PORTFOLIO PROJECT 2 : GLASSDOOR and data cleaning

Description of the dataset

Derived from Glassdoor, a prominent job listing platform, this dataset compiles information from diverse data science job postings. It encompasses vital details such as job titles, company names, comprehensive descriptions, qualifications, salaries, and locations. Valuable for job seekers, employers, and researchers, this dataset illuminates trends within the data science job market, aiding decision-making in various domains.

Goal

The objective is to perform data cleaning operations, including checking the top row, determining the number of columns and rows, identifying columns with null values and their data types using the "info()" function, and checking for non-null values and duplicate entries in the dataset. If the dataset contains null values or duplicate entries, rows with such issues should be removed using the "drop" function from the dataset.

Importing dataset

In [67]:
```
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
```

In [68]:
```
df_raw_data=pd.read_csv('Uncleaned_DS_jobs.csv')
```

1) Overview

In [69]:
```
df_raw_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 672 entries, 0 to 671
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   index              672 non-null    int64  
 1   Job Title          672 non-null    object 
 2   Salary Estimate    672 non-null    object 
 3   Job Description    672 non-null    object 
 4   Rating             672 non-null    float64
 5   Company Name       672 non-null    object 
 6   Location           672 non-null    object 
```

```
 7   Headquarters       672 non-null    object
 8   Size               672 non-null    object
 9   Founded            672 non-null    int64
 10  Type of ownership  672 non-null    object
 11  Industry           672 non-null    object
 12  Sector             672 non-null    object
 13  Revenue            672 non-null    object
 14  Competitors        672 non-null    object
dtypes: float64(1), int64(2), object(12)
memory usage: 78.9+ KB
```

In [70]:
```
df_raw_data.head(3)
```

Out[70]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarte
0	0	Sr Data Scientist	$137K-$171K (Glassdoor est.)	Description\n\nThe Senior Data Scientist is re...	3.1	Healthfirst\n3.1	New York, NY	New York, N
1	1	Data Scientist	$137K-$171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech\n4.2	Chantilly, VA	Herndon, V
2	2	Data Scientist	$137K-$171K (Glassdoor est.)	Overview\n\n\nAnalysis Group is one of the lar...	3.8	Analysis Group\n3.8	Boston, MA	Boston, M

In [71]:
```
df_raw_data.columns
```

Out[71]:
```
Index(['index', 'Job Title', 'Salary Estimate', 'Job Description', 'Rating',
       'Company Name', 'Location', 'Headquarters', 'Size', 'Founded',
       'Type of ownership', 'Industry', 'Sector', 'Revenue', 'Competitors'],
      dtype='object')
```

2) Data Cleaning

In [72]:
```
df_raw_data.drop(labels='index',axis=1,inplace =True)
```

In [73]:
```
df_raw_data.head(3)
```

Out[73]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	
0	Sr Data Scientist	$137K-$171K (Glassdoor est.)	Description\n\nThe Senior Data Scientist is re...	3.1	Healthfirst\n3.1	New York, NY	New York, NY	em
1	Data Scientist	$137K-$171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech\n4.2	Chantilly, VA	Herndon, VA	em

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	
2	Data Scientist	$137K-$171K (Glassdoor est.)	Overview\n\n\nAnalysis Group is one of the lar...	3.8	Analysis Group\n3.8	Boston, MA	Boston, MA	em

In [74]:
```
df_raw_data[df_raw_data.duplicated()==True]
```

Out[74]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarter
135	Machine Learning Engineer	$90K-$109K (Glassdoor est.)	Role Description\nTriplebyte screens and evalu...	3.2	Triplebyte\n3.2	Remote	San Francisc C
136	Senior Data Engineer	$90K-$109K (Glassdoor est.)	Lendio is looking to fill a position for a Sen...	4.9	Lendio\n4.9	Lehi, UT	Lehi, U
358	Data Scientist	$122K-$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
359	Data Scientist	$122K-$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
360	Data Scientist	$122K-$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
361	Data Scientist	$122K-$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
362	Data Scientist	$122K-$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
389	Data Scientist	$110K-$163K (Glassdoor est.)	Job Description\nAs a Data Scientist, you will...	-1.0	HireAi	San Francisco, CA	-
496	Data Scientist	$95K-$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
497	Data Scientist	$95K-$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
498	Data Scientist	$95K-$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
499	Data Scientist	$95K-$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-
500	Data Scientist	$95K-$119K (Glassdoor	Job Overview: The Data Scientist is a key	-1.0	Hatch Data Inc	San Francisco,	-

Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarter
	est.)	memb...			CA	

In [75]:
```
print("dataframe shape considering all rows : "+ str(df_raw_data.shape))
print("dataframe shape considering ony duplicated rows : "+str(df_raw_data[df_raw_da
```

```
dataframe shape considering all rows : (672, 14)
dataframe shape considering ony duplicated rows : (13, 14)
```

In [76]:
```
#Let's drop duplicated rows
df_raw_data.drop_duplicates(inplace=True)
print("dataframe shape after deleting duplicated rows : "+str(df_raw_data.shape))
```

```
dataframe shape after deleting duplicated rows : (659, 14)
```

In [77]:
```
# Correcting the company name
df_raw_data['Company Name']=df_raw_data['Company Name'].apply(lambda x:x.split("\n")
df_raw_data.head(3)
```

Out[77]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Si
0	Sr Data Scientist	$137K-$171K (Glassdoor est.)	Description\n\nThe Senior Data Scientist is re...	3.1	Healthfirst	New York, NY	New York, NY	1001 50 employe
1	Data Scientist	$137K-$171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 100 employe
2	Data Scientist	$137K-$171K (Glassdoor est.)	Overview\n\n\nAnalysis Group is one of the lar...	3.8	Analysis Group	Boston, MA	Boston, MA	1001 50 employe

In [78]:
```
#Let's work on the salary column
for i in range(len(df_raw_data)):

    #Low salary
    try:
        df_raw_data.loc[i,'lowest salary']=df_raw_data.loc[i,'Salary Estimate'].spli
    except :
        df_raw_data.loc[i,'lowest salary']="NaN"

    # High salary
    try:
        df_raw_data.loc[i,'highest salary']=df_raw_data.loc[i,'Salary Estimate'].spl
    except :
        df_raw_data.loc[i,'highest salary']="NaN"
```

In [79]:
```
df_raw_data.head(3)
```

Out[79]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Si
0	Sr Data Scientist	$137K-$171K (Glassdoor	Description\n\nThe Senior Data Scientist is	3.1	Healthfirst	New York, NY	New York, NY	1001 50

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Si
		est.)	re...					employe
1	Data Scientist	$137K-$171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 100 employe
2	Data Scientist	$137K-$171K (Glassdoor est.)	Overview\n\n\nAnalysis Group is one of the lar...	3.8	Analysis Group	Boston, MA	Boston, MA	1001 50 employe

In [80]:
```
df_raw_data["lowest salary"]=df_raw_data["lowest salary"].apply(lambda x : str(x).re
df_raw_data["lowest salary"]=df_raw_data["lowest salary"].apply(lambda x : str(x).re
df_raw_data["highest salary"]=df_raw_data["highest salary"].apply(lambda x : str(x).
df_raw_data["highest salary"]=df_raw_data["highest salary"].apply(lambda x : str(x).
```

In [81]:
```
df_raw_data.head(3)
```

Out[81]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Si
0	Sr Data Scientist	$137K-$171K (Glassdoor est.)	Description\n\nThe Senior Data Scientist is re...	3.1	Healthfirst	New York, NY	New York, NY	1001 50 employe
1	Data Scientist	$137K-$171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 100 employe
2	Data Scientist	$137K-$171K (Glassdoor est.)	Overview\n\n\nAnalysis Group is one of the lar...	3.8	Analysis Group	Boston, MA	Boston, MA	1001 50 employe

In [82]:
```
#Setting the correct type for the newly created salary column
df_raw_data["lowest salary"]=df_raw_data["lowest salary"].apply(lambda x : x.replace
df_raw_data["highest salary"]=df_raw_data["highest salary"].apply(lambda x : x.repla
df_raw_data["lowest salary"]=df_raw_data["lowest salary"].apply(lambda x : x.replace
df_raw_data["highest salary"]=df_raw_data["highest salary"].apply(lambda x : x.repla

df_raw_data["lowest salary"]=df_raw_data["lowest salary"].astype('int')
df_raw_data["highest salary"]=df_raw_data["highest salary"].astype('int')
df_raw_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 672 entries, 0 to 500
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Job Title         659 non-null    object
 1   Salary Estimate   659 non-null    object
 2   Job Description   659 non-null    object
 3   Rating            659 non-null    float64
 4   Company Name      659 non-null    object
 5   Location          659 non-null    object
 6   Headquarters      659 non-null    object
 7   Size              659 non-null    object
 8   Founded           659 non-null    float64
```

```
 9   Type of ownership  659 non-null    object
 10  Industry           659 non-null    object
 11  Sector             659 non-null    object
 12  Revenue            659 non-null    object
 13  Competitors        659 non-null    object
 14  lowest salary      672 non-null    int32
 15  highest salary     672 non-null    int32
dtypes: float64(2), int32(2), object(12)
memory usage: 100.2+ KB
```

2) Statistical study and useful information

In [83]:
```
df_raw_data.head(3)
```

Out[83]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Si
0	Sr Data Scientist	$137K-$171K (Glassdoor est.)	Description\n\nThe Senior Data Scientist is re...	3.1	Healthfirst	New York, NY	New York, NY	1001 50 employe
1	Data Scientist	$137K-$171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 100 employe
2	Data Scientist	$137K-$171K (Glassdoor est.)	Overview\n\n\nAnalysis Group is one of the lar...	3.8	Analysis Group	Boston, MA	Boston, MA	1001 50 employe

List of all the positions :

In [84]:
```
sorted(df_raw_data['Job Title'].astype('str').unique().tolist())
```

Out[84]:
```
['(Sr.) Data Scientist -',
 'AI Data Scientist',
 'AI Ops Data Scientist',
 'AI/ML - Machine Learning Scientist, Siri Understanding',
 'Analytics - Business Assurance Data Analyst',
 'Analytics Manager',
 'Analytics Manager - Data Mart',
 'Applied AI Scientist / Engineer',
 'Applied Computer Scientist',
 'Applied Technology Researcher / Data Scientist',
 'Associate Data Scientist',
 'Aviation AI/ML Data Scientist',
 'Big Data Engineer',
 'Business Data Analyst',
 'Business Intelligence Analyst',
 'Business Intelligence Analyst I- Data Insights',
 'COMPUTER SCIENTIST - ENGINEER - RESEARCH COMPUTER SCIENTIST - SIGNAL PROCESSING',
 'COMPUTER SCIENTIST - ENGINEER - RESEARCH COMPUTER SCIENTIST - TRANSPORTATION TECHN
OLOGY',
 'Chief Scientist',
 'Clinical Data Analyst',
 'Cloud Data Engineer (Azure)',
 'Computational Behavioral Scientist',
 'Computational Scientist',
 'Computational Scientist, Machine Learning',
 'Computer Scientist 1',
```

```
 'Computer Vision / Deep Learning Scientist',
 'Data & Machine Learning Scientist',
 'Data Analyst',
 'Data Analyst - Unilever Prestige',
 'Data Analyst I',
 'Data Analyst II',
 'Data Analyst/Engineer',
 'Data Analytics Engineer',
 'Data Architect',
 'Data Engineer',
 'Data Engineer (Analytics, SQL, Python, AWS)',
 'Data Engineer (Remote)',
 'Data Engineer - Kafka',
 'Data Engineer, Digital & Comp Pathology',
 'Data Engineer, Enterprise Analytics',
 'Data Integration and Modeling Engineer',
 'Data Modeler',
 'Data Modeler (Analytical Systems)',
 'Data Science All Star Program - Data Engineer Track',
 'Data Science Analyst',
 'Data Science Instructor',
 'Data Science Manager',
 'Data Science Manager, Payment Acceptance - USA',
 'Data Science Software Engineer',
 'Data Scientist',
 'Data Scientist (TS/SCI w/ Poly)',
 'Data Scientist (TS/SCI)',
 'Data Scientist - Contract',
 'Data Scientist - Image and Video Analytics',
 'Data Scientist - Intermediate',
 'Data Scientist - Machine Learning',
 'Data Scientist - Risk',
 'Data Scientist - Statistics, Early Career',
 'Data Scientist - Statistics, Mid-Career',
 'Data Scientist - TS/SCI FSP or CI Required',
 'Data Scientist - TS/SCI Required',
 'Data Scientist / Applied Mathematician',
 'Data Scientist / Machine Learning Expert',
 'Data Scientist 3 (718)',
 'Data Scientist Machine Learning',
 'Data Scientist Technical Specialist',
 'Data Scientist(s)/Machine Learning Engineer',
 'Data Scientist, Applied Machine Learning - Bay Area',
 'Data Scientist, Kinship - NYC/Portland',
 'Data Scientist- Industrial Discrete Sector Industry',
 'Data Scientist-Human Resources',
 'Data Scientist/Data Analytics Practitioner',
 'Data Scientist/Machine Learning',
 'Data Solutions Engineer - Data Modeler',
 'Decision Scientist',
 'Developer III - Data Science',
 'Development Scientist, Voltaren',
 'Director of Data Science',
 'Diversity and Inclusion Data Analyst',
 'E-Commerce Data Analyst',
 'ELISA RESEARCH SCIENTIST (CV-15)',
 'ENGINEER - COMPUTER SCIENTIST - RESEARCH COMPUTER SCIENTIST - SIGNAL PROCESSING -
SAN ANTONIO OR',
 'Enterprise Data Analyst (Enterprise Portfolio Management Office)',
 'Environmental Data Science',
 'Equity Data Insights Analyst - Quantitative Analyst',
 'Experienced Data Scientist',
 'Geospatial Data Scientist',
 'Global Data Analyst',
```

```
 'Health Data Scientist - Biomedical/Biostats',
 'Health Plan Data Analyst, Sr',
 'Human Factors Scientist',
 'Hydrogen/Tritium Materials Scientist (Experienced)',
 'IT Partner Digital Health Technology and Data Science',
 'In-Line Inspection Data Analyst',
 'Information Systems Engineering Specialist (Engineering Scientist)',
 'Intelligence Data Analyst, Senior',
 'Jr. Business Data Analyst (position added 6/12/2020)',
 'Jr. Data Engineer',
 'Lead Certified Clinical Laboratory Scientist - Saturday - Tuesday, 8:00pm - 6:30am
shift',
 'Lead Data Scientist',
 'Lead Data Scientist – Network Analysis and Control',
 'Machine Learning Engineer',
 'Machine Learning Engineer, Sr.',
 'Machine Learning Engineer/Scientist',
 'Machine Learning Scientist - Bay Area, CA',
 'Machine Learning Scientist / Engineer',
 'Manager / Lead, Data Science & Analytics',
 'Manager, Field Application Scientist, Southeast',
 'Market Research Data Scientist',
 'Medical Lab Scientist',
 'NGS Scientist',
 'Operations Data Analyst',
 'Patient Safety- Associate Data Scientist',
 'Principal Data & Analytics Platform Engineer',
 'Principal Data Scientist',
 'Principal Data Scientist - Machine Learning',
 'Principal Machine Learning Scientist',
 'Principal Scientist/Associate Director, Quality Control and Analytical Technologie
s',
 'Product Data Scientist - Ads Data Science',
 'Production Engineer - Statistics/Data Analysis',
 'Purification Scientist',
 'RFP Data Analyst',
 'Real World Evidence (RWE) Scientist',
 'Real World Science, Data Scientist',
 'Report Writer-Data Analyst',
 'Research Scientist - Patient-Centered Research (Remote)',
 'Research Scientist Patient Preferences (Remote)',
 'Say Business Data Analyst',
 'Scientist - Biomarker and Flow Cytometry',
 'Scientist - Machine Learning',
 'Scientist - Molecular Biology',
 'Scientist / Group Lead, Cancer Biology',
 'Scientist/Research Associate-Metabolic Engineering',
 'Senior Analyst/Data Scientist',
 'Senior Business Intelligence Analyst',
 'Senior Clinical Data Scientist Programmer',
 'Senior Data & Machine Learning Scientist',
 'Senior Data Analyst',
 'Senior Data Analyst - Finance & Platform Analytics',
 'Senior Data Engineer',
 'Senior Data Scientist',
 'Senior Data Scientist - Algorithms',
 'Senior Data Scientist - R&D Oncology',
 'Senior Data Scientist – Image Analytics, Novartis AI Innovation Lab',
 'Senior Machine Learning Engineer',
 'Senior Machine Learning Scientist - Bay Area, CA',
 'Senior Principal Data Scientist (Python/R)',
 'Senior Research Statistician- Data Scientist',
 'Senior Scientist - Toxicologist - Product Integrity (Stewardship)',
 'Software Data Engineer',
```

```
 'Software Engineer (Data Scientist, C,C++,Linux,Unix) - SISW - MG',
 'Software Engineer - Data Science',
 'Software Engineer - Machine Learning & Data Science (Applied Intelligence Services
Team)',
 'Sr Data Analyst',
 'Sr Data Engineer (Sr BI Developer)',
 'Sr Data Scientist',
 'Sr Scientist - Extractables & Leachables',
 'Sr. Data Analyst',
 'Sr. Data Scientist',
 'Sr. Data Scientist II',
 'Sr. ML/Data Scientist - AI/NLP/Chatbot',
 'Sr. Research Associate/ Scientist, NGS prep & Molecular Genomics',
 'Staff BI and Data Engineer',
 'Staff Data Scientist',
 'Staff Data Scientist - Analytics',
 'Staff Data Scientist - Pricing',
 'Staff Scientist- Upstream PD',
 'Statistical Scientist',
 'Tableau Data Engineer 20-0117',
 'VP, Data Science',
 'Vice President, Biometrics and Clinical Data Management',
 'Weapons and Sensors Engineer/Scientist',
 'nan']
```

In [85]:
```
df_raw_data['Salary Estimate'].value_counts()
```

Out[85]:
```
$75K-$131K (Glassdoor est.)     32
$79K-$131K (Glassdoor est.)     32
$99K-$132K (Glassdoor est.)     32
$137K-$171K (Glassdoor est.)    30
$90K-$109K (Glassdoor est.)     28
$56K-$97K (Glassdoor est.)      22
$79K-$106K (Glassdoor est.)     22
$90K-$124K (Glassdoor est.)     22
$92K-$155K (Glassdoor est.)     21
$138K-$158K (Glassdoor est.)    21
$128K-$201K (Glassdoor est.)    21
$212K-$331K (Glassdoor est.)    21
$69K-$116K (Glassdoor est.)     21
$124K-$198K (Glassdoor est.)    21
$112K-$116K (Glassdoor est.)    21
$91K-$150K (Glassdoor est.)     21
$101K-$165K (Glassdoor est.)    21
$110K-$163K (Glassdoor est.)    20
$79K-$147K (Glassdoor est.)     20
$145K-$225K(Employer est.)      20
$31K-$56K (Glassdoor est.)      20
$141K-$225K (Glassdoor est.)    20
$66K-$112K (Glassdoor est.)     20
$80K-$132K (Glassdoor est.)     20
$87K-$141K (Glassdoor est.)     20
$105K-$167K (Glassdoor est.)    20
$79K-$133K (Glassdoor est.)     19
$71K-$123K (Glassdoor est.)     19
$122K-$146K (Glassdoor est.)    16
$95K-$119K (Glassdoor est.)     16
Name: Salary Estimate, dtype: int64
```

In [86]:
```
df_raw_data.describe()
```

Out[86]:

	Rating	Founded	lowest salary	highest salary
count	659.000000	659.000000	672.000000	672.000000
mean	3.592413	1661.701062	95.119048	142.361607
std	1.295563	733.544565	38.068385	55.732664
min	-1.000000	-1.000000	0.000000	0.000000
25%	3.400000	1932.500000	79.000000	116.000000
50%	3.800000	1995.000000	90.000000	132.000000
75%	4.300000	2009.000000	122.000000	163.000000
max	5.000000	2019.000000	212.000000	331.000000

In [87]:
```
df_raw_data.columns
```

Out[87]:
```
Index(['Job Title', 'Salary Estimate', 'Job Description', 'Rating',
       'Company Name', 'Location', 'Headquarters', 'Size', 'Founded',
       'Type of ownership', 'Industry', 'Sector', 'Revenue', 'Competitors',
       'lowest salary', 'highest salary'],
      dtype='object')
```

a) Average minimum wage by Location

In [88]:
```
# Salary study
df_salary=df_raw_data[['Job Title','Company Name','Location','Founded','lowest salar
df_salary=df_salary[df_salary['lowest salary']!=0]
df_salary_by_location=df_salary.groupby("Location")[["lowest salary"]].agg(lambda x:
df_salary_by_location=df_salary_by_location.sort_values(by=['lowest salary'], ascend
df_salary_by_location['lowest salary']=df_salary_by_location['lowest salary'].astype
df_salary_by_location=df_salary_by_location.rename(columns={'lowest salary':'lowest
df_salary_by_location
```

Out[88]:

Location	lowest salary(K$)
Colorado Springs, CO	31
Tulsa, OK	55
Southfield, MI	55
Harrisburg, PA	56
Oak Ridge, TN	63
...	...
Dayton, OH	155
Lexington Park, MD	158
Wilmington, DE	212
Fort Sam Houston, TX	212
Pleasanton, CA	212

202 rows × 1 columns

Interpretation : We can see that the top three locations that have the highest minimum salary offer for a job are Wilmington, DE,Fort Sam Houston, TX and Pleasanton, CA with a value of 212 K$ for each one of them

NOTE : At this point we just focused our study on the location but note that the same study could be done focusing on other parameters such as date of foundation, the sector...

b) Average maximum wage by Location

In [89]:
```
# Salary study
df_salary=df_raw_data[['Job Title','Company Name','Location','Founded','highest sala
df_salary=df_salary[df_salary['highest salary']!=0]
df_salary_by_location=df_salary.groupby("Location")[["highest salary"]].agg(lambda x
df_salary_by_location=df_salary_by_location.sort_values(by=['highest salary'], ascen
df_salary_by_location['highest salary']=df_salary_by_location['highest salary'].asty
df_salary_by_location=df_salary_by_location.rename(columns={'highest salary':'highes
df_salary_by_location
```

Out[89]:

Location	highest salary(K$)
Colorado Springs, CO	56
Tulsa, OK	81
Southfield, MI	94
Harrisburg, PA	97
Maple Plain, MN	102
...	...
Dayton, OH	231
Lexington Park, MD	249
Pleasanton, CA	331
Wilmington, DE	331
Fort Sam Houston, TX	331

202 rows × 1 columns

Interpretation : We can see that the top three locations that have the highest maximum salary offer for a job are Wilmington, DE,Fort Sam Houston, TX and Pleasanton, CA with a value of 331 K$ for each one of them

c) Number of offers by Location

In [93]:
```
df_offers_by_location=df_salary.groupby("Location").count()
df_offers_by_location=df_offers_by_location[['Job Title']]
df_offers_by_location=df_offers_by_location.rename(columns={'Job Title':'Nb of Job O
df_offers_by_location.sort_values("Nb of Job Offer")
```

Out[93]:

Location	Nb of Job Offer
Adelphi, MD	1
Norfolk, VA	1
New Orleans, LA	1
Mountain View, CA	1
Monterey, CA	1
...	...
Chicago, IL	22
Boston, MA	23
Washington, DC	26
New York, NY	49
San Francisco, CA	56

202 rows × 1 columns

In [99]:
```
df_offers_by_location.describe().astype('int')
```

Out[99]:

	Nb of Job Offer
count	202
mean	3
std	6
min	1
25%	1
50%	2
75%	3
max	56

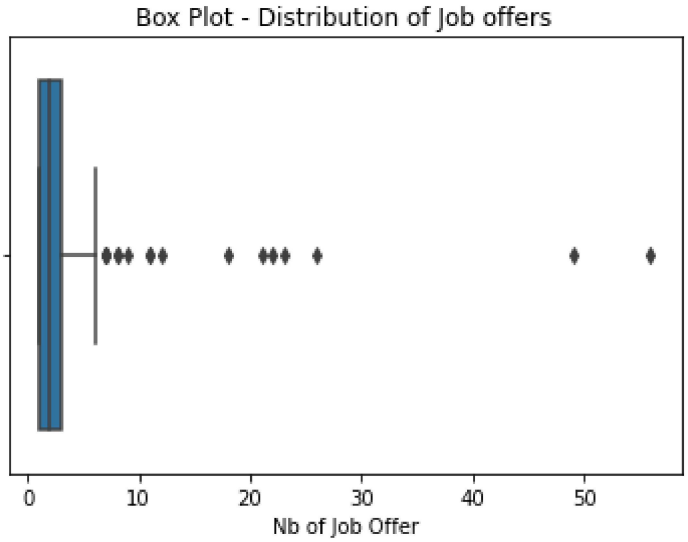
Interpretation :

These data show the number of different values taken by each feactures for the corresponding area. A such visual allow one to establish some characteristics like the number of job offers by region(columnn Job Title), The number of compagny name.

Visualization

In [105…
```
sb.boxplot(x=df_offers_by_location['Nb of Job Offer']).set_title('Box Plot - Distrib
```

Out[105…
```
Text(0.5, 1.0, 'Box Plot - Distribution of Job offers')
```



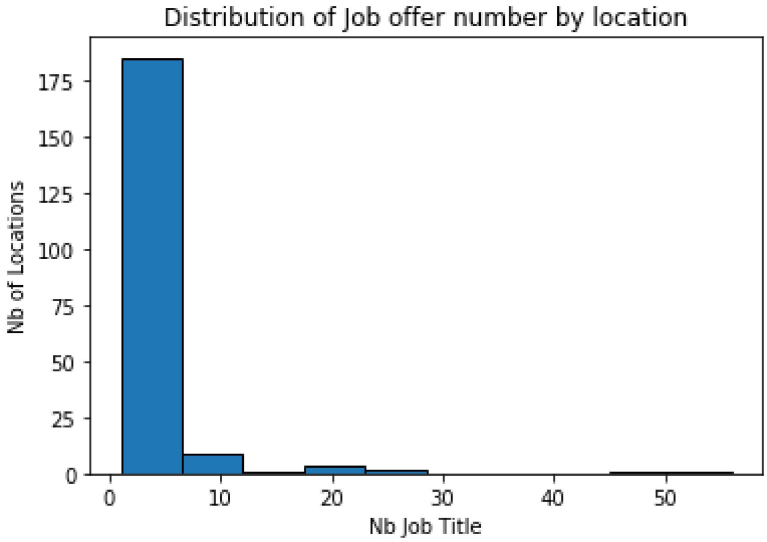
In [98]:

```
# Créer un histogramme
plt.hist(df_offers_by_location['Nb of Job Offers'], bins=10, edgecolor='black')

# Ajouter des labels et un titre
plt.xlabel('Nb Job Title')
plt.ylabel('Nb of Locations')
plt.title('Distribution of Job offer number by location')

#plt.xticks(range(min(df_offers_by_location['Nb of Job Offer']), max(df_offers_by_lo

# Afficher le graphique
plt.show()
```



Visualization interpratation:
We notice that most of the locations have between 1 and 15 job offers. We then see that a second group of location offer between 15 and 30 job offers. Finally, we have few location(not a lot) that have more than 6 job offers.

Interpretation : When one is looking for a job it is important to consider the location. The number of job offer by location can have an impact on wether someone would want to move out in a specific area in order to have more chance to find a certain type of job. In addition, this indicator can demonstrate that some fields may be more representred in a

particual region than in an other.

In this case, we see that San Francisco has more job offer than any other Location, followed by New York(NY) and Washington