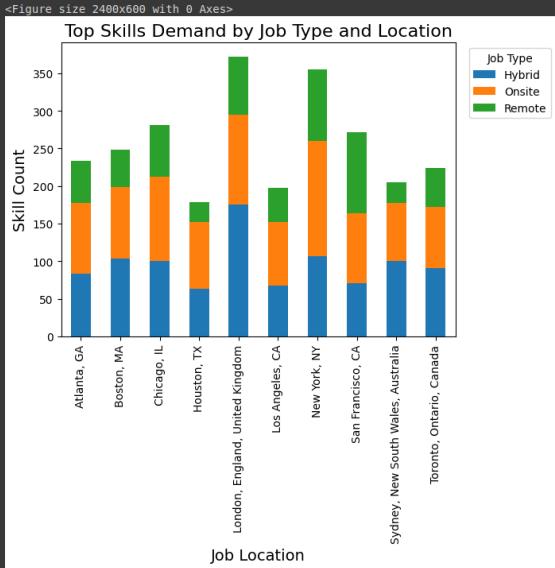
- This visualization serves as a strategic guide for job seekers and companies to understand and adapt to evolving employment trends. *italicized text*

```
[ ] top_locations = df_master['job_location'].value_counts().head(10).index

pivot_data = pd.pivot_table(df_master[df_master['job_location'].isin(top_locations)],
                           index='job_location',
                           columns='job_type',
                           values='token_number_after_lem',
                           aggfunc='count',
                           fill_value=0)

pivot_data = pivot_data.reset_index().sort_values(by='job_location')

plt.figure(figsize=(24, 6))
pivot_data.set_index('job_location')[sorted(pivot_data.columns[1:])].plot(kind='bar', stacked=True)
plt.title('Top Skills Demand by Job Type and Location', fontsize=16)
plt.xlabel('Job Location', fontsize=14)
plt.ylabel('Skill Count', fontsize=14)
plt.legend(title='Job Type', bbox_to_anchor=(1.02, 1), loc='upper left')
plt.xticks(rotation=90)
plt.show()
```



Visualization 4: Shahbaz Syed

- The word cloud displayed visualizes the frequency of various skills demanded in the top five job categories.
 - Prominent terms like 'Software Development,' AWS, and 'Data Engineering' stand out, suggesting their critical importance in the job market.
 - This offers a quick and intuitive way to understand the skill sets that are currently in high demand.

```
[ ] from wordcloud import WordCloud

top_job_titles = df_master['job_title'].value_counts().index[:5]
df_top_jobs = df_master[df_master['job_title'].isin(top_job_titles)]

skills_string = " ".join(skill for skill in df_top_jobs['job_skills'])

wordcloud = WordCloud(background_color='white').generate(skills_string)

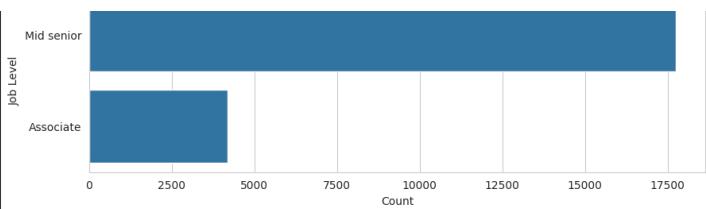
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.title('Word Cloud for Job Skills for top 5 type of jobs')
plt.axis('off')
plt.show()
```



Visualization 5: VVS Phani

- The bar chart illustrates the distribution of job levels, comparing the prevalence of 'Associate' versus 'Mid Senior' level positions.
 - The data shows a significantly higher count of 'Mid Senior' level jobs, highlighting a potential market trend favoring more experienced professionals.
 - This insight could be critical for both job seekers and employers in understanding the current employment landscape.

```
[ ] sns.set_style("whitegrid")
plt.figure(figsize=(10, 3))
sns.countplot(y='job level', data=df_master, order = df_master['job level'].value_counts().index)
plt.title('Distribution of Job Levels')
plt.xlabel('Count')
plt.ylabel('Job Level')
plt.show()
```

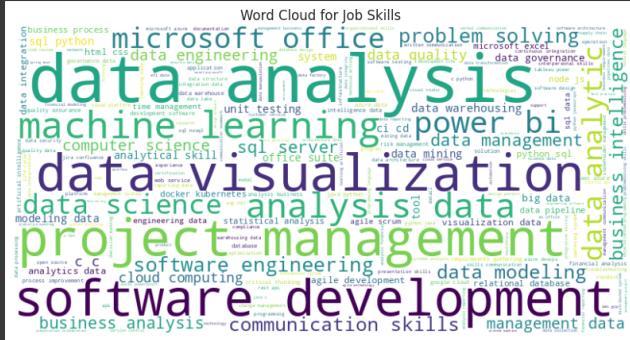


Visualization 6: VVS Phani

- The word cloud depicted focuses on key skills relevant to the job market, with terms like 'data analysis,' 'machine learning,' 'project management,' and 'software development' being especially prominent.
- This suggests a high demand for these skills across various industries in general.
- The graphic serves as a visual guide for individuals looking to refine their expertise to align with market needs.

```
[ ] skills_series = df_master['job_skills'].dropna().apply(lambda x: ' '.join(x.lower().split(',')))
all_skills = ' '.joinskills_series)
wordcloud = WordCloud(width=800, height=400, background_color = 'white').generate(all_skills)

plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Word Cloud for Job Skills')
plt.show()
```



Visualization 7: VVS Phani

- The heatmap visualizes a cross-tabulation of job levels against job types. - Mid-senior level roles dominate, especially in the onsite category, indicating a mature job market.
- Associate roles are fewer in comparison, with hybrid roles outnumbering remote opportunities at this level.
- This data could inform job seekers about the most prevalent job types at different career stages.

```
[ ] plt.figure(figsize=(10, 6))
job_level_type_cross = pd.crosstab(df_master['job_level'], df_master['job_type'])
sns.heatmap(job_level_type_cross, annot=True, fmt="d", cmap="BuPu")
plt.title('Job Level and Type Cross-tabulation')
plt.xlabel('Job Type')
plt.ylabel('Job Level')
plt.show()
```



[]

Dropping features that are not required now as they are redundant and other feature carry the inherent information they provide.

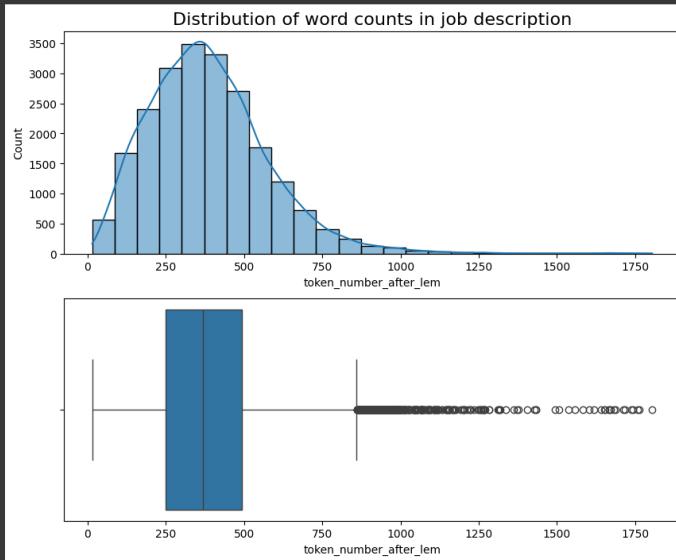
```
[ ] df_master.drop(['latitude_longitude', 'POS', 'POS_clean', 'clean_token', 'token_number', 'clean_job_desc', 'job_summary', 'latitude', 'longitude'], axis=1, inplace=True)
[ ] df_master.head(2)
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_skills	city	normalized_job_title	job_skills_summary	job_descлем	token_number_after_l
0	Market Research & Insights Analyst	Indiana University Foundation	Bloomington, IN	https://www.linkedin.com/jobs/view/market-rese...	Bloomington	United States	Mid senior	Onsite	Data analysis, Market research, Survey develop...	Bloomington	Market Research Insights Analyst	data analysis, market research, survey deve...	data analysis market research survey develop...	5
1	Business Systems Analyst 1	Cook Medical	Bloomington, IN	https://www.linkedin.com/jobs/view/business-sy...	Bloomington	United States	Mid senior	Onsite	Business Analysis, Technical Writing, Software...	Bloomington	Business Systems Analyst 1	business analysis, technical writing, sof...	business analysis technical writing software t...	3

This gives us the spread of token across the data.

- Though outliers are present (Jobs with more than 800 tokens in job summary), we are not removing them or truncating as they can contain rich-information that can help the model.
- This may lead to sparsity but we are not concerned with that at this point in time.

```
[ ] from scipy.stats import normaltest  
  
df_master['token_number_after_lem'].describe()  
fig, axes = plt.subplots(nrows=2, ncols=1, figsize=(10, 8))  
  
histplot = sns.histplot(  
    data=df_master,  
    x='token_number_after_lem',  
    bins=25,  
    kde=True,  
    ax=axes[0]);  
histplot.set_title("Distribution of word counts in job description", fontsize=16);  
  
boxplot = sns.boxplot(  
    data=df_master,  
    x='token_number_after_lem',  
    orient="h",  
    width=0.9,  
    ax=axes[1]);
```



```
[ ]
```

```
[ ] from collections import Counter, defaultdict  
  
df_master[df_master['token_number_after_lem'] > 800]['job_desc_lem'].count()  
630  
  
[ ] # Not removing the jobs as outliers with more token. May contain important information needed.  
  
[ ] df_master["normalized_job_title"][5]  
'Business Analyst'  
  
[ ] def normalize_skills(skill_string):  
    return [skill.strip().lower() for skill in skill_string.split(',') if skill]  
  
df_master['normalized_job_skills'] = df_master['job_skills'].apply(lambda x: normalize_skills(x))  
  
def top_n_skills(skill_lists, n=10):  
    aggregated_skills = Counter([skill for sublist in skill_lists for skill in sublist if skill])  
    return aggregated_skills.most_common(n)  
  
skill_counts_per_job_title = df_master.groupby('normalized_job_title')[['normalized_job_skills']].agg(lambda x: top_n_skills(x, n=10)).reset_index(name='top_skills')  
top_20_titles = df_master['normalized_job_title'].value_counts().head(20).index.tolist()  
  
for title in top_20_titles:  
    top_skills = skill_counts_per_job_title.loc[skill_counts_per_job_title['normalized_job_title'] == title, 'top_skills'].iloc[0]  
    print(f'Top 10 skills for {title}: {top_skills}')  
  
Top 10 skills for Senior Software Engineer: [('python', 796), ('java', 725), ('aws', 690), ('sql', 577), ('javascript', 510), ('agile', 489), ('docker', 402), ('go', 370), ('kubernetes', 369), ('typescript', 351)]  
Top 10 skills for Software Engineer: [('python', 357), ('java', 282), ('javascript', 254), ('aws', 235), ('software engineering', 222), ('git', 210), ('c++', 208), ('sql', 203), ('c#', 194), ('react', 176)]  
Top 10 skills for Business Analyst: [('business analysis', 296), ('project management', 197), ('data analysis', 164), ('communication', 129), ('requirements gathering', 108), ('agile', 103), ('sql', 101), ('excel', 97)]  
Top 10 skills for Senior Data Engineer: [('python', 427), ('sql', 364), ('data engineering', 247), ('aws', 195), ('snowflake', 159), ('spark', 153), ('data modeling', 152), ('etl', 140), ('data warehousing', 137), ('data湖', 136)]  
Top 10 skills for Data Engineer: [('python', 283), ('sql', 279), ('data engineering', 198), ('aws', 139), ('etl', 122), ('data warehousing', 118), ('data modeling', 104), ('azure', 91), ('data pipelines', 89), ('spark', 86)]  
Top 10 skills for Data Scientist: [('python', 345), ('machine learning', 291), ('data science', 260), ('sql', 230), ('r', 206), ('data visualization', 147), ('data analysis', 135), ('statistics', 114), ('data mining', 109), ('data science', 108), ('data engineering', 104), ('data architecture', 101), ('data湖', 97)]  
Top 10 skills for Data Analyst: [('data analysis', 266), ('sql', 257), ('data visualization', 178), ('tableau', 144), ('python', 129), ('power bi', 110), ('excel', 108), ('data mining', 78), ('r', 74), ('communication', 71), ('data science', 68), ('data engineering', 65), ('data architecture', 62), ('data湖', 59)]  
Top 10 skills for Senior Financial Analyst: [('financial analysis', 214), ('budgeting', 181), ('forecasting', 173), ('accounting', 162), ('financial modeling', 145), ('excel', 140), ('data analysis', 124), ('financial modeling', 118), ('data science', 109), ('data engineering', 108), ('data architecture', 107), ('data湖', 106), ('data mining', 105), ('data visualization', 104), ('data science', 103), ('data engineering', 102), ('data architecture', 101), ('data湖', 100)]  
Top 10 skills for Senior Data Scientist: [('python', 303), ('machine learning', 284), ('sql', 220), ('data science', 197), ('r', 166), ('data visualization', 109), ('statistics', 100), ('data analysis', 97), ('aws', 94), ('data engineering', 91), ('data architecture', 88), ('data湖', 85), ('data mining', 82), ('data science', 80), ('data engineering', 78), ('data architecture', 76), ('data湖', 74), ('data mining', 72), ('data science', 70), ('data engineering', 68), ('data architecture', 66), ('data湖', 64), ('data mining', 62), ('data science', 60), ('data engineering', 58), ('data architecture', 56), ('data湖', 54), ('data mining', 52), ('data science', 50), ('data engineering', 48), ('data architecture', 46), ('data湖', 44)]  
Top 10 skills for Senior Data Analyst: [('python', 251), ('data analysis', 193), ('python', 181), ('tableau', 170), ('data visualization', 161), ('power bi', 117), ('r', 108), ('data analytics', 85), ('excel', 75), ('tableau', 74), ('data visualization', 73), ('power bi', 72), ('r', 71), ('data analytics', 68), ('excel', 67), ('tableau', 66), ('data visualization', 65), ('power bi', 64), ('r', 63), ('data analytics', 62), ('excel', 61), ('tableau', 60), ('data visualization', 59), ('power bi', 58), ('r', 57), ('data analytics', 56), ('excel', 55), ('tableau', 54), ('data visualization', 53), ('power bi', 52), ('r', 51), ('data analytics', 50), ('excel', 49), ('tableau', 48), ('data visualization', 47), ('power bi', 46), ('r', 45), ('data analytics', 44), ('excel', 43), ('tableau', 42), ('data visualization', 41), ('power bi', 40), ('r', 39), ('data analytics', 38), ('excel', 37), ('tableau', 36), ('data visualization', 35), ('power bi', 34), ('r', 33), ('data analytics', 32), ('excel', 31), ('tableau', 30), ('data visualization', 29), ('power bi', 28), ('r', 27), ('data analytics', 26), ('excel', 25), ('tableau', 24), ('data visualization', 23), ('power bi', 22), ('r', 21), ('data analytics', 20), ('excel', 19), ('tableau', 18), ('data visualization', 17), ('power bi', 16), ('r', 15), ('data analytics', 14), ('excel', 13), ('tableau', 12), ('data visualization', 11), ('power bi', 10), ('r', 9), ('data analytics', 8), ('excel', 7), ('tableau', 6), ('data visualization', 5), ('power bi', 4), ('r', 3), ('data analytics', 2), ('excel', 1)]  
Top 10 skills for Software Developer: [('c++', 138), ('javascript', 115), ('sql', 89), ('software development', 81), ('java', 81), ('python', 75), ('git', 70), ('css', 63), ('net', 62), ('agile', 50)]  
Top 10 skills for Senior Analyst: [('data analysis', 95), ('sql', 81), ('excel', 79), ('communication', 74), ('project management', 63), ('tableau', 61), ('data visualization', 47), ('problem solving', 43), ('python', 41), ('data science', 39), ('data engineering', 37), ('data architecture', 35), ('data湖', 33), ('data mining', 31), ('data science', 29), ('data engineering', 27), ('data architecture', 25), ('data湖', 23), ('data mining', 21), ('data science', 19), ('data engineering', 17), ('data architecture', 15), ('data湖', 13), ('data mining', 11), ('data science', 9), ('data engineering', 7), ('data architecture', 5), ('data湖', 3)]  
Top 10 skills for Business Systems Analyst: [('sql', 96), ('project management', 83), ('business analysis', 75), ('communication', 65), ('data analysis', 59), ('agile', 51), ('business systems analysis', 46), ('problem solving', 43), ('python', 41), ('data science', 39), ('data engineering', 37), ('data architecture', 35), ('data湖', 33), ('data mining', 31), ('data science', 29), ('data engineering', 27), ('data architecture', 25), ('data湖', 23), ('data mining', 21), ('data science', 19), ('data engineering', 17), ('data architecture', 15), ('data湖', 13), ('data mining', 11), ('data science', 9), ('data engineering', 7), ('data architecture', 5), ('data湖', 3)]  
Top 10 skills for Lead Software Engineer: [('java', 161), ('aws', 157), ('python', 139), ('sql', 134), ('javascript', 105), ('cloud computing', 105), ('agile', 96), ('go', 94), ('typescript', 93), ('microservices', 91), ('node.js', 87), ('aws', 86), ('aws', 85), ('aws', 84), ('aws', 83), ('aws', 82), ('aws', 81), ('aws', 80), ('aws', 79), ('aws', 78), ('aws', 77), ('aws', 76), ('aws', 75), ('aws', 74), ('aws', 73), ('aws', 72), ('aws', 71), ('aws', 70), ('aws', 69), ('aws', 68), ('aws', 67), ('aws', 66), ('aws', 65), ('aws', 64), ('aws', 63), ('aws', 62), ('aws', 61), ('aws', 60), ('aws', 59), ('aws', 58), ('aws', 57), ('aws', 56), ('aws', 55), ('aws', 54), ('aws', 53), ('aws', 52), ('aws', 51), ('aws', 50), ('aws', 49), ('aws', 48), ('aws', 47), ('aws', 46), ('aws', 45), ('aws', 44), ('aws', 43), ('aws', 42), ('aws', 41), ('aws', 40), ('aws', 39), ('aws', 38), ('aws', 37), ('aws', 36), ('aws', 35), ('aws', 34), ('aws', 33), ('aws', 32), ('aws', 31), ('aws', 30), ('aws', 29), ('aws', 28), ('aws', 27), ('aws', 26), ('aws', 25), ('aws', 24), ('aws', 23), ('aws', 22), ('aws', 21), ('aws', 20), ('aws', 19), ('aws', 18), ('aws', 17), ('aws', 16), ('aws', 15), ('aws', 14), ('aws', 13), ('aws', 12), ('aws', 11), ('aws', 10), ('aws', 9), ('aws', 8), ('aws', 7), ('aws', 6), ('aws', 5), ('aws', 4), ('aws', 3), ('aws', 2), ('aws', 1), ('aws', 0)]  
Top 10 skills for Business Intelligence Analyst: [('sql', 181), ('business intelligence', 89), ('data visualization', 86), ('data analysis', 81), ('tableau', 70), ('power bi', 62), ('python', 45), ('data modeling', 42), ('data science', 39), ('data engineering', 37), ('data architecture', 35), ('data湖', 33), ('data mining', 31), ('data science', 29), ('data engineering', 27), ('data architecture', 25), ('data湖', 23), ('data mining', 21), ('data science', 19), ('data engineering', 17), ('data architecture', 15), ('data湖', 13), ('data mining', 11), ('data science', 9), ('data engineering', 7), ('data architecture', 5), ('data湖', 3)]  
Top 10 skills for Staff Software Engineer: [('python', 51), ('java', 48), ('aws', 40), ('javascript', 34), ('react', 31), ('kubernetes', 29), ('software engineering', 28), ('sql', 27), ('agile', 22), ('c++', 21)]  
Top 10 skills for Embedded Software Engineer: [('c++', 53), ('python', 38), ('c', 37), ('embedded systems', 29), ('linux', 25), ('i2c', 22), ('git', 21), ('embedded software development', 19), ('c++', 19), ('c', 18), ('i2c', 17), ('git', 16), ('embedded systems', 15), ('linux', 14), ('i2c', 13), ('git', 12), ('embedded software development', 11), ('c++', 11), ('c', 10), ('i2c', 9), ('git', 8), ('embedded systems', 7), ('linux', 6), ('i2c', 5), ('git', 4), ('embedded software development', 3), ('c++', 3), ('c', 2), ('i2c', 1), ('git', 1)]  
Top 10 skills for Senior Business Analyst: [('business analysis', 67), ('project management', 42), ('data analysis', 33), ('sql', 26), ('communication', 26), ('requirements gathering', 22), ('stakeholder management', 21), ('data mining', 20), ('data science', 19), ('data engineering', 17), ('data architecture', 15), ('data湖', 13), ('data mining', 11), ('data science', 9), ('data engineering', 7), ('data architecture', 5), ('data湖', 3)]  
Top 10 skills for Senior Embedded Software Engineer: [('c++', 46), ('python', 35), ('c', 32), ('git', 24), ('linux', 23), ('embedded linux', 22), ('rtos', 21), ('embedded systems', 20), ('spi', 20), ('c programming', 19), ('c++', 18), ('c', 17), ('i2c', 16), ('git', 15), ('embedded systems', 14), ('linux', 13), ('i2c', 12), ('git', 11), ('embedded systems', 10), ('linux', 9), ('i2c', 8), ('git', 7), ('embedded systems', 6), ('linux', 5), ('i2c', 4), ('git', 3), ('embedded systems', 2), ('linux', 1)]
```

Visualization 8: Usha

- The bar chart ranks the top 10 normalized job skills in frequency, with 'SQL' and 'Python' leading the chart, indicative of their high demand in the tech industry.
- Skills in 'data analysis' and 'java' are also prevalent, while 'communication' skills are emphasized as essential.
- This information is crucial for job seekers aiming to prioritize their learning and development focus areas.

```
[ ] sns.set_style("whitegrid")  
all_skills_dict = {}  
for title in top_20_titles:
```

```

for skill in skills_list:
    if skill in all_skills_dict:
        all_skills_dict[skill] += 1
    else:
        all_skills_dict[skill] = 1

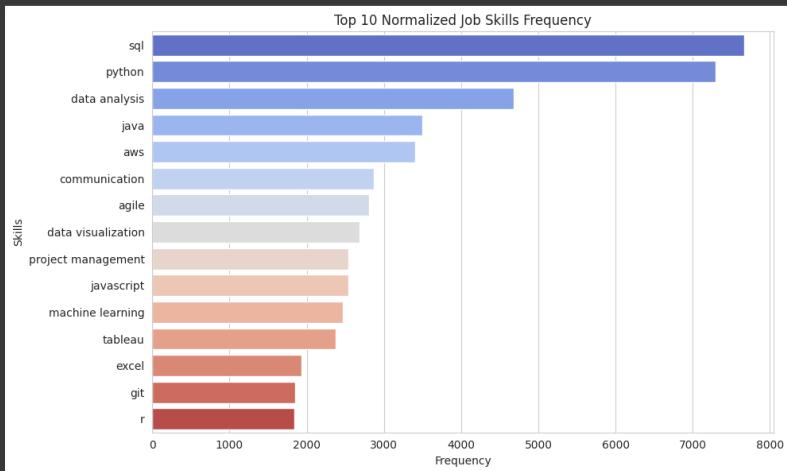
top_skills_dict = dict(sorted(all_skills_dict.items(), key=lambda item: item[1], reverse=True)[:15])

top_skills_df = pd.DataFrame(list(top_skills_dict.items()), columns=['Skill', 'Frequency'])

plt.figure(figsize=(10, 6))
sns.barplot(x='Frequency', y='Skill', data=top_skills_df, palette='coolwarm')
plt.title('Top 10 Normalized Job Skills Frequency')
plt.xlabel('Frequency')
plt.ylabel('Skills')
plt.tight_layout()

plt.show()

```



Visualization 9: Saurav Joshi

- After having the final data ready, this visualization helps us know the most common skills required in jobs.
- This can directly help a candidate to align themselves with these skills based on the job they are looking for.
- Also it can be seen as learning Python and SQL has the best advantage as it is present across job titles.

```

[ ] sns.set_style("whitegrid")

ncols = 2
num_titles_to_plot = 10
nrows = num_titles_to_plot // ncols + (num_titles_to_plot % ncols > 0)
fig, axs = plt.subplots(nrows=nrows, ncols=ncols, figsize=(20, 4 * nrows), constrained_layout=True)

palette = sns.color_palette("coolwarm", n_colors=10)
axs = axs.flatten()

for i, title in enumerate(top_20_titles[:num_titles_to_plot]):
    top_skills = skill_counts_per_job_title.loc[skill_counts_per_job_title['normalized_job_title'] == title, 'top_skills'].iloc[0]
    skills, counts = zip(*top_skills)

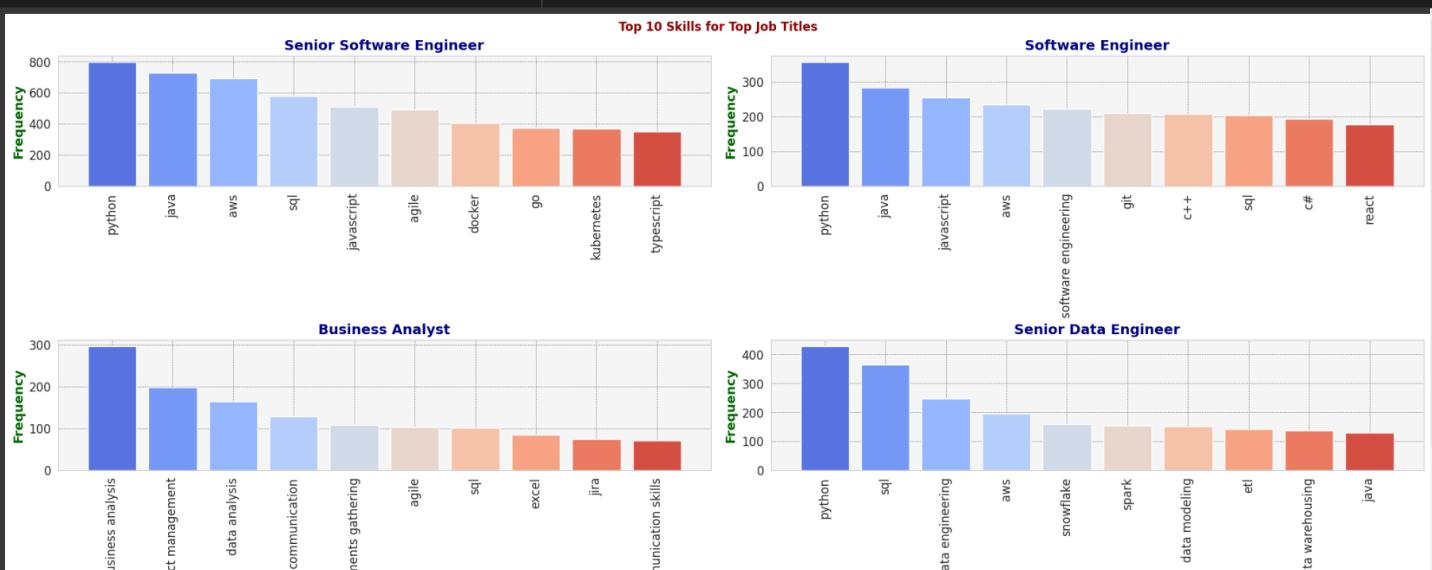
    axs[i].bar(skills, counts, color=palette)
    axs[i].set_title(title, fontsize=14, fontweight='bold', color='navy')
    axs[i].tick_params(axis='x', rotation=90, labelsize=12)
    axs[i].tick_params(axis='y', labelsize=12)
    axs[i].set_ylabel('Frequency', fontsize=13, fontweight='bold', color='darkgreen')

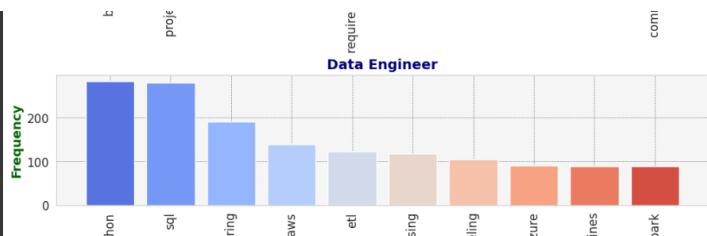
    axs[i].grid(True, which='major', linestyle='--', linewidth='0.5', color='grey')
    axs[i].set_facecolor('whitesmoke')

for i in range(len(top_20_titles[:num_titles_to_plot]), len(axs)):
    axs[i].axis('off')

plt.suptitle('Top 10 Skills for Top Job Titles', fontsize=12, fontweight='bold', color='darkred')
plt.show()

```





VISUALIZATION - 10: Venkata Pidaparthi

Geographical Distribution of Remote vs Onsite Jobs

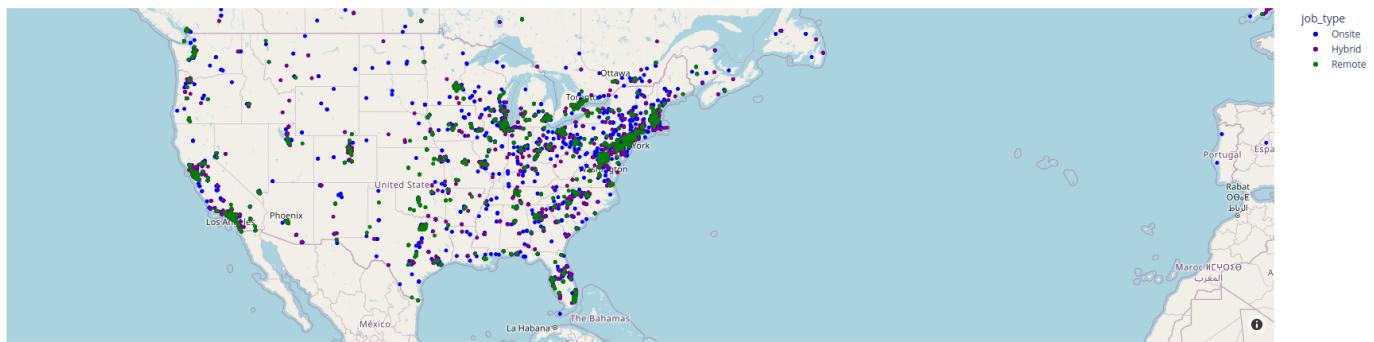
Purpose and Insight: This map visually differentiates between remote, onsite, and hybrid jobs based on their locations, highlighting regional preferences or necessities for job types. It's a tool for understanding the job market's geographical nuances, offering insights for job seekers, companies, and analysts.

Uses:

- Helps job seekers find areas with a high density of preferred job types.
- Aids companies in identifying regions with the workforce they need.
- Provides market insights for strategic decisions and trend analysis.

```
[ ] fig = px.scatter_mapbox(df_master, lat="latitude", lon="longitude", color="job_type",
                           hover_name="job_title", hover_data=["company", "job_location"],
                           color_discrete_map={"Remote": "green", "Onsite": "blue", "Hybrid": "purple"},
                           zoom=3, height=600, title="Geographical Distribution of Remote vs Onsite Jobs")
fig.update_layout(mapbox_style="open-street-map")
fig.show()
```

Geographical Distribution of Remote vs Onsite Jobs



Modelling

- Use **CountVectorizer** and **TfidfVectorizer** to convert text (processed job summary) into vectors.
- Using n-gram range from 1-2 for CountVectorizer as words like "machine learning" are relatively rich in information than machine and learning independently.
- Avoiding words that are over 95% and less than 5 in frequency.
- Getting cosine similarity for a sample resume and the vectors generated.

Saurav Joshi

```
[ ] from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()

def preprocess_text(text):
    text = text.lower()
    text = re.sub(r'[\w\s]', ' ', text)
    tokens = nltk.word_tokenize(text)
    tokens = [lemmatizer.lemmatize(token) for token in tokens if token not in stop_words]
    processed_text = ' '.join(tokens)
    return processed_text

# Baseline model setup
count_vect = CountVectorizer(ngram_range=(1,2), min_df=5, max_df=0.95)
count_job_desc = count_vect.fit_transform(df_master["job_desc_lem"])

# TF-IDF model setup
tfidf_vect = TfidfVectorizer(min_df=5, max_df=0.95)
tfidf_job_desc = tfidf_vect.fit_transform(df_master["job_desc_lem"])
```

A common function to process similarities based on tfidVec or countVec

- It is interesting to note that CountVectorizer's top 5 jobs possess more similarity with the resume than TfidfVectorizer.
 - However, this is because of the way **TfidfVectorizer** calculates the numerical value for a given vector. It adjusts for the frequency of the word in the corpus of documents. This means that common words across all documents are given less weight, and unique words are given more weight.
- The **CountVectorizer** model may be overemphasizing the importance of frequent terms, which can artificially inflate the similarity score, even if those terms aren't particularly meaningful.
- Thus, a higher cosine similarity score does not always mean a better match in terms of relevancy.

```
[ ] def jobs_to_resume(raw_resume_text, model_type='tfidf'):
    resume_text = preprocess_text(raw_resume_text)

    if model_type == 'tfidf':
        resume_vector = tfidf_vect.transform([resume_text])
        job_desc_vector = tfidf_job_desc
    else:
        resume_vector = count_vect.transform([resume_text])
        job_desc_vector = count_job_desc
```

```
cosine_similarities = cosine_similarity(resume_vector, job_desc_vector).flatten()
top_5_job_indices = cosine_similarities.argsort()[-5:][::-1]

top_5_jobs = df_master.iloc[top_5_job_indices]
job_indices = top_5_jobs.index

return top_5_jobs, job_indices, cosine_similarities[top_5_job_indices], resume_vector

# This is for now. This shall be updated to a resume parser.
raw_parsed_resume_text = "Saul Joshi 347-466-8549 | sjoshi50@uic.edu | linkedin.com/in/sauravjoshi27 | github.com/sauravjoshi | sauravjoshi.dev Versatile Software Engineer: Transforming Ideas into Reality with 3+ years of experience. I have a strong background in software engineering, machine learning, and data science. My expertise includes building efficient web applications using Python, Django, and React, as well as developing machine learning models using TensorFlow, PyTorch, and Scikit-learn. I am also experienced in working with large datasets using pandas, NumPy, and various database systems like MySQL and PostgreSQL. I have a passion for solving complex problems and a commitment to continuous learning and improvement. I am currently looking for opportunities to apply my skills and knowledge to challenging projects and contribute to the success of my team. I am available for remote or hybrid work arrangements and am excited about the prospect of working with a dynamic and supportive team." + "\n\n"

recommended_jobs_tfidf, job_indices_tfidf, scores_tfidf, resume_vector_tfidf = jobs_to_resume(raw_parsed_resume_text, 'tfidf')

recommended_jobs_count, job_indices_count, scores_count, resume_vector_count = jobs_to_resume(raw_parsed_resume_text, 'count')

print("TF-IDF Model Recommendations:")
for job_idx, score in zip(job_indices_tfidf, scores_tfidf):
    print(f"Job Index: {job_idx}, Score: {(score:.4f)}")
    print(f"Job Description: {df_master.loc[job_idx, 'job_desc_lem']}")
    print("----")

print("\nCount Vectorizer Model Recommendations:")
for job_idx, score in zip(job_indices_count, scores_count):
    print(f"Job Index: {job_idx}, Score: {(score:.4f)}")
    print(f"Job Description: {df_master.loc[job_idx, 'job_desc_lem']}")
    print("----")

TF-IDF Model Recommendations:
Job Index: 19972, Score: 0.2335
Job Description: ai data science team leadership nlp computer vision ml python pytorch tensorflow currently partnered exciting well funded fintech start greater san diego area looking expand data science team responsible for ...
-----
Job Index: 18653, Score: 0.2250
Job Description: data science machine learning artificial intelligence natural language processing audio signal processing video signal processing speech processing large language model llm python aws cloud service ...
-----
Job Index: 19906, Score: 0.2133
Job Description: data science machine learning natural language processing speech processing audio processing python py torch tensor flow tiny ml panda num py spark dask ray aws azure gcloud linux ssh department bsc ...
-----
Job Index: 18400, Score: 0.2090
Job Description: nlp machine learning python tensorflow py torch scikit learn num py panda matplotlib seaborn jupyter notebook google colab bert lla predictive modeling transfer learning big data deep learning model ...
-----
Job Index: 19298, Score: 0.2070
Job Description: python programming tensor flow kera py torch panda numpy scikit learn nltk scipy transformer based deep learning nlp model llm prompt engineering finetuning llm training llm model rag based business ...
-----

Count Vectorizer Model Recommendations:
Job Index: 19904, Score: 0.3167
Job Description: programming data science engineering scalable application web application visualization interface web framework apis database visualization framework backend language python java script html cs angular ...
-----
Job Index: 19906, Score: 0.3049
Job Description: data science machine learning natural language processing speech processing audio processing python py torch tensor flow tiny ml panda num py spark dask ray aws azure gcloud linux ssh department bsc ...
-----
Job Index: 17184, Score: 0.3033
Job Description: full stack development web development software engineering aws api development unit testing integration testing endto end testing java script type script html cs react angular vue javascript oauth jwt git ...
-----
Job Index: 12104, Score: 0.2945
Job Description: software engineering research development embedded system backend development database architecture data warehouse strategy data governance data security data modeling data mart data ingestion etl ...
-----
Job Index: 12866, Score: 0.2870
Job Description: angular java script html typescript cs .js web development software development machine learning user interface design data visualization data synchronization unit testing diamond kinetics diamond ...
-----
```

```
[ ] recommended_jobs_count.head(3)
```

	job_title	company	job_location	job_link	search_city	search_country	job_level	job_type	job_skills	city	normalized_job_title	job_skills_summary	job_descлем	token_number_after_lem
19904	Data Scientist, Senior Analyst	University of Chicago	Chicago, IL	https://www.linkedin.com/jobs/view/data-scient...	Chicago	United States	Mid senior	Onsite	Programming, Data science, Engineering, Scalab...	Chicago	Data Scientist	programming, data science, engineering, sc...	programming data science engineering scalable ...	7
19906	Sr. Data Scientist	University of Chicago	Chicago, IL	https://www.linkedin.com/jobs/view/sr-data-sci...	Chicago	United States	Mid senior	Onsite	Data science, Machine learning, Natural langu...	Chicago	Senior Data Scientist	data science, machine learning, natural lan...	data science machine learning natural language...	7
17184	Software Developer	Project Canary	Denver, CO	https://www.linkedin.com/jobs/view/software-de...	Boulder	United States	Mid senior	Onsite	Full Stack Development, Web Development, Softw...	Denver	Software Developer	full stack development, web development, ...	full stack development web development softwar...	3

Next we employ a KNN-based recommender system to match job descriptions with a candidate's resume.

- Unlike relying solely on term frequency metrics from `CountVectorizer` or `TfidfVectorizer`, our KNN approach captures the nuanced similarities within the feature space.
 - This method enhances the recommendation quality by considering a wider context and providing more personalized job matches.
 - It proves advantageous as it factors in the multidimensional relationship between data points, offering a sophisticated alternative to simple similarity scoring.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.neighbors import NearestNeighbors

def get_recommendations(model, job_desc_matrix, resume_vector, n_neighbors=5):
    model.fit(job_desc_matrix)

    distances, indices = model.kneighbors(resume_vector, n_neighbors=n_neighbors)

    recommendations = []
    for dist, idx in zip(distances[0], indices[0]):
        job_info = df_master.iloc[idx].to_dict()
        job_info['Match Distance'] = dist
        recommendations.append(job_info)
    return pd.DataFrame(recommendations)

knn = NearestNeighbors(n_neighbors=5, metric='cosine')

count_recommendations = get_recommendations(knn, count_job_desc, resume_vector_count, n_neighbors=5)
tfidf_recommendations = get_recommendations(knn, tfidf_job_desc, resume_vector_tfidf, n_neighbors=5)
```

- It is interesting to note that the Match distance for CountVectorizer base recommendations is low in comparison with TfidfVectorizer which clearly illustrated the shortcoming of Tfidf punishing resume description keywords occurrence across the corpus.
 - We thus aim to move towards more contextually rich embeddings word2vec, glove, or use state-of-the-art open source Embedding models or APIs.

```
[1] import plotly.graph_objects as go
```

```
x_tfidf = [0] * len(tfidf_recommendations['Match Distance'])
x_count = [1] * len(count_recommendations['Match Distance'])
```

```

fig = go.Figure()

fig.add_trace(go.Scatter(
    x=x_tfidf,
    y=tfidf_recommendations['Match Distance'],
    mode='markers',
    name='TF-IDF',
    marker=dict(color='skyblue', size=12, line=dict(color='black', width=1)),
    text=tfidf_recommendations['job_title'],
    hoverinfo='text+y'
))

fig.add_trace(go.Scatter(
    x=x_count,
    y=count_recommendations['Match Distance'],
    mode='markers',
    name='Count Vectorizer',
    marker=dict(color='lightgreen', size=12, line=dict(color='black', width=1)),
    text=count_recommendations['job_title'],
    hoverinfo='text+y'
))

fig.update_layout(
    title='Detailed Job Recommendations Comparison',
    xaxis=dict(
        tickmode='array',
        tickvals=[0, 1],
        ticktext=['TF-IDF', 'Count Vectorizer']
    ),
    yaxis_title="Match Distance",
    margin=dict(l=20, r=20, t=30, b=20)
)

fig.show()

```



Visualization 11: Saurav Joshi

- Trying to visualize how the recommended jobs are closer to resume in reduced dimensionality space (from original tfidf vector [21926 X 18017]).
- Leverage T-SNE for two-dimensional t-SNE space.
- The cluster of blue dots near the 'X' suggests that the top recommendations are indeed closely aligned with the resume's features.

```

[ ] tfidf_job_desc.shape
(21926, 18017)

[ ] from sklearn.manifold import TSNE

# Use t-SNE to reduce the tf-idf vectors to two dimensions
tsne = TSNE(n_components=2, random_state=42)
tfidf_reduced_tsne = tsne.fit_transform(tfidf_job_desc.toarray())

[ ] combined_tfidf = np.vstack((tfidf_job_desc.toarray(), resume_vector_tfidf.toarray()))
combined_reduced_tsne = tsne.fit_transform(combined_tfidf)

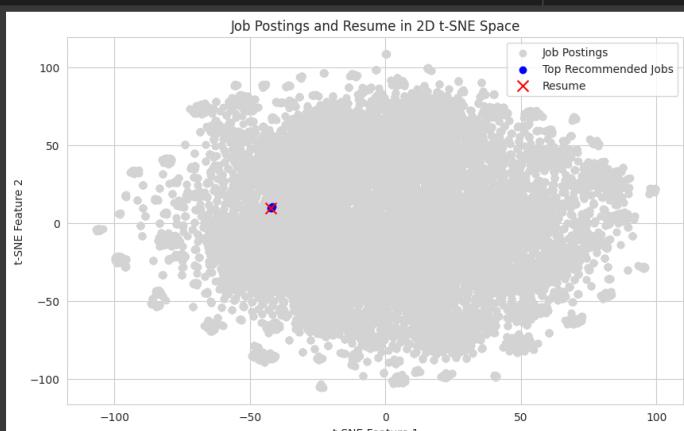
resume_reduced_tsne = combined_reduced_tsne[-1]
top_jobs_reduced_tsne = combined_reduced_tsne[job_indices_tfidf]

[ ] plt.figure(figsize=(10, 6))

plt.scatter(tfidf_reduced_tsne[:, 0], tfidf_reduced_tsne[:, 1], c='lightgray', label='Job Postings')
plt.scatter(top_jobs_reduced_tsne[:, 0], top_jobs_reduced_tsne[:, 1], c='blue', label='Top Recommended Jobs')
plt.scatter(resume_reduced_tsne[0], resume_reduced_tsne[1], c='red', marker='x', s=100, label='Resume')

plt.xlabel('t-SNE Feature 1')
plt.ylabel('t-SNE Feature 2')
plt.title('Job Postings and Resume in 2D t-SNE Space')
plt.legend()
plt.show()

```



Measurements

1. Creating 2 ground truth for the resume.
 - Job Titles relevant to the resume
 - Job Skills relevant to the resume
2. Finally looking into the recommended jobs based on both Count and TF-IDF vector to check similarity with the ground truth labels.

```
[ ] # Measuring How good the recommendations are.
# Using fuzzy matching instead of direct equivalent keyword matching.
!pip install fuzzywuzzy

[ ] from fuzzywuzzy import fuzz

base_job_titles = ['Data Scientist', 'Machine Learning Engineer', 'Software Engineer']
base_skills = ['Python', 'Machine Learning', 'Deep Learning', 'C++', 'Node.JS', 'JavaScript', 'MongoDB']

def count_relevant_titles(recommendations, base_titles):
    count = 0
    for title in recommendations['normalized_job_title']:
        if any(fuzz.partial_ratio(title.lower(), ref.lower()) > 80 for ref in base_titles):
            count += 1
    return count

def count_relevant_skills(recommendations, base_skills):
    count = 0
    for skills in recommendations['normalized_job_skills']:
        individual_skills = skills
        for skill in individual_skills:
            for skill in individual_skills:
                if any(fuzz.partial_ratio(skill.strip().lower(), ref.lower()) > 90 for ref in base_skills):
                    count += 1
                    break
    return count

relevant_titles_tfidf = count_relevant_titles(tfidf_recommendations, base_job_titles)
relevant_titles_count = count_relevant_titles(count_recommendations, base_job_titles)

relevant_skills_tfidf = count_relevant_skills(tfidf_recommendations, base_skills)
relevant_skills_count = count_relevant_skills(count_recommendations, base_skills)

print("TF-IDF Recommendations:")
print(f"Relevant Titles: {relevant_titles_tfidf} / {len(tfidf_recommendations)}")
print(f"Relevant Skills: {relevant_skills_tfidf} / {len(tfidf_recommendations)}")

print("\nCount Vectorizer Recommendations:")
print(f"Relevant Titles: {relevant_titles_count} / {len(count_recommendations)}")
print(f"Relevant Skills: {relevant_skills_count} / {len(count_recommendations)}")

TF-IDF Recommendations:
Relevant Titles: 5 / 5
Relevant Skills: 5 / 5

Count Vectorizer Recommendations:
Relevant Titles: 4 / 5
Relevant Skills: 5 / 5
```

Project Reflection

Challenges Encountered

- **Data Procurement and Processing:** Securing appropriate data, particularly managing and deriving insights from textual content, alongside the complexities of modeling based on this data.

Initial Insights

- **Model Performance:** The existing model is adept at providing job recommendations, with plans to further explore and utilize advanced embedding techniques for future model iterations.

Concrete Results

- **Relevancy of Recommendations:** The model's job recommendations align closely with the baseline resume, showcasing the practical effectiveness of the current system.

Future Hurdles

- **Enhanced Embeddings:** The primary focus is now on incorporating embeddings that capture richer contextual information to refine the recommendation process further.

Project Trajectory

- **On Track:** The project is progressing as planned, aligning with the goals and deliverables outlined during the initial pitch presentation.

Forward Strategy

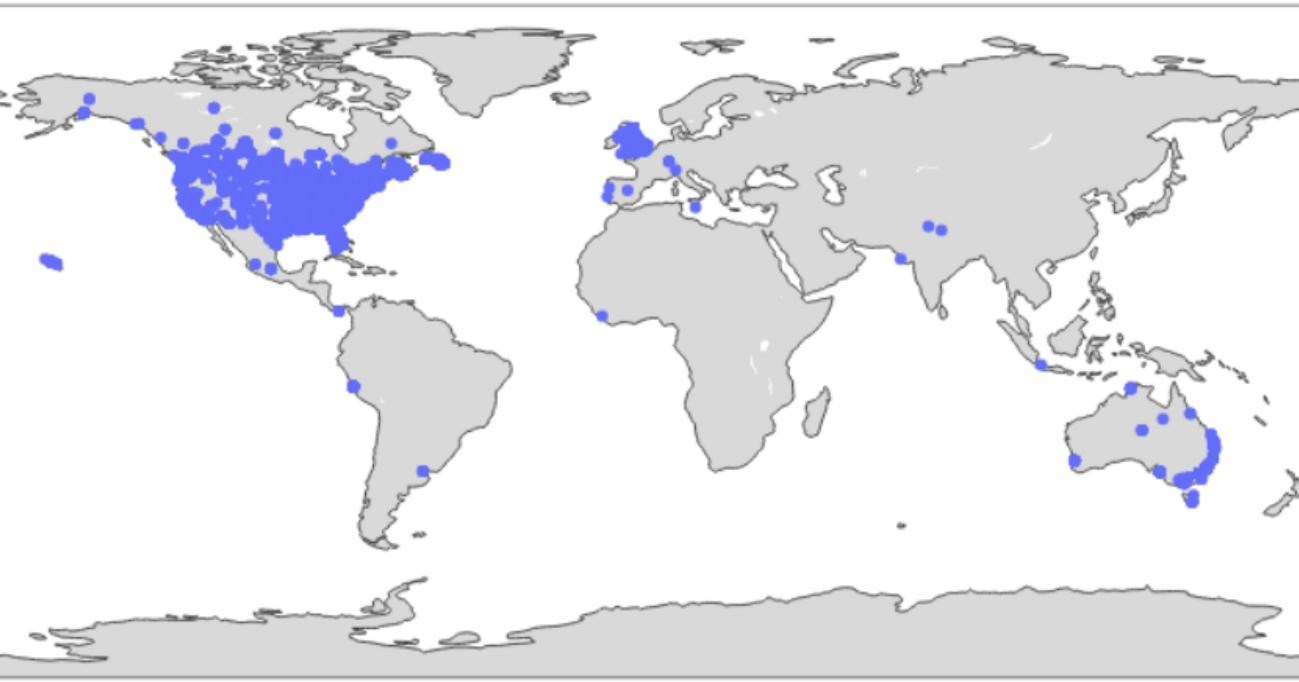
- **Viability and Advancements:** The positive outcomes justify the project's continuation, with an immediate strategy to adopt state-of-the-art embedding models to improve the system's recommendation capabilities.

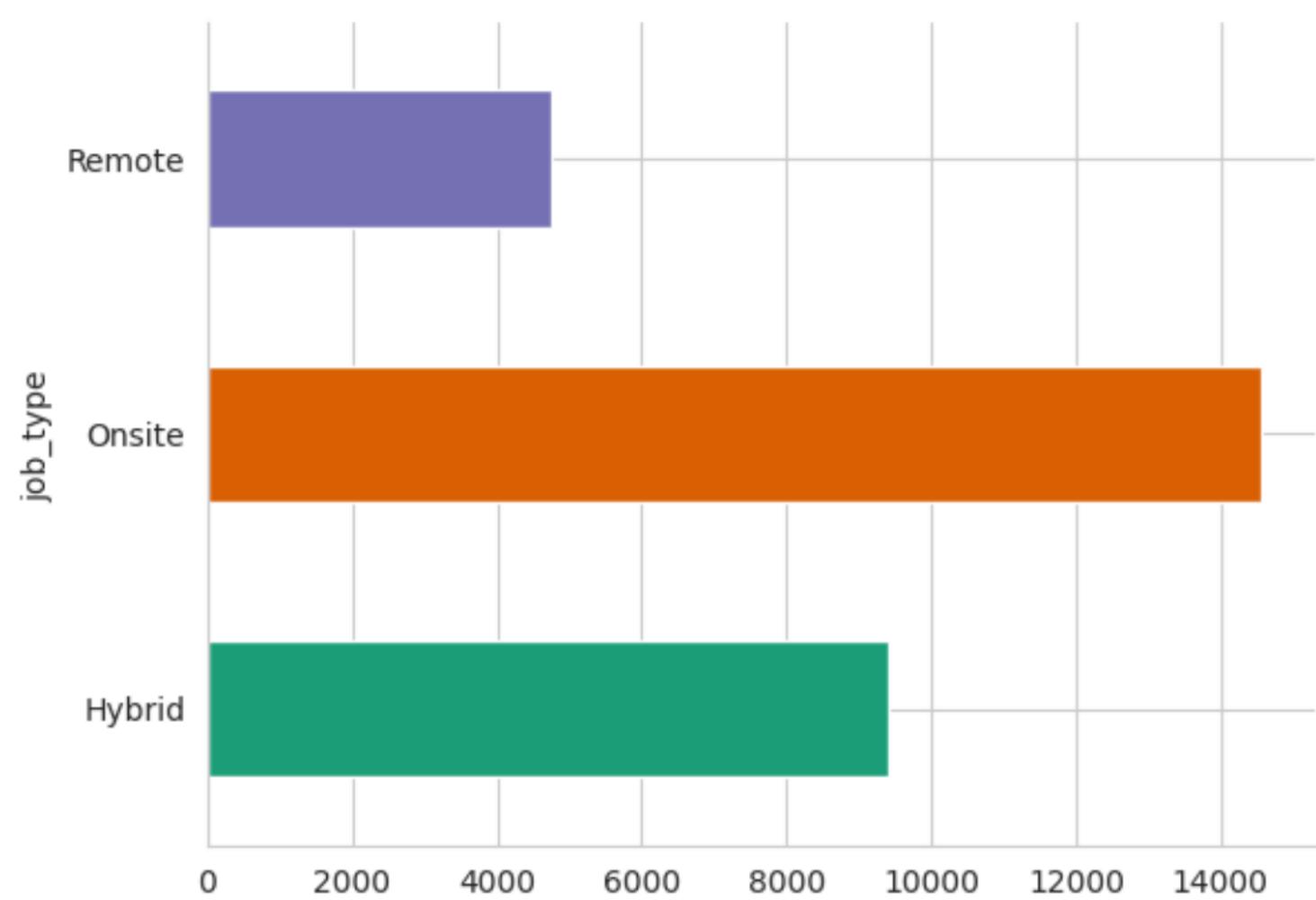
Next Steps:

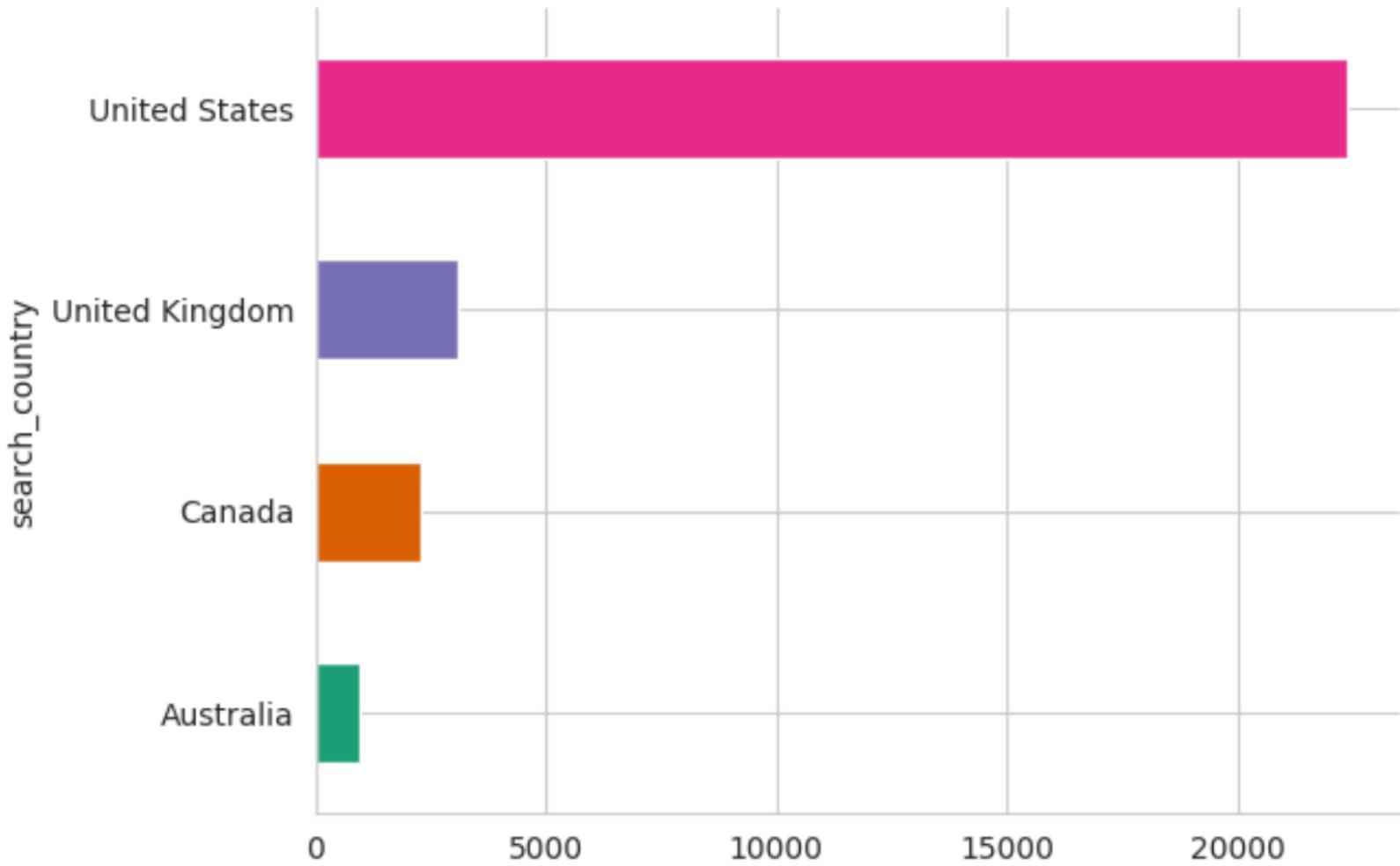
1. We will be building dynamic web-application that takes in your resume, stores it in your machine (keeping your sensitive data to you).
2. Using Clustering and KNN interpolated to recommend top-K relevant jobs amongst the most fit cluster.
3. Moving to advanced embedding techniques as discussed above.



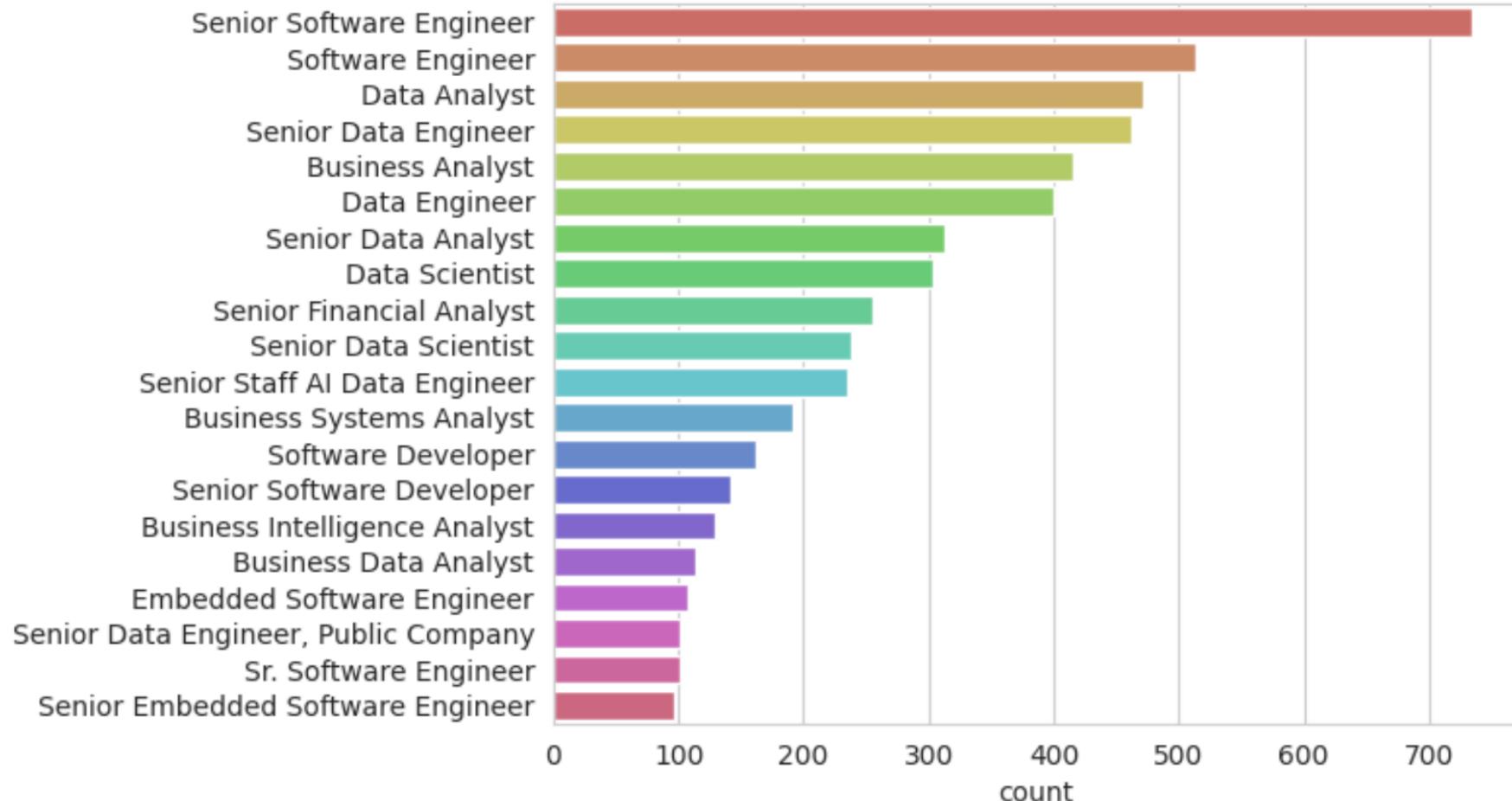
Geographical Plot of Cities



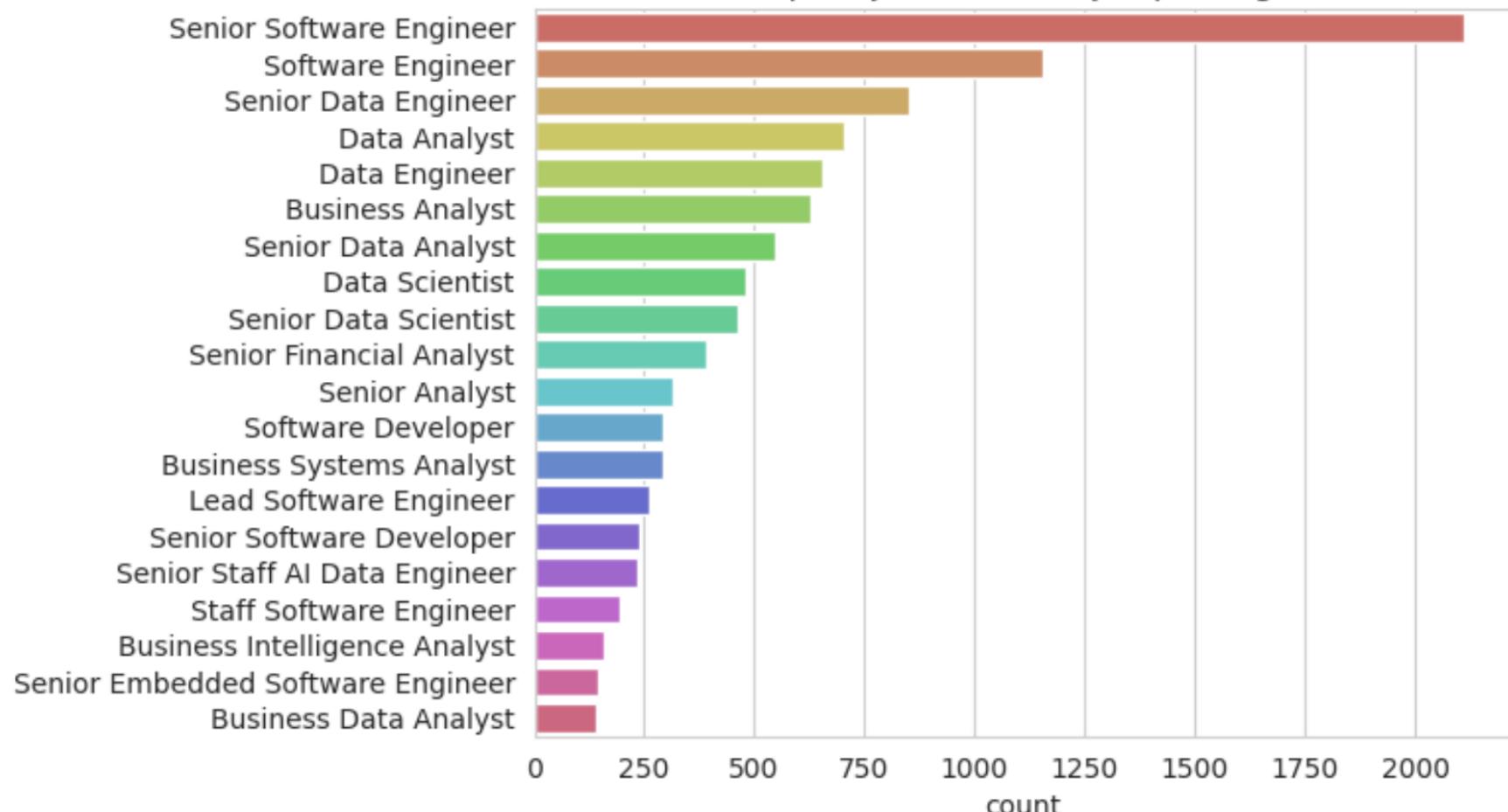


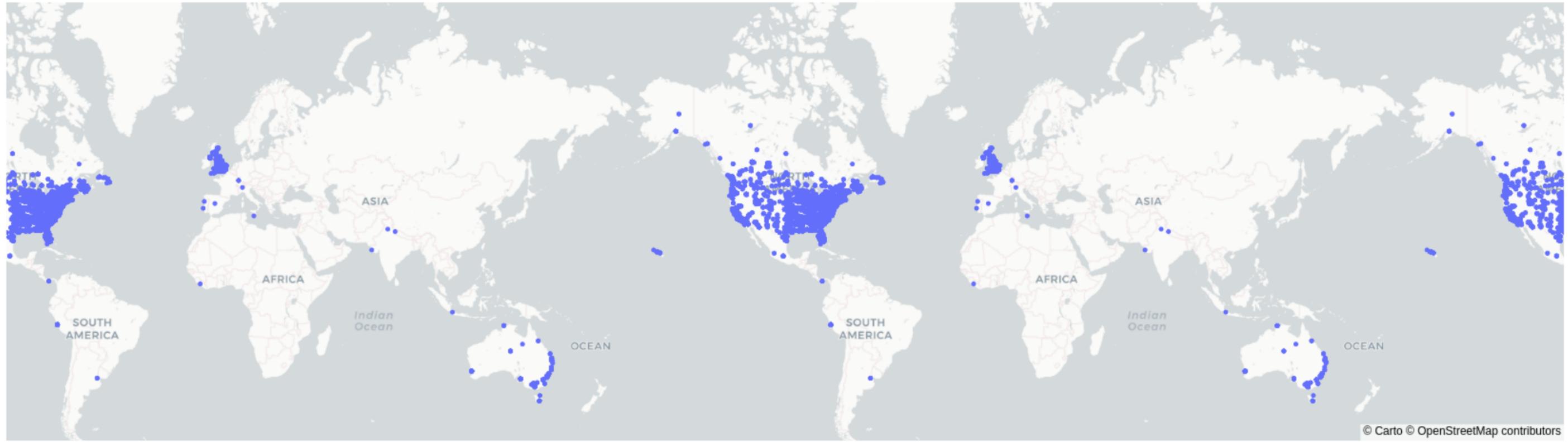


Top-20 job title from job postings

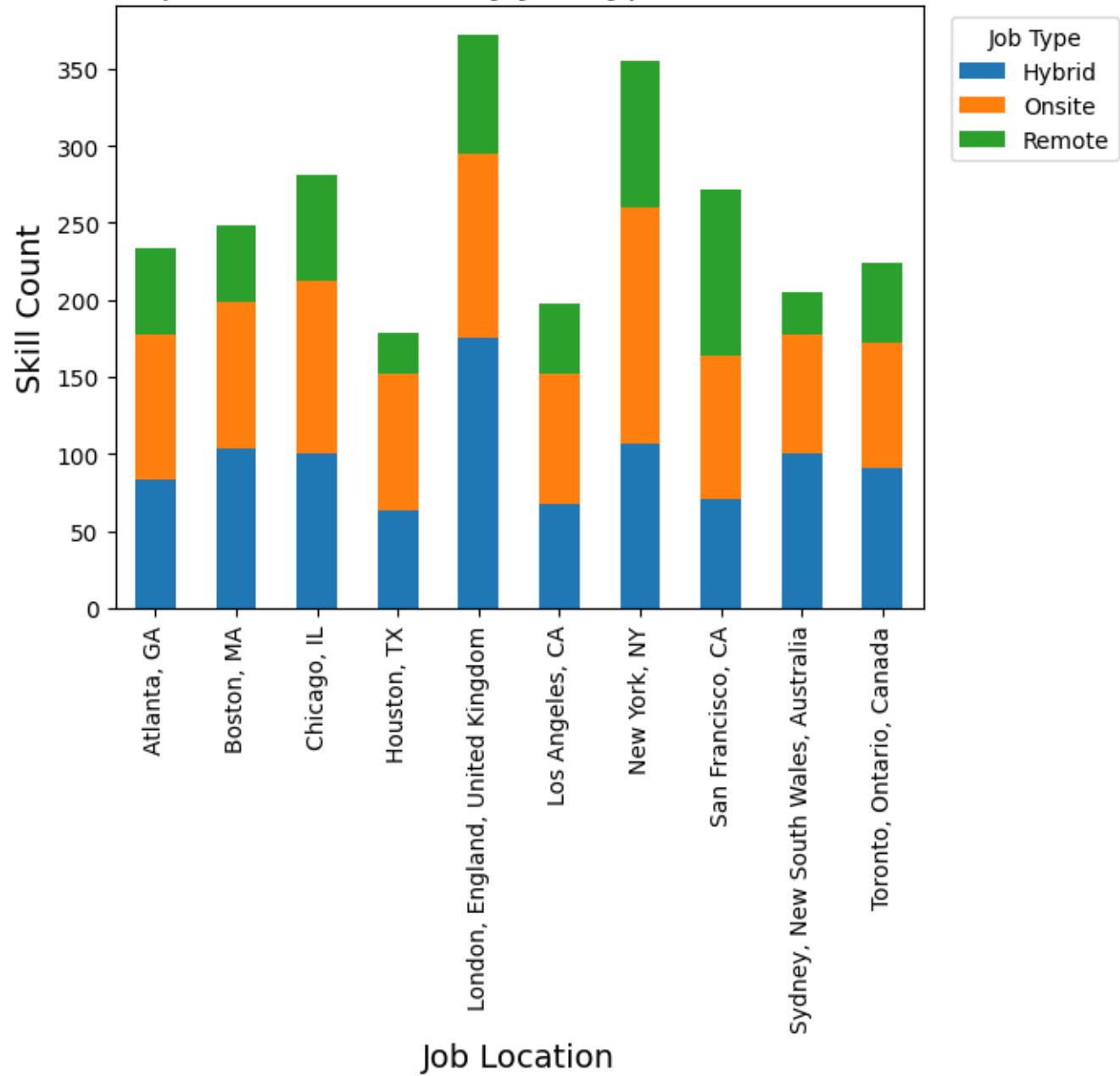


Top-20 job title from job postings





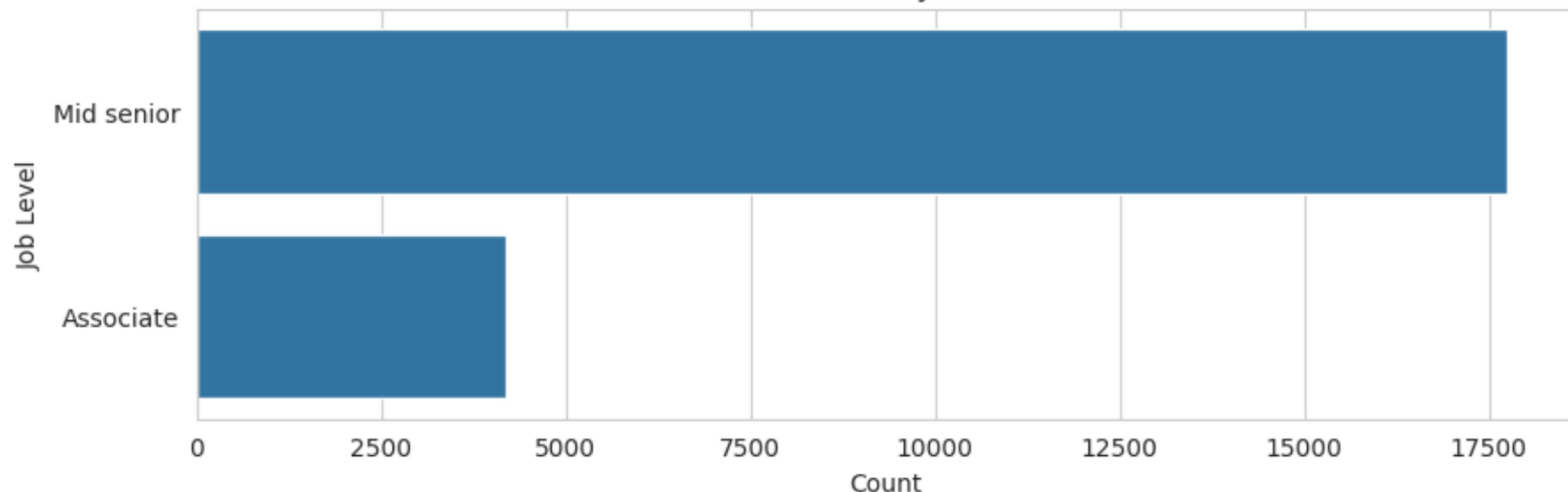
Top Skills Demand by Job Type and Location



Word Cloud for Job Skills for top 5 type of jobs



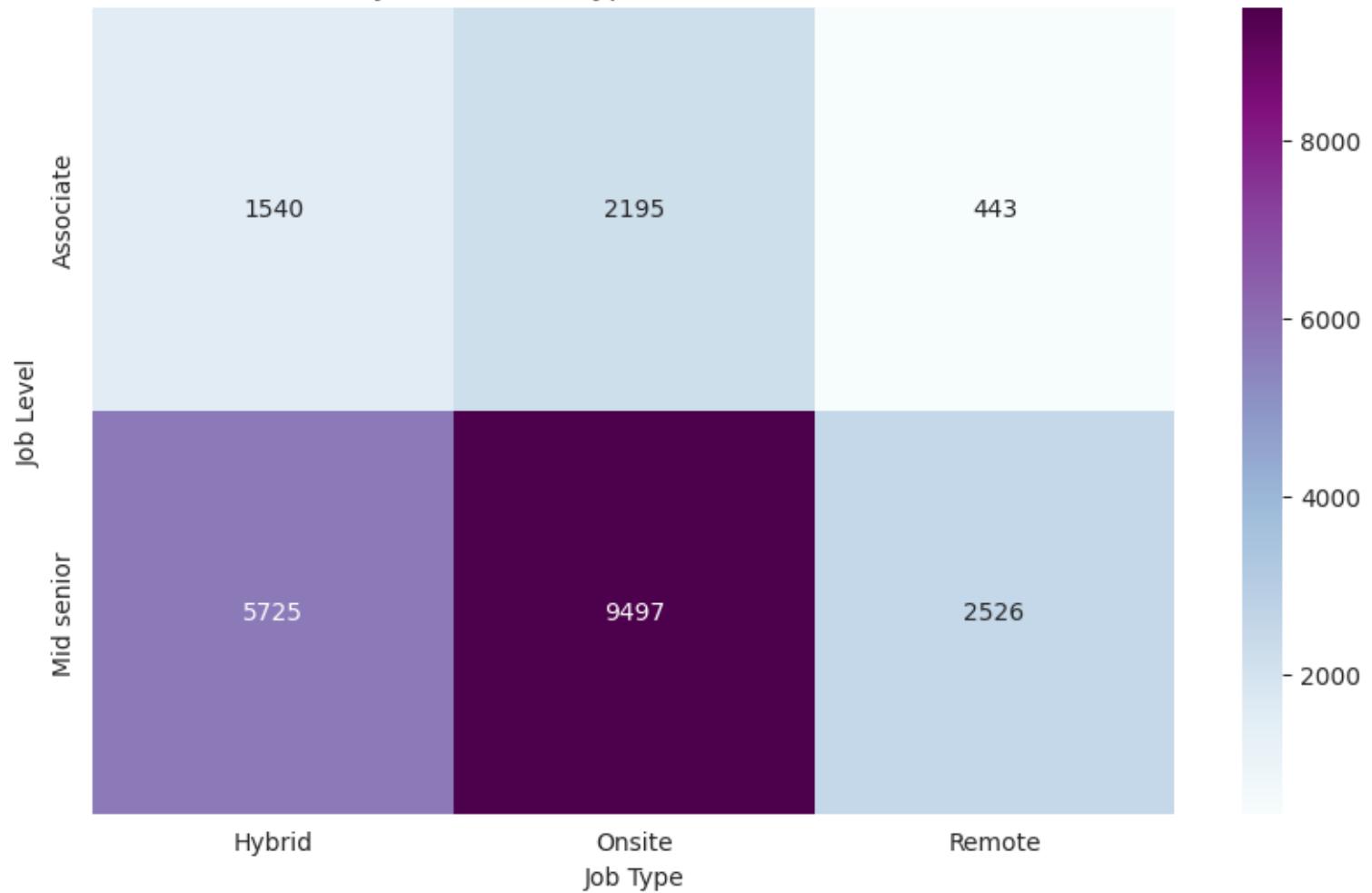
Distribution of Job Levels



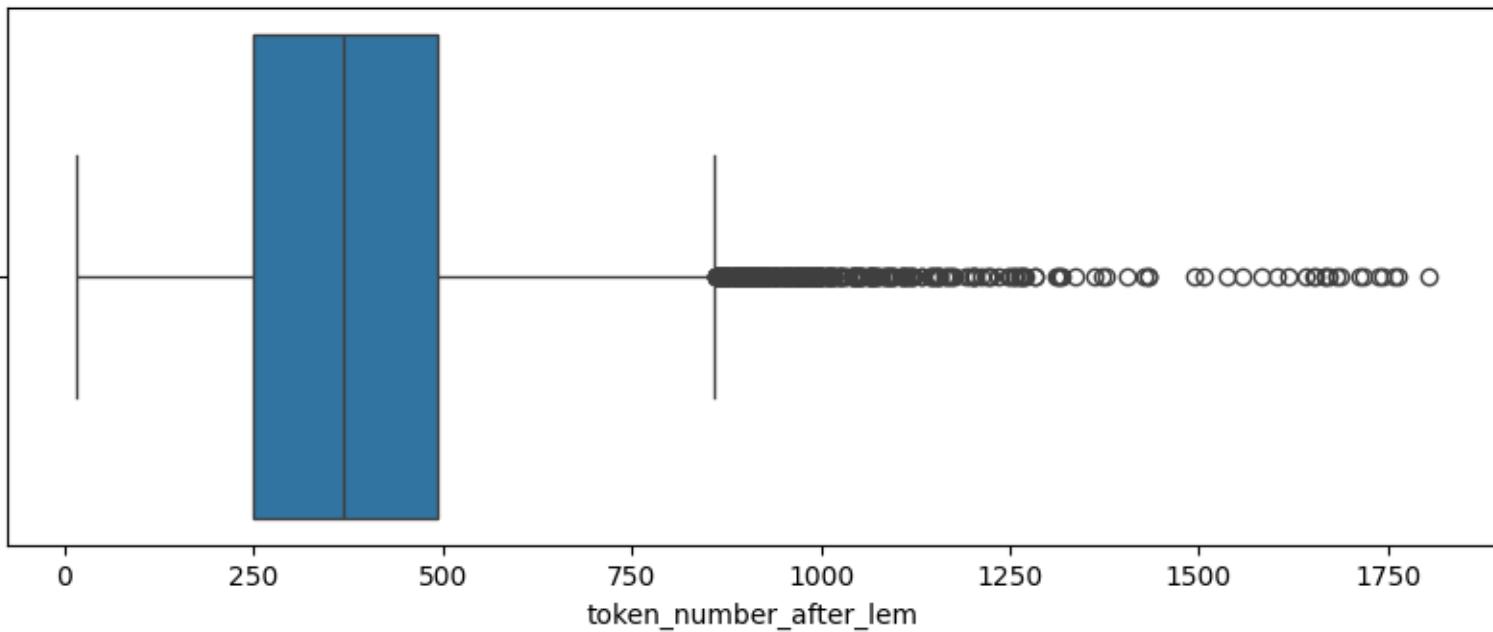
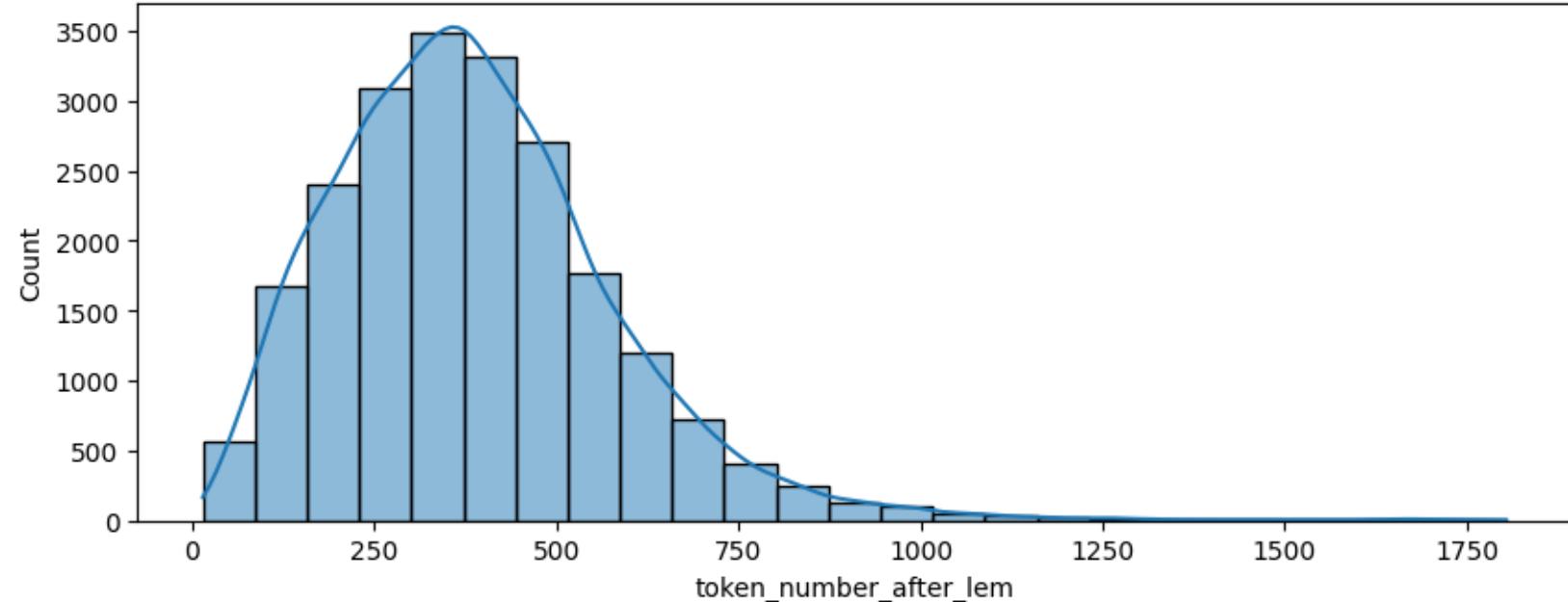
Word Cloud for Job Skills



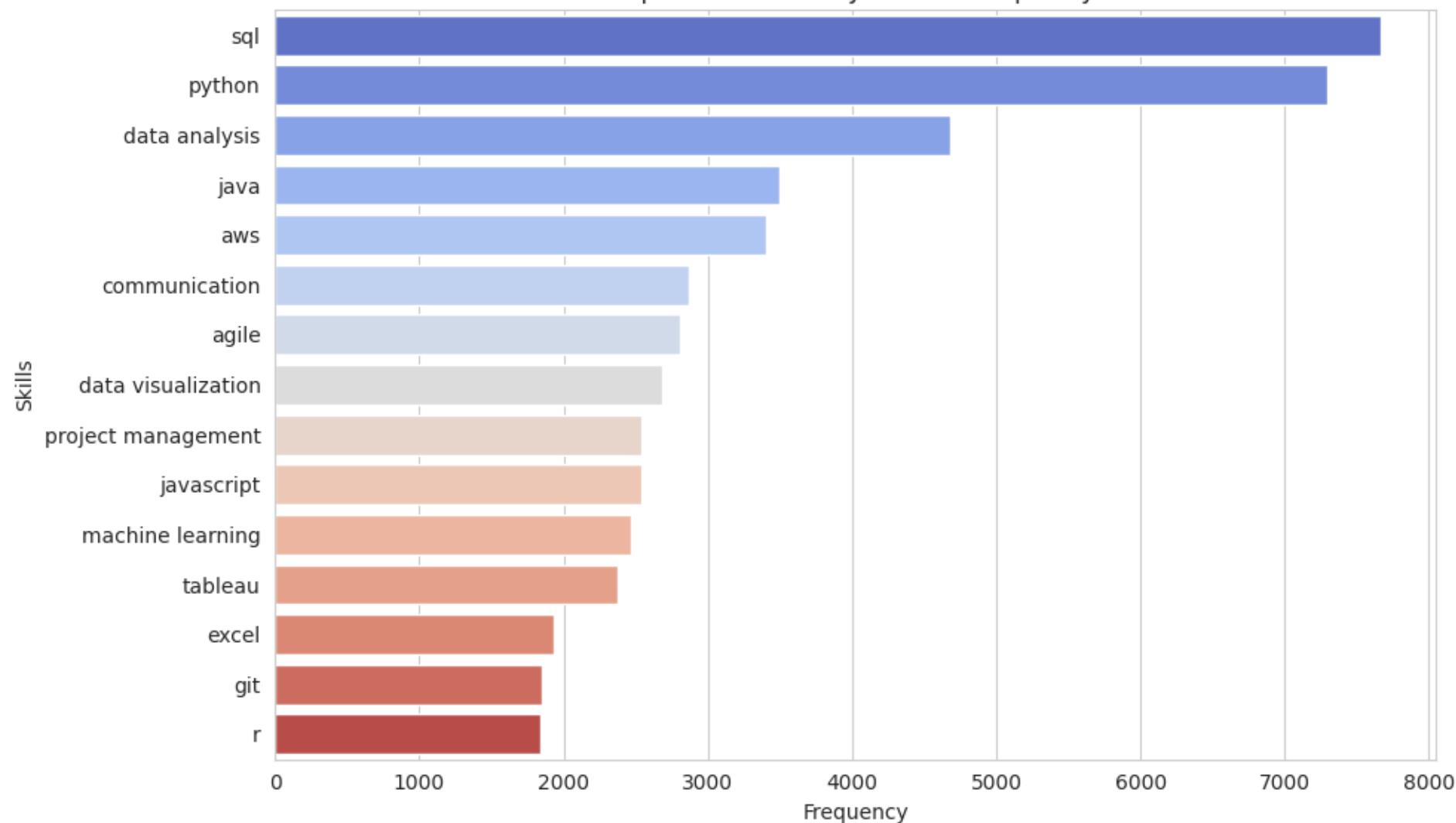
Job Level and Type Cross-tabulation

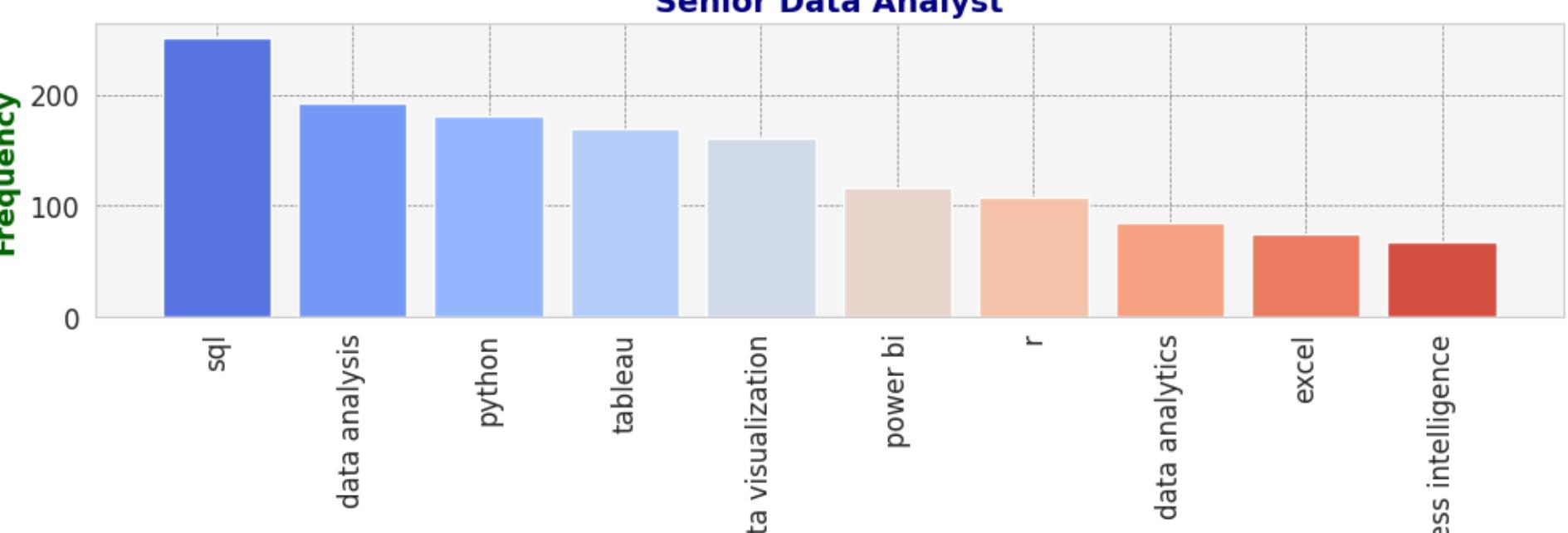
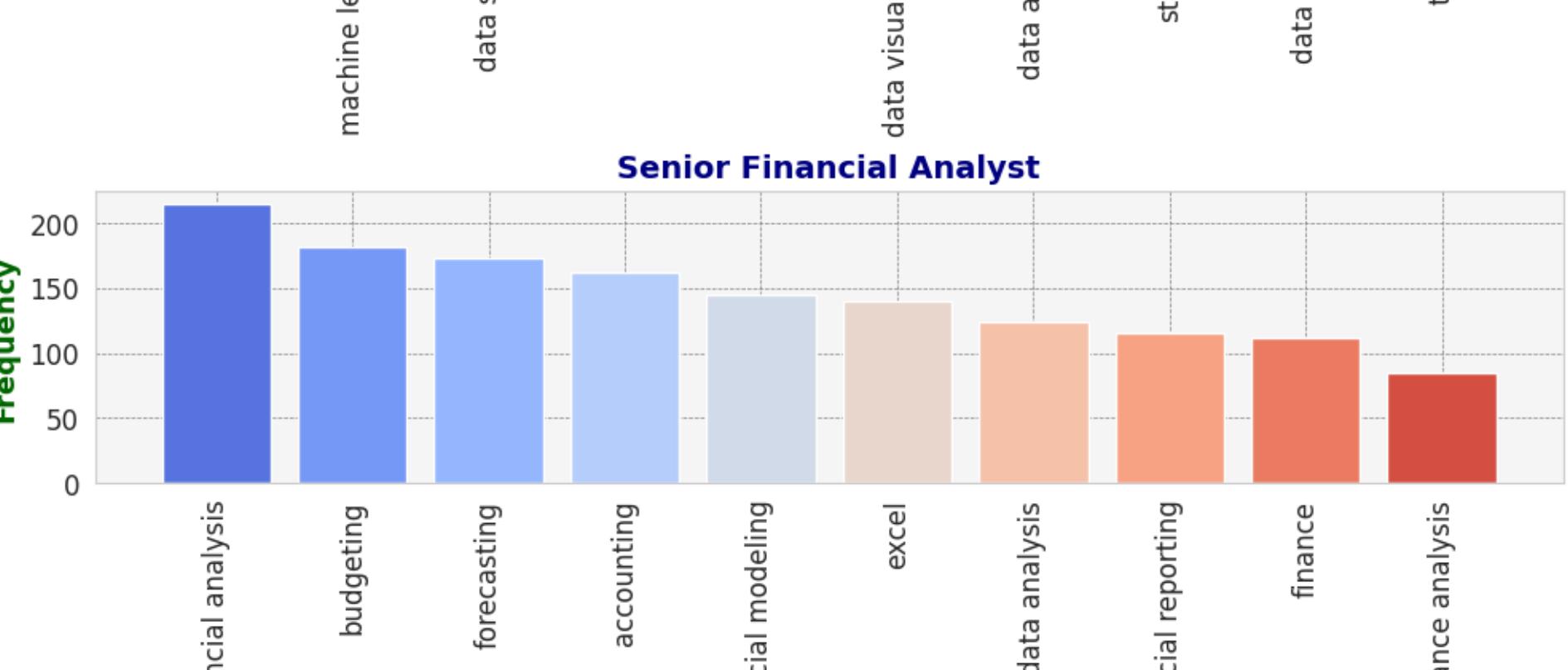
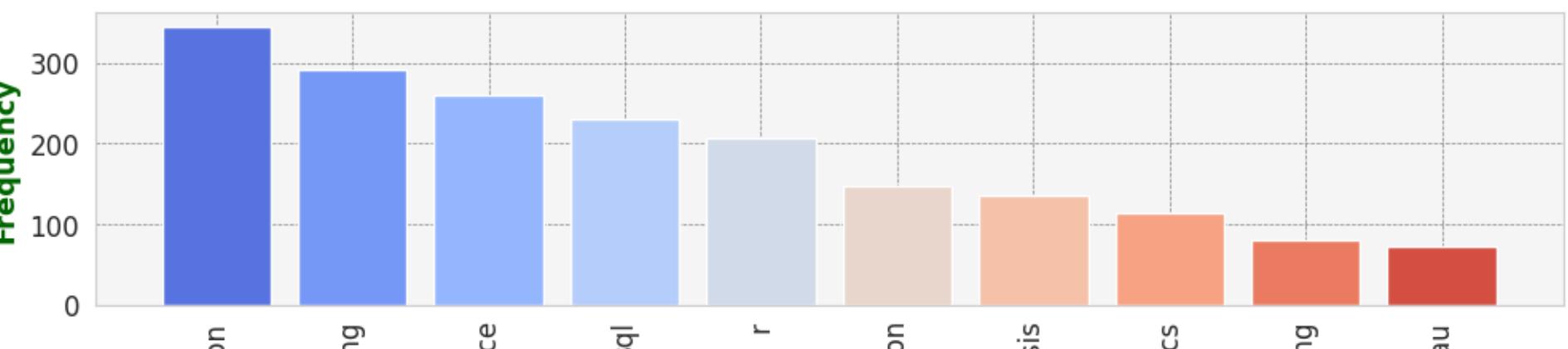
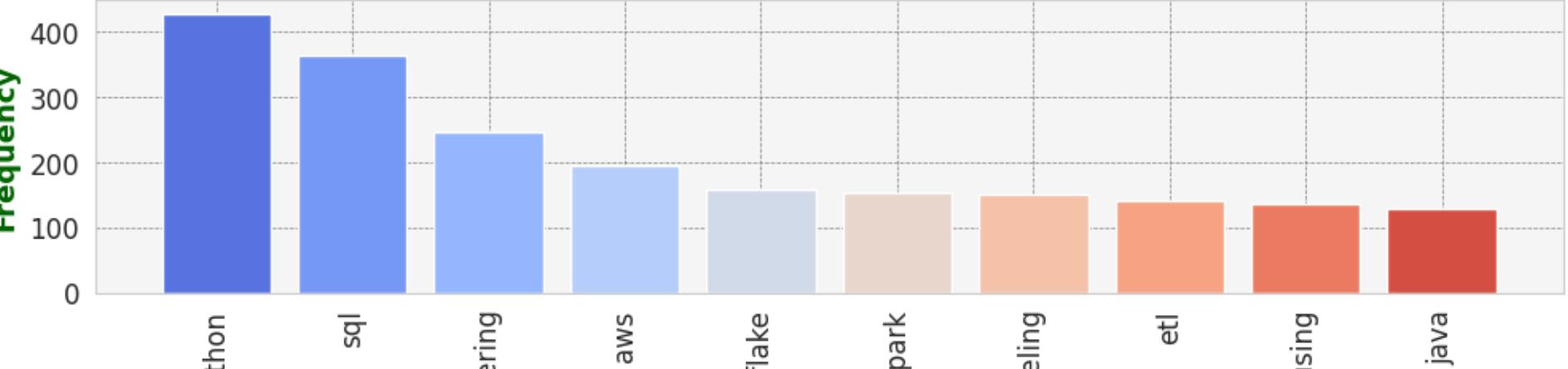
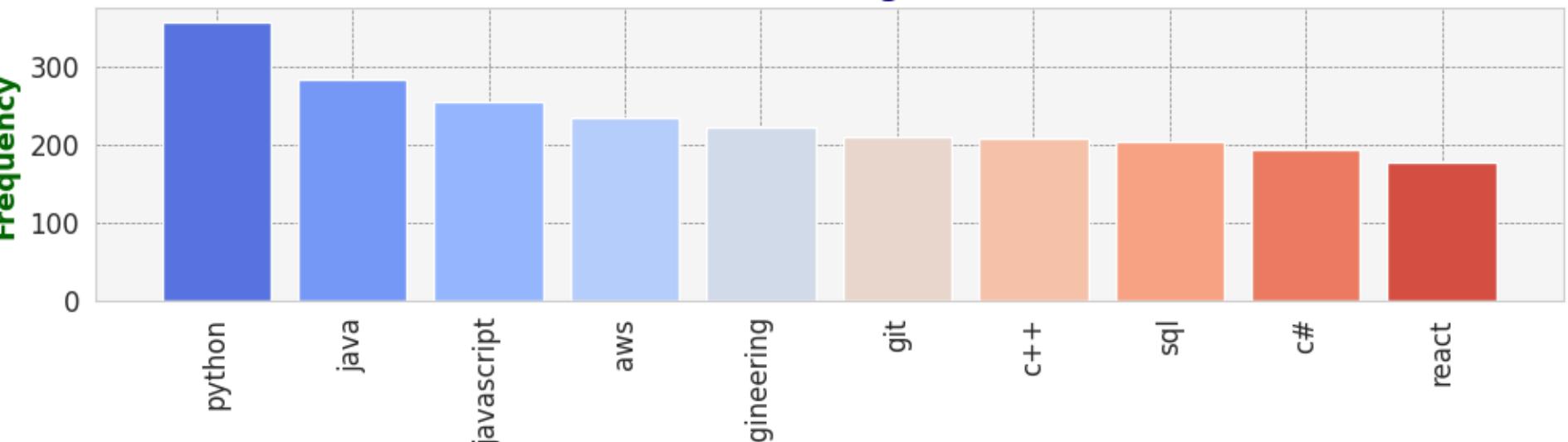
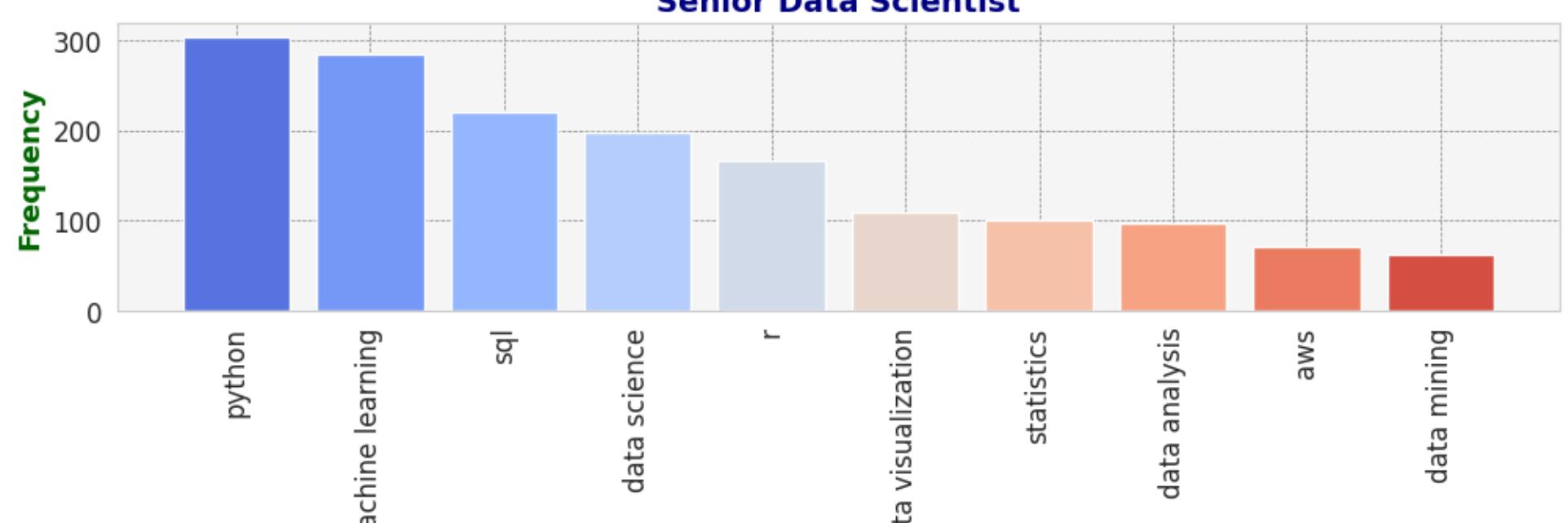
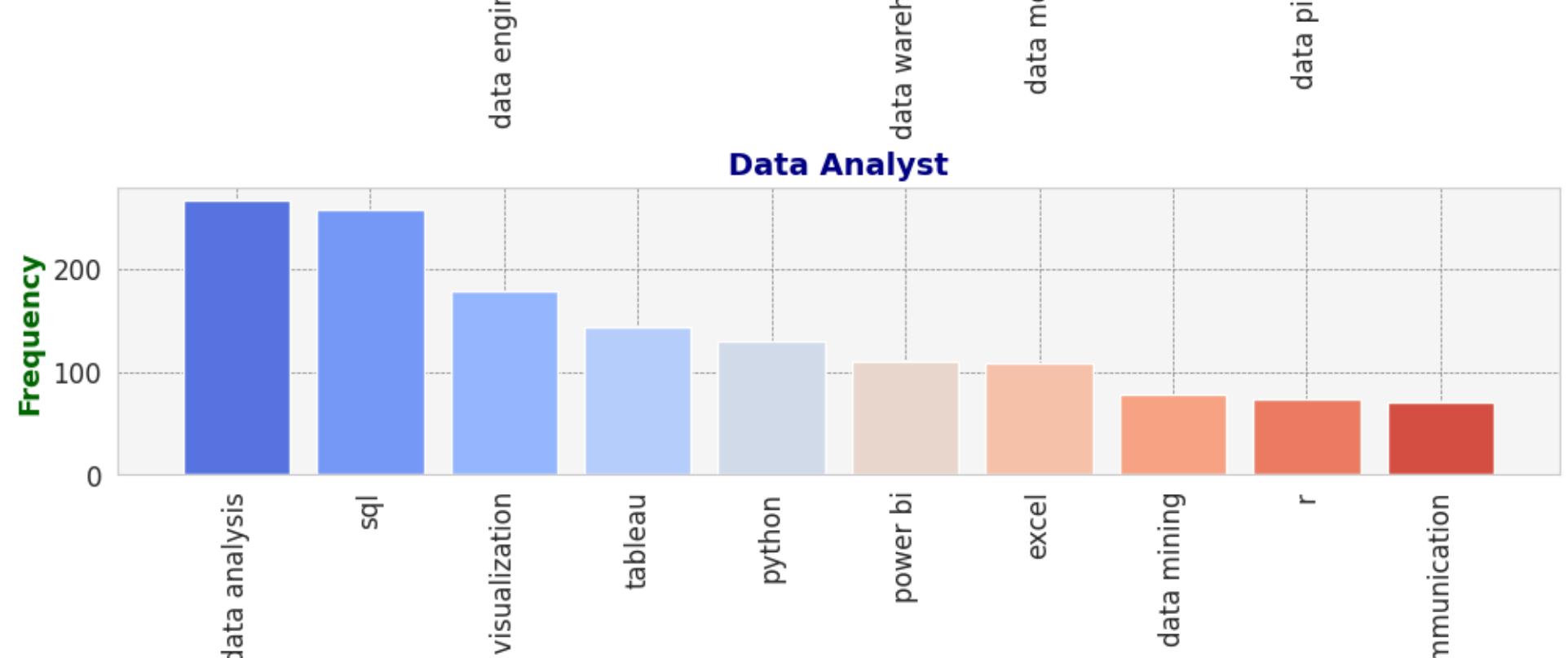
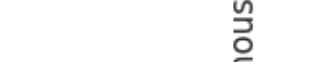
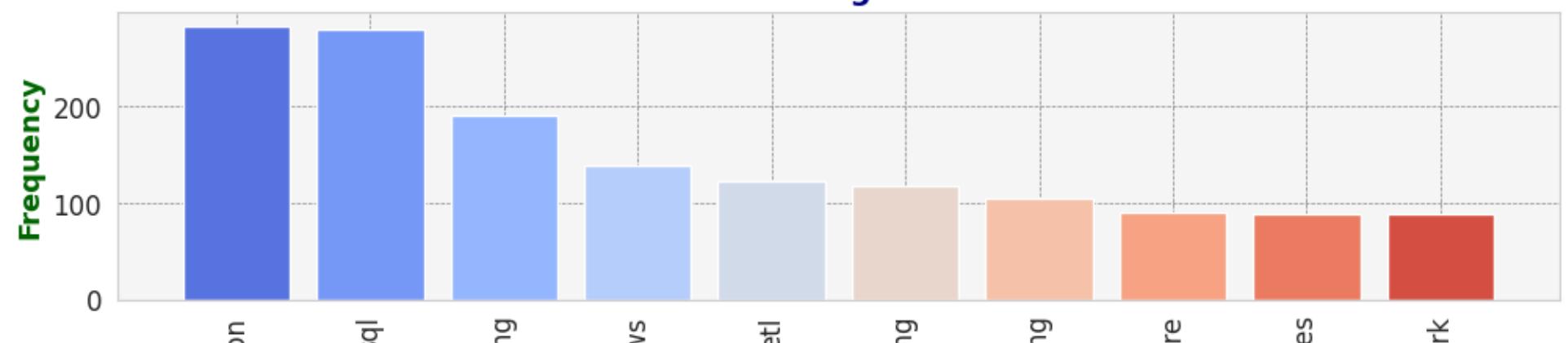
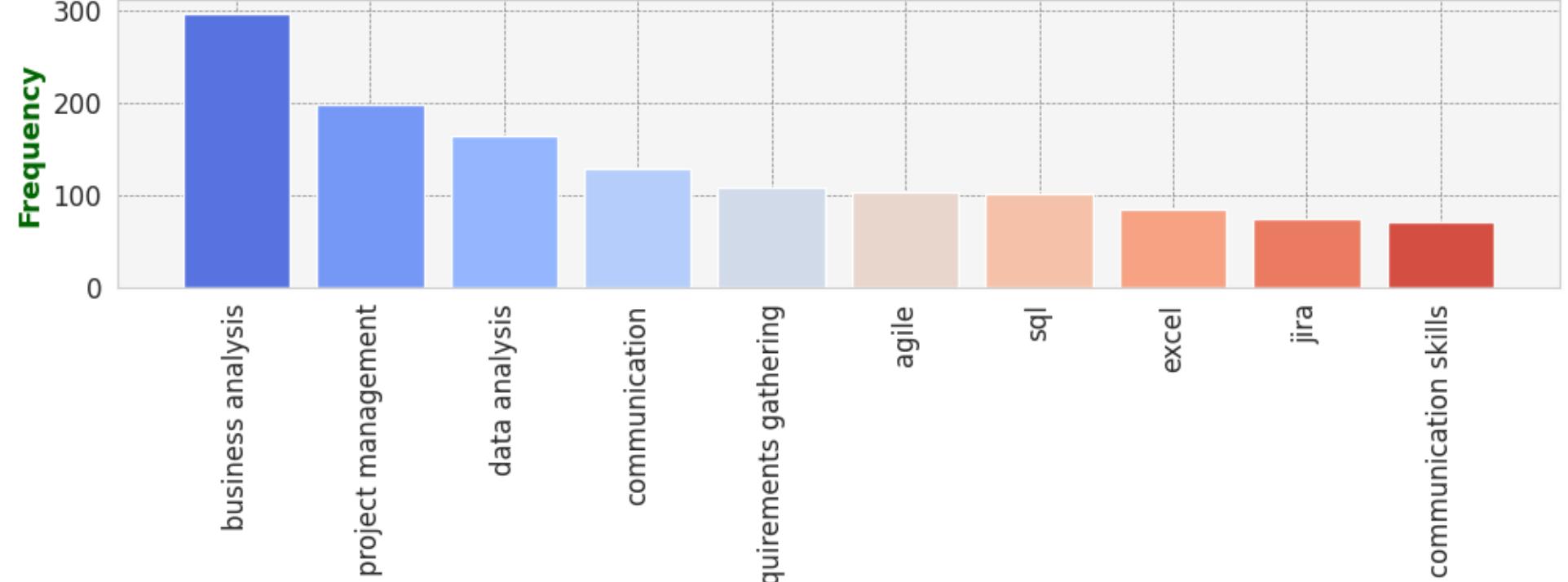
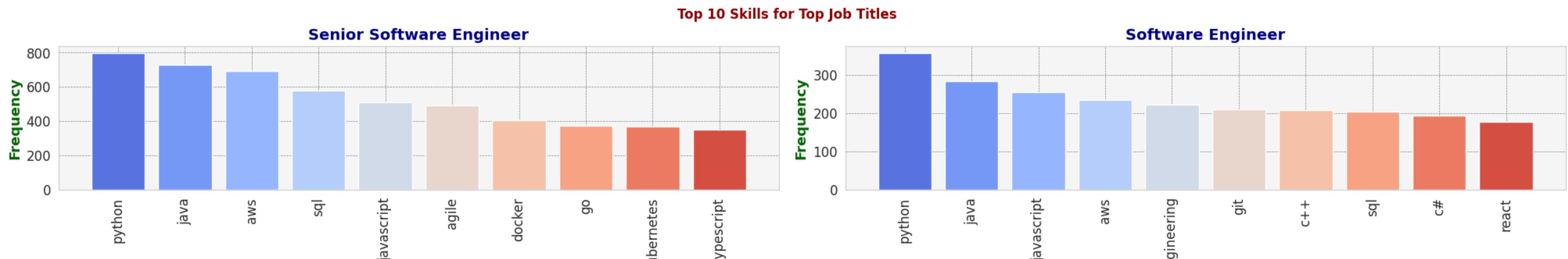


Distribution of word counts in job description

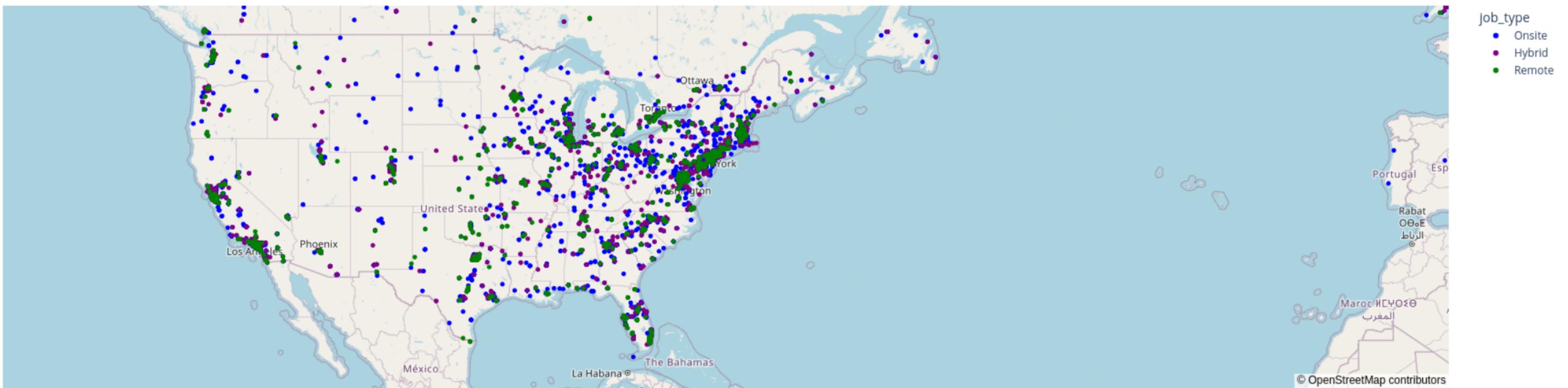


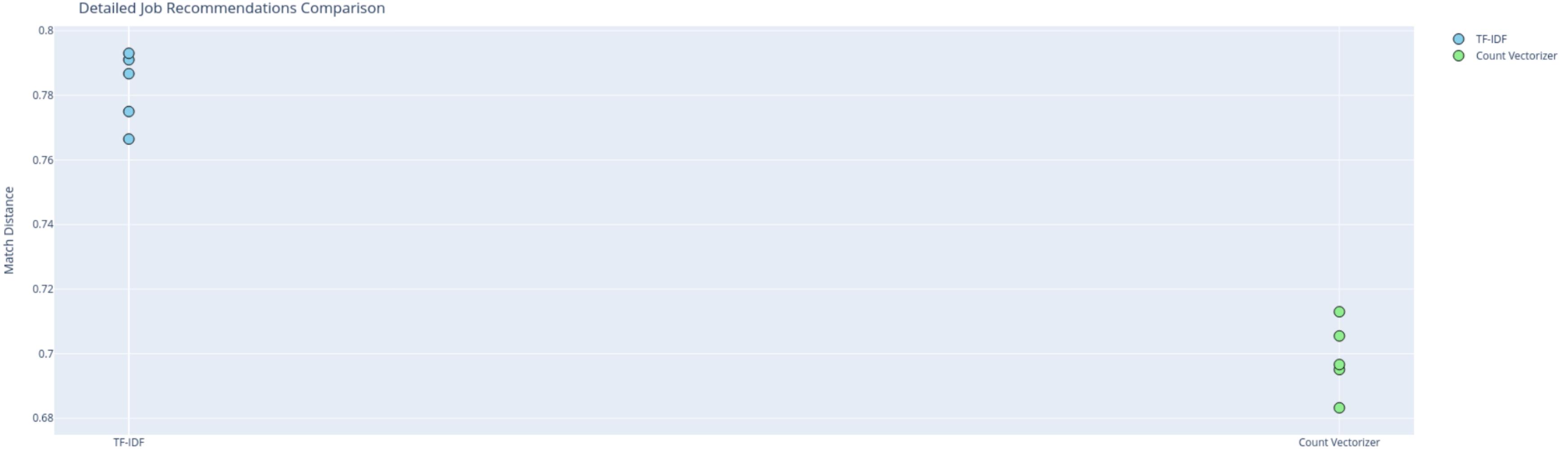
Top 10 Normalized Job Skills Frequency





Geographical Distribution of Remote vs Onsite Jobs





Job Postings and Resume in 2D t-SNE Space

