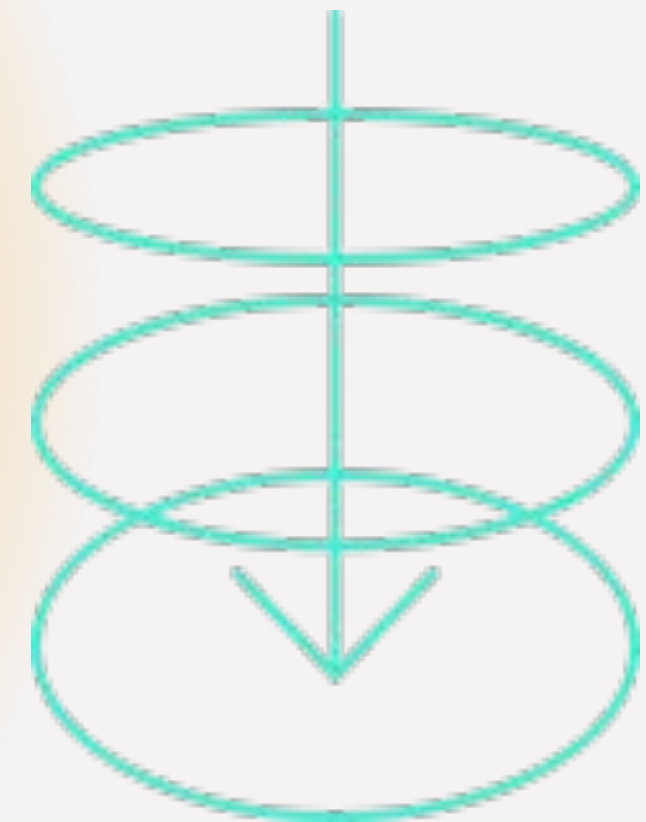


Job Recommender

kmeans classifier



Presenter:

Ahmad Salim AHMADI

16/02/2024

- 01 - Introduction
- 02 - Data Collection
- 03 - Pre Processing
- 04 - Relational Database
- 05 - REST API
- 06 - Data Analysis
- 07 - Classifier
- 08 - Conclusion
- 09 - Challenges



01 - Introduction

- Effect of Covid-19 in job market
- Data related job market after Covid-19
- Job Recommender Can help with finding a job
- Data Analyst, Data Engineer and Data Scientist



02 - Data Collection

- www.indeed.com
- BeautifulSoup
- Selenium

Job Details

- Job title
- URL
- Job ID
- Hiring Company
- Location
- Posted Date
- Salary

Job Requirements

- Job ID
- Job Details/Descriptions

Job Details

```
(705,
[['Data Engineer/Sr Data Engineer',
 'https://www.indeed.com/rc/clk?jk=ffa7381a28d82bad&bb=qs2YMGDdzQ6o9KTF3c38a6Lm6bsyjewJG9wW5pVNa1pDtdbwDdOXPAk9xcKsTLsVb_e
 waDrhdFUHldvRpJ6MgGGREdSwSezomYxJSQHz_Pk%3D&xkcb=SoBJ67M3EV2r59WbHZ0LbzkdCdPP&fccid=621355d16566db19&vjs=3',
 'job_ffa7381a28d82bad',
 'DNV',
 'Remote in Seattle, WA',
 'Data Engineer/Sr Data Engineer',
 'Posted\nJust posted',
 'https://www.indeed.com/rc/clk?jk=ffa7381a28d82bad&bb=qs2YMGDdzQ6o9KTF3c38a6Lm6bsyjewJG9wW5pVNa1pDtdbwDdOXPAk9xcKsTLsVb_e
 waDrhdFUHldvRpJ6MgGGREdSwSezomYxJSQHz_Pk%3D&xkcb=SoBJ67M3EV2r59WbHZ0LbzkdCdPP&fccid=621355d16566db19&vjs=3',
 '$105,000 - $150,000 a year']])
```

Job Requirements

















```
(702,
[{'job_id': 'job_f3f82b5077f60352',
 'job_description': 'About us\nWe are the independent expert in assurance and risk management. Driven by our purpose, to s
afeguard life, property, and the environment, we empower our customers and their stakeholders with facts and reliable insigh
ts so that critical decisions can be made with confidence.\nAs a trusted voice for many of the world's most successful organ
izations, we use our knowledge to advance safety and performance, set industry benchmarks, and inspire and invent solutions
to tackle global transformations.\nAbout the role\nEVOLVE Intelligence accelerates the transition towards a carbon-free futu
re through software and analytics.\nWe are looking for a Data Engineer/Sr. Data Engineer to help us accomplish this missio
n.\nThe Analytics & Data Science team in DNV - Energy Management's Technology group is a remote-first team. We offer more th
an just a job; we provide a community where you can learn, grow, and thrive your way. Join a dynamic and diverse technology
team that values relationships and the environment as much as results. Help us create software that empowers utility clean e
nergy customers to combat climate change!\nThis is a remote position open to any location in the continental United State
s.\nWhat You'll Do:\nAs a Data Engineer/Sr. Data Engineer, you will design, develop, and maintain data architecture, pipelin
es, and systems that play a vital role in how our utility partners steward clean energy programs. Your impact will be immedi
ate and will directly enable pathways to decarbonization through energy efficiency, demand response, storage, electric vehic
les, and renewable energy technologies. You will solve a variety of problems that leverage your deep understanding of data e
ngineering principles.\nHow You'll Succeed:\nCollaborate with cross-functional teams, including machine learning engineers,
software developers, analytics engineers, and product managers to translate business requirements into highly available data
solutions\nLeverage your creative problem-solving skills to architect, develop, and maintain scalable and efficient data pro
cessing pipelines using PySpark and other distributed computing technologies\nCreate and optimize data models that support r
eporting, analytics, artificial intelligence, and software\nOptimize data storage and retrieval by designing and implementin
g efficient storage systems that use technologies such as Timescale and Apache Spark\nApply data validation, data profiling,
```


02 - Data Collection

One week of Scraping

- Arizona
- California
- Michigan
- Ohio
- Washington

 washington_da.csv
 washington_da_jd.csv
 washington_de.csv
 washington_de_jd.csv
 washington_ds.csv
 washington_ds_jd.csv

 arizona_da.csv
 arizona_da_jd.csv
 arizona_de.csv
 arizona_de_jd.csv
 arizona_ds.csv
 arizona_ds_jd.csv
 clustered_skills.csv
 IndeedBot.log
 kmeans_model.pickle
 kmeans_model.pkl
 los_angeles_da.csv
 los_angeles_da_jd.csv
 los_angeles_de.csv
 los_angeles_de_jd.csv
 los_angeles_ds.csv
 los_angeles_ds_jd.csv

 michigan_da.csv
 michigan_da_jd.csv
 michigan_de.csv
 michigan_de_jd.csv
 michigan_ds.csv
 michigan_ds_jd.csv
 new_york1.csv
 ohio_da.csv
 ohio_da_jd.csv
 ohio_de.csv
 ohio_de_jd.csv
 ohio_ds.csv
 ohio_ds_jd.csv
 orlando1.csv
 orlando1_jd.csv

03 - Pre Process

- Handling duplicates
- Handling null values
- Handle date formats
- Filtering values
- Renaming columns
- Creating columns
- Dropping columns when necessary
- Dropping rows when necessary
- Verifying that the datatypes in the DataFrame were accurate, if not, handle conversion
- String formatting (strip, replace methods)
- Concatenation for joining multiple DataFrames
- Indexing

	job_title	url	job_id	company	location	additional_info	posted_date	full_url	salary	state	days_ago	average_salary
0	Board Certified Behavior Analyst (FT BCBA)	https://www.indeed.com/rc/clk?jk=ebd87097db3af...	job_ebd87097db3afc8e	Autism Spectrum Therapies	Pomona, CA 91766	Board Certified Behavior Analyst (FT BCBA)	Posted\nJust posted	https://www.indeed.com/rc/clk?jk=ebd87097db3af...	71,250—82,000 a year	CA	0	76625.0
1	Data Analyst - Health, Principal	https://www.indeed.com/rc/clk?jk=96b9457e9bc11...	job_96b9457e9bc1131e	Blue Shield of California	Woodland Hills, CA 91367	Data Analyst - Health, Principal	Posted\nToday	https://www.indeed.com/rc/clk?jk=96b9457e9bc11...	136,400—204,600 a year	CA	0	170500.0
2	nCino Business Analyst Senior	https://www.indeed.com/rc/clk?jk=f1bee190e67cbf0f	job_f1bee190e67cbf0f	City National Bank	Los Angeles, CA 90071	nCino Business Analyst Senior	Posted\nToday	https://www.indeed.com/rc/clk?jk=f1bee190e67cb...	92,114—156,880 a year	CA	0	124497.0
3	Research Analyst, Marketing Sciences, Innovation	https://www.indeed.com/rc/clk?jk=dc0507a8cb2c1...	job_dc0507a8cb2c1b3a	Ipsos	Culver City, CA	Research Analyst, Marketing Sciences, Innovation	Posted\nToday	https://www.indeed.com/rc/clk?jk=dc0507a8cb2c1...	66,000—68,500 a year	CA	0	67250.0
4	Data Quality Analyst	https://www.indeed.com/rc/clk?jk=02e70cbcff3d8c2c	job_02e70cbcff3d8c2c	Prime Healthcare Management Inc	Ontario, CA 91764	Data Quality Analyst	Posted\nToday	https://www.indeed.com/rc/clk?jk=02e70cbcff3d8...	NaN	CA	0	NaN

	job_id	job_description
0	job_96b9457e9bc1131e	JOB DESCRIPTION\nYour Role\nThe Network Perfor...
1	job_f1bee190e67cbf0f	Overview:\nNCINO BUSINESS ANALYST SENIOR\n\nWH...
2	job_ebd87097db3afc8e	Overview:\nWe're looking for...\n\nBright,\n\n...
3	job_02e70cbcff3d8c2c	Overview:\nPrime Healthcare is an award-winnin...
4	job_1907904fc85ed964	About Codazen\nWant to apply technology in way...

03 - Pre Process

```

job_id      object
sql         int64
python      int64
power_bi    int64
tableau     int64
api         int64
etl         int64
elt         int64
nosql       int64
docker      int64
hadoop      int64
excel       int64
nlp         int64
experience  int64
data_warehouse int64
spark       int64
kafka       int64
airflow     int64
linux       int64
ml          int64
azure       int64
aws         int64
google_cloud int64
bdegree     int64
mdegree     int64
dtype: object

```

	job_id	sql	python	power_bi	tableau	api	etl	elt	nosql	docker	...	spark	kafka	airflow	linux	ml	azure	aws	google_cloud	bdegree	mdegree
697	job_627e86f337a568f5	1	1	0	0	0	1	0	1	1	...	1	1	0	0	1	1	1	0	1	1
698	job_81f380eb69c9453c	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
699	job_4e796c8c0eedbfed	1	1	0	0	0	0	0	0	0	...	1	0	0	0	1	0	0	0	0	0
700	job_238ec6b39c6f242c	0	1	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
701	job_f5d6430ea04a9d9b	0	1	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	1

5 rows x 25 columns

	job_id	job_description
0	job_96b9457e9bc1131e	JOB DESCRIPTION\nYour Role\nThe Network Perfor...
1	job_f1bee190e67cbf0f	Overview:\nNCINO BUSINESS ANALYST SENIOR\n\nWH...
2	job_ebd87097db3afc8e	Overview:\nWe're looking for...\n\nBright,\n\n...
3	job_02e70cbcff3d8c2c	Overview:\nPrime Healthcare is an award-winnin...
4	job_1907904fc85ed964	About Codazen\nWant to apply technology in way...

04 - Database

```
mysql> CREATE TABLE job_announcement (  
-> job_id VARCHAR(255) PRIMARY KEY,  
-> job_title VARCHAR(255),  
-> url VARCHAR(255),  
-> company VARCHAR(255),  
-> location VARCHAR(255),  
-> additional_info VARCHAR(255),  
-> posted_date VARCHAR(255),  
-> full_url VARCHAR(255),  
-> job_details TEXT,  
-> days_ago INT(11)  
-> );
```

Query OK, 0 rows affected (0.02 sec)

```
mysql>
```

```
mysql> CREATE TABLE job_description (  
-> job_id VARCHAR(255) PRIMARY KEY,  
-> job_description TEXT  
-> );
```

Query OK, 0 rows affected (0.02 sec)

```
mysql>
```

```
mysql> CREATE TABLE job_skills (  
-> job_id VARCHAR(255) PRIMARY KEY,  
-> sql_skill INT,  
-> mysql INT,  
-> python INT,  
-> power_bi INT,  
-> tableau INT,  
-> api INT,  
-> etl INT,  
-> elt INT,  
-> nosql INT,  
-> docker INT,  
-> hadoop INT,  
-> eda INT,  
-> machine_learning INT,  
-> excel INT,  
-> nlp INT,  
-> experience INT  
-> );
```

Query OK, 0 rows affected (0.02 sec)

```
mysql> CREATE TABLE job_states (  
-> job_id varchar(255) PRIMARY KEY,  
-> state varchar(24)  
-> );
```

Query OK, 0 rows affected (0.02 sec)

```
mysql> show tables;
```

```
+-----+  
| Tables_in_kweekly_ironhack |  
+-----+  
| job_announcement           |  
| job_description            |  
| job_skills                  |  
| job_states                  |  
+-----+  
4 rows in set (0.00 sec)
```

```
mysql>
```

```
mysql> select count(*) from job_announcement;  
+-----+  
| count(*) |  
+-----+  
|      3105 |  
+-----+  
1 row in set (0.01 sec)
```

```
mysql> select count(*) from job_skills;  
+-----+  
| count(*) |  
+-----+  
|      2752 |  
+-----+  
1 row in set (0.00 sec)
```


04 - Database

📌 kweekly_ironhack job_announcement
🔑 job_id : varchar(255)
📄 job_title : varchar(255)
📄 url : varchar(255)
📄 company : varchar(255)
📄 location : varchar(255)
📄 additional_info : varchar(255)
📄 posted_date : varchar(255)
📄 full_url : varchar(255)
average_salary : float
days_ago : int(11)
📄 state : varchar(255)
📄 salary : varchar(255)

📌 kweekly_ironhack job_description
🔑 job_id : varchar(255)
📄 job_description : text

📌 kweekly_ironhack job_states
🔑 job_id : varchar(255)
📄 state : varchar(24)

📌 kweekly_ironhack job_skills
🔑 job_id : varchar(255)
sql_skill : int(11)
python : int(11)
power_bi : int(11)
tableau : int(11)
api : int(11)
etl : int(11)
elt : int(11)
nosql : int(11)
docker : int(11)
hadoop : int(11)
ml : int(11)
excel : int(11)
nlp : int(11)
experience : int(11)
data_warehouse : int(11)
spark : int(11)
kafka : int(11)
airflow : int(11)
linux : int(11)
azure : int(11)
aws : int(11)
google_cloud : int(11)
bdegree : int(11)
mdegree : int(11)



05 - REST API

End Points:

1. /jobs
2. /jobs/job_id

1. /jobs?page=18&items_per_page=2

```
127.0.0.1:8080/jobs?page=18&items_per_page=2

{
  "jobs": [
    {
      "days_ago": 0,
      "job_id": "job_95c81037cc63ef9f",
      "job_title": "Data Engineer III",
      "location": "Hudson, OH 44236",
      "skills": [
        "sql_skill",
        "excel",
        "experience",
        "data_warehouse",
        "azure",
        "bdegree"
      ],
      "url": "https://www.indeed.com/rc/clk?jk=95c81037cc63ef9f&bb=eYPFN53zFyoNtuXhGn7Zf_HxaqWlpPrHGq8rR9cgezKOKnyneIw5_MLz1dMh1IX_XCzbxQjn9RHpKMkcL73d1Y8qsJbD1U6cScnpj59-AcE%3D&xkcb=SoAi67M3ERAqoCgLAB0IbzkCdPP&fccid=7a5ca6215f560ef3&vjs=3"
    },
    {
      "days_ago": 0,
      "job_id": "job_53ca8933e4149d0d",
      "job_title": "Data Engineer",
      "location": "Cincinnati, OH",
      "skills": [
        "api",
        "etl",
        "nosql",
        "ml",
        "experience",
        "bdegree"
      ],
      "url": "https://www.indeed.com/rc/clk?jk=53ca8933e4149d0d&bb=eYPFN53zFyoNtuXhGn7Zf80HC1CZW0rmQzEv-B1A0Uh4MeIrnT30tCI6A8HyZTdP4UjZ-FQpAyg-eJ6yQPd9NYxwzvfH4Xgd0TMRmyBpdg%3D&xkcb=SoCs67M3ERAqoCgLAB0PbzkCdPP&fccid=f51b75d687e40b14&vjs=3"
    }
  ],
  "last_page": "/jobs?page=1&items_per_page=2",
  "next_page": null
}
```


05 - REST API

End Points:

1. /jobs
2. /jobs/job_id

1. /jobs?page=18&items_per_page=2
2. 127.0.0.1:8080/jobs/job_95c81037cc63ef9f

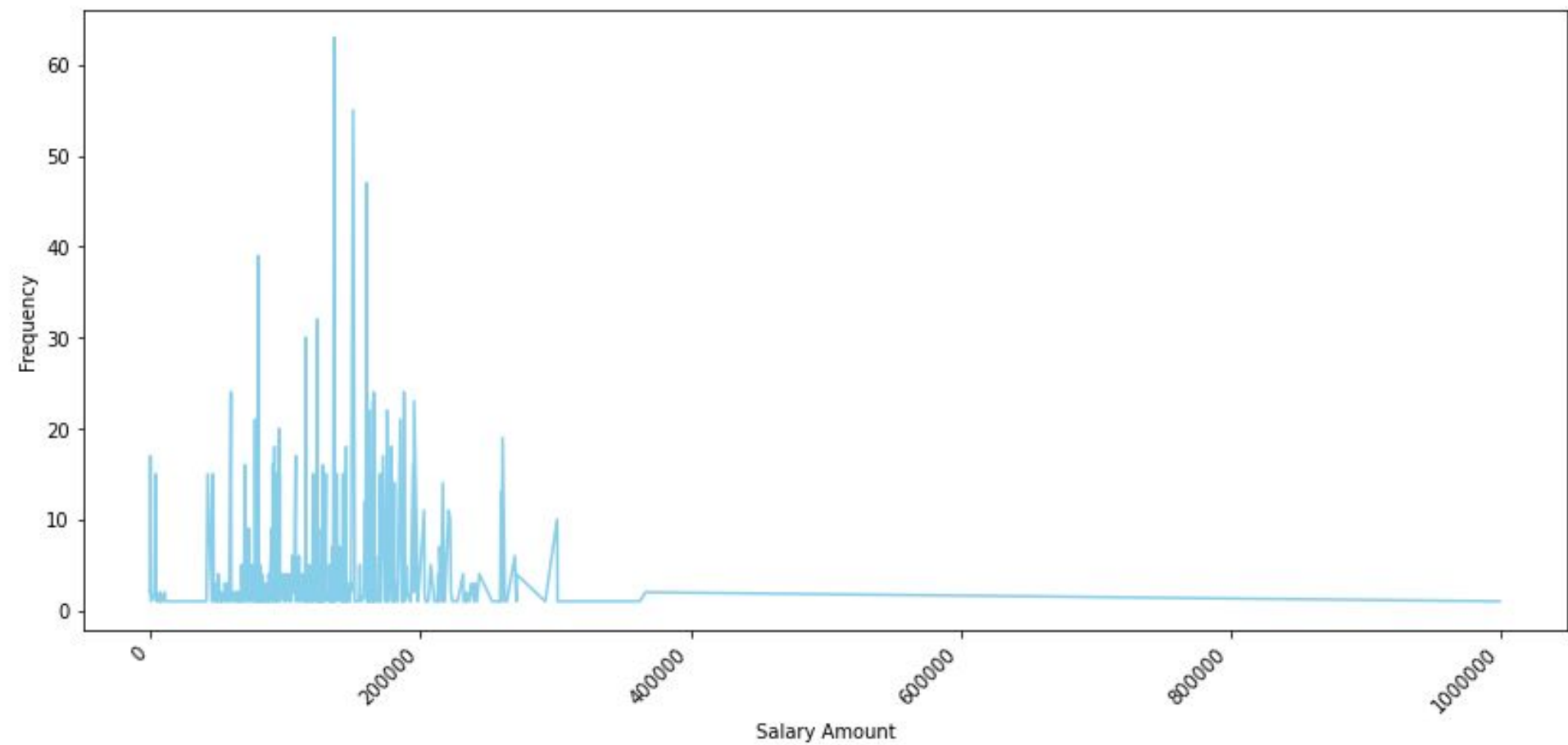


The screenshot shows a web browser window with the address bar displaying the URL `127.0.0.1:8080/jobs/job_95c81037cc63ef9f`. The browser's taskbar at the top includes icons for WhatsApp, a POST request, OFPRA, AfghanTV, a Watch icon, Domicile, A/B Testing?, and a recipe site. The main content area displays a JSON response from the API, which is a job listing for a 'Data Engineer III' position in Hudson, OH. The response includes details about the job title, location, skills (SQL, Excel, Experience, Data Warehouse, Azure, and Bachelor's degree), and a link to the job on Indeed.

```
{
  "days_ago": 0,
  "job_title": "Data Engineer III",
  "location": "Hudson, OH 44236",
  "skills": [
    "sql_skill",
    "excel",
    "experience",
    "data_warehouse",
    "azure",
    "bdegree"
  ],
  "url": "https://www.indeed.com/rc/clk?jk=95c81037cc63ef9f&bb=eYPFN53zFyoNtuXhGn7Zf_HxaqWlpPrHGq8rR9cgezKOKnyneIw5_MLz1dMh1IX_XCzbxQjn9RHpKMkcL73d1Y8qsJbD1U6cScnpj59-AcE%3D&xkcb=SoAi67M3ERAqoCgLAB0IbzkdCdPP&fccid=7a5ca6215f560ef3&vjs=3"
}
```

06 - Data Analysis

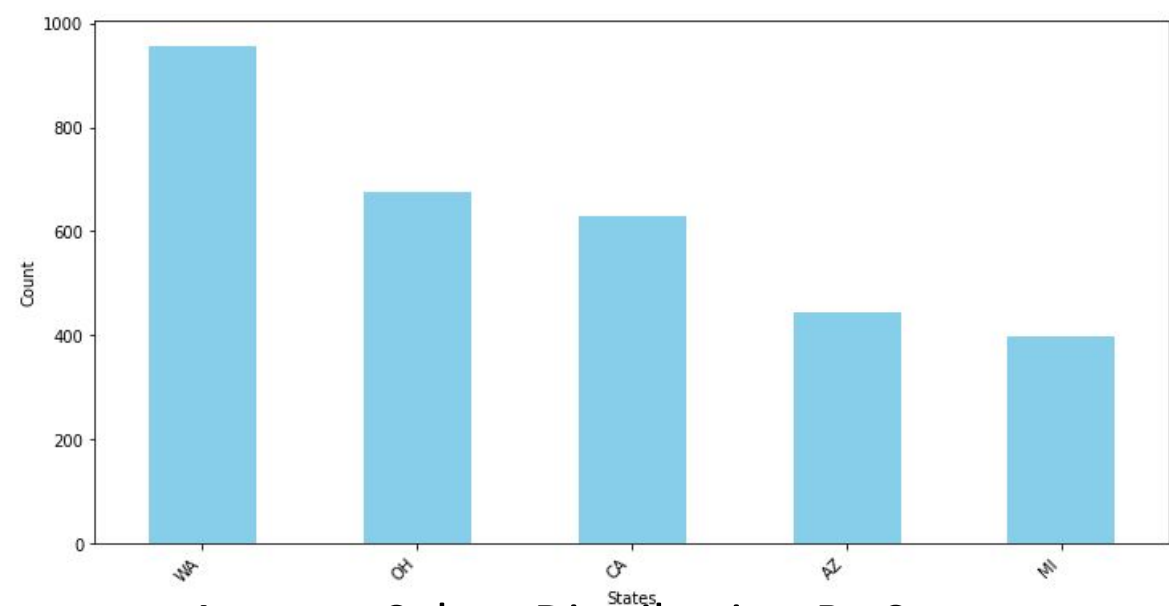
Salary Distribution



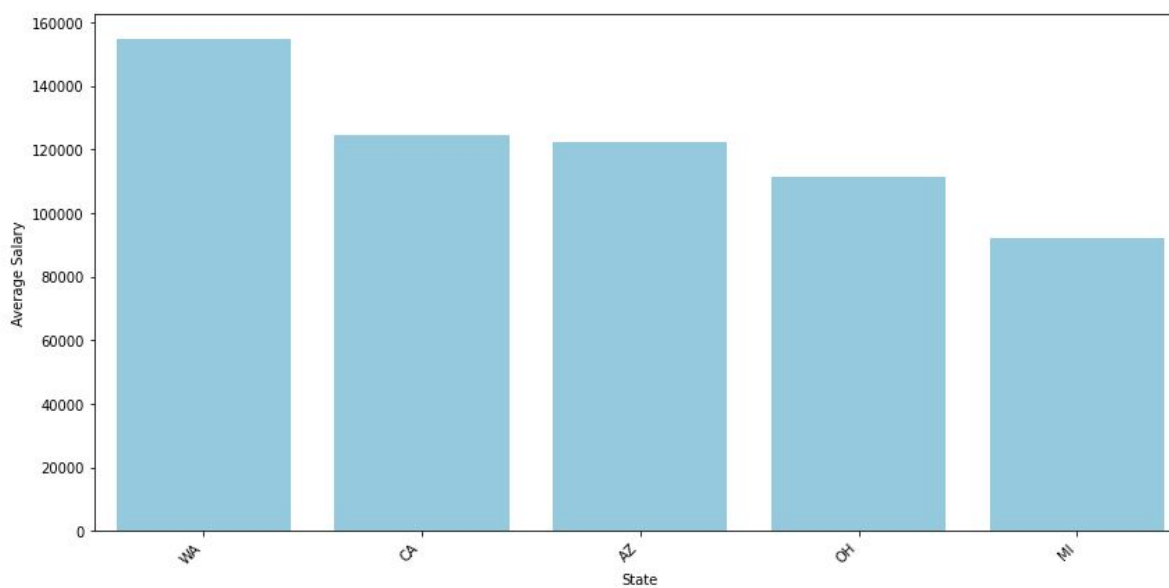
```
df[df['average_salary'] > 400000][['job_id', 'job_title', 'average_salary']]
```

	job_id	job_title	average_salary
88	job_5933df26f5f39396	Vice President, Data Center and AI, General Ma...	999499.5

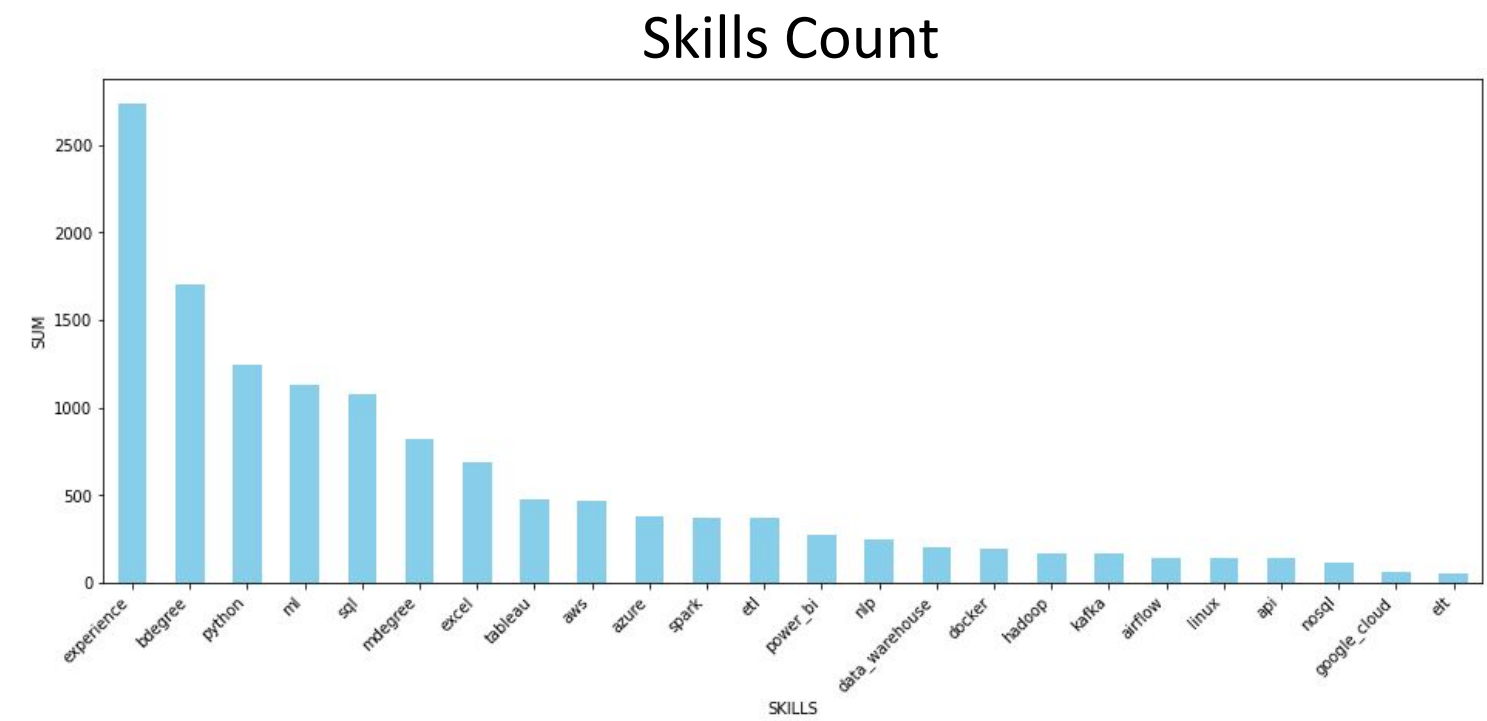
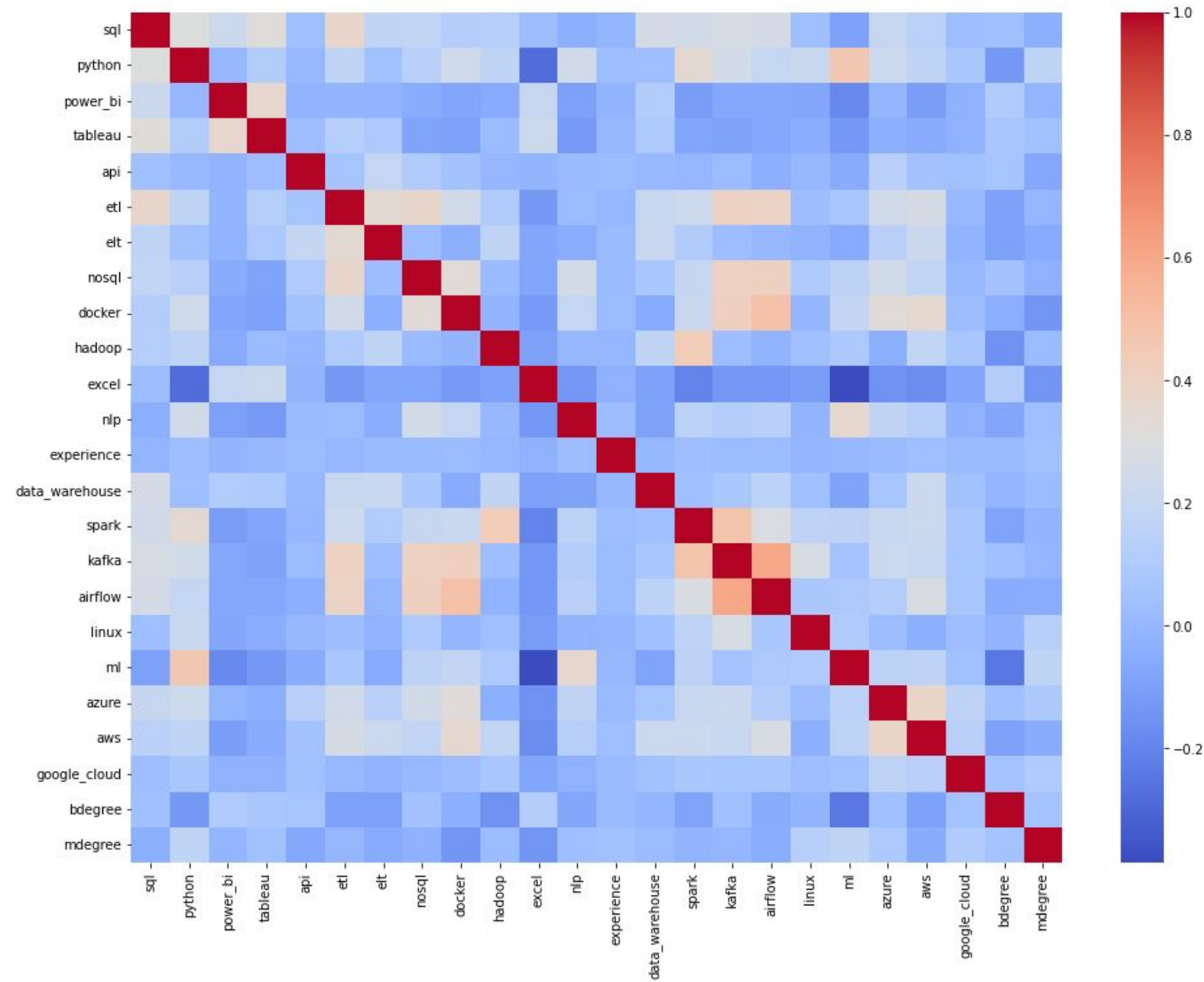
Job Distribution By State



Average Salary Distribution By State

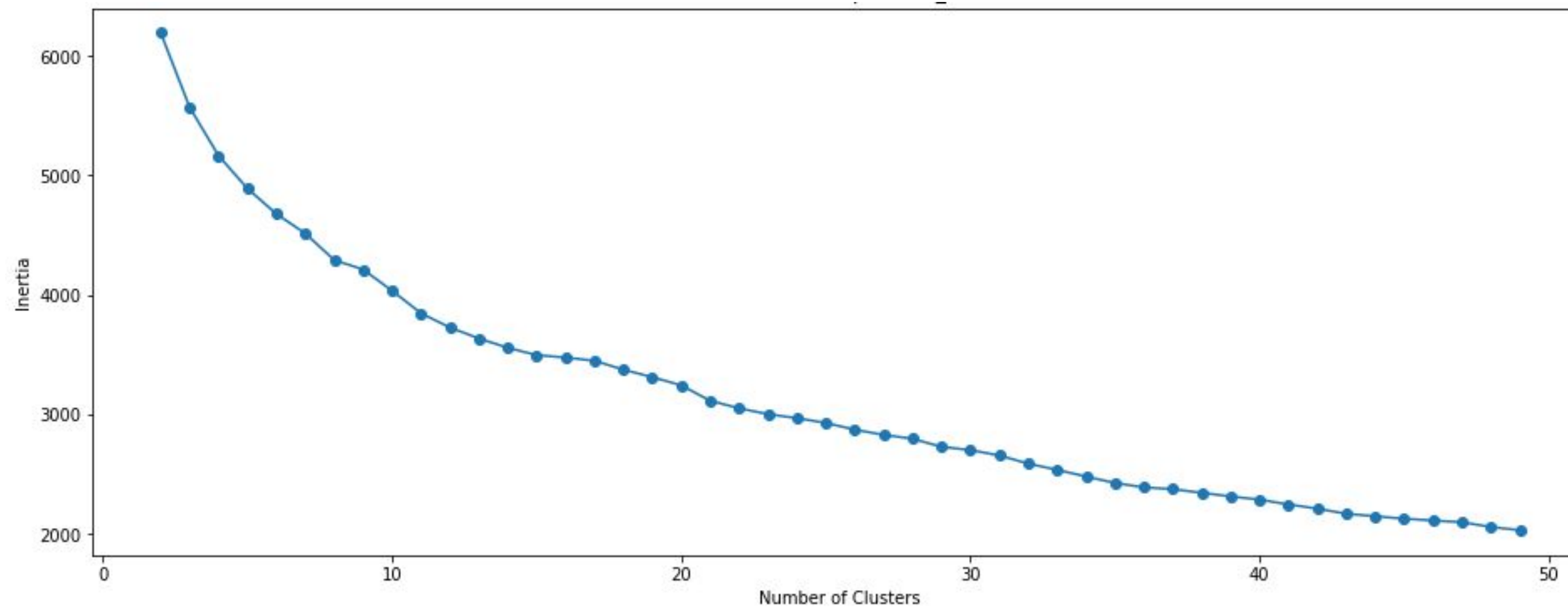


06 - Data Analysis



07 - Classifier - Optimal Cluster Number

Elbow Method



Silhouette Score

Calinski-Harabasz Index	calinski_harabasz_score_value
2.0	499.280658
3.0	431.482742
4.0	381.910028
5.0	341.124189
6.0	310.062916
7.0	284.255411
8.0	276.580740
9.0	252.848819
10.0	248.055304
11.0	247.518181
12.0	240.143619
13.0	231.352025
14.0	222.775903
15.0	213.872843
16.0	201.780875
24.0	173.949537
25.0	170.465831
26.0	168.976386
27.0	166.502035
28.0	163.335568
29.0	163.775668
30.0	160.590173
31.0	159.356432
32.0	160.582391
33.0	160.432979
34.0	160.927820
35.0	161.359555
36.0	160.281468
37.0	157.252907
38.0	155.930542
39.0	154.692390

07 - Classifier - kmeans

Type the numbers corresponding to your skills (e.g., 1, 3, 5):

Enter your skills separated by commas:

- 1. sql
- 2. python
- 3. power_bi
- 4. tableau
- 5. api
- 6. etl
- 7. elt
- 8. nosql
- 9. docker
- 10. hadoop
- 11. excel
- 12. nlp
- 13. experience
- 14. data_warehouse
- 15. spark
- 16. kafka
- 17. airflow
- 18. linux
- 19. ml
- 20. azure
- 21. aws
- 22. google_cloud
- 23. bdegree
- 24. mdegree

Type the numbers corresponding to your skills (e.g., 1, 3, 5): 1, 2, 3, 4, 11
Top 3 recommended job_ids: ['job_9744e7bcb2fdc67a', 'job_ec66f16be7abecca', 'job_9c251dc167832c6a']

	job_title	company	location	url
224	Management Analyst (Public Works)	City of Cypress, CA	Cypress, CA	https://www.indeed.com/rc/clk?jk=ec66f16be7abe...
225	Business Analyst - Planning	Funko	Burbank, CA 91505	https://www.indeed.com/rc/clk?jk=9744e7bcb2fdc...
227	Senior Analyst, Digital & Customer Insights an...	Chipotle	Newport Beach, CA 92660	https://www.indeed.com/rc/clk?jk=9c251dc167832...



Input skills

Euclidean Distance to Item 1: 1.0
Euclidean Distance to Item 2: 1.0
Euclidean Distance to Item 3: 1.0



recommended 3 job_id



details of 3 job_id

07 - Classifier - kmeans

Type the numbers corresponding to your skills (e.g., 1, 3, 5):

Enter your skills separated by commas:

1. sql
2. python
3. power_bi
4. tableau
5. api
6. etl
7. elt
8. nosql
9. docker
10. hadoop
11. excel
12. nlp
13. experience
14. data_warehouse
15. spark
16. kafka
17. airflow
18. linux
19. ml
20. azure
21. aws
22. google_cloud
23. bdegree
24. mdegree



Input skills

Euclidean Distance to Item 1: 3.3166247903554
Euclidean Distance to Item 2: 3.3166247903554
Euclidean Distance to Item 3: 3.3166247903554



recommended 3 job_id



details of 3 job_id

Type the numbers corresponding to your skills (e.g., 1, 3, 5): 1, 6, 8, 9, 16, 17, 23
Top 3 recommended job_ids: ['job_1540b8845a60d26a', 'job_0df4cfd9767d3b59', 'job_a7d515dac2c449c3']

	job_title	company	location	url
20	Machine Learning Algorithm Engineer	Transit Pro Tech Inc	West Covina, CA 91791	https://www.indeed.com/rc/clk?jk=1540b8845a60d...
108	Technical Artist, Synthetic Data	Meta	Los Angeles, CA 90066 \n(Del Rey area)	https://www.indeed.com/rc/clk?jk=0df4cfd9767d3...
138	Machine Learning Research Scientist	d-Matrix	Hybrid remote in Ontario, CA	https://www.indeed.com/rc/clk?jk=a7d515dac2c44...

08 - Conclusion

- A tense and challenging project
- Potential enhancements or future directions for job recommender system.
 - Input CV instead of skills
 - Change level of importance of each feature and avoid binary features
 - streamlit or web app

09 - Challenges

- Lack of time
- Scraping was quite slow
- Scraping gives lots of errors
- Scraping data comes with lots of noises hence cleaning is time consuming comparing to API

Thank You!

