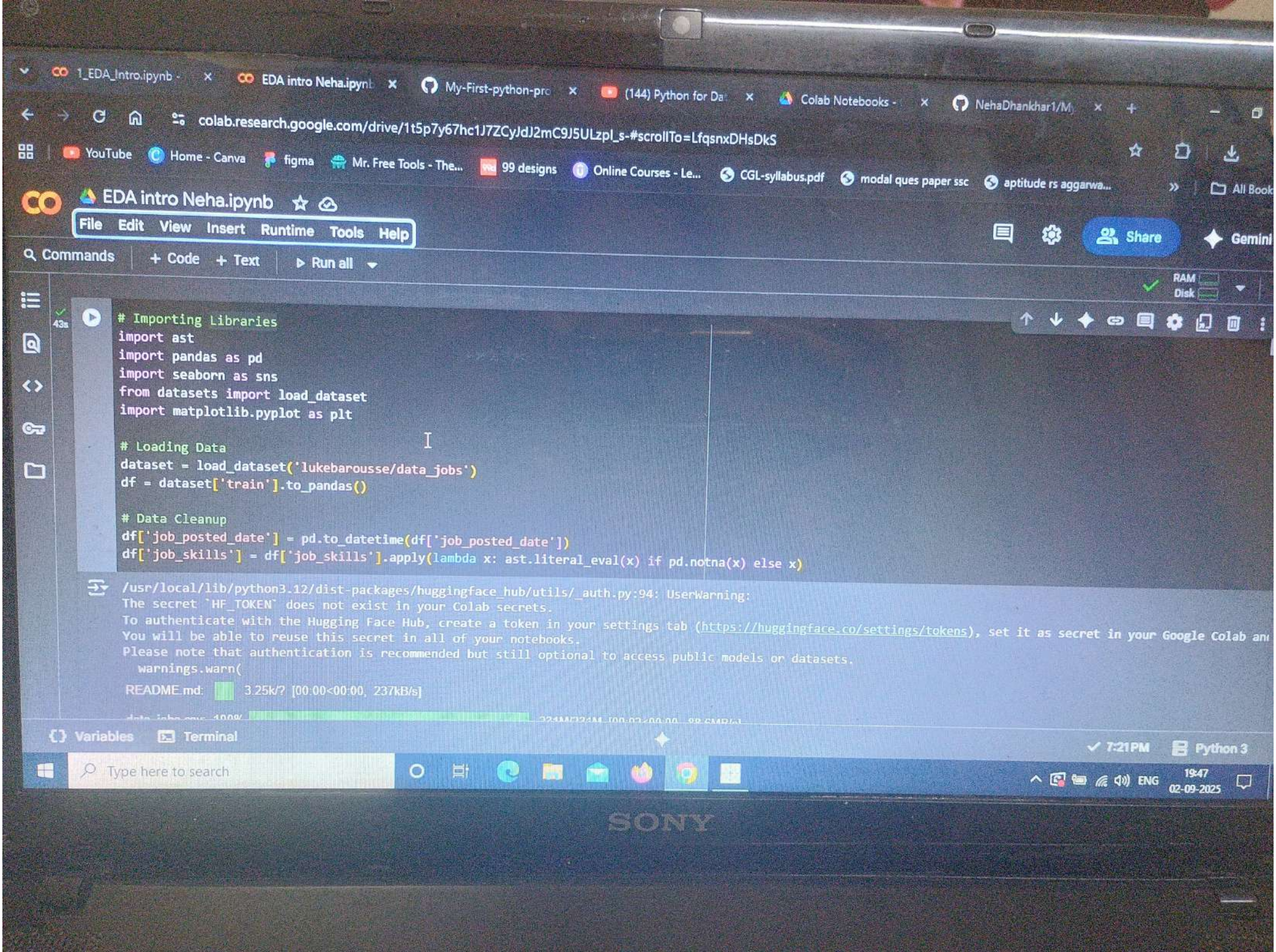

1_EDA_Intro.ipynb
EDA intro Neha.ipynb
My-First-python-pro
(144) Python for Da
Colab Notebooks -
NehaDhankhar1/M
colab.research.google.com/drive/1t5p7y67hc1J7ZCyJdJ2mC9J5ULzpI_s-#scrollTo=LfqsnxDHsDkS
YouTube
Home - Canva
figma
Mr. Free Tools - The...
99 designs
Online Courses - Le...
CGL-syllabus.pdf
modal ques paper ssc
aptitude rs aggarwa...
All Book
EDA intro Neha.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
43s
# Importing Libraries
import ast
import pandas as pd
import seaborn as sns
from datasets import load_dataset
import matplotlib.pyplot as plt
# Loading Data
dataset = load_dataset('lukebarousse/data_jobs')
df = dataset['train'].to_pandas()
# Data Cleanup
df['job_posted_date'] = pd.to_datetime(df['job_posted_date'])
df['job_skills'] = df['job_skills'].apply(lambda x: ast.literal_eval(x) if pd.notna(x) else x)
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab an
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
warnings.warn(
README.md: 3.25k/? [00:00<00:00, 237kB/s]
Variables
Terminal
7:21PM
Python 3
Type here to search
ENG
19:47
02-09-2025
SONY

1_EDA_Intro.ipynb - | EDA intro Neha.ipyn | My-First-python-pr | (144) Python for Da | Colab Notebooks - | NehaDhankhar1/M...

colab.research.google.com/drive/1t5p7y67hc1J7ZCyJdJ2mC9J5ULzpI_s-#scrollTo=LfqsnxDHsDkS

YouTube | Home - Canva | figma | Mr. Free Tools - The... | 99 designs | Online Courses - Le... | CGL-syllabus.pdf | modal ques paper ssc | aptitude rs aggarwa...

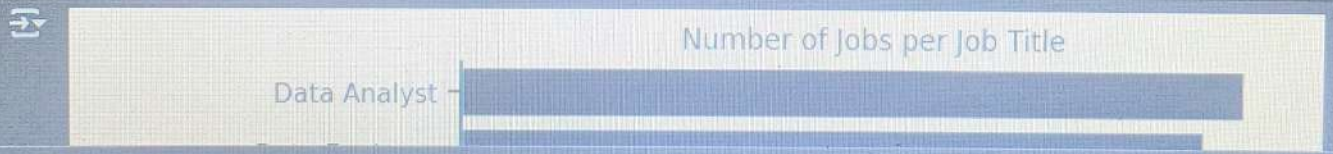
EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

Commands | + Code | + Text | Run all

```
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
README.md: 3.25k/? [00:00<00:00, 237kB/s]
data_jobs.csv: 100% 231M/231M [00:02<00:00, 88.6MB/s]
Generating train split: 100% 785741/785741 [00:14<00:00, 101457.14 examples/s]
```

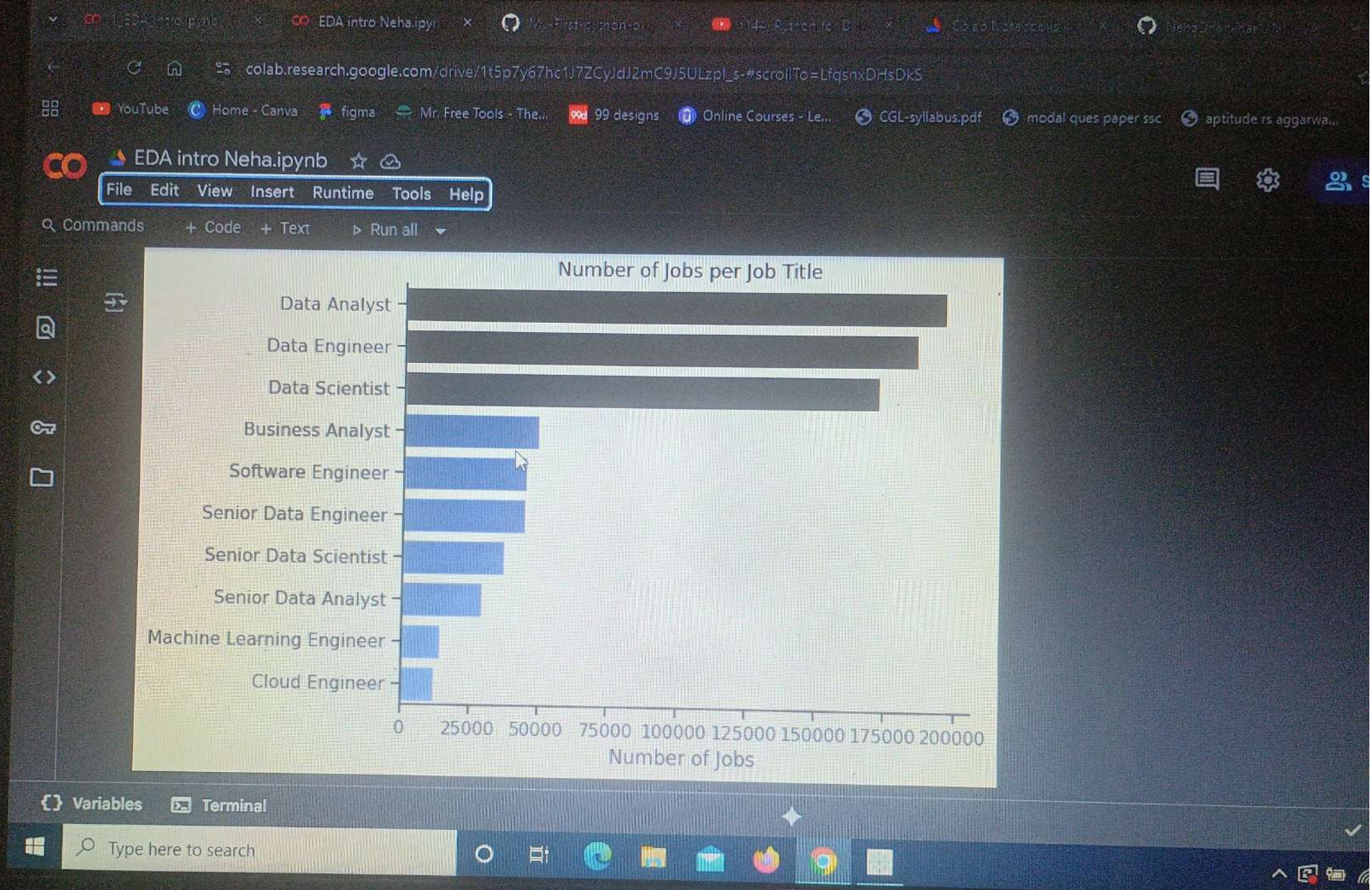
```
df_plot = df['job_title_short'].value_counts().to_frame()

sns.set_theme(style='ticks')
sns.barplot(data=df_plot, x='count', y='job_title_short', hue='count', palette='dark:b_r', legend=False)
sns.despine()
plt.title('Number of Jobs per Job Title')
plt.xlabel('Number of Jobs')
plt.ylabel('')
plt.show()
```



Variables | Terminal

Type here to search

SONY


EDA intro Neha.ipynb
colab.research.google.com/drive/1t5p7y67hc1J7ZCyJdJ2mC9J5ULzpl_s-#scrollTo=LfqsnxDHsDkS
EDA intro Neha.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
Number of Jobs per Job Title
Data Analyst
Data Engineer
Data Scientist
Business Analyst
Software Engineer
Senior Data Engineer
Senior Data Scientist
Senior Data Analyst
Machine Learning Engineer
Cloud Engineer
0 25000 50000 75000 100000 125000 150000 175000 200000
Number of Jobs
Variables Terminal
Type here to search

colab.research.google.com/drive/1t5p7y87hc1U7ZCyUd...

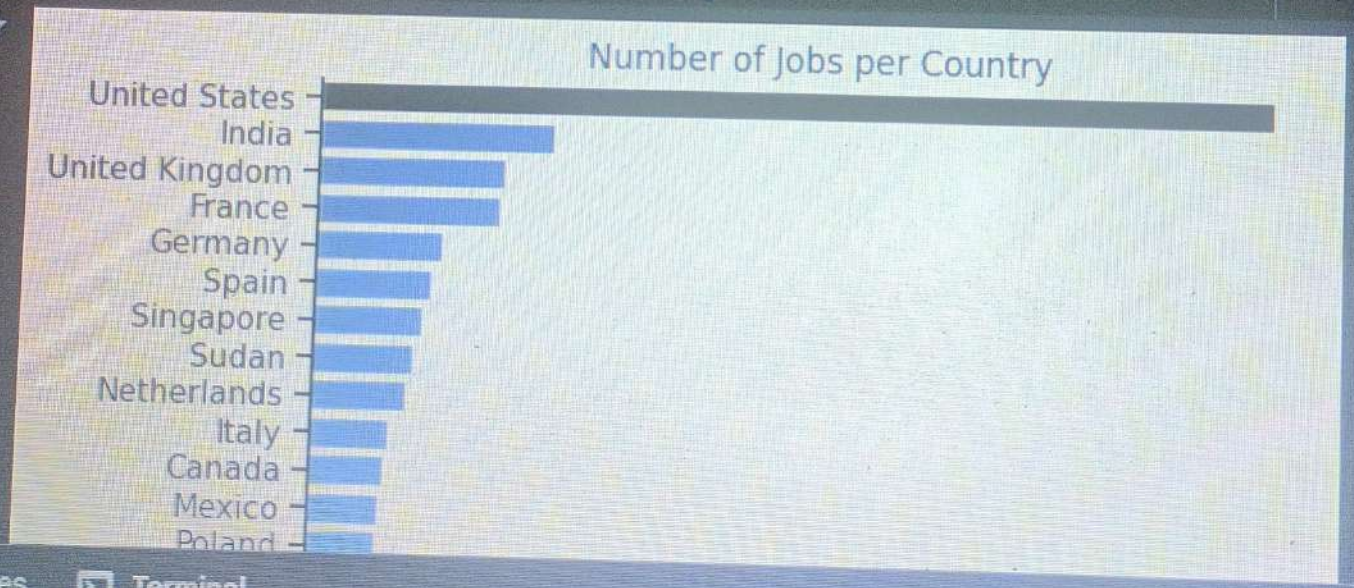
YouTube | Home - Canva | figma | Mr. Free Tools - The... | 99 designs | Online Courses - Le... | CGL-syllabus.pdf | modal ques paper ssc

EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text ▷ Run all

```
df_plot = df['job_country'].value_counts().to_frame().head(20)

sns.set_theme(style='ticks')
sns.barplot(data=df_plot, x='count', y='job_country', hue='count', palette='dark:b_r', legend=False)
sns.despine()
plt.title('Number of Jobs per Country')
plt.xlabel('Number of Jobs')
plt.ylabel('')
plt.show()
```



Variables Terminal

Type here to search

SONY

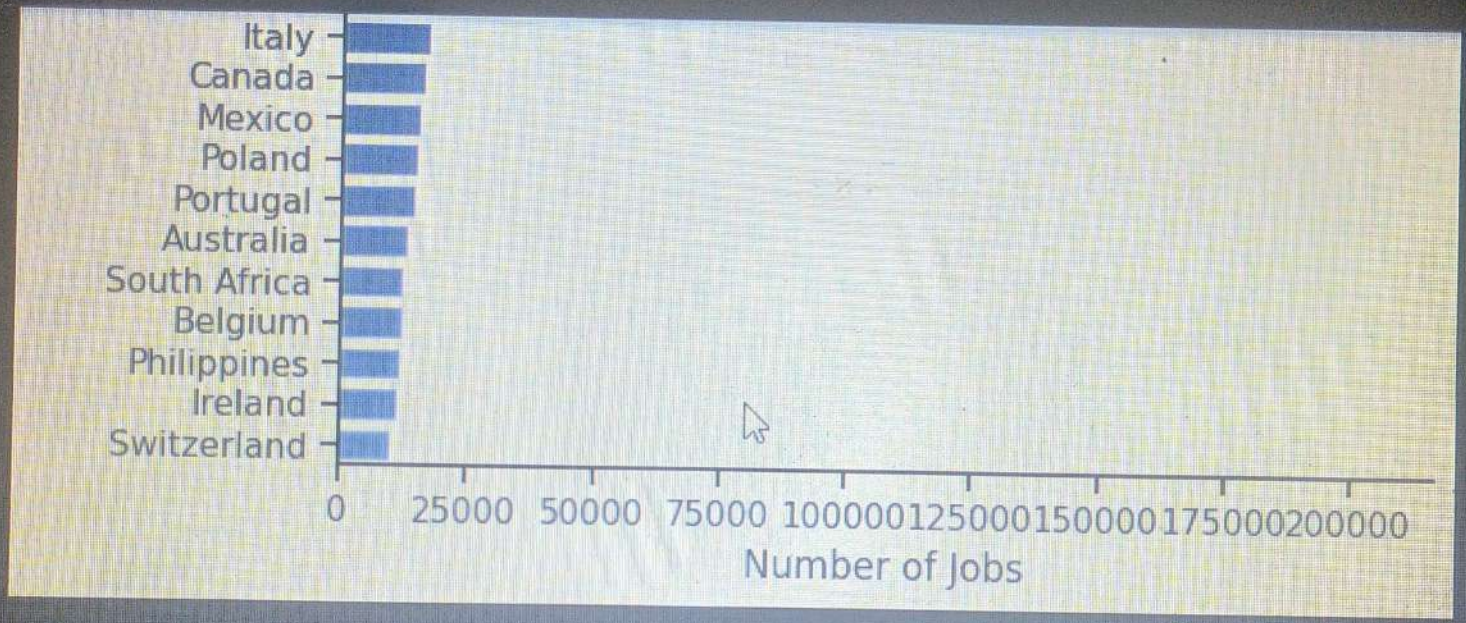
colab.research.google.com/drive/1t5p7y67hc1J7ZCyJdJ2mC9J5ULzpI_s-#scrollTo=LfqsnxDHsDkS

YouTube | Home - Canva | figma | Mr. Free Tools - The... | 99 designs | Online Courses - Le... | CGL-syllabus.pdf | modal qu

EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all



[4]
```
df_plot = df['company_name'].value_counts().to_frame()[1:].head(20)

sns.set_theme(style='ticks')
sns.barplot(data=df_plot, x='count', y='company_name', hue='count', palette='dark:b_r', legend=False)
sns.despine()
plt.title('Number of Jobs per Company')
plt.xlabel('Number of Jobs')
plt.ylabel('')
plt.show()
```

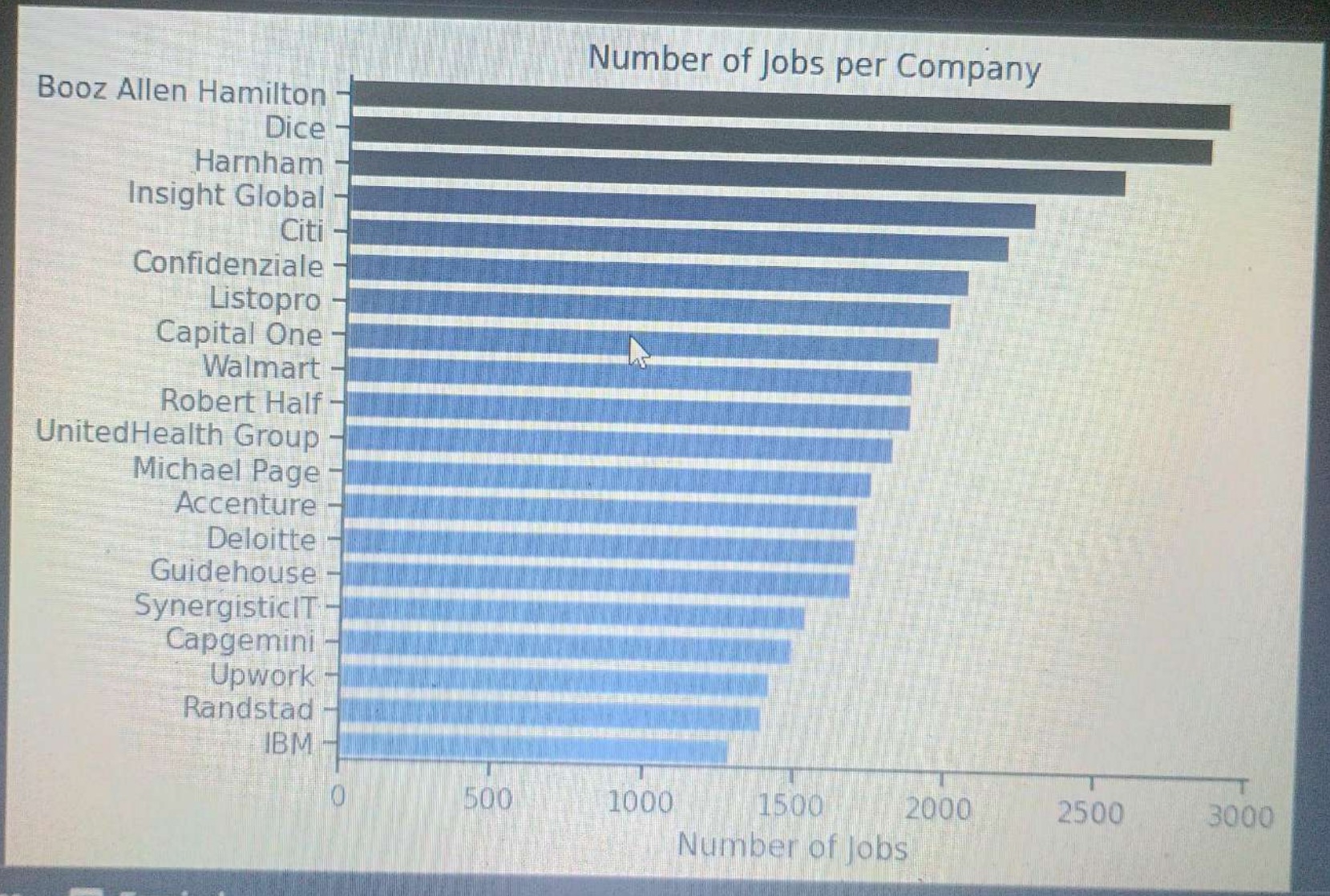
Variables Terminal

Type here to search

EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all



Variables Terminal

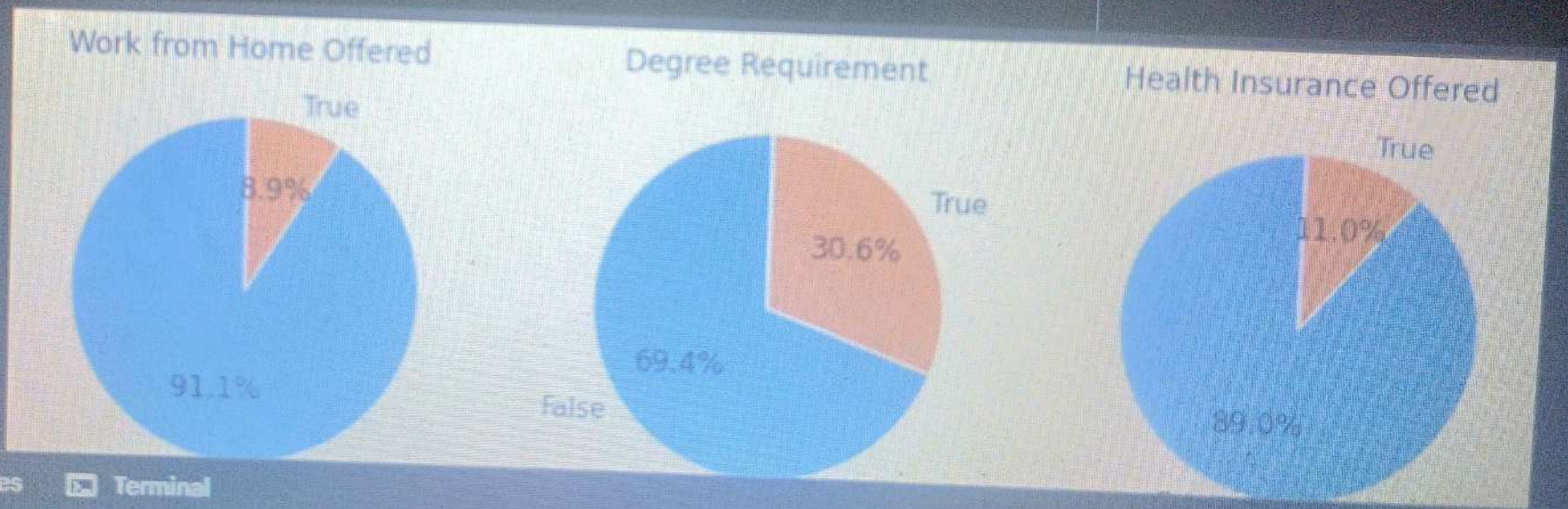
YouTube | Home - Canva | figma | Mr. Free Tools - The... | 99 designs | Online Courses - Le... | CGL-syllabus.pdf

EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all

```
dict_column = {
    'job_work_from_home': 'Work from Home Offered',
    'job_no_degree_mention': 'Degree Requirement',
    'job_health_insurance': 'Health Insurance Offered'
}

fig, ax = plt.subplots(1, 3, figsize=(11, 3.5))

for i, (column, title) in enumerate(dict_column.items()):
    ax[i].pie(df[column].value_counts(), labels=['False', 'True'], autopct='%1.1f%%', startangle=90)
    ax[i].set_title(title)

plt.show()
```



Variables Terminal

Type here to search

EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all

```python
# Importing Libraries
import ast
import pandas as pd
import seaborn as sns
from datasets import load_dataset
import matplotlib.pyplot as plt

# Loading Data
dataset = load_dataset('lukebarousse/data_jobs')
df = dataset['train'].to_pandas()

# Data Cleanup
df['job_posted_date'] = pd.to_datetime(df['job_posted_date'])
df['job_skills'] = df['job_skills'].apply(lambda x: ast.literal_eval(x) if pd.notna(x) else x)
```

```python
df_DA_US = df[(df['job_country'] == 'United States') & (df['job_title_short'] == 'Data Analyst')]
```

```python
df_plot = df_DA_US['job_location'].value_counts().head(10).to_frame()

sns.set_theme(style='ticks')
sns.barplot(data=df_plot, x='count', y='job_location', hue='count', palette='dark:b_r', legend=False)
sns.despine()
plt.title('Counts of Job Locations for Data Analyst in the US')
plt.xlabel('Number of Jobs')
plt.ylabel('')
plt.show()
```

Variables Terminal

Type here to search

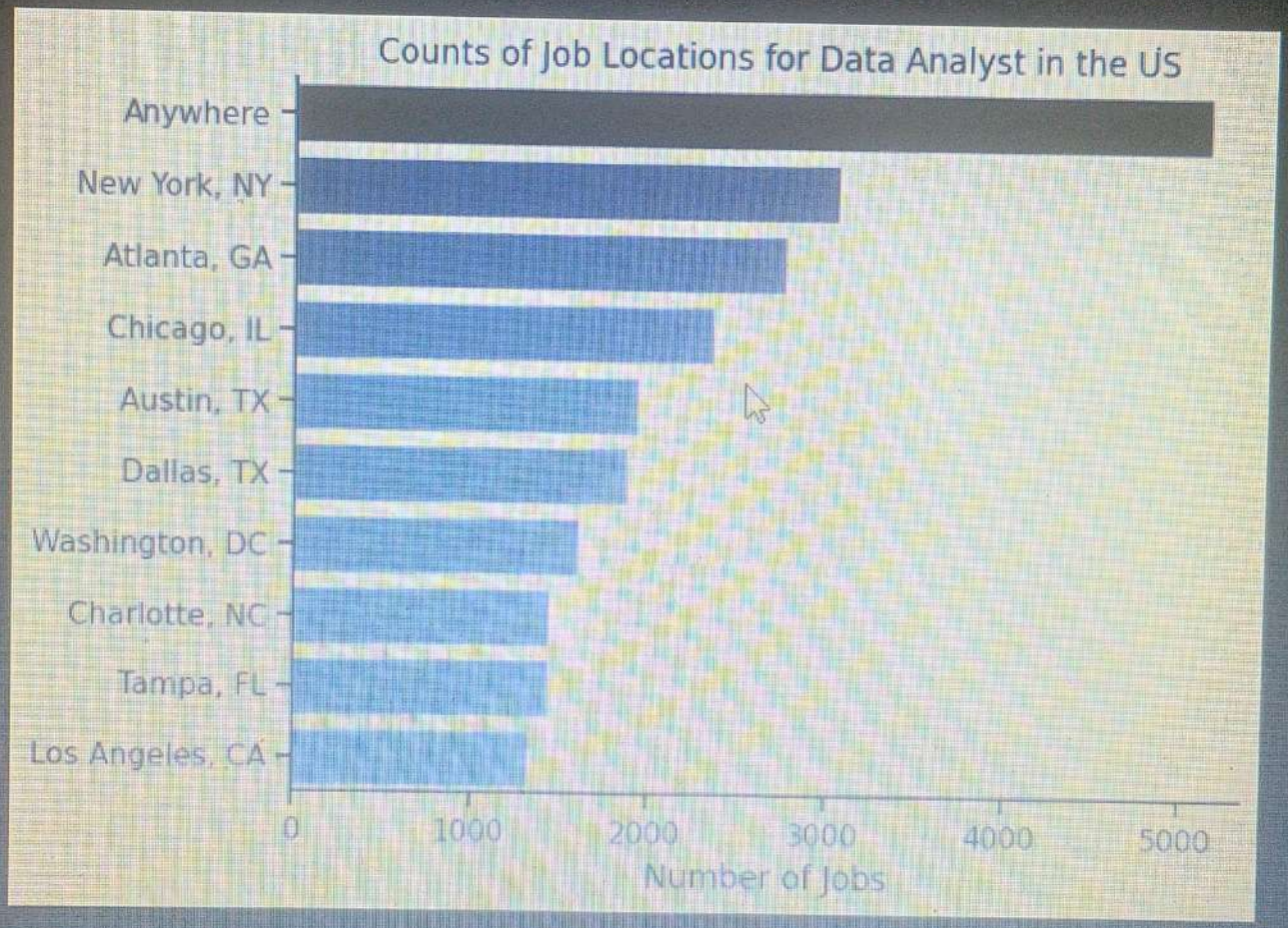
colab.research.google.com/drive/1t5p7y67hc1J7ZCyJdJ2mC9J5ULzpl_s-#scrollTo=LfqsnxDHsDkS

YouTube | Home - Canva | figma | Mr. Free Tools - The... | 99 designs | Online Courses - Le... | CGL-syllabus.pdf

EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all



[9] # rewrite the above with a for loop

Variables Terminal

Type here to search

colab.research.google.com/drive/1t5p7y67hc1J7ZCyJdJ2mC9J5ULzpl_s-#scrollTo=LfqsnxDHsDkS

YouTube | Home - Canva | figma | Mr. Free Tools - The... | 99 designs | Online Courses - Le... | CGL-syllabus.pdf | modal ques

EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all

```
# rewrite the above with a for loop
dict_column = {
    'job_work_from_home': 'Work from Home Offered',
    'job_no_degree_mention': 'Degree Requirement',
    'job_health_insurance': 'Health Insurance Offered'
}

fig, ax = plt.subplots(1, 3)
fig.set_size_inches((12, 5))

for i, (column, title) in enumerate(dict_column.items()):
    ax[i].pie(df_DA_US[column].value_counts(), labels=['False', 'True'], autopct='%1.1f%%', startangle=90)
    ax[i].set_title(title)

# plt.suptitle('Benefit Analysis of Data Jobs', fontsize=16)
plt.show()
```

Variables Terminal

Type here to search

colab.research.google.com/drive/1t5p7y67hc1J7ZCyJdJ2mC9J5ULzpL_s-#scrollTo=LfqsnxDHsDkS

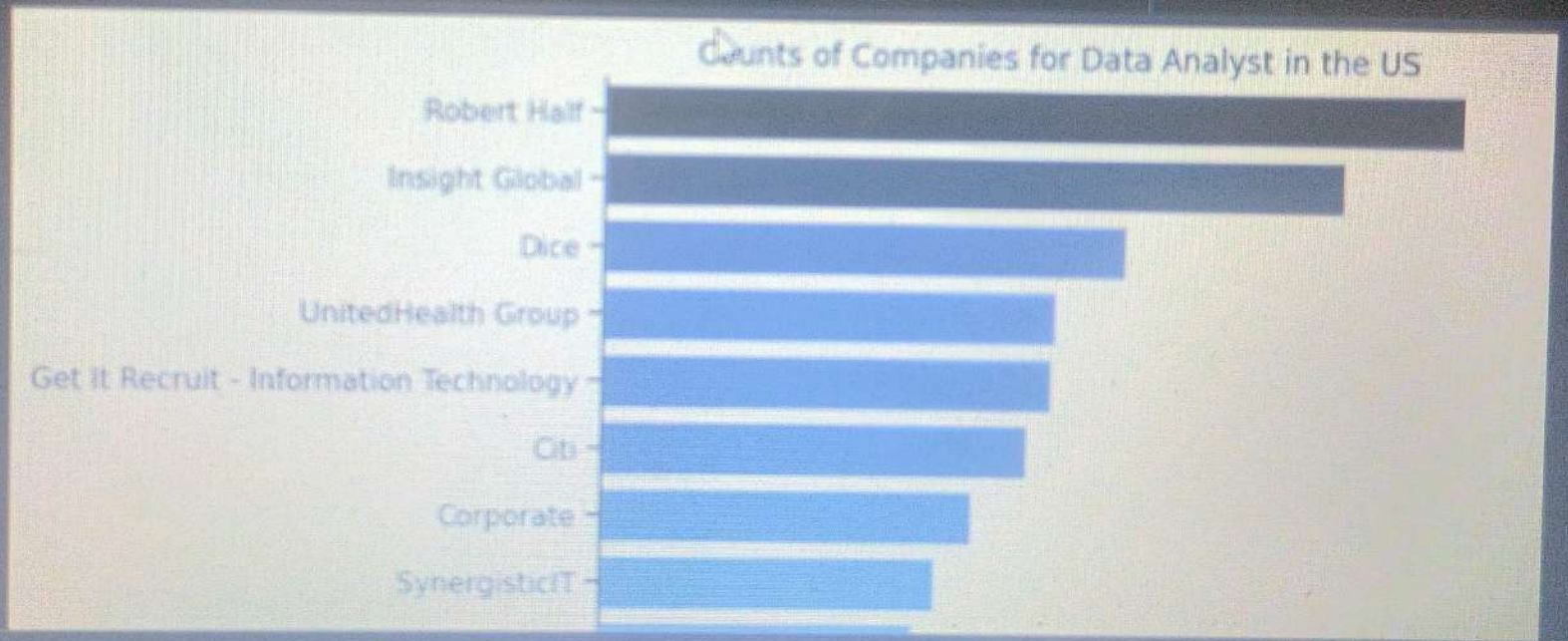
YouTube | Home - Canva | figma | Mr. Free Tools - The... | 99 designs | Online Courses - Le... | CGL-syllabus.pdf | modal ques pa

EDA intro Neha.ipynb

File Edit View Insert Runtime Tools Help

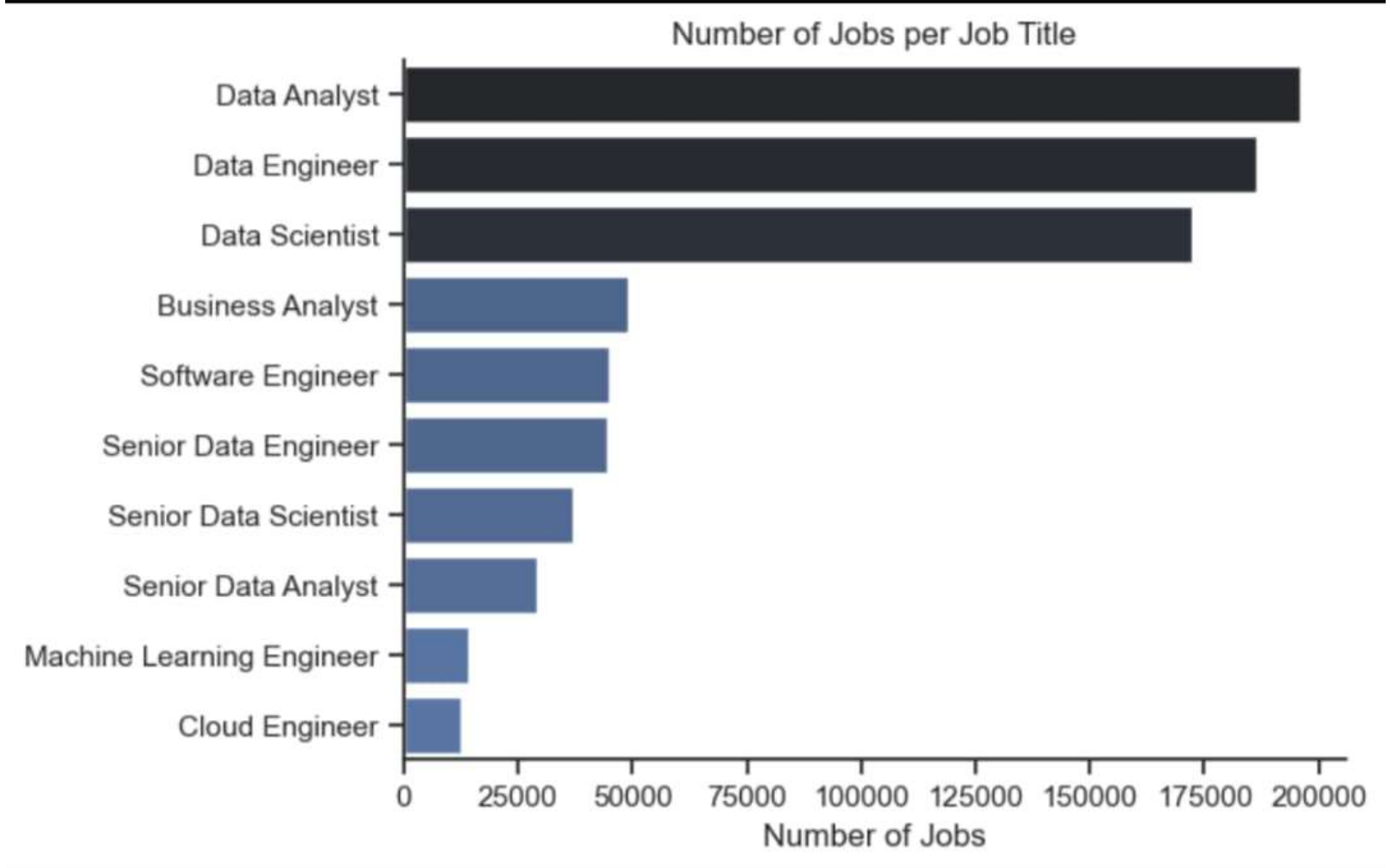
Commands + Code + Text ▶ Run all

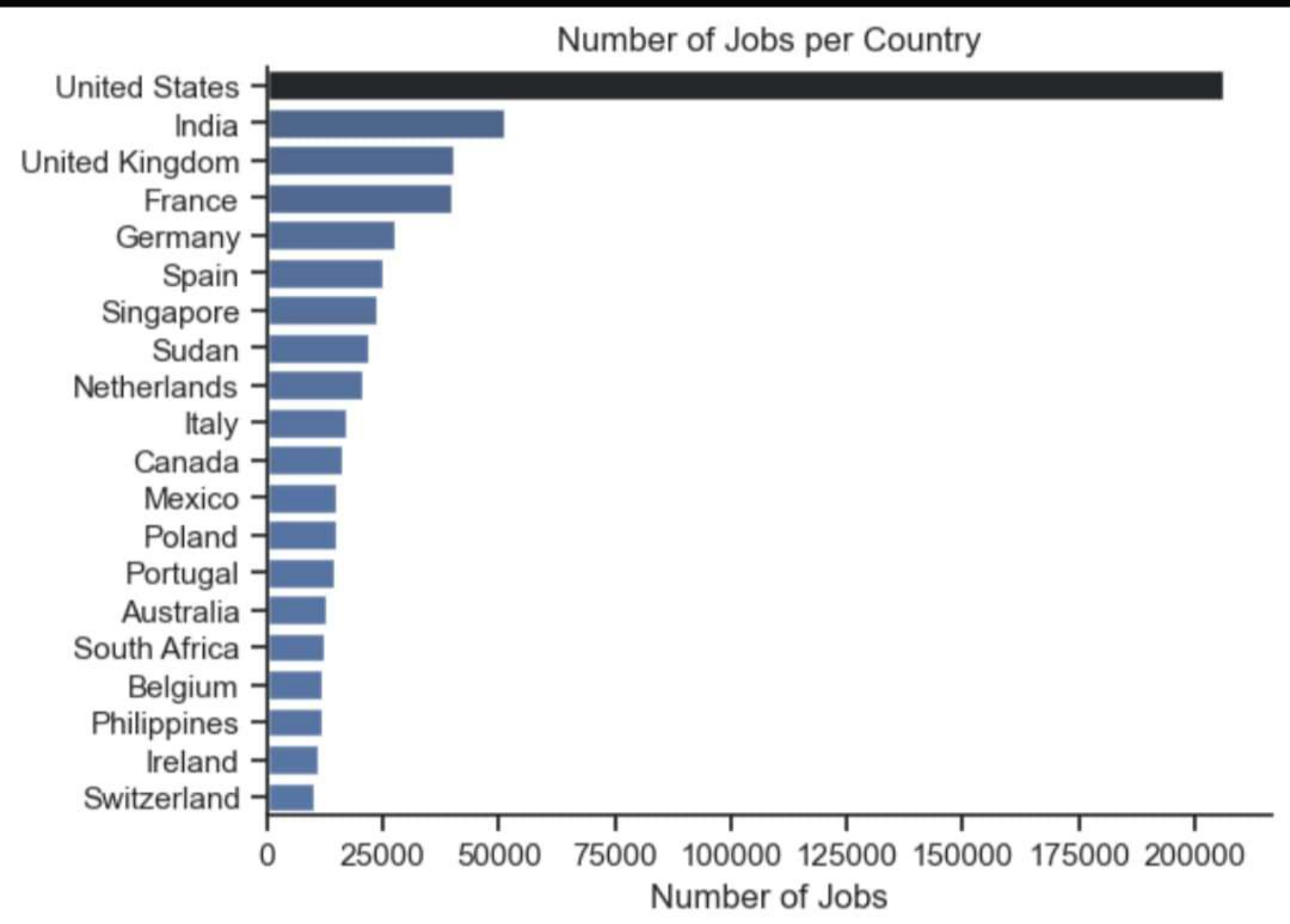
```
df_plot = df_DA_US['company_name'].value_counts().head(10).to_frame()

sns.set_theme(style='ticks')
sns.barplot(data=df_plot, x='count', y='company_name', hue='count', palette='dark:b_r', legend=False)
sns.despine()
plt.title('Counts of Companies for Data Analyst in the US')
plt.xlabel('Number of Jobs')
plt.ylabel('')
plt.show()
```

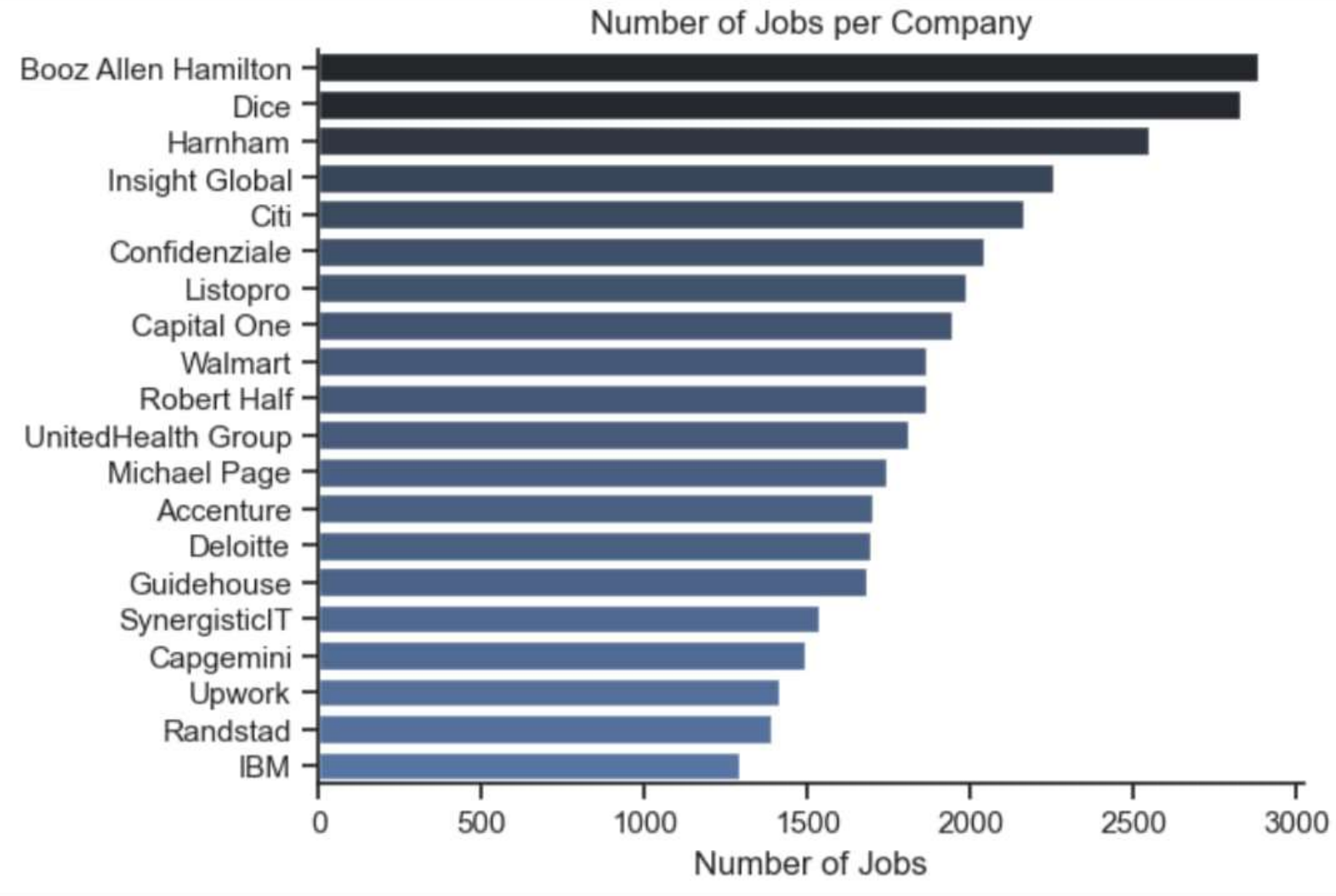


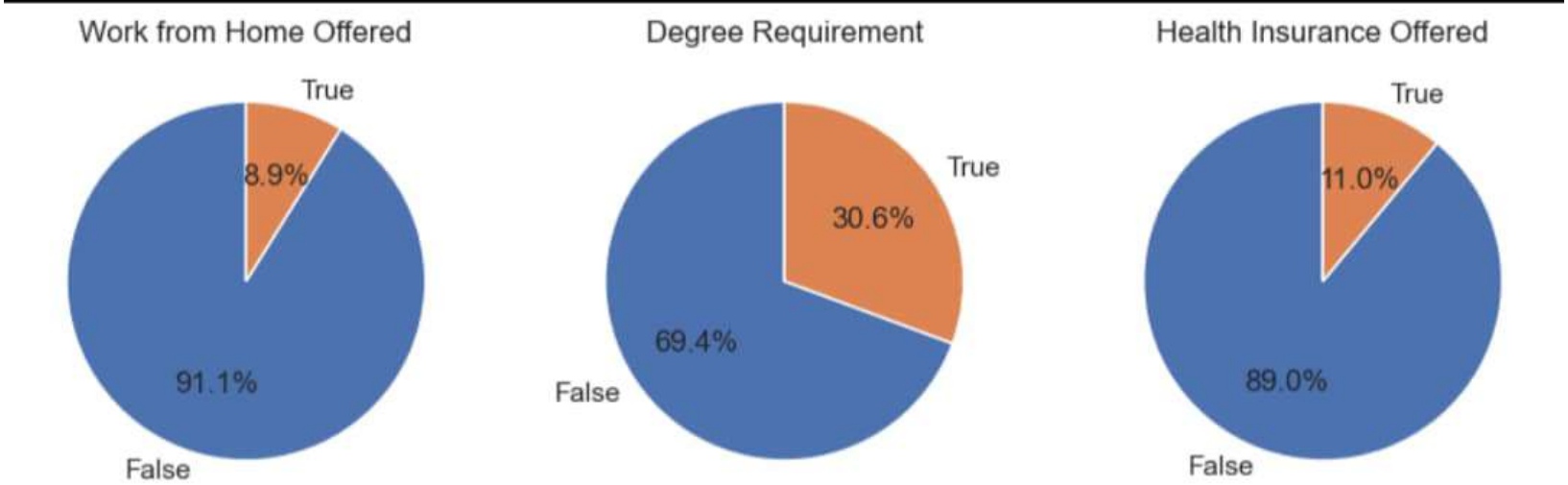
Variables Terminal

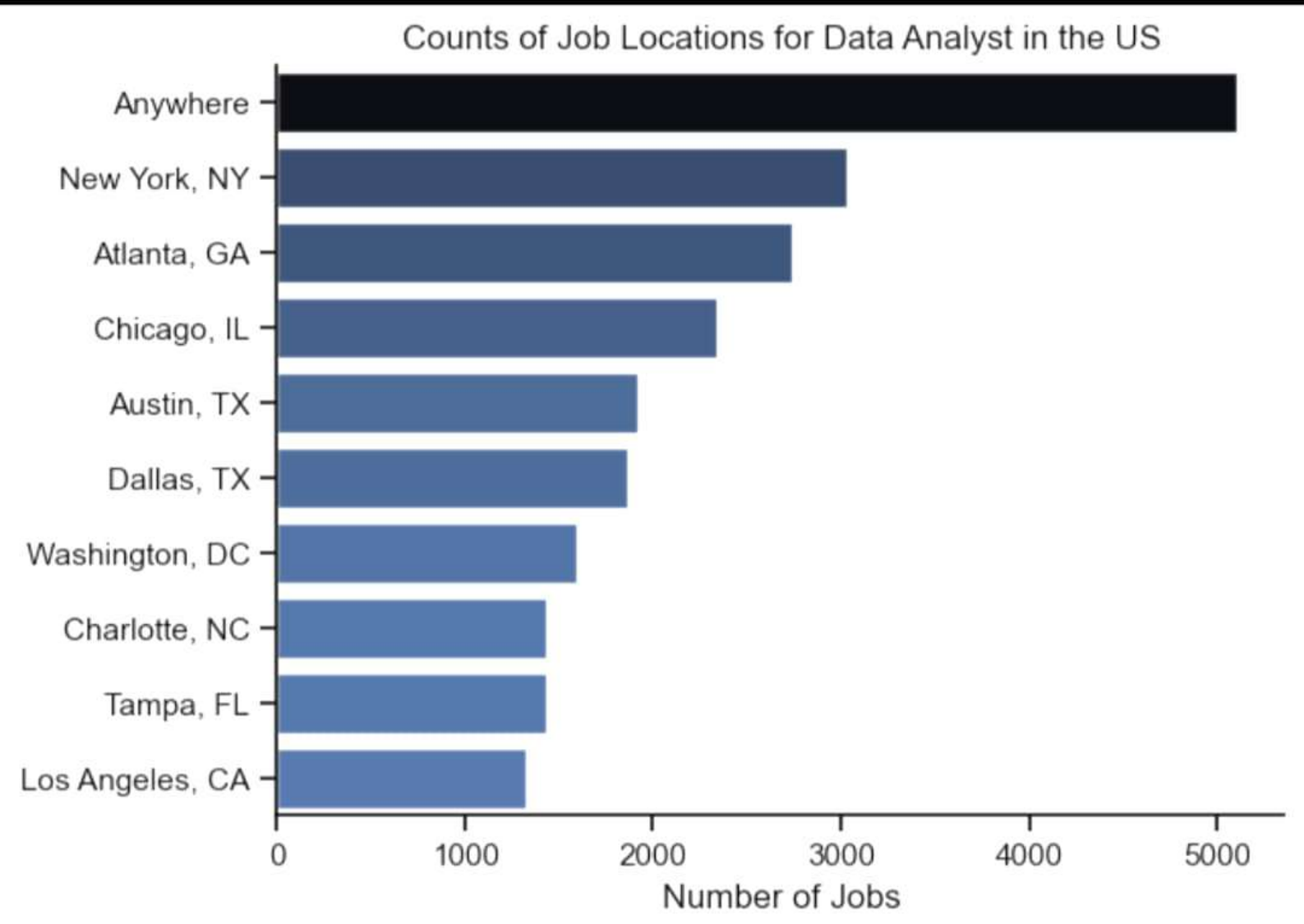
Type here to search

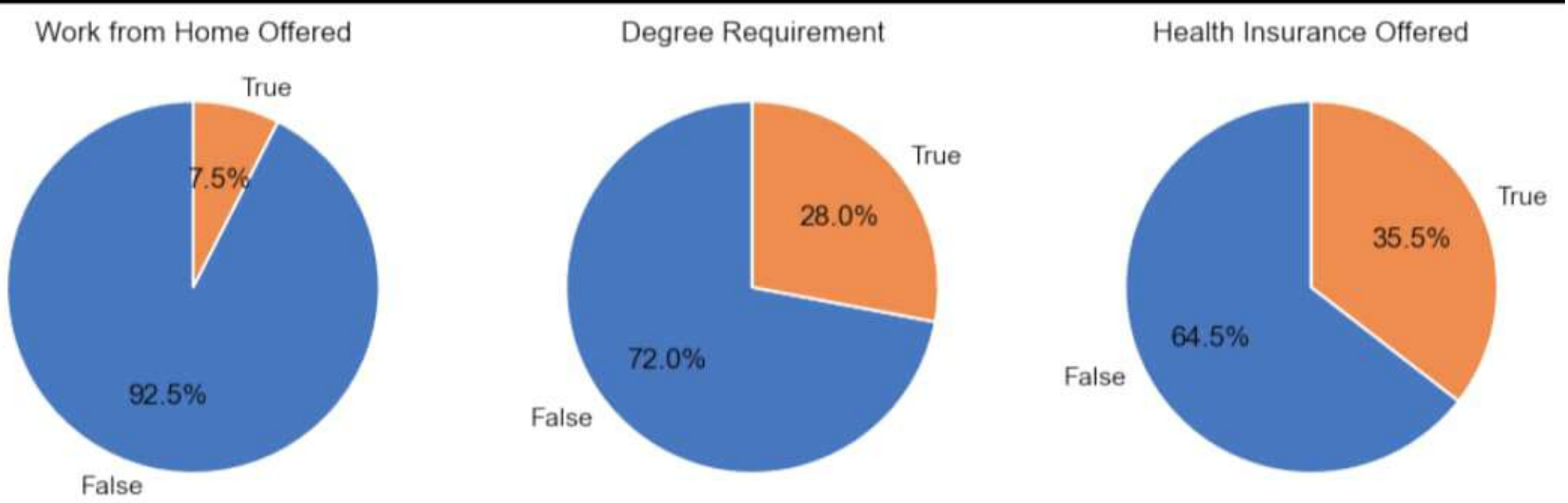

Number of Jobs per Job Title
Data Analyst
Data Engineer
Data Scientist
Business Analyst
Software Engineer
Senior Data Engineer
Senior Data Scientist
Senior Data Analyst
Machine Learning Engineer
Cloud Engineer
0
25000
50000
75000
100000
125000
150000
175000
200000
Number of Jobs

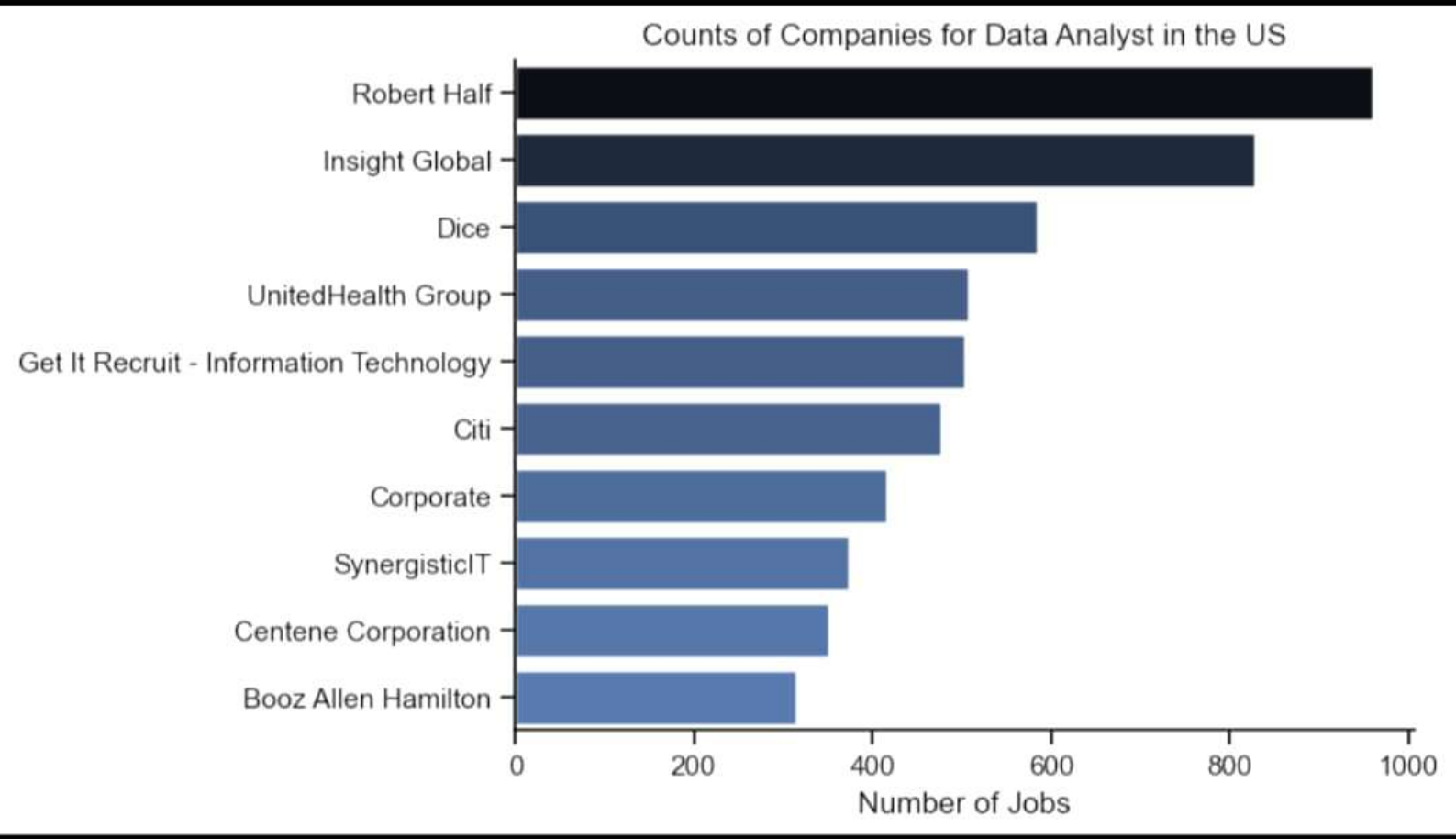

Number of Jobs per Country
United States
India
United Kingdom
France
Germany
Spain
Singapore
Sudan
Netherlands
Italy
Canada
Mexico
Poland
Portugal
Australia
South Africa
Belgium
Philippines
Ireland
Switzerland
0
25000
50000
75000
100000
125000
150000
175000
200000
Number of Jobs


Number of Jobs per Company
Booz Allen Hamilton
Dice
Harnham
Insight Global
Citi
Confidenziale
Listopro
Capital One
Walmart
Robert Half
UnitedHealth Group
Michael Page
Accenture
Deloitte
Guidehouse
SynergisticIT
Capgemini
Upwork
Randstad
IBM
0
500
1000
1500
2000
2500
3000
Number of Jobs


Work from Home Offered
True
8.9%
91.1%
False
Degree Requirement
True
30.6%
69.4%
False
Health Insurance Offered
True
11.0%
89.0%
False


Counts of Job Locations for Data Analyst in the US
Anywhere
New York, NY
Atlanta, GA
Chicago, IL
Austin, TX
Dallas, TX
Washington, DC
Charlotte, NC
Tampa, FL
Los Angeles, CA
0
1000
2000
3000
4000
5000
Number of Jobs


Work from Home Offered
True
7.5%
92.5%
False
Degree Requirement
True
28.0%
72.0%
False
Health Insurance Offered
True
35.5%
64.5%
False


Counts of Companies for Data Analyst in the US
Robert Half
Insight Global
Dice
UnitedHealth Group
Get It Recruit - Information Technology
Citi
Corporate
SynergisticIT
Centene Corporation
Booz Allen Hamilton
0
200
400
600
800
1000
Number of Jobs