1. Importing Libraries

```python
from bs4 import BeautifulSoup
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
import numpy as np
import pandas as pd
import time
import requests
import warnings as wn
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.offline as py
from wordcloud import WordCloud,STOPWORDS,ImageColorGenerator
import string
from collections import Counter
import nltk
from nltk.corpus import stopwords
warnings.filterwarnings("ignore")
import plotly.express as px
```

2. Installing Chrome Driver for Selenium

```python
driver = webdriver.Chrome(ChromeDriverManager().install())
```

3. Iterating through indeed job portal and fetching title ,company name and link for job description of all Data Scientist Jobs.

```python
Titles
joblist = pd.DataFrame()
for start in range(0,1500,10):
    sleep(randint(2,10))
    url = 'https://in.indeed.com/jobs?q=data%20scientist&start=' + str(start)
    driver.get(url)
    soup = BeautifulSoup(driver.page_source, 'html.parser')
    jobs = soup.find_all('a', class_='jcs-JobTitle')
    for job in jobs:
        title = job.find('span').text
        link = job.get('href')
        company = job.find('span', class_='companyName').text
        joblist = joblist.append({'title':title, 'link':link, 'company':company}, ignore_index=True)
```


	title	link	company
0	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Oracle
1	Data scientist	https://in.indeed.com/rc/clk?jk=...	Zensar
2	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Apple
3	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Accenture
4	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Amazon
...	...	...	...
2668	Senior Data Scientist	https://in.indeed.com/rc/clk?jk=...	Fractal Analytics
2669	Senior Business Scientist	https://in.indeed.com/rc/clk?jk=...	ZS
2670	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Publicis Sapient
2671	Machine Learning Engineer	https://in.indeed.com/rc/clk?jk=...	Lululemon India Tech Hub
2672	Lead Quantitative Analytics Specialist	https://in.indeed.com/rc/clk?jk=...	Wells Fargo

2662 rows × 3 columns

4. Saving the dataframe to a csv file and reusing it to fetch Job description.

```python
joblist.to_csv('joblist.csv')
```

```python
joblist2 = pd.read_csv('joblist.csv')
```

	Unnamed: 0	title	link	company
0	0	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Oracle
1	1	Data scientist	https://in.indeed.com/rc/clk?jk=...	Zensar
2	2	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Apple
3	3	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Accenture
4	4	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Amazon
...	...	...	...	...
2668	2668	Senior Data Scientist	https://in.indeed.com/rc/clk?jk=...	Fractal Analytics
2669	2669	Senior Business Scientist	https://in.indeed.com/rc/clk?jk=...	ZS
2670	2670	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Publicis Sapient
2671	2671	Machine Learning Engineer	https://in.indeed.com/rc/clk?jk=...	Lululemon India Tech Hub
2672	2672	Lead Quantitative Analytics Specialist	https://in.indeed.com/rc/clk?jk=...	Wells Fargo

2662 rows × 4 columns

5. Cleaning link column as some of the links does not start with https so removing those from our analysis.

```python
joblist3 = joblist2[joblist2['link'].str.contains('https')==True].reset_index()
```

	index	Unnamed: 0	title	link	company
0	1	1	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Zensar
1	2	2	Data scientist	https://in.indeed.com/rc/clk?jk=...	Apple
2	3	3	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Accenture
3	4	4	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Honeywell
4	5	5	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Accenture
...	...	...	...	...	...
1667	2664	2664	Machine Learning Engineer	https://in.indeed.com/rc/clk?jk=...	Lululemon India Tech Hub
1668	2667	2667	Applied Scientist India Machine Learning	https://in.indeed.com/rc/clk?jk=...	Amazon
1669	2668	2668	Senior Data Scientist	https://in.indeed.com/rc/clk?jk=...	Fractal Analytics
1670	2669	2669	Senior Business Scientist	https://in.indeed.com/rc/clk?jk=...	ZS
1671	2670	2670	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Publicis Sapient

1672 rows × 5 columns

6. Fetching job description of all jobs by looping through 'link' column.

```python
description = []
for link in joblist3new['link']:
    sleep(1)
    driver.get(link)
    time.sleep(3)
    detailsoup = BeautifulSoup(driver.page_source, 'html.parser')
    description = detailsoup.find('div', class_='jobsearch-jobDescriptionText').text.replace('\n',' ')
    description.append(description)
joblist3new['Description'] = description
```

```
IndexError                    Traceback (most recent call last)
<ipython-input-18-...> in <module>
----> 1 for link in joblist3new['link']:
      2    sleep(1)
      3    driver.get(link)
```

There was error while fetching job description and we were able to fetch 638 records only so we will do analysis on the same.

7. Creating new dataframe with 638 job list and setting all words of Description to lowercase for further comparison.

```python
joblist4 = joblist3new[:638]
joblist4['Description'] = joblist4['Description'].str.lower()
joblist4
```

	index	Unnamed: 0	title	link	company	Description
0	1	1	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Zensar	design, develop and programs methods, proces...
1	2	2	Data scientist	https://in.indeed.com/rc/clk?jk=...	Apple	responsibilities/problem-solving you should ha...
2	3	3	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Accenture	overview i mentate we believe in the power of...
3	4	4	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Honeywell	mission is emerging as the innovation of a bu...
4	5	5	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Accenture	about accenture accenture is a global prof...
...	...	...	...	...	...	...
633	701	701	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Maersk	sample text
634	702	702	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Accenture	sample text
635	703	703	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Angel Broking	sample text
636	704	704	Data Scientist	https://in.indeed.com/rc/clk?jk=...	MindTree	sample text
637	705	705	Data Scientist	https://in.indeed.com/rc/clk?jk=...	Accenture	sample text

637 rows × 6 columns

8. Created Skills list and then adding corresponding columns for each skill in our dataframe.

```python
skills = ['statistics','python','sql','r','c','java','scala','sas','tableau','excel','hadoop','spark','machine learning','deep learning','nlp','pytorch','tensorflow','keras','optimization','visualization','classification','regression','clustering','aws','azure','gcp','probability','research','modeling','domain']
```

```python
for i in skills:
    joblist4[i] = joblist4['Description'].str.find(i)
joblist4.loc[:,'statistics':].head()
```

```python
# Count of each skill in the description
words_all = pd.DataFrame(dict(ProductTable))
table_field = {"Statistics":"Count"}
for i in words:
    table.add(words[i].sum())
table_reverse = True
print(table)
```

```
Skill        | Count
---------------|-------
statistics   | 638
python       | 638
sql          | 638
...
```

8.1 Creating dictionary for each skill and count of those skill as values.

```python
wordDict={}
for i in words:
    dict[i]=joblist4[i].values.sum()
```

8.2 Plotting Wordcloud for skill using generate from frequencies of "WordCloud".

```python
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
wc = WordCloud(background_color="white",
               stopwords=STOPWORDS,
               width=1000,height=1000).generate_from_frequencies(wordDict)
plt.subplots(figsize=(12,10))
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```


8.3 Plotting Bar Chart using plotly for top 15 skills.

```python
words15 = dict(sorted(wordDict.items(), key = lambda x: x[1], reverse = True))
first_n = dict(list(words15.items())[:15])
names = list(first_n.keys())
values = list(first_n.values())
data = [ go.Bar(x=first_n.keys(),y=first_n.values(),orientation='v',marker=dict(color='rgb(bww)'),
               opacity=0.8)]
layout = go.Layout(title='Top 15 most popular skills for a Data Scientist',titlex,
               yaxis=dict(title='Frequency in Job Description'),
               xaxis=dict(title='Skill Name'))
fig = go.Figure(data=data, layout=layout)
fig.update_layout(template='plotly_dark')
fig.update_layout(template='plotly_dark')
py.iplot(fig)
```


9. Created Soft Skills list and then adding corresponding columns for each soft skill in our dataframe.

```python
soft_skills = ['communication','curiosity','storytelling','critical thinking','team player','problem solving','business','creative','presentation','collaboration','writing','analytical','domain','research','nlp']
for i in soft_skills:
    joblist4[i] = joblist4['Description'].str.find(i)
```

9.1 Creating dictionary for each soft skill and count of those skills as values.

```python
wordDict1={}
for i in soft_skills:
    dict[i]=joblist4[i].values.sum()
```

9.2 Plotting Wordcloud for skill using generate from frequencies of "WordCloud".

```python
wc = WordCloud(background_color="white",
               stopwords=STOPWORDS,
               width=1000,height=1000).generate_from_frequencies(wordDict1)
plt.subplots(figsize=(12,10))
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```


9.3 Plotting Bar Chart using plotly for top 15 soft skills.

```python
sorted_dict = dict(sorted(wordDict1.items(), key = lambda x: x[1], reverse = True))
first_n = dict(list(sorted_dict.items())[:15])
names = list(first_n.keys())
values = list(first_n.values())
data = [ go.Bar(x=first_n.keys(),y=first_n.values(),orientation='v',marker=dict(color='rgb(bww)'),
               opacity=0.8)]
layout = go.Layout(title='Top 15 most popular soft skills for a Data Scientist',titlex,
               yaxis=dict(title='Frequency in Job Description'),
               xaxis=dict(title='Soft Skill Name'))
fig = go.Figure(data=data, layout=layout)
fig.update_layout(template='plotly_dark')
py.iplot(fig)
```