In [1]:
```
import pandas as pd
from  sqlalchemy import create_engine
import numpy as np
import matplotlib.pyplot  as plt
```

In [2]:
```
# csv_file = r"E:\My Docs\DATA learn\Internship Projects\fascinating ones\Data Anal
# df= pd.read_csv(csv_file)
```

In [3]:
```
conn  =create_engine("mysql+mysqlconnector://root:Passakr3@localhost/DA_jobs")
```

In [4]:
```
# df.to_sql('jobs_table', con=conn, if_exists = 'replace', index=False)
```

In [5]:
```
jobs = pd.read_sql('select * from jobs_table', conn)
```

In [6]:
```
jobs.shape
```

Out[6]:
```
(2252, 16)
```

In [7]:
```
jobs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2252 entries, 0 to 2251
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   S.no.              2252 non-null   int64  
 1   Job Title          2252 non-null   object 
 2   Salary Estimate    2252 non-null   object 
 3   Job Description    2252 non-null   object 
 4   Rating             2252 non-null   float64
 5   Company Name       2252 non-null   object 
 6   Location           2252 non-null   object 
 7   Headquarters       2252 non-null   object 
 8   Size               2252 non-null   object 
 9   Founded            2252 non-null   int64  
 10  Type of ownership  2252 non-null   object 
 11  Industry           2252 non-null   object 
 12  Sector             2252 non-null   object 
 13  Revenue            2252 non-null   object 
 14  Competitors        2252 non-null   object 
 15  Easy Apply         2252 non-null   object 
dtypes: float64(1), int64(2), object(13)
memory usage: 281.6+ KB
```

In [8]:
```
pd.read_sql_query('select * from jobs_table limit 1',  conn)
```

Out[8]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters
0	0	Data Analyst, Center on Immigration and Justic...	$37K–66K (Glassdoor est.)	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	New York, NY	New York, NY

In [9]:
```
pd.read_sql_query('select * from  jobs_table  where `Company Name` is null', conn)
```

Out[9]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Fou

In [10]:
```
jobs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2252 entries, 0 to 2251
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   S.no.              2252 non-null   int64  
 1   Job Title          2252 non-null   object 
 2   Salary Estimate    2252 non-null   object 
 3   Job Description    2252 non-null   object 
 4   Rating             2252 non-null   float64
 5   Company Name       2252 non-null   object 
 6   Location           2252 non-null   object 
 7   Headquarters       2252 non-null   object 
 8   Size               2252 non-null   object 
 9   Founded            2252 non-null   int64  
 10  Type of ownership  2252 non-null   object 
 11  Industry           2252 non-null   object 
 12  Sector             2252 non-null   object 
 13  Revenue            2252 non-null   object 
 14  Competitors        2252 non-null   object 
 15  Easy Apply         2252 non-null   object 
dtypes: float64(1), int64(2), object(13)
memory usage: 281.6+ KB
```

In [11]:
```
jobs[jobs.duplicated(keep=False)]
```

Out[11]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Fou

Checking for the duplicates

In [12]:
```
jobs[jobs.duplicated(subset=['Company Name', 'Job Description','Location'], keep=Fa
```

Out[12]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Fou

In [13]:
```
jobs.describe()
```

Out[13]:

	S.no.	Rating	Founded
count	2252.000000	2252.000000	2252.000000
mean	1125.674067	3.162478	1399.144316
std	650.489856	1.663286	901.646960
min	0.000000	-1.000000	-1.000000
25%	562.750000	3.100000	-1.000000
50%	1125.500000	3.600000	1979.000000
75%	1688.250000	4.000000	2002.000000
max	2252.000000	5.000000	2019.000000

In [14]:
```
jobs.head()
```

Out[14]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location
0	0	Data Analyst, Center on Immigration and Justic...	$37K–66K (Glassdoor est.)	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	New York, NY
1	1	Quality Data Analyst	$37K–66K (Glassdoor est.)	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New York, NY
2	2	Senior Data Analyst, Insights & Analytics Team...	$37K–66K (Glassdoor est.)	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	New York, NY
3	3	Data Analyst	$37K–66K (Glassdoor est.)	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New York, NY
4	4	Reporting Data Analyst	$37K–66K (Glassdoor est.)	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel	New York, NY

Jobs Cleaned

it is cleaned dataset minus some columns

In [15]:
```
jobs_cleaned = jobs.drop(['Competitors','Easy Apply', 'Founded'], axis=1).copy()
```

In [16]:
```
jobs_cleaned
```

Out[16]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Lo
0	0	Data Analyst, Center on Immigration and Justic...	$37K–66K (Glassdoor est.)	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	Ne
1	1	Quality Data Analyst	$37K–66K (Glassdoor est.)	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	Ne
2	2	Senior Data Analyst, Insights & Analytics Team...	$37K–66K (Glassdoor est.)	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	Ne
3	3	Data Analyst	$37K–66K (Glassdoor est.)	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	Ne
4	4	Reporting Data Analyst	$37K–66K (Glassdoor est.)	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel	Ne
...	...	...	...	...	...	...	
2247	2248	RQS - IHHA - 201900004460 -1q Data Security An...	$78K–104K (Glassdoor est.)	Maintains systems to protect data from unautho...	2.5	Avacend, Inc.	Den
2248	2249	Senior Data Analyst (Corporate Audit)	$78K–104K (Glassdoor est.)	Position:\nSenior Data Analyst (Corporate Audi...	2.9	Arrow Electronics	Cen
2249	2250	Technical Business Analyst (SQL, Data analytic...	$78K–104K (Glassdoor est.)	Title: Technical Business Analyst (SQL, Data a...	-1.0	Spiceorb	Den
2250	2251	Data Analyst 3, Customer Experience	$78K–104K (Glassdoor est.)	Summary\n\nResponsible for working cross-funct...	3.1	Contingent Network Services	Cen
2251	2252	Senior Quality Data Analyst	$78K–104K (Glassdoor est.)	You.\n\nYou bring your body, mind, heart and s...	3.4	SCL Health	Broc

2252 rows × 13 columns

Rows having nulls

In [17]:
```
jobs_cleaned.replace([-1,'-1'],np.nan, inplace=True)
```

In [18]:
```
jobs_cleaned.isna().sum()
```

Out[18]:
```
S.no.                 0
Job Title             0
Salary Estimate       1
Job Description       0
Rating              271
Company Name          0
Location              0
Headquarters        171
Size                162
Type of ownership   162
Industry            352
Sector              352
Revenue             162
dtype: int64
```

In [19]:
```
# rows_having_nulls = jobs_cleaned[jobs_cleaned.eq(-1).any(axis=1)]
```

In [20]:
```
jobs_cleaned
```

Out[20]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Lo
0	0	Data Analyst, Center on Immigration and Justic...	$37K–66K (Glassdoor est.)	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	Ne
1	1	Quality Data Analyst	$37K–66K (Glassdoor est.)	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	Ne
2	2	Senior Data Analyst, Insights & Analytics Team...	$37K–66K (Glassdoor est.)	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	Ne
3	3	Data Analyst	$37K–66K (Glassdoor est.)	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	Ne
4	4	Reporting Data Analyst	$37K–66K (Glassdoor est.)	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel	Ne
...	...	...	...	...	...	...	
2247	2248	RQS - IHHA - 201900004460 -1q Data Security An...	$78K–104K (Glassdoor est.)	Maintains systems to protect data from unautho...	2.5	Avacend, Inc.	Den
2248	2249	Senior Data Analyst (Corporate Audit)	$78K–104K (Glassdoor est.)	Position:\nSenior Data Analyst (Corporate Audi...	2.9	Arrow Electronics	Cen
2249	2250	Technical Business Analyst (SQL, Data analytic...	$78K–104K (Glassdoor est.)	Title: Technical Business Analyst (SQL, Data a...	NaN	Spiceorb	Den
2250	2251	Data Analyst 3, Customer Experience	$78K–104K (Glassdoor est.)	Summary\n\nResponsible for working cross-funct...	3.1	Contingent Network Services	Cen
2251	2252	Senior Quality Data Analyst	$78K–104K (Glassdoor est.)	You.\n\nYou bring your body, mind, heart and s...	3.4	SCL Health	Broc

2252 rows × 13 columns

In [21]:
```
jobs_cleaned['Salary Estimate'] = jobs_cleaned['Salary Estimate'].str.split(' ').st
jobs_cleaned['Salary Estimate'] = jobs_cleaned['Salary Estimate'].str.strip()
jobs_cleaned['Salary Estimate'] = jobs_cleaned['Salary Estimate'].str.replace('$','
```

In [22]:
```
x = jobs_cleaned.copy()
```

In [23]:
```
jobs_cleaned[['min_salary','max_salary']] = jobs_cleaned['Salary Estimate'].str.spl
```

In [24]:
```
def clean_sal_col(z, dataset):
    dataset[z] = dataset[z].replace('', np.nan)  # there was a row where x['a'] was
    dataset[z] = dataset[z].str.replace('K','', case=False).str.strip()
    dataset[z] = dataset[z].astype(float)
    dataset[z] = dataset[z]*1000  # as we had removed K
    dataset[z].fillna(0)
    return dataset
```

In [25]:
```
clean_sal_col('max_salary', jobs_cleaned)
```

Out[25]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Loc
0	0	Data Analyst, Center on Immigration and Justic...	37K-66K	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	New
1	1	Quality Data Analyst	37K-66K	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New
2	2	Senior Data Analyst, Insights & Analytics Team...	37K-66K	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	New
3	3	Data Analyst	37K-66K	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New
4	4	Reporting Data Analyst	37K-66K	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel	New
...	...	...	...	...	...	...	
2247	2248	RQS - IHHA - 201900004460 -1q Data Security An...	78K-104K	Maintains systems to protect data from unautho...	2.5	Avacend, Inc.	Denve
2248	2249	Senior Data Analyst (Corporate Audit)	78K-104K	Position:\nSenior Data Analyst (Corporate Audi...	2.9	Arrow Electronics	Cente
2249	2250	Technical Business Analyst (SQL, Data analytic...	78K-104K	Title: Technical Business Analyst (SQL, Data a...	NaN	Spiceorb	Denve
2250	2251	Data Analyst 3, Customer Experience	78K-104K	Summary\n\nResponsible for working cross-funct...	3.1	Contingent Network Services	Cente
2251	2252	Senior Quality Data Analyst	78K-104K	You.\n\nYou bring your body, mind, heart and s...	3.4	SCL Health	Broom

2252 rows × 15 columns

In [26]:
```
clean_sal_col('min_salary',  jobs_cleaned)
```

Out[26]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Loc
0	0	Data Analyst, Center on Immigration and Justic...	37K-66K	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	New
1	1	Quality Data Analyst	37K-66K	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New
2	2	Senior Data Analyst, Insights & Analytics Team...	37K-66K	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	New
3	3	Data Analyst	37K-66K	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New
4	4	Reporting Data Analyst	37K-66K	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel	New
...	...	...	...	...	...	...	
2247	2248	RQS - IHHA - 201900004460 -1q Data Security An...	78K-104K	Maintains systems to protect data from unautho...	2.5	Avacend, Inc.	Denve
2248	2249	Senior Data Analyst (Corporate Audit)	78K-104K	Position:\nSenior Data Analyst (Corporate Audi...	2.9	Arrow Electronics	Cente
2249	2250	Technical Business Analyst (SQL, Data analytic...	78K-104K	Title: Technical Business Analyst (SQL, Data a...	NaN	Spiceorb	Denve
2250	2251	Data Analyst 3, Customer Experience	78K-104K	Summary\n\nResponsible for working cross-funct...	3.1	Contingent Network Services	Cente
2251	2252	Senior Quality Data Analyst	78K-104K	You.\n\nYou bring your body, mind, heart and s...	3.4	SCL Health	Broom

2252 rows × 15 columns

In [27]:
```
jobs_cleaned['avg_salary'] = jobs_cleaned[['min_salary','max_salary']].mean(axis=1)
```

In [28]:
```
jobs_cleaned = jobs_cleaned.dropna(subset=['min_salary','max_salary'])
```

In [29]:
```
jobs_cleaned
```

Out[29]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Loc
0	0	Data Analyst, Center on Immigration and Justic...	37K-66K	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	New
1	1	Quality Data Analyst	37K-66K	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New
2	2	Senior Data Analyst, Insights & Analytics Team...	37K-66K	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	New
3	3	Data Analyst	37K-66K	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New
4	4	Reporting Data Analyst	37K-66K	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel	New
...	...	...	...	...	...	...	
2247	2248	RQS - IHHA - 201900004460 -1q Data Security An...	78K-104K	Maintains systems to protect data from unautho...	2.5	Avacend, Inc.	Denve
2248	2249	Senior Data Analyst (Corporate Audit)	78K-104K	Position:\nSenior Data Analyst (Corporate Audi...	2.9	Arrow Electronics	Cente
2249	2250	Technical Business Analyst (SQL, Data analytic...	78K-104K	Title: Technical Business Analyst (SQL, Data a...	NaN	Spiceorb	Denve
2250	2251	Data Analyst 3, Customer Experience	78K-104K	Summary\n\nResponsible for working cross-funct...	3.1	Contingent Network Services	Cente
2251	2252	Senior Quality Data Analyst	78K-104K	You.\n\nYou bring your body, mind, heart and s...	3.4	SCL Health	Broom

2251 rows × 16 columns

In [30]:
```
x[['a','b']] = x['Salary Estimate'].str.split('-',expand=True)
```

In [31]:
```
x
```

Out[31]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Loc
0	0	Data Analyst, Center on Immigration and Justic...	37K-66K	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	New
1	1	Quality Data Analyst	37K-66K	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New
2	2	Senior Data Analyst, Insights & Analytics Team...	37K-66K	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	New
3	3	Data Analyst	37K-66K	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New
4	4	Reporting Data Analyst	37K-66K	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel	New
...	...	...	...	...	...	...	
2247	2248	RQS - IHHA - 201900004460 -1q Data Security An...	78K-104K	Maintains systems to protect data from unautho...	2.5	Avacend, Inc.	Denve
2248	2249	Senior Data Analyst (Corporate Audit)	78K-104K	Position:\nSenior Data Analyst (Corporate Audi...	2.9	Arrow Electronics	Cente
2249	2250	Technical Business Analyst (SQL, Data analytic...	78K-104K	Title: Technical Business Analyst (SQL, Data a...	NaN	Spiceorb	Denve
2250	2251	Data Analyst 3, Customer Experience	78K-104K	Summary\n\nResponsible for working cross-funct...	3.1	Contingent Network Services	Cente
2251	2252	Senior Quality Data Analyst	78K-104K	You.\n\nYou bring your body, mind, heart and s...	3.4	SCL Health	Broom

2252 rows × 15 columns

In [32]:
```
pd.to_pickle(jobs_cleaned,'jobs_cleaned')
```

In [33]:
```
jobs_cleaned.groupby('Type of ownership')
```

Out[33]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x0000026004892F80>
```

In [34]:
```
jobs_cleaned['Type of ownership'].value_counts()
```

Out[34]:
```
Type of ownership
Company - Private                    1272
Company - Public                      452
Nonprofit Organization                124
Subsidiary or Business Segment         89
Government                             37
College / University                   34
Hospital                               19
Unknown                                16
Other Organization                     13
Contract                               11
School / School District                9
Private Practice / Firm                 9
Self-employed                           2
Franchise                               2
Name: count, dtype: int64
```

In [ ]: