Data Science

Hayk Minasyan

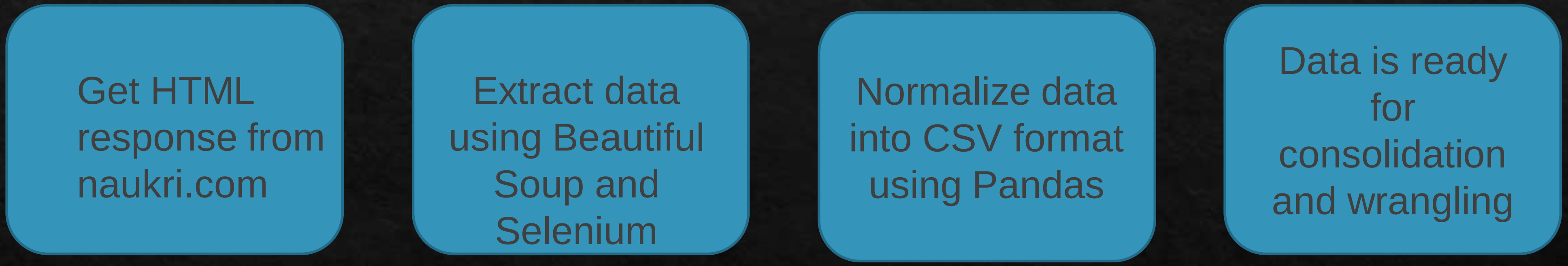
Outline

Data scraping

Data cleaning

Data visualization

Data scraping



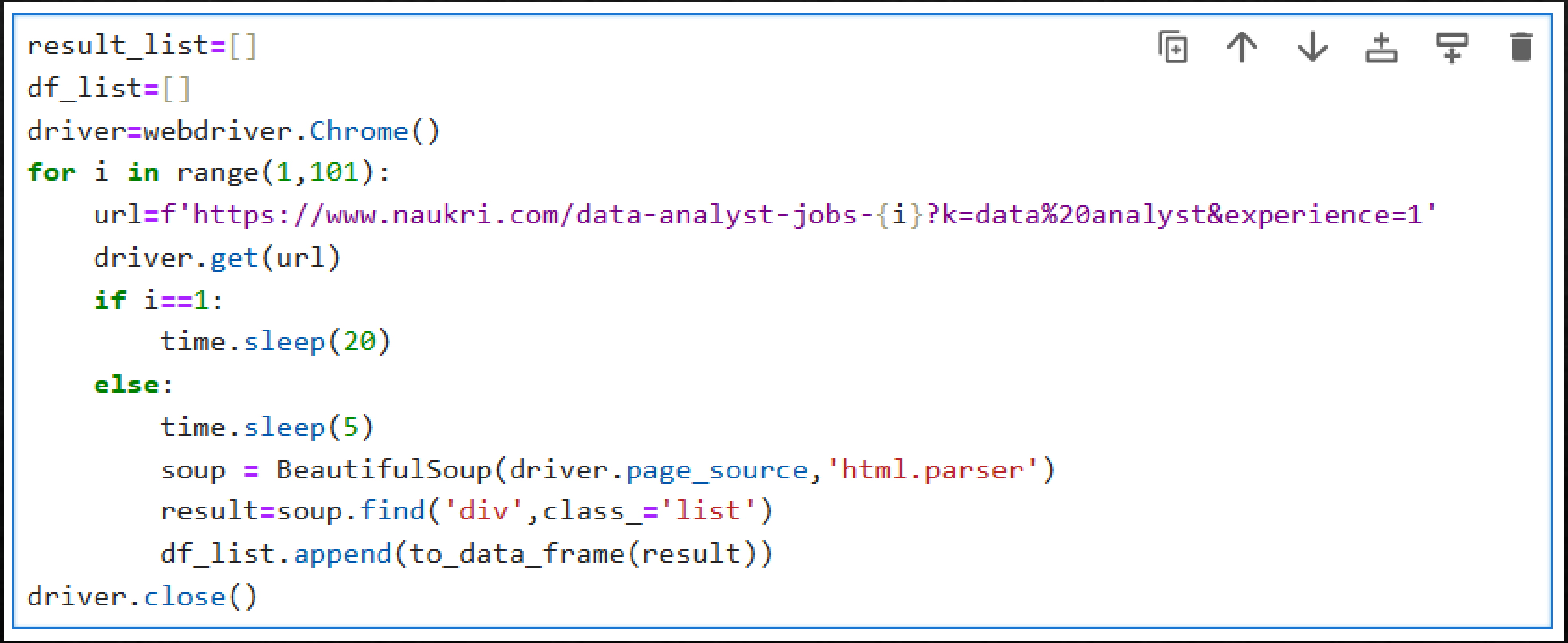
Get HTML response from naukri.com

Extract data using Beautiful Soup and Selenium

Normalize data into CSV format using Pandas

Data is ready for consolidation and wrangling

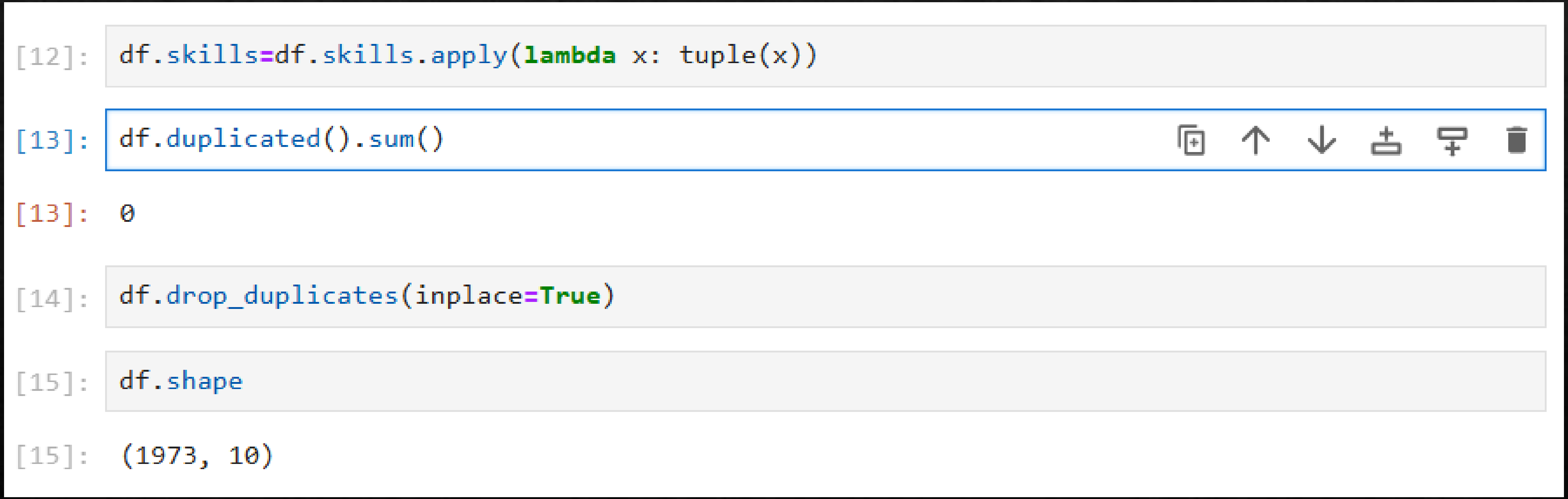
Collecting web page data using Selenium and Beautiful soup



```
result_list=[]
df_list=[]
driver=webdriver.Chrome()
for i in range(1,101):
    url=f'https://www.naukri.com/data-analyst-jobs-{i}?k=data%20analyst&experience=1'
    driver.get(url)
    if i==1:
        time.sleep(20)
    else:
        time.sleep(5)
        soup = BeautifulSoup(driver.page_source,'html.parser')
        result=soup.find('div',class_='list')
        df_list.append(to_data_frame(result))
driver.close()
```

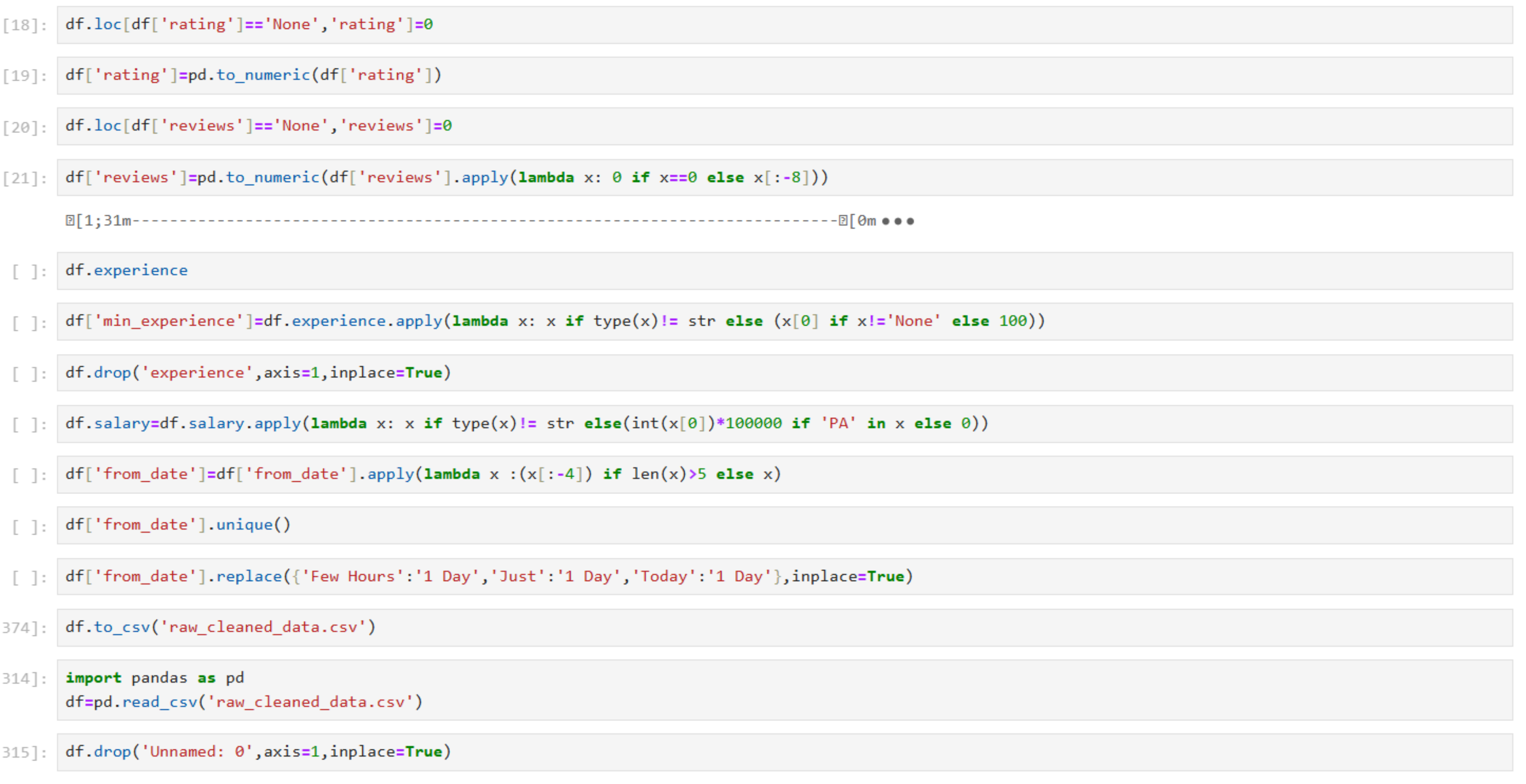
Data cleaning

- Dealing with duplicates



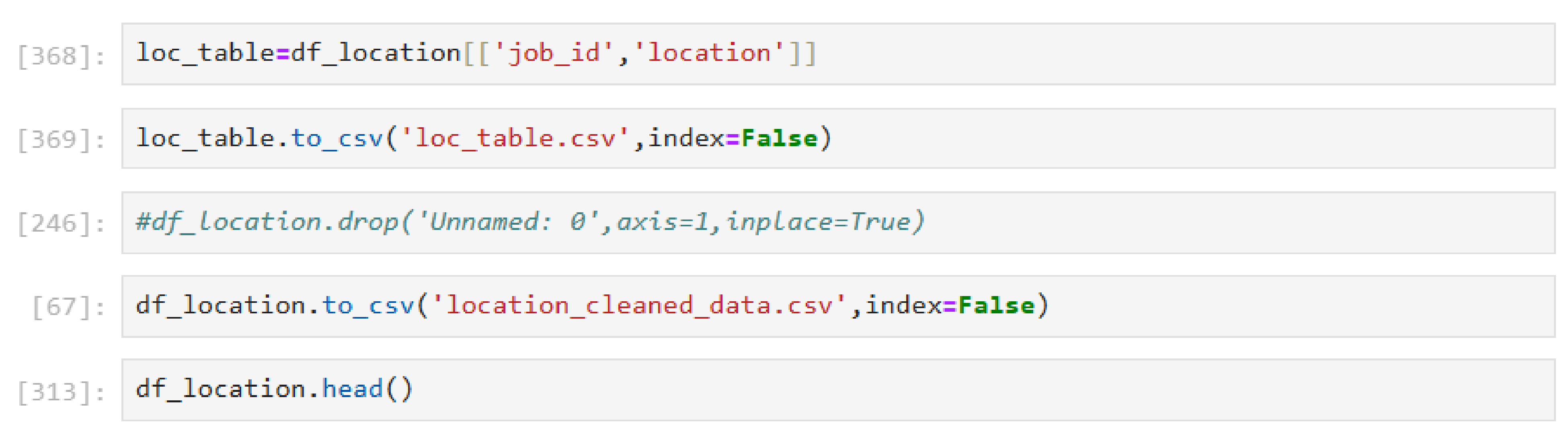
```
[12]: df.skills=df.skills.apply(lambda x: tuple(x))

[13]: df.duplicated().sum()

[13]: 0

[14]: df.drop_duplicates(inplace=True)

[15]: df.shape

[15]: (1973, 10)
```

Changing the data types



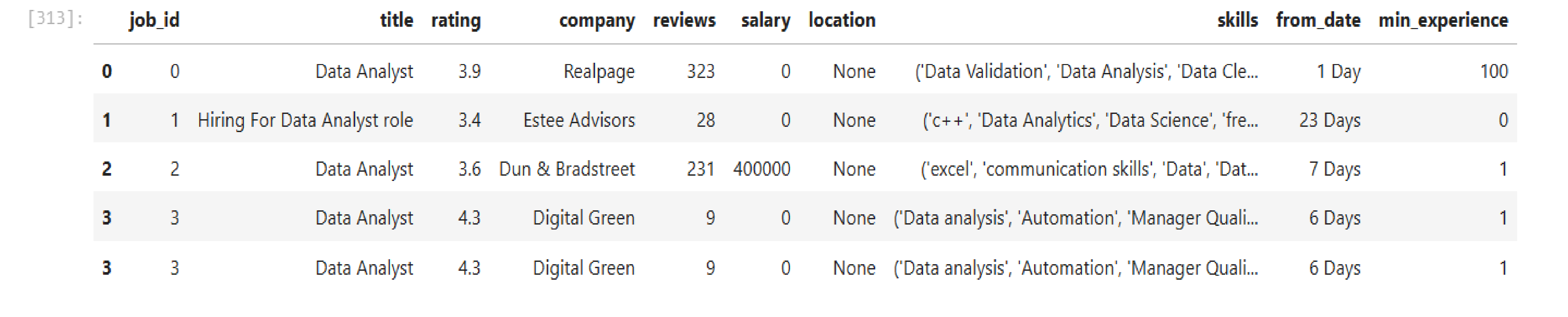
```
[18]: df.loc[df['rating']=='None','rating']=0

[19]: df['rating']=pd.to_numeric(df['rating'])

[20]: df.loc[df['reviews']=='None','reviews']=0

[21]: df['reviews']=pd.to_numeric(df['reviews'].apply(lambda x: 0 if x==0 else x[:-8]))

      □[1;31m---------------------------------------------------------------------------□[0m •••

[ ]: df.experience

[ ]: df['min_experience']=df.experience.apply(lambda x: x if type(x)!= str else (x[0] if x!='None' else 100))

[ ]: df.drop('experience',axis=1,inplace=True)

[ ]: df.salary=df.salary.apply(lambda x: x if type(x)!= str else(int(x[0])*100000 if 'PA' in x else 0))

[ ]: df['from_date']=df['from_date'].apply(lambda x :(x[:-4]) if len(x)>5 else x)

[ ]: df['from_date'].unique()

[ ]: df['from_date'].replace({'Few Hours':'1 Day','Just':'1 Day','Today':'1 Day'},inplace=True)

374]: df.to_csv('raw_cleaned_data.csv')

314]: import pandas as pd
      df=pd.read_csv('raw_cleaned_data.csv')

315]: df.drop('Unnamed: 0',axis=1,inplace=True)
```

Creating a new csv file wit job id and locations



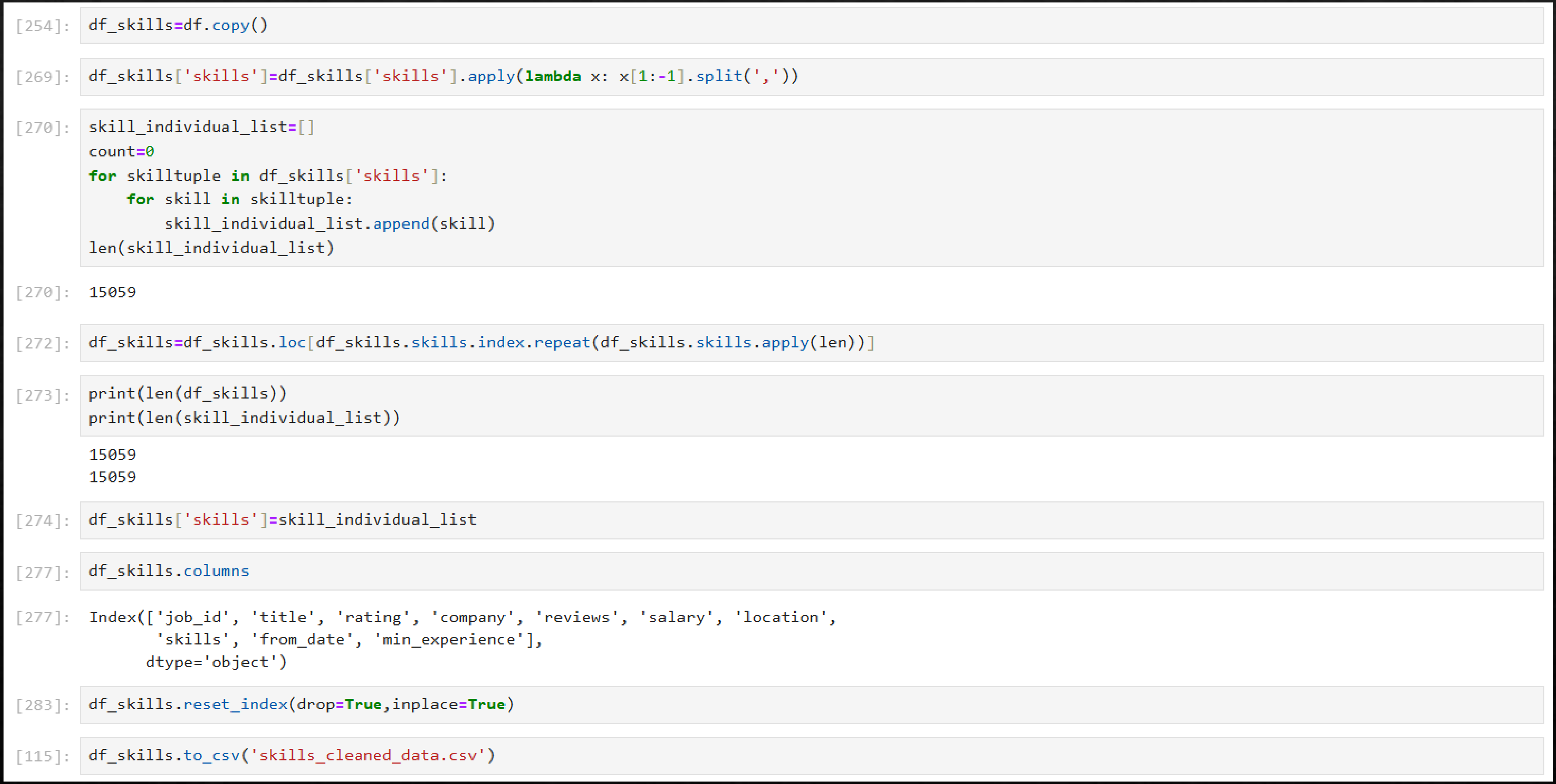
```
[368]: loc_table=df_location[['job_id','location']]

[369]: loc_table.to_csv('loc_table.csv',index=False)

[246]: #df_location.drop('Unnamed: 0',axis=1,inplace=True)

 [67]: df_location.to_csv('location_cleaned_data.csv',index=False)

[313]: df_location.head()
```

Split skills to individual raws

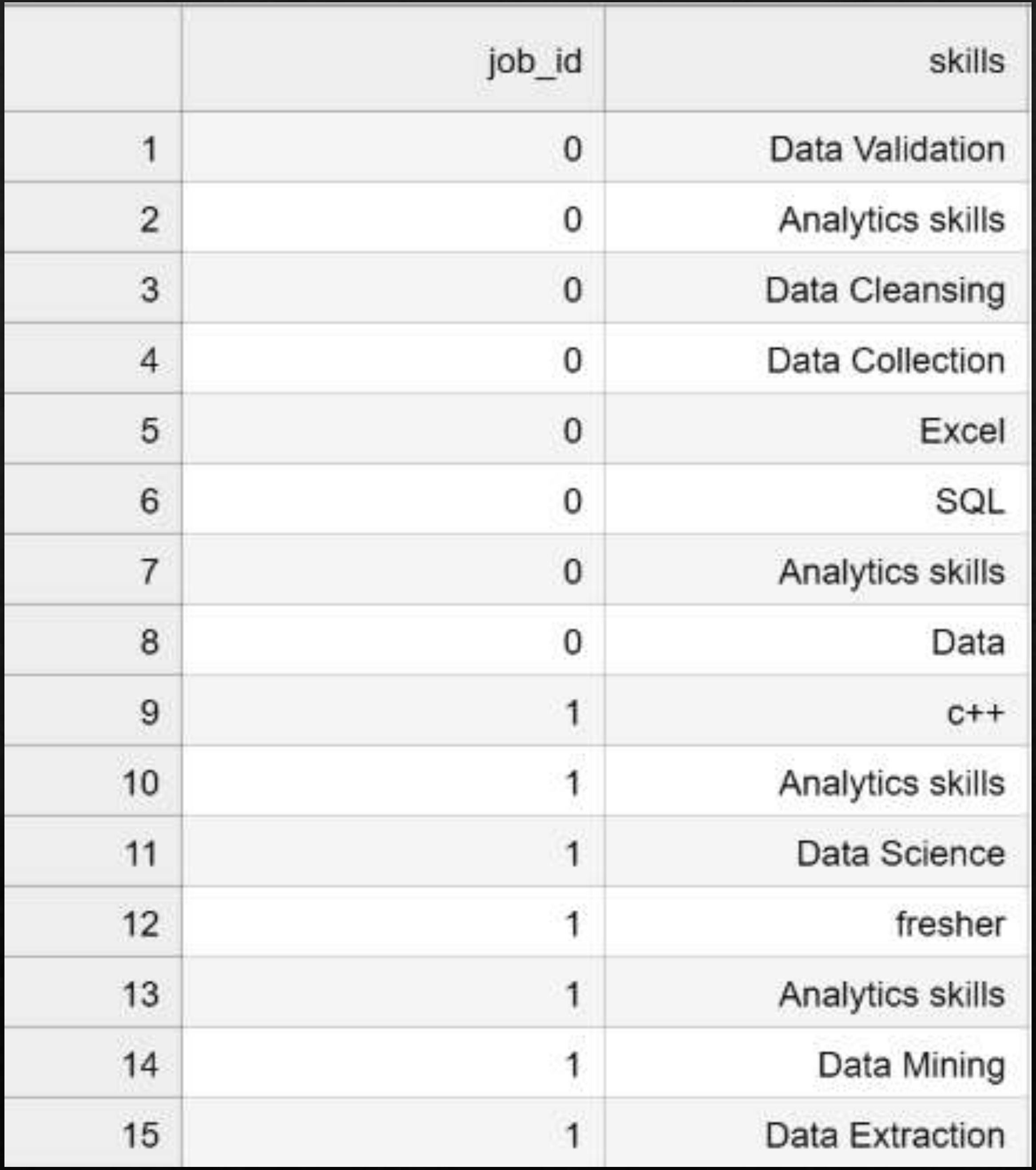


[313]:

	job_id	title	rating	company	reviews	salary	location	skills	from_date	min_experience
0	0	Data Analyst	3.9	Realpage	323	0	None	('Data Validation', 'Data Analysis', 'Data Cle...	1 Day	100
1	1	Hiring For Data Analyst role	3.4	Estee Advisors	28	0	None	('c++', 'Data Analytics', 'Data Science', 'fre...	23 Days	0
2	2	Data Analyst	3.6	Dun & Bradstreet	231	400000	None	('excel', 'communication skills', 'Data', 'Dat...	7 Days	1
3	3	Data Analyst	4.3	Digital Green	9	0	None	('Data analysis', 'Automation', 'Manager Quali...	6 Days	1
3	3	Data Analyst	4.3	Digital Green	9	0	None	('Data analysis', 'Automation', 'Manager Quali...	6 Days	1

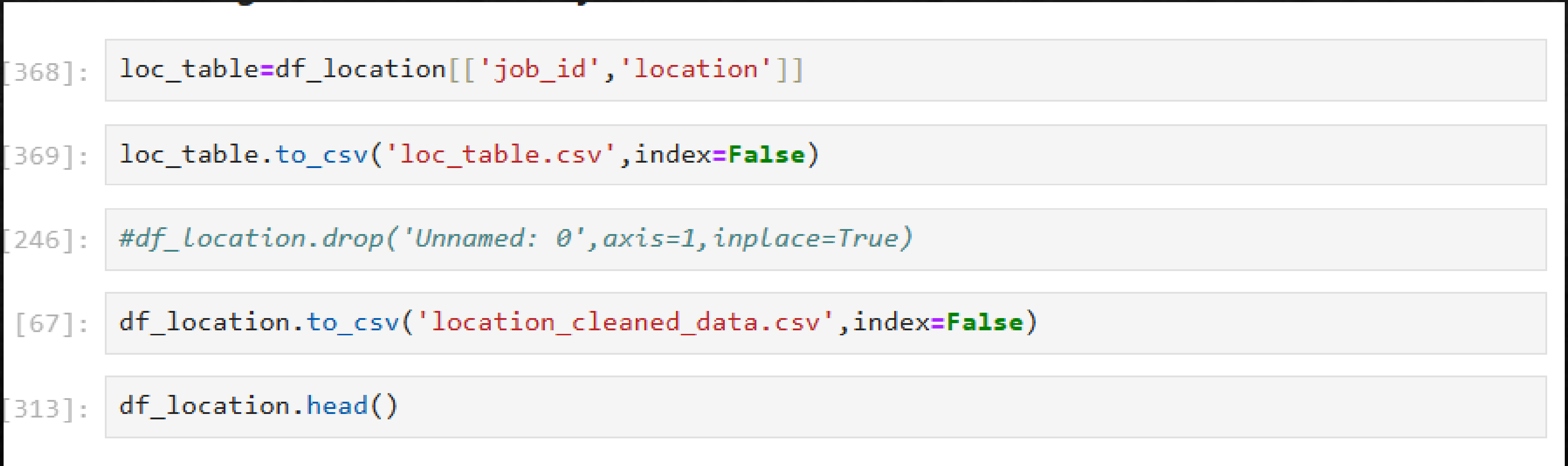


```
[254]: df_skills=df.copy()

[269]: df_skills['skills']=df_skills['skills'].apply(lambda x: x[1:-1].split(','))

[270]: skill_individual_list=[]
       count=0
       for skilltuple in df_skills['skills']:
           for skill in skilltuple:
               skill_individual_list.append(skill)
       len(skill_individual_list)

[270]: 15059

[272]: df_skills=df_skills.loc[df_skills.skills.index.repeat(df_skills.skills.apply(len))]

[273]: print(len(df_skills))
       print(len(skill_individual_list))

       15059
       15059

[274]: df_skills['skills']=skill_individual_list

[277]: df_skills.columns

[277]: Index(['job_id', 'title', 'rating', 'company', 'reviews', 'salary', 'location',
              'skills', 'from_date', 'min_experience'],
             dtype='object')

[283]: df_skills.reset_index(drop=True,inplace=True)

[115]: df_skills.to_csv('skills_cleaned_data.csv')
```

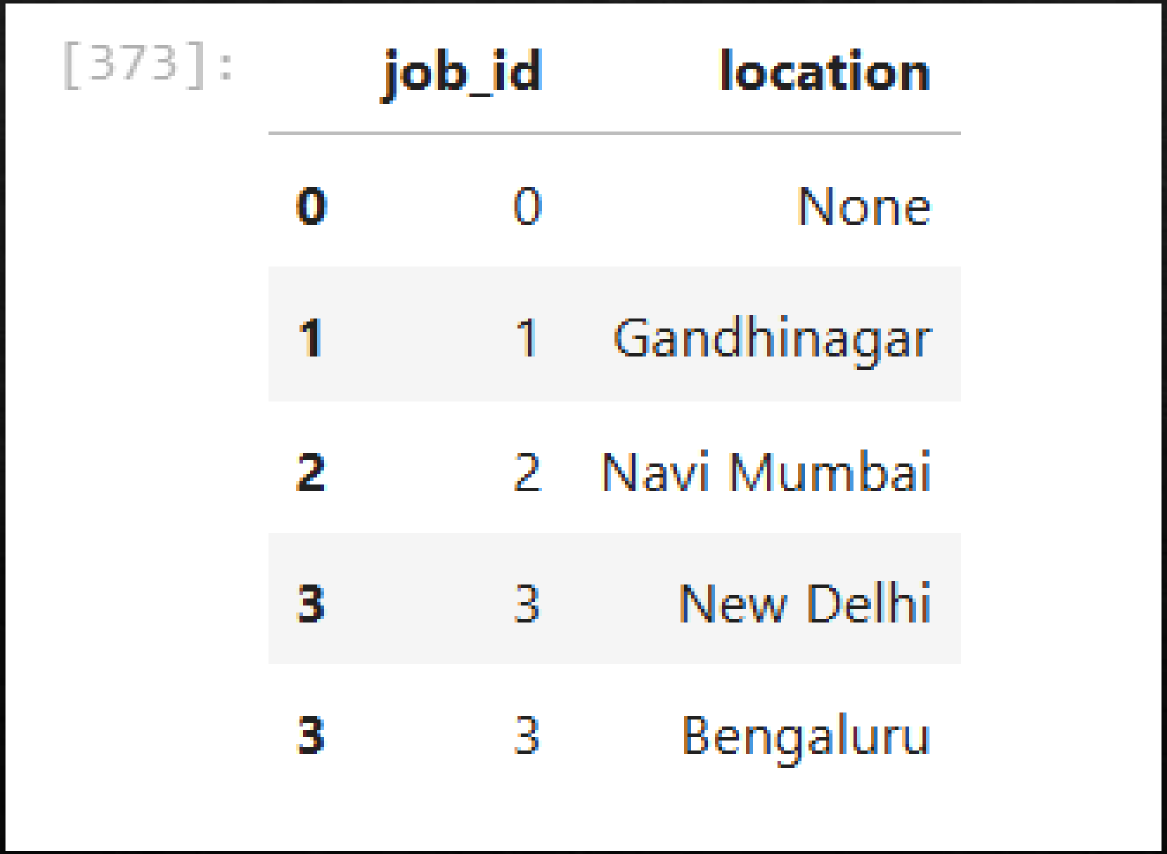


	job_id	skills
1	0	Data Validation
2	0	Analytics skills
3	0	Data Cleansing
4	0	Data Collection
5	0	Excel
6	0	SQL
7	0	Analytics skills
8	0	Data
9	1	c++
10	1	Analytics skills
11	1	Data Science
12	1	fresher
13	1	Analytics skills
14	1	Data Mining
15	1	Data Extraction

Creating a new csv file with job id and locations



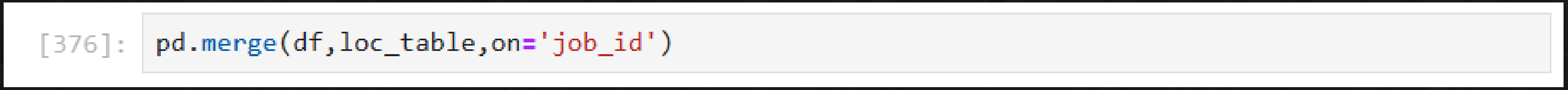
```
[368]: loc_table=df_location[['job_id','location']]

[369]: loc_table.to_csv('loc_table.csv',index=False)

[246]: #df_location.drop('Unnamed: 0',axis=1,inplace=True)

 [67]: df_location.to_csv('location_cleaned_data.csv',index=False)

[313]: df_location.head()
```



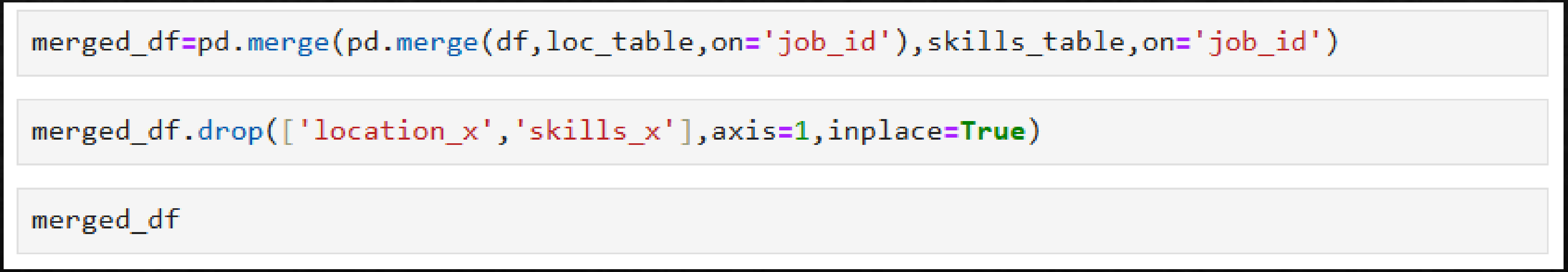
[373]:

	job_id	location
0	0	None
1	1	Gandhinagar
2	2	Navi Mumbai
3	3	New Delhi
3	3	Bengaluru

Joining the data frames



```
[376]: pd.merge(df,loc_table,on='job_id')
```



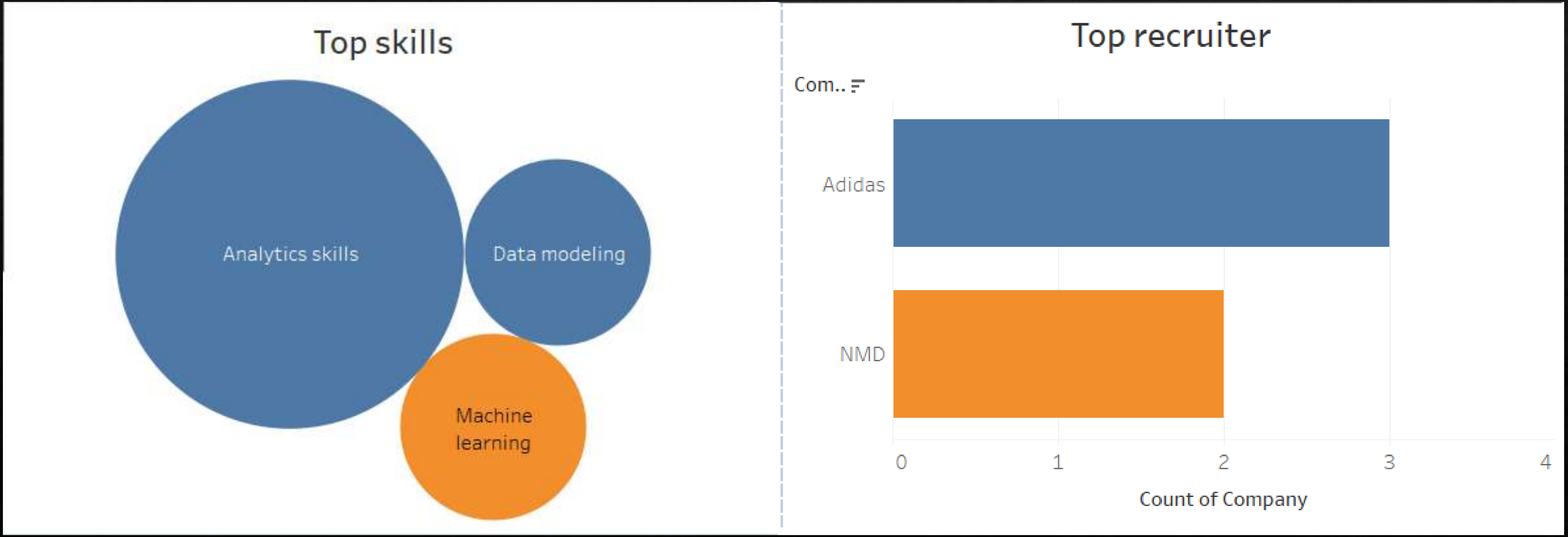
```
merged_df=pd.merge(pd.merge(df,loc_table,on='job_id'),skills_table,on='job_id')
```

```
merged_df.drop(['location_x','skills_x'],axis=1,inplace=True)
```

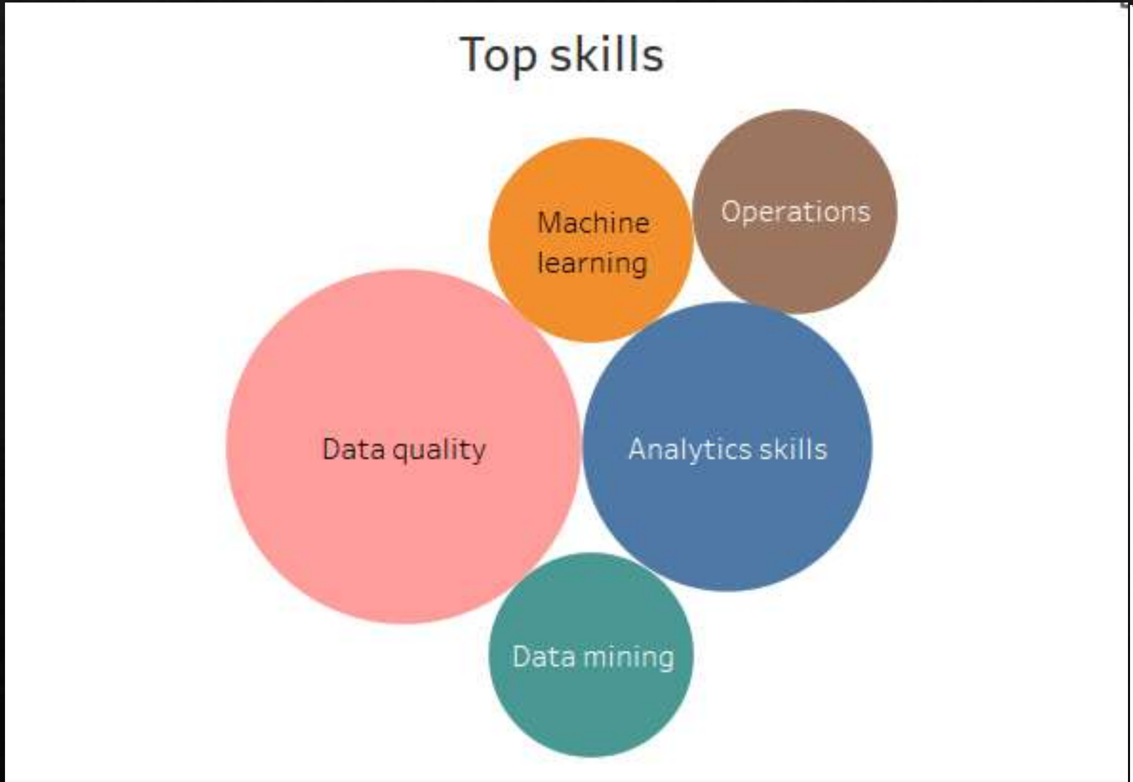
```
merged_df
```

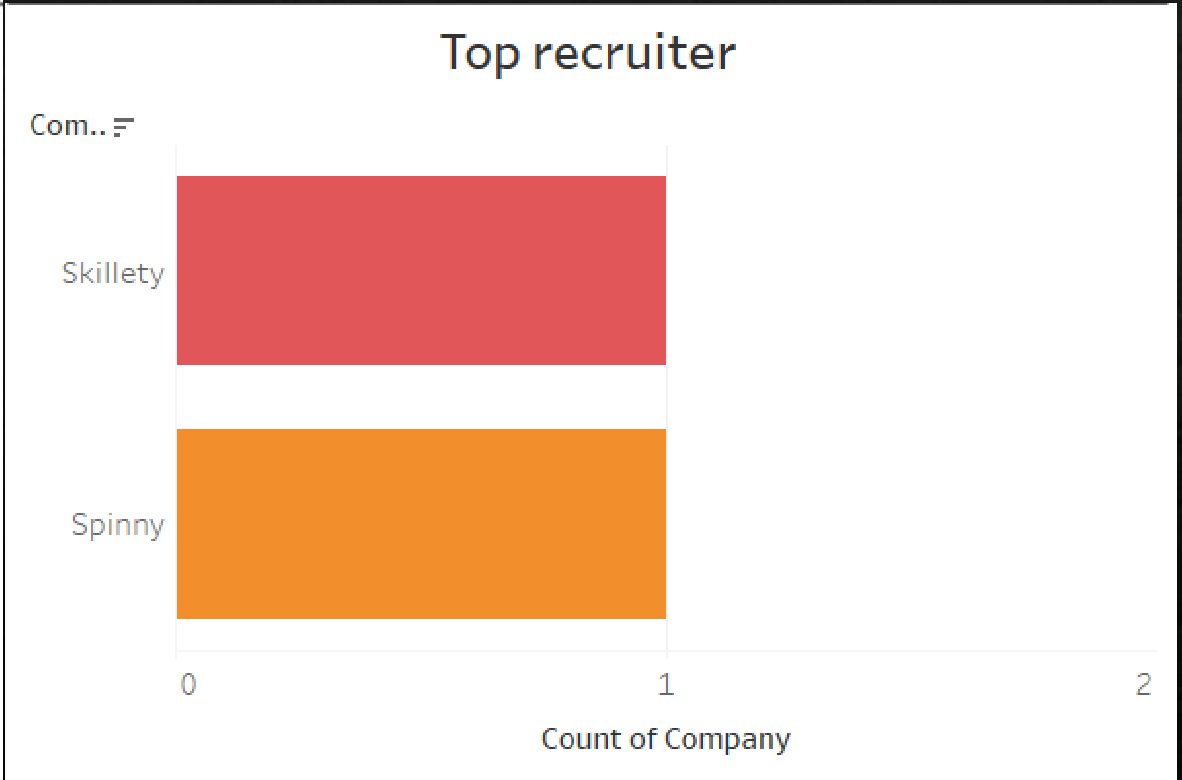
Data Visualization with Tabuleau

Juniur Data sciencist

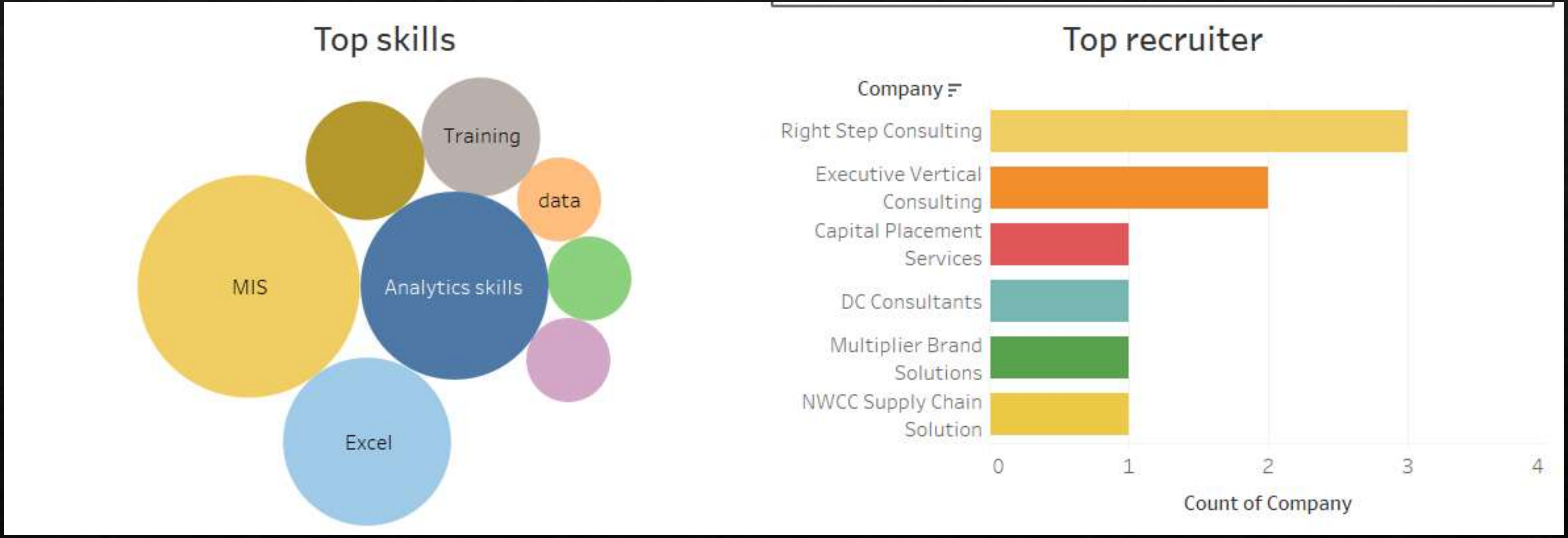


Data engineer


Top skills


Top recruiter

MIS Executive



Top skills

Top recruiter

Thank you