Introduction

Welcome to this EDA project exercise

Data description

The dataset is a CSV. file scrapped from the Glassdoor website of every job offer for the *Data scientist* job.

It is composed of one .CSV file containing 672 rows and 14 columns:

Column name	Description
Job Title	Job title
Salary Estimate	Estimate salary by Glassdoor in K. US dollars
Job Description	Description of the job offer
Rating	Rating of the company
Company Name	Name of the company
Location	Language of the audiobook
Headquarters	Rating and number of ratings of the audiobook
Size	Size of the company in number of employees
Founded	Year where the company has been founded
Type of ownership	Type of ownership of the company
Industry	Industry of the company
Sector	Sector of the company
Revenue	Revenue of the company in US dollars
Competitors	Direct competitors if existing

Steps to perform

The **'Salary Estimate'** column can't be analyzed as it is. Therefore split the maximum and minimum salary.

Regarding the **'Headquarters'** and **'Location'** columns we want to have the city and state in a separate columns.

As in data set company dataset is having **name** and **rating** together seperate it using seperator

Replace *'Unknown / Non-Applicable'* into *'N/A'* for more readibility:

Plot graphs

Plot bar chart for **Number of job offers per Seniority** using columns no_of_job_offers and job_seniority.

Plot bar chart for **Number of job offers per Seniority and Sector** using columns no_of_job_offers and job_seniority, Sector.

Plot bar chart for **Average estimate salary per Sector** using column Average estimate salary, sector.

Plot Scatter plot for **Correlation between the average salary and the company rating** use columns Average estimate salary, Rating, Seniority.

Plot histplot **Distribution of the average estimate salary per type of ownership** using Avg_salary_estimate, Type of ownership

Plot boxplot for **Distribution of the average salary in each company size** using Size, Avg_salary_estimate,

Plot countplot for **Number of job offers per Company size** using Size, job

Provide ths Summary

On analysis

In [1]:

```
#Importing libararies
import pandas as pd
import numpy as np
import datetime as dt
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
```

In [2]:

```
#Importing the data
Dataset = pd.read_csv(r"C:\Users\sanme\Downloads\archive\Uncleaned_DS_jobs.csv" , index_col=False)
df = Dataset.copy()
df
```

Out[2]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry
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
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
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
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
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
...	...	...	...	...	...	...	...	...	...	...	...	...
667	667	Data Scientist	$105K–$167K (Glassdoor est.)	Summary\n\nWe're looking for a data scientist ...	3.6	TRANZACT\n3.6	Fort Lee, NJ	Fort Lee, NJ	1001 to 5000 employees	1989	Company - Private	Advertising & Marketing
668	668	Data Scientist	$105K–$167K (Glassdoor est.)	Job Description\nBecome a thought leader withi...	-1.0	JKGT	San Francisco, CA	-1	-1	-1	-1	-1
669	669	Data Scientist	$105K–$167K (Glassdoor est.)	Join a thriving company that is changing the w...	-1.0	AccessHope	Irwindale, CA	-1	-1	-1	-1	-1
670	670	Data Scientist	$105K–$167K (Glassdoor est.)	100 Remote Opportunity As an AINLP Data Scient...	5.0	ChaTeck Incorporated\n5.0	San Francisco, CA	Santa Clara, CA	1 to 50 employees	-1	Company - Private	Advertising & Marketing
671	671	Data Scientist	$105K–$167K (Glassdoor est.)	Description\n\nThe Data Scientist will be part...	2.7	1-800-Flowers\n2.7	New York, NY	Carle Place, NY	1001 to 5000 employees	1976	Company - Public	Wholesale

672 rows × 15 columns

In [3]:

```
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 [4]:

```
df.columns
```

Out[4]:

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

Data Preprocessing

Selecting the columns that to keep in analysis:-

1) The 'index' column is a duplicate so it has to drop.

2) The 'Job Description' won't be accurate neither as each row contains a high amount of text therefore impossible to convert in numeric values.

3) The 'Competitors' column contains mainly '-1' values so we won't be able to analyze it.

In [5]:

```
df.columns
```

Out[5]:

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

In [6]:

```
df.drop(['Competitors','Job Description','index'], axis=1, inplace=True)
```

Now we want to remove every rows containing the '-1' values in the 'Headquarters', 'Founded' and 'Industry' columns:-

In [7]:

```
df_new=df.drop(df[(df.Headquarters == '-1') | (df.Founded == -1) | (df.Industry == -1) | (df.Sector == '-1') | (df.Rating == -1)].ind
```

Then we remove the duplicates

In [8]:

```
df_new.dropna(inplace=True)
df_new.drop_duplicates(inplace=True)
```

Steps to Perform

1) Salary Estimate

We can notice undesired characters at the end of each value combined with the rating of the company in the 'Company Name'.

Similarly, we don't need the information '(Glassdor est.)' in the 'Salary Estimate' column.

In [9]:

```
df_new['Company Name']=df_new['Company Name'].str.split('\n').str[0]
df_new['Salary Estimate']=df_new['Salary Estimate'].str.split('(').str[0]
```

In [10]:

```
df_new
```

Out[10]:

	Job Title	Salary Estimate	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue
0	Sr Data Scientist	$137K–171K	3.1	Healthfirst	New York, NY	New York, NY	1001 to 5000 employees	1993	Nonprofit Organization	Insurance Carriers	Insurance	Unknown / Non-Applicable
1	Data Scientist	$137K–171K	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 to 10000 employees	1968	Company - Public	Research & Development	Business Services	$1 to $2 billion (USD)
2	Data Scientist	$137K–171K	3.8	Analysis Group	Boston, MA	Boston, MA	1001 to 5000 employees	1981	Private Practice / Firm	Consulting	Business Services	$100 to $500 million (USD)
3	Data Scientist	$137K–171K	3.5	INFICON	Newton, MA	Bad Ragaz, Switzerland	501 to 1000 employees	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturing	$100 to $500 million (USD)
4	Data Scientist	$137K–171K	2.9	Affinity Solutions	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable
...	...	...	...	...	...	...	...	...	...	...	...	...
663	Data Scientist	$105K–167K	4.1	A-Line Staffing Solutions	Durham, NC	Utica, MI	501 to 1000 employees	2004	Company - Private	Staffing & Outsourcing	Business Services	Unknown / Non-Applicable
665	Data Scientist	$105K–167K	3.8	Criterion Systems, Inc.	Vienna, VA	Vienna, VA	201 to 500 employees	2005	Company - Private	IT Services	Information Technology	$50 to $100 million (USD)
666	Data Scientist	$105K–167K	4.0	Foundation Medicine	Boston, MA	Cambridge, MA	1001 to 5000 employees	2010	Company - Public	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals	$100 to $500 million (USD)
667	Data Scientist	$105K–167K	3.6	TRANZACT	Fort Lee, NJ	Fort Lee, NJ	1001 to 5000 employees	1989	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable
671	Data Scientist	$105K–167K	2.7	1-800-Flowers	New York, NY	Carle Place, NY	1001 to 5000 employees	1976	Company - Public	Wholesale	Business Services	$1 to $2 billion (USD)

548 rows × 12 columns

The 'Salary Estimate' column can't be analyzed as it is.

Therefore we can split the maximum and minimum salary using the str.split method.

Also we can do this for 'Headquarters' column as well to have the city and the state separated:

In [11]:

```
df_new2=df_new
df_new2[['Min_salary_estimate','Max_salary_estimate']] = df_new2['Salary Estimate'].str.split('-',expand=True)
```

In [12]:

```
df_new2.drop(['Salary Estimate'],axis=1,inplace=True)
df_new2
```

Out[12]:

	Job Title	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Min_salary_estim
0	Sr Data Scientist	3.1	Healthfirst	New York, NY	New York, NY	1001 to 5000 employees	1993	Nonprofit Organization	Insurance Carriers	Insurance	Unknown / Non-Applicable	$13
1	Data Scientist	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 to 10000 employees	1968	Company - Public	Research & Development	Business Services	$1 to 2 billion (USD)	$13
2	Data Scientist	3.8	Analysis Group	Boston, MA	Boston, MA	1001 to 5000 employees	1981	Private Practice / Firm	Consulting	Business Services	$100 to 500 million (USD)	$13
3	Data Scientist	3.5	INFICON	Newton, MA	Bad Ragaz, Switzerland	501 to 1000 employees	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturing	$100 to 500 million (USD)	$13
4	Data Scientist	2.9	Affinity Solutions	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	$13
...	...	...	...	...	...	...	...	...	...	...	...	
663	Data Scientist	4.1	A-Line Staffing Solutions	Durham, NC	Utica, MI	501 to 1000 employees	2004	Company - Private	Staffing & Outsourcing	Business Services	Unknown / Non-Applicable	$10
665	Data Scientist	3.8	Criterion Systems, Inc.	Vienna, VA	Vienna, VA	201 to 500 employees	2005	Company - Private	IT Services	Information Technology	$50 to 100 million (USD)	$10
666	Data Scientist	4.0	Foundation Medicine	Boston, MA	Cambridge, MA	1001 to 5000 employees	2010	Company - Public	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals	$100 to 500 million (USD)	$10
667	Data Scientist	3.6	TRANZACT	Fort Lee, NJ	Fort Lee, NJ	1001 to 5000 employees	1989	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	$10
671	Data Scientist	2.7	1-800-Flowers	New York, NY	Carle Place, NY	1001 to 5000 employees	1976	Company - Public	Wholesale	Business Services	$1 to 2 billion (USD)	$10

548 rows × 13 columns

Both columns still contain undesired string values. We can use the str.replace method to get rid of those:

In [13]:

```
df_new2['Min_salary_estimate'] = df_new2['Min_salary_estimate'].str.replace(r'\D',"")
df_new2['Max_salary_estimate'] = df_new2['Max_salary_estimate'].str.replace(r'\D',"")
```

```
C:\Users\sanme\AppData\Local\Temp\ipykernel_3532\476796256.py:1: FutureWarning: The default value of regex will change from
True to False in a future version.
  df_new2['Min_salary_estimate'] = df_new2['Min_salary_estimate'].str.replace(r'\D',"")
C:\Users\sanme\AppData\Local\Temp\ipykernel_3532\476796256.py:2: FutureWarning: The default value of regex will change from
True to False in a future version.
  df_new2['Max_salary_estimate'] = df_new2['Max_salary_estimate'].str.replace(r'\D',"")
```

And convert it to numeric values

In [14]:

```
df_new2['Min_salary_estimate']=df_new2['Min_salary_estimate'].astype(np.int64)
df_new2['Max_salary_estimate']=df_new2['Max_salary_estimate'].astype(np.int64)
```

In [15]:

```
df_new2['Min_salary_estimate'] = df_new2['Min_salary_estimate'].multiply(1000)
df_new2['Max_salary_estimate'] = df_new2['Max_salary_estimate'].multiply(1000)
df_new2
```

Out[15]:

	Job Title	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Min_salary_estima
0	Sr Data Scientist	3.1	Healthfirst	New York, NY	New York, NY	1001 to 5000 employees	1993	Nonprofit Organization	Insurance Carriers	Insurance	Unknown / Non-Applicable	1370
1	Data Scientist	4.2	ManTech	Chantilly, VA	Herndon, VA	5001 to 10000 employees	1968	Company - Public	Research & Development	Business Services	$1 to 2$ billion (USD)	1370
2	Data Scientist	3.8	Analysis Group	Boston, MA	Boston, MA	1001 to 5000 employees	1981	Private Practice / Firm	Consulting	Business Services	$100 to 500$ million (USD)	1370
3	Data Scientist	3.5	INFICON	Newton, MA	Bad Ragaz, Switzerland	501 to 1000 employees	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturing	$100 to 500$ million (USD)	1370
4	Data Scientist	2.9	Affinity Solutions	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	1370
...	...	...	...	...	...	...	...	...	...	...	...	
663	Data Scientist	4.1	A-Line Staffing Solutions	Durham, NC	Utica, MI	501 to 1000 employees	2004	Company - Private	Staffing & Outsourcing	Business Services	Unknown / Non-Applicable	1050
665	Data Scientist	3.8	Criterion Systems, Inc.	Vienna, VA	Vienna, VA	201 to 500 employees	2005	Company - Private	IT Services	Information Technology	$50 to 100$ million (USD)	1050
666	Data Scientist	4.0	Foundation Medicine	Boston, MA	Cambridge, MA	1001 to 5000 employees	2010	Company - Public	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals	$100 to 500$ million (USD)	1050
667	Data Scientist	3.6	TRANZACT	Fort Lee, NJ	Fort Lee, NJ	1001 to 5000 employees	1989	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	1050
671	Data Scientist	2.7	1-800-Flowers	New York, NY	Carle Place, NY	1001 to 5000 employees	1976	Company - Public	Wholesale	Business Services	$1 to 2$ billion (USD)	1050

548 rows × 13 columns

2) Headquarters and Location

Regarding the 'Headquarters' and 'Location' columns we want to have the city and state in a separate location:

In [16]:

```
df_new2[['HQ_city','HQ_state']]=df_new2['Headquarters'].str.split(',',expand=True)
df_new2['Job_city'] = df_new2.Location.str.split(',', expand = True)[0]
df_new2['Job_state'] = df_new2.Location.str.split(',', expand = True)[1]
df_new2.drop(['Headquarters','Location'], axis=1, inplace=True)
df_new2
```

Out[16]:

	Job Title	Rating	Company Name	Size	Founded	Type of ownership	Industry	Sector	Revenue	Min_salary_estimate	Max_salary_estimate
0	Sr Data Scientist	3.1	Healthfirst	1001 to 5000 employees	1993	Nonprofit Organization	Insurance Carriers	Insurance	Unknown / Non-Applicable	137000	171000
1	Data Scientist	4.2	ManTech	5001 to 10000 employees	1968	Company - Public	Research & Development	Business Services	$1 to 2 billion (USD)	137000	171000
2	Data Scientist	3.8	Analysis Group	1001 to 5000 employees	1981	Private Practice / Firm	Consulting	Business Services	$100 to 500 million (USD)	137000	171000
3	Data Scientist	3.5	INFICON	501 to 1000 employees	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturing	$100 to 500 million (USD)	137000	171000
4	Data Scientist	2.9	Affinity Solutions	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	137000	171000
...	...	...	...	...	...	...	...	...	...	...	...
663	Data Scientist	4.1	A-Line Staffing Solutions	501 to 1000 employees	2004	Company - Private	Staffing & Outsourcing	Business Services	Unknown / Non-Applicable	105000	167000
665	Data Scientist	3.8	Criterion Systems, Inc.	201 to 500 employees	2005	Company - Private	IT Services	Information Technology	$50 to 100 million (USD)	105000	167000
666	Data Scientist	4.0	Foundation Medicine	1001 to 5000 employees	2010	Company - Public	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals	$100 to 500 million (USD)	105000	167000
667	Data Scientist	3.6	TRANZACT	1001 to 5000 employees	1989	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	105000	167000
671	Data Scientist	2.7	1-800-Flowers	1001 to 5000 employees	1976	Company - Public	Wholesale	Business Services	$1 to 2 billion (USD)	105000	167000

548 rows × 15 columns

3) Size

In [17]:

```
df_new2['Size']=df_new2['Size'].str.replace(' employees','')
df_new2['Size']=df_new2['Size'].str.replace(' to ','-')
df_new2
```

Out[17]:

	Job Title	Rating	Company Name	Size	Founded	Type of ownership	Industry	Sector	Revenue	Min_salary_estimate	Max_salary_estimate	H
0	Sr Data Scientist	3.1	Healthfirst	1001-5000	1993	Nonprofit Organization	Insurance Carriers	Insurance	Unknown / Non-Applicable	137000	171000	N
1	Data Scientist	4.2	ManTech	5001-10000	1968	Company - Public	Research & Development	Business Services	1to2 billion (USD)	137000	171000	H
2	Data Scientist	3.8	Analysis Group	1001-5000	1981	Private Practice / Firm	Consulting	Business Services	100to500 million (USD)	137000	171000	
3	Data Scientist	3.5	INFICON	501-1000	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturing	100to500 million (USD)	137000	171000	
4	Data Scientist	2.9	Affinity Solutions	51-200	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	137000	171000	N
...	...	...	...	...	...	...	...	...	...	...	...	
663	Data Scientist	4.1	A-Line Staffing Solutions	501-1000	2004	Company - Private	Staffing & Outsourcing	Business Services	Unknown / Non-Applicable	105000	167000	
665	Data Scientist	3.8	Criterion Systems, Inc.	201-500	2005	Company - Private	IT Services	Information Technology	50to100 million (USD)	105000	167000	
666	Data Scientist	4.0	Foundation Medicine	1001-5000	2010	Company - Public	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals	100to500 million (USD)	105000	167000	Ca
667	Data Scientist	3.6	TRANZACT	1001-5000	1989	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	105000	167000	F
671	Data Scientist	2.7	1-800-Flowers	1001-5000	1976	Company - Public	Wholesale	Business Services	1to2 billion (USD)	105000	167000	

548 rows × 15 columns

Founded

In [18]:

```
today=dt.date.today()
df_new2['Company_age']=today.year-df_new2['Founded']
df_new2
```

Out[18]:

	Job Title	Rating	Company Name	Size	Founded	Type of ownership	Industry	Sector	Revenue	Min_salary_estimate	Max_salary_estimate	H
0	Sr Data Scientist	3.1	Healthfirst	1001-5000	1993	Nonprofit Organization	Insurance Carriers	Insurance	Unknown / Non-Applicable	137000	171000	N
1	Data Scientist	4.2	ManTech	5001-10000	1968	Company - Public	Research & Development	Business Services	$1 to $2 billion (USD)	137000	171000	H
2	Data Scientist	3.8	Analysis Group	1001-5000	1981	Private Practice / Firm	Consulting	Business Services	$100 to $500 million (USD)	137000	171000	
3	Data Scientist	3.5	INFICON	501-1000	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturing	$100 to $500 million (USD)	137000	171000	
4	Data Scientist	2.9	Affinity Solutions	51-200	1998	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	137000	171000	N
...	...	...	...	...	...	...	...	...	...	...	...	
663	Data Scientist	4.1	A-Line Staffing Solutions	501-1000	2004	Company - Private	Staffing & Outsourcing	Business Services	Unknown / Non-Applicable	105000	167000	
665	Data Scientist	3.8	Criterion Systems, Inc.	201-500	2005	Company - Private	IT Services	Information Technology	$50 to $100 million (USD)	105000	167000	
666	Data Scientist	4.0	Foundation Medicine	1001-5000	2010	Company - Public	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals	$100 to $500 million (USD)	105000	167000	Ca
667	Data Scientist	3.6	TRANZACT	1001-5000	1989	Company - Private	Advertising & Marketing	Business Services	Unknown / Non-Applicable	105000	167000	F
671	Data Scientist	2.7	1-800-Flowers	1001-5000	1976	Company - Public	Wholesale	Business Services	$1 to $2 billion (USD)	105000	167000	

548 rows × 16 columns

Revenue

The value 'Unknown / Non-Applicable' can be converted into 'N/A' for more readibility:

In [19]:

```
df_new2['Revenue'].value_counts()
```

Out[19]:

```
Unknown / Non-Applicable            169
$100 to $500 million (USD)           91
$10+ billion (USD)                   61
$10 to $25 million (USD)             40
$1 to $2 billion (USD)               36
$25 to $50 million (USD)             35
$2 to $5 billion (USD)               33
$50 to $100 million (USD)            30
$500 million to $1 billion (USD)     19
$1 to $5 million (USD)               18
$5 to $10 billion (USD)               7
$5 to $10 million (USD)               7
Less than $1 million (USD)            2
Name: Revenue, dtype: int64
```

In [20]:

```
df_new2['Revenue'] = df_new2['Revenue'].str.replace('Unknown / Non-Applicable',"N/A")
df_new2['Revenue'] = df_new2['Revenue'].str.replace('USD','')
df_new2['Revenue'] = df_new2['Revenue'].str.replace('[()]','')
df_new2['Revenue'] = df_new2['Revenue'].str.replace('to','-')
df_new2
```

```
C:\Users\sanme\AppData\Local\Temp\ipykernel_3532\2720762483.py:3: FutureWarning: The default value of regex will change fro
m True to False in a future version.
  df_new2['Revenue'] = df_new2['Revenue'].str.replace('[()]','')
```

Out[20]:

	Job Title	Rating	Company Name	Size	Founded	Type of ownership	Industry	Sector	Revenue	Min_salary_estimate	Max_salary_estimate	H
0	Sr Data Scientist	3.1	Healthfirst	1001-5000	1993	Nonprofit Organization	Insurance Carriers	Insurance	N/A	137000	171000	Ne
1	Data Scientist	4.2	ManTech	5001-10000	1968	Company - Public	Research & Development	Business Services	1−2 billion	137000	171000	H
2	Data Scientist	3.8	Analysis Group	1001-5000	1981	Private Practice / Firm	Consulting	Business Services	100−500 million	137000	171000	
3	Data Scientist	3.5	INFICON	501-1000	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturing	100−500 million	137000	171000	
4	Data Scientist	2.9	Affinity Solutions	51-200	1998	Company - Private	Advertising & Marketing	Business Services	N/A	137000	171000	Ne
...	...	...	...	...	...	...	...	...	...	...	...	
663	Data Scientist	4.1	A-Line Staffing Solutions	501-1000	2004	Company - Private	Staffing & Outsourcing	Business Services	N/A	105000	167000	
665	Data Scientist	3.8	Criterion Systems, Inc.	201-500	2005	Company - Private	IT Services	Information Technology	50−100 million	105000	167000	
666	Data Scientist	4.0	Foundation Medicine	1001-5000	2010	Company - Public	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals	100−500 million	105000	167000	Cam
667	Data Scientist	3.6	TRANZACT	1001-5000	1989	Company - Private	Advertising & Marketing	Business Services	N/A	105000	167000	F
671	Data Scientist	2.7	1-800-Flowers	1001-5000	1976	Company - Public	Wholesale	Business Services	1−2 billion	105000	167000	

548 rows × 16 columns

Now let's create a new column named 'Avg_salary_estimate' to get the average salary estimated:

In [21]:

```
df_new2['Avg_salary_estimate']=df_new2[['Min_salary_estimate','Max_salary_estimate']].mean(axis=1)
df_new2['Avg_salary_estimate']=df_new2['Avg_salary_estimate'].astype('int')
```

Now we notice the information 'Senior' in certain rows on the 'Job Title' column.

We can therefore create a new column 'Seniority' to identify which job offer requires to be Senior or Junior:

In [22]:

```
df_new2['Seniority'] = df_new2['Job Title'].str.contains('Senior',case=False)
df_new2['Seniority'] = df_new2['Seniority'].astype(str)
df_new2['Seniority'] = df_new2['Seniority'].str.replace('False','Junior')
df_new2['Seniority'] = df_new2['Seniority'].str.replace('True','Senior')
```

In [23]:

```
df_new2['Seniority'].value_counts()
```

Out[23]:

```
Junior    497
Senior     51
Name: Seniority, dtype: int64
```

In [24]:

```
df_finaldata=df_new2
df_finaldata
```

Out[24]:

	Job Title	Rating	Company Name	Size	Founded	Type of ownership	Industry	Sector	Revenue	Min_salary_estimate	Max_salary_estimate	H
0	Sr Data Scientist	3.1	Healthfirst	1001-5000	1993	Nonprofit Organization	Insurance Carriers	Insurance	N/A	137000	171000	Ne
1	Data Scientist	4.2	ManTech	5001-10000	1968	Company - Public	Research & Development	Business Services	1−2 billion	137000	171000	H
2	Data Scientist	3.8	Analysis Group	1001-5000	1981	Private Practice / Firm	Consulting	Business Services	100−500 million	137000	171000	
3	Data Scientist	3.5	INFICON	501-1000	2000	Company - Public	Electrical & Electronic Manufacturing	Manufacturing	100−500 million	137000	171000	
4	Data Scientist	2.9	Affinity Solutions	51-200	1998	Company - Private	Advertising & Marketing	Business Services	N/A	137000	171000	Ne
...	...	...	...	...	...	...	...	...	...	...	...	
663	Data Scientist	4.1	A-Line Staffing Solutions	501-1000	2004	Company - Private	Staffing & Outsourcing	Business Services	N/A	105000	167000	
665	Data Scientist	3.8	Criterion Systems, Inc.	201-500	2005	Company - Private	IT Services	Information Technology	50−100 million	105000	167000	
666	Data Scientist	4.0	Foundation Medicine	1001-5000	2010	Company - Public	Biotech & Pharmaceuticals	Biotech & Pharmaceuticals	100−500 million	105000	167000	Can
667	Data Scientist	3.6	TRANZACT	1001-5000	1989	Company - Private	Advertising & Marketing	Business Services	N/A	105000	167000	F
671	Data Scientist	2.7	1-800-Flowers	1001-5000	1976	Company - Public	Wholesale	Business Services	1−2 billion	105000	167000	

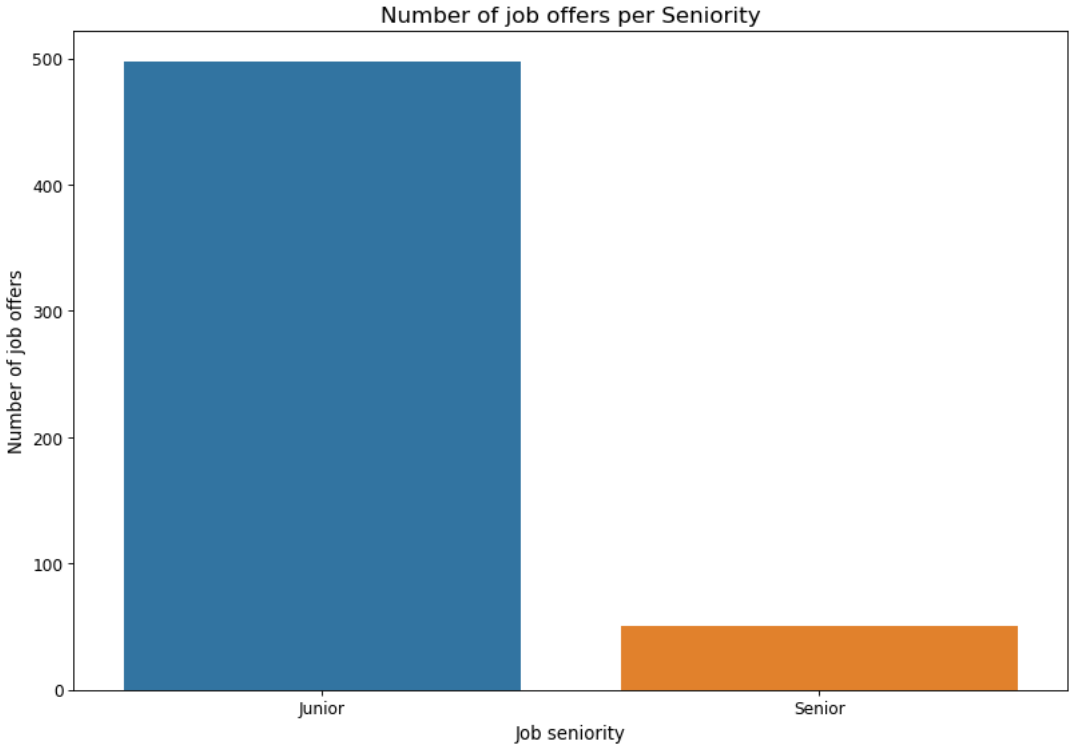
548 rows × 18 columns

Plotting Graphs

1) Number of job offers per Seniority

In [25]:

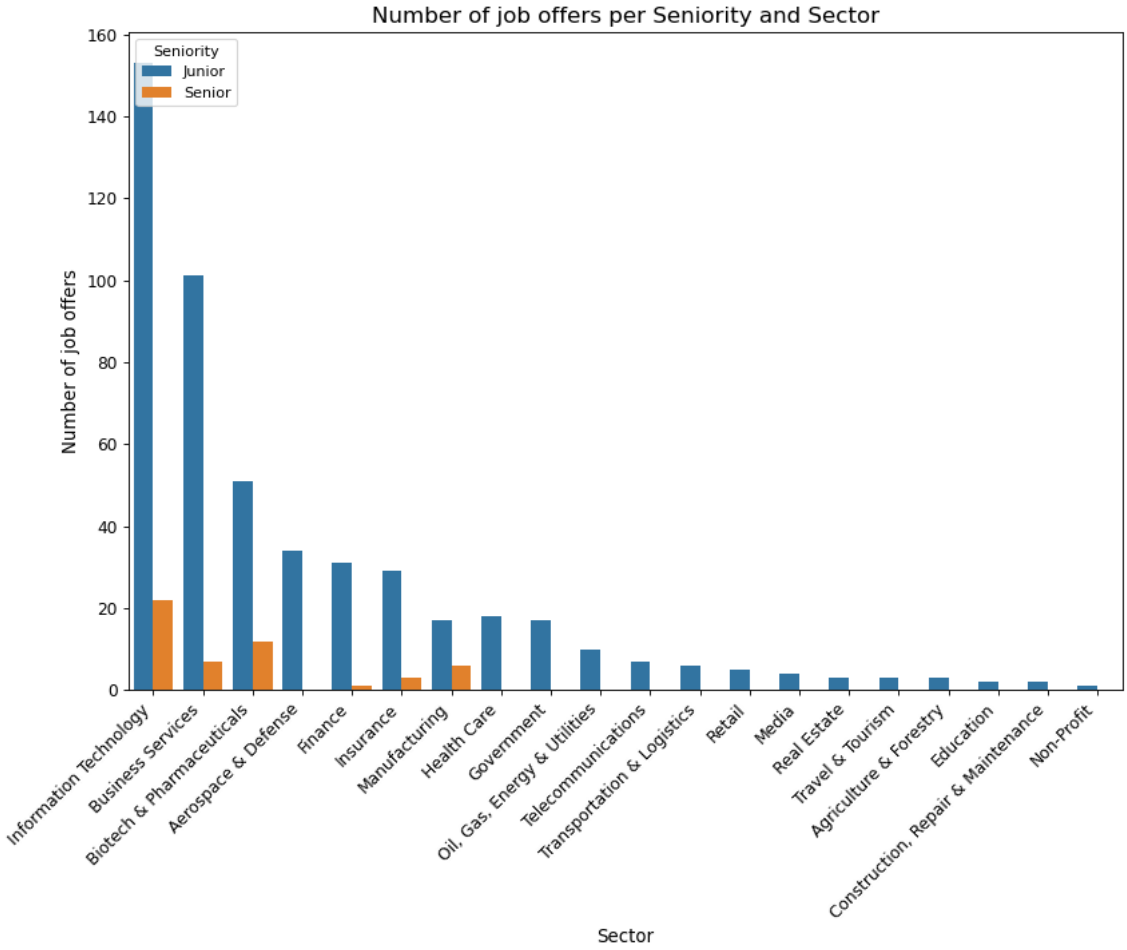
```
plt.figure(figsize=(12, 8), dpi=80)
sns.countplot(x='Seniority', data=df_finaldata)
plt.xlabel("Job seniority", fontsize=12)
plt.ylabel("Number of job offers", fontsize=12)
plt.title("Number of job offers per Seniority", fontsize=15)
plt.xticks(fontsize=11)
plt.yticks(fontsize=11)
plt.show()
```



2) Number of job offers per Seniority and Sector

In [26]:

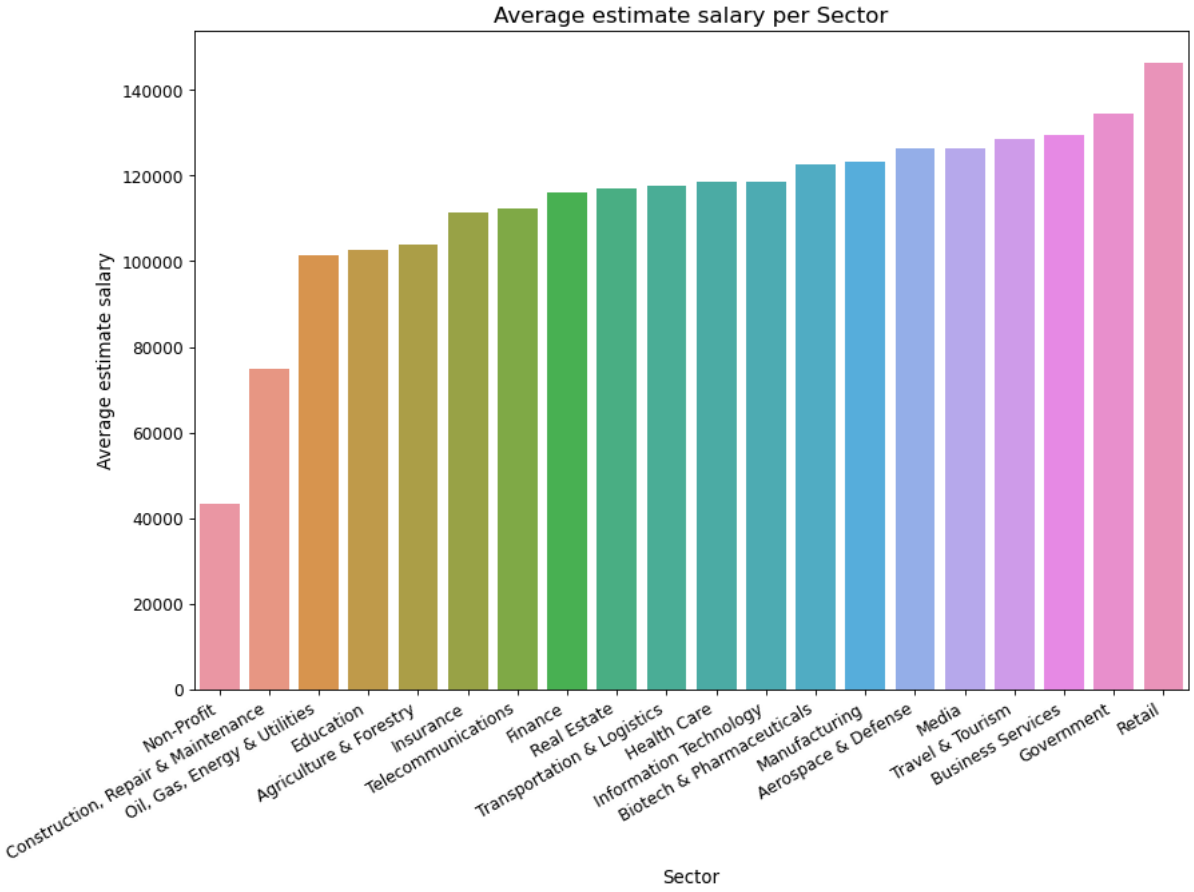
```
figure(figsize=(12, 8), dpi=80)
sns.countplot(data=df_finaldata, x=df_finaldata['Sector'],hue=df_finaldata['Seniority'], order=df_finaldata['Sector'].value_counts().
plt.xticks(rotation=45,ha='right', va='top', fontsize=11)
plt.yticks(fontsize=11)
plt.xlabel("Sector",fontsize=12)
plt.ylabel("Number of job offers",fontsize=12)
plt.title("Number of job offers per Seniority and Sector",fontsize=15)
plt.show()
```



3) Average estimate salary per Sector

In [27]:

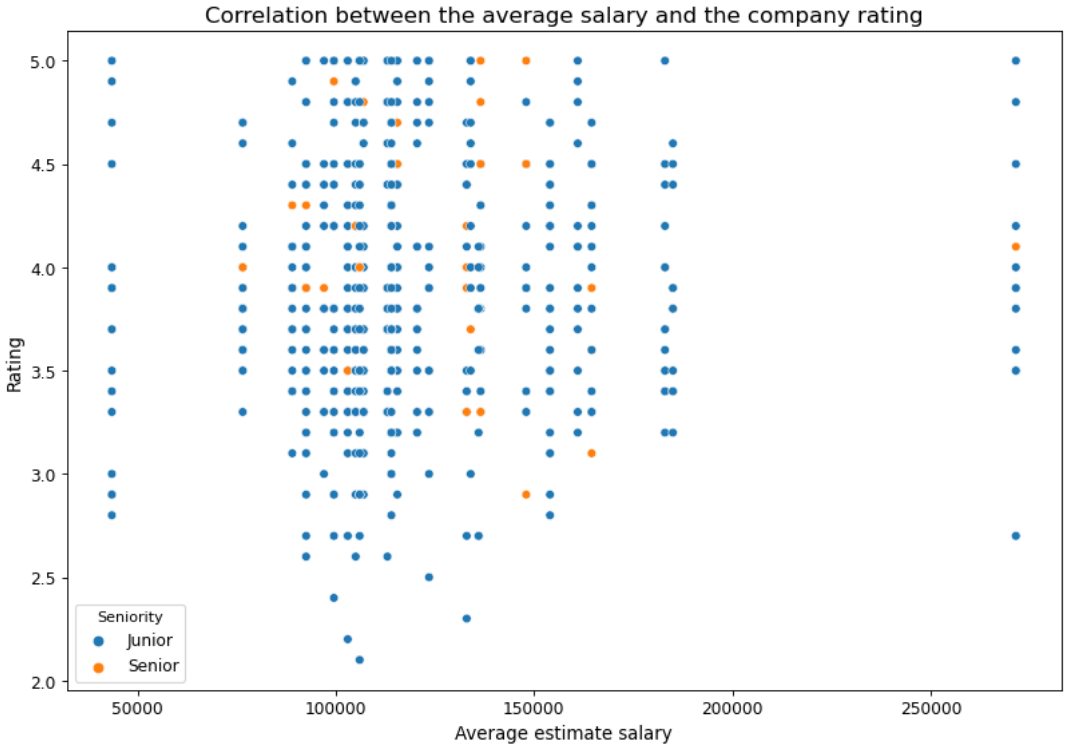
```
plt.figure(figsize=(12, 8), dpi=80)
sns.barplot(x='Sector', y='Avg_salary_estimate', data=df_finaldata, order=df_finaldata.groupby(["Sector"])['Avg_salary_estimate'].mea
plt.xlabel("Sector", fontsize=12)
plt.ylabel("Average estimate salary", fontsize=12)
plt.title("Average estimate salary per Sector", fontsize=15)
plt.xticks(rotation=30, ha='right', va='top', fontsize=11)
plt.yticks(fontsize=11)
plt.show()
```



4) Correlation between the average salary and the company rating

In [28]:

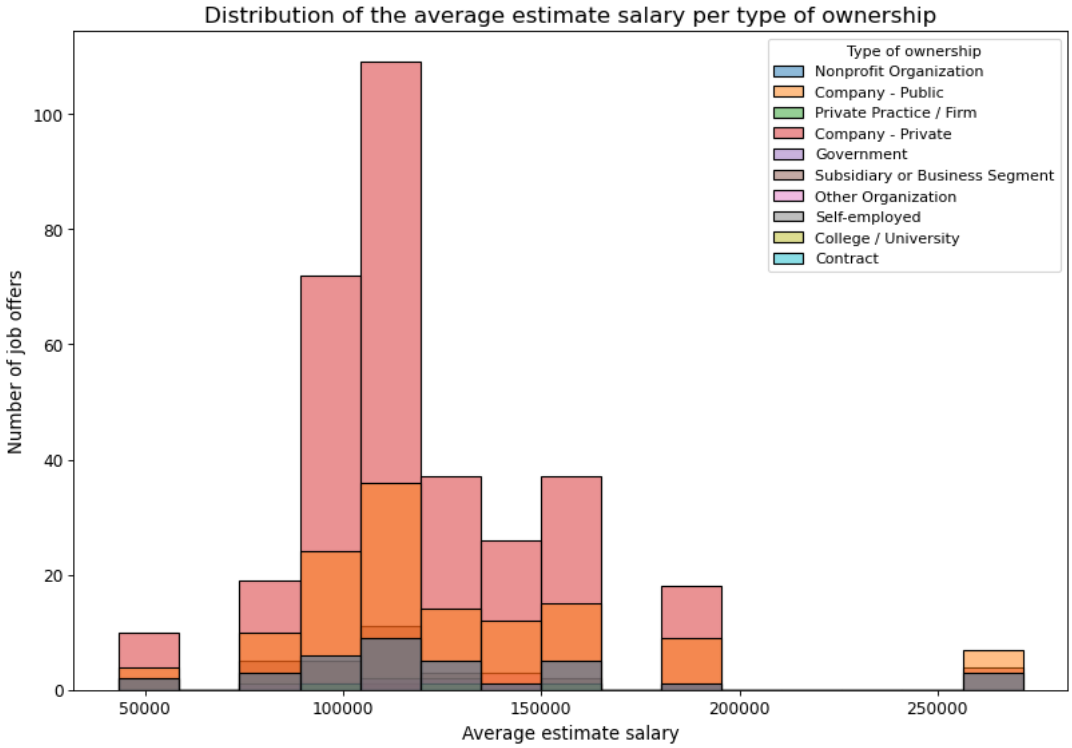
```
figure(figsize=(12, 8), dpi=80)
sns.scatterplot(data=df_finaldata, x='Avg_salary_estimate', y='Rating', hue='Seniority')
plt.xlabel("Average estimate salary", fontsize=12)
plt.ylabel("Rating", fontsize=12)
plt.title("Correlation between the average salary and the company rating", fontsize=15)
plt.xticks(fontsize=11)
plt.yticks(fontsize=11)
plt.legend(title='Seniority', fontsize=11)
plt.show()
```



5) Distribution of the average estimate salary per type of ownership

In [29]:

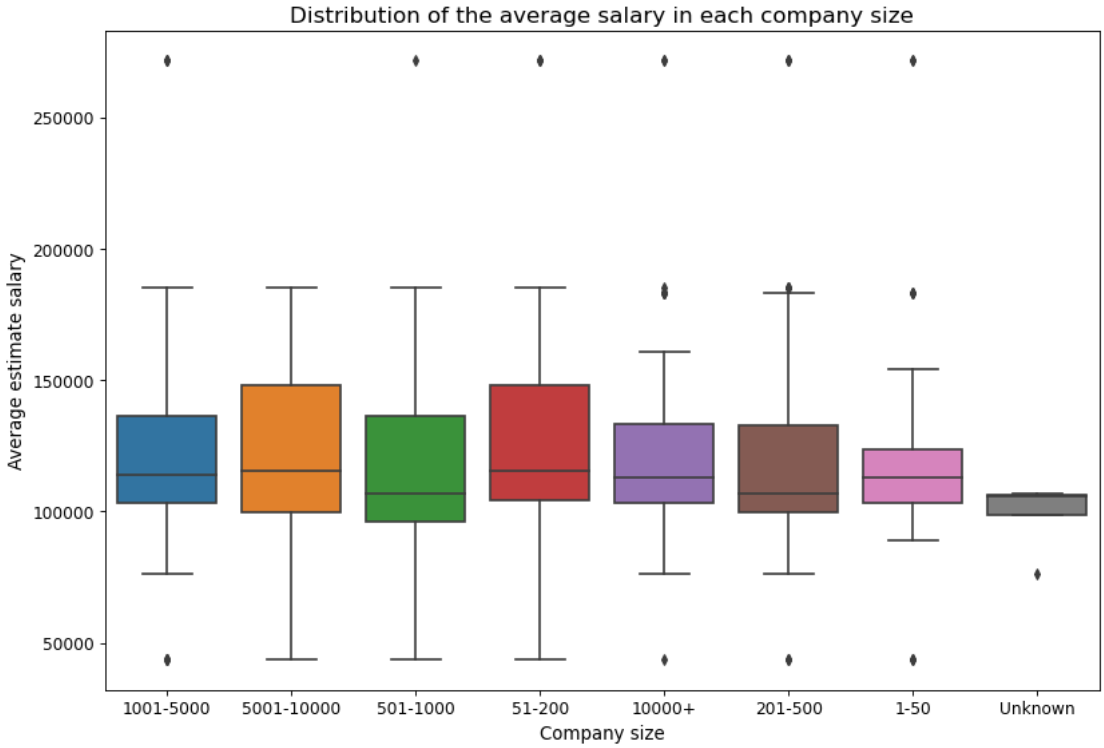
```
figure(figsize=(12, 8), dpi=80)
sns.histplot(df_finaldata,x=df_finaldata['Avg_salary_estimate'], bins=15, hue=df_finaldata['Type of ownership'])
plt.xlabel("Average estimate salary",fontsize=12)
plt.ylabel("Number of job offers",fontsize=12)
plt.title("Distribution of the average estimate salary per type of ownership",fontsize=15)
plt.xticks(fontsize=11)
plt.yticks(fontsize=11)
plt.show()
```



6) Distribution of the average salary in each company size

In [30]:

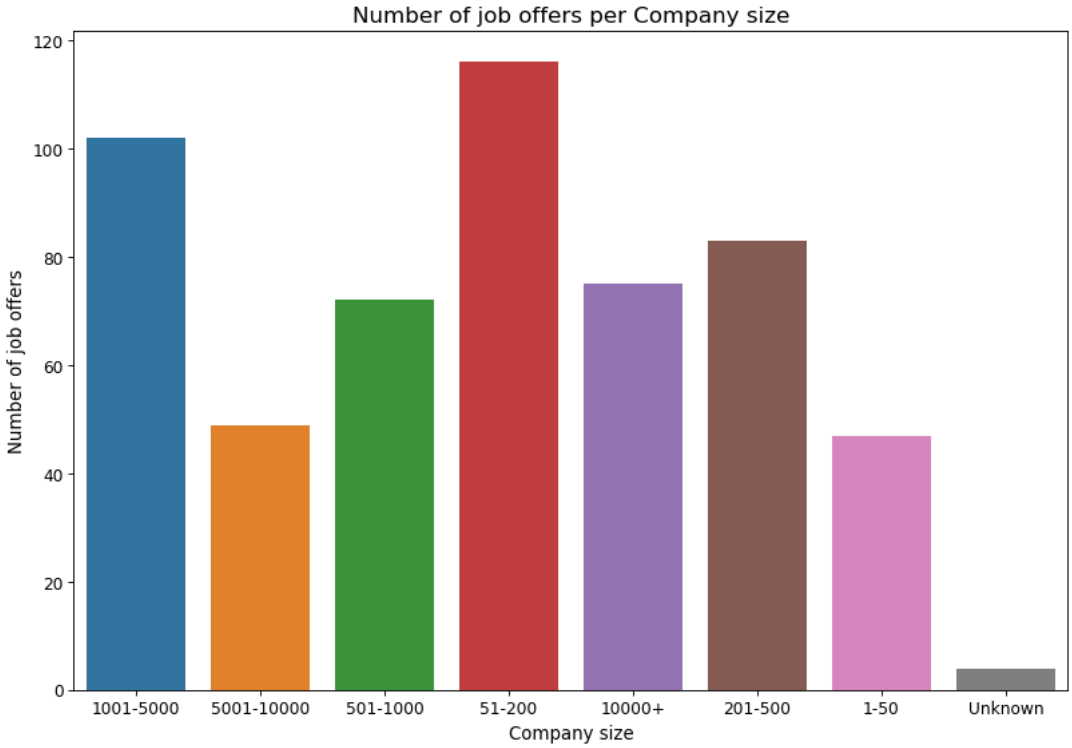
```
figure(figsize=(12, 8), dpi=80)
sns.boxplot(x=df_finaldata['Size'], y=df_finaldata['Avg_salary_estimate'])
plt.xlabel("Company size", fontsize=12)
plt.ylabel("Average estimate salary", fontsize=12)
plt.title("Distribution of the average salary in each company size", fontsize=15)
plt.xticks(fontsize=11)
plt.yticks(fontsize=11)
plt.show()
```



7) Number of job offers per Company size

In [31]:

```
plt.figure(figsize=(12, 8), dpi=80)
sns.countplot(x=df_finaldata['Size'])
plt.xlabel("Company size", fontsize=12)
plt.ylabel("Number of job offers", fontsize=12)
plt.title("Number of job offers per Company size", fontsize=15)
plt.xticks(fontsize=11)
plt.yticks(fontsize=11)
plt.show()
```



Report on Analysis

After performing the analysis and plotting the graphs, we have gained insights into the data and divided into two parts:

A) Jobs Offered:-

- The first two bar charts provided information on the number of job offers per seniority and sector.
- Hence, going through it we can see there are much more junior job offers than senior job offers, hence quite encouraging for youngest Data Scientists who are looking to get a job for the same.
- While in 2nd plot we can see that the more no. of jobs which recrute Data Scientist are in IT sector which is as expected.
- Finally the Senior Data Scientist those who are searching for jobs they can basically focus on these sectors : Information Technology, Biotech & Pharmaceuticals and Business Services as there are more opportunities.

B) Salary:-

- The next graphs which can be analysis are Average estimate Salary per sector, correlation between the average salary and company rating based on different job seniorities and distribution of the average estimate salary per type of ownership.
- Firstly we see that the Retail sector is the sector that offers the highest average entry level salary while, the Non-profit sector is the one with the lowest average salary.
- Sometimes the job seekers thought that if the Company rating is high then the avg salary for the post will also be great. So we need to analyse that if there is any corelation between the company's ratinng and the avg salary offered. Hence by seeing the scatter plot we can summarize that there is no corelation.
- Finally we can see that most of the job offers are rewarded with an average estimate salary between 90,000 and 1,15,000 US dollars, with a minimum at 50,000 US dollars and 2,60,000 US dollars.
- The boxplot depicted the distribution of average salaries in each company size.
- Lastly, the countplot displayed the number of job offers per company size based on different seniority levels.

These visualizations give a comprehensive understanding of the data and can help in making data-driven decisions related to Job Offers, Salaries, and Company Preferences in the Data Science job market.

In [35]:

```
df_finaldata.to_csv('Cleaned_DS_Jobs_Glassdoor.csv', index=False)
```