DATA ANALYTICS

SALARIES PREDICTION

University of Miami
Data Analytics Bootcamp

OUR PROJECT

We aim to make a prediction about the salaries that the graduates from the Data Analytics Bootcamp could obtain, taking in consideration the years of experience and each company’s salary level.

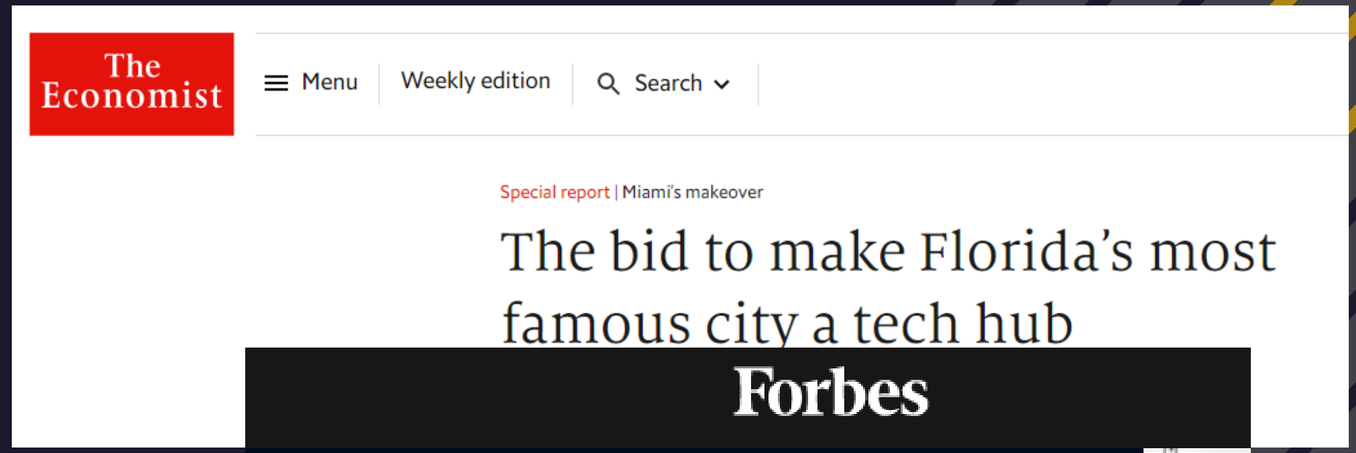
University of Miami

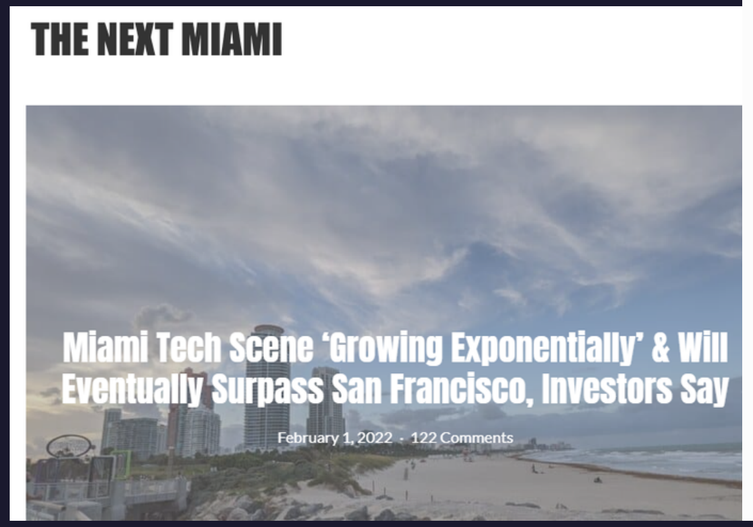
GUIDING QUESTIONS

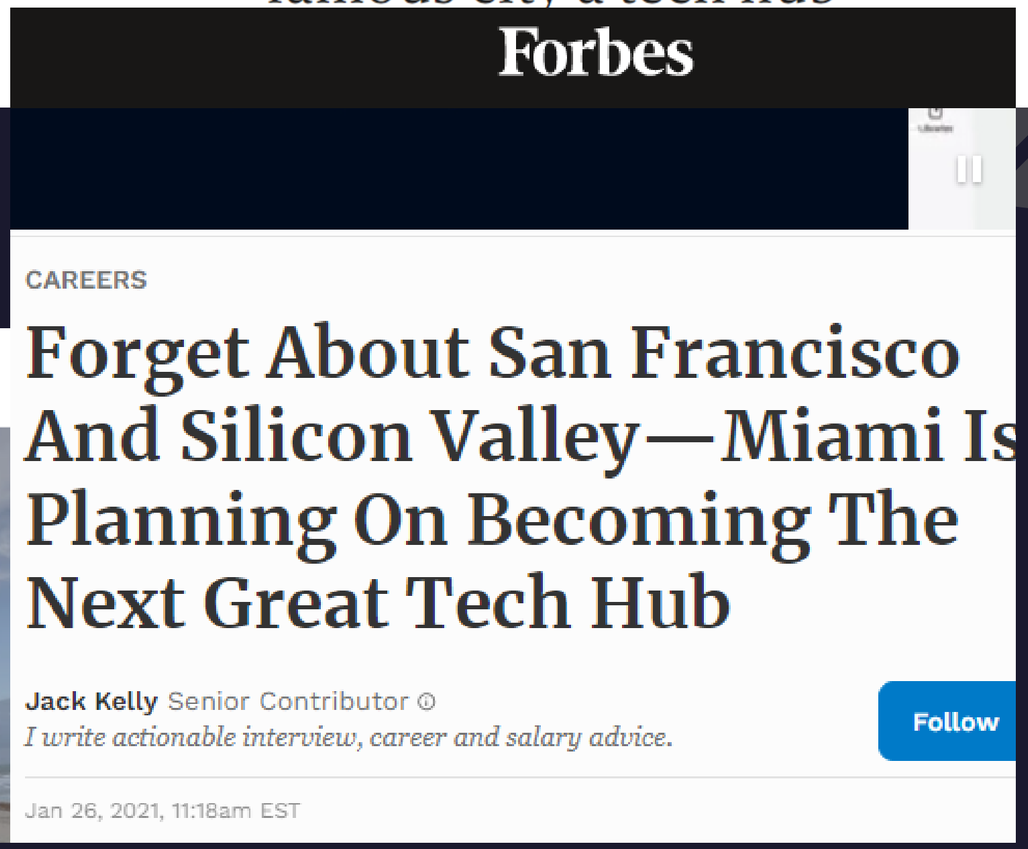
- Based of their experience, what salary can the graduates expect?
- What would that salary be in the Florida area?
- In the case of a remote job, what would the salary be for other estates in the US?
- Do the years of experience at the same company have an effect on the salary increase?

Note

We weren't able to analyze Data Science salaries within the Florida State, since there were very few entries related to this state.
However, recent news articles suggest that Florida is becoming more attractive for Tech companies.

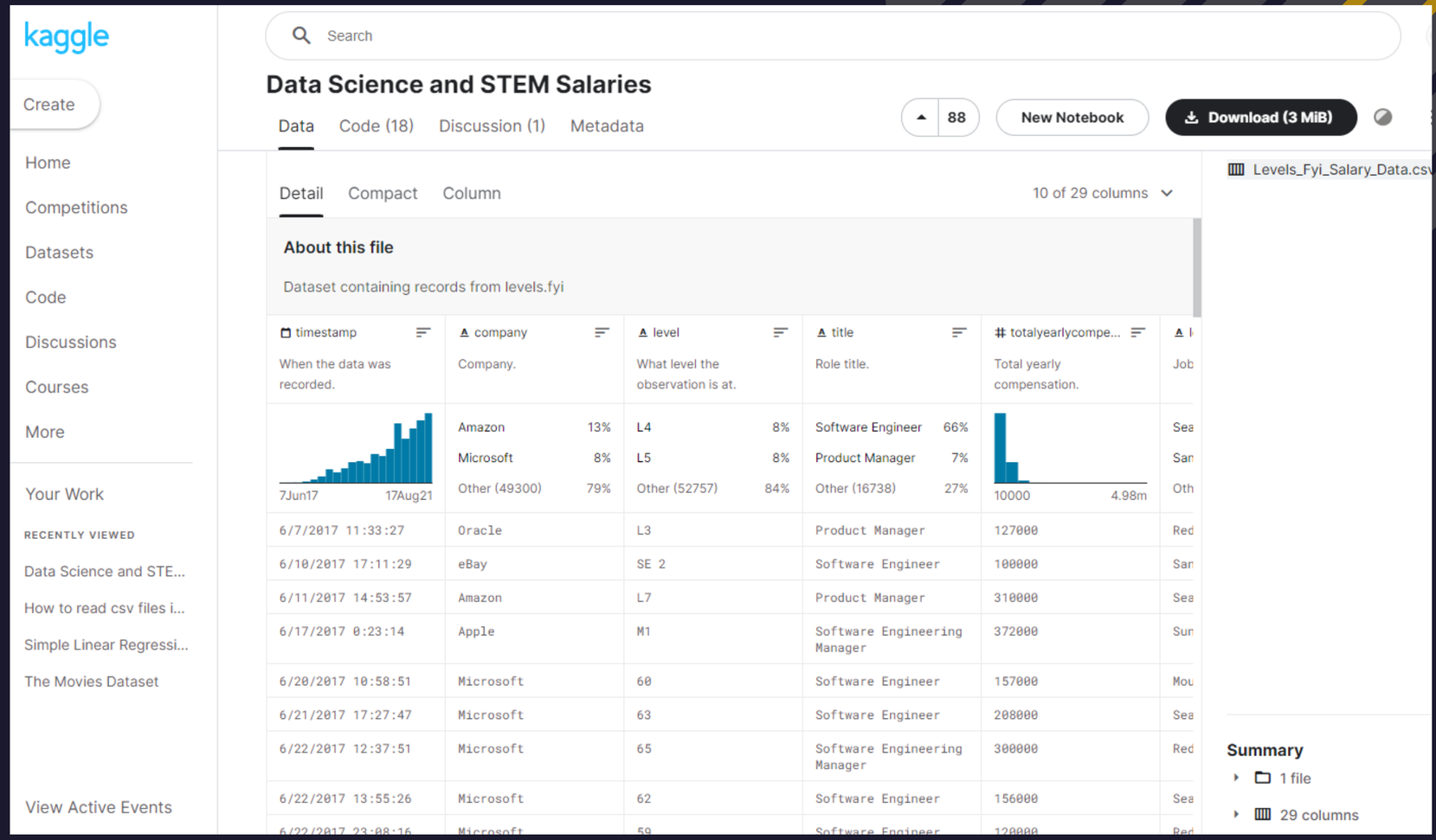






THE DATASET

The source of the data is Kaggle.com. The raw data contains information about more than 60k employees from top tech companies. This significant amount of data makes our analysis possible. Not only does it include information about the base salary, but also about bonus, stocks, years of experience, company salary levels and location.



University of Miami

```
'Gender' columns has 68% of missing value

alaries_df.drop(['gender'], axis = 1, inpl

'cityid' column is not relevant

alaries_df.drop(['cityid'], axis = 1, inpl

'rowNumber' column is not relevant

alaries_df.drop(['rowNumber'], axis = 1, i

'dmaid' column is not relevant

alaries_df.drop(['dmaid'], axis = 1, inpla

drop all columns related to race, as it i

alaries_df.drop(['Race_Asian'], axis = 1,
alaries_df.drop(['Race_White'], axis = 1,
alaries_df.drop(['Race_Two_Or_More'], axis
alaries_df.drop(['Race_Black'], axis = 1,
alaries_df.drop(['Race_Hispanic'], axis =
alaries_df.drop(['Race'], axis = 1, inplac
```

Delete Columns

```
1 salaries_df.head()
```

	timestamp	company	level	ti
0	6/7/2017 11:33	Oracle	L3	Prod Manag
1	6/10/2017 17:11	eBay	SE 2	Softwa Engine
2	6/11/2017 14:53	Amazon	L7	Prod Manag
3	6/17/2017 0:23	Apple	M1	Softwa Engineeri Manag
4	6/20/2017 10:58	Microsoft	60	Softwa Engine

rows × 28 columns

Standarize

Year_Column	INT
Company	VARCHAR
City	VARCHAR
Title	VARCHAR
Salary_Level	INT
Area	VARCHAR
Total_Annual_Compensation	INT
Years_of_Experience	NUMERIC
Years_At_Company	NUMERIC
Base_Salary	INT
Stock_Grant_Value	INT
Bonus	INT
Masters_Degree	INT
Bachelors_Degree	INT
Doctorate_Degree	INT

Create SQL Tables

```
from sqlalchemy import create_engine
from config import db_password
```

```
# Connect to SQL Database

db_string = f"postgresql://postgres:{db_pa
engine = create_engine(db_string)
```

```
# Create a pandas df from the SQL table, w

ba_salaries_df = pd.read_sql_table(
    "ba_level1_salaries",
    con=engine)
```

```
ba_salaries_df.head()
```

	year_column	company	city	state1	title	s
0	2020	UBS	Krakow	MA	Business Analyst	
1	2021	HCA Healthcare	Nashville	TN	Business Analyst	
2	2021	Qualtrics	Provo	UT	Business Analyst	
3	2021	Clearwater	Boise	ID	Business	

Join SQL with Pandas

DATA CLEANING

```
pd.DataFrame(np.c_[Y_test , pred , diff] , columns=['Actual','Predicted','Difference'])
```

	Actual	Predicted	Difference
0	1716000.0	508067.359096	1.207933e+06
1	411000.0	245029.870544	1.659701e+05
2	333000.0	288704.143903	4.429586e+04
3	410000.0	221484.259844	1.885157e+05
4	140000.0	179858.633920	-3.985863e+04
...	...	...	...
10563	242000.0	243810.184647	-1.810185e+03
10564	140000.0	224492.243903	-8.449224e+04
10565	152000.0	203457.774601	-5.145777e+04
10566	239000.0	185978.482513	5.302152e+04
10567	210000.0	175630.963964	3.436904e+04

10568 rows × 3 columns

```
lr.score(X_test , Y_test)
```

```
0.22222032534605973
```

```
mean_squared_error(Y_test , pred, squared=False)
```

```
124042.05187670092
```

```
r2_score(Y_test , pred)
```

```
0.22222032534605973
```

MACHINE LEARNING

Linear Regression Model on complete dataset

University of Miami

```
pd.DataFrame(np.c_[Y_test , pred , diff] , columns=['Actual','Predicted','Difference'])
```

	Actual	Predicted	Difference
0	150000	80000	70000
1	96000	90000	6000
2	142000	350000	-208000
3	96000	92000	4000
4	400000	173000	227000
...	...	...	...
243	81000	90000	-9000
244	138000	138000	0
245	153000	120000	33000
246	167000	191000	-24000
247	87000	125000	-38000

248 rows × 3 columns

```
rf.score(X_test , Y_test)
```

```
0.020161290322580645
```

```
mean_squared_error(Y_test , pred, squared=False)
```

```
62304.40361384915
```

```
r2_score(Y_test , pred)
```

```
-0.10145820676197603
```

MACHINE LEARNING

Random Forest Model on Data Science positions

University of Miami

```
pd.DataFrame(np.c_[Y_test , pred , diff] , columns=['Actual','Predicted','Difference'])
```

	Actual	Predicted	Difference
0	130000	296000	-166000
1	130000	108000	22000
2	100000	185000	-85000
3	165000	130000	35000
4	169000	139000	30000
...	...	...	...
94	107000	155000	-48000
95	80000	80000	0
96	85000	78000	7000
97	131000	275000	-144000
98	125000	110000	15000

99 rows × 3 columns

```
rf.score(X_test , Y_test)
```

```
0.030303030303030304
```

```
mean_squared_error(Y_test , pred, squared=False)
```

```
56574.97037318953
```

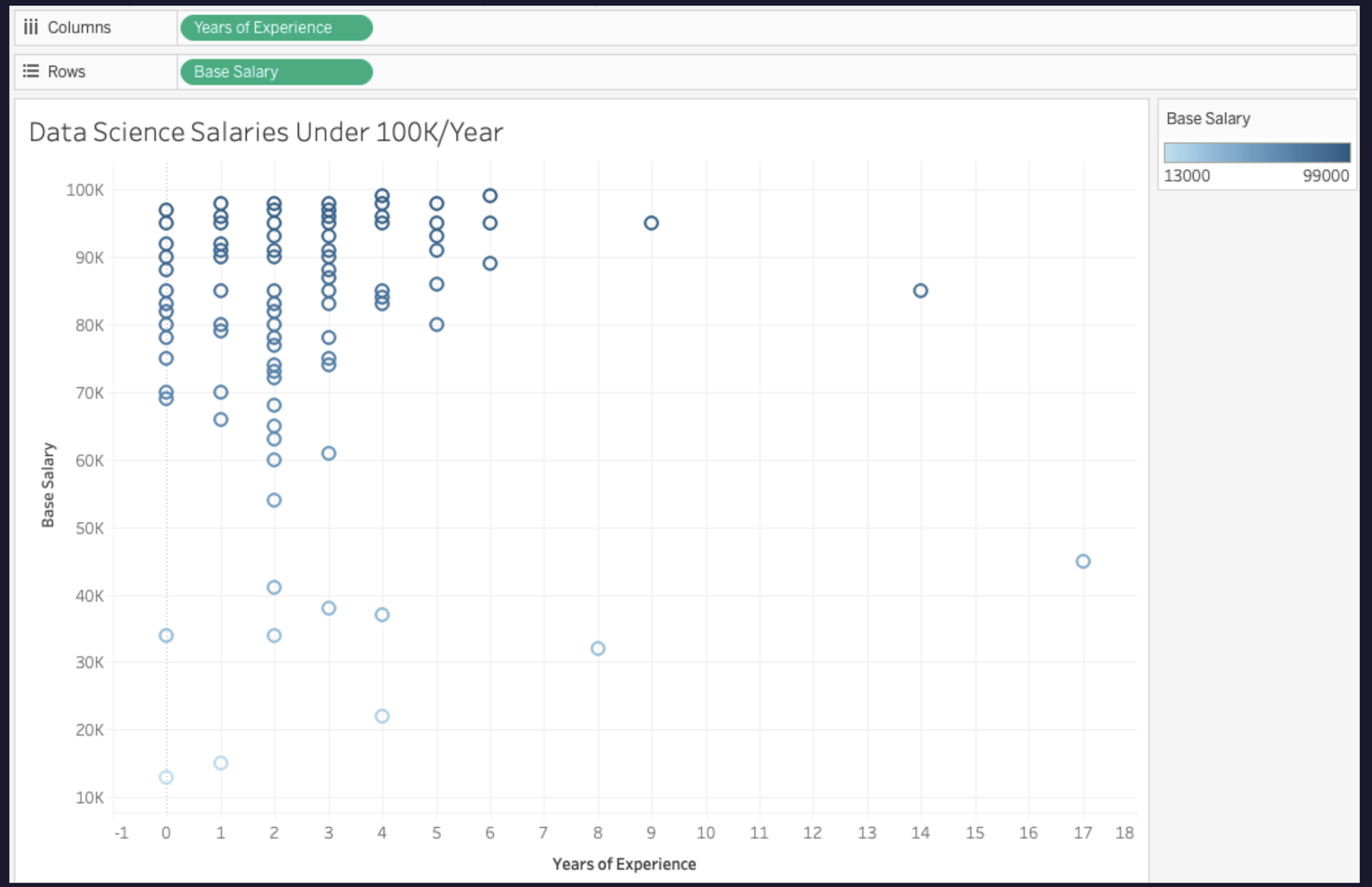
```
r2_score(Y_test , pred)
```

```
-0.03185276782647617
```

MACHINE LEARNING

Random Forest Model on Business Analyst positions

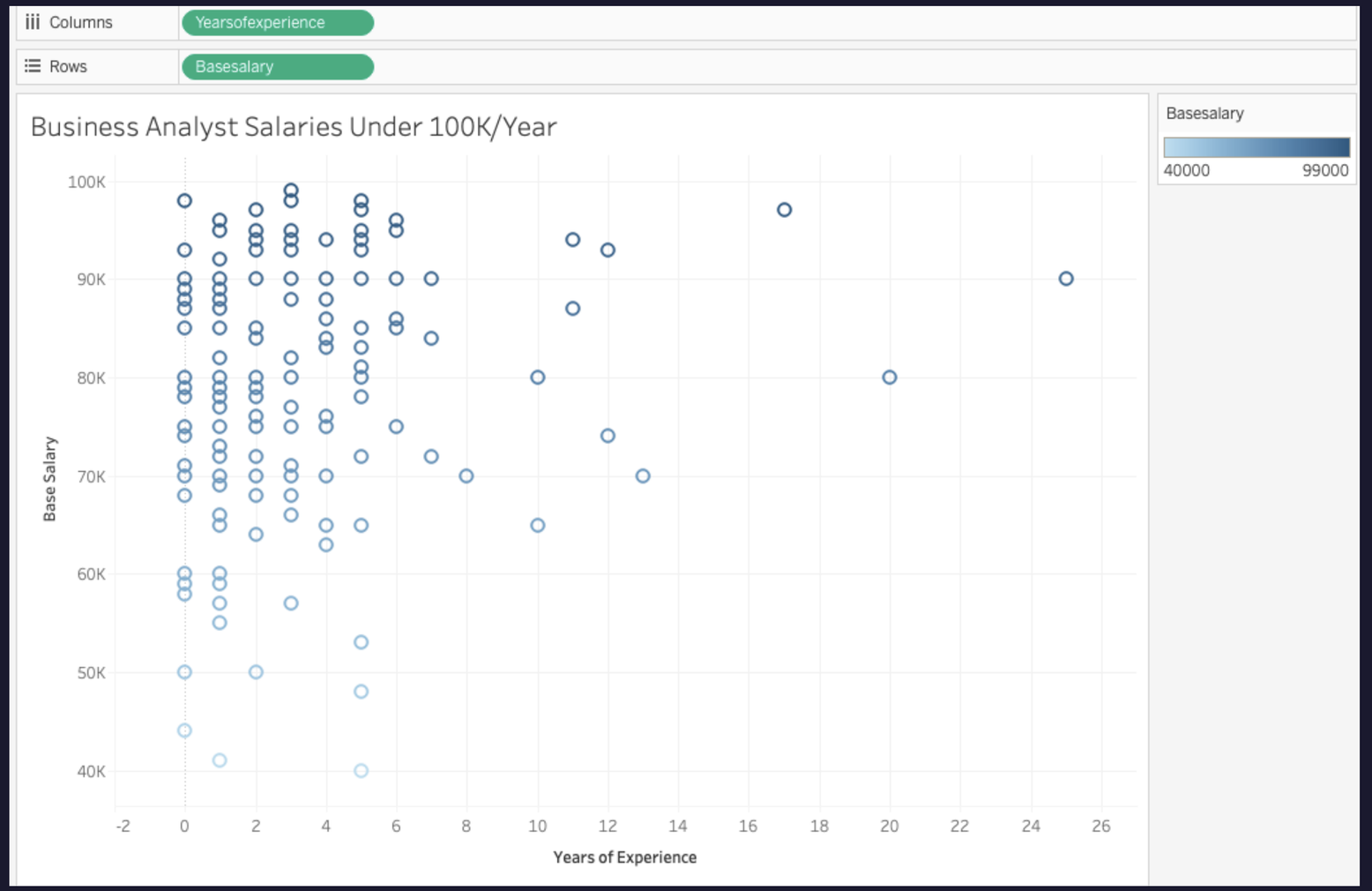
University of Miami


Columns
Years of Experience
Rows
Base Salary
Data Science Salaries Under 100K/Year
Base Salary
13000
99000
100K
90K
80K
70K
60K
50K
40K
30K
20K
10K
Base Salary
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Years of Experience

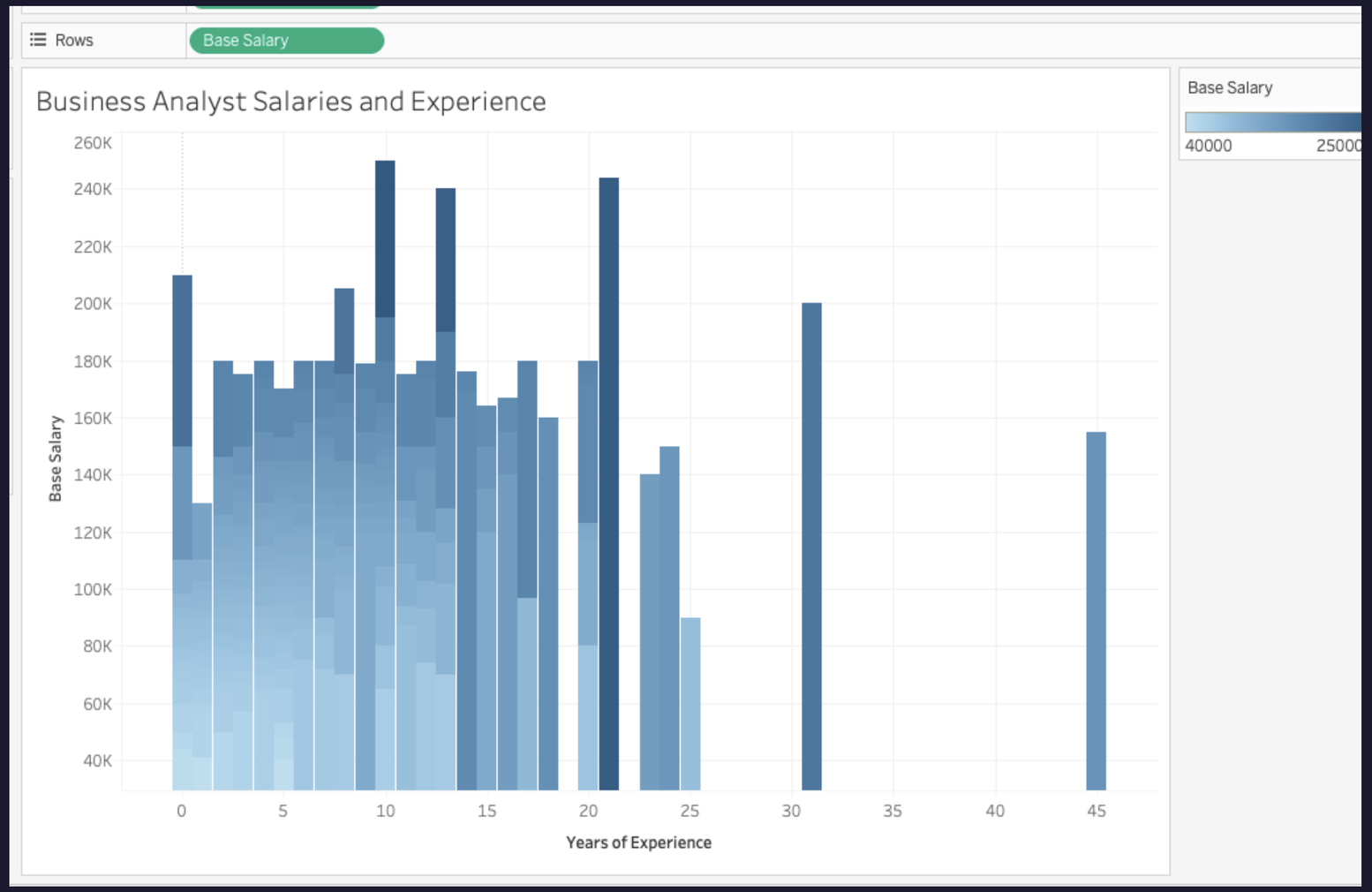
Tableau


Columns
Years of Experience
Rows
Base Salary
Data Science Salaries Jobs
Base Salary
600K
500K
400K
300K
200K
100K
0K
-2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
Years of Experience
Base Salary
13000
630000

Tableau


Columns
Yearsofexperience
Rows
Basesalary
Business Analyst Salaries Under 100K/Year
Basesalary
40000
99000
100K
90K
80K
70K
60K
50K
40K
Base Salary
-2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
Years of Experience

Tableau


Rows
Base Salary
Business Analyst Salaries and Experience
Base Salary
40000
25000
260K
240K
220K
200K
180K
160K
140K
120K
100K
80K
60K
40K
Base Salary
0
5
10
15
20
25
30
35
40
45
Years of Experience

Tableau

THANK YOU!

University of Miami