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

In [2]:
```
salary=pd.read_csv('data_scientist_salary.csv')
```

In [3]:
```
salary.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 [4]:
```
salary.describe()
```

Out[4]:

	index	Rating	Founded	Hourly	Employer provided	Lower Salary	Upper Salary	S
count	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742
mean	469.129380	3.618868	1837.154987	0.032345	0.022911	74.754717	128.214286	101
std	279.793117	0.801210	497.183763	0.177034	0.149721	30.945892	45.128650	37
min	0.000000	-1.000000	-1.000000	0.000000	0.000000	15.000000	16.000000	15
25%	221.500000	3.300000	1939.000000	0.000000	0.000000	52.000000	96.000000	73
50%	472.500000	3.700000	1988.000000	0.000000	0.000000	69.500000	124.000000	97
75%	707.750000	4.000000	2007.000000	0.000000	0.000000	91.000000	155.000000	122
max	955.000000	5.000000	2019.000000	1.000000	1.000000	202.000000	306.000000	254

8 rows × 25 columns

In [5]:
```
salary.shape
```

Out[5]:
```
(742, 42)
```

In [6]:
```
salary.isnull().sum()
```

...

In [7]:
```
salary.query('company_txt == "Reynolds American"')
```

Out[7]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Head
331	417	Scientist Manufacturing - Kentucky BioProcessing	$68K-$139K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Owensboro, KY	Sa
358	458	Scientist - Analytical Services	$65K-$134K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa
364	466	Senior Scientist - Regulatory Submissions	$80K-$155K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa
393	501	Scientist Manufacturing Pharma - Kentucky BioP...	$68K-$139K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Owensboro, KY	Sa
413	528	Senior Scientist - Toxicologist - Product Inte...	$47K-$101K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa
421	539	Senior Scientist - Biostatistician	$65K-$96K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa
478	610	Scientist Manufacturing - Kentucky BioProcessing	$68K-$139K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Owensboro, KY	Sa
534	683	Scientist - Analytical Services	$65K-$134K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa
543	695	Senior Scientist - Regulatory Submissions	$80K-$155K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa
596	756	Scientist Manufacturing Pharma - Kentucky BioP...	$68K-$139K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Owensboro, KY	Sa
622	791	Senior Scientist - Toxicologist - Product Inte...	$47K-$101K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa
636	812	Senior Scientist - Biostatistician	$65K-$96K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa
715	926	Scientist - Analytical Services	$65K-$134K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Head
725	938	Senior Scientist - Regulatory Submissions	$80K-$ 155K (Glassdoor est.)	British American Tobacco\nReynolds American In...	3.1	Reynolds American\n3.1	Winston-Salem, NC	Sa

14 rows × 42 columns

In [8]:
```
salary=salary.drop_duplicates(['company_txt', 'Upper Salary'])
```

In [9]:
```
salary
```

Out[9]:

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headc
0	0	Data Scientist	$53K−91K (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Go
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	Baltim
2	2	Data Scientist	$80K−90K (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Cle
3	3	Data Scientist	$56K−97K (Glassdoor est.)	*Organization and Job ID**\nJob ID: 310709\n\n...	3.8	PNNL\n3.8	Richland, WA	Richl
4	4	Data Scientist	$86K−143K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New Y
...	...	...	...	...	...	...	...	
695	896	Data Architect / Data Modeler	$63K−110K (Glassdoor est.)	Medidata: Conquering Diseases Together\n\nMedi...	4.3	Medidata Solutions\n4.3	New York, NY	New Y
700	901	Data Scientist	$65K−113K (Glassdoor est.)	DatamanUSA has an exciting opportunity for a D...	3.4	DatamanUSA, LLC\n3.4	Olympia, WA	Cer
716	928	Associate Scientist / Sr. Associate Scientist,...	$59K−125K (Glassdoor est.)	Who we are\n\n23andMe, the leading consumer ge...	4.0	23andMe\n4.0	South San Francisco, CA	Su
732	945	Machine Learning Engineer (NLP)	$80K−142K (Glassdoor est.)	CK-12's mission is to provide free access to o...	4.1	CK-12 Foundation\n4.1	Palo Alto, CA	Palo
735	948	Data Engineer	$62K−113K (Glassdoor est.)	Do you find data architecture exciting? Does b...	3.9	Fivestars\n3.9	San Francisco, CA	Franci

455 rows × 42 columns

In [10]:
```
columns_deleted = ['Job Title', 'Salary Estimate', 'Job Description', 'Rating',
                   'Industry', 'Revenue', 'Competitors', 'Hourly', 'Employer pr
salary.drop(columns = columns_deleted, axis = 1, inplace = True)

change_name_column = {'job_title_sim': 'job_title'}
salary.rename(columns = change_name_column, inplace = True)

salary.head()
```

Out[10]:

	index	Location	Size	Type of ownership	Sector	Lower Salary	Upper Salary	Avg Salary(K)	company_txt	L
0	0	Albuquerque, NM	501 - 1000	Company - Private	Aerospace & Defense	53	91	72.0	Tecolote Research	
1	1	Linthicum, MD	10000+	Other Organization	Health Care	63	112	87.5	University of Maryland Medical System	
2	2	Clearwater, FL	501 - 1000	Company - Private	Business Services	80	90	85.0	KnowBe4	
3	3	Richland, WA	1001 - 5000	Government	Oil, Gas, Energy & Utilities	56	97	76.5	PNNL	
4	4	New York, NY	51 - 200	Company - Private	Business Services	86	143	114.5	Affinity Solutions	

5 rows × 29 columns

In [11]:
```
salary.shape
```

Out[11]:
```
(455, 29)
```

In [12]:
```
salary[['Upper Salary','Lower Salary']].describe()
```

Out[12]:

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

In [13]:
```
job_titles=salary[['Upper Salary','Lower Salary','job_title']]
```

In [14]:
```
salary['job_title'].value_counts()
```

Out[14]:
```
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 [15]:
```
high_qualified=['director','Data scientist project manager']
```

In [16]:
```
high_jobs=salary[['Upper Salary','Lower Salary','job_title']]
high_jobs=high_jobs.query('job_title in @high_qualified')
```

In [17]:
```
high_jobs.groupby('job_title').mean()
```

Out[17]:

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

In [18]:
```
high_jobs.groupby('job_title').median()
```

Out[18]:

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

In [19]:
```
seniors=salary[['Upper Salary','Lower Salary','job_title','seniority_by_title']
sr=['sr']
seniors=seniors.query('seniority_by_title in @sr')
```

In [20]:
```
seniors['job_title'].value_counts()
```

Out[20]:
```
data scientist                58
other scientist               23
analyst                       18
data engineer                 18
machine learning engineer      2
na                             2
Name: job_title, dtype: int64
```

In [21]:
```
seniors
```

Out[21]:

	Upper Salary	Lower Salary	job_title	seniority_by_title
21	119	73	data scientist	sr
38	180	115	data scientist	sr
44	150	110	data scientist	sr
46	211	158	data scientist	sr
60	132	82	data scientist	sr
...	...	...	...	...
642	89	50	analyst	sr
643	129	68	na	sr
649	135	71	data engineer	sr
693	140	120	data scientist	sr
716	125	59	other scientist	sr

121 rows × 4 columns

In [22]:
```
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[22]:

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

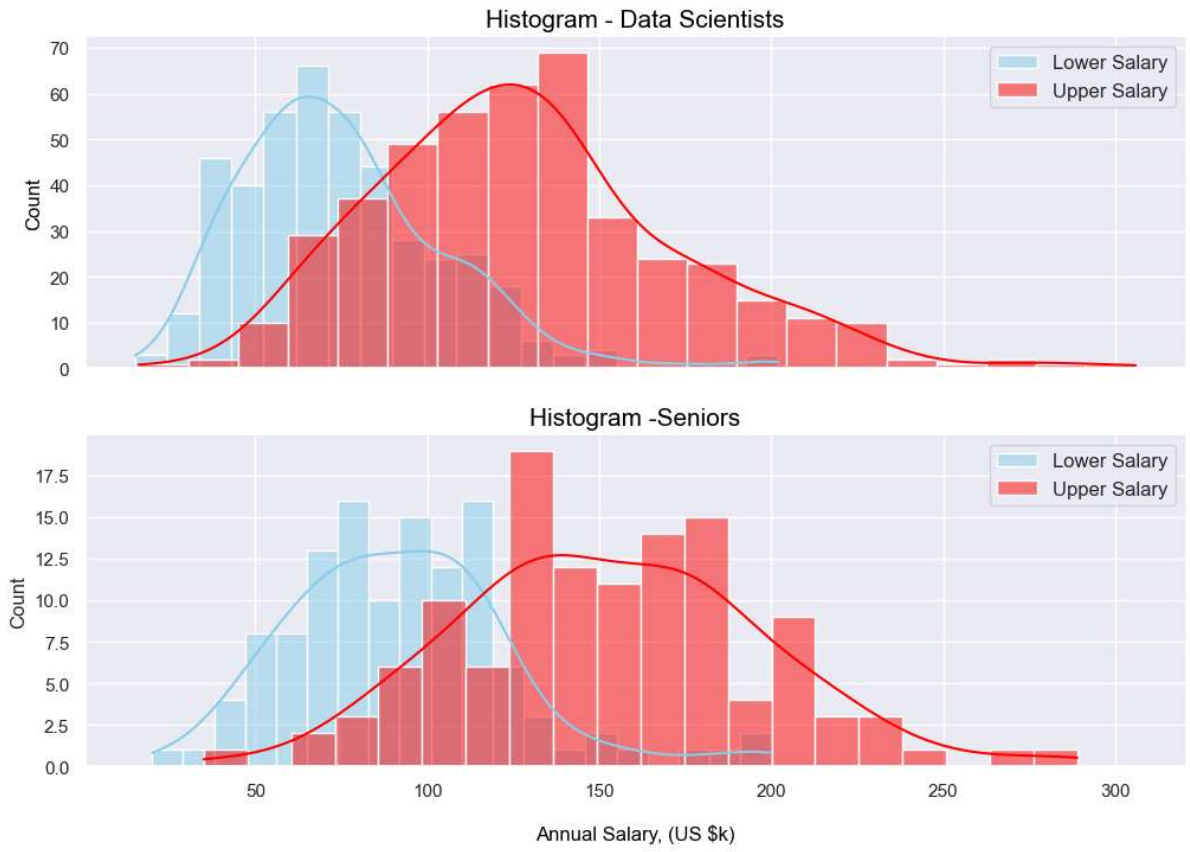
438 rows × 3 columns

In [23]:
```
comparison = data_scientist.copy()
comparison = comparison.rename(columns={'Lower Salary': 'Lower Salary - Data Sc
                                        'Upper Salary': 'Upper Salary - Data Sc
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()
```

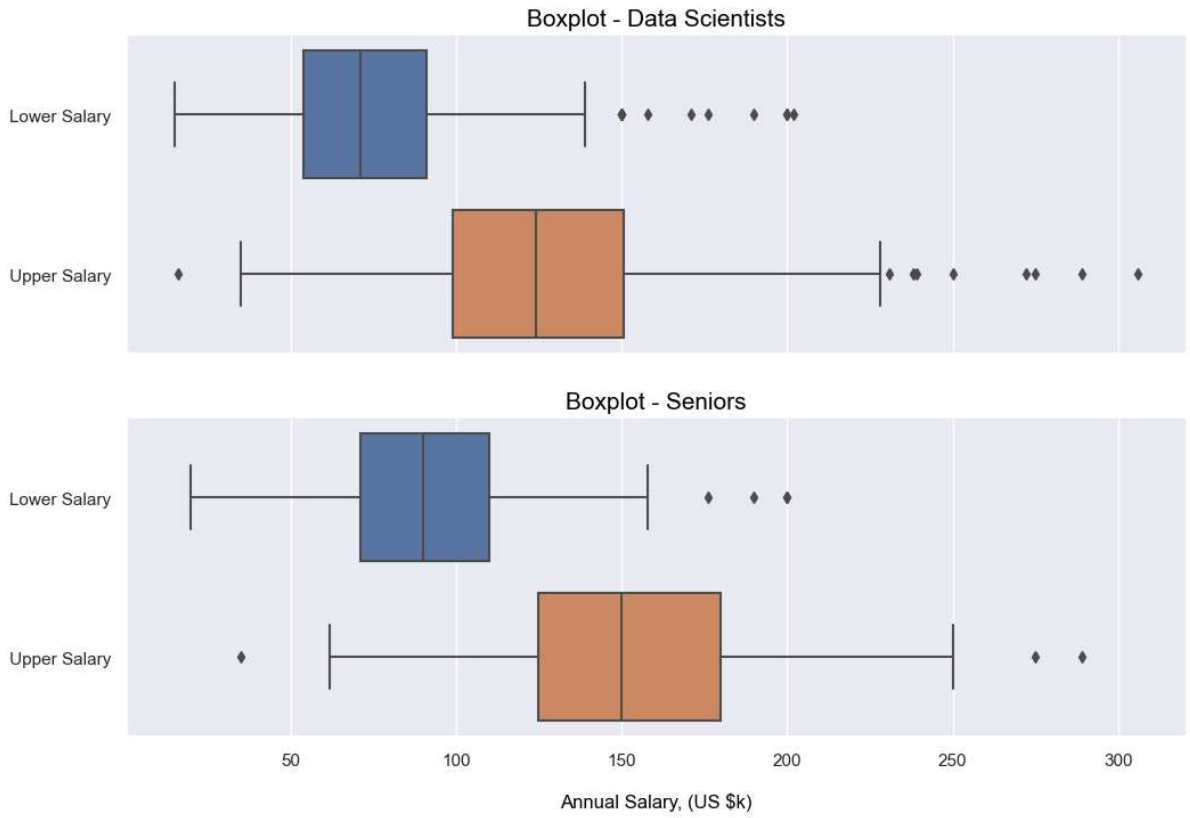
Out[23]:

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

In [24]:
```
sns.set(style = 'darkgrid')

fig, axes = plt.subplots(2, 1, sharex=True, figsize=(12,8))

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

ax1 = sns.histplot(ax = axes[0], x = data_scientist['Lower Salary'], color = 's
ax1 = sns.histplot (ax = axes[0], x = data_scientist['Upper Salary'], color =
axes[0].set_title('Histogram - Data Scientists', fontsize = 15, color = 'black'
axes[0].set_ylabel('Count', color = 'black', fontsize = 12)
ax1.legend(fontsize = 12)


ax2=sns.histplot(x=seniors['Lower Salary'],ax=axes[1], color = 'skyblue', label
ax2=sns.histplot(x=seniors['Upper Salary'],ax=axes[1], color = 'red', label =
ax2.legend(fontsize = 12)
axes[1].set_title('Histogram -Seniors', fontsize = 15, color = 'black')
plt.show()
```



In [25]:
```
sns.set(style = 'darkgrid')

fig, axes = plt.subplots(2, 1, sharex=True, figsize=(12,8))

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

ax1 = sns.boxplot(ax = axes[0], data = data_scientist, order = ['Lower Salary',
axes[0].set_title('Boxplot - Data Scientists', fontsize = 15, color = 'black')
ax2 = sns.boxplot(ax = axes[1], data = seniors, order = ['Lower Salary', 'Upper
axes[1].set_title('Boxplot - Seniors', fontsize = 15, color = 'black')

plt.show()
```



In [26]:
```
data_scientist.corr().round(4)
```

Out[26]:

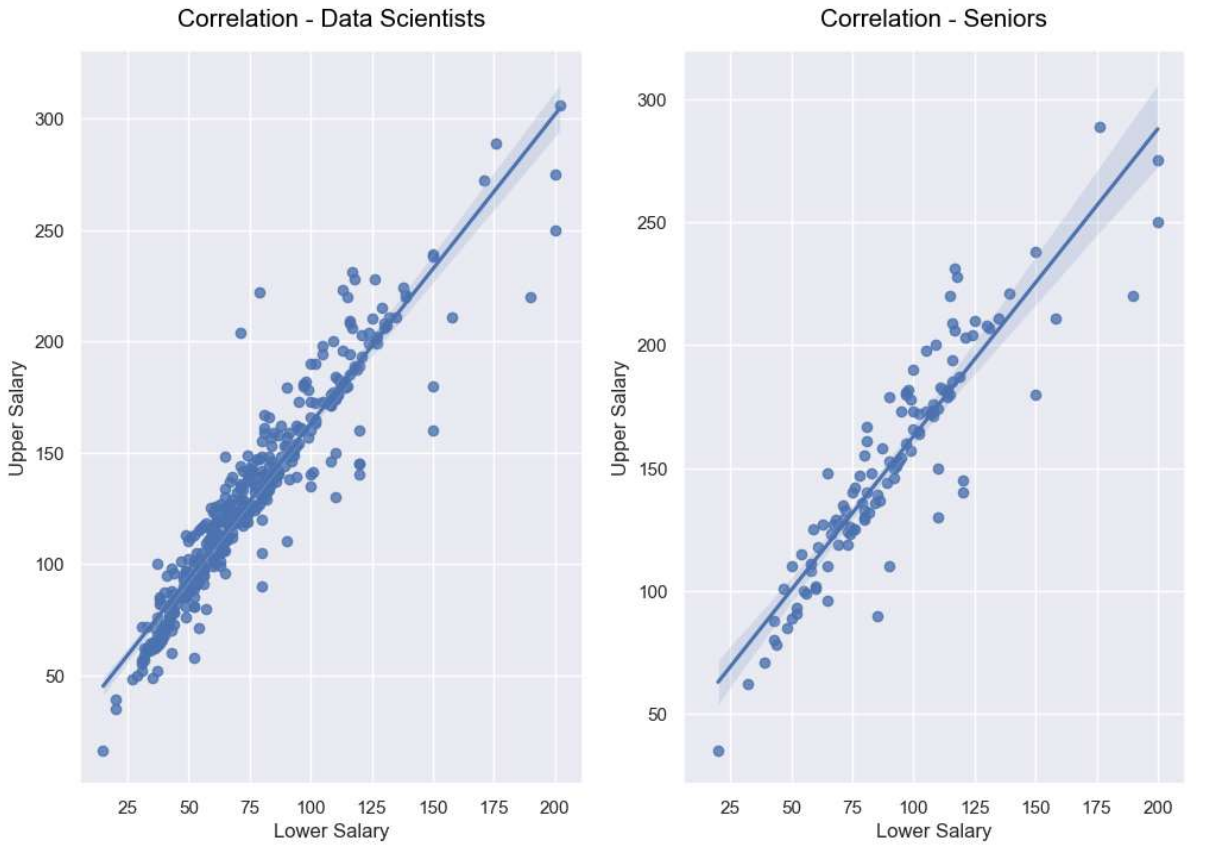
	Upper Salary	Lower Salary
Upper Salary	1.0000	0.9403
Lower Salary	0.9403	1.0000

In [27]:
```
seniors.corr().round(4)
```

Out[27]:

	Upper Salary	Lower Salary
Upper Salary	1.0000	0.9108
Lower Salary	0.9108	1.0000

In [28]:
```
sns.set(style = 'darkgrid')

fig, axes = plt.subplots(1,2, sharex=True, figsize=(12,8))

ax1= sns.regplot(ax = axes[0], x="Lower Salary", y="Upper Salary", data=data_sc
ax1.set_title('Correlation - Data Scientists', color = 'black', fontsize = 15,
ax2=sns.regplot(ax = axes[1],x="Lower Salary", y="Upper Salary",data=seniors)
ax2.set_title('Correlation - Seniors', color = 'black', fontsize = 15, y =1.02)

plt.show()
```



There's a clearly high correlation between the salary range - Upper and Lower - among different companies. The highest correlation is observed in the 'Data Scientist' dataset. By the way, what are the top 10 sectors of the companies in the USA hiring Data Scientists?

Barplot of Sectors and Skills

In [29]:
```
sectors=salary.copy()
```

In [30]:
```
sectors['Sector'].unique()
```

Out[30]:
```
array(['Aerospace & Defense', 'Health Care', 'Business Services',
       'Oil, Gas, Energy & Utilities', 'Real Estate', 'Finance',
       'Information Technology', 'Retail', 'Biotech & Pharmaceuticals',
       'Media', 'Insurance', 'Transportation & Logistics',
       'Telecommunications', '-1', 'Manufacturing', 'Mining & Metals',
       'Government', 'Education', 'Agriculture & Forestry',
       'Travel & Tourism', 'Non-Profit',
       'Arts, Entertainment & Recreation',
       'Construction, Repair & Maintenance', 'Accounting & Legal',
       'Consumer Services'], dtype=object)
```

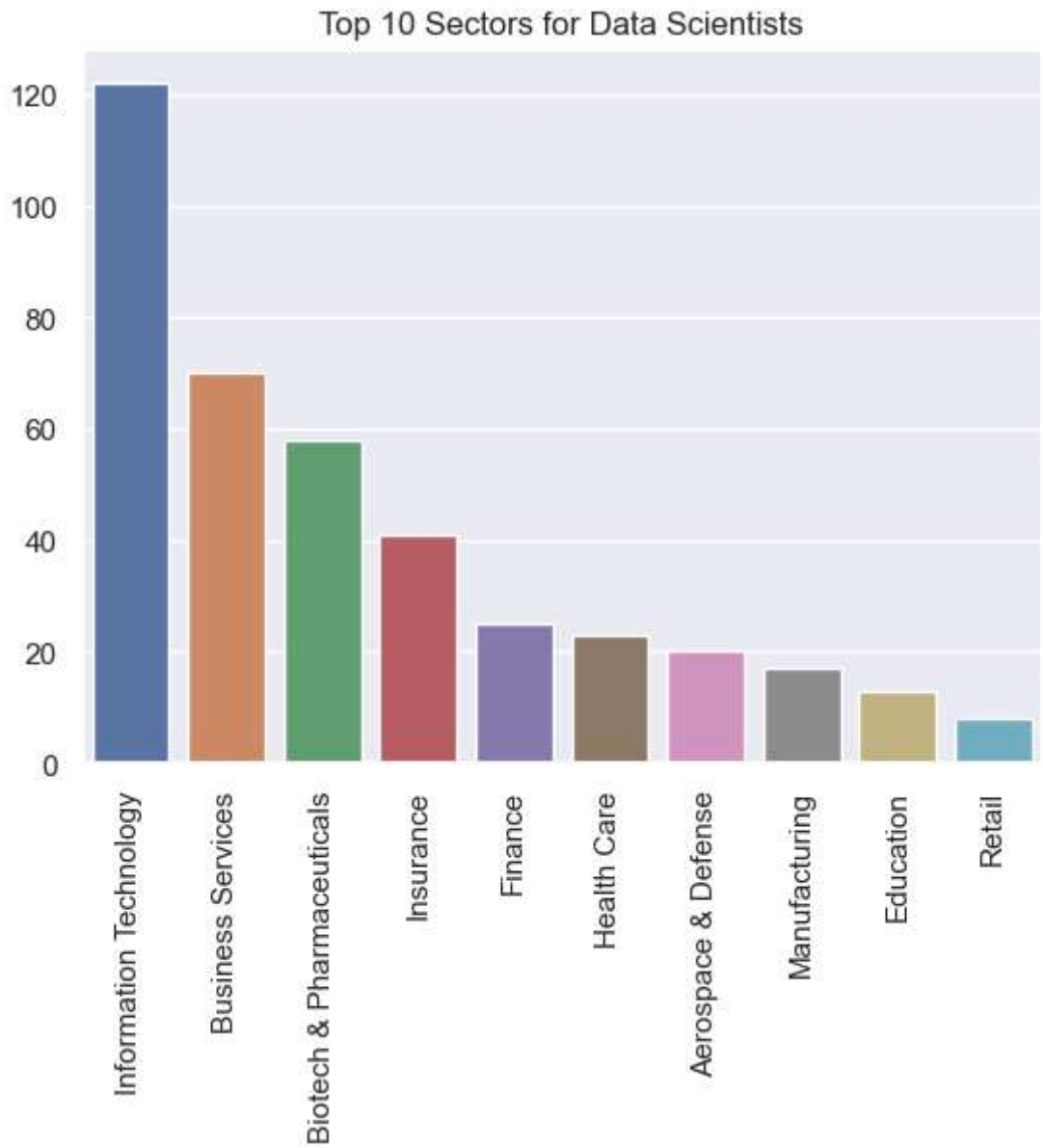
In [31]:
```
sectors = sectors[sectors['Sector'] != '-1']
```

In [32]:
```
sectors=sectors[sectors['Sector'] != '-1']
```

In [33]:
```
top10_sectors=sectors['Sector'].value_counts().iloc[:10]
top10_sectors
```

Out[33]:
```
Information Technology       122
Business Services             70
Biotech & Pharmaceuticals     58
Insurance                     41
Finance                       25
Health Care                   23
Aerospace & Defense           20
Manufacturing                 17
Education                     13
Retail                         8
Name: Sector, dtype: int64
```

In [34]:
```
plt.title('Top 10 Sectors for Data Scientists')
ax=sns.barplot(x=top10_sectors.index,y=top10_sectors.values)
plt.xticks(rotation=90)


plt.show()
```



In [35]:
```
skills=salary.copy()
skills=skills.loc[:,'Python':'google_an']
skills.value_counts()
```

Out[35]:
```
Python  spark  aws  excel  sql  sas  keras  pytorch  scikit  tensor  hadoop
tableau  bi  flink  mongo  google_an
0       0      0    1      0    0    0      0        0       0       0
0        0   0      0      0            52
                    0      0    0    0      0        0       0       0
0        0   0      0      0            43
1       0      0    0      0    0    0      0        0       0       0
0        0   0      0      0            14
                    1      0    0      0        0       0       0
0        0   0      0      0            14
                    1      0    0    0      0        0       0       0
0        0   0      0      0            13

..
                    1      0    0      0        0       0       0
0        1   0      0      0              1
                    0      1    0      0        0       0       0
1        1   0      0      0              1

0        0   0      0      0              1
                           0    1      0        0       1       0
0        0   0      0      0              1
        1      1    1      1    1    0      0        1       0       1
1        0   0      0      0              1
Length: 172, dtype: int64
```

In [36]:
```
list_0 = []
list_1 = []
for column in skills:
    series_count = skills[column].value_counts()
    list_0.append(series_count[0])
    list_1.append(series_count[1])

skills = pd.DataFrame(data = (list_0, list_1), index = ['not required', 'requir
skills = skills.T
skills['total'] = skills['not required'] + skills['required']
skills['required (%)'] = round(100 * skills['required'] / skills['total'], 2)
skills = skills.sort_values("required", ascending=False)
skills
```

Out[36]:

	not required	required	total	required (%)
Python	197	258	455	56.70
sql	201	254	455	55.82
excel	208	247	455	54.29
aws	344	111	455	24.40
spark	345	110	455	24.18
tableau	356	99	455	21.76
hadoop	373	82	455	18.02
tensor	408	47	455	10.33
sas	414	41	455	9.01
bi	416	39	455	8.57
scikit	418	37	455	8.13
pytorch	431	24	455	5.27
mongo	431	24	455	5.27
keras	436	19	455	4.18
google_an	447	8	455	1.76
flink	448	7	455	1.54

In [37]:
```
skills
```

Out[37]:

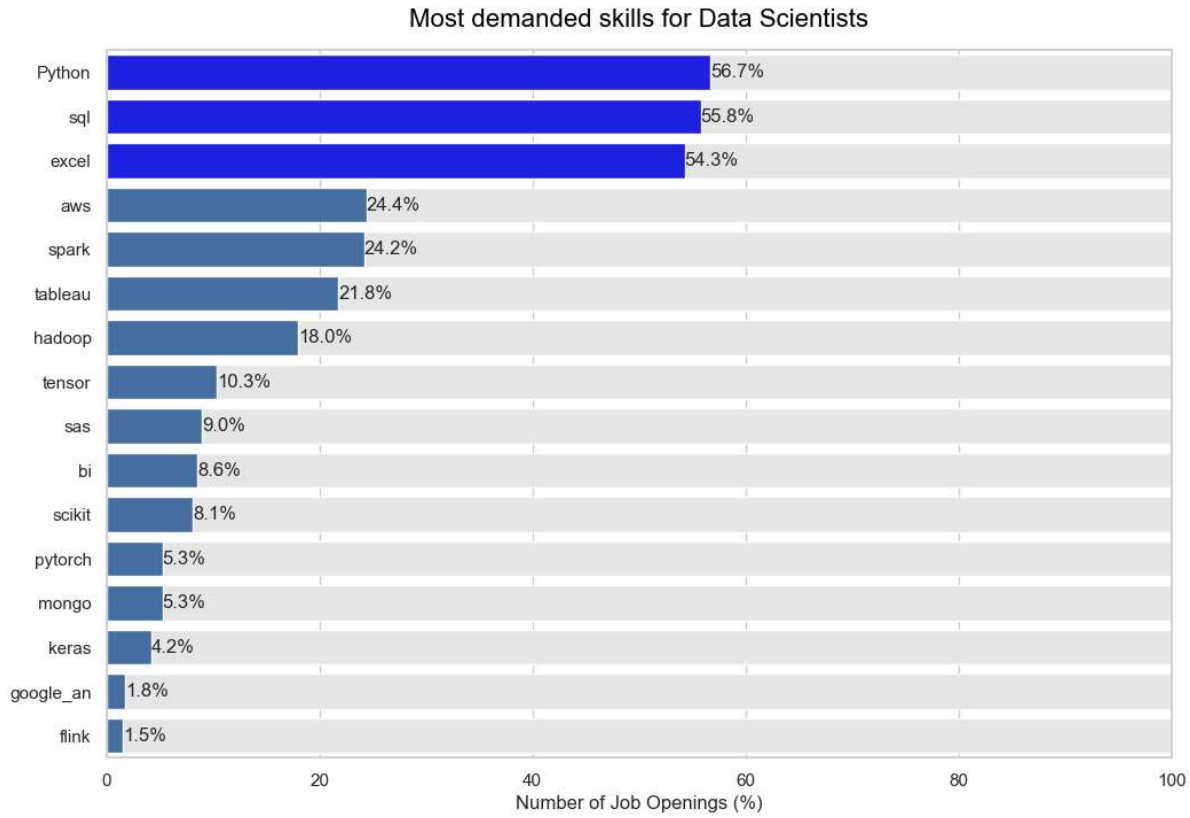
	not required	required	total	required (%)
Python	197	258	455	56.70
sql	201	254	455	55.82
excel	208	247	455	54.29
aws	344	111	455	24.40
spark	345	110	455	24.18
tableau	356	99	455	21.76
hadoop	373	82	455	18.02
tensor	408	47	455	10.33
sas	414	41	455	9.01
bi	416	39	455	8.57
scikit	418	37	455	8.13
pytorch	431	24	455	5.27
mongo	431	24	455	5.27
keras	436	19	455	4.18
google_an	447	8	455	1.76
flink	448	7	455	1.54

In [38]:
```
skills
```

Out[38]:

	not required	required	total	required (%)
Python	197	258	455	56.70
sql	201	254	455	55.82
excel	208	247	455	54.29
aws	344	111	455	24.40
spark	345	110	455	24.18
tableau	356	99	455	21.76
hadoop	373	82	455	18.02
tensor	408	47	455	10.33
sas	414	41	455	9.01
bi	416	39	455	8.57
scikit	418	37	455	8.13
pytorch	431	24	455	5.27
mongo	431	24	455	5.27
keras	436	19	455	4.18
google_an	447	8	455	1.76
flink	448	7	455	1.54

In [45]:
```
sns.set_theme(style = 'whitegrid')

f, ax = plt.subplots(figsize = (12, 8))

sns.set_color_codes('pastel')
sns.barplot(x = 'total', y = skills.index, data = skills, color = '0.9')

sns.set_color_codes('muted')
top3 = ['blue' if (x > 50) else '#386cb0'for x in skills['required (%)'] ]
sns.barplot(x = 'required (%)', y = skills.index, data = skills,
            label = 'Top 3 demanded skills', palette = top3)


ax.set(xlim = (0, 100), ylabel = "", xlabel = 'Number of Job Openings (%)')
ax.set_title('Most demanded skills for Data Scientists', color = 'black', fonts
ax.bar_label(ax.containers[1], fmt = '%1.1f%%')

plt.show()
```



In [ ]:
```
```