Data scientist salary

Libraries

```
In [61]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns
```

Load the dataset

```
In [4]: DS = pd.read_csv('Data Scientist Salary EDA.csv',encoding = 'latin1')
```

In [5]: DS

Out[5]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	...	tenso
0	0	Data Scientist	$53K-91K$ (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 - 1000	1973	...	
1	1	Healthcare Data Scientist	$63K-112K$ (Glassdoor est.)	What You Will Do:\n\nI. General Summary\n\nThe...	3.4	University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+	1984	...	
2	2	Data Scientist	$80K-90K$ (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 - 1000	2010	...	
3	3	Data Scientist	$56K-97K$ (Glassdoor est.)	*Organization and Job ID**\nJob ID: 310709\n\n...	3.8	PNNL\n3.8	Richland, WA	Richland, WA	1001 - 5000	1965	...	
4	4	Data Scientist	$86K-143K$ (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 - 200	1998	...	
...	...	...	...	...	...	...	...	...	...	...	...	
737	950	Sr Scientist, Immuno-Oncology - Oncology	$58K-111K$ (Glassdoor est.)	Site Name: USA - Massachusetts - Cambridge\nPo...	3.9	GSK\n3.9	Cambridge, MA	Brentford, United Kingdom	10000+	1830	...	
738	951	Senior Data Engineer	$72K-133K$ (Glassdoor est.)	THE CHALLENGE\nEventbrite has a world-class da...	4.4	Eventbrite\n4.4	Nashville, TN	San Francisco, CA	1001 - 5000	2006	...	
739	952	Project Scientist - Auton Lab, Robotics Institute	$56K-91K$ (Glassdoor est.)	The Auton Lab at Carnegie Mellon University is...	2.6	Software Engineering Institute\n2.6	Pittsburgh, PA	Pittsburgh, PA	501 - 1000	1984	...	
740	953	Data Science Manager	$95K-160K$ (Glassdoor est.)	Data Science ManagerResponsibilities:\n\nOvers...	3.2	Numeric, LLC\n3.2	Allentown, PA	Chadds Ford, PA	1 - 50	-1	...	

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	...	tens
741	955	Research Scientist âªª Security and Privacy	$61K - 126K$ (Glassdoor est.)	Returning Candidate? Log back in to the Career...	3.6	Riverside Research Institute\n3.6	Beavercreek, OH	Arlington, VA	501 - 1000	1967	...	

742 rows × 42 columns

Explore the data

In [6]: `DS.head()`

Out[6]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	...	tensor	hadoop	tablea
0	0	Data Scientist	$53K-91K$ (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 - 1000	1973	...	0	0
1	1	Healthcare Data Scientist	$63K-$ 112K (Glassdoor est.)	What You Will Do:\n\nI. General Summary\n\nThe...	3.4	University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+	1984	...	0	0
2	2	Data Scientist	$80K-90K$ (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 - 1000	2010	...	0	0
3	3	Data Scientist	$56K-97K$ (Glassdoor est.)	*Organization and Job ID**\nJob ID: 310709\n\n...	3.8	PNNL\n3.8	Richland, WA	Richland, WA	1001 - 5000	1965	...	0	0
4	4	Data Scientist	$86K-$ 143K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 - 200	1998	...	0	0

5 rows × 42 columns

In [7]: `DS.tail()`

Out[7]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	...	tensor
737	950	Sr Scientist, Immuno-Oncology - Oncology	$58K-111K$ (Glassdoor est.)	Site Name: USA - Massachusetts - Cambridge\nPo...	3.9	GSK\n3.9	Cambridge, MA	Brentford, United Kingdom	10000+	1830	...	0
738	951	Senior Data Engineer	$72K-133K$ (Glassdoor est.)	THE CHALLENGE\nEventbrite has a world-class da...	4.4	Eventbrite\n4.4	Nashville, TN	San Francisco, CA	1001 - 5000	2006	...	0
739	952	Project Scientist - Auton Lab, Robotics Institute	$56K-91K$ (Glassdoor est.)	The Auton Lab at Carnegie Mellon University is...	2.6	Software Engineering Institute\n2.6	Pittsburgh, PA	Pittsburgh, PA	501 - 1000	1984	...	0
740	953	Data Science Manager	$95K-160K$ (Glassdoor est.)	Data Science ManagerResponsibilities:\n\nOvers...	3.2	Numeric, LLC\n3.2	Allentown, PA	Chadds Ford, PA	1 - 50	-1	...	0
741	955	Research Scientist â□□ Security and Privacy	$61K-126K$ (Glassdoor est.)	Returning Candidate? Log back in to the Career...	3.6	Riverside Research Institute\n3.6	Beavercreek, OH	Arlington, VA	501 - 1000	1967	...	0

5 rows × 42 columns

In [8]: `DS.shape`

Out[8]: `(742, 42)`

In [9]: `DS.sample()`

Out[9]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	...	tensor	hadoop	tableau
219	267	Lead Data Scientist	$124K-$204K (Glassdoor est.)	Job Description\n\nSince 1851, MassMutuals com...	3.6	MassMutual\n3.6	Boston, MA	Springfield, MA	5001 - 10000	1851	...	0	1	0

1 rows × 42 columns

In [12]: `DS.columns`

Out[12]:
```
Index(['index', 'Job Title', 'Salary Estimate', 'Job Description', 'Rating',
       'Company Name', 'Location', 'Headquarters', 'Size', 'Founded',
       'Type of ownership', 'Industry', 'Sector', 'Revenue', 'Competitors',
       'Hourly', 'Employer provided', 'Lower Salary', 'Upper Salary',
       'Avg Salary(K)', 'company_txt', 'Job Location', 'Age', 'Python',
       'spark', 'aws', 'excel', 'sql', 'sas', 'keras', 'pytorch', 'scikit',
       'tensor', 'hadoop', 'tableau', 'bi', 'flink', 'mongo', 'google_an',
       'job_title_sim', 'seniority_by_title', 'Degree'],
      dtype='object')
```

In [13]: `DS.mean()`

C:\Users\abhim\AppData\Local\Temp\ipykernel_10116\258044430.py:1: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid columns or specify the value of numeric_only to silence this warning.
  DS.mean()

Out[13]:
```
index               469.129380
Rating                3.618868
Founded            1837.154987
Hourly                0.032345
Employer provided     0.022911
Lower Salary         74.754717
Upper Salary        128.214286
Avg Salary(K)       101.484501
Age                  47.524259
Python                0.528302
spark                 0.225067
aws                   0.237197
excel                 0.522911
sql                   0.512129
sas                   0.088949
keras                 0.039084
pytorch               0.052561
scikit                0.072776
tensor                0.097035
hadoop                0.167116
tableau               0.199461
bi                    0.075472
flink                 0.013477
mongo                 0.049865
google_an             0.018868
dtype: float64
```

In [14]: `DS.median()`

C:\Users\abhim\AppData\Local\Temp\ipykernel_10116\3307674007.py:1: FutureWarning: The default value of numeric_only in DataFrame.median is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid columns or specify the value of numeric_only to silence this warning.
  DS.median()

Out[14]:
```
index               472.5
Rating                3.7
Founded            1988.0
Hourly                0.0
Employer provided     0.0
Lower Salary         69.5
Upper Salary        124.0
Avg Salary(K)        97.5
Age                  25.0
Python                1.0
spark                 0.0
aws                   0.0
excel                 1.0
sql                   1.0
sas                   0.0
keras                 0.0
pytorch               0.0
scikit                0.0
tensor                0.0
hadoop                0.0
tableau               0.0
bi                    0.0
flink                 0.0
mongo                 0.0
google_an             0.0
dtype: float64
```

In [10]: `DS.describe()`

Out[10]:

	index	Rating	Founded	Hourly	Employer provided	Lower Salary	Upper Salary	Avg Salary(K)	Age	Python	...	keras
count	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	...	742.000000
mean	469.129380	3.618868	1837.154987	0.032345	0.022911	74.754717	128.214286	101.484501	47.524259	0.528302	...	0.039084
std	279.793117	0.801210	497.183763	0.177034	0.149721	30.945892	45.128650	37.482449	53.839080	0.499535	...	0.193925
min	0.000000	-1.000000	-1.000000	0.000000	0.000000	15.000000	16.000000	15.500000	-1.000000	0.000000	...	0.000000
25%	221.500000	3.300000	1939.000000	0.000000	0.000000	52.000000	96.000000	73.500000	12.000000	0.000000	...	0.000000
50%	472.500000	3.700000	1988.000000	0.000000	0.000000	69.500000	124.000000	97.500000	25.000000	1.000000	...	0.000000
75%	707.750000	4.000000	2007.000000	0.000000	0.000000	91.000000	155.000000	122.500000	60.000000	1.000000	...	0.000000
max	955.000000	5.000000	2019.000000	1.000000	1.000000	202.000000	306.000000	254.000000	277.000000	1.000000	...	1.000000

8 rows × 25 columns

In [11]: `DS.describe(include = 'object')`

Out[11]:

	Job Title	Salary Estimate	Job Description	Company Name	Location	Headquarters	Size	Type of ownership	Industry	Sector	Revenue
count	742	742	742	742	742	742	742	742	742	742	742
unique	264	416	463	343	200	198	8	9	60	25	13
top	Data Scientist	$49K-113K (Glassdoor est.)	Description\nMedical Laboratory Scientist - Te...	MassMutual\n3.6	New York, NY	New York, NY	1001 - 5000	Company - Private	Biotech & Pharmaceuticals	Information Technology	Unknown / Non-Applicable
freq	131	6	4	14	55	52	150	410	112	180	204

In [15]: DS.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 742 entries, 0 to 741
Data columns (total 42 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   index             742 non-null    int64
 1   Job Title         742 non-null    object
 2   Salary Estimate   742 non-null    object
 3   Job Description   742 non-null    object
 4   Rating            742 non-null    float64
 5   Company Name      742 non-null    object
 6   Location          742 non-null    object
 7   Headquarters      742 non-null    object
 8   Size              742 non-null    object
 9   Founded           742 non-null    int64
 10  Type of ownership 742 non-null    object
 11  Industry          742 non-null    object
 12  Sector            742 non-null    object
 13  Revenue           742 non-null    object
 14  Competitors       742 non-null    object
 15  Hourly            742 non-null    int64
 16  Employer provided 742 non-null    int64
 17  Lower Salary      742 non-null    int64
 18  Upper Salary      742 non-null    int64
 19  Avg Salary(K)     742 non-null    float64
 20  company_txt       742 non-null    object
 21  Job Location      742 non-null    object
 22  Age               742 non-null    int64
 23  Python            742 non-null    int64
 24  spark             742 non-null    int64
 25  aws               742 non-null    int64
 26  excel             742 non-null    int64
 27  sql               742 non-null    int64
 28  sas               742 non-null    int64
 29  keras             742 non-null    int64
 30  pytorch           742 non-null    int64
 31  scikit            742 non-null    int64
 32  tensor            742 non-null    int64
 33  hadoop            742 non-null    int64
 34  tableau           742 non-null    int64
 35  bi                742 non-null    int64
```

```
 36   flink                742 non-null     int64
 37   mongo                742 non-null     int64
 38   google_an            742 non-null     int64
 39   job_title_sim        742 non-null     object
 40   seniority_by_title   742 non-null     object
 41   Degree               742 non-null     object
dtypes: float64(2), int64(23), object(17)
memory usage: 243.6+ KB
```

In [16]:
```python
DS.isnull().sum()
```

```
Out[16]:  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
          Hourly                  0
          Employer provided       0
          Lower Salary            0
          Upper Salary            0
          Avg Salary(K)           0
          company_txt             0
          Job Location            0
          Age                     0
          Python                  0
          spark                   0
          aws                     0
          excel                   0
          sql                     0
          sas                     0
          keras                   0
          pytorch                 0
          scikit                  0
          tensor                  0
          hadoop                  0
          tableau                 0
          bi                      0
          flink                   0
          mongo                   0
          google_an               0
          job_title_sim           0
          seniority_by_title      0
```

```
Degree                   0
dtype: int64
```

Data cleaning

In [23]:
```python
Columns_deleted = ['Job Title', 'Salary Estimate', 'Job Description', 'Rating','Company Name', 'Headquarters', 'Founde
                   'Industry', 'Revenue', 'Competitors', 'Hourly', 'Employer provided', 'Age']
```

In [26]:
```python
DS.drop(columns = Columns_deleted,axis=1,inplace = True)
```

```
C:\Users\abhim\AppData\Local\Temp\ipykernel_10116\1920359739.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning
-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-
a-copy)
  DS.drop(columns = Columns_deleted,axis=1,inplace = True)
```

In [27]:
```python
DS = DS.drop_duplicates(['company_txt','Upper Salary'])
```

In [28]:
```python
change_name_column = {'job_title_sim': 'job_title'}
DS.rename(columns = change_name_column,inplace=True)
```

In [29]: `DS.head()`

Out[29]:

pe of rship	Sector	Lower Salary	Upper Salary	Avg Salary(K)	company_txt	Job Location	...	tensor	hadoop	tableau	bi	flink	mongo	google_an	job_title	seniority_by_
any - ivate	Aerospace & Defense	53	91	72.0	Tecolote Research	NM	...	0	0	1	1	0	0	0	data scientist	
Other ation	Health Care	63	112	87.5	University of Maryland Medical System	MD	...	0	0	0	0	0	0	0	data scientist	
any - ivate	Business Services	80	90	85.0	KnowBe4	FL	...	0	0	0	0	0	0	0	data scientist	
ment	Oil, Gas, Energy & Utilities	56	97	76.5	PNNL	WA	...	0	0	0	0	0	0	0	data scientist	
any - ivate	Business Services	86	143	114.5	Affinity Solutions	NY	...	0	0	0	0	0	0	0	data scientist	

In [30]: `DS.shape`

Out[30]: `(455, 29)`

valuable insights for analysis

In [31]: `DS[['Lower Salary', 'Upper Salary']].describe()`

Out[31]:

	Lower Salary	Upper Salary
count	455.000000	455.000000
mean	74.885714	128.274725
std	30.023781	44.146848
min	15.000000	16.000000
25%	54.000000	98.000000
50%	71.000000	124.000000
75%	90.500000	150.000000
max	202.000000	306.000000

In [32]: `DS['job_title'].unique()`

Out[32]:
```
array(['data scientist', 'other scientist', 'analyst', 'data engineer',
       'data analitics', 'na', 'data modeler',
       'Data scientist project manager', 'machine learning engineer',
       'director'], dtype=object)
```

In [34]:
```
job_titles = DS[['Lower Salary', 'Upper Salary', 'job_title']]
# Filtering job title
```

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

Out[36]:

	Lower Salary	Upper Salary	job_title
0	53	91	data scientist
1	63	112	data scientist
2	80	90	data scientist
3	56	97	data scientist
4	86	143	data scientist

In [37]: `job_titles['job_title'].value_counts()`

Out[37]:
```
data scientist                 206
data engineer                   75
other scientist                 69
analyst                         69
machine learning engineer       10
Data scientist project manager   8
na                               6
data analitics                   5
data modeler                     4
director                         3
Name: job_title, dtype: int64
```

In [46]:
```python
data_career = ['data scientist', 'other scientist', 'analyst', 'data engineer','data analitics', 'data modeler',
               'machine learning engineer']

data_scientist = job_titles.query('job_title in @data_career')
data_scientist
```

Out[46]:

	Lower Salary	Upper Salary	job_title
0	53	91	data scientist
1	63	112	data scientist
2	80	90	data scientist
3	56	97	data scientist
4	86	143	data scientist
...	...	...	...
695	63	110	data modeler
700	65	113	data scientist
716	59	125	other scientist
732	80	142	machine learning engineer
735	62	113	data engineer

438 rows × 3 columns

In [47]:
```python
data_scientist[['Lower Salary', 'Upper Salary']].describe()
```

Out[47]:

	Lower Salary	Upper Salary
count	438.000000	438.000000
mean	75.283105	128.744292
std	30.003543	44.275024
min	15.000000	16.000000
25%	54.000000	99.000000
50%	71.000000	124.000000
75%	91.000000	150.750000
max	202.000000	306.000000

In [48]:
```python
highly_qualified = ['Data scientist project manager', 'director']
managerial_jobs = DS[['Lower Salary', 'Upper Salary', 'job_title']]
managerial_jobs = managerial_jobs.query('job_title in @highly_qualified')
managerial_jobs.head()
```

Out[48]:

	Lower Salary	Upper Salary	job_title
127	26	55	Data scientist project manager
153	85	134	Data scientist project manager
169	59	116	Data scientist project manager
200	42	86	Data scientist project manager
325	67	135	director

In [49]: `managerial_jobs.groupby('job_title').mean()`

Out[49]:

job_title	Lower Salary	Upper Salary
Data scientist project manager	51.125000	98.000000
director	101.666667	173.666667

In [50]: `managerial_jobs.groupby('job_title').median()`

Out[50]:

job_title	Lower Salary	Upper Salary
Data scientist project manager	47.5	95.0
director	102.0	178.0

In [53]:
```python
seniors = DS[['Lower Salary', 'Upper Salary', 'job_title',
              'seniority_by_title']]

sr = ['sr']

seniors = seniors.query('seniority_by_title in @sr')
seniors.head()
```

Out[53]:

	Lower Salary	Upper Salary	job_title	seniority_by_title
21	73	119	data scientist	sr
38	115	180	data scientist	sr
44	110	150	data scientist	sr
46	158	211	data scientist	sr
60	82	132	data scientist	sr

In [54]: `seniors['job_title'].value_counts()`

Out[54]:
```
data scientist              58
other scientist             23
analyst                     18
data engineer               18
machine learning engineer    2
na                           2
Name: job_title, dtype: int64
```

In [55]:
```python
comparison = data_scientist.copy()
comparison = comparison.rename(columns={'Lower Salary': 'Lower Salary - Data Scientists',
                                        'Upper Salary': 'Upper Salary - Data Scientists'})
comparison.drop(columns = 'job_title', axis = 1, inplace = True)
comparison['Lower Salary - Seniors'] = seniors['Lower Salary']
comparison['Upper Salary - Seniors'] = seniors['Upper Salary']
comparison.describe()
```

Out[55]:

	Lower Salary - Data Scientists	Upper Salary - Data Scientists	Lower Salary - Seniors	Upper Salary - Seniors
count	438.000000	438.000000	119.000000	119.000000
mean	75.283105	128.744292	92.403361	153.882353
std	30.003543	44.275024	32.506927	44.257945
min	15.000000	16.000000	20.000000	35.000000
25%	54.000000	99.000000	71.000000	125.000000
50%	71.000000	124.000000	92.000000	151.000000
75%	91.000000	150.750000	110.500000	180.500000
max	202.000000	306.000000	200.000000	289.000000

Visualization

```python
In [106]: fig, axes = plt.subplots(2, 1, sharex=True, figsize=(12,8))

          plt.xlabel('Annual Salary, (US $k)', color = 'black', fontsize = 12)

          ax1 = sns.histplot(ax = axes[0], x = data_scientist['Lower Salary'],color = 'blue', label = 'Lower Salary', kde = True
          ax1 = sns.histplot (ax = axes[0], x = data_scientist['Upper Salary'],color = 'gold', label = 'Upper Salary', kde = Tru
          axes[0].set_title('Data Scientists', fontsize = 15,color = 'black')
          axes[0].set_ylabel('Count', color = 'black', fontsize = 12)

          ax2 = sns.histplot(ax = axes[1], x = seniors['Lower Salary'],color = 'red', label = 'Lower Salary', kde = True)
          ax2 = sns.histplot (ax = axes[1], x = seniors['Upper Salary'], color = 'yellow', label = 'Upper Salary', kde = True)
          axes[1].set_title('Seniors', fontsize = 15, color = 'black')
          axes[1].set_ylabel('Count', color = 'black', fontsize = 12)

          plt.show()
```
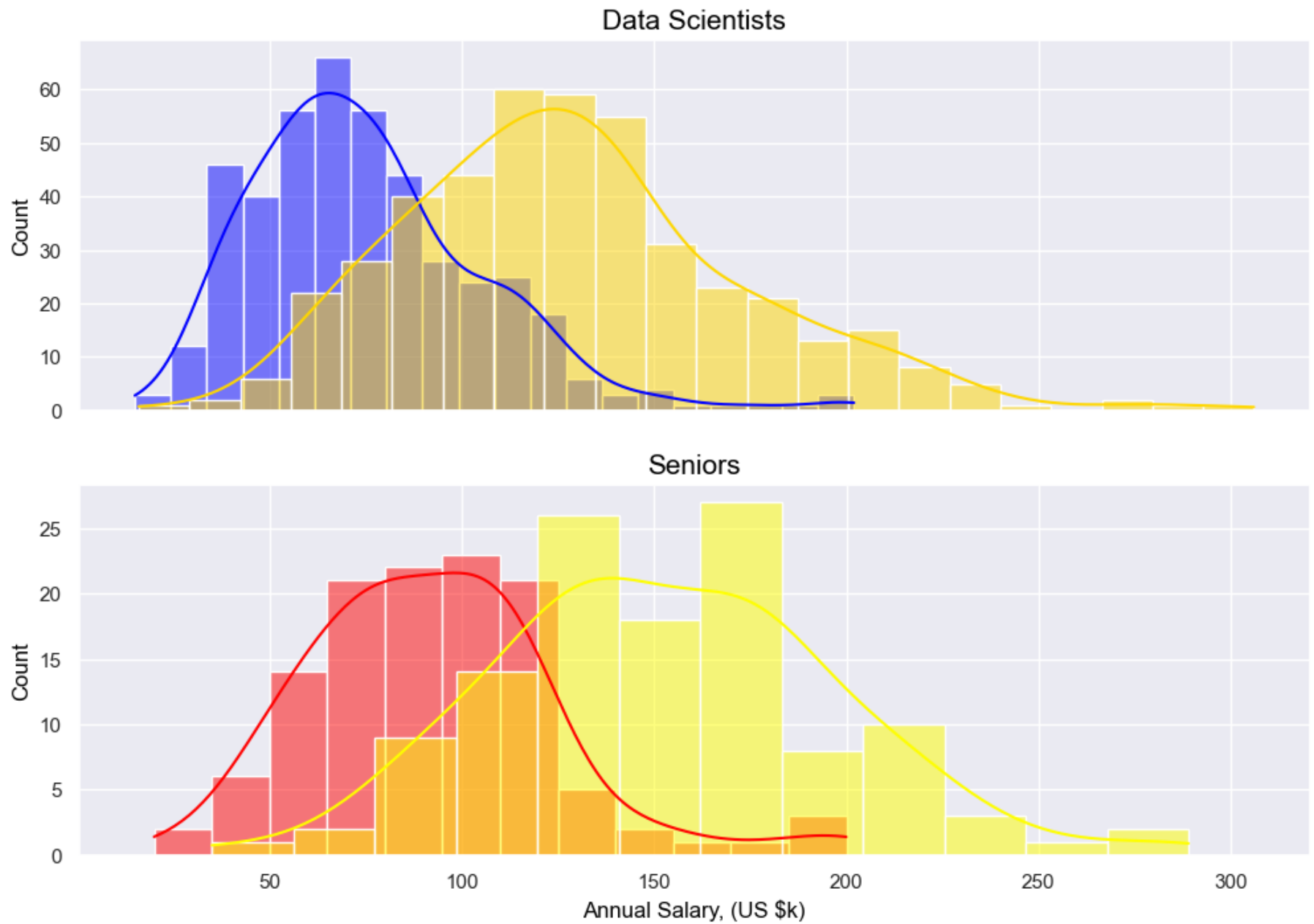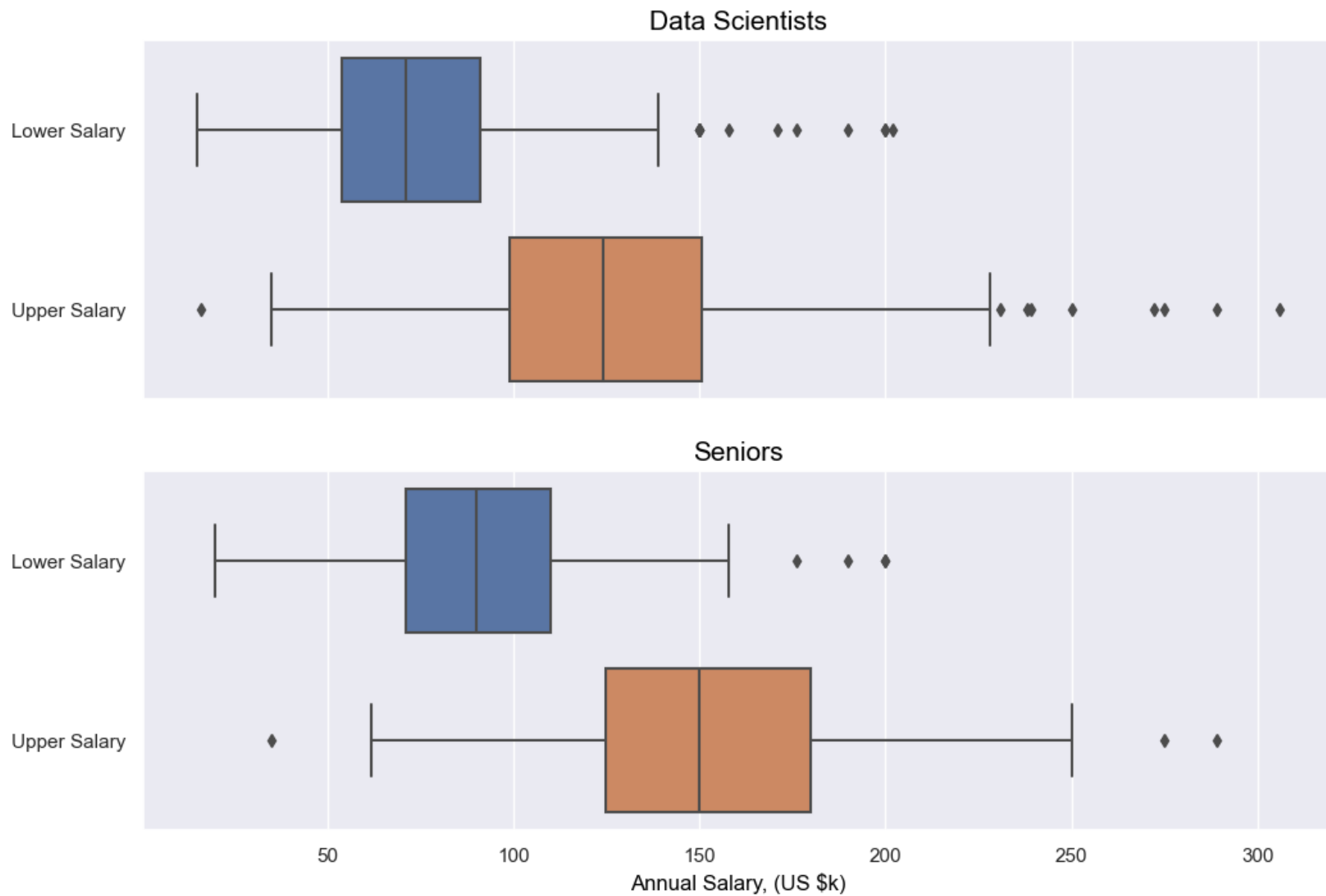
Data Scientists

Seniors

Annual Salary, (US $k)

In [ ]:

In [99]:
```python
fig, axes = plt.subplots(2, 1, sharex=True, figsize=(12,8))

plt.xlabel('Annual Salary, (US $k)', color = 'black', fontsize = 12)

ax1 = sns.boxplot(ax = axes[0], data = data_scientist,order = ['Lower Salary', 'Upper Salary'], orient = 'h')
axes[0].set_title('Data Scientists', fontsize = 15, color = 'black')

ax2 = sns.boxplot(ax = axes[1], data = seniors,order = ['Lower Salary', 'Upper Salary'], orient = 'h')
axes[1].set_title('Seniors', fontsize = 15, color = 'black')

plt.show()
```
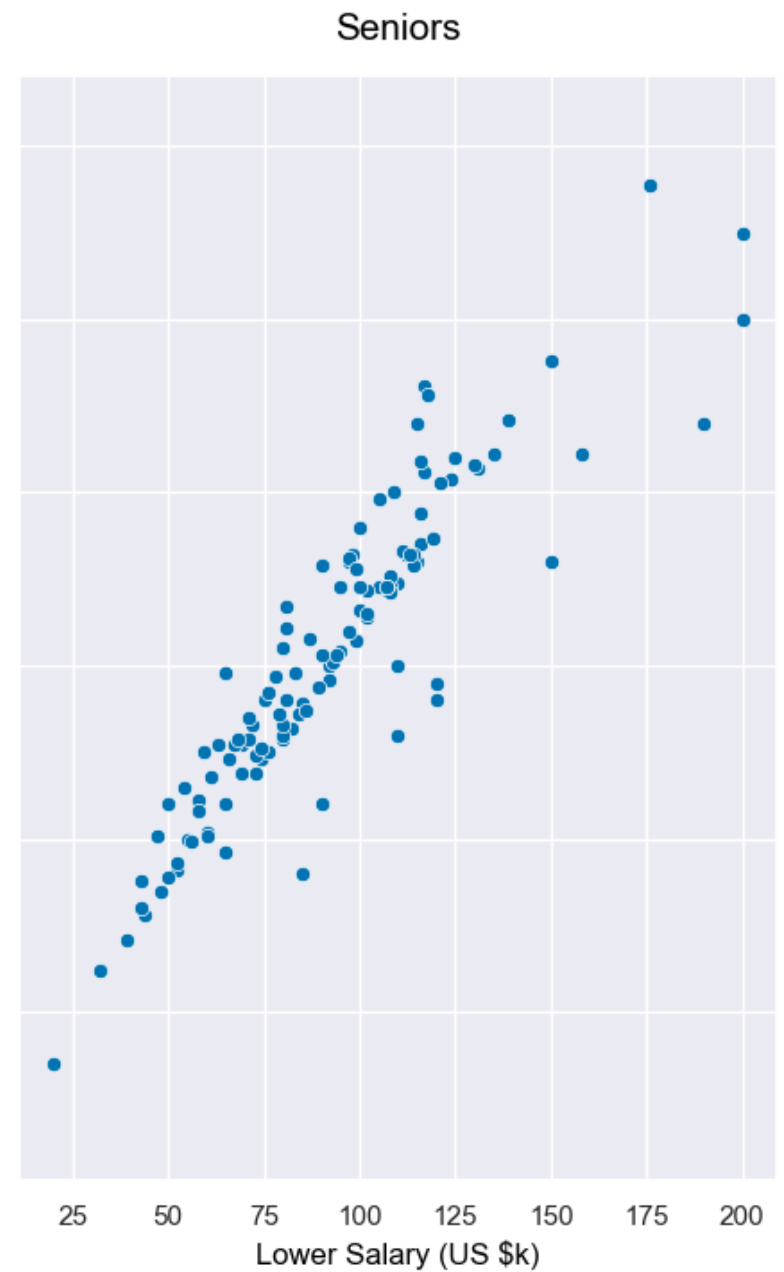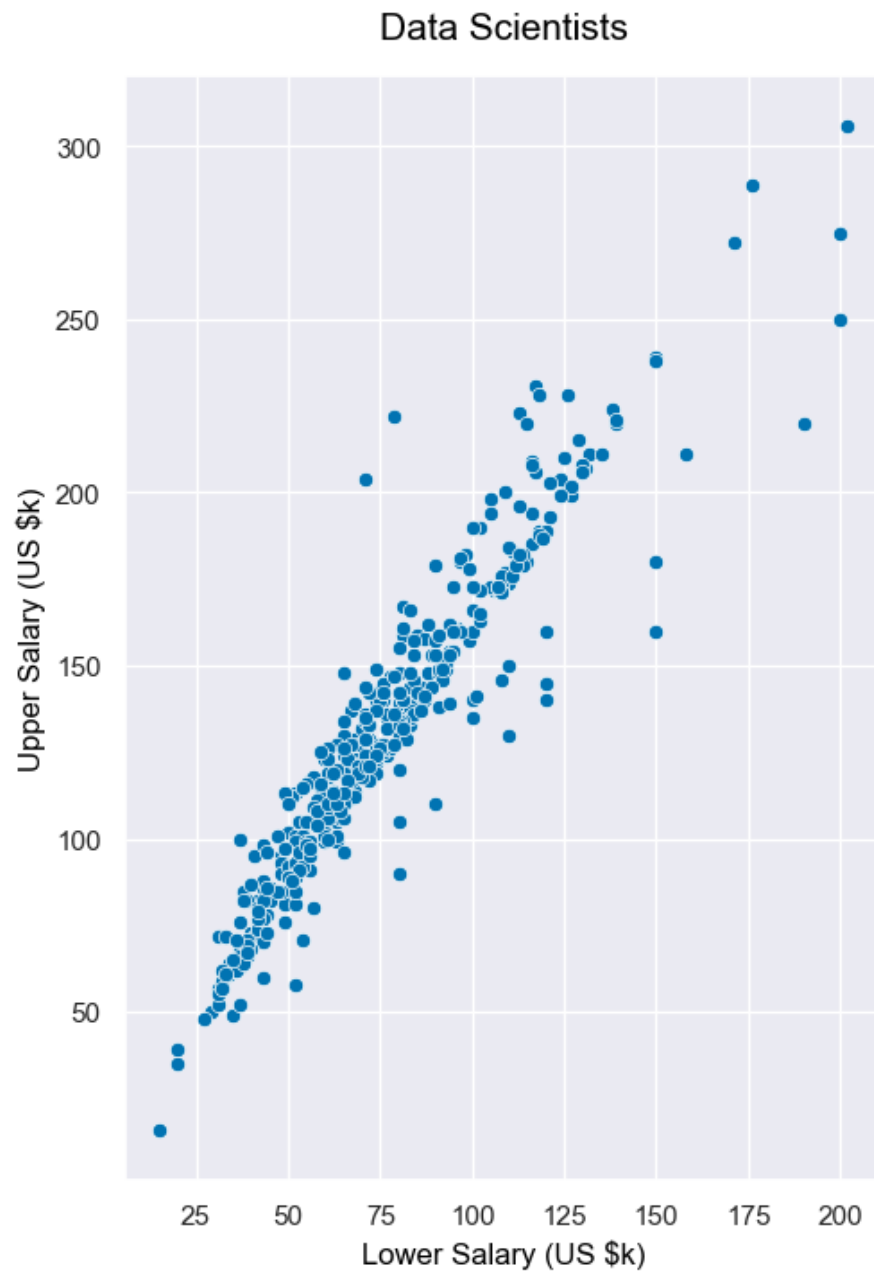
In [ ]:

```python
fig, axes = plt.subplots(1, 2, sharey=True, figsize=(12, 8))

ax1 = sns.scatterplot(ax=axes[0], x="Lower Salary", y="Upper Salary",data=data_scientist)
ax1.set_title('Data Scientists', color='black', fontsize=15, y=1.02)
axes[0].set_xlabel('Lower Salary (US $k)', color='black', fontsize=12)
axes[0].set_ylabel('Upper Salary (US $k)', color='black', fontsize=12)

ax2 = sns.scatterplot(ax=axes[1], x="Lower Salary", y="Upper Salary",data=seniors)
ax2.set_title('Seniors', color='black', fontsize=15, y=1.02)
axes[1].set_xlabel('Lower Salary (US $k)', color='black', fontsize=12)
axes[1].set_ylabel('', color='black', fontsize=12)

plt.show()
```

--END--

In [ ]: