PROJECT 3 – DATA SCIENCE JOBS MARKET ANALYSIS

Collaborators:

Chuck Bui

Amanuel Lebassi

Beau Massie

Christopher Turner

Project 3 – Data Science Jobs Market Analysis

Scope and Criteria

The purpose of this project is to design and implement an ETL pipeline that processes our data.

Answer the questions:

- Are there any specific data science roles that are more prevalent?
- How do these roles differ in terms of salary and location?
- How does salary vary by experience level
- Is there a correlation between location and salary range offered?
- Are there any trends in salary offered based on company rating?

Key Metrics:

Available positions, salaries, location, company rating

Resources:

Kaggle data set with 485 rows of data in 8 columns

QUICK TAKE

- The "Cleaned Data Science Job Market & Salaries 2024" was chosen because it had many relevant key attributes that we were interested in, such as salary data, job titles, company names and ratings.
- We cleaned our database and created new columns like Job Category from inside the 'Job Title' column and
- Created an API call for OMDB to cross reference with the cleaned data set by utilizing IMDB ID's, which is a universal standard for movie data. This was needed for ratings and number of votes data.

LOAD THE DATA AND BEGIN CLEANING

[2]:
```
# import libraries
import csv
import pandas as pd
```

[3]:
```
file_path = 'data/clean_data.csv'
df = pd.read_csv(file_path)
df.head()
```

[3]:

	Job Title	Company Name	Location	Date	Job Link	Company Rating	Min Salary	Max Salary
0	Associate Stop Loss Underwriter	The Insurance Center\n2.7	Onalaska, WI	30d+	https://www.glassdoor.com/partner/jobListing.h...	2.7	57.0	84.0
1	Manager of Data Science	Nuvative, Inc.\n3.4	Wichita, KS	30d+	https://www.glassdoor.com/partner/jobListing.h...	3.4	106.0	157.0
2	Senior Data Product Manager	ProviderTrust\n4.2	Nashville, TN	11d	https://www.glassdoor.com/partner/jobListing.h...	4.2	105.0	141.0
3	Oncology Nurse Navigator	Inizio Engage\n3.6	Portland, OR	1d	https://www.glassdoor.com/partner/jobListing.h...	3.6	90.0	113.0
4	Head of Artificial Intelligence – Americas Region	Covestro\n3.6	Pittsburgh, PA	30d+	https://www.glassdoor.com/partner/jobListing.h...	3.6	89.0	148.0

[4]:
```
#checking if any non listed company rating docs have it in the compnay name
no_rating_df = df[df['Company Rating'].isnull()]
no_rating_df
```

[4]:

	Job Title	Company Name	Location	Date	Job Link	Company Rating	Min Salary	Max Salary
5	IT Manager	Western Welding Academy	Gillette, WY	26d	https://www.glassdoor.com/partner/jobListing.h...	NaN	NaN	NaN
6	Data Center Services Technician - IT, Crypto M...	Growler Mining	Tuscaloosa, AL	30d+	https://www.glassdoor.com/partner/jobListing.h...	NaN	NaN	NaN
7	Data Center Construction Coordinator	Applied Digital	Ellendale, ND	23d	https://www.glassdoor.com/partner/jobListing.h...	NaN	58.0	80.0
9	Principal Data Scientist	HelloGov	Tampa, FL	13d	https://www.glassdoor.com/partner/jobListing.h...	NaN	166.0	245.0
15	Sr. Digital Signal Processing Engineer	Cherish Health	Boston, MA	18d	https://www.glassdoor.com/partner/jobListing.h...	NaN	153.0	200.0
17	Senior Predictive Modeler	Sitewise Analytics	Dallas, TX	30d+	https://www.glassdoor.com/partner/jobListing.h...	NaN	NaN	NaN

[5]:
```
#removing the ratings from the company names
df['Company Name'] = df['Company Name'].str.split('\n').str[0]
df
```

[5]:

	Job Title	Company Name	Location	Date	Job Link	Company Rating	Min Salary	Max Salary
0	Associate Stop Loss Underwriter	The Insurance Center	Onalaska, WI	30d+	https://www.glassdoor.com/partner/jobListing.h...	2.7	57.0	84.0
1	Manager of Data Science	Nuvative, Inc.	Wichita, KS	30d+	https://www.glassdoor.com/partner/jobListing.h...	3.4	106.0	157.0
2	Senior Data Product Manager	ProviderTrust	Nashville, TN	11d	https://www.glassdoor.com/partner/jobListing.h...	4.2	105.0	141.0
3	Oncology Nurse Navigator	Inizio Engage	Portland, OR	1d	https://www.glassdoor.com/partner/jobListing.h...	3.6	90.0	113.0
4	Head of Artificial Intelligence – Americas Region	Covestro	Pittsburgh, PA	30d+	https://www.glassdoor.com/partner/jobListing.h...	3.6	89.0	148.0
...	...	...	...	...	...	...	...	...
480	Cloud Administrator	GM Financial	Arlington, TX	25d	https://www.glassdoor.com/partner/jobListing.h...	4.0	NaN	NaN
481	Robotics Engineer (AI)	Alpha Net Consulting	United States	4d	https://www.glassdoor.com/partner/jobListing.h...	NaN	NaN	NaN
482	Tchr of English- Newark School of Data Science...	Newark Board of Education	Newark, NJ	30d+	https://www.glassdoor.com/partner/jobListing.h...	3.3	62.0	107.0
483	Statistician	Sciome LLC	Research Triangle Park, NC	30d+	https://www.glassdoor.com/partner/jobListing.h...	NaN	NaN	NaN
484	Quantitative Analytics Manager - Data Modeling...	Freddie Mac	McLean, VA	5d	https://www.glassdoor.com/partner/jobListing.h...	3.6	140.0	210.0

485 rows × 8 columns

[6]:
```
# Dropping the Job Link column and creating a new DataFrame
df = df.drop('Job Link', axis=1)

# Display the cleaned DataFrame
df.head()
```

[6]:

	Job Title	Company Name	Location	Date	Company Rating	Min Salary	Max Salary
0	Associate Stop Loss Underwriter	The Insurance Center	Onalaska, WI	30d+	2.7	57.0	84.0
1	Manager of Data Science	Nuvative, Inc.	Wichita, KS	30d+	3.4	106.0	157.0
2	Senior Data Product Manager	ProviderTrust	Nashville, TN	11d	4.2	105.0	141.0
3	Oncology Nurse Navigator	Inizio Engage	Portland, OR	1d	3.6	90.0	113.0
4	Head of Artificial Intelligence – Americas Region	Covestro	Pittsburgh, PA	30d+	3.6	89.0	148.0

[7]:
```
# Remove 'd' and 'd+' from the days_listed column
df['Date'] = df['Date'].str.replace('d\+?', '', regex=True)
df['Date'] = df['Date'].str.replace('24h', '1', regex=False)
# Display the updated DataFrame
df.head()
```

CLEANING AND CREATING COLUMNS

[8]:
```
# Renaming the column
df.rename(columns={'Date': 'Days Listed'}, inplace=True)

# Display the updated DataFrame
df.head()
```

[8]:

	Job Title	Company Name	Location	Days Listed	Company Rating	Min Salary	Max Salary
0	Associate Stop Loss Underwriter	The Insurance Center	Onalaska, WI	30	2.7	57.0	84.0
1	Manager of Data Science	Nuvative, Inc.	Wichita, KS	30	3.4	106.0	157.0
2	Senior Data Product Manager	ProviderTrust	Nashville, TN	11	4.2	105.0	141.0
3	Oncology Nurse Navigator	Inizio Engage	Portland, OR	1	3.6	90.0	113.0
4	Head of Artificial Intelligence – Americas Region	Covestro	Pittsburgh, PA	30	3.6	89.0	148.0

[9]:
```
df = df[df['Max Salary'].notnull() & df['Min Salary'].notnull()]
df.head()
```

[9]:

	Job Title	Company Name	Location	Days Listed	Company Rating	Min Salary	Max Salary
0	Associate Stop Loss Underwriter	The Insurance Center	Onalaska, WI	30	2.7	57.0	84.0
1	Manager of Data Science	Nuvative, Inc.	Wichita, KS	30	3.4		
2	Senior Data Product Manager	ProviderTrust	Nashville, TN	11	4.2		
3	Oncology Nurse Navigator	Inizio Engage	Portland, OR	1	3.6		
4	Head of Artificial Intelligence – Americas Region	Covestro	Pittsburgh, PA	30	3.6		

10]:
```
# adding a average column by averaging the min and max
df['Average Salary'] = df[['Min Salary', 'Max Salary']].mean(axis=1)
df.head()
```

10]:

	Job Title	Company Name	Location	Days Listed	Company Rating	Min
0	Associate Stop Loss Underwriter	The Insurance Center	Onalaska, WI	30	2.7	
1	Manager of Data Science	Nuvative, Inc.	Wichita, KS	30	3.4	
2	Senior Data Product Manager	ProviderTrust	Nashville, TN	11	4.2	
3	Oncology Nurse Navigator	Inizio Engage	Portland, OR	1	3.6	
4	Head of Artificial Intelligence – Americas Region	Covestro	Pittsburgh, PA	30	3.6	

11]:
```
df=df[df['Company Rating'].notnull()]
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 348 entries, 0 to 484
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Job Title       348 non-null    object
 1   Company Name    348 non-null    object
 2   Location        348 non-null    object
```

[12]:
```
# Create a 'remote' column by checking if 'Remote' is in either 'Location' or 'Job Title'
df['Remote'] = df['Location'].str.contains('Remote', case=False, regex=True) | df['Job Title'].str.contains('Remote', case=False, regex=True)
df
```

[12]:

	Job Title	Company Name	Location	Days Listed	Company Rating	Min Salary	Max Salary	Average Salary	Remote
0	Associate Stop Loss Underwriter	The Insurance Center	Onalaska, WI	30	2.7	57.0	84.0	70.5	False
1	Manager of Data Science	Nuvative, Inc.	Wichita, KS	30	3.4	106.0	157.0	131.5	False
2	Senior Data Product Manager	ProviderTrust	Nashville, TN	11	4.2	105.0	141.0	123.0	False
3	Oncology Nurse Navigator	Inizio Engage	Portland, OR	1	3.6	90.0	113.0	101.5	False
4	Head of Artificial Intelligence – Americas Region	Covestro	Pittsburgh, PA	30	3.6	89.0	148.0	118.5	False
...	...	...	...	...	...	...	...	...	...
474	Principal Data Analyst	Citizens	Irving, TX	1	3.6	110.0	160.0	135.0	False
476	Manager, Data Science Platform Engineering	Dave Inc.	Remote	30	4.0	169.0	271.0	220.0	True
479	Analytics Solution Owner	Network Coverage	Remote	30	4.4	75.0	125.0	100.0	True
482	Tchr of English- Newark School of Data Science...	Newark Board of Education	Newark, NJ	30	3.3	62.0	107.0	84.5	False
484	Quantitative Analytics Manager - Data Modeling...	Freddie Mac	McLean, VA	5	3.6	140.0	210.0	175.0	False

348 rows × 9 columns

CREATING 'JOB CATEGORY' AND 'COMPANY RATING'

[17]:
```
# Create a function to categorize jobs
def categorize_job(title):
    title = title.lower()
    if 'analyst' in title:
        return 'Analyst'
    elif 'analytics' in title:
        return 'Analyst'
    elif 'engineer' in title:
        return 'Engineer'
    elif 'scientist' in title:
        return 'Scientist'
    elif 'science' in title:
        return 'Scientist'
    elif 'manager' in title:
        return 'Manager'
    else:
        return 'Other'

# Apply the function to create a new column
clean_df['Job Category'] = clean_df['Job Title'].apply(categorize_job)
clean_df
```

[21]:
```
def categorize_rating(rating):
    if pd.isna(rating):
        return None
    elif 0 <= rating < 1.1:
        return '0-1'
    elif 1.1 <= rating < 2.1:
        return '1-2'
    elif 2.1 <= rating < 3.5:
        return '2-3'
    elif 3.5 <= rating <= 5:
        return '4-5'
    else:
        return None
# Apply the function to create a new column 'Rating Range'
clean_df.loc[:, 'Rating Range'] = clean_df['Company Rating'].apply(categorize_rating)
```

[22]:

	Job Title	Company Name	Days Listed	Company Rating	Min Salary	Max Salary	Average Salary	Remote	City	State	Job Category	Rating Range
0	Associate Stop Loss Underwriter	The Insurance Center	30	2.7	57.0	84.0	70.5	False	Onalaska	WI	Other	2-3
1	Manager of Data Science	Nuvative, Inc.	30	3.4	106.0	157.0	131.5	False	Wichita	KS	Scientist	2-3
2	Senior Data Product Manager	ProviderTrust	11	4.2	105.0	141.0	123.0	False	Nashville	TN	Manager	4-5
3	Oncology Nurse Navigator	Inizio Engage	1	3.6	90.0	113.0	101.5	False	Portland	OR	Other	4-5
4	Head of Artificial Intelligence – Americas Region	Covestro	30	3.6	89.0	148.0	118.5	False	Pittsburgh	PA	Other	4-5
...	...	...	...	...	...	...	...	...	...	...	...	...
474	Principal Data Analyst	Citizens	1	3.6	110.0	160.0	135.0	False	Irving	TX	Analyst	4-5
476	Manager, Data Science Platform Engineering	Dave Inc.	30	4.0	169.0	271.0	220.0	True	Remote	Remote	Engineer	4-5
479	Analytics Solution Owner	Network Coverage	30	4.4	75.0	125.0	100.0	True	Remote	Remote	Analyst	4-5
482	Tchr of English- Newark School of Data Science...	Newark Board of Education	30	3.3	62.0	107.0	84.5	False	Newark	NJ	Scientist	2-3
484	Quantitative Analytics Manager - Data Modeling...	Freddie Mac	5	3.6	140.0	210.0	175.0	False	McLean	VA	Analyst	4-5

339 rows × 12 columns

[17]:

	Job Title	Company Name	Days Listed	Company Rating	Min Salary	Max Salary	Average Salary	Remote	City	State	Job Category
0	Associate Stop Loss Underwriter	The Insurance Center	30	2.7	57.0	84.0	70.5	False	Onalaska	WI	Other
1	Manager of Data Science	Nuvative, Inc.	30	3.4	106.0	157.0	131.5	False	Wichita	KS	Scientist
2	Senior Data Product Manager	ProviderTrust	11	4.2	105.0	141.0	123.0	False	Nashville	TN	Manager
3	Oncology Nurse Navigator	Inizio Engage	1	3.6	90.0	113.0	101.5	False	Portland	OR	Other
4	Head of Artificial Intelligence – Americas Region	Covestro	30	3.6	89.0	148.0	118.5	False	Pittsburgh	PA	Other
...	...	...	...	...	...	...	...	...	...	...	...
474	Principal Data Analyst	Citizens	1	3.6	110.0	160.0	135.0	False	Irving	TX	Analyst
476	Manager, Data Science Platform Engineering	Dave Inc.	30	4.0	169.0	271.0	220.0	True	Remote	None	Engineer
479	Analytics Solution Owner	Network Coverage	30	4.4	75.0	125.0	100.0	True	Remote	None	Analyst
482	Tchr of English- Newark School of Data Science...	Newark Board of Education	30	3.3	62.0	107.0	84.5	False	Newark	NJ	Scientist
484	Quantitative Analytics Manager - Data Modeling...	Freddie Mac	5	3.6	140.0	210.0	175.0	False	McLean	VA	Analyst

348 rows × 11 columns

CREATING 'POSTINGS' COLUMN AND OUR DATAFRAMES

```
[23]: # create a grouby for state locations - we'll use this for the size of our scatter bubbles
location_postings = clean_df.groupby('State').size().reset_index(name='Postings')
```

```
[24]: # merge locations_postings with our df to get postings column
clean_df = clean_df.merge(location_postings, on='State', how='left')
```

```
[25]: # check it
clean_df
```

[25]:

	Job Title	Company Name	Days Listed	Company Rating	Min Salary	Max Salary	Average Salary	Remote	City	State	Job Category	Rating Range	Postings
0	Associate Stop Loss Underwriter	The Insurance Center	30	2.7	57.0	84.0	70.5	False	Onalaska	WI	Other	2-3	4
1	Manager of Data Science	Nuvative, Inc.	30	3.4	106.0	157.0	131.5	False	Wichita	KS	Scientist	2-3	1
2	Senior Data Product Manager	ProviderTrust	11	4.2	105.0	141.0	123.0	False	Nashville	TN	Manager	4-5	4
3	Oncology Nurse Navigator	Inizio Engage	1	3.6	90.0	113.0	101.5	False	Portland	OR	Other	4-5	1
4	Head of Artificial Intelligence – Americas Region	Covestro	30	3.6	89.0	148.0	118.5	False	Pittsburgh	PA	Other	4-5	11
...	...	...	...	...	...	...	...	...	...	...	...	...	...
334	Principal Data Analyst	Citizens	1	3.6	110.0	160.0	135.0	False	Irving	TX	Analyst	4-5	30
335	Manager, Data Science Platform Engineering	Dave Inc.	30	4.0	169.0	271.0	220.0	True	Remote	Remote	Engineer	4-5	21
336	Analytics Solution Owner	Network Coverage	30	4.4	75.0	125.0	100.0	True	Remote	Remote	Analyst	4-5	21
337	Tchr of English- Newark School of Data Science...	Newark Board of Education	30	3.3	62.0	107.0	84.5	False	Newark	NJ	Scientist	2-3	19
338	Quantitative Analytics Manager - Data Modeling...	Freddie Mac	5	3.6	140.0	210.0	175.0	False	McLean	VA	Analyst	4-5	28

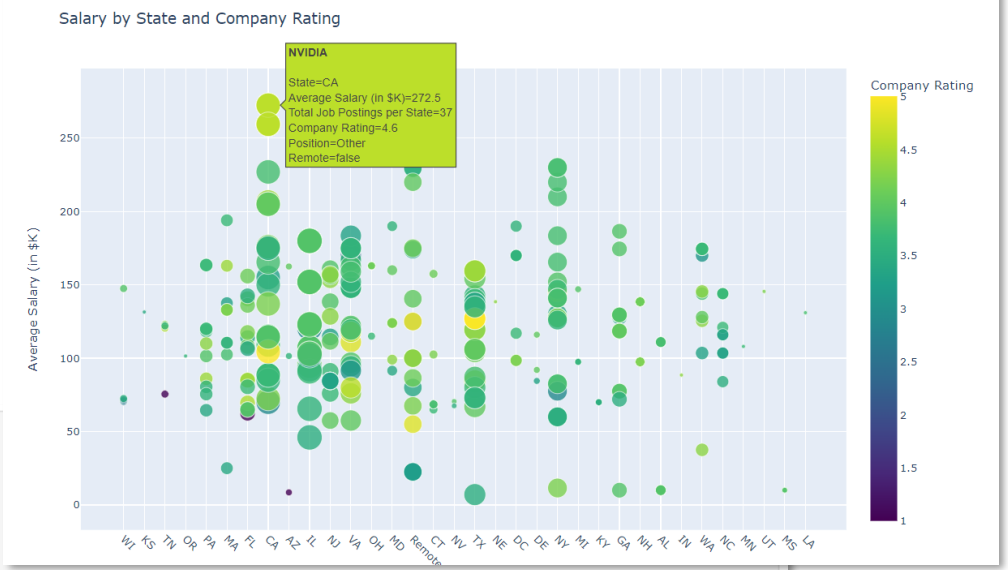
339 rows × 13 columns

```
[28]: # create a remote df
remote_df = clean_df[clean_df['Remote']== True]

# create a non remote df
non_remote_df = clean_df[clean_df['Remote']== False]

# create df's for each position
sci_df = clean_df[clean_df['Job Category'] == 'Scientist']
eng_df = clean_df[clean_df['Job Category'] == 'Engineer']
mgt_df = clean_df[clean_df['Job Category'] == 'Manager']
ana_df = clean_df[clean_df['Job Category'] == 'Analyst']
other_df = clean_df[clean_df['Job Category'] == 'Other']
```

```
# plotly express scatter
fig = px.scatter(
    clean_df,
    x='State', y='Average Salary',
    size='Postings', color='Company Rating',
    hover_name='Company Name',
    hover_data=['Company Rating', 'Job Category', 'Remote'],
    title='Salary by State and Company Rating',
    labels={'State': 'State', 'Job Category': 'Position', 'Postings':'Total Job Postings per State','Average Salary': 'Average Salary (in $K)','Remote':'
    color_continuous_scale=px.colors.sequential.Viridis,
    height=700
)
# add interaction popup with a white border
fig.update_traces(marker=dict(opacity=.8,
                              line=dict(width=.7,
                                        color='white')))
fig.update_xaxes(tickangle=45)
fig.show()
```



CREATING TABLES IN POSTGRESQL

```
# run this in bash to install psycopg2: pip install sqlalchemy psycopg2
```

```
# job_data: this is the full data table - use psycopg2 to create tables in PostgreSQL
from sqlalchemy import create_engine

# Create an engine to connect to PostgreSQL
engine = create_engine('postgresql://postgres:postgres@localhost:5432/job_data')

data = clean_df

# Writing the data to a new table in the PostgreSQL database
data.to_sql('job_data', engine, if_exists='replace', index=False)

print("Data written to PostgreSQL successfully!")
```

```
Data written to PostgreSQL successfully!
```

```
# non_remote_job_data- use psycopg2 to create tables in PostgreSQL

from sqlalchemy import create_engine

# Create an engine to connect to PostgreSQL
engine = create_engine('postgresql://postgres:postgres@localhost:5432/job_data')

data = non_remote_df

# Writing the data to a new table in the PostgreSQL database
data.to_sql('non_remote_job_data', engine, if_exists='replace', index=False)

print("Data written to PostgreSQL successfully!")
```

```
Data written to PostgreSQL successfully!
```

```
# remote_job_data - use psycopg2 to create tables in PostgreSQL
from sqlalchemy import create_engine

# Create an engine to connect to PostgreSQL
engine = create_engine('postgresql://postgres:postgres@localhost:5432/job_data')

data = remote_df

# Writing the data to a new table in the PostgreSQL database
data.to_sql('remote_job_data', engine, if_exists='replace', index=False)

print("Data written to PostgreSQL successfully!")
```

```
Data written to PostgreSQL successfully!
```

```
# scientist_job_data - use psycopg2 to create tables in PostgreSQL
from sqlalchemy import create_engine

# Create an engine to connect to PostgreSQL
engine = create_engine('postgresql://postgres:postgres@localhost:5432/job_data')

data = sci_df
```

TITLE

-
-

CONCLUSION

-
-
-
-