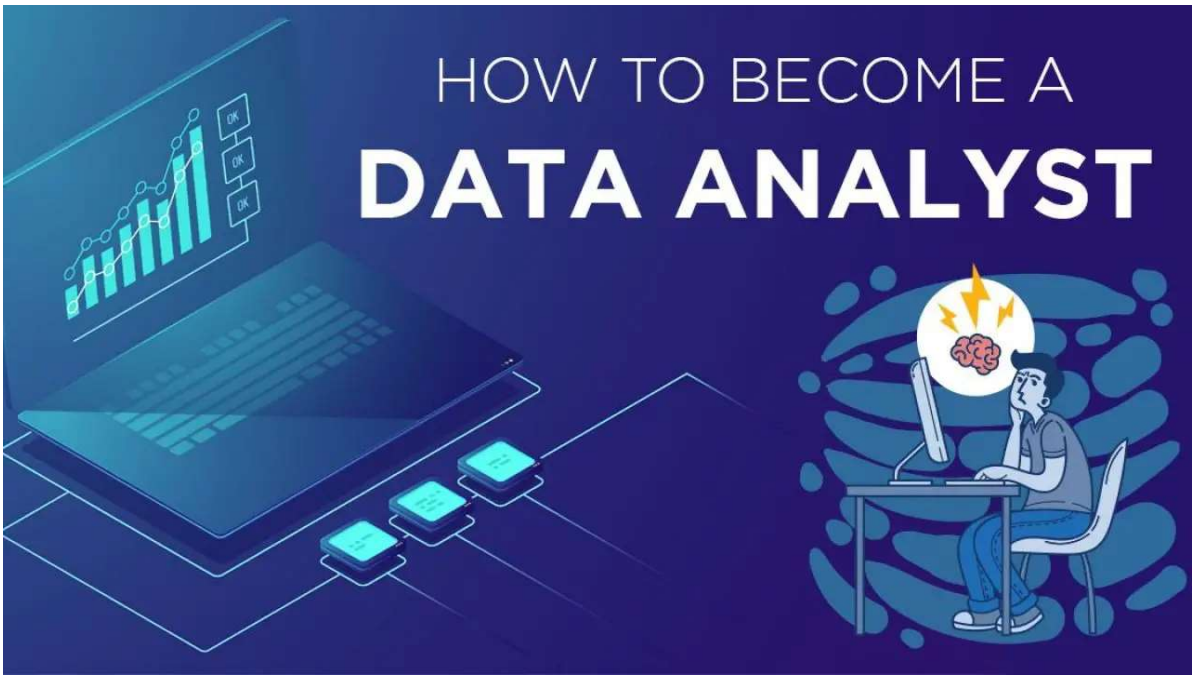
Data Analyst Jobs Prediction

NAME : Bishwajit Sen Project : DATA ANALYST JOBS PREDICTION (USA)



Introduction to the "Data Analyst Jobs" Dataset

The "Data Analyst Jobs" dataset is a comprehensive collection of job postings in the field of Data Analysis, providing valuable insights into the diverse and dynamic job market. This dataset offers a rich assortment of information, ranging from job titles and descriptions to salary details, company sizes, and industry sectors.

With the ever-increasing demand for data-driven decision-making and the rapid growth of data-driven industries, data analysts play a crucial role in extracting meaningful insights from vast datasets, identifying trends, and making recommendations. As a result, this dataset serves as a valuable resource for job seekers, career advisors, and analysts interested in exploring the current landscape of data analyst job opportunities.

You can find here a broad spectrum of job titles, from entry-level positions to senior roles, allowing users to explore the salary distributions and company profiles across various job levels. Additionally, the dataset's industry-specific information empowers researchers and professionals to gain a deeper understanding of the sectors where data analysts are most in demand.

Importing Libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [2]:
```
df = pd.read_csv(r"C:\Users\Dell\OneDrive\Desktop\PRACTICE\data analyst job\archive (6
df.head()
```

Out[2]:

	Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	H
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\n3.2	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\n3.8	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\n3.4	New York, NY	
3	3	Data Analyst	$37K–66K (Glassdoor est.)	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity\n4.1	New York, NY	
4	4	Reporting Data Analyst	$37K–66K (Glassdoor est.)	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel\n3.9	New York, NY	

In [3]:
```
df
```

Out[3]:

	Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Loc
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\n3.2	New
1	1	Quality Data Analyst	$37K–66K (Glassdoor est.)	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York\n3.8	New
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\n3.4	New
3	3	Data Analyst	$37K–66K (Glassdoor est.)	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity\n4.1	New
4	4	Reporting Data Analyst	$37K–66K (Glassdoor est.)	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel\n3.9	New
...	...	...	...	...	...	...	
2248	2248	RQS - IHHA - 201900004460 -1q Data Security An...	$78K–104K (Glassdoor est.)	Maintains systems to protect data from unautho...	2.5	Avacend, Inc.\n2.5	Denve
2249	2249	Senior Data Analyst (Corporate Audit)	$78K–104K (Glassdoor est.)	Position:\nSenior Data Analyst (Corporate Audi...	2.9	Arrow Electronics\n2.9	Cente
2250	2250	Technical Business Analyst (SQL, Data analytic...	$78K–104K (Glassdoor est.)	Title: Technical Business Analyst (SQL, Data a...	-1.0	Spiceorb	Denve
2251	2251	Data Analyst 3, Customer Experience	$78K–104K (Glassdoor est.)	Summary\n\nResponsible for working cross-funct...	3.1	Contingent Network Services\n3.1	Cente
2252	2252	Senior Quality Data Analyst	$78K–104K (Glassdoor est.)	You.\n\nYou bring your body, mind, heart and s...	3.4	SCL Health\n3.4	Broom

2253 rows × 16 columns

In [4]:
```
# Drop specified columns from the DataFrame, ignoring errors if they don't exist

df.drop(['Job Description', 'Company Name', 'Headquarters', 'Easy Apply', 'Competitors
```

In [5]:
```
# check for empty rows

df.isnull().sum()
```

Out[5]:
```
Unnamed: 0           0
Job Title            0
Salary Estimate      0
Rating               0
Location             0
Size                 0
Type of ownership    0
Industry             0
Sector               0
Revenue              0
dtype: int64
```

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2253 entries, 0 to 2252
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         2253 non-null   int64  
 1   Job Title          2253 non-null   object 
 2   Salary Estimate    2253 non-null   object 
 3   Rating             2253 non-null   float64
 4   Location           2253 non-null   object 
 5   Size               2253 non-null   object 
 6   Type of ownership  2253 non-null   object 
 7   Industry           2253 non-null   object 
 8   Sector             2253 non-null   object 
 9   Revenue            2253 non-null   object 
dtypes: float64(1), int64(1), object(8)
memory usage: 176.1+ KB
```

In [7]:
```
df.head()
```

Out[7]:

	Unnamed: 0	Job Title	Salary Estimate	Rating	Location	Size	Type of ownership	Industry	Se
0	0	Data Analyst, Center on Immigration and Justic...	$37K–66K (Glassdoor est.)	3.2	New York, NY	201 to 500 employees	Nonprofit Organization	Social Assistance	Non-F
1	1	Quality Data Analyst	$37K–66K (Glassdoor est.)	3.8	New York, NY	10000+ employees	Nonprofit Organization	Health Care Services & Hospitals	Health
2	2	Senior Data Analyst, Insights & Analytics Team...	$37K–66K (Glassdoor est.)	3.4	New York, NY	1001 to 5000 employees	Company - Private	Internet	Informa Techno
3	3	Data Analyst	$37K–66K (Glassdoor est.)	4.1	New York, NY	201 to 500 employees	Subsidiary or Business Segment	IT Services	Informa Techno
4	4	Reporting Data Analyst	$37K–66K (Glassdoor est.)	3.9	New York, NY	501 to 1000 employees	Company - Private	Sports & Recreation	Entertainr & Recrea

In [9]:
```
# replacing Job Titles to avoid duplicates

df['Job Title'] = df['Job Title'].replace(['Sr. Data Analyst', 'sr. data analyst', 'Sr
                                           'Senior Data Analyst', regex=True)
df['Job Title'] = df['Job Title'].replace(['Data Analyst I', 'data analyst i', 'Data A
                                           'Junior Data Analyst', 'Junior Data Analyst
df['Job Title'] = df['Job Title'].replace(['Data Analyst II', 'data analyst ii', 'Midd
                                           'Middle Data Analyst', regex=True)
```

In [10]:
```
df.info
```

Out[10]:
```
<bound method DataFrame.info of       Unnamed: 0
Job Title  \
0              0  Data Analyst, Center on Immigration and Justic...
1              1                                Quality Data Analyst
2              2  Senior Data Analyst, Insights & Analytics Team...
3              3                                        Data Analyst
4              4                             Reporting Data Analyst
...          ...                                                ...
2248        2248  RQS - IHHA - 201900004460 -1q Data Security An...
2249        2249              Senior Data Analyst (Corporate Audit)
2250        2250  Technical Business Analyst (SQL, Data analytic...
2251        2251                Data Analyst 3, Customer Experience
2252        2252                        Senior Quality Data Analyst

                  Salary Estimate  Rating        Location  \
0      $37K-$66K (Glassdoor est.)     3.2    New York, NY
1      $37K-$66K (Glassdoor est.)     3.8    New York, NY
2      $37K-$66K (Glassdoor est.)     3.4    New York, NY
3      $37K-$66K (Glassdoor est.)     4.1    New York, NY
4      $37K-$66K (Glassdoor est.)     3.9    New York, NY
...                           ...     ...             ...
2248  $78K-$104K (Glassdoor est.)     2.5      Denver, CO
2249  $78K-$104K (Glassdoor est.)     2.9  Centennial, CO
2250  $78K-$104K (Glassdoor est.)    -1.0      Denver, CO
2251  $78K-$104K (Glassdoor est.)     3.1  Centennial, CO
2252  $78K-$104K (Glassdoor est.)     3.4  Broomfield, CO

                        Size               Type of ownership  \
0       201 to 500 employees          Nonprofit Organization
1           10000+ employees          Nonprofit Organization
2     1001 to 5000 employees               Company - Private
3       201 to 500 employees  Subsidiary or Business Segment
4      501 to 1000 employees               Company - Private
...                      ...                             ...
2248     51 to 200 employees               Company - Private
2249        10000+ employees                Company - Public
2250                      -1                              -1
2251    201 to 500 employees               Company - Private
2252        10000+ employees          Nonprofit Organization

                                    Industry  \
0                           Social Assistance
1            Health Care Services & Hospitals
2                                    Internet
3                                IT Services
4                          Sports & Recreation
...                                        ...
2248                   Staffing & Outsourcing
2249                                Wholesale
2250                                       -1
2251  Enterprise Software & Network Solutions
2252         Health Care Services & Hospitals

                                Sector                      Revenue
0                           Non-Profit  $100 to $500 million (USD)
1                          Health Care      $2 to $5 billion (USD)
2              Information Technology    Unknown / Non-Applicable
3              Information Technology   $50 to $100 million (USD)
4     Arts, Entertainment & Recreation  $100 to $500 million (USD)
...                                ...                         ...
```

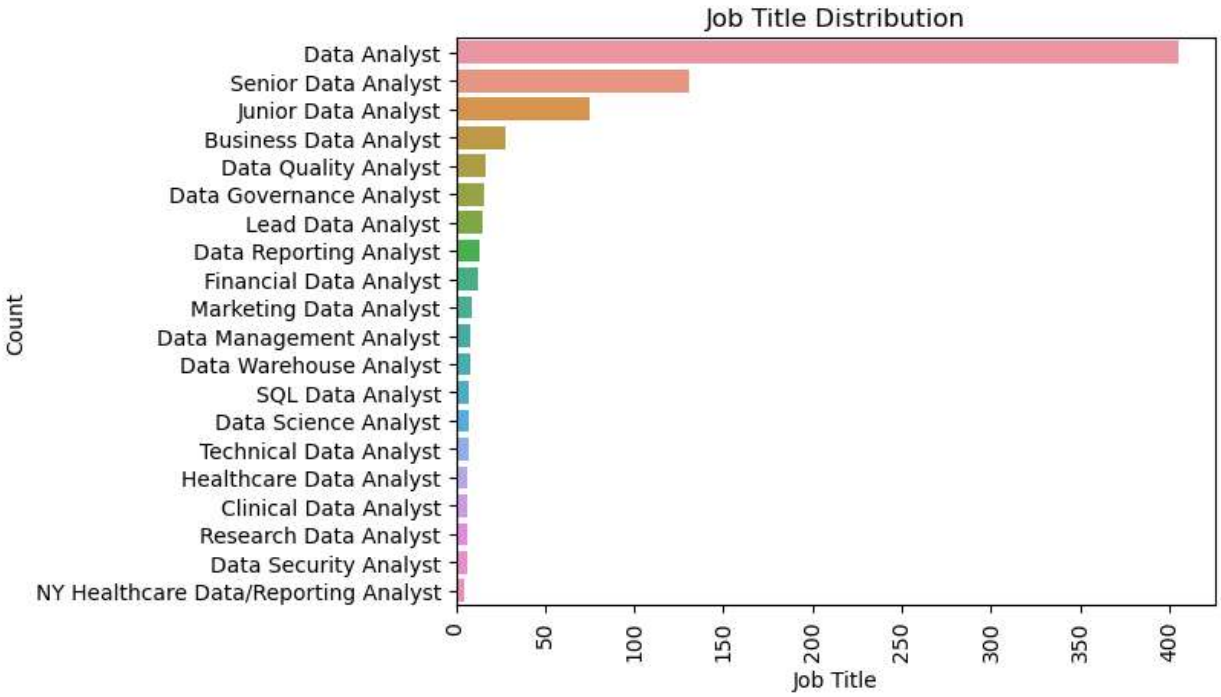
```
2248                 Business Services    Unknown / Non-Applicable
2249                 Business Services          $10+ billion (USD)
2250                                -1                          -1
2251           Information Technology    $25 to $50 million (USD)
2252                       Health Care      $2 to $5 billion (USD)

[2253 rows x 10 columns]>
```

In [11]:
```
# that's better

df['Job Title'].value_counts().head(20)
```

Out[11]:
```
Data Analyst                              405
Senior Data Analyst                       131
Junior Data Analyst                        75
Business Data Analyst                      28
Data Quality Analyst                       17
Data Governance Analyst                    16
Lead Data Analyst                          15
Data Reporting Analyst                     13
Financial Data Analyst                     12
Marketing Data Analyst                      9
Data Management Analyst                     8
Data Warehouse Analyst                      8
SQL Data Analyst                            7
Data Science Analyst                        7
Technical Data Analyst                      7
Healthcare Data Analyst                     6
Clinical Data Analyst                       6
Research Data Analyst                       6
Data Security Analyst                       6
NY Healthcare Data/Reporting Analyst        5
Name: Job Title, dtype: int64
```

Analysis

In [12]:
```
# viz

df_title = df['Job Title'].value_counts().head(20)

sns.barplot(x=df_title.values, y=df_title.index)

plt.xlabel('Job Title')
plt.ylabel('Count')
plt.title('Job Title Distribution')
plt.xticks(rotation=90)
plt.show()
```

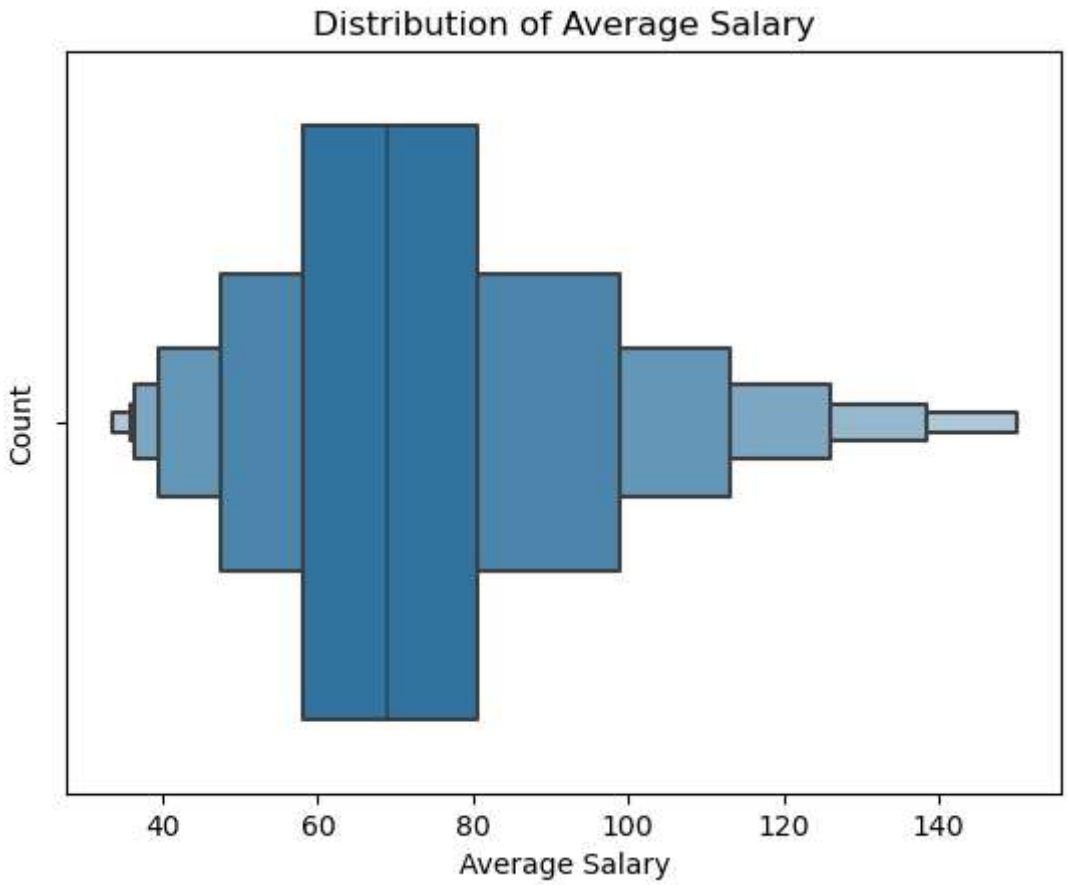


In [13]:
```
## Changing Salary column to int for better calculation

df[['MinSalary', 'MaxSalary']] = df['Salary Estimate'].str.extract(r'\$(\d+)K-\$(\d+)K

df['MinSalary'] = pd.to_numeric(df['MinSalary'])
df['MaxSalary'] = pd.to_numeric(df['MaxSalary'])
```

In [14]:
```
# chnging format to float

df['MinSalary'] = df['MinSalary'].astype(float)
df['MaxSalary'] = df['MaxSalary'].astype(float)

df['AverageSalary'] = (df['MaxSalary'] + df['MinSalary']) / 2
```

In [15]:
```
# Droping unuseful columns

df.drop(['Salary Estimate', 'MinSalary', 'MaxSalary'], axis=1, inplace=True)
```

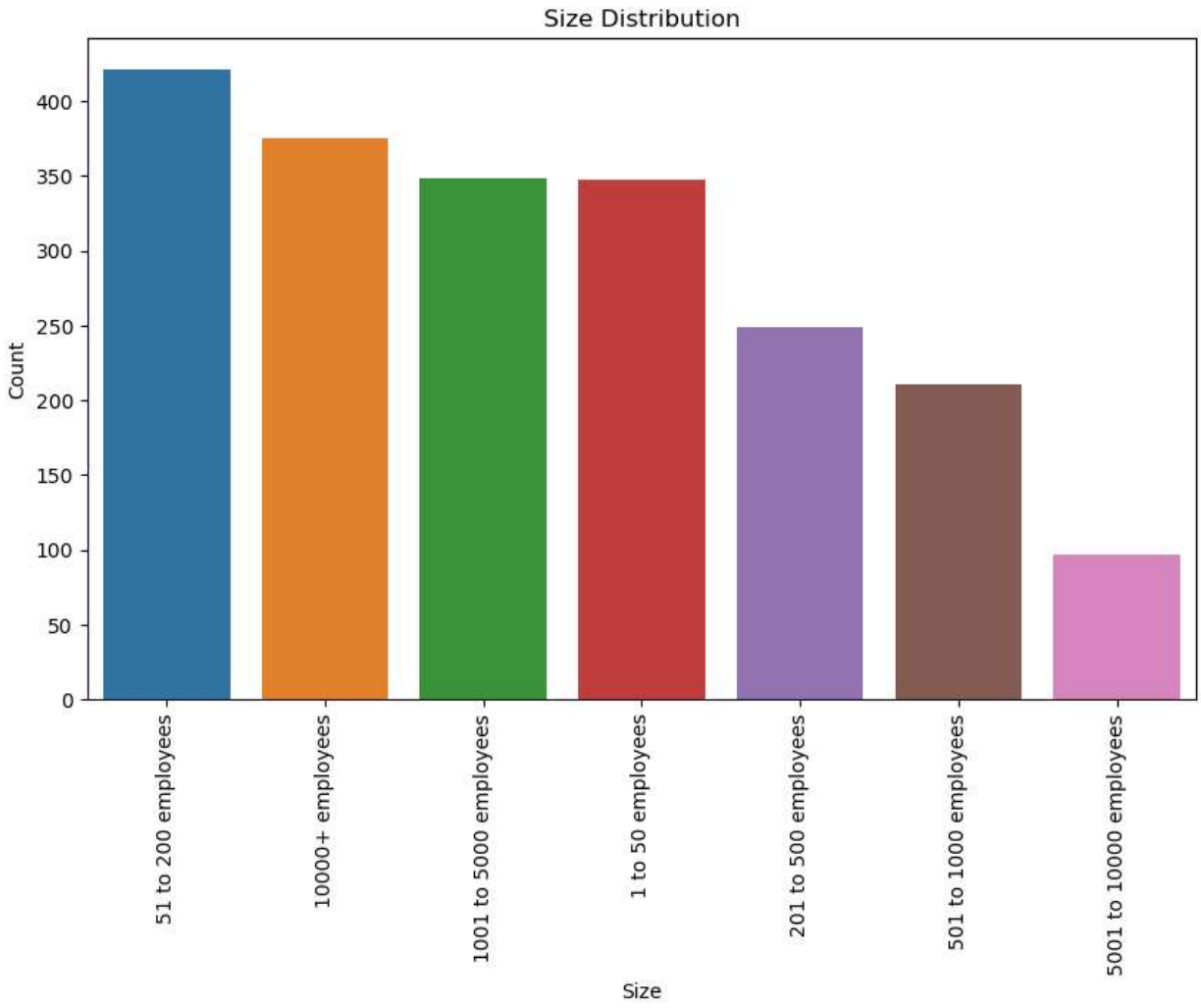
In [16]:
```
# Average Salary

sns.boxenplot(data=df, x='AverageSalary')
plt.xlabel('Average Salary')
plt.ylabel('Count')
plt.title('Distribution of Average Salary')
plt.show()
```



Key:

Average Salary on the market is about 60-80 thousands dollars per year.

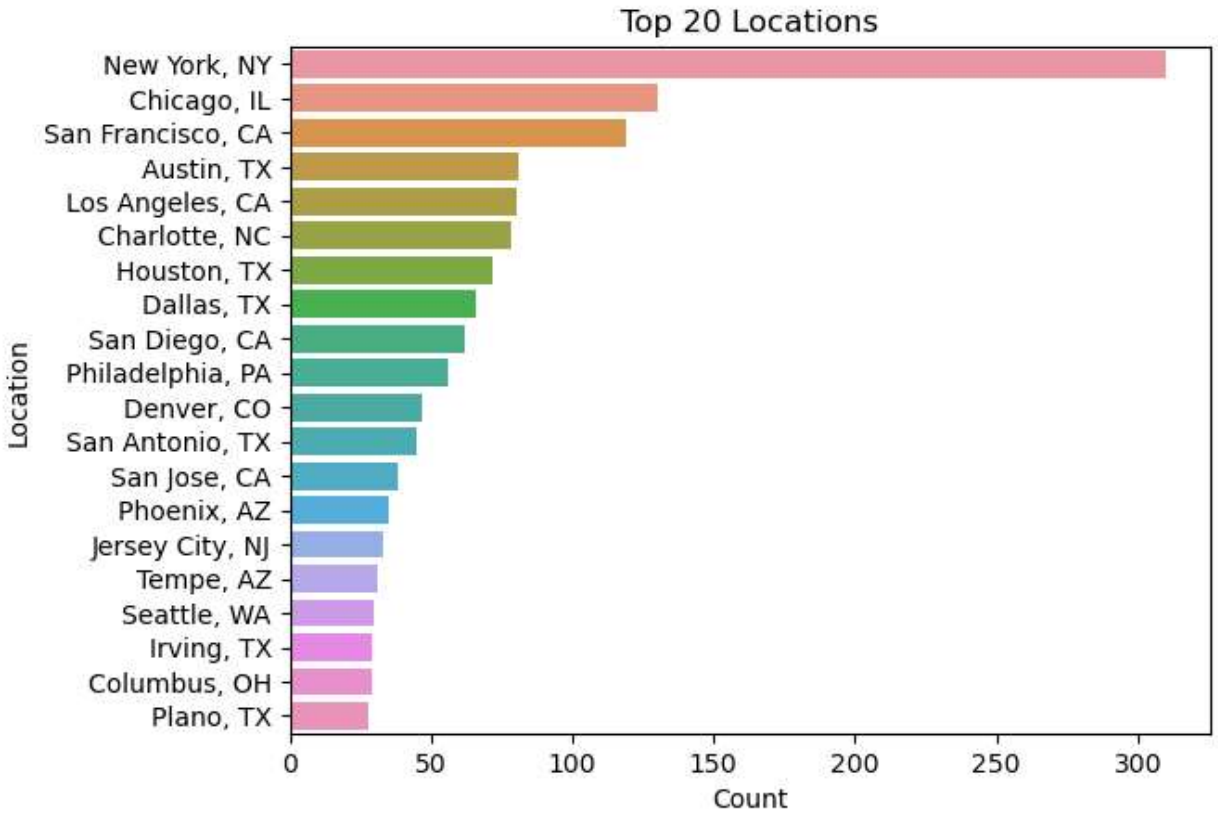
In [17]:
```
# Companies by Amount of Employees

filtered_size = df[(df['Size'] != '-1') & (df['Size'] != 'Unknown')]
df_size = filtered_size['Size'].value_counts().head(20)

plt.figure(figsize=(10, 6))
sns.barplot(x=df_size.index, y=df_size.values)
plt.xlabel('Size')
plt.ylabel('Count')
plt.title('Size Distribution')
plt.xticks(rotation=90)
plt.show()
```



Key:

Most of the interviewers work in small companies/ start-ups. Large Corporations are falling right behind.

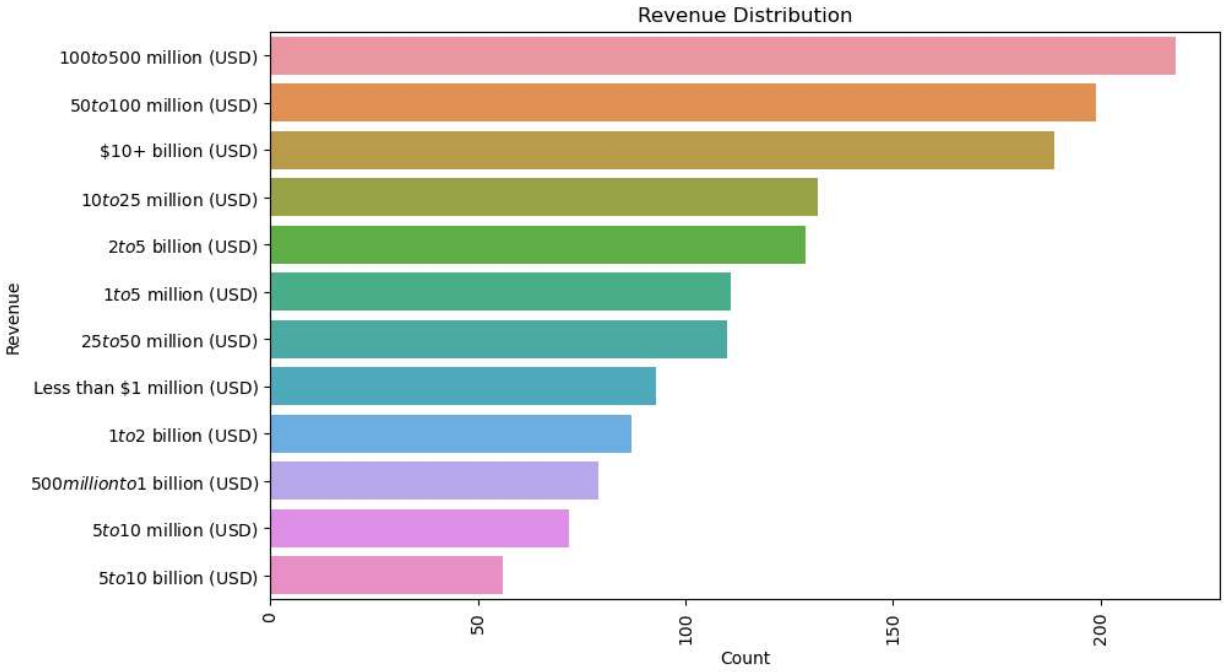
In [18]:
```
# Top work locations among interviewed

top_locations = df['Location'].value_counts().head(20)

sns.barplot(x=top_locations.values, y=top_locations.index)

plt.xlabel('Count')
plt.ylabel('Location')
plt.title('Top 20 Locations')
plt.show()
```



Key:

Employees are mostly located in New York. Also many analysts work in States of California and Texas.

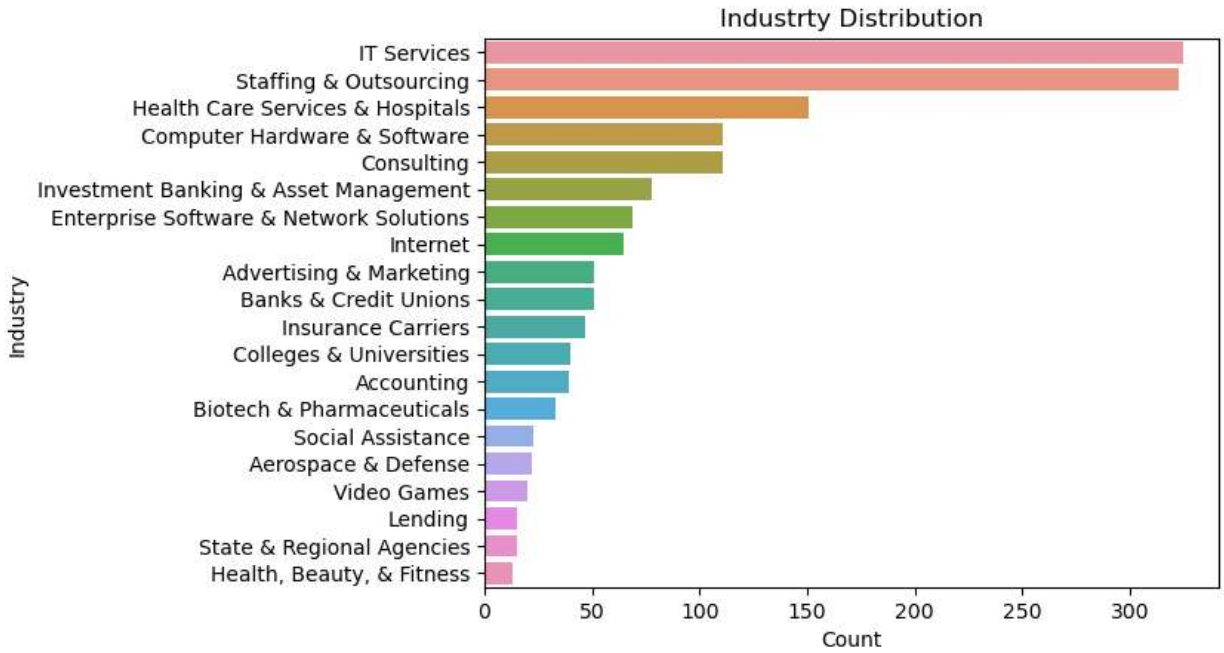
In [19]:
```
# Revenue distribution

filtered_revenue = df[(df['Revenue'] != '-1') & (df['Revenue'] != 'Unknown / Non-Appli
df_revenue = filtered_revenue['Revenue'].value_counts()

plt.figure(figsize=(10, 6))
sns.barplot(x=df_revenue.values, y=df_revenue.index)
plt.xlabel('Count')
plt.ylabel('Revenue')
plt.title('Revenue Distribution')
plt.xticks(rotation=90)
plt.show()
```



Key:

Most of the interviewed work in high revenue companies.

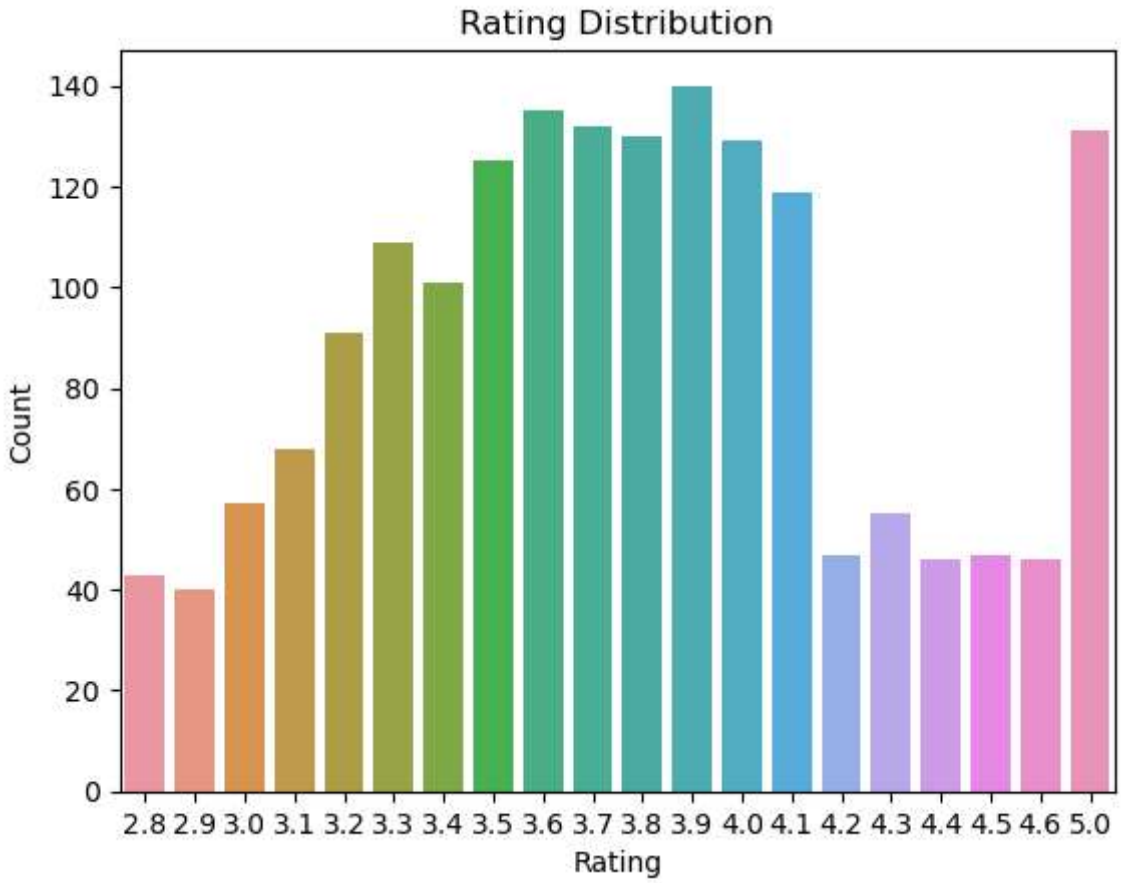
In [20]:
```
# Employees by Industry

df_industry = df[df['Industry'] != '-1']['Industry'].value_counts().head(20)

sns.barplot(x=df_industry.values, y=df_industry.index)
plt.xlabel('Count')
plt.ylabel('Industry')
plt.title('Industrty Distribution')
plt.show()
```



Key:

IT services and Staffing & Outsourcing are dominant industries.

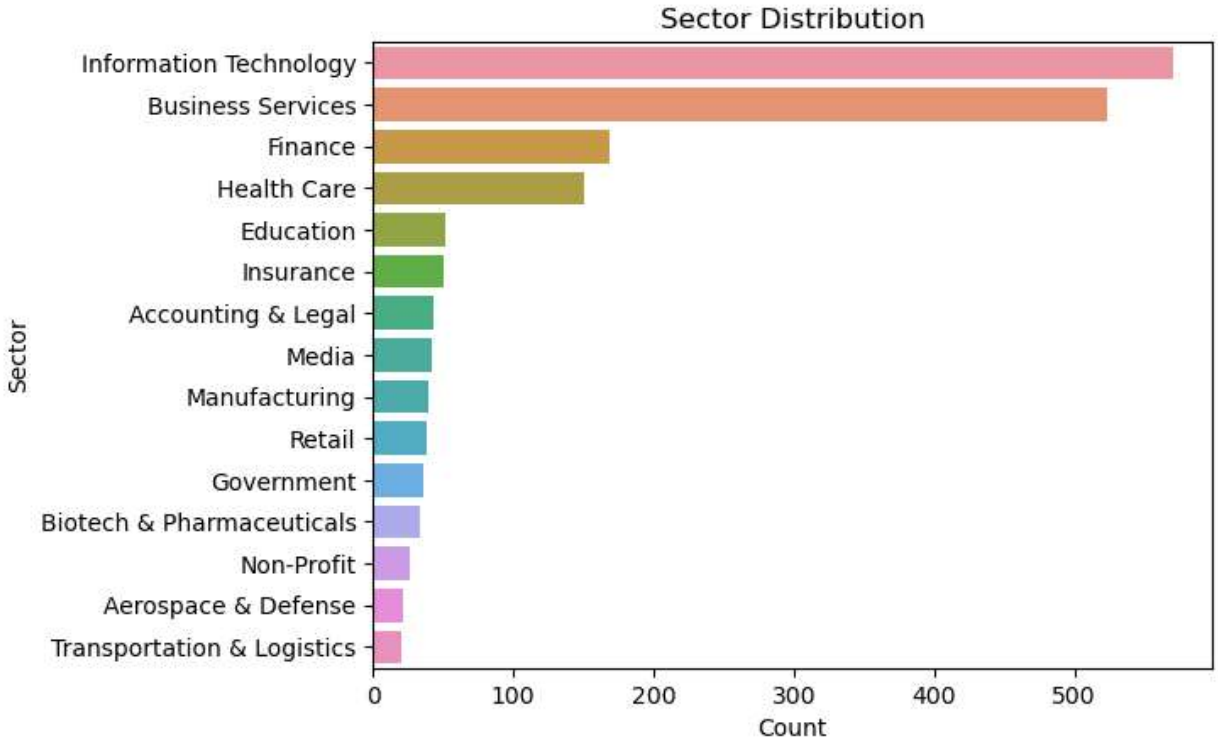
In [21]:
```
# Rating distribution

df_rating = df[df['Rating'] != -1]['Rating'].value_counts().head(20)
sns.barplot(x=df_rating.index, y=df_rating.values)
plt.xlabel('Rating')
plt.ylabel('Count')
plt.title('Rating Distribution')
plt.show()
```



Key:

Average Rating of companies is between 3.5 - 4.1 points, with another splash at 5 points.

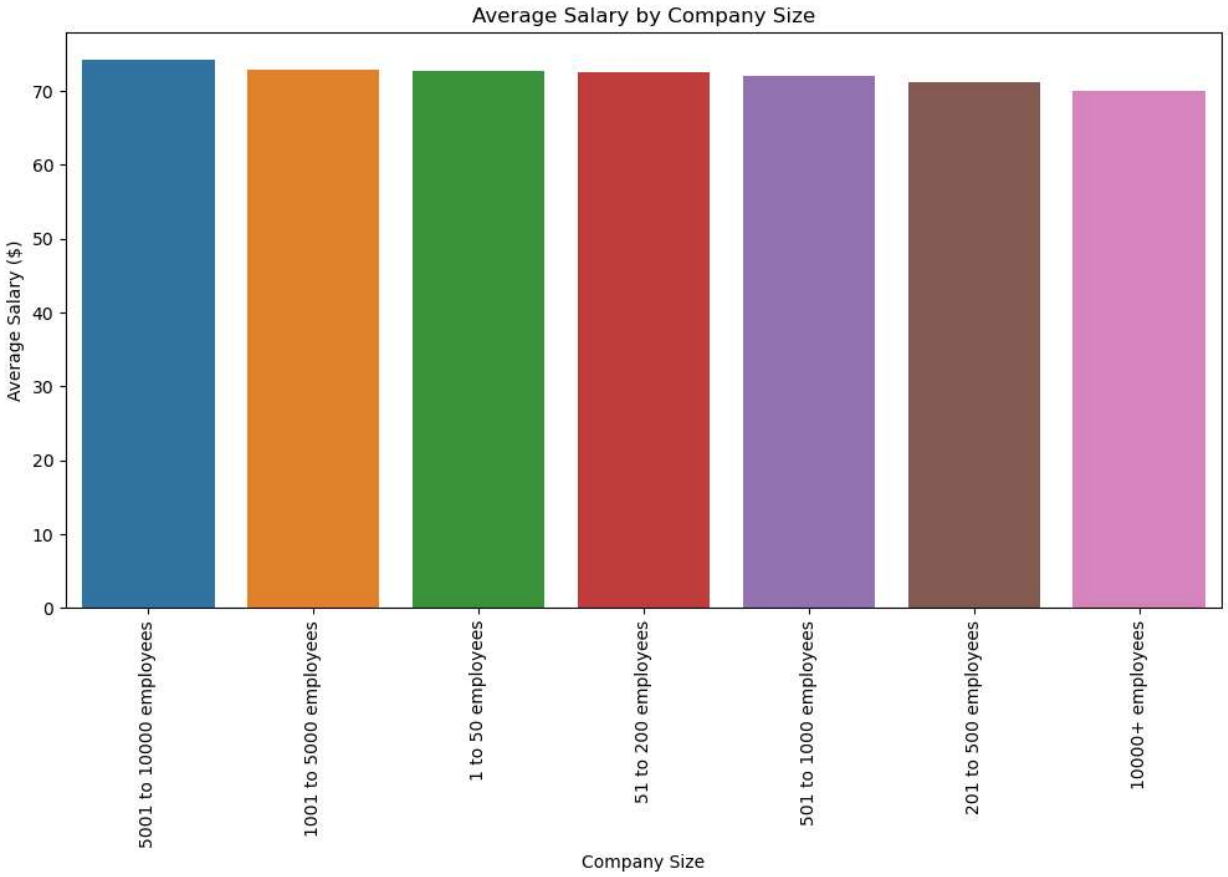
In [22]:
```
# Sector distribution

df_sector = df[df['Sector'] != '-1']['Sector'].value_counts().head(15)

sns.barplot(x=df_sector.values, y=df_sector.index)
plt.xlabel('Count')
plt.ylabel('Sector')
plt.title('Sector Distribution')
plt.show()
```



Key:

IT and Business Sevices are dominant.

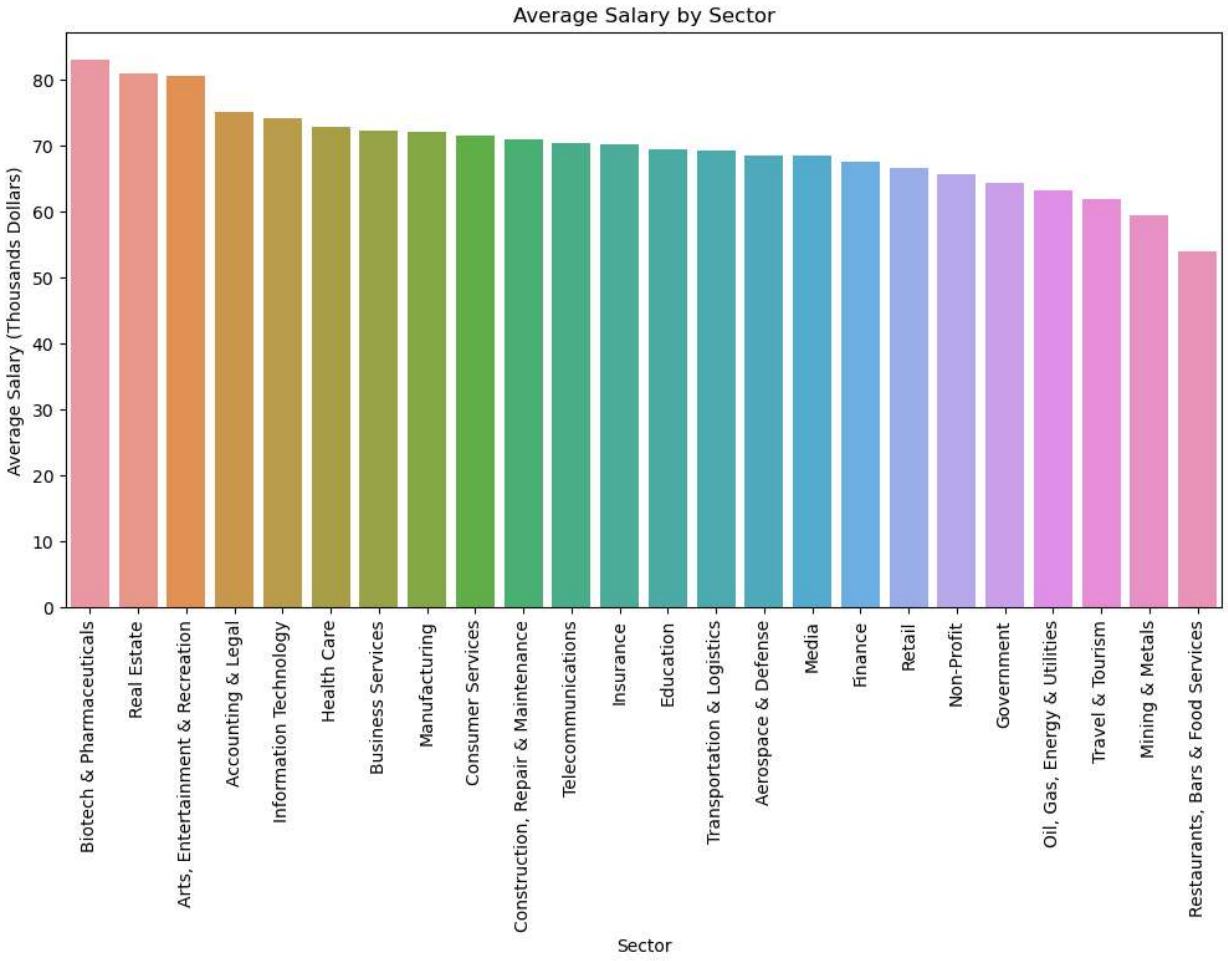
In [23]:
```
# Salary by Company Size

df_filtered = df[(df['Size'] != '-1') & (df['Size'] != 'Unknown')]
df_sizeXsalary = df_filtered.groupby('Size')['AverageSalary'].mean().reset_index()

# Sort the DataFrame by 'AverageSalary' in descending order
df_sizeXsalary = df_sizeXsalary.sort_values(by='AverageSalary', ascending=False)

# Plot the bar chart
plt.figure(figsize=(12, 6))
sns.barplot(x='Size', y='AverageSalary', data=df_sizeXsalary)
plt.xlabel('Company Size')
plt.ylabel('Average Salary ($)')
plt.title('Average Salary by Company Size')
plt.xticks(rotation=90)

plt.show()
```



Key:

Size of a company doesn't affect paycheck.

In [24]:
```
# Salary by Sector

average_salary_by_sector = df[df['Sector'] != '-1'].groupby('Sector')['AverageSalary']

average_salary_by_sector = average_salary_by_sector.sort_values(by='AverageSalary', as

plt.figure(figsize=(12, 6))
sns.barplot(x='Sector', y='AverageSalary', data=average_salary_by_sector)
plt.xticks(rotation=90)
plt.xlabel('Sector')
plt.ylabel('Average Salary (Thousands Dollars)')
plt.title('Average Salary by Sector')
plt.show()
```

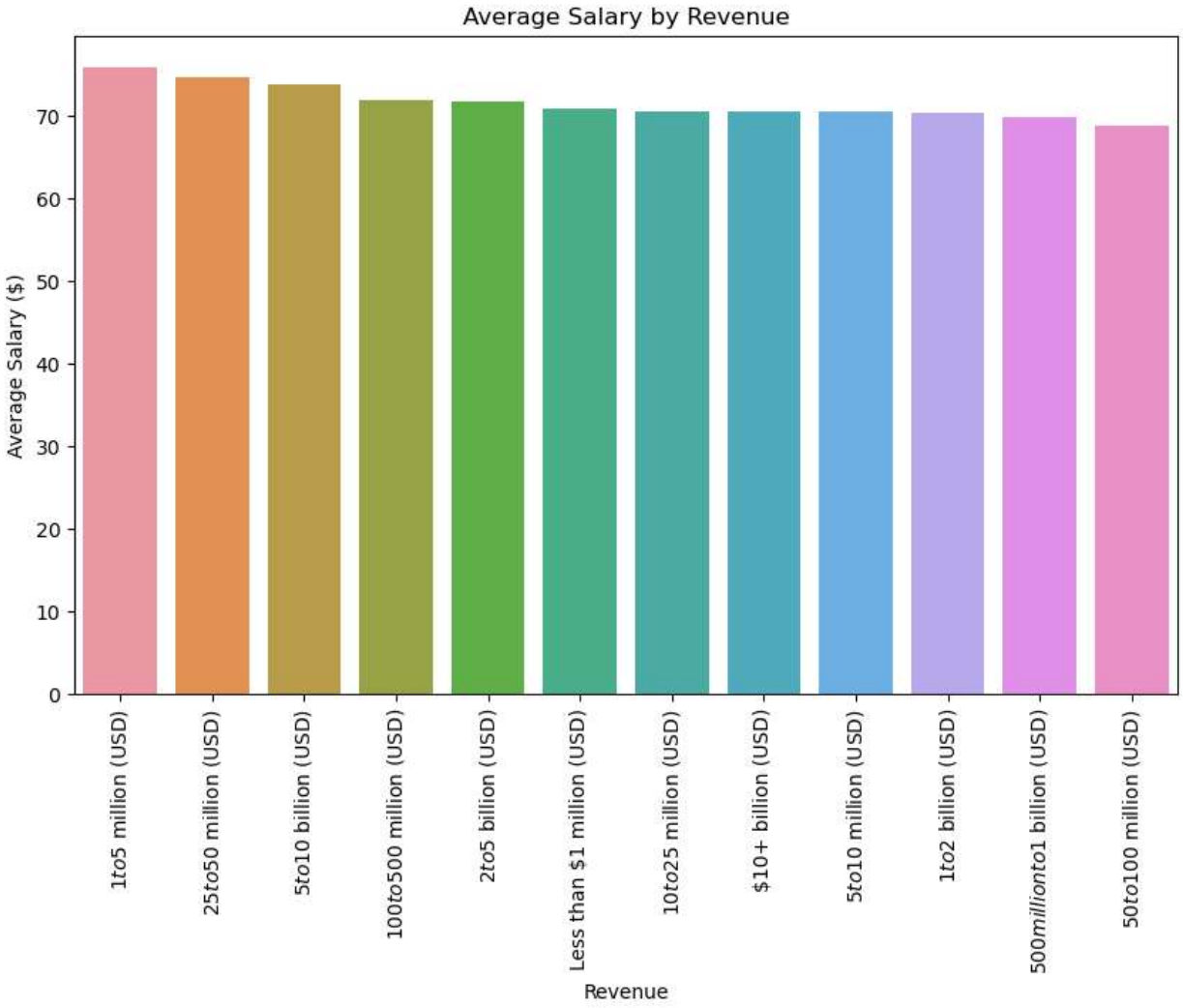


Key:

Top 3 Sectors by Salary: Biotech & Pharma, Real Estate and Arts, Entertaiment & Recreation;

Tail 3 Sectors by Salary: Tourism, Mining and Restaraunts & Food.

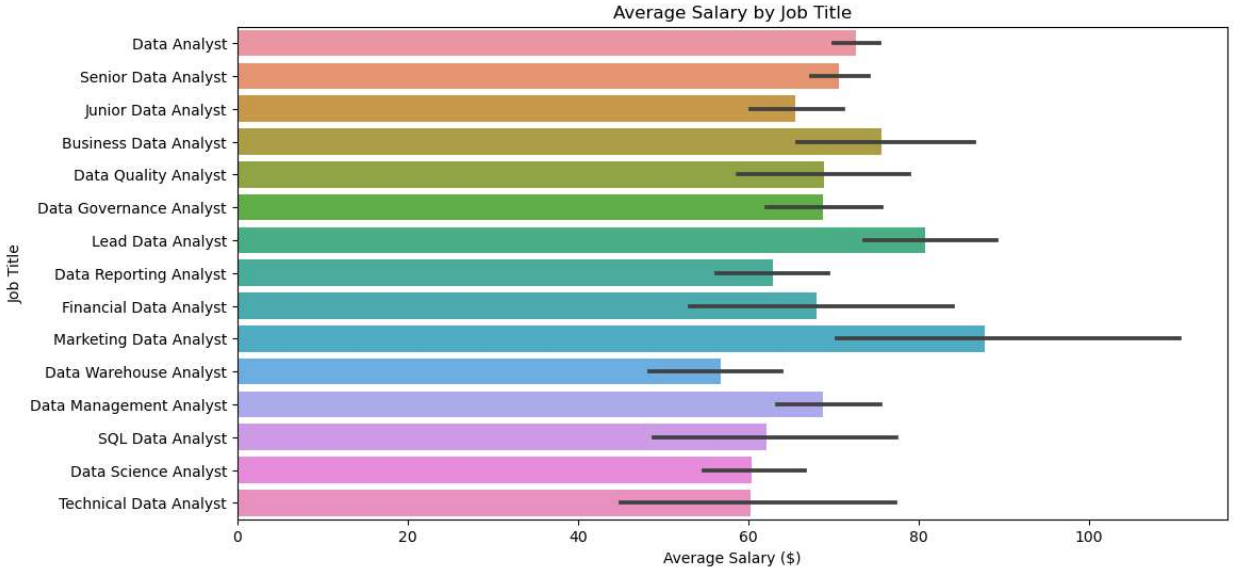
In [25]:
```
# Salary by Revenue

df_salaryXrevenue = filtered_revenue.groupby('Revenue')['AverageSalary'].mean().reset_
df_salaryXrevenue = df_salaryXrevenue.sort_values(by='AverageSalary', ascending=False)

plt.figure(figsize=(10, 6))
sns.barplot(x='Revenue', y='AverageSalary', data=df_salaryXrevenue)
plt.xlabel('Revenue')
plt.ylabel('Average Salary ($)')
plt.title('Average Salary by Revenue')
plt.xticks(rotation=90)
plt.show()
```



Key:

Revenue of a company doesn't affect Salary.

In [26]:
```
# Salary and Job Title

df_sorted = df.sort_values(by='AverageSalary', ascending=False)

plt.figure(figsize=(12, 6))
sns.barplot(x='AverageSalary', y='Job Title', data=df_sorted, orient='h', order=df_sor
plt.xlabel('Average Salary ($)')
plt.ylabel('Job Title')
plt.title('Average Salary by Job Title')
plt.show()
```

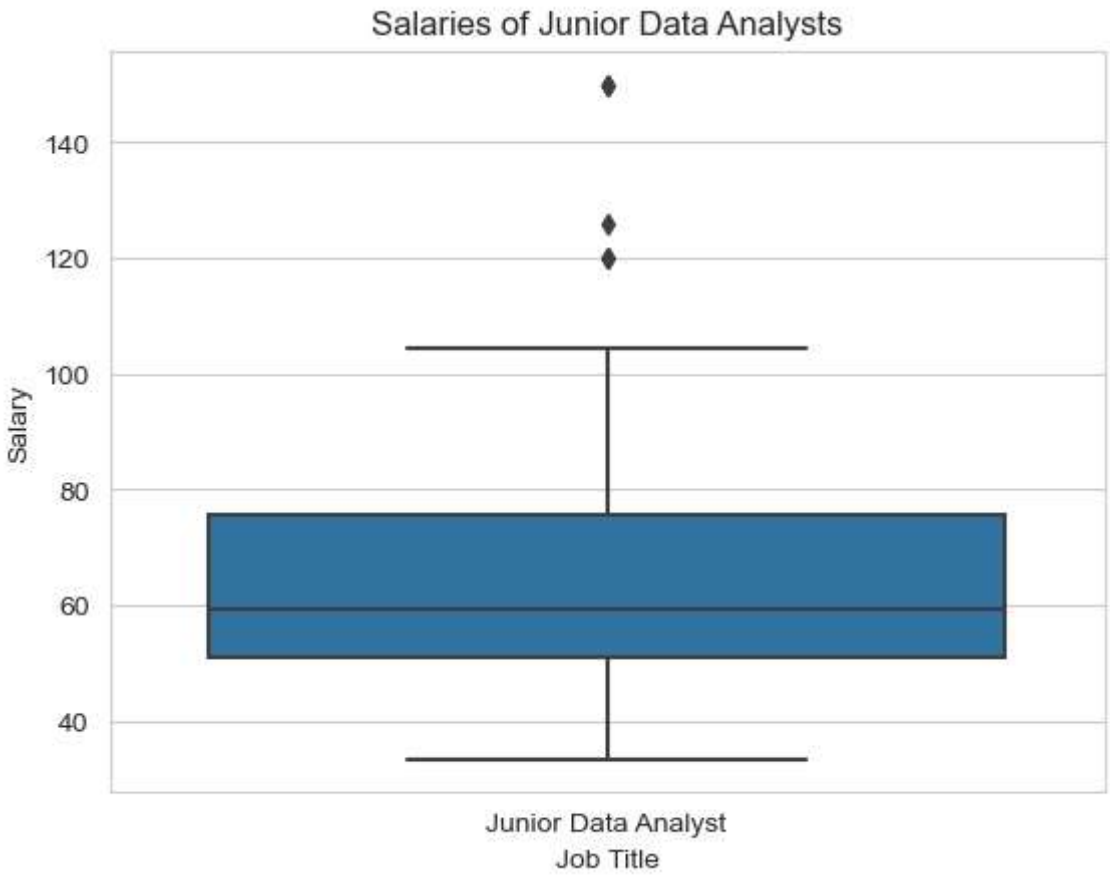


Key:

Top 3 titles by salary: Marketing Data Analysts, Lead Data Analysts and Business Data Analysts;

Last 3 are: Data Warehouse Analysts, SQL Data Analysts and Techical Data Analysts.

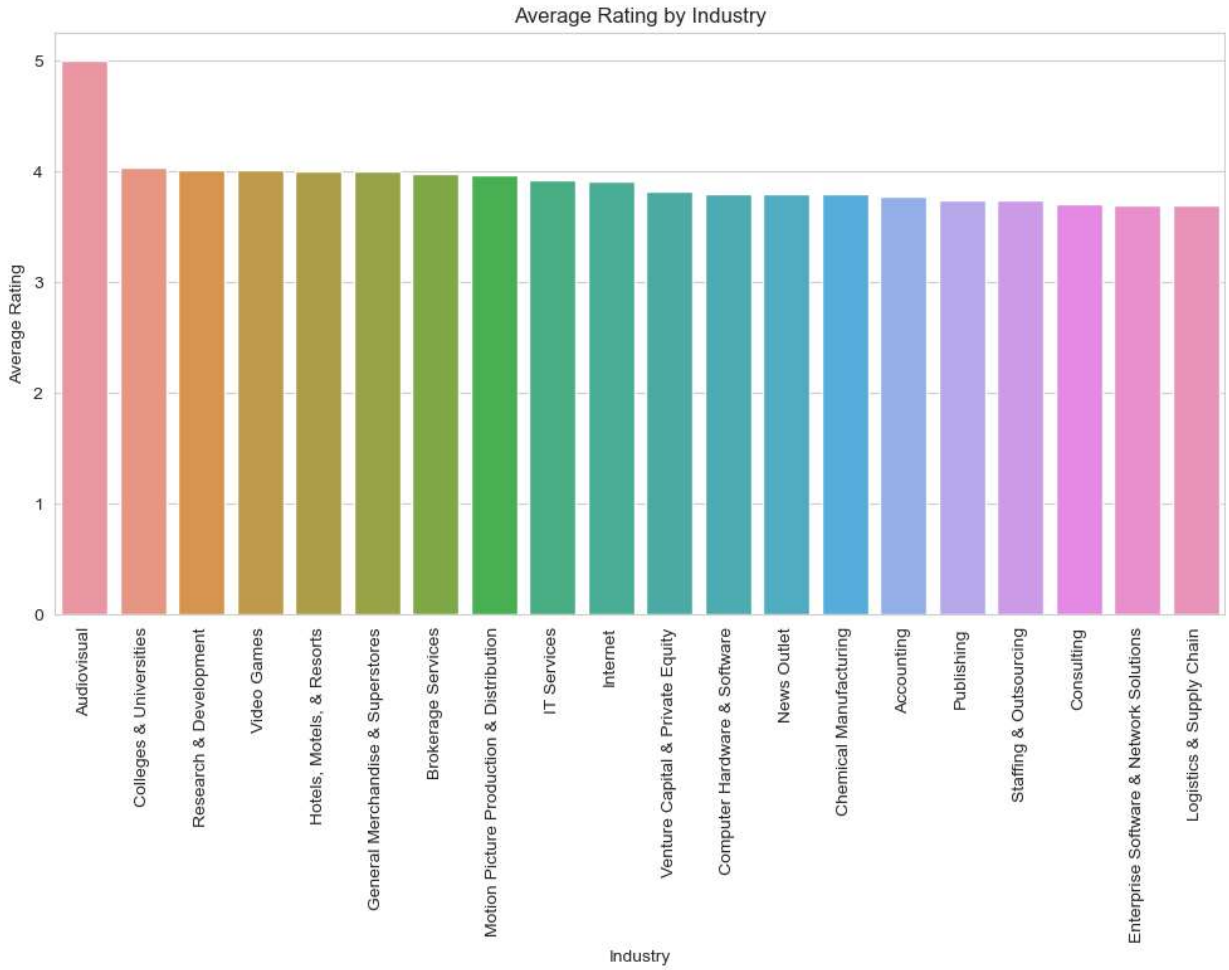
In [27]:
```
junior_data_analyst = df[df['Job Title'] == 'Junior Data Analyst']

sns.set_style('whitegrid')
sns.boxplot(x='Job Title', y='AverageSalary', data=junior_data_analyst)
plt.xlabel('Job Title')
plt.ylabel('Salary')
plt.title('Salaries of Junior Data Analysts')
plt.show()
```



Key:

Juniors might start with slightly below Average Salary.

In [28]:
```
# Rating by Industry

average_rating_by_industry = df[df['Industry'] != '-1'].groupby('Industry')['Rating'].

# Sort by average rating in descending order and take the top 20
average_rating_by_industry = average_rating_by_industry.sort_values(by='Rating', ascer

# Plot the bar plot
plt.figure(figsize=(12, 6))
sns.barplot(x='Industry', y='Rating', data=average_rating_by_industry)
plt.xticks(rotation=90)
plt.xlabel('Industry')
plt.ylabel('Average Rating')
plt.title('Average Rating by Industry')
plt.show()
```



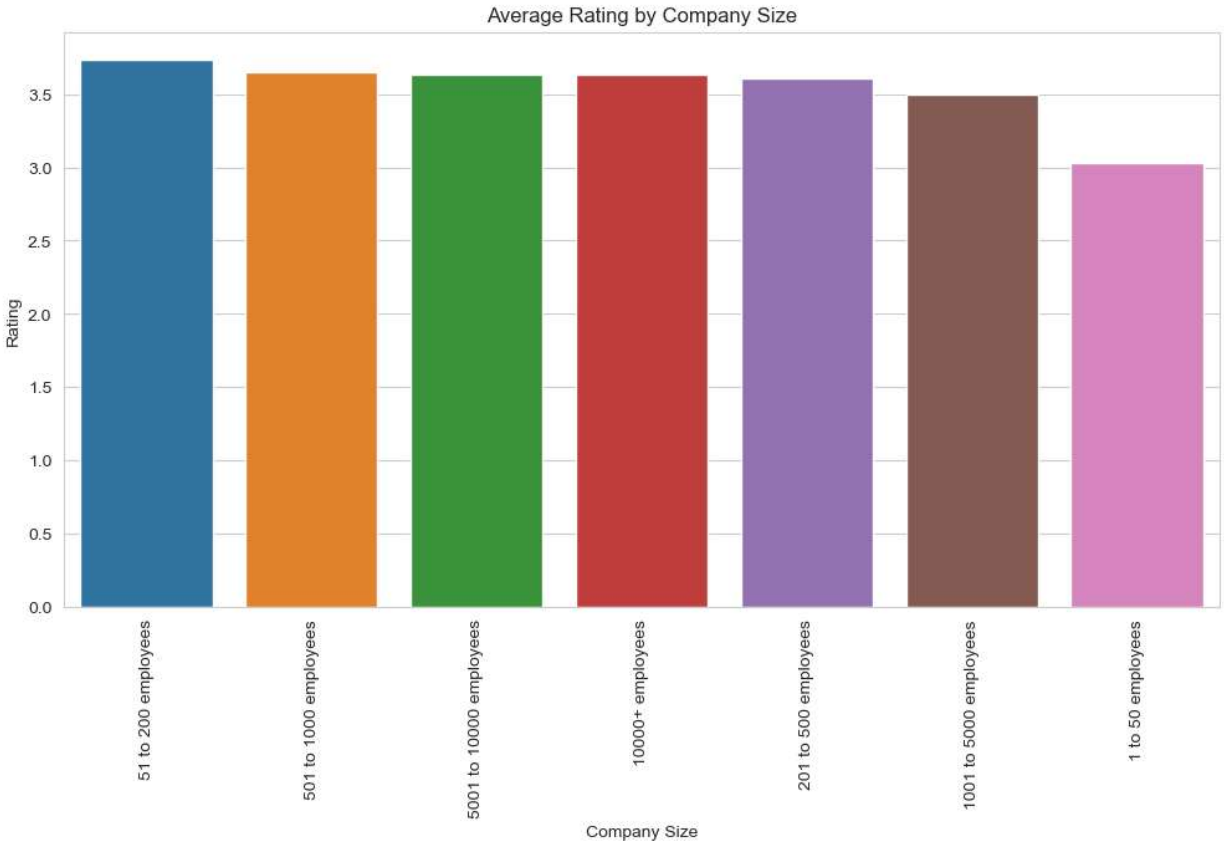
In [29]:
```
count_audiovisual = df['Industry'].value_counts().get('Audiovisual', 0)

print(count_audiovisual)
```

1

Key:

There is only one person who works in Audiovisual, so we don't count that. Average Rating is not affected by Industry.

In [30]:
```
# Rating by Size

df_sizeXrating = df[(df['Size'] != '-1') & (df['Size'] != 'Unknown')].groupby('Size')[
df_sizeXrating = df_sizeXrating.sort_values(by='Rating', ascending=False)

plt.figure(figsize=(12, 6))
sns.barplot(x='Size', y='Rating', data=df_sizeXrating)  # <- Correct DataFrame name he
plt.xlabel('Company Size')
plt.ylabel('Rating')
plt.title('Average Rating by Company Size')
plt.xticks(rotation=90)
plt.show()
```



Key:

Only those, who works in small start-ups are less satisfied with their company. In other sizes ratings are the same.

Rating by Revenue

In [31]:
```
df_revenueXrating = filtered_revenue.groupby('Revenue')['Rating'].mean().reset_index()
df_revenueXrating = df_revenueXrating.sort_values(by='Rating', ascending=False)

plt.figure(figsize=(10, 6))
sns.barplot(x='Revenue', y='Rating', data=df_revenueXrating)
plt.xlabel('Revenue')
plt.ylabel('Average Rating')
plt.title('Average Rating by Revenue')
plt.xticks(rotation=90)
plt.show()
```

Key:

Interviewers are mostly satisfied in companies with Revenue from 50 to 100 million dollars;

Less satisfied with Revenue from 1 to 5 million.

In [32]:
```
df.shape
```

Out[32]:
```
(2253, 10)
```

In [33]:
```
df.columns
```

Out[33]:
```
Index(['Unnamed: 0', 'Job Title', 'Rating', 'Location', 'Size',
       'Type of ownership', 'Industry', 'Sector', 'Revenue', 'AverageSalary'],
      dtype='object')
```

In [34]:
```
df.duplicated().sum()
```

Out[34]:
```
0
```

In [35]:
```
df.describe()
```

Out[35]:

	Unnamed: 0	Rating	AverageSalary
count	2253.0000	2253.000000	2252.000000
mean	1126.0000	3.160630	72.123002
std	650.5294	1.665228	23.600734
min	0.0000	-1.000000	33.500000
25%	563.0000	3.100000	58.000000
50%	1126.0000	3.600000	69.000000
75%	1689.0000	4.000000	80.500000
max	2252.0000	5.000000	150.000000

In [36]:
```
df.describe(include=object)
```

Out[36]:

	Job Title	Location	Size	Type of ownership	Industry	Sector	Revenue
count	2253	2253	2253	2253	2253	2253	2253
unique	1266	253	9	15	89	25	14
top	Data Analyst	New York, NY	51 to 200 employees	Company - Private	-1	Information Technology	Unknown / Non-Applicable
freq	405	310	421	1273	353	570	615

In [37]:
```
df.dtypes
```

Out[37]:
```
Unnamed: 0           int64
Job Title           object
Rating             float64
Location            object
Size                object
Type of ownership   object
Industry            object
Sector              object
Revenue             object
AverageSalary      float64
dtype: object
```

In [38]:
```
df.isnull().sum()
```

Out[38]:
```
Unnamed: 0         0
Job Title          0
Rating             0
Location           0
Size               0
Type of ownership  0
Industry           0
Sector             0
Revenue            0
AverageSalary      1
dtype: int64
```

Conclusions

Average Salary of Data Analyst is about 60-80 thousands dollars per year. It may be a little bit lower for Junior Data Analyst; Mostly interwviewers work in Start-ups or Big Corporations; Work offices are mostly based in states of NY, CA and TX; Most of the interviewed doesn't work in a specific industry or Sector, but focus on software development, consulting, and technical support to businesses in vast areas; Interviewers are mostly satisfied with companies, they work in; Company's Size doesn't affect it's raiting or Salary for Analysts; Company's Revenue doesn't affect it's raiting or Salary for Analysts; One of the lowest Rating among Analytics are Start-up / small companies; Top 3 Sectors by Salary: Biotech & Pharma, Real Estate and Arts, Entertaiment & Recreation. 3 Sectors with lowest Salary: Tourism, Mining and Restaraunts & Food; Top 3 Job titles by Salary: Marketing Data Analysts, Lead Data Analysts and Business Data Analysts. Last 3 are: Data Warehouse Analysts, SQL Data Analysts and Techical Data Analysts.