Job Salary Prediction

Data Robot

Dhanashree Mane
Sanya Purwar
Arjoo Gangwal

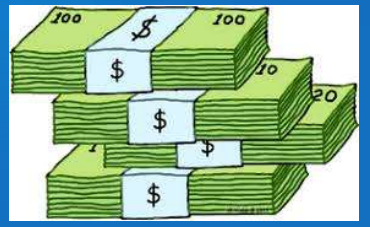
What are we doing?

- Salary is a **big deciding factor** for a majority of people
- We are trying to build a model that would **predict the average salary based on**
  - **Job title**
  - **Location**
  - **Years of experience**
  - **Company Size**
  - **Company Sector**

Sanya has 2 years of experience in Software Industry

Offer 1:
Location: San Francisco
Company-Size: Small-size
Company-Sector: Software/IT

Offer 2:
Location: Washington
Company-Size: Mid-size
Company-Sector: Software/IT



Project Pipeline

Collecting the data

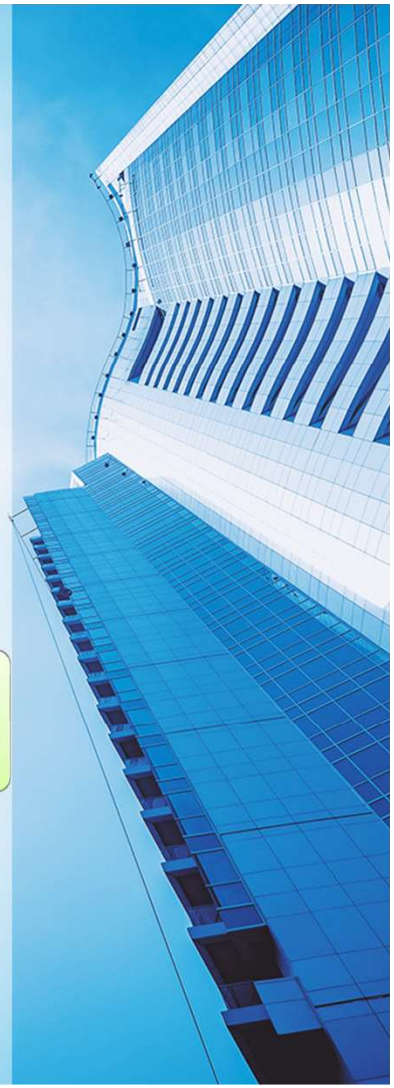
- We scraped data from Indeed
- Used BeautifulSoup to scrape.

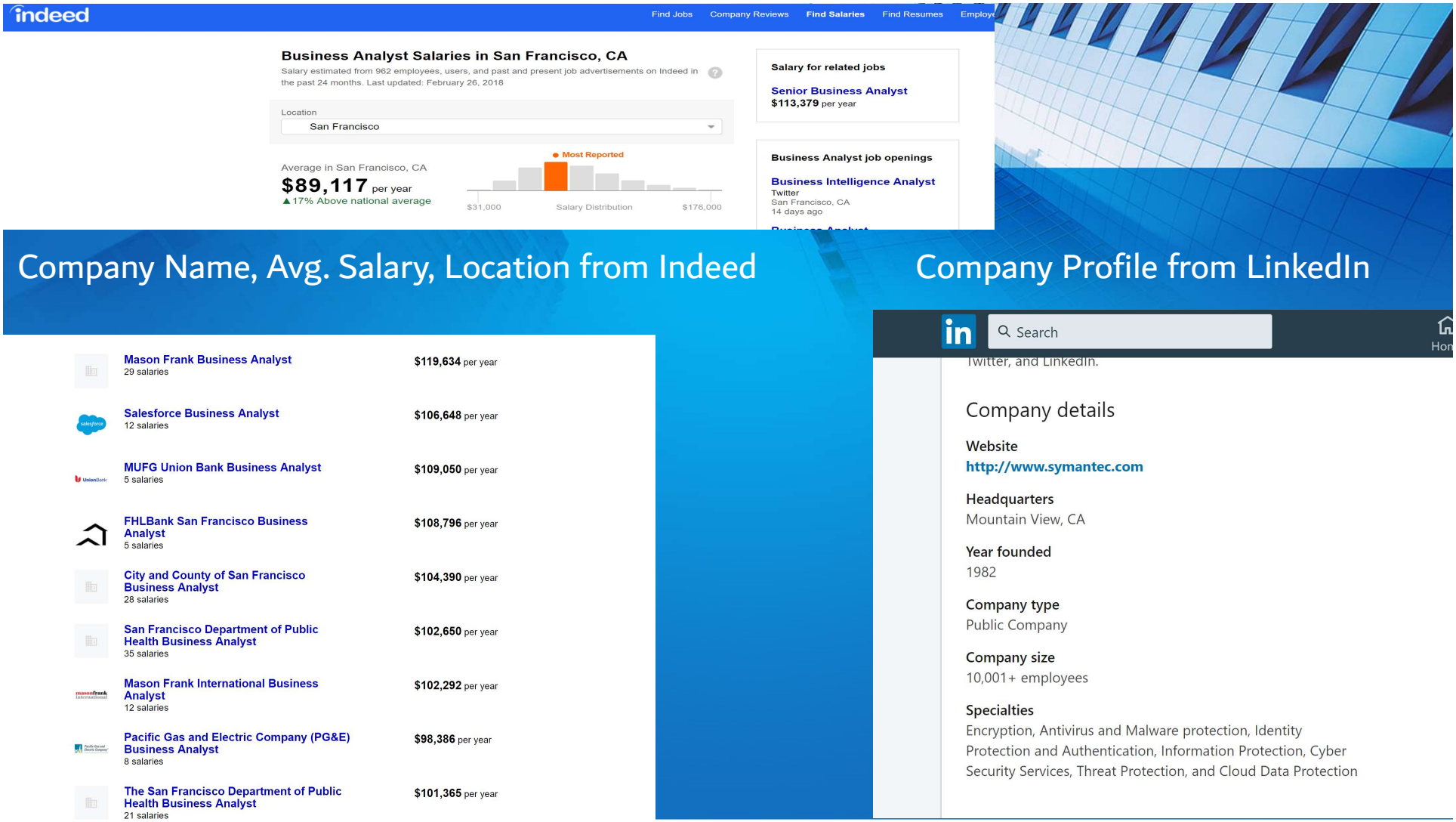
Data Cleaning

- The Job title, Location and Salaries were mapped to the company size and Sector.
- We handled Outliers and dealt with incomplete rows.

Data Correlation

- We used **Multiple Linear Regression** to fit our model.
- Features incorporated: Location, Years of experience, Company size and Sector.




indeed
Find Jobs
Company Reviews
Find Salaries
Find Resumes
Business Analyst Salaries in San Francisco, CA
Salary estimated from 962 employees, users, and past and present job advertisements on Indeed in the past 24 months. Last updated: February 26, 2018
Location
San Francisco
Average in San Francisco, CA
$89,117 per year
▲17% Above national average
Most Reported
$31,000
Salary Distribution
$176,000
Salary for related jobs
Senior Business Analyst
$113,379 per year
Business Analyst job openings
Business Intelligence Analyst
Twitter
San Francisco, CA
14 days ago
Company Name, Avg. Salary, Location from Indeed
Company Profile from LinkedIn
Mason Frank Business Analyst
29 salaries
$119,634 per year
Salesforce Business Analyst
12 salaries
$106,648 per year
MUFG Union Bank Business Analyst
5 salaries
$109,050 per year
FHLBank San Francisco Business Analyst
5 salaries
$108,796 per year
City and County of San Francisco Business Analyst
28 salaries
$104,390 per year
San Francisco Department of Public Health Business Analyst
35 salaries
$102,650 per year
Mason Frank International Business Analyst
12 salaries
$102,292 per year
Pacific Gas and Electric Company (PG&E) Business Analyst
8 salaries
$98,386 per year
The San Francisco Department of Public Health Business Analyst
21 salaries
$101,365 per year
Search
Twitter, and LinkedIn.
Company details
Website
http://www.symantec.com
Headquarters
Mountain View, CA
Year founded
1982
Company type
Public Company
Company size
10,001+ employees
Specialties
Encryption, Antivirus and Malware protection, Identity Protection and Authentication, Information Protection, Cyber Security Services, Threat Protection, and Cloud Data Protection

Challenges We Faced

1

Had to manually map the company name with their respective company size and sector by taking information about each company name from LinkedIn

2

Had to manually match the salary with the Years of experience taking references from Glassdoor.com

Our Data Looks like

	Company Name	Salary	Years of Experience	Location	Company Size	Sector	Job Title
2	3DI INC	167000	7 to 10	Los Angeles	Small-size	Software/IT	Business Analyst
3	3DI INC	123683	4 to 6	Los Angeles	Small-size	Software/IT	Business Analyst
4	3DI INC	80366	1 to 3	Los Angeles	Small-size	Software/IT	Business Analyst
5	3EDGEUSAGROUP LLC	121100	7 to 10	New York	Small-size	LAW	Business Analyst
6	3EDGEUSAGROUP LLC	101727.5	4 to 6	New York	Small-size	LAW	Business Analyst
7	3EDGEUSAGROUP LLC	82355	1 to 3	New York	Small-size	LAW	Business Analyst
8	4C Connect Inc.	149950	7 to 10	Atlanta	Small-size	Accounting	Business Analyst
9	4C Connect Inc.	112475	4 to 6	Atlanta	Small-size	Accounting	Business Analyst
10	4C Connect Inc.	75000	1 to 3	Atlanta	Small-size	Accounting	Business Analyst
11	A.T.KearneyInc.	150000	7 to 10	Atlanta	Mid-size	Consulting	Business Analyst
12	A.T.KearneyInc.	110731.5	4 to 6	Atlanta	Mid-size	Consulting	Business Analyst
13	A.T.KearneyInc.	71463	1 to 3	Atlanta	Mid-size	Consulting	Business Analyst
14	Absolute Opportunities	151920	7 to 10	Washington	Small-size	Education	Business Analyst
15	Absolute Opportunities	127960	4 to 6	Washington	Small-size	Education	Business Analyst
16	Absolute Opportunities	104000	1 to 3	Washington	Small-size	Education	Business Analyst
17	Acadia Technologies	140000	7 to 10	Atlanta	Small-size	Software/IT	Business Analyst
18	Acadia Technologies	97500	4 to 6	Atlanta	Small-size	Software/IT	Business Analyst
19	Acadia Technologies	55000	1 to 3	Atlanta	Small-size	Software/IT	Business Analyst

Why Multiple linear regression?

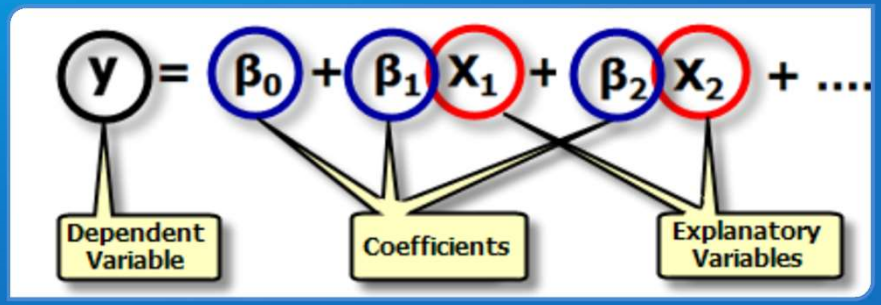
Location

Years of Experience

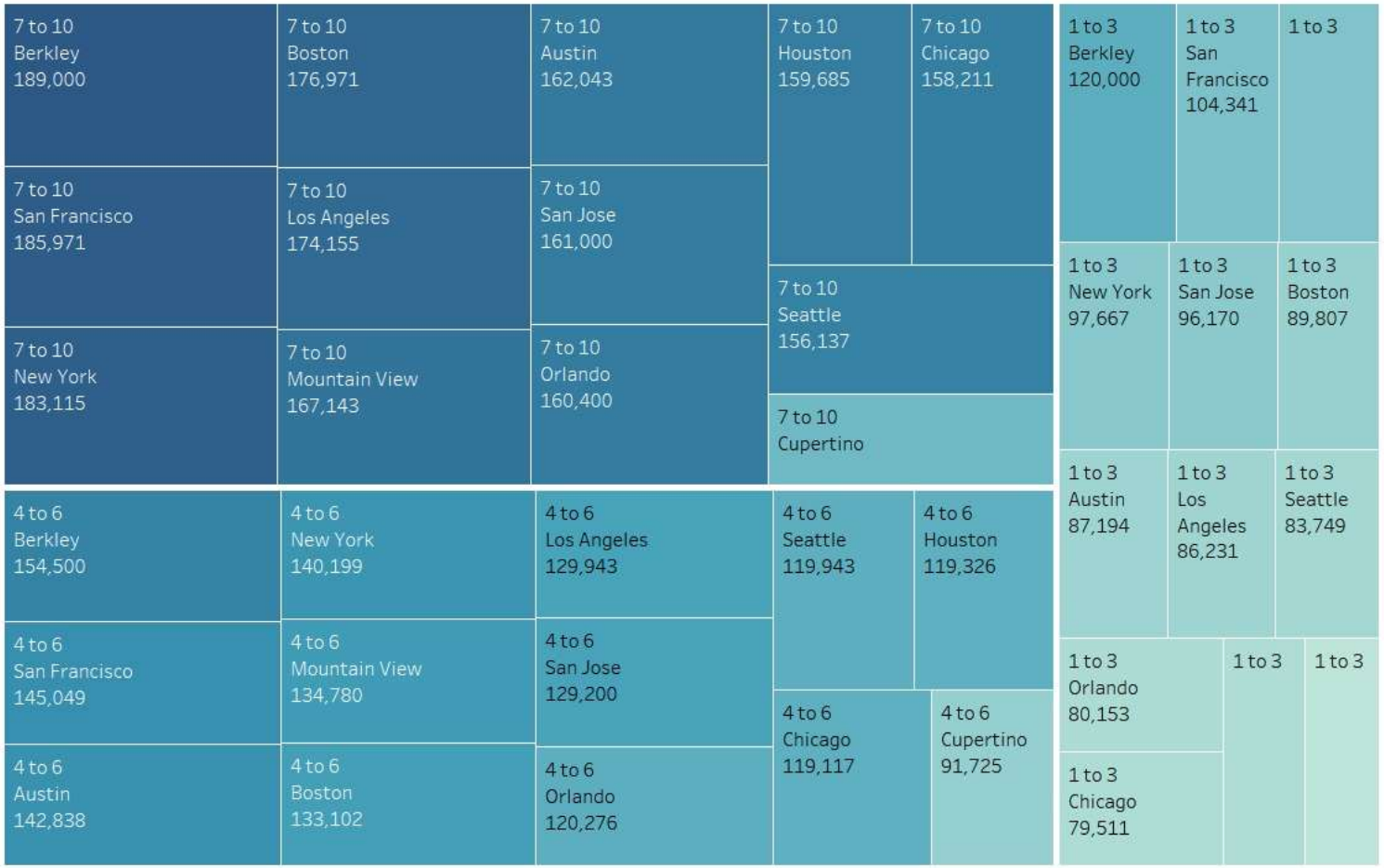
Company Size

Company Sector

- Independent Variables(X1,X2,..Xn): Location, Yrs. of Experience, Company Size, Sector.
- Dependent Variable(y): Salary



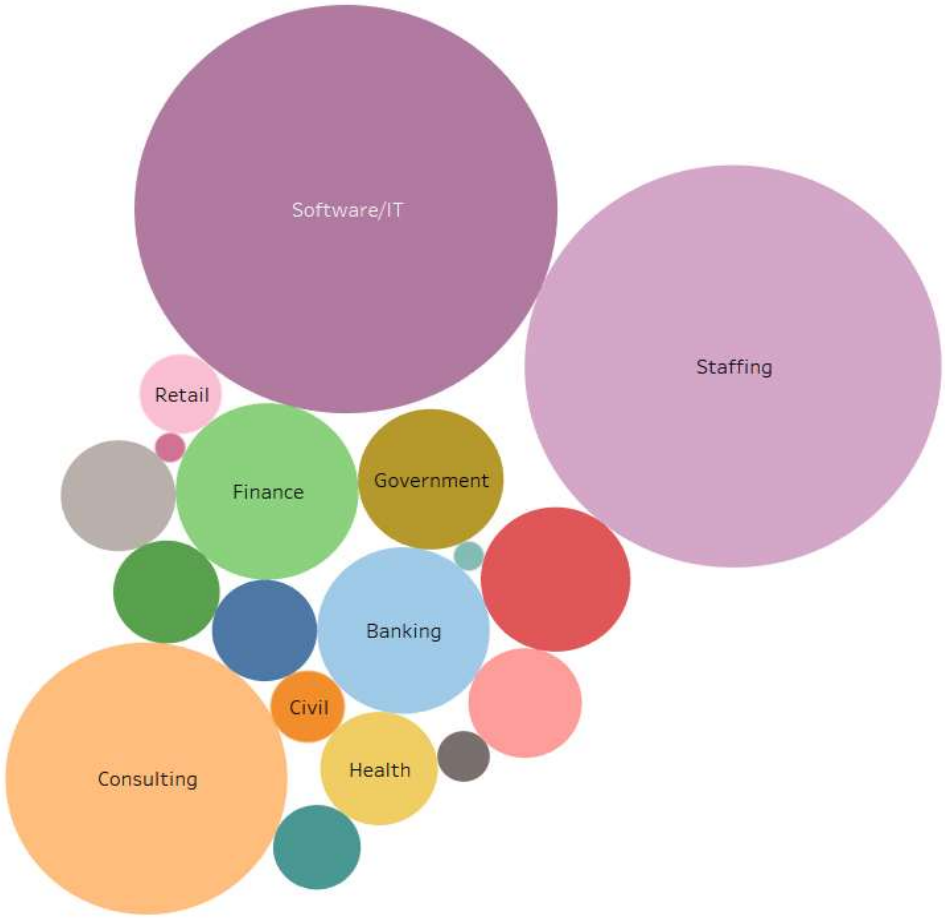
Location and Yrs of Experience wise Salary



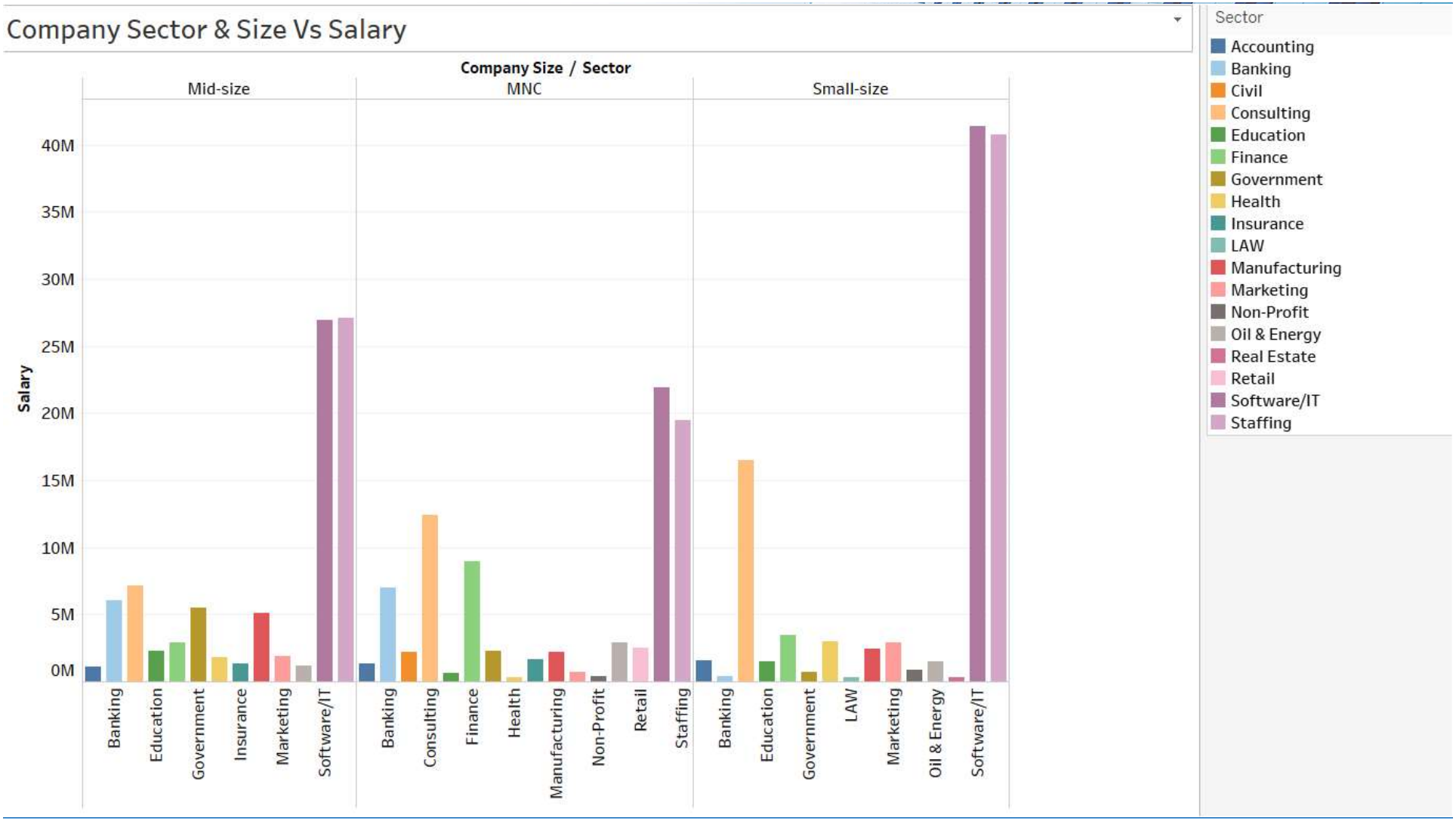
AVG(Salary)

72,750 189,000

Sector Vs Salary



Sector
- Accounting
- Banking
- Civil
- Consulting
- Education
- Finance
- Government
- Health
- Insurance
- LAW
- Manufacturing
- Marketing
- Non-Profit
- Oil & Energy
- Real Estate
- Retail
- Software/IT
- Staffing


Company Sector & Size Vs Salary
Company Size / Sector
Mid-size
MNC
Small-size
Salary
40M
35M
30M
25M
20M
15M
10M
5M
0M
Banking
Education
Government
Insurance
Marketing
Software/IT
Banking
Consulting
Finance
Health
Manufacturing
Non-Profit
Retail
Staffing
Banking
Education
Government
LAW
Marketing
Oil & Energy
Software/IT
Sector
Accounting
Banking
Civil
Consulting
Education
Finance
Government
Health
Insurance
LAW
Manufacturing
Marketing
Non-Profit
Oil & Energy
Real Estate
Retail
Software/IT
Staffing

Company Location

In [10]:
```
# Mapping the Location

city = {"Austin": 0, "Berkley":1, "Boston": 2, "Chicago": 3, "Cupertino": 4, "Houston": 5, "Los Angeles
        "Mountain View":7, "New York": 8, "Orlando": 9, "San Francisco":10, "San Jose": 11, "Seattle":
       "Cleveland": 14, "MenloPark": 15, "OAKLAND": 16, "PaloAlto": 17, "REDWOODCITY": 18, "Sacramento"
       "SanMateo": 21, "Sunnyvale": 22, "Washington": 23, "SantaClara": 24, "Minneapolis": 25}
data["Location"] = data["Location"].map(city)
data.head()
```

Out[10]:

	Company Name	Salary	Years of Experience	Location	Company Size	Sector	Job Title
0	3DI INC	167000.0	2	6	Small-size	Software/IT	Business Analyst
1	3DI INC	123683.0	1	6	Small-size	Software/IT	Business Analyst
2	3DI INC	80366.0	0	6	Small-size	Software/IT	Business Analyst
3	3EDGEUSAGROUP LLC	121100.0	2	8	Small-size	LAW	Business Analyst
4	3EDGEUSAGROUP LLC	101727.5	1	8	Small-size	LAW	Business Analyst

Company Size

Out[234]:

	Company Name	Salary	Years of Experience	Location	Company Size	Sector	Job Title
0	3DI INC	167000.0	2	6	Small-size	Software/IT	Business Analyst
1	3DI INC	123683.0	1	6	Small-size	Software/IT	Business Analyst
2	3DI INC	80366.0	0	6	Small-size	Software/IT	Business Analyst
3	3EDGEUSAGROUP LLC	121100.0	2	8	Small-size	LAW	Business Analyst
4	3EDGEUSAGROUP LLC	101727.5	1	8	Small-size	LAW	Business Analyst

In [235]:

```
# Mapping the Company Size

size = {'Small-size': 0, 'Mid-size': 1, 'MNC': 2}
data['Company Size'] = data['Company Size'].map(size)
data.head()
```

Out[235]:

	Company Name	Salary	Years of Experience	Location	Company Size	Sector	Job Title
0	3DI INC	167000.0	2	6	0	Software/IT	Business Analyst
1	3DI INC	123683.0	1	6	0	Software/IT	Business Analyst
2	3DI INC	80366.0	0	6	0	Software/IT	Business Analyst
3	3EDGEUSAGROUP LLC	121100.0	2	8	0	LAW	Business Analyst
4	3EDGEUSAGROUP LLC	101727.5	1	8	0	LAW	Business Analyst

Sector

In [12]:
```
# Mapping the Sector

sector = {"Oil & Energy": 0, "Accounting": 1, "Banking": 2, "Civil": 3, "Consulting":4, "Education":5,
          "Government":7, "Health": 8, "Insurance": 9, "LAW": 10, "Manufacturing": 11, "Marketing": 12,
          "Software/IT": 14, "Staffing": 15, "Real Estate": 16, "Retail": 17}
data["Sector"] = data["Sector"].map(sector)
data.head()
```

Out[12]:

	Company Name	Salary	Years of Experience	Location	Company Size	Sector	Job Title
0	3DI INC	167000.0	2	6	0	14	Business Analyst
1	3DI INC	123683.0	1	6	0	14	Business Analyst
2	3DI INC	80366.0	0	6	0	14	Business Analyst
3	3EDGEUSAGROUP LLC	121100.0	2	8	0	10	Business Analyst
4	3EDGEUSAGROUP LLC	101727.5	1	8	0	10	Business Analyst

Split the data for training and testing

In [30]:
```
feature_col = ["Location", "Years of Experience","Company Size","Sector" ]
X = data[feature_col]
y = data["Salary"]

xtrain, xtest, ytrain, ytest = train_test_split(X, y, test_size=0.10, random_state = 52)

train_x = xtrain
train_y = ytrain
test_x = xtest
test_y = ytest

from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
train_x = sc.fit_transform(train_x)
test_x= sc.transform(test_x)
```

Train the model

In [31]:
```
regr = linear_model.LinearRegression()

# Train the model using the training sets
clf = regr.fit(train_x, train_y)

# Make predictions using the testing set
Sal_y_pred = regr.predict(test_x)

#Accuracy:
clf.score(test_x,test_y)
```

Out[31]: 0.70769752746380643

Model Accuracy : 70.76%

Predicting Salary

Prediction

In [15]:
```
# Location: Mountain View, Yrs of exp: 4 to 6, size: MNC, Sector: Consulting

regr.predict([[7,1,2,4]])
```

Out[15]:
```
array([ 142767.42008853])
```

Save the model in Pickle

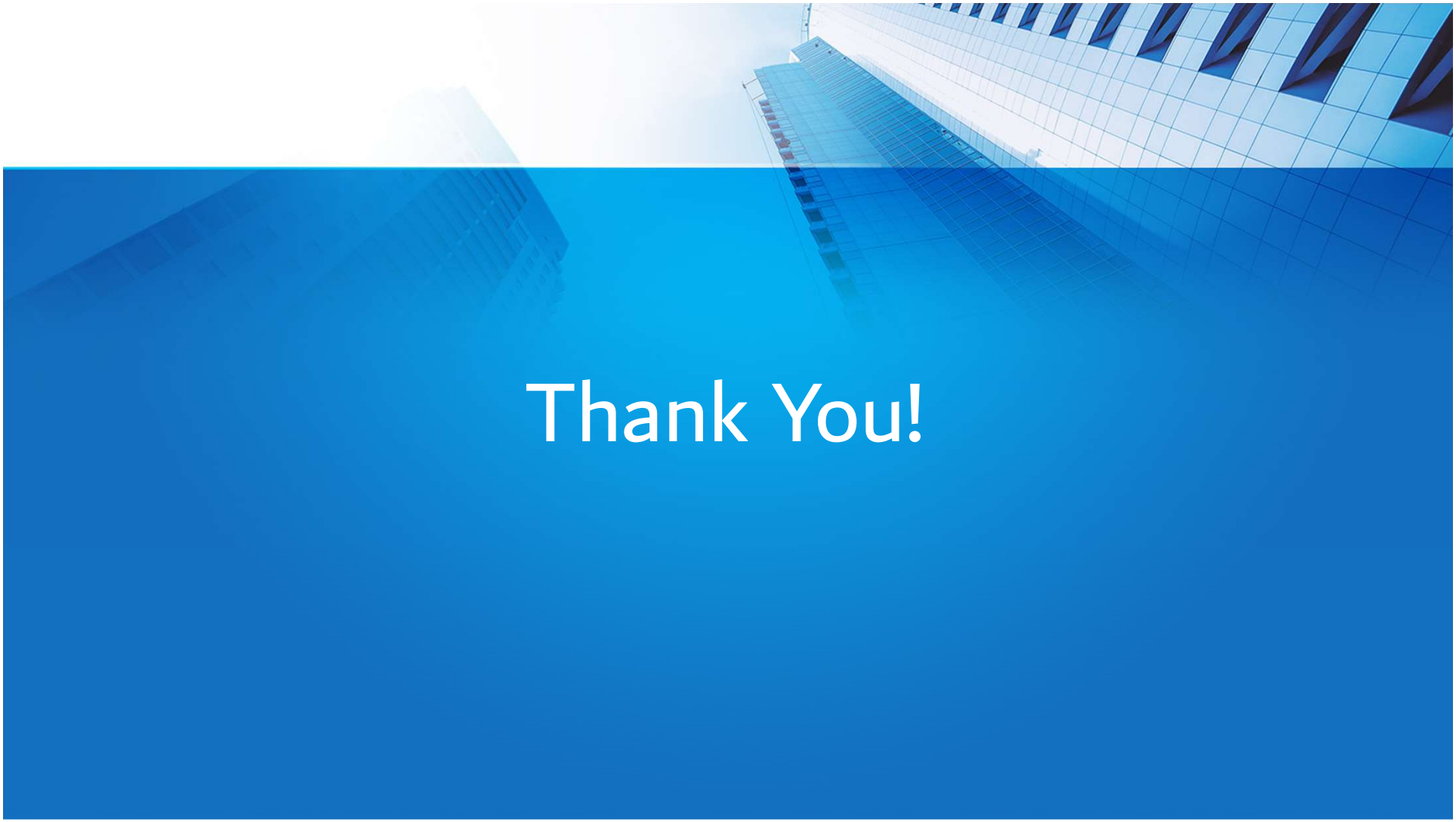
In [ ]:
```
# Pickling the model
```

In [32]:
```
import pickle

with open("python_lin_reg_model.pkl", "wb") as file_handler:
    pickle.dump(regr, file_handler)
with open("python_lin_reg_model.pkl", "rb") as file_handler:
    loaded_pickle = pickle.load(file_handler)
```

Future Enhancements

- Currently we have data for **Business Analyst** only.
- Addition of Job titles like **Project Manager, Software Engineer, Operations Manager, Data Scientist.**
- Currently we have collected the salaries in **26 cities.**
- Addition of more cities.


Thank You!