MY Frist Project As Data Analyst

1- importing & Cleaning & Transformation Data

IMPORT REQUIUED LIBRARY& MODULES (ALSO DATASET)

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

#Loading_file & file_path

file_path ='Uncleaned_DS_jobs.csv'

uncleaned_df= pd.read_csv(file_path)
```

overview on dataset

In [2]:
```
uncleaned_df.shape
```

Out[2]: (672, 15)

In [3]:
```
uncleaned_df.head()
```

Out[3]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sect
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, NY	1001 to 5000 employees	1993	Nonprofit Organization	Insurance Carriers	Insuran
1	1	Data Scientist	137K–171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech\n4.2	Chantilly, VA	Herndon, VA	5001 to 10000 employees	1968	Company - Public	Research & Development	Busine Servic
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, MA	1001 to 5000 employees	1981	Private Practice / Firm	Consulting	Busine Servic
3	3	Data Scientist	137K–171K (Glassdoor est.)	JOB DESCRIPTION:\n\nDo you have a passion for ...	3.5	INFICON\n3.5	Newton, MA	Bad Ragaz, Switzerland	501 to 1000 employees	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturi
4	4	Data Scientist	137K–171K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Busine Servic

In [4]:
```
uncleaned_df.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 [5]:
```
uncleaned_df.isnull().sum()
```

Out[5]:
```
index              0
Job Title          0
Salary Estimate    0
Job Description    0
Rating             0
Company Name       0
Location           0
Headquarters       0
Size               0
Founded            0
Type of ownership  0
Industry           0
Sector             0
Revenue            0
Competitors        0
dtype: int64
```

I'm droping 'index',Because i am trying to Finding Duplicate values,When i don't remove the index python can understand they are unique row,So at this time we are using this method,Drop('index',aixs=1, inplace= True)

In [6]:
```
uncleaned_df.drop('index',axis=1,inplace=True)
uncleaned_df.shape
```

Out[6]: (672, 14)

now show number of duplicated values and rows that have duplicates values

In [7]:
```
print(uncleaned_df[uncleaned_df.duplicated()].shape)

uncleaned_df[uncleaned_df.duplicated()]
```

```
(13, 14)
```

Out[7]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenu
135	Machine Learning Engineer	$90K–$109K (Glassdoor est.)	Role Description\nTriplebyte screens and evalu...	3.2	Triplebyte\n3.2	Remote	San Francisco, CA	51 to 200 employees	2015	Company - Private	Computer Hardware & Software	Information Technology	Unknow / Nor Applicabl
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, UT	201 to 500 employees	2011	Company - Private	Lending	Finance	$50 to $10 millio (USD
358	Data Scientist	$122K–$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
359	Data Scientist	$122K–$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
360	Data Scientist	$122K–$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
361	Data Scientist	$122K–$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
362	Data Scientist	$122K–$146K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
389	Data Scientist	$110K–$163K (Glassdoor est.)	Job Description\nAs a Data Scientist, you will...	-1.0	HireAi	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
496	Data Scientist	$95K–$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
497	Data Scientist	$95K–$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
498	Data Scientist	$95K–$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
499	Data Scientist	$95K–$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-
500	Data Scientist	$95K–$119K (Glassdoor est.)	Job Overview: The Data Scientist is a key memb...	-1.0	Hatch Data Inc	San Francisco, CA	-1	-1	-1	-1	-1	-1	-

Drop Doplicates Values

In [8]:
```
uncleaned_df.drop_duplicates(inplace=True)
uncleaned_df.shape
```

Out[8]: (659, 14)

In [9]:
```
uncleaned_df.sample(10)
```

Out[9]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Se
79	Data Scientist 3 (718)	$79K–$131K (Glassdoor est.)	Amyris has developed a high-throughput genetic...	3.3	Amyris\n3.3	Emeryville, CA	Emeryville, CA	201 to 500 employees	2003	Company - Public	Biotech & Pharmaceuticals	Biote Pharmaceuti
115	Data Scientist	$99K–$132K (Glassdoor est.)	***Please note: All hiring and recruitment at ...	3.6	Spring Health\n3.6	New York, NY	New York, NY	1 to 50 employees	2016	Company - Private	Health Care Services & Hospitals	Health (
248	Data Scientist	$90K–$124K (Glassdoor est.)	Descript is a new kind of audio/video creation...	4.3	Descript\n4.3	San Francisco, CA	Houston, TX	1 to 50 employees	-1	Company - Private	-1	
120	Data Scientist	$99K–$132K (Glassdoor est.)	TACG is a driven Alaskan Native 8(a) professio...	4.5	TACG Solutions\n4.5	Dayton, OH	Beavercreek, OH	51 to 200 employees	2006	Company - Private	Consulting	Busir Serv
231	Research Scientist - Patient-Centered Research...	$71K–$123K (Glassdoor est.)	*This position can be in either of our Evidera...	3.8	Evidera\n3.8	Bethesda, MD	Bethesda, MD	501 to 1000 employees	2013	Subsidiary or Business Segment	Biotech & Pharmaceuticals	Biote Pharmaceuti
289	Data Scientist	$141K–$225K (Glassdoor est.)	We're looking for data scientists to work on o...	3.4	Mackin\n3.4	Menlo Park, CA	Pittsburgh, PA	51 to 200 employees	1960	Company - Private	Architectural & Engineering Services	Busir Serv
563	Data Science Manager	$128K–$201K (Glassdoor est.)	Job Requisition ID #\n20WD38093\nJob Title\nDa...	4.0	Autodesk\n4.0	San Francisco, CA	San Rafael, CA	5001 to 10000 employees	1982	Company - Public	Computer Hardware & Software	Informa Techno
610	Data Scientist	$80K–$132K (Glassdoor est.)	Tygart is currently seeking Data Scientist to ...	4.7	Tygart Technology, Inc\n4.7	Washington, DC	Fairmont, WV	1 to 50 employees	-1	Company - Private	-1	
504	Data Scientist	$95K–$119K (Glassdoor est.)	Job Description\nWorking at Sophinea\n\nSophin...	-1.0	Sophinea	Chantilly, VA	-1	1 to 50 employees	-1	Unknown	-1	
288	Data Scientist	$141K–$225K (Glassdoor est.)	About Joby\nLocated in Northern California, th...	4.3	Joby Aviation\n4.3	San Carlos, CA	Santa Cruz, CA	51 to 200 employees	-1	Company - Private	-1	

After Remove Null-values,Duplicate (DataSet)
Update Columns Names

In [10]:
```
new_columns_uncleaned  = ['Job_title','Salary_estimate_range','Job_Description','Job_post_rating',
                'Company_Name','Job_Location','Company_Headquarters','Company_size','Company_founded_at',
                'Company_ownership','Company_industry','Company_Sector','Comapny_Revenue_range','Company_competitors']
uncleaned_df.columns = new_columns_uncleaned
uncleaned_df.head()
```

Out[10]:

	Job_title	Salary_estimate_range	Job_Description	Job_post_rating	Company_Name	Job_Location	Company_Headquarters	Company_size	Company_fou
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	1001 to 5000 employees	
1	Data Scientist	137K–171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech\n4.2	Chantilly, VA	Herndon, VA	5001 to 10000 employees	
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	1001 to 5000 employees	
3	Data Scientist	137K–171K (Glassdoor est.)	JOB DESCRIPTION:\n\nDo you have a passion for ...	3.5	INFICON\n3.5	Newton, MA	Bad Ragaz, Switzerland	501 to 1000 employees	
4	Data Scientist	137K–171K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 to 200 employees	

Salary Estimate Column

Check Values in ('Salary Estimate').column

Create New columns (min salary,max salary,avg salary)

Remove Extra symbols And character in Values

In [11]:
```
uncleaned_df['Salary_estimate_range'].value_counts()
```

Out[11]:
```
$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_range, dtype: int64
```

In [37]:
```
uncleaned_df['min_salary']=0
uncleaned_df['max_salary']=0
uncleaned_df['avg_salary']=0
for i in range(len(uncleaned_df)):
    try:
        uncleaned_df.loc[i,"min_salary"]=int(uncleaned_df['Salary_estimate_range'][i].split(" ")[0].split("-")[0].replace("$","")
        uncleaned_df.loc[i,"max_salary"]=int(uncleaned_df['Salary_estimate_range'][i].split(" ")[0].split("-")[1].replace("$","")
    except:
        uncleaned_df.loc[i,"min_salary"]=int(uncleaned_df['Salary_estimate_range'][i].split("(E")[0].split("-")[0].replace("$","'
        uncleaned_df.loc[i,"max_salary"]=int(uncleaned_df['Salary_estimate_range'][i].split("(E")[0].split("-")[1].replace("$","'
    finally:
        uncleaned_df.loc[i,"Salary_estimate_range"]=str(uncleaned_df.loc[i,"min_salary"])+"-"+str(uncleaned_df.loc[i,"max_salary"
        uncleaned_df.loc[i,"avg_salary"]=np.mean([uncleaned_df.loc[i,"min_salary"],uncleaned_df.loc[i,"max_salary"]])
```

Removing Number in Company Name columns

In [12]:
```
uncleaned_df['Company_Name']=uncleaned_df['Company_Name'].apply(lambda x:x.split("\n")[0])
```

**Exctract Data Job Description **

Create New Columns Based on Skill required for Job (Python,Excel,Sql,Tableau,Sparx,mechine learing,Aws) If Job description Contain Those skill in column print '1' otherwise '0' will print

In [13]:
```
uncleaned_df['Job_Description'][0].split("\n\n")
```

Out[13]:
```
['Description',
 'The Senior Data Scientist is responsible for defining, building, and improving statistical models to improve business process
es and outcomes in one or more healthcare domains such as Clinical, Enrollment, Claims, and Finance. As part of the broader ana
lytics team, Data Scientist will gather and analyze data to solve and address complex business problems and evaluate scenarios
to make predictions on future outcomes and work with the business to communicate and support decision-making. This position req
uires strong analytical skills and experience in analytic methods including multivariate regressions, hierarchical linear model
s, regression trees, clustering methods and other complex statistical techniques.',
 'Duties & Responsibilities:',
 '• Develops advanced statistical models to predict, quantify or forecast various operational and performance metrics in multip
le healthcare domains\n• Investigates, recommends, and initiates acquisition of new data resources from internal and external s
ources\n• Works with multiple teams to support data collection, integration, and retention requirements based on business needs
\n• Identifies critical and emerging technologies that will support and extend quantitative analytic capabilities\n• Collaborat
```

```
es with business subject matter experts to select relevant sources of information\n• Develops expertise with multiple machine l
earning algorithms and data science techniques, such as exploratory data analysis and predictive modeling, graph theory, recomm
ender systems, text analytics and validation\n• Develops expertise with Healthfirst datasets, data repositories, and data movem
ent processes\n• Assists on projects/requests and may lead specific tasks within the project scope\n• Prepares and manipulates
data for use in development of statistical models\n• Other duties as assigned',
 'Minimum Qualifications:',
 "-Bachelor's Degree",
 'Preferred Qualifications:',
 '- Master’s degree in Computer Science or Statistics\nFamiliarity with major cloud platforms such as AWS and Azure\nHealthcare
Industry Experience',
 'Minimum Qualifications:',
 "-Bachelor's Degree",
 'Preferred Qualifications:',
 '- Master’s degree in Computer Science or Statistics\nFamiliarity with major cloud platforms such as AWS and Azure\nHealthcare
Industry Experience',
 'WE ARE AN EQUAL OPPORTUNITY EMPLOYER. Applicants and employees are considered for positions and are evaluated without regard
to mental or physical disability, race, color, religion, gender, national origin, age, genetic information, military or veteran
status, sexual orientation, marital status or any other protected Federal, State/Province or Local status unrelated to the perf
ormance of the work involved.',
 'If you have a disability under the Americans with Disability Act or a similar law, and want a reasonable accommodation to ass
ist with your job search or application for employment, please contact us by sending an email to careers@Healthfirst.org or cal
ling 212-519-1798 . In your email please include a description of the accommodation you are requesting and a description of the
position for which you are applying. Only reasonable accommodation requests related to applying for a position within Healthfir
st Management Services will be reviewed at the e-mail address and phone number supplied. Thank you for considering a career wit
h Healthfirst Management Services.\nEEO Law Poster and Supplement',
 ']]>']
```

In [14]:
```
uncleaned_df['python']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "python" in x.lower() else 0)
uncleaned_df['excel']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "excel" in x.lower() else 0)
uncleaned_df['Sql']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "sql" in x.lower() else 0)
uncleaned_df['no Sql']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "nosql" in x.lower() else 0)
uncleaned_df['Cloud']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "cloud" in x.lower() else 0)
uncleaned_df['Visualization']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "visualization" in x.lower() else 0)
uncleaned_df['BI']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "powerbi" in x.lower() else 0)
uncleaned_df['tableau']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "tableau" in x.lower() else 0)
uncleaned_df['Sparx']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "sparx" in x.lower() else 0)
uncleaned_df['ML']=uncleaned_df['Job_Description'].apply(lambda x: 1 if ("machine learning") in x.lower() else 0)
uncleaned_df['aws']=uncleaned_df['Job_Description'].apply(lambda x: 1 if "aws" in x.lower() else 0)
uncleaned_df.sample(20)
```

Out[14]:

	Job_title	Salary_estimate_range	Job_Description	Job_post_rating	Company_Name	Job_Location	Company_Headquarters	Company_size	Comp
503	Data Scientist	$95K−119K (Glassdoor est.)	Title: Data Scientist\n\nLocation: Chicago or ...	4.0	Two95 International Inc.	Chicago, IL	Cherry Hill, NJ	1 to 50 employees	
363	Senior Data Scientist	$122K−146K (Glassdoor est.)	Do you have a head for numbers? Like turning r...	3.5	Maxar Technologies	Herndon, VA	Westminster, CO	5001 to 10000 employees	
501	Data Scientist	$95K−119K (Glassdoor est.)	Job Description\nData Scientist\n\nLocation: T...	4.6	Murray Resources	The Woodlands, TX	Houston, TX	1 to 50 employees	
143	AI Data Scientist	$90K−109K (Glassdoor est.)	A chance to provide active support to our spon...	3.3	MITRE	McLean, VA	Bedford, MA	5001 to 10000 employees	
352	Decision Scientist	$122K−146K (Glassdoor est.)	Are you passionate about Decision Science?\n\n...	4.5	Johns Hopkins University Applied Physics Labor...	Laurel, MD	Laurel, MD	5001 to 10000 employees	
559	Data Scientist	$128K−201K (Glassdoor est.)	Data Works is an employee-focused technology c...	4.5	E3 Federal Solutions	Chantilly, VA	McLean, VA	501 to 1000 employees	
321	Development Scientist, Voltaren	$145K−225K(Employer est.)	Site Name: Richmond Sherwood\nPosted Date: Mar...	3.9	GSK	Richmond, VA	Brentford, United Kingdom	10000+ employees	
14	Data Scientist	$137K−171K (Glassdoor est.)	Position Description:\n\nWant to make a differ...	3.4	Mathematica Policy Research	Washington, DC	Princeton, NJ	1001 to 5000 employees	
602	Decision Scientist	$80K−132K (Glassdoor est.)	Are you passionate about Decision Science?\n\n...	4.5	Johns Hopkins University Applied Physics Labor...	Laurel, MD	Laurel, MD	5001 to 10000 employees	
297	Data Analyst	$141K−225K (Glassdoor est.)	ShorePoint is a cybersecurity services firm wi...	4.5	ShorePoint	Reston, VA	Washington, DC	1 to 50 employees	
385	Senior Data Scientist	$110K−163K (Glassdoor est.)	Klaviyo is looking for Senior Data Scientists ...	4.8	Klaviyo	Boston, MA	Boston, MA	201 to 500 employees	
128	Data Engineer	$90K−109K (Glassdoor est.)	IZEA was built to connect the world's top bran...	4.2	IZEA	Winter Park, FL	Winter Park, FL	51 to 200 employees	
491	Data Analyst	$95K−119K (Glassdoor est.)	Nolij Consulting LLC is a certified Women-Owne...	3.9	Nolij Consulting	Falls Church, VA	Vienna, VA	51 to 200 employees	
163	Data Analyst	$101K−165K (Glassdoor est.)	We are currently seeking an experienced, dynam...	2.7	USAC	Washington, DC	Washington, DC	501 to 1000 employees	
235	Data Scientist	$71K−123K (Glassdoor est.)	Job Description\nKey Job Duties and Responsibi...	3.3	II-VI Incorporated	Champaign, IL	Saxonburg, PA	10000+ employees	
341	Data Scientist	$79K−147K (Glassdoor est.)	Hi,\n\nGreetings of the Day!!!\n\nLooking for ...	4.8	TechProjects	New York, NY	North Brunswick, NJ	1 to 50 employees	
492	Data Scientist	$95K−119K (Glassdoor est.)	COMPANY OVERVIEW INFICON is a growing, global,...	3.4	Dice.com	Newton, MA	Denver, CO	201 to 500 employees	
289	Data Scientist	$141K−225K (Glassdoor est.)	We're looking for data scientists to work on o...	3.4	Mackin	Menlo Park, CA	Pittsburgh, PA	51 to 200 employees	
59	Data Scientist - Intermediate	$75K−131K (Glassdoor est.)	Job description:\n\nCreate and maintain code f...	4.5	Envision LLC	Saint Louis, MO	Saint Louis, MO	201 to 500 employees	
666	Data Scientist	$105K−167K (Glassdoor est.)	About Foundation Medicine:\n\nFoundation Medic...	4.0	Foundation Medicine	Boston, MA	Cambridge, MA	1001 to 5000 employees	

20 rows × 25 columns

Create New Column State Useing ['Location'] column

In [15]:
```
uncleaned_df['Job_Location'].apply(lambda x:x.split(",")[-1]).value_counts()
```

Out[15]:
```
CA    154
VA     89
MA     62
NY     52
MD     40
TI     30
```

```
 DC                26
 TX                17
 WA                16
 OH                14
 PA                12
 MO                12
United States      11
 NJ                10
 CO                10
 NC                 9
 GA                 9
 TN                 8
 FL                 8
 OK                 6
 WI                 6
Remote             5
 IN                5
 MI                5
 CT                4
 AL                4
 MN                4
 AZ                4
 NE                3
 IA                3
 RI                2
New Jersey         2
 SC                2
 OR                2
 UT                2
Utah               2
 NH                2
 MS                1
 LA                1
 KS                1
Texas              1
 DE                1
California         1
 WV                1
Name: Job_Location, dtype: int64
```

In [16]:
```
uncleaned_df['State']= uncleaned_df['Job_Location'].apply(lambda x:x.split(",")[-1].strip())
uncleaned_df['State'].replace(['United States','New Jersey','California','Texas ','Remote','Utah'],["US","NJ","CA","Tx",np.nan,np
uncleaned_df.dropna(axis=0,inplace= True)
uncleaned_df.reset_index(inplace= True)
```

Create New Column Employeetype Based on ['Job_Titile']

Based on the job title tag like sr, senior,lead,vp...etc find the employee is Senior or junior

In [17]:
```
uncleaned_df['Job_title'].value_counts()
```

Out[17]:
```
Data Scientist                                    326
Data Engineer                                      26
Senior Data Scientist                              19
Machine Learning Engineer                          14
Data Analyst                                       12
                                                 ...
Business Data Analyst                               1
Purification Scientist                              1
Data Science Instructor                             1
Data Engineer, Enterprise Analytics                 1
AI/ML - Machine Learning Scientist, Siri Understanding      1
Name: Job_title, Length: 168, dtype: int64
```

In [18]:
```
def employeetype(job_title):
    job_title=job_title.lower()
    emp=['sr','senior','lead','principal','vp','vice president','director']
    for i in emp:
        if i in job_title:
            return "Senior_emp"
        else:
            return "junior_emp"
    return NaN

uncleaned_df['Employee expriece']= uncleaned_df['Job_title'].apply(employeetype)
uncleaned_df.head()
```

Out[18]:

	index	Job_title	Salary_estimate_range	Job_Description	Job_post_rating	Company_Name	Job_Location	Company_Headquarters	Company_size	Comp
0	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 to 5000 employees	
1	1	Data Scientist	137K–171K (Glassdoor est.)	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 to 10000 employees	
2	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 to 5000 employees	
3	3	Data Scientist	137K–171K (Glassdoor est.)	JOB DESCRIPTION:\n\nDo you have a passion for ...	3.5	INFICON	Newton, MA	Bad Ragaz, Switzerland	501 to 1000 employees	
4	4	Data Scientist	137K–171K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions	New York, NY	New York, NY	51 to 200 employees	

5 rows × 28 columns

Export File into PC In CSV format after Cleaning & Transformation

In [19]:
```
uncleaned_df.to_csv('Data_science_Job2.csv',index=False)
```

2- Data Exploration & some Visualization

In [20]:
```
df = pd.read_csv('Cleaned_DS_Jobs.csv')
df.head()
```

Out[20]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Type of ownership	Industry	...	company_age	python	excel

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Type of ownership	Industry	...	avg_salary	job_state	same_state	company_age
0	Sr Data Scientist	137-171	Description\n\nThe Senior Data Scientist is re...	3.1	Healthfirst	New York, NY	New York, NY	1001 to 5000 employees	Nonprofit Organization	Insurance Carriers	...	27	0	0	
1	Data Scientist	137-171	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 to 10000 employees	Company - Public	Research & Development	...	52	0	0	
2	Data Scientist	137-171	Overview\n\n\nAnalysis Group is one of the lar...	3.8	Analysis Group	Boston, MA	Boston, MA	1001 to 5000 employees	Private Practice / Firm	Consulting	...	39	1	1	
3	Data Scientist	137-171	JOB DESCRIPTION:\n\nDo you have a passion for ...	3.5	INFICON	Newton, MA	Bad Ragaz, Switzerland	501 to 1000 employees	Company - Public	Electrical & Electronic Manufacturing	...	20	1	1	
4	Data Scientist	137-171	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions	New York, NY	New York, NY	51 to 200 employees	Company - Private	Advertising & Marketing	...	22	1	1	

5 rows × 27 columns

In [21]:
```
print(df.columns)
print(df.shape)
```

```
Index(['Job Title', 'Salary Estimate', 'Job Description', 'Rating',
       'Company Name', 'Location', 'Headquarters', 'Size', 'Type of ownership',
       'Industry', 'Sector', 'Revenue', 'min_salary', 'max_salary',
       'avg_salary', 'job_state', 'same_state', 'company_age', 'python',
       'excel', 'hadoop', 'spark', 'aws', 'tableau', 'big_data', 'job_simp',
       'seniority'],
      dtype='object')
(660, 27)
```

In [22]:
```
df['job_simp'].value_counts()
```

Out[22]:
```
data scientist    447
na                 68
analyst            55
data engineer      46
mle                34
manager             7
director            3
Name: job_simp, dtype: int64
```

In [23]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 660 entries, 0 to 659
Data columns (total 27 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Job Title          660 non-null    object
 1   Salary Estimate    660 non-null    object
 2   Job Description    660 non-null    object
 3   Rating             660 non-null    float64
 4   Company Name       660 non-null    object
 5   Location           660 non-null    object
 6   Headquarters       660 non-null    object
 7   Size               660 non-null    object
 8   Type of ownership  660 non-null    object
 9   Industry           660 non-null    object
 10  Sector             660 non-null    object
 11  Revenue            660 non-null    object
 12  min_salary         660 non-null    int64
 13  max_salary         660 non-null    int64
 14  avg_salary         660 non-null    int64
 15  job_state          660 non-null    object
 16  same_state         660 non-null    int64
 17  company_age        660 non-null    int64
 18  python             660 non-null    int64
 19  excel              660 non-null    int64
 20  hadoop             660 non-null    int64
 21  spark              660 non-null    int64
 22  aws                660 non-null    int64
 23  tableau            660 non-null    int64
 24  big_data           660 non-null    int64
 25  job_simp           660 non-null    object
 26  seniority          660 non-null    object
dtypes: float64(1), int64(12), object(14)
memory usage: 139.3+ KB
```

In [24]:
```
df.isnull().sum()
```

Out[24]:
```
Job Title            0
Salary Estimate      0
Job Description      0
Rating               0
Company Name         0
Location             0
Headquarters         0
Size                 0
Type of ownership    0
Industry             0
Sector               0
Revenue              0
min_salary           0
max_salary           0
avg_salary           0
job_state            0
same_state           0
company_age          0
python               0
excel                0
hadoop               0
spark                0
aws                  0
tableau              0
big_data             0
job_simp             0
seniority            0
dtype: int64
```

In [25]:
```
df.job_state.value_counts()
```

Out[25]:
```
CA    165
VA     89
MA     62
NY     52
MD     40
TI     30
```

```
DC     26
TX     17
WA     16
OH     14
MO     12
PA     12
US     11
CO     10
NJ     10
NC      9
GA      9
FL      8
TN      8
OK      6
WI      6
IN      5
MI      5
AZ      4
AL      4
MN      4
CT      4
UT      3
NE      3
IA      3
OR      2
SC      2
RI      2
NH      2
LA      1
MS      1
KS      1
DE      1
WV      1
Name: job_state, dtype: int64
```

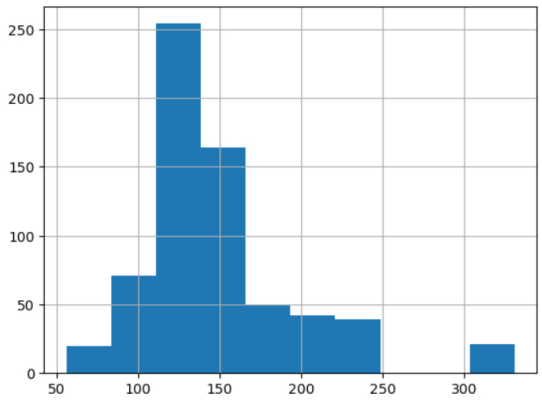
In [26]:
```
df.describe()
```

Out[26]:

	Rating	min_salary	max_salary	avg_salary	same_state	company_age	python	excel	hadoop	spark	aws	tableau	66
count	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	
mean	3.587424	99.296970	148.301515	123.612121	0.407576	29.736364	0.730303	0.440909	0.212121	0.281818	0.260606	0.184848	
std	1.183540	33.161485	48.264588	39.786698	0.491756	39.763033	0.444139	0.496873	0.409120	0.450226	0.439298	0.388469	
min	0.000000	31.000000	56.000000	43.000000	0.000000	-1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	3.300000	79.000000	119.000000	103.000000	0.000000	5.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	3.800000	91.000000	133.000000	114.000000	0.000000	16.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	4.300000	122.000000	165.000000	136.000000	1.000000	37.250000	1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	
max	5.000000	212.000000	331.000000	271.000000	1.000000	239.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

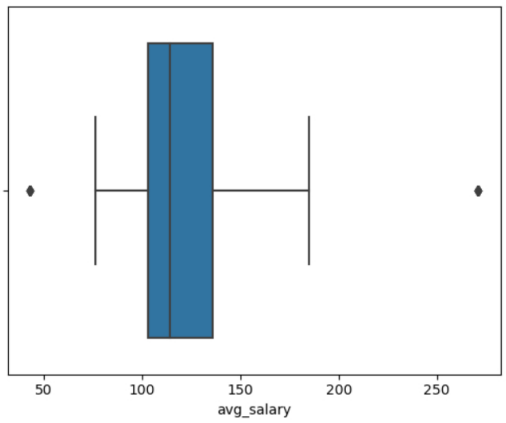
In [27]:
```
df.max_salary.hist()
```

Out[27]: <Axes: >



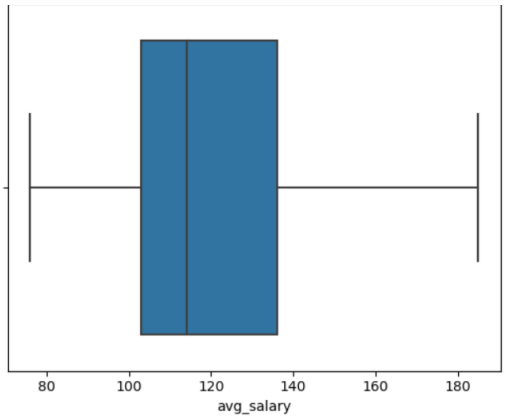
In [28]:
```
sns.boxplot(x=df["avg_salary"])
```

Out[28]: <Axes: xlabel='avg_salary'>



In [29]:
```
df = df[(df.avg_salary>50)&(df.avg_salary<200)]
sns.boxplot(x=df["avg_salary"])
```

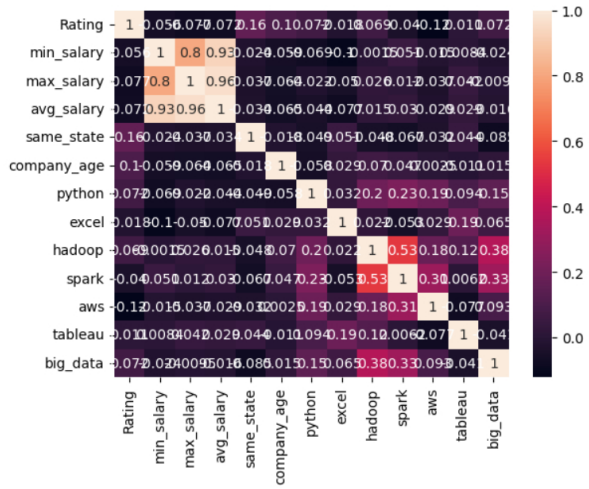
Out[29]: <Axes: xlabel='avg_salary'>



In [30]:
```
sns.heatmap(df.corr(),annot=True)
```

```
C:\Users\BAHGAT\AppData\Local\Temp\ipykernel_27140\4277794465.py:1: FutureWarning: The default value of numeric_only in DataFra
me.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric
_only to silence this warning.
  sns.heatmap(df.corr(),annot=True)
```

Out[30]: <Axes: >



In [31]:
```
pd.pivot_table(df, index='job_simp', values='avg_salary')
```

Out[31]:

job_simp	avg_salary
analyst	118.264151
data engineer	115.355556
data scientist	122.791866
director	127.000000
manager	115.666667
mle	115.625000
na	120.516129

In [32]:
```
pd.pivot_table(df, index=['job_simp','seniority'], values='avg_salary').sort_values('avg_salary', ascending =False)
```

Out[32]:

job_simp	seniority	avg_salary
director	senior	130.500000
na	senior	129.000000
mle	senior	126.666667
data scientist	na	123.746631
analyst	senior	121.125000
director	na	120.000000
manager	senior	120.000000
na	na	119.436364
analyst	na	118.166667
data engineer	na	116.051282
data scientist	senior	115.255319
manager	na	114.800000
mle	na	111.304348
data engineer	senior	110.833333
analyst	jr	76.000000

In [33]:
```
pd.pivot_table(df, index='job_state', values='avg_salary').sort_values('avg_salary', ascending =False)
```

Out[33]:

job_state	avg_salary
WI	144.333333

AZ	140.750000
IA	140.666667
MI	137.750000
NC	134.750000
MS	133.000000
NH	132.000000
OR	131.500000
AL	131.000000
RI	130.500000
NY	129.791667
TX	127.500000
PA	125.583333
VA	125.152941
IL	123.379310
CA	121.425000
US	120.500000
NJ	120.400000
DC	118.857143
FL	118.250000
MA	118.067797
GA	117.750000
OH	115.666667
WA	115.071429
MO	114.750000
WV	114.000000
IN	113.200000
MD	111.594595
MN	110.666667
CO	110.111111
NE	107.666667
OK	105.250000
UT	104.333333
LA	103.000000
KS	103.000000
TN	101.875000
CT	98.000000
SC	95.500000

In [34]:
```
pd.pivot_table(df, index=['job_state', 'job_simp'], values='avg_salary', aggfunc='count').sort_values('job_state', ascending =Fal
```

Out[34]:

		avg_salary
job_state	job_simp	
US	data scientist	7
	data engineer	1
WV	data scientist	1
WI	na	1
	data scientist	4
...	...	...
AZ	data scientist	2
	data engineer	1
AL	na	2
	data scientist	1
	data engineer	1

106 rows × 1 columns

In [35]:
```
pd.pivot_table(df[df.job_simp=='data scientist'], index='job_state', values='avg_salary').sort_values('avg_salary', ascending =Fa
```

Out[35]:

	avg_salary
job_state	
NH	161.000000
MI	150.666667
AZ	147.000000
WI	144.000000
PA	136.600000
NC	136.000000
NY	131.342105
TX	126.181818
FL	125.500000
CA	125.061947
VA	125.032258
NJ	124.250000
US	123.000000
IL	122.631579
OH	119.750000
MA	118.333333
DC	118.150000
GA	117.750000

WA	116.000000
UT	115.000000
MD	114.000000
WV	114.000000
MO	112.888889
NE	110.000000
CO	109.500000
AL	107.000000
IN	107.000000
MN	106.000000
KS	103.000000
TN	101.800000
OR	99.000000
CT	94.500000

In [36]:
```
df.head()
```

Out[36]:

	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Type of ownership	Industry	...	company_age	python	excel
0	Sr Data Scientist	137-171	Description\n\nThe Senior Data Scientist is re...	3.1	Healthfirst	New York, NY	New York, NY	1001 to 5000 employees	Nonprofit Organization	Insurance Carriers	...	27	0	0
1	Data Scientist	137-171	Secure our Nation, Ignite your Future\n\nJoin ...	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 to 10000 employees	Company - Public	Research & Development	...	52	0	0
2	Data Scientist	137-171	Overview\n\n\nAnalysis Group is one of the lar...	3.8	Analysis Group	Boston, MA	Boston, MA	1001 to 5000 employees	Private Practice / Firm	Consulting	...	39	1	1
3	Data Scientist	137-171	JOB DESCRIPTION:\n\nDo you have a passion for ...	3.5	INFICON	Newton, MA	Bad Ragaz, Switzerland	501 to 1000 employees	Company - Public	Electrical & Electronic Manufacturing	...	20	1	1
4	Data Scientist	137-171	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions	New York, NY	New York, NY	51 to 200 employees	Company - Private	Advertising & Marketing	...	22	1	1

5 rows × 27 columns