Homework 3

Due: Monday Oct 14, at 11:59pm via Blackboard

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

plt.style.use('ggplot')
plt.rcParams["figure.figsize"]=10,6
```

Problem 1: Peformance of Large vs. Small Companies

Companies vary greatly in size. This variation can hide how well a company is performing. Rather than looking at the raw profit numbers, analysts consider financial ratios that adjust for the size of the company. A popular ratio is the return on assets, defined as:

Return on Assets $= NetIncome/TotalAssets$

Net income is another name for profits, and the total assets of a company is the value of everything it owns that is used to produce profits. The return on assets indicates how much profit the company generates relative to the amount that it invested to make that profit. A company with losses rather than profits has a negative return on assets.

Data: The data set `Company.csv` gives the company name, total assets (in Millions $), net income (in Millions $), and the number of employees reported by 167 retailers in the United States.

In the following questions, you will be performing an **exploratory data analysis (EDA)** for the given companies data.

In [2]:
```
#Read the data
```

Out[2]:
```
(167, 4)
```

In [5]:
```

```

Out[5]:

	Company Name	Total Assets (M$)	Net Income (M$)	# Employees
0	1-800-FLOWERS.COM	256	-4	2200
1	99 CENTS ONLY STORES	824	74	12000
2	A.C. MOORE ARTS & CRAFTS INC	237	-30	4710
3	ABERCROMBIE & FITCH -CL A	2948	150	85000
4	ADVANCE AUTO PARTS INC	3354	346	51017

In [10]:
```

```

Out[10]:

	Total Assets (M$)	Net Income (M$)	# Employees	Return on Assets
count	167.0	167.0	167.0	167.0
mean	5287.0	334.0	49385.0	0.0
std	16120.0	1385.0	173006.0	0.0
min	102.0	-1510.0	193.0	-1.0
25%	348.0	2.0	4120.0	0.0
50%	992.0	34.0	12700.0	0.0
75%	3040.0	191.0	35300.0	0.0
max	180663.0	16389.0	2100000.0	0.0

1a. (2 points) Compute and report (in a short paragraph of text) the following summary statistics for the `Net Income (M\$)` data (round your values to the nearest integer). Hint: Use the Round function

- Mean
- Median
- Standard Deviation
- Range
- IQR

In [3]:
```
# mean
```

Out[3]: 334

In [4]:
```
# median
```

Out[4]: 34

In [5]:
```
# standard deviation
```

Out[5]: 1385

In [6]:
```
# range
```

Out[6]: 17899

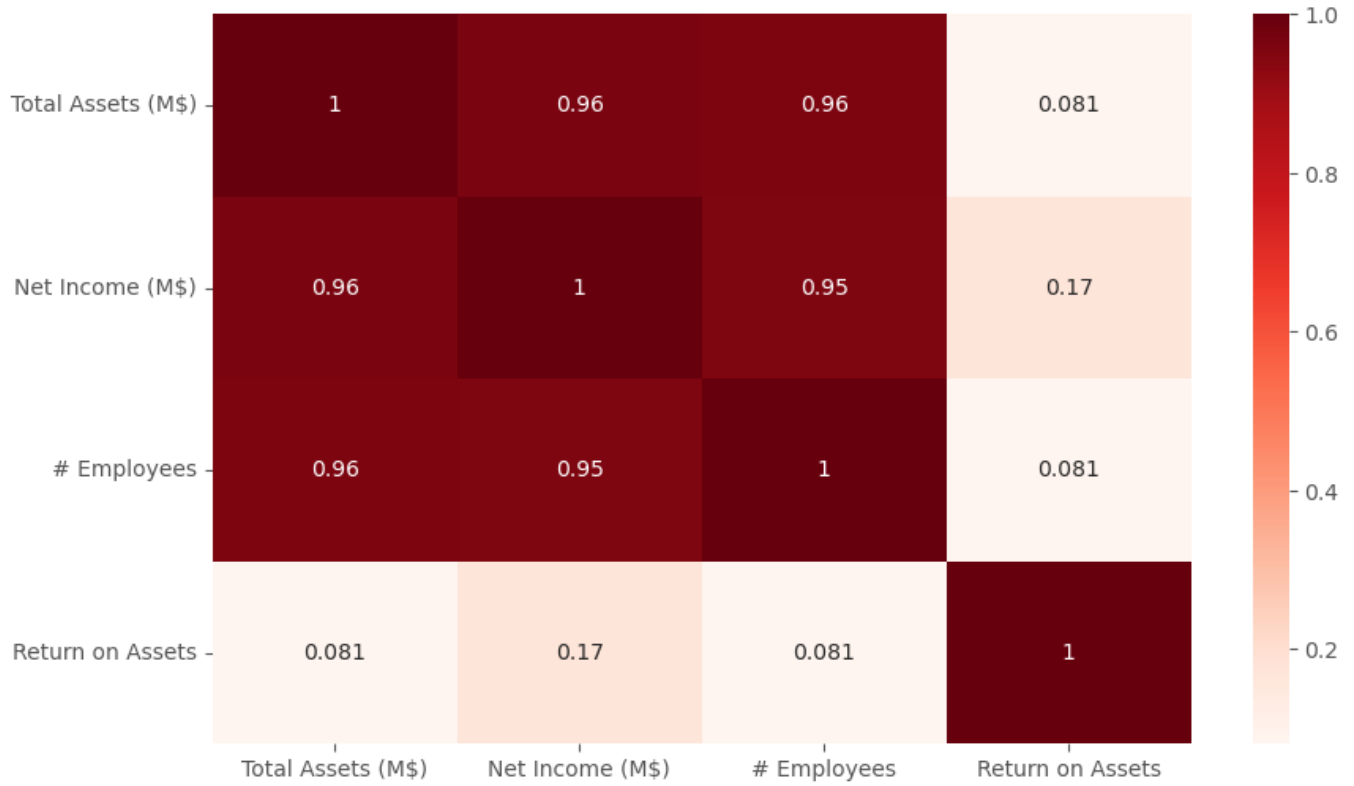
In [7]:
```
# IQR
```

Out[7]: 188

(2 points) Create a heatmap for the dataset company. Can Net income be a factor determining Return on Assets? Briefly Explain

In [7]:
```
```

Out[7]: <Axes: >



1b. (2 points) Report the proportion of companies that inccured losses. For this question, you are expected to add a new categorical variable to the dataset (call it `Profit`) with two levels: `PROFIT` if the net income is above zero (net income >= 0) and `LOSS` if the net income is below zero (net income < 0).

In [12]:

```
```

Out[12]:

```
PROFIT    0.766467
LOSS      0.233533
Name: Profit, dtype: float64
```

In [13]:

```
#OR Using List comprehension
```

Out[13]:

	Company Name	Total Assets (M$)	Net Income (M$)	# Employees	Profit	ProfitA
0	1-800-FLOWERS.COM	256	-4	2200	LOSS	Loss
1	99 CENTS ONLY STORES	824	74	12000	PROFIT	Profit
2	A.C. MOORE ARTS & CRAFTS INC	237	-30	4710	LOSS	Loss
3	ABERCROMBIE & FITCH -CL A	2948	150	85000	PROFIT	Profit
4	ADVANCE AUTO PARTS INC	3354	346	51017	PROFIT	Profit

In [8]:

```
```

Out[8]:

```
Profit    0.766467
Loss      0.233533
Name: ProfitA, dtype: float64
```

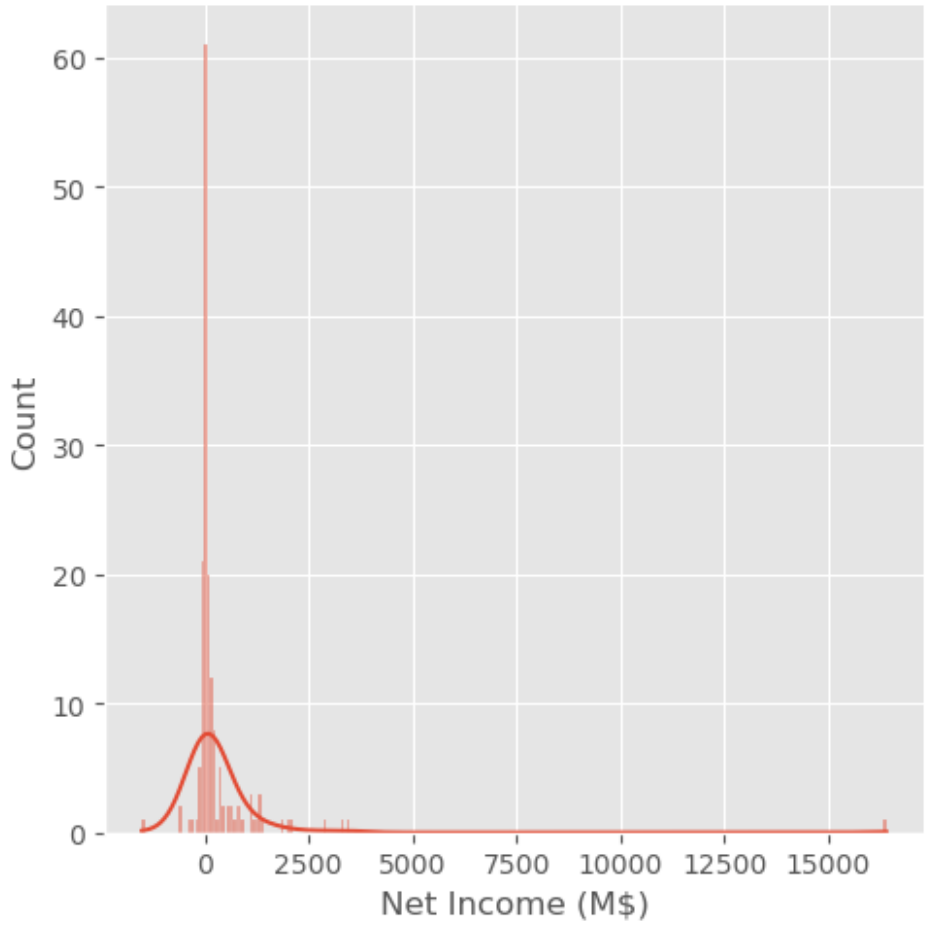
In [ ]:

```
```

1c. (2 points) What is the shape of the distribution of the variable `Net Income (M$)` ? For this question, you are expected to create **both** a histogram and a boxplot, and comment about the shape of the distribution and if there are any companies with an outlier net income.

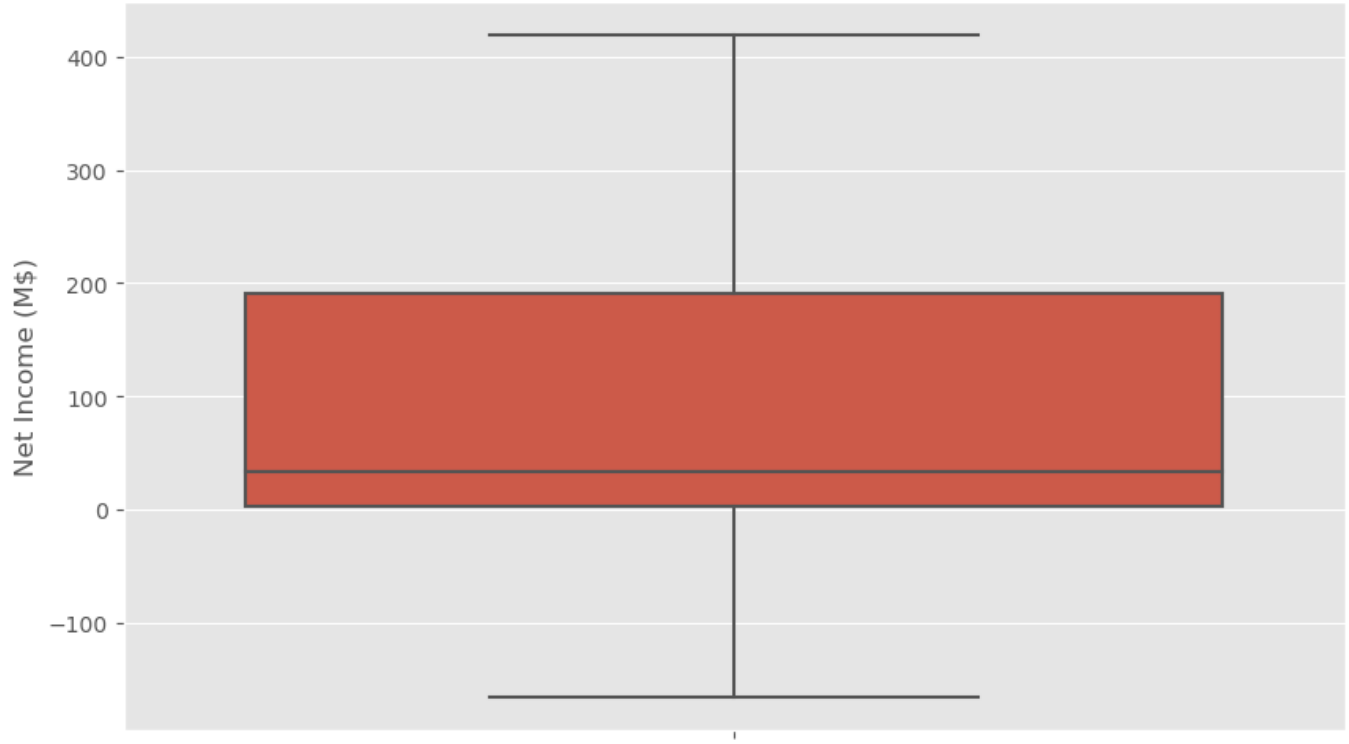
In [16]:

Out[16]: <seaborn.axisgrid.FacetGrid at 0x1d10d7ed990>



In [15]:

Out[15]: <Axes: ylabel='Net Income (M$)'>



1d. (2 points) A company that has more than 5000 employees is considered a large one, otherwise it is cosidered small. Create a new categroical variable (call it `Company Size`) with two levels: `LARGE` if the number of employees is greater than 5000 (employees > 5000), and `SMALL` otherwise (employees <=5000). What is the % of large and small companies in the dataset?

In [17]:

Out[17]:
```
LARGE    0.688623
SMALL    0.311377
Name: Company Size, dtype: float64
```

In [18]:
```
#OR using Lamdba
```

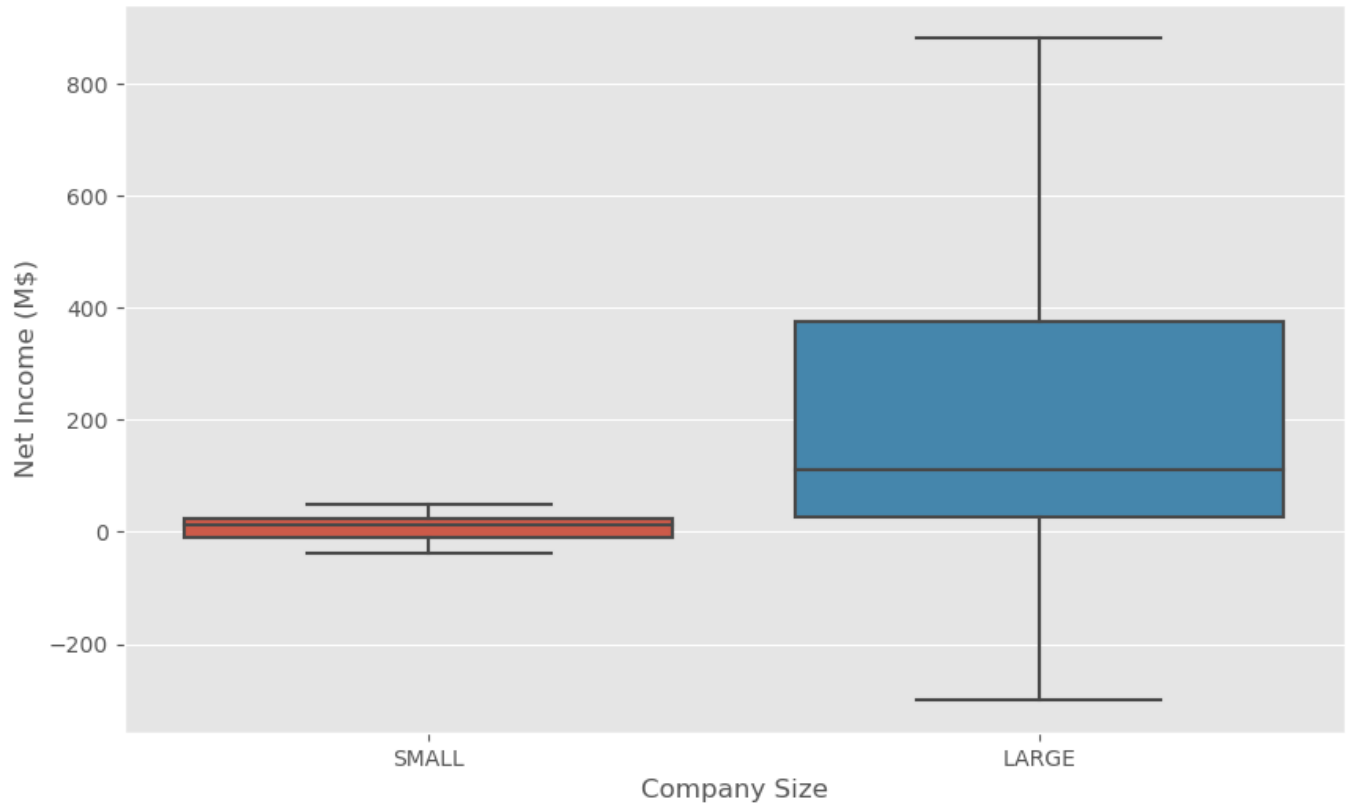
Out[18]:

	Company Name	Total Assets (M$)	Net Income (M$)	# Employees	Profit	ProfitA	ProfitB	Company Size	Company SizeA
0	1-800-FLOWERS.COM	256	-4	2200	LOSS	Loss	Loss	SMALL	Small
1	99 CENTS ONLY STORES	824	74	12000	PROFIT	Profit	Profit	LARGE	Large
2	A.C. MOORE ARTS & CRAFTS INC	237	-30	4710	LOSS	Loss	Loss	SMALL	Small
3	ABERCROMBIE & FITCH -CL A	2948	150	85000	PROFIT	Profit	Profit	LARGE	Large
4	ADVANCE AUTO PARTS INC	3354	346	51017	PROFIT	Profit	Profit	LARGE	Large

1e. (2 points) Create a side-by-side boxplot to compare the distribtution of `Net Income (M$)` for both Large and Small companies eliminating the outliers. What does this graph tell you about the net income for both types of companies?

In [19]:

Out[19]: `<Axes: xlabel='Company Size', ylabel='Net Income (M$)'>`



In [ ]:

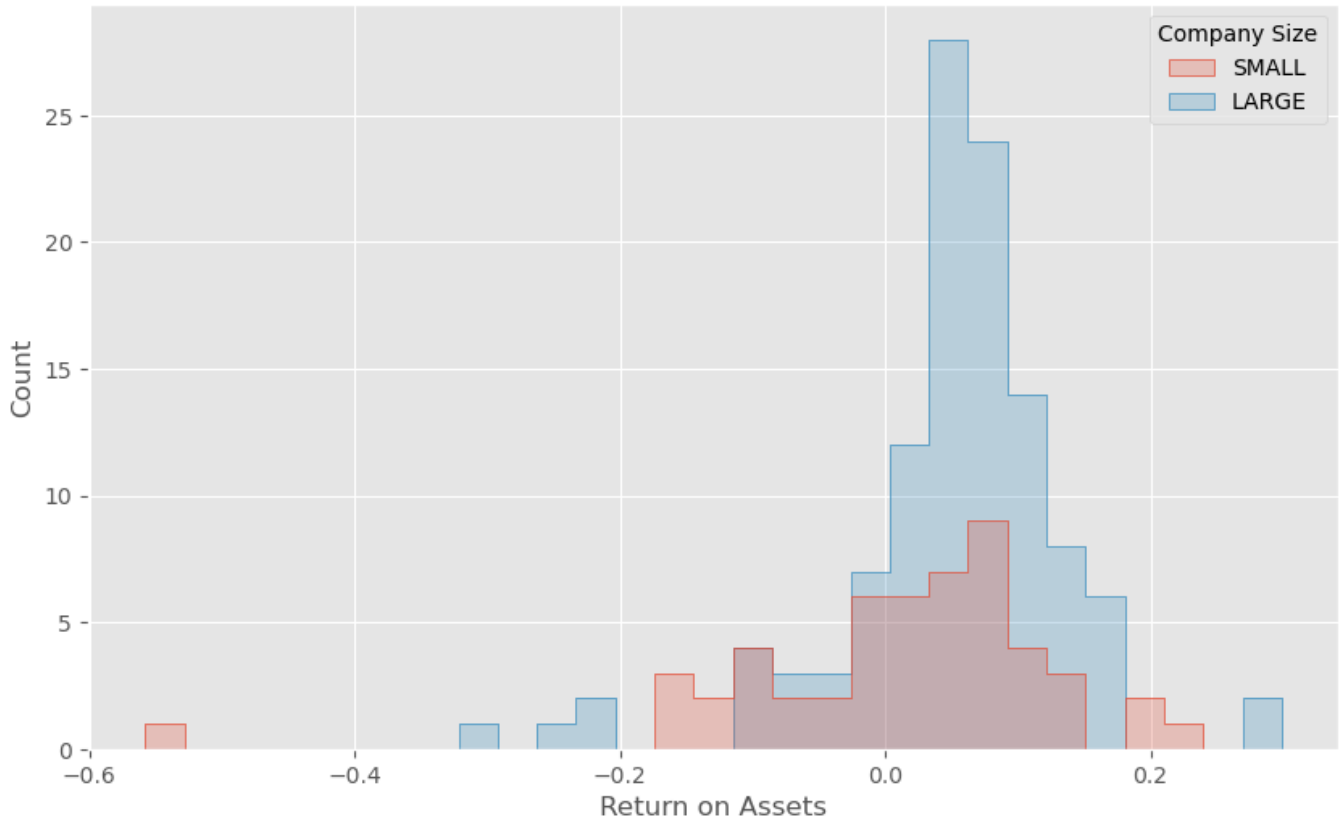
1f. (3 points) A better way to assess the performance of companies is to look at their `Return on Assets` instead of looking only at net income. The return on assets indicates how much profit the company generates relative to the amount that it invested to make profits.

- Create a new numerical variable (call it `Return on Assets`) based on the formula: **Return on Assets = Net Income/Total Assets**.
- What is the shape of the distribution of the variable `Return on Assets` ? For this question, you are expected to create **both** a histogram, using Seaborn's histplot and a boxplot, to distinguish between large and small companies, and comment about the shape of the distribution and if there are any companies with an outlier return on assets value.
- Create a side-by-side boxplot to compare the distribtution of `Return on Assets` for both Large and Small companies. What does this graph tell you about the return on assets for both types of companies?

In [4]:

In [21]:

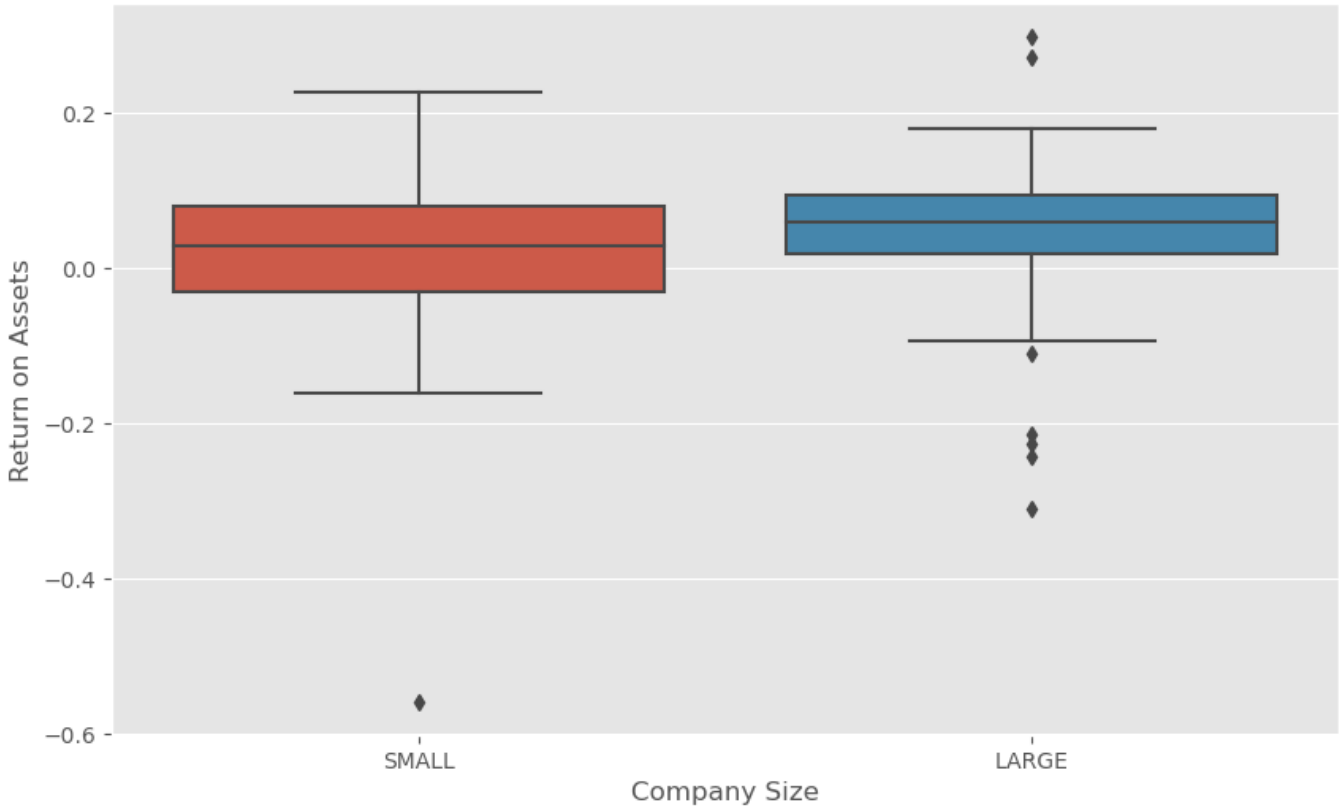
Out[21]: <Axes: xlabel='Return on Assets', ylabel='Count'>



In [ ]:

In [23]:

Out[23]: <Axes: xlabel='Company Size', ylabel='Return on Assets'>



1g. (3 points) Create a scatterplot of Total Assets (x) against Net Income (y),

For Company size, distinguis between Small and Large companies using a different color.

-Add horizontal and vertical lines to your graph to correspond to the mean Net Income (horizontal) and mean Total Assests (vertical), selecting orange as the line color and 'dashed' as the linesyle

-Add the title "Total Assets vs. Net Income" with a fontsize of 14 and locate the title to the center

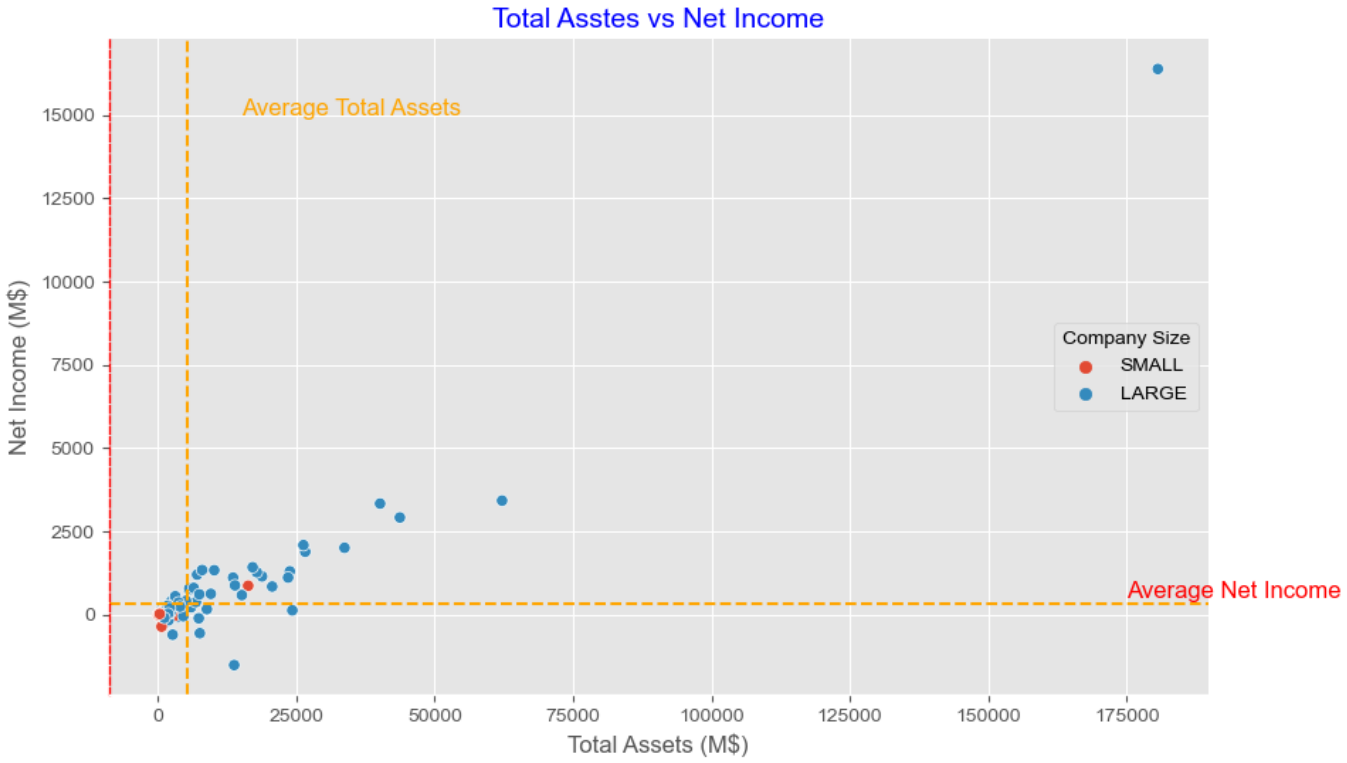
-Eliminate the top and tight spines, and set the color of the left spine to red and 'dashed' as the linestyle

-Add text "Average Total Assets" to your graph at xy position(15000,15000) in orange and fontsize of 12

-Add text "Average Net Income" to your graph at xy position(175000,500) in red and fontsize of 12

-set the grid to white

In [24]:



1h. (1 point) Which company has the least return on assets?

In [6]:

Out[6]:

	Company Name	Total Assets (M$)	Net Income (M$)	# Employees	Return on Assets
123	SCHOOL SPECIALTY INC	638	-356	1919	-0.557994

1j. (1 point) Which company is the outlier on the plot? Hint: FInd the company with has the highest total assets?

In [8]:

Out[8]:

	Company Name	Total Assets (M$)	Net Income (M$)	# Employees
159	WAL-MA2:A111ART STORES INC	180663	16389	2100000

Problem 2: Data Analytics Jobs in the USA

Soon you will start getting ready to explore the job market for data analyst/data scientist positions (internship and full time). In this case study, we will assess the job market in the USA, and in particular, we are interested to learn which business sectors and companies are looking to hire data analysts in different US states.

The data set (`DataAnalyst.csv`) is available for download from blackboard. It is scrapped and cleaned from GlassDoor using this web scrapper.

The dataset has a sample of 2,253 job listings. The following table describes some of the variables necessary to answer the questions in this quiz:

Variables	Explanation
`Job Title`	listing's job title
`Job Description`	listing's job description
`Rating`	the company's rating on Glassdoor
`Company Name`	the listing company's name
`City`	city location of the company
`State`	state location of the company
`Size`	number of employees in the company
`Founded`	the year the company was founded
`Type of ownership`	is the company private, public, non-profit, etc.?
`Industry`	primary business activity
`Sector`	economic sector classification for the company
`Revenue`	company's income generated from business operations
`Competitors`	the company's list of competitors
`Min_Salary`	the minimum salary listing for the position
`Max_Salary`	the maximum salary listing for the position

In this homework, we assume that the sample of 2,253 job listings is a representative of the population of job listings in the USA.

In [1]:
```
# read the data
```

In [22]:
```
data.head()
```

Out[22]:

	Job Title	Job Description	Rating	Company Name	City	State	Size	Founded	Typ owner
0	Data Analyst, Center on Immigration and Justic...	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice	New York	NY	201 to 500 employees	1961.0	Nonp Organiza
1	Quality Data Analyst	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New York	NY	10000+ employees	1893.0	Nonp Organiza
2	Senior Data Analyst, Insights & Analytics Team...	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	New York	NY	1001 to 5000 employees	2003.0	Compa Pr
3	Data Analyst	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New York	NY	201 to 500 employees	2002.0	Subsi or Busi Segr
4	Reporting Data Analyst	ABOUT FANDUEL GROUP\n\nFanDuel Group is a worl...	3.9	FanDuel	New York	NY	501 to 1000 employees	2009.0	Compa Pr

2a. (1 point) What are the top 4 sectors with the highest count of job listings?

In [62]:

```
```

Out[62]:

```
Information Technology    570
Business Services         524
Finance                   169
Health Care               151
Name: Sector, dtype: int64
```

2b. (2 point) Suppose that you want to focus your job search in the following sectors (Information Technology, Business Services, Finance, Health Care). Create a subset of the given dataset that include only these 4 sectors with their data (include all variables).

Name the subset dataframe `mydata` .

In [29]:

```
```

Out[29]:

```
array(['Health Care', 'Information Technology', 'Finance',
       'Business Services'], dtype=object)
```

In [30]:

```
mydata.head()
```

Out[30]:

	Job Title	Job Description	Rating	Company Name	City	State	Size	Founded	Type ownershi
1	Quality Data Analyst	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New York	NY	10000+ employees	1893.0	Nonpro Organizatic
2	Senior Data Analyst, Insights & Analytics Team...	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	New York	NY	1001 to 5000 employees	2003.0	Company Priva
3	Data Analyst	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New York	NY	201 to 500 employees	2002.0	Subsidia or Busine Segme
5	Data Analyst	About Cubist\nCubist Systematic Strategies is ...	3.9	Point72	New York	NY	1001 to 5000 employees	2014.0	Company Priva
6	Business/ Data Analyst (FP&A)	Two Sigma is a different kind of investment ma...	4.4	Two Sigma	New York	NY	1001 to 5000 employees	2001.0	Company Priva

2c. (2 points) You are given the range of salary for each job listing (minimum and maximum salary). Add a new variable to `mydata` to estimate the salary of the for each of the listing in the dataset. The estimate salary is the average of the given minimum and maximum salary. #Hint Create a copy of the dataset mydata to avoid the "warning message"

Name the the new column `Est_Salary` .

What is the **average**, and **standard deviation** for the estimated salary among the 4 sectors listed in `mydata` dataframe?

In [43]:

Out[43]:

	Est_Salary	
	mean	std
Sector		
Business Services	72.135496	22.411196
Finance	67.644970	22.545747
Health Care	72.807947	26.554150
Information Technology	74.247368	25.520887

In [39]:

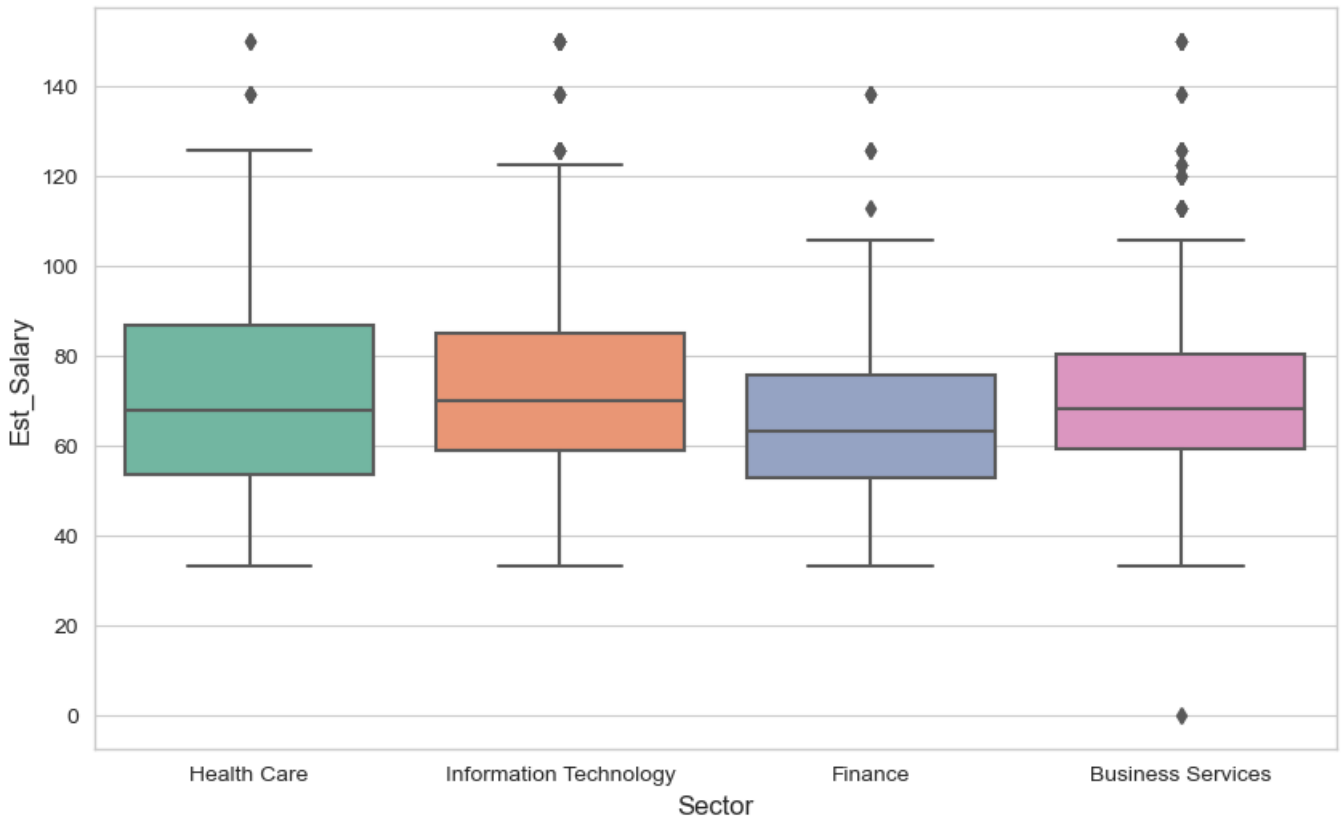
Out[39]:

	Job Title	Job Description	Rating	Company Name	City	State	Size	Founded	Type o ownershi
1	Quality Data Analyst	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New York	NY	10000+ employees	1893.0	Nonprof Organizatio
2	Senior Data Analyst, Insights & Analytics Team...	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace	New York	NY	1001 to 5000 employees	2003.0	Company Privat
3	Data Analyst	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New York	NY	201 to 500 employees	2002.0	Subsidiar or Busines Segmer

2d. (2 points) Create a side-by-side boxplot to show the distribution of salaries among the four hiring sectors (listed in `mydata`). Use "Set2 as the palette colors.

In [40]:

Out[40]:
```
<Axes: xlabel='Sector', ylabel='Est_Salary'>
```



(1 point) What does the boxplot tell you about the salaries in these industries for data analysts?

In [ ]:

2e. (2 points) List the company names (unique) in the **Information Technology** sector that has job postings with estimated salaries above 100K dollars?

In [70]:

Out[70]:
```
array(['Criteo', 'Tekfortune Inc.', 'Staffigo Technical Services, LLC',
       '8K Miles Software Services, Inc.', 'VTS',
       'RMS Computer Corporation', 'Reliable Software Resources',
       'Oracle', 'Avani Technology Solutions', 'Primesoft',
       'Systemart LLC', 'TechProjects', 'Information Technology Partners',
       'TikTok', 'Synchronous Solutions, Inc', 'HR Pundits',
       'Softpath System LLC', 'Motorola Solutions', 'Capgemini', 'NVIDIA',
       'Risk Management Solutions (RMS)', 'LeanData', 'Alteryx',
       'L&T Infotech', 'IntraEdge', 'Joomag, Inc.', 'Moveworks', 'Ursus',
       'Nuro', 'TalentBurst, Inc.', 'BayOne Solutions', 'Logic Planet',
       'Netflix', 'Diverse Lynx', 'Adwait Algorithm', 'Netflix, Inc.',
       'Apple', 'Collabera', 'Crystal Equation', 'Frontend Arts',
       'Poshmark', 'Zolon Tech Solutions Inc.', 'Lodestone', 'SAP',
       'Calsoft Labs', 'Coinbase', 'Trifacta', 'Wilbur Labs',
       'User Testing', 'Priceonomics', 'BOLD', 'Flatiron Health',
       'Twitter', 'Evolver, Inc.', 'Lyft', 'Scale AI', 'Softova Inc',
       'LeadStack', 'TaskRabbit'], dtype=object)
```

In [ ]:

2f. (2 points) List the company names (distinct) in the **Information Technology** or **Finance** sector that have job postings with estimated salaries above 100K dollars?

In [71]:

Out[71]:
```
array(['Criteo', 'Tekfortune Inc.', 'Intercontinental Exchange, Inc.',
       'Staffigo Technical Services, LLC',
       '8K Miles Software Services, Inc.', 'VTS',
       'RMS Computer Corporation', 'J.P. Morgan',
       'Sumitomo Mitsui Banking Corporation (SMBC)', 'Geller & Company',
       'Reliable Software Resources', 'The Bank of New York Mellon',
       'Oracle', 'Avani Technology Solutions', 'Primesoft',
       'Systemart LLC', 'TechProjects', 'Information Technology Partners',
       'TikTok', 'Synchronous Solutions, Inc', 'HR Pundits',
       'Softpath System LLC', 'Motorola Solutions', 'Capgemini', 'Tempus',
       'NVIDIA', 'Risk Management Solutions (RMS)', 'LeanData', 'Alteryx',
       'L&T Infotech', 'IntraEdge', 'Joomag, Inc.', 'Moveworks', 'Ursus',
       'Nuro', 'TalentBurst, Inc.', 'BayOne Solutions', 'Logic Planet',
       'Netflix', 'Diverse Lynx', 'Adwait Algorithm', 'Netflix, Inc.',
       'Apple', 'Collabera', 'Crystal Equation', 'Frontend Arts',
       'Poshmark', 'Zolon Tech Solutions Inc.', 'Lodestone', 'SAP',
       'Calsoft Labs', 'Veem', 'Coinbase', 'Trifacta', 'Wilbur Labs',
       'User Testing', 'Upstart', 'Credible', 'Priceonomics', 'BOLD',
       'Flatiron Health', 'The Voleon Group', 'Twitter',
       'Turn/River Capital', 'Evolver, Inc.', 'Lyft',
       'First Republic Bank', 'Scale AI', 'Softova Inc', 'LeadStack',
       'Chime', 'TaskRabbit'], dtype=object)
```

In [ ]:

2g. (2 points) Create a new variable, using Lambda, to re-classigy ownership into 'NonProfit' if the companies are "Nonprofit Organization" or, "College / University", "Govt" if they are "Government" and all others as "For Profit." Hint: Create a copy of the dataset mydata to avoid the "warning message"

In [41]:
```
#
```

In [42]:
```

```

Out[42]:

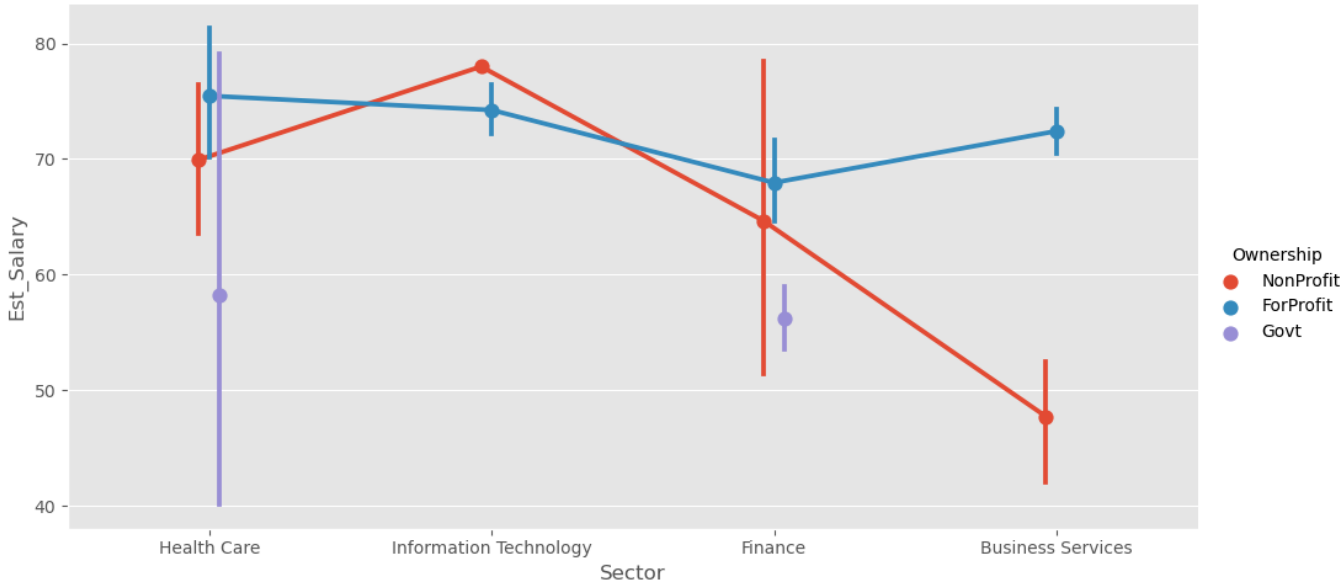
	Job Title	Job Description	Rating	Company Name	City	State	Size	Founded	Type ownershi
1	Quality Data Analyst	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York	New York	NY	10000+ employees	1893.0	Nonpro Organizatic
2	Senior Data Analyst, Insights & Analytics Team...	We’re looking for a Senior Data Analyst who ha...	3.4	Squarespace	New York	NY	1001 to 5000 employees	2003.0	Company Priva
3	Data Analyst	Requisition NumberRR-0001939\nRemote:Yes\nWe c...	4.1	Celerity	New York	NY	201 to 500 employees	2002.0	Subsidia or Busine Segme
5	Data Analyst	About Cubist\nCubist Systematic Strategies is ...	3.9	Point72	New York	NY	1001 to 5000 employees	2014.0	Company Priva
6	Business/ Data Analyst (FP&A)	Two Sigma is a different kind of investment ma...	4.4	Two Sigma	New York	NY	1001 to 5000 employees	2001.0	Company Priva

2h. (3 points) Using Seaborn, create a point-plot to show the Est_Salary by sectors(x axis) and distinguished by "Ownership."

In [44]:
```

```

Out[44]:
```
<seaborn.axisgrid.FacetGrid at 0x1a3141e3a50>
```



Which sector can be expected to have the greatest variation in estimated salaries, and which ownership-type has the gratest variation in estimated salaries?

2i. (3 points) Use the dataset with the 4 sectors (`mydata`) to create a dot plot (lollipop plot) that shows the top 15 states with the highest average salaries.

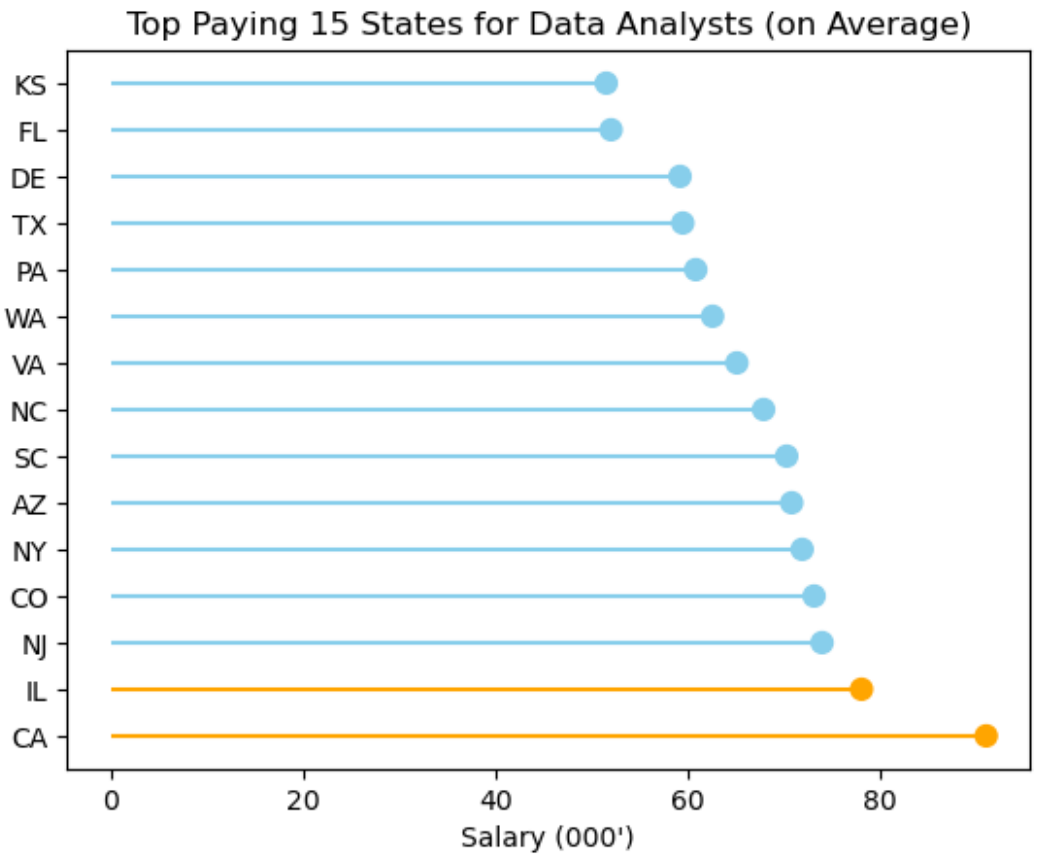
Name the dataframe `top15states`

The resulting dataframe should have two columns (`State` , `Avg Salary` *), where* `Avg Salary` *is the mean salary in the corresponding* `State`

Use two different colors of your choice to distingusih between the states with avegrage salary larger than $75K and thos with average salary less than $75K.

In [12]:

In [13]:


Top Paying 15 States for Data Analysts (on Average)
KS
FL
DE
TX
PA
WA
VA
NC
SC
AZ
NY
CO
NJ
IL
CA
0
20
40
60
80
Salary (000')

In [38]:

In [ ]: