In [2]:
```
import pickle  as pkl
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

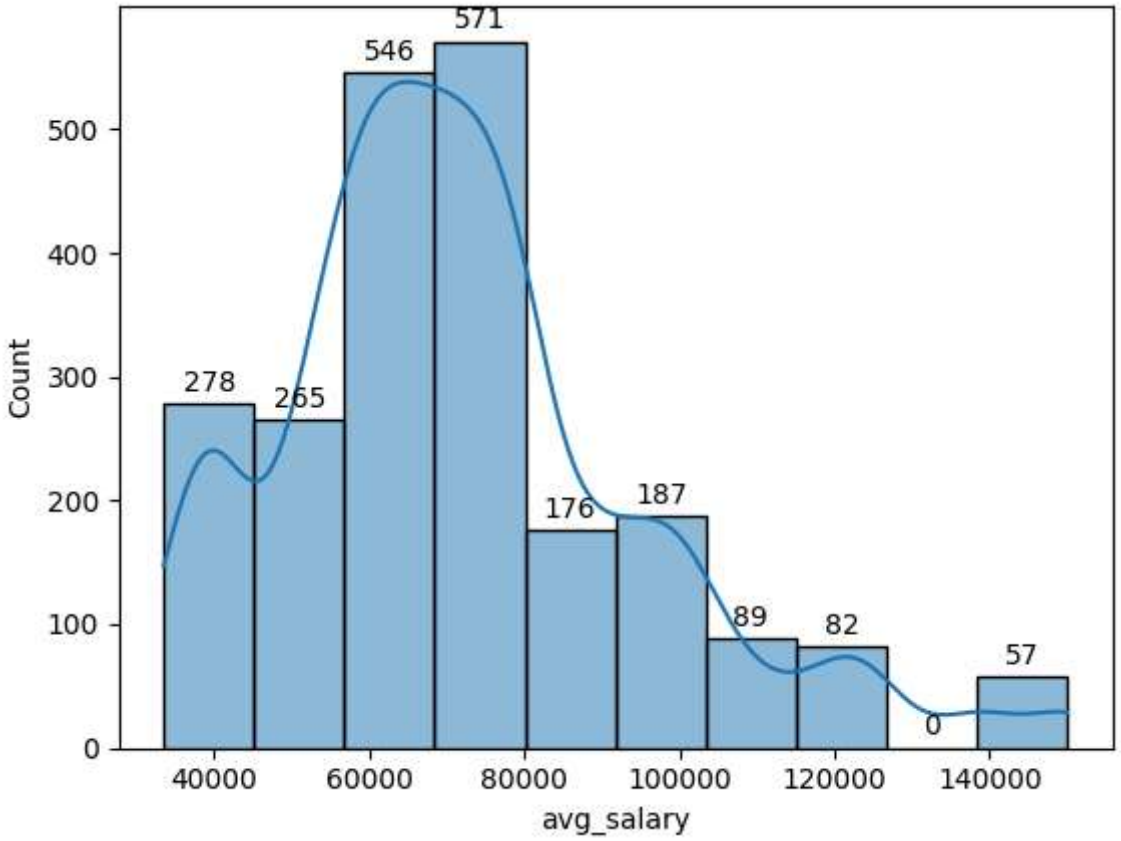
In [3]:
```
jobs_cleaned = pd.read_pickle('jobs_cleaned')
```

In [4]:
```
jobs_cleaned.head(2)
```

Out[4]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headqua
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 York, NY	New Yo
1	1	Quality Data Analyst	37K-66K	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New York, NY	New Yo

In [5]:
```
fg  = sns.histplot(data=jobs_cleaned['avg_salary'],bins=10, kde=True)
for patch in fg.patches:
    height = patch.get_height()
    width = patch.get_width()
    fg.text(x=patch.get_x()+width/2,
                y=height+10,
                s=height,
                ha='center')
```



In [6]:
```
jobs_cleaned.columns
```

Out[6]:
```
Index(['S.no.', 'Job Title', 'Salary Estimate', 'Job Description', 'Rating',
       'Company Name', 'Location', 'Headquarters', 'Size', 'Type of ownership',
       'Industry', 'Sector', 'Revenue', 'min_salary', 'max_salary',
       'avg_salary'],
      dtype='object')
```

In [7]:
```
jobs_cleaned['Job Title'].value_counts()
jobs_cleaned['Job Title'].nunique()
```

Out[7]:
```
1267
```

In [8]:
```
jobs_cleaned[jobs_cleaned['Job Title'].str.contains('data analyst', case=False)]
```

Out[8]:

	S.no.	Job Title	Salary Estimate	Job Description	Rating	Com N
0	0	Data Analyst, Center on Immigration and Justic...	37K-66K	Are you eager to roll up your sleeves and harn...	3.2	Institu Ju
1	1	Quality Data Analyst	37K-66K	Overview\n\nProvides analytical and technical ...	3.8	Vis N Servi New
2	2	Senior Data Analyst, Insights & Analytics Team...	37K-66K	We're looking for a Senior Data Analyst who ha...	3.4	Squares
3	3	Data Analyst	37K-66K	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Ce
4	4	Reporting Data Analyst	37K-66K	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	Far
...	...	...	...	...	...	
2244	2245	Data Analyst Supporting the DEA #20-242	78K-104K	Salary:\nPublished Job Title:\nData Analyst Su...	2.8	Forfe Sup Assoc
2246	2247	Marketing/Communications - Data Analyst-Marketing	78K-104K	Job Description\nJob Title: Marketing/Communic...	4.1	Soft Service
2248	2249	Senior Data Analyst (Corporate Audit)	78K-104K	Position:\nSenior Data Analyst (Corporate Audi...	2.9	A Electr
2250	2251	Data Analyst 3, Customer Experience	78K-104K	Summary\n\nResponsible for working cross-funct...	3.1	Contir Net Ser
2251	2252	Senior Quality Data Analyst	78K-104K	You.\n\nYou bring your body, mind, heart and s...	3.4	SCL H

1670 rows × 16 columns

In [ ]:

In [9]:
```
jobs_cleaned[jobs_cleaned['Job Title'].str.contains('sr data analyst',case=False)][
```

Out[9]:
```
Series([], Name: count, dtype: int64)
```

In [10]:
```
jobs_cleaned[jobs_cleaned['Job Title'].str.contains('senior data analyst',case=Fals
```

Out[10]:

Job Title	Company Name	Headquarters	Industry	Job Description	Location	Rating	Revenue
20-63 Flood Planning Data Analyst (Senior Data Analyst)	1	1	1	1	1	1	1
Application Programmer V/ Senior Data Analyst	1	1	1	1	1	1	1
Bilingual Senior Data Analyst (Japanese / English)	1	1	1	1	1	1	1
Business Senior Data Analyst	1	1	1	1	1	1	1
Business Senior Data Analyst - RQS 2018	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...
Senior Data Analyst/Data Warehouse consultant with Financial Healthcare systems	1	1	0	1	1	0	1
Senior Data Analysts (Banking Domain)	1	1	1	1	1	1	1
Senior data analyst	1	1	1	1	1	1	1
Software Engineer - Senior Data Analyst	1	1	1	1	1	1	1
Strategic Senior Data Analyst -	1	1	1	1	1	1	1

Job Title	Company Name	Headquarters	Industry	Job Description	Location	Rating	Revenue
Strategic Planning Divsion							

89 rows × 15 columns

In [11]:
```
jobs_cleaned['Job Title'] = jobs_cleaned['Job Title'].replace(['Sr. Data Analyst','
```

In [ ]:
```
```

In [12]:
```
# most_common_data_jobs.reset_index()
# most_common_data_jobs.reset_index()
jobs_cleaned['job_title'] =jobs_cleaned['Job Title'].replace(['Data Analyst I', 'da
jobs_cleaned['Job Title']= jobs_cleaned['Job Title'].replace('Data Analyst Junior',
```

In [13]:
```
jobs_cleaned['Job Title'] = jobs_cleaned['Job Title'].replace(['Data Analyst II', '
```

In [14]:
```
most_common_data_jobs = jobs_cleaned['Job Title'].value_counts().nlargest(5).reset_
# most_common_data_jobs.reset_index()
most_common_data_jobs
```

Out[14]:

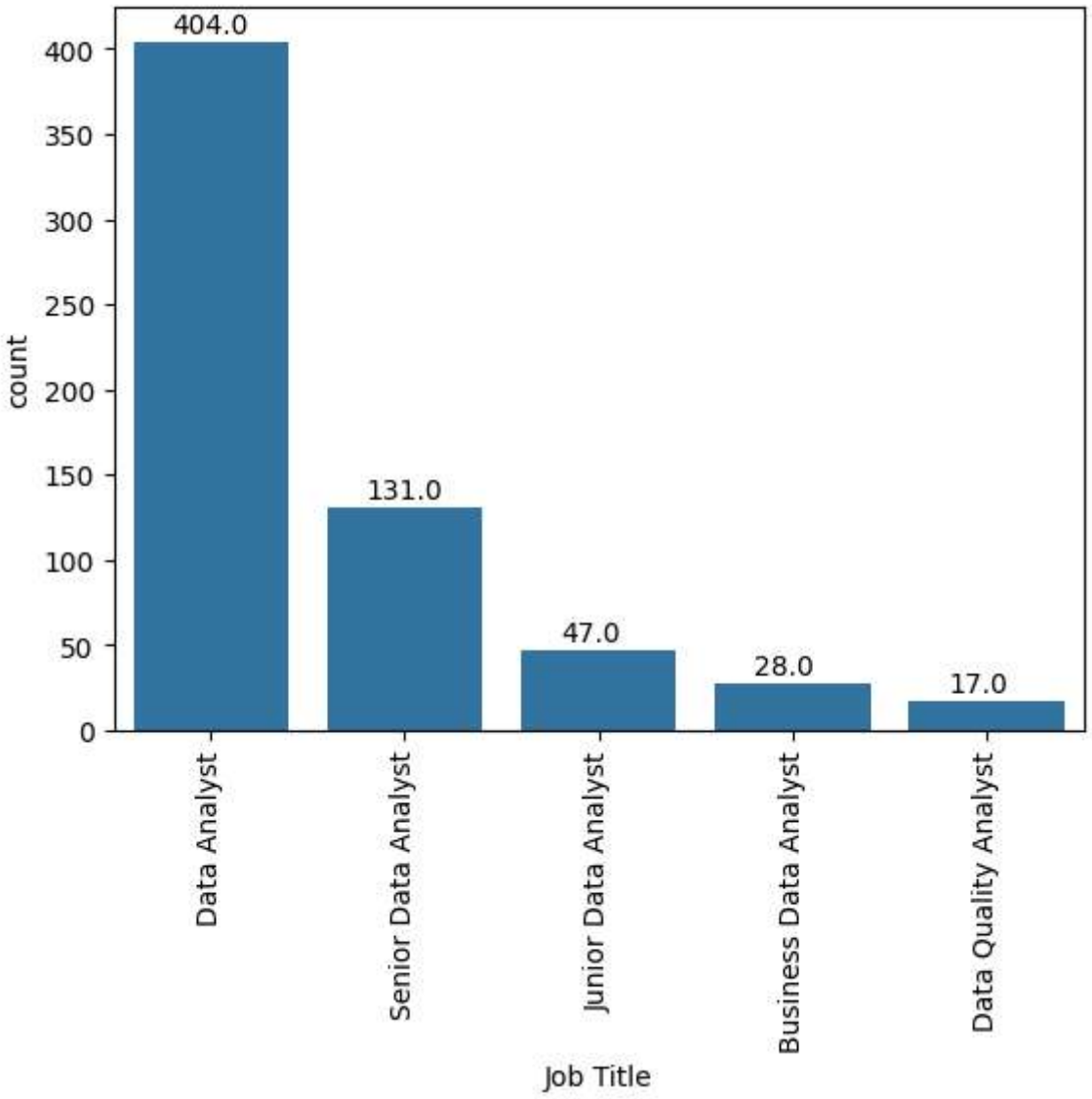
	Job Title	count
0	Data Analyst	404
1	Senior Data Analyst	131
2	Junior Data Analyst	47
3	Business Data Analyst	28
4	Data Quality Analyst	17

In [15]:
```
jobs_cleaned[jobs_cleaned['Job Title']=='Data Analyst Junior']
```

Out[15]:

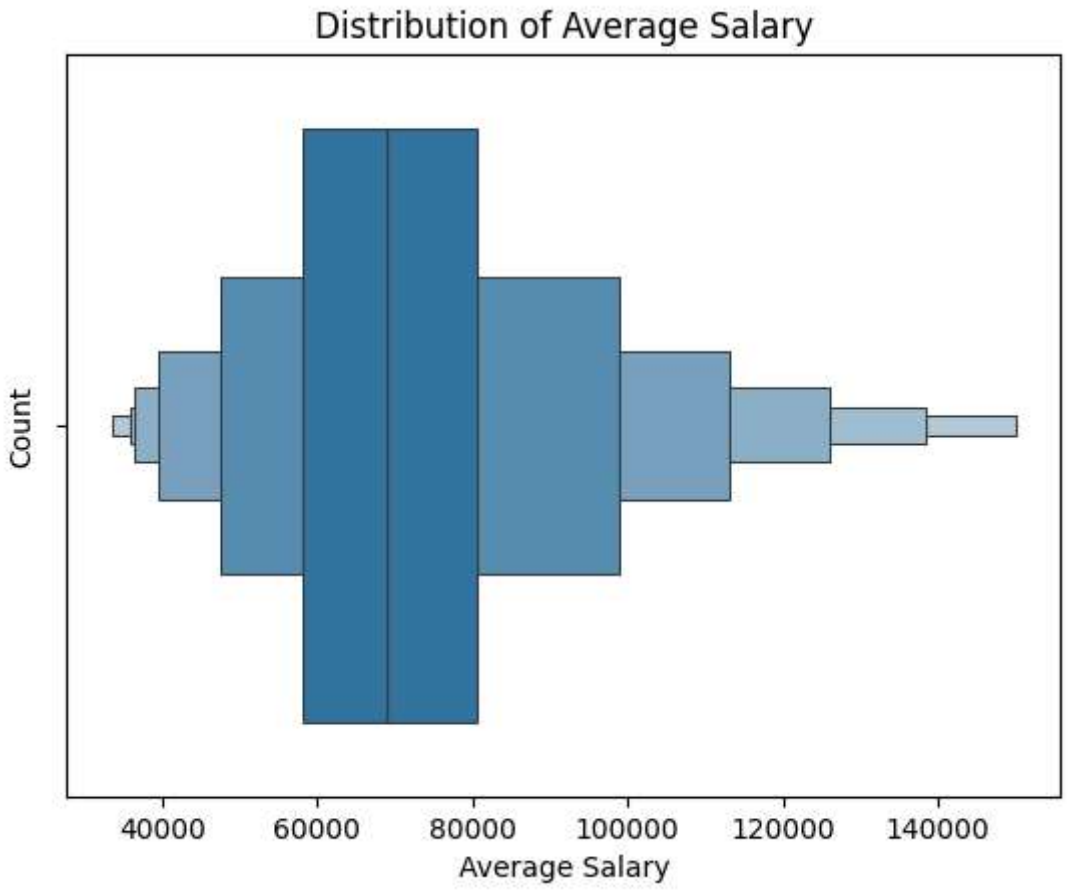
	S.no.	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	T own

In [16]:
```
plot = sns.barplot(data=most_common_data_jobs, x='Job Title', y='count')
plt.xticks(rotation=90)
for i in plot.patches:
    plt.text(i.get_x()+i.get_width()/4,i.get_height()+5, s = i.get_height())
```



In [17]:
```
jobs_cleaned = jobs_cleaned.drop('job_title', axis=1)
```

In [18]:
```
sns.boxenplot(data=jobs_cleaned, x='avg_salary')
plt.xlabel("Average Salary")
plt.ylabel("Count")
plt.title('Distribution of Average Salary')
plt.show()
```



In [19]:
```
jobs_cleaned['avg_salary']
```

Out[19]:
```
0       51500.0
1       51500.0
2       51500.0
3       51500.0
4       51500.0
         ...
2247    91000.0
2248    91000.0
2249    91000.0
2250    91000.0
2251    91000.0
Name: avg_salary, Length: 2251, dtype: float64
```

In [20]:
```
pd.to_pickle(jobs_cleaned, 'jobs_cleaned' )
```

In [ ]:
```

```