Way to Google Data Analysis

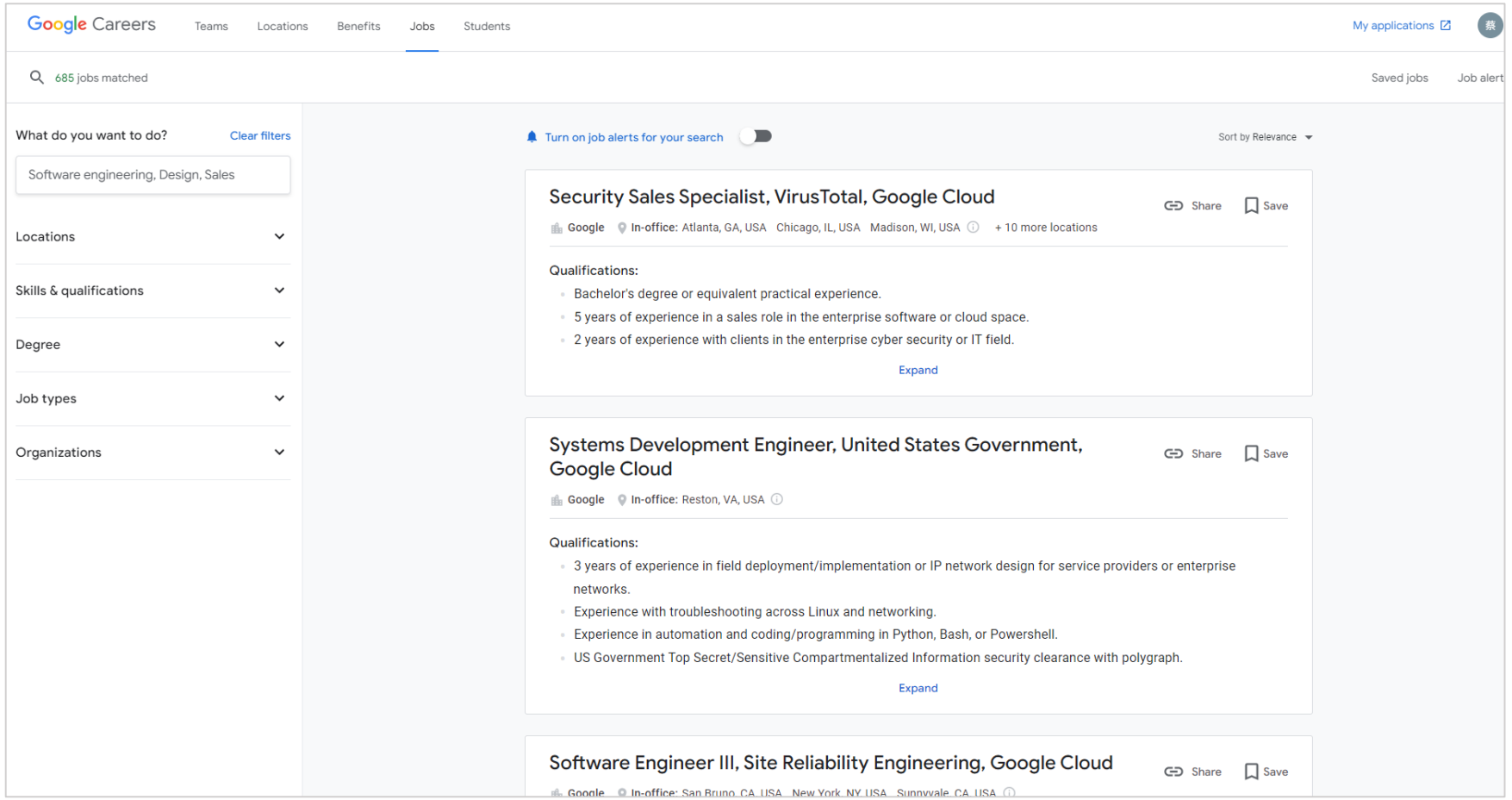
Python Data Analysis EDA Project

By Michael Tsai

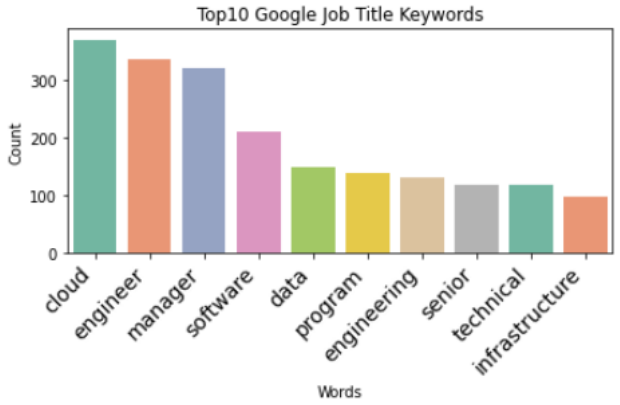
Quick View

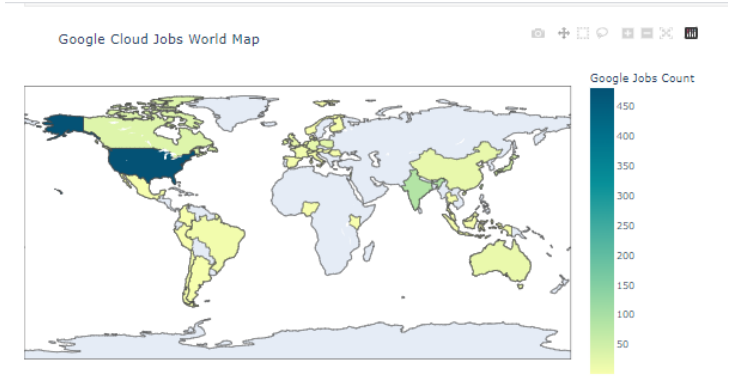
Demo of Project

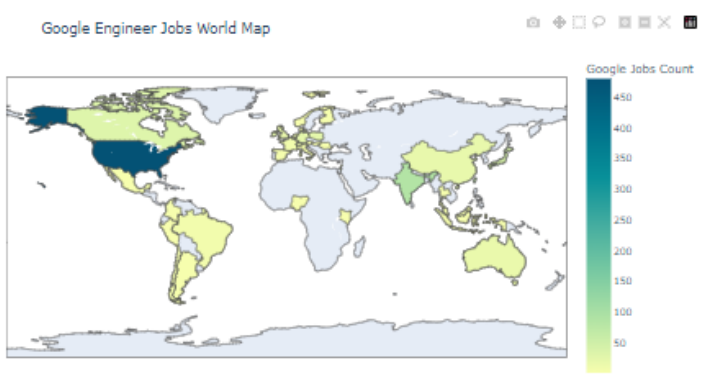
Quick View – From Google Jobs

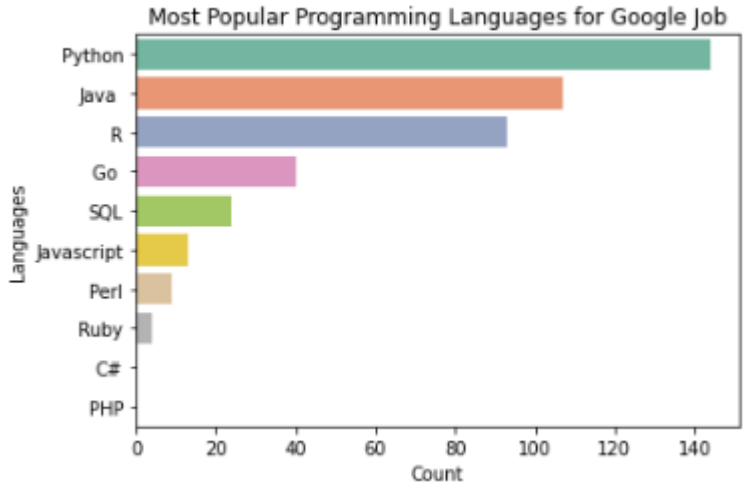


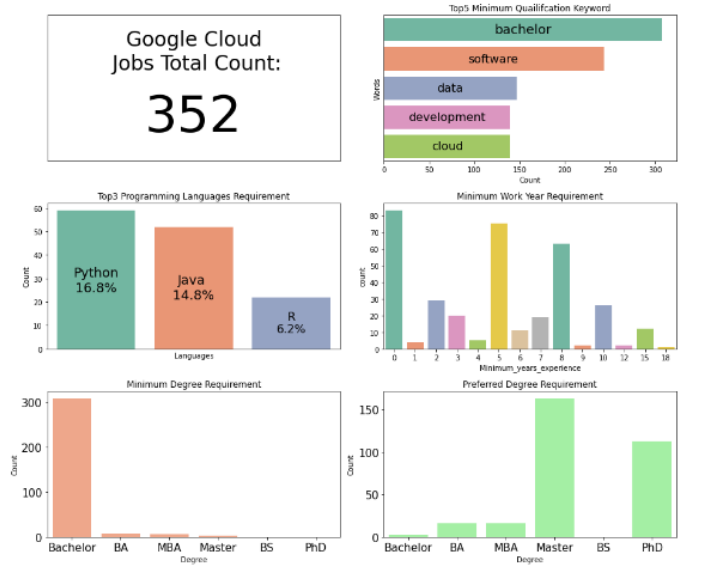
Quick View – To Jobs Analysis Dashboard

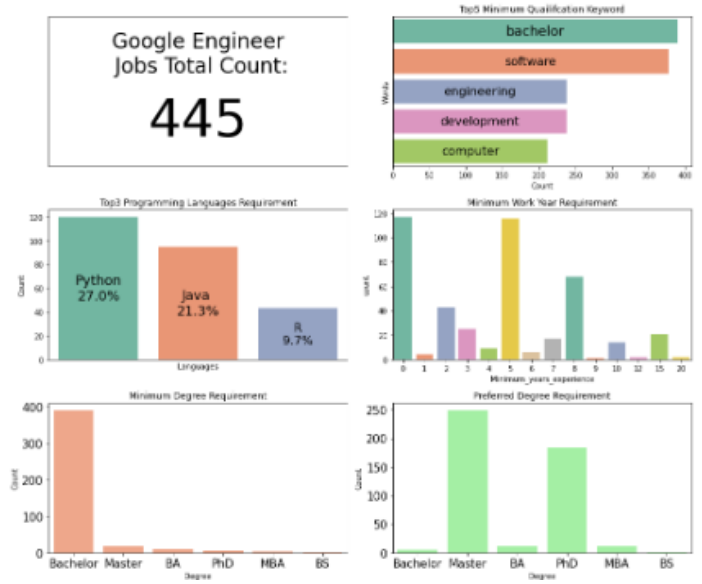












Summary

One page project summary

Way to Google!

Goolge Jobs Analysis Project

Purpose:

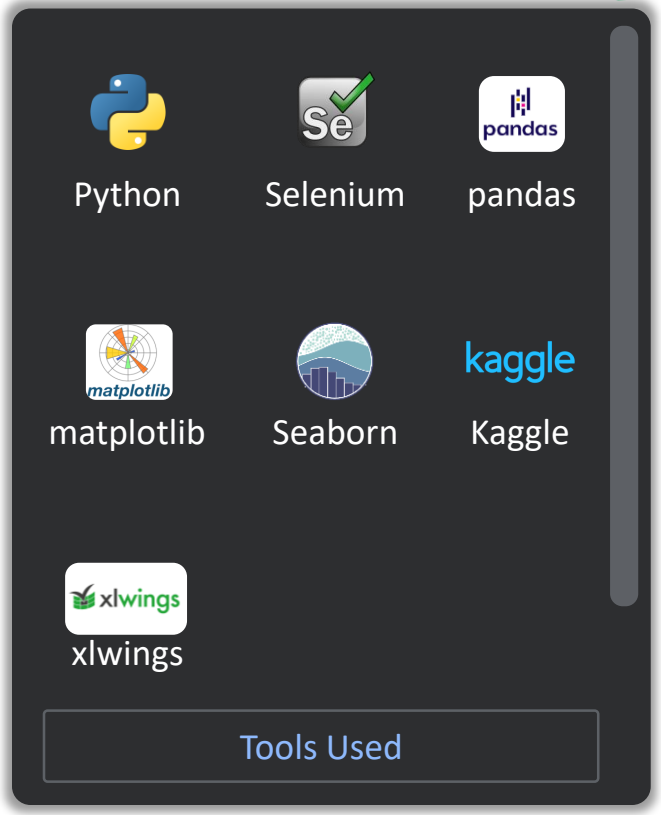
Google is my dream company. I believe lots of data analyst want to join Google, just like me. I wonder what will it require to work in Google.

Questions/Tasks:

- ☐ What Programming languages does Google request the most?
- ☐ What about the requirement for degrees?
- ☐ Does work experience important? If so, how many years work experience is generally required?
- ☐ What is the most popular job type in Google now?

* View on GitHub

Summary



Skills:

- Selenium web scraping
- pandas data cleaning &processing
- matplotlib data visualization

Data Preparation

Date Cleaning & Data Processing

Data Source

Data Name: Google Jobs All Information

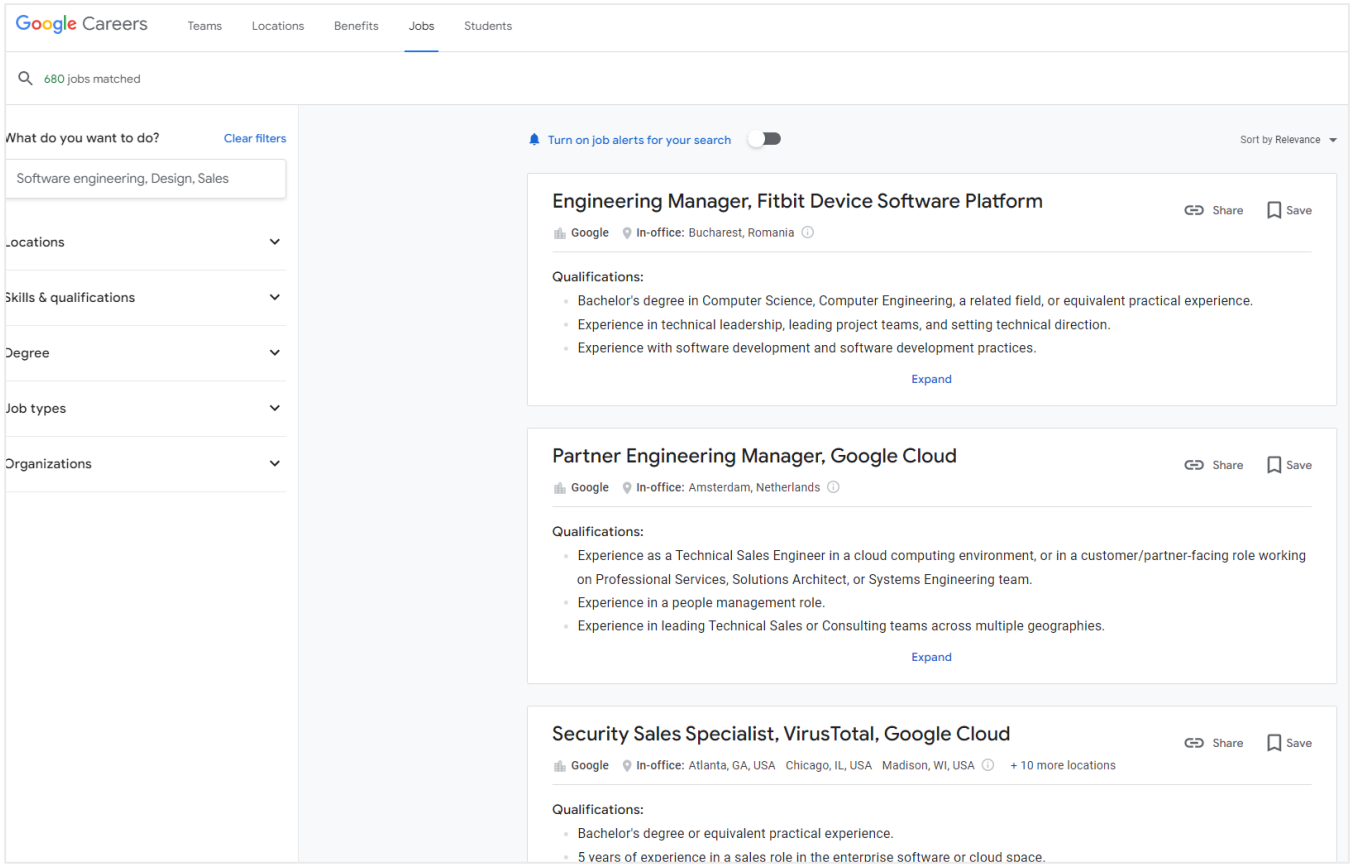
Source: [Google Careers Jobs]

Data Range: All the posts on January 21st 2023.

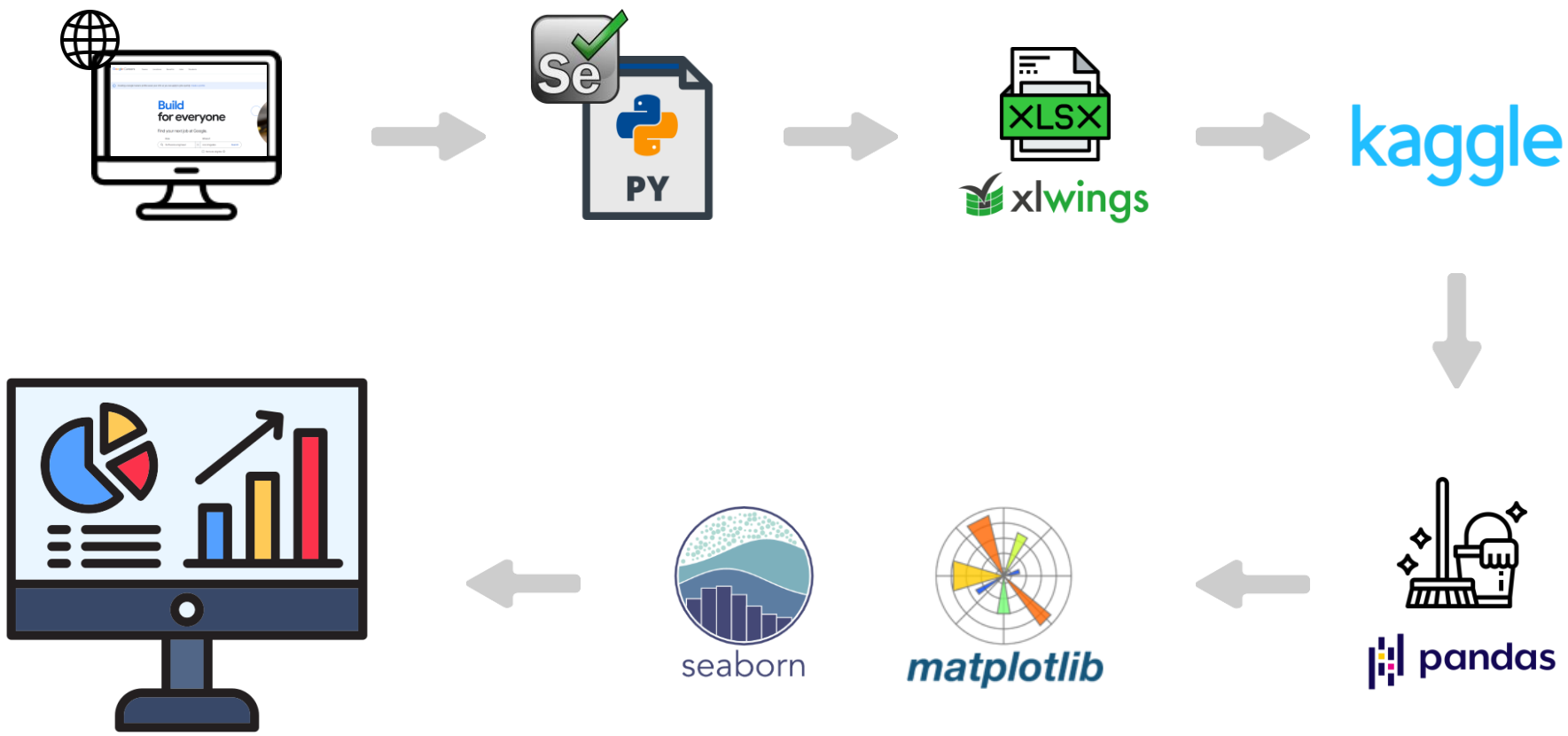
Data Size: 1K rows Data

Data information:

- Company
- Job title
- Location
- Post Date
- Link
- Minimum Qualificatoin
- Preferred Qualification
- Responsibilities
- About Job



Data Process



Python Code - Web Scraping

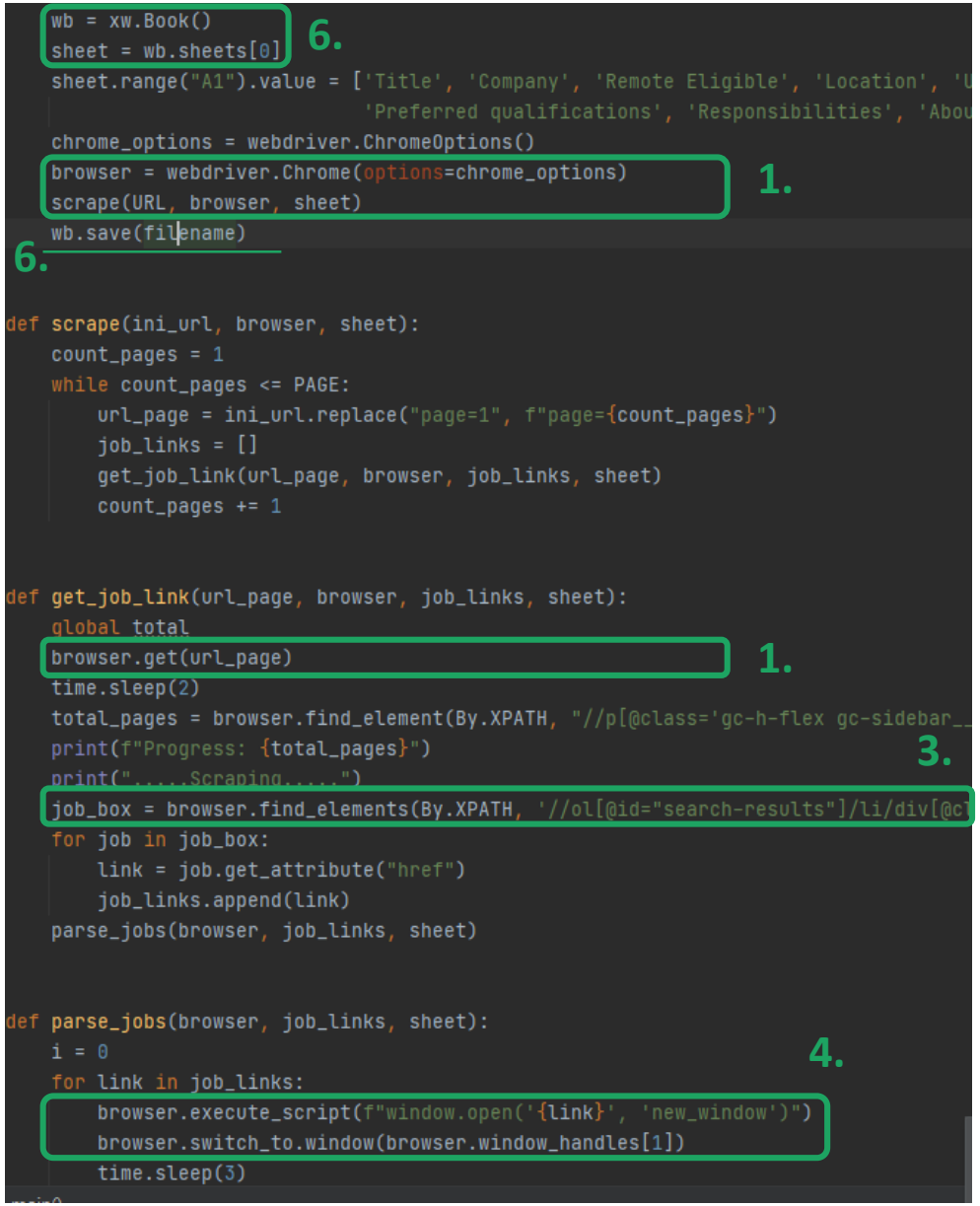
Target:

Simulate user enter Google careers website, scrap the information we need save the data in XLSX.

Python Script: Link

Key Skills:

1. Create ***Selenium Chrome driver*** and get in to assigned url
2. Pause by ***time.sleep()*** to avoid crash
3. Use Selenium ***find_elements() with X.path*** to find all the job link in current page
4. Create a new tab by ***Selenium executing java_script*** command.
5. Find information we need by Selenium find_elements() with X.path and add into dataframe
6. Use Xlwings to create XLSX file, write data in it and save it.



Python Code – Data Cleaning

Target:

Simulate user enter Google careers website, scrap the information we need save the data in XLSX.

Full Python Script on Kaggle: Link

Key Skills:

1. With ***str.contains()*** and ***re.findall()***, I can Identify and drop NA, incorrect and space only data.
2. Organize the string into correct format by using ***column.apply(lambda…)***
3. Extract the data (ex: country, keyword) we need from strings by using ***word_tokenize***, ***nltk.stopwords***, ***country_converter***…*etc* through ***.apply(lambda…)***

1.

Check if any NA in the datasets

```
pd.isnull(df).sum()
```

```
Title                      0
Company                    0
Location                   0
Update_Time                0
Minimum_Qualifications     0
Preferred_Qualifications   1
Responsibilities           0
About_Job                  0
Link                       0
dtype: int64
```

But, is it really only 1 null data? Let's check if there're sti

```
import re
df = df.replace(r'^\s+$', np.nan, regex=True)
pd.isnull(df).sum()
```

```
df = df.dropna(how="any", axis = "rows")
pd.isnull(df).sum()
```

```
Title                      0
Company                    0
Location                   0
Update_Time                0
Minimum_Qualifications     0
Preferred_Qualifications   0
Responsibilities           0
About_Job                  0
Link                       0
dtype: int64
```

2.

```
def string_manipulation(text):
    text = str(text).replace("\t", "").replace("\n", " ").replace("\r", "").replace("¡¦s", "'s'")
    return text

df["Location"] = df.Location.apply(lambda x : string_manipulation(x))
df["Update_Time"] = df.Update_Time.apply(lambda x : x[:10])
df["Update_Time"] = pd.to_datetime(df["Update_Time"])
df["Minimum_Qualifications"] = df.Minimum_Qualifications.apply(lambda x : string_manipulation(x))
df["Preferred_Qualifications"] = df.Preferred_Qualifications.apply(lambda x : string_manipulation(x))
df["Responsibilities"] = df.Responsibilities.apply(lambda x : string_manipulation(x))
df.head()
```

3.

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

stop_words = set(stopwords.words('english'))
exclude_list = ["years", "experience", "degree", "equivalent", "practical", "technical",
"role"]
df['Responsibilities'] = df.Responsibilities.apply(lambda x: word_tokenize(x))
df['Responsibilities'] = df.Responsibilities.apply(lambda x: [w for w in x if w not in s
top_words])
df['Responsibilities'] = df.Responsibilities.apply(lambda x: ' '.join(x))

df['Minimum_Qualifications'] = df.Minimum_Qualifications.apply(lambda x: word_tokenize
(x))
df['Minimum_Qualifications'] = df.Minimum_Qualifications.apply(lambda x: [w for w in x i
f w not in stop_words and w.lower() not in exclude_list])
df['Minimum_Qualifications'] = df.Minimum_Qualifications.apply(lambda x: ' '.join(x))

df['Preferred_Qualifications'] = df.Preferred_Qualifications.apply(lambda x: word_tokeni
ze(x))
df['Preferred_Qualifications'] = df.Preferred_Qualifications.apply(lambda x: [w for w in
x if w not in stop_words])
df['Preferred_Qualifications'] = df.Preferred_Qualifications.apply(lambda x: ' '.join
(x))
df.head()
```

Python Code – Data Analysis and Visualization (1/3)

Target: Analyze the data and visualize the insight

Full Analysis Process on Kaggle : Link

Key Skills:

1. Using ***Counter()*** from collections module to count **the keyword appearance** from mass data for further analysis, such as degrees, programming languages, title keyword…
2. Using ***df.merge()*** to join related table and showing chart with ***plt.subplots()*** for easy compare.
3. Using **Seaborn** module to generate more recognizable and visually appealing charts
4. Using ***plotly.choroplethplotl()*** to generate geo map.
5. Defining a **huge function** to allow user to **generate a dashboard** based on the assigned position keyword, which will contain all the statistical data related to the job.

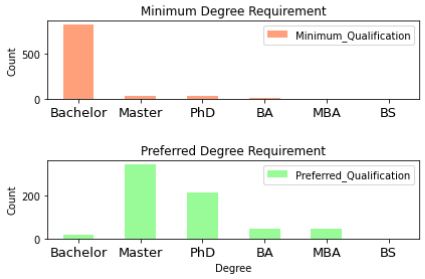
1.

```
programming_languages = ['Python', 'Java ','C#', 'PHP', 'Javascript', 'Ruby', 'Perl', 'SQL', 'Go ', "R"]
languages = {}
for pl in programming_languages:
    count = df['Minimum_Qualifications'].str.contains(pl).sum()
    languages[pl] = count
languages = sorted(languages.items(), key=lambda x: x[1], reverse=True)
print(languages)
```

```
[('Python', 144), ('Java ', 107), ('R', 93), ('Go ', 40), ('SQL', 24), ('Javascript', 13), ('Perl', 9), ('Ruby', 4), ('C#', 0), ('PHP', 0)]
```

2.

```
fig, axes = plt.subplots(2,1)
df_degree.plot.bar(x="Degree", y="Minimum_Qualification", ax=axes[0], color="lightsalmon", rot=0)
df_degree.plot.bar(x="Degree", y="Preferred_Qualification", ax=axes[1], color="palegreen", rot=0)
axes[0].title.set_text("Minimum Degree Requirement")
axes[0].set_ylabel("Count")
axes[0].set_xlabel(" ")
axes[0].tick_params(axis='x', labelsize=13)
axes[1].title.set_text("Preferred Degree Requirement")
axes[1].set_ylabel("Count")
axes[1].tick_params(axis='x', labelsize=13)
fig.tight_layout()
```



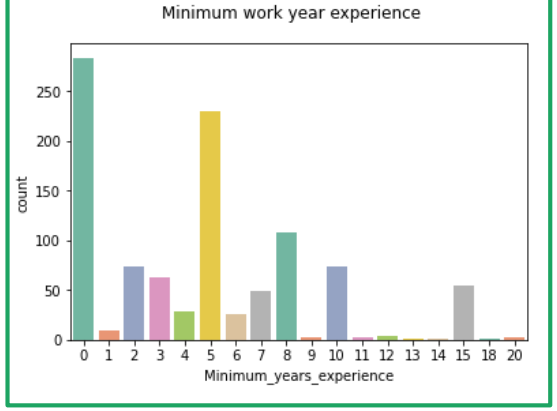

```
df_degree = df_mini.merge(df_pre, on=["Degree"])
df_degree
```

	Degree	Minimum_Qualification	Preferred_Qualification
0	Bachelor	821	17
1	Master	40	346
2	PhD	33	213
3	BA	14	44
4	MBA	7	44
5	BS	2	1

3.

```
import seaborn as sns
sns.countplot(x=df["Minimum_years_experience"]
              ,palette="Set2")
plt.suptitle('Minimum work year experience')
```

```
Text(0.5, 0.98, 'Minimum work year experience')
```



Python Code – Data Analysis and Visualization (2/3)

Target: Analyze the data and visualize the insight

Full Analysis Process on Kaggle : Link

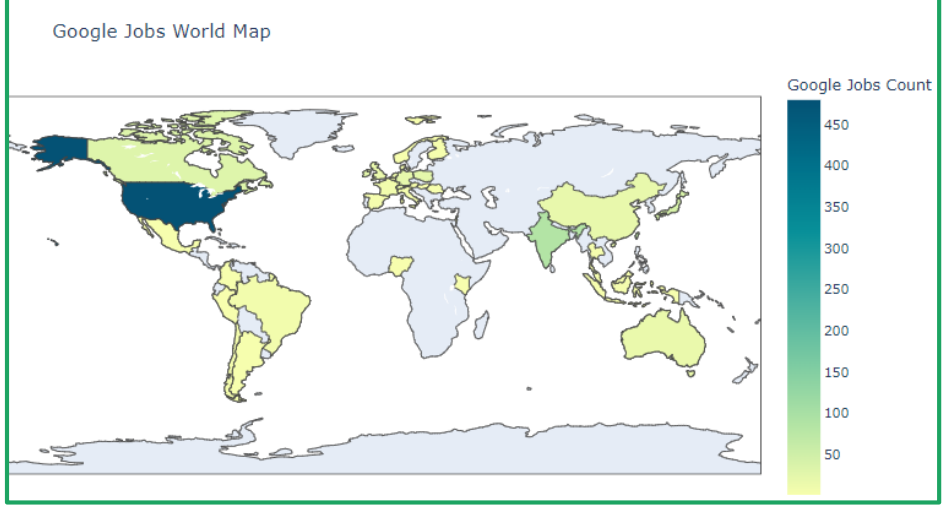
Key Skills:

1. **Counting the keyword occurrence** from mass data for further analysis, such as degrees, programming languages…
2. Using ***df.merge()*** to join related table and showing chart with ***plt.subplots()*** for easy compare.
3. Using **Seaborn** module to generate more recognizable and visually appealing charts
4. Using ***plotly.choroplethplotl()*** to generate geo map to display job allocation.
5. Defining a **huge function** to allow user to **generate a dashboard in one picture** based on the assigned position keyword, which will contain all the statistical data related to the job.

4.

```
import plotly.express as px
database = px.data.gapminder().query('year == 2007')
df_country_list = pd.merge(database, df_country, how='inner', on='country')
url = (
    "https://raw.githubusercontent.com/python-visualization/folium/master/examples/data"
)

fig = px.choropleth(df_country_list,
                    locations="country",#"iso_alpha",
                    locationmode="country names",#"ISO-3",
                    geojson = f"{url}/world-countries.json",
                    color="count",
                    color_continuous_scale="Bluyl",
                    labels={'count':'Google Jobs Count'},
                    title=f"Google Jobs World Map",
                   )
fig.update_layout(
autosize=False,
width=850,
height=500,
margin={"r":0,"t":50,"l":0,"b":50, "pad": 4})
fig.show()
```



Python Code – Data Analysis and Visualization (3/3)

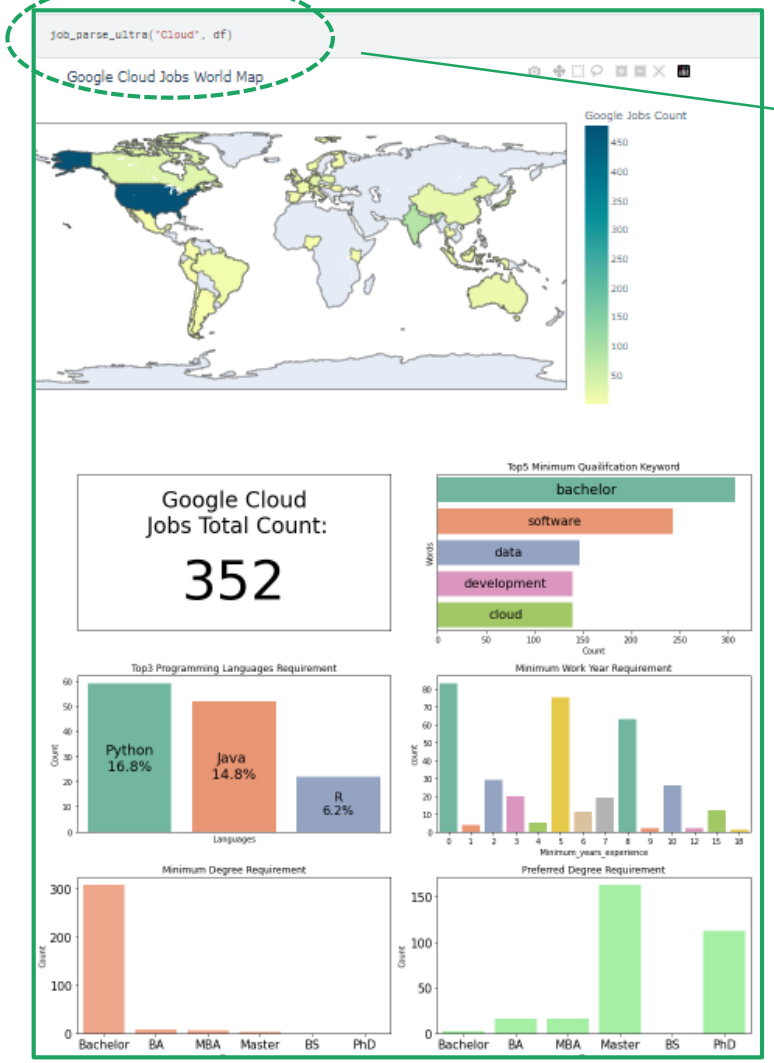
Target: Analyze the data and visualize the insight

Full Analysis Process on Kaggle : Link

Key Skills:

1. **Counting the keyword occurrence** from mass data for further analysis, such as degrees, programming languages…
2. Using ***df.merge()*** to join related table and showing chart with ***plt.subplots()*** for easy compare.
3. Using **Seaborn** module to generate more recognizable and visually appealing charts
4. Using ***plotly.choroplethplotl()*** to generate geo map to display job allocation.
5. Defining a **huge function** to allow user to **generate a dashboard in one picture** based on the assigned position keyword, which will contain all the statistical data related to the job.

5.




```
job_parse_ultra("Cloud", df)
```

Insight

Insight Finding during EDA

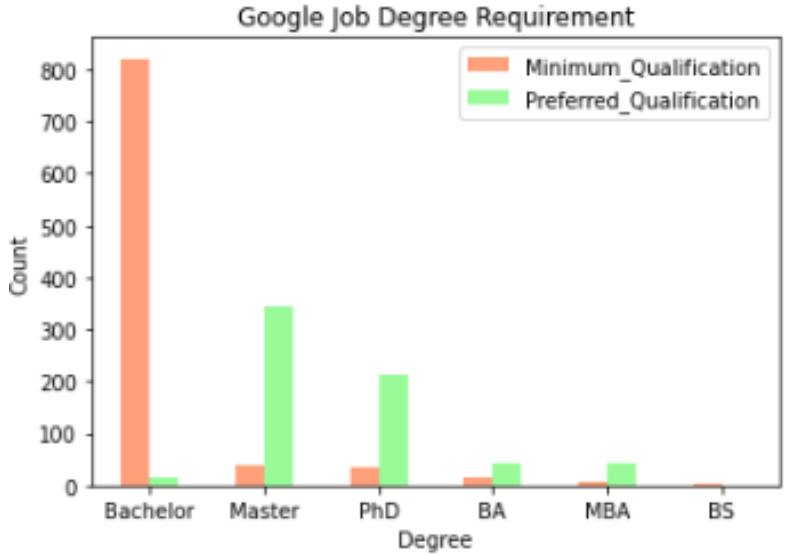
Insight 1 – Degree & Programming Languages

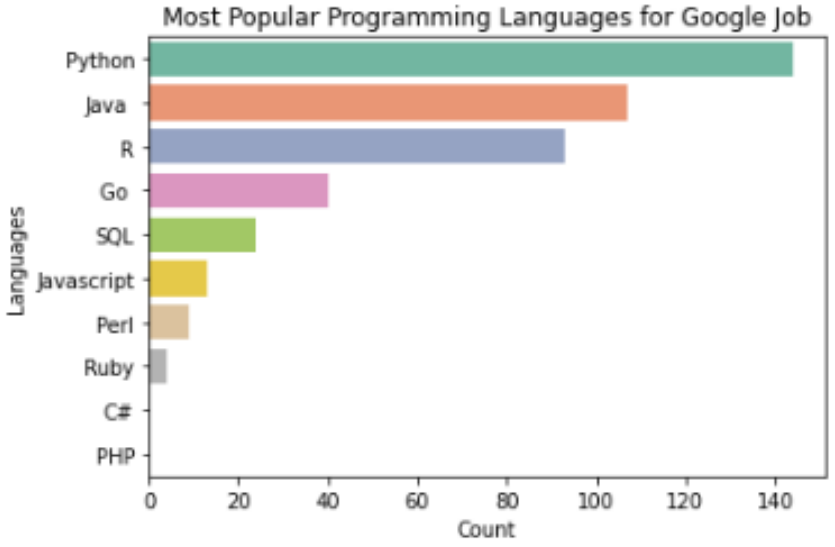
Degree Requirement:

You need **at least bachelor degree** to get in Google. **Master and MBA** make you **more competitive** than other candidates.

Programming Languages Requirement:

Python, **Java**, and **R** are the 3 most popular programming languages. Furthermore, the popularity of these 3 programming languages are significantly ahead of other languages.

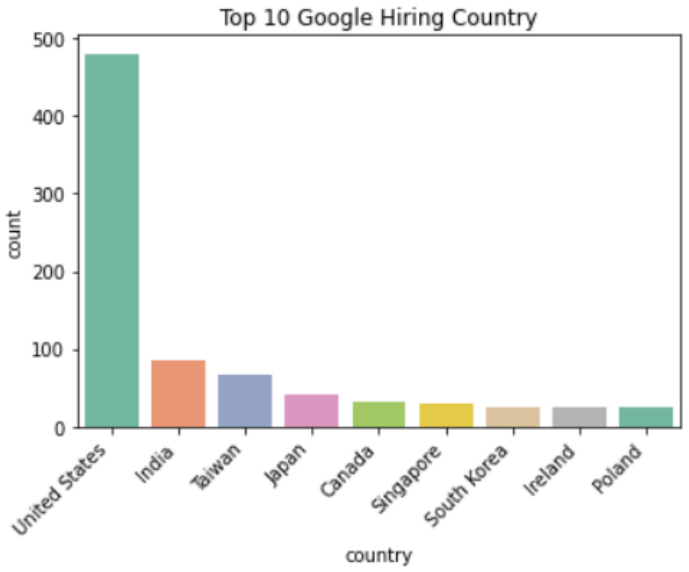


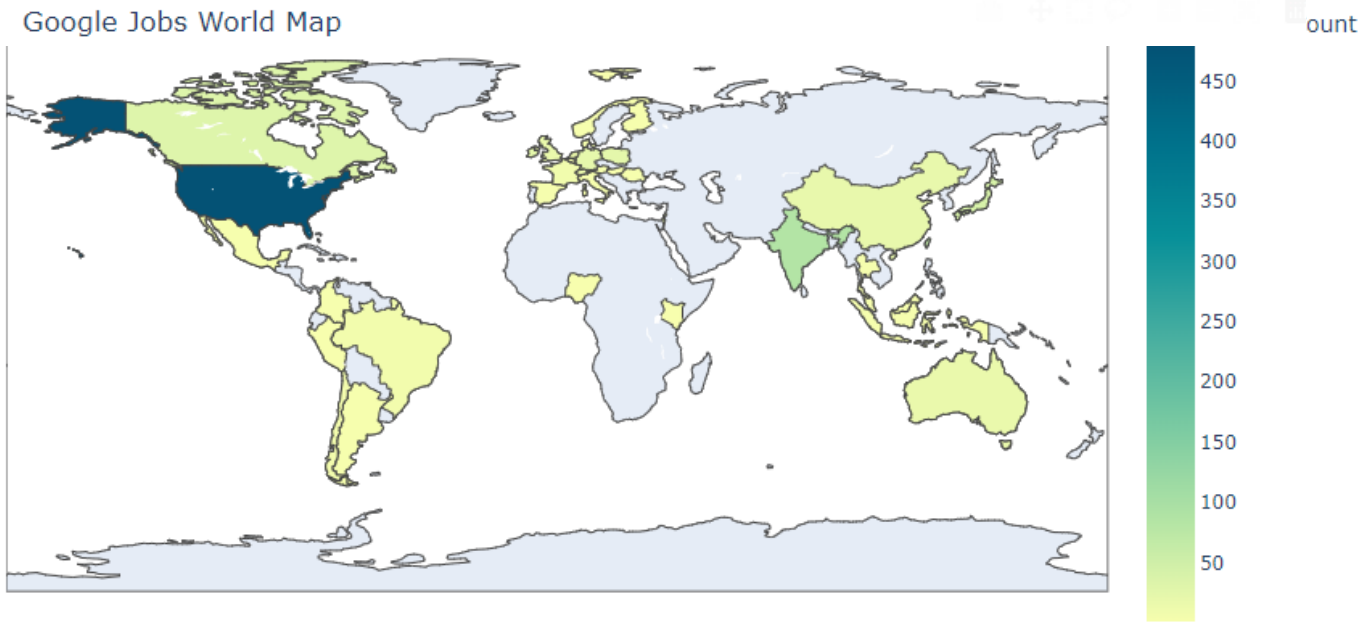


Insight 2 – Recruiting Location

In Google's global recruitment, **the United States stands out significantly** from other countries.

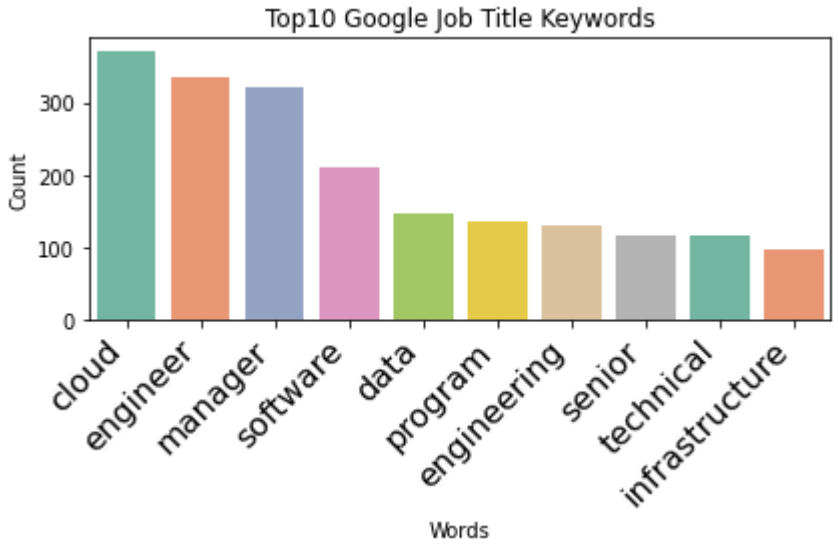
On top of that, **50% of the top 10 recruiting countries are in Asia**, which shows that Google also place a strong emphasis on the talent market in Asia, particularly in India, Taiwan, and Japan.

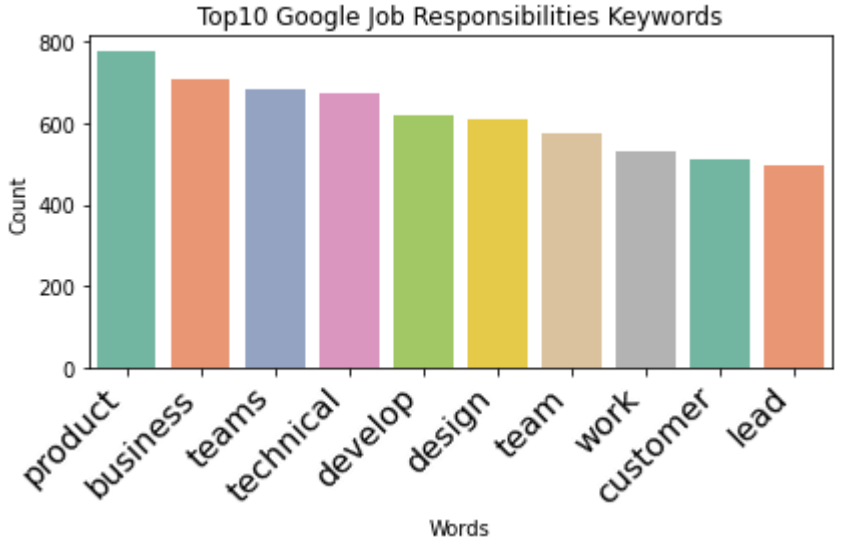




Insight 3 – Interesting Fact

Despite most of the top 10 Google job opening title keywords are related to technical field, the top 10 responsibilities keywords are mostly related to leadship and teamwork. This shows **equiping technical skills alone is not enough to excel the job at Google**. Soft skills such as **communication, teamwork and business awareness are also required.**





Conclusion

Answers for Questions/Tasks:

What Programming languages does Google request the most?

- ✓ **Python** , **Java** and **R** are three popular programming languages for Google Jobs.

What about the requirement for degrees?

- ✓ You need at least **bachelor degree** to get in Google. **Master and MBA** make you more competitive than other candidates.

Does work experience important? If so, how many years work experience is generally required?

- ✓ Most of the jobs did not mentioned work experience. Other than this, most of the works require **5 years work experience.**

What is the most popular job type in Google now?

- ✓ By searching title keywords, Google has the most in-demand talent in **Cloud related field**.

Recommendation

- ✓ If you are planning to learn a programming language before deciding on a career direction, **learning Python** will increase your chances of meeting the requirements for Google job openings.

- ✓ If you are looking for the country with the most job openings, you can refer **to the United States,** as it is the place where Google have the **highest talent.** On top of that, place a strong emphasis on the talent market in Asia, particularly in **India, Taiwan, and Japan.**

- ✓ Although demonstrating technical skills in a resument is important, **Google also values teamwork and communication skills**. Therefore, do not forget to show these skills as well.

Thank You!


MichaelLTsai


Michael Tsai


Michael Tsai


Michael Tsai


h94xup6