



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



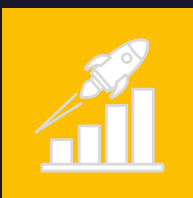
GUIDING QUESTIONS



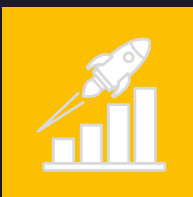
Based on their experience, what salary can the graduates expect?



What would that salary be in the Florida area?



In the case of a remote job, what would the salary be for other states in the US?



Do the years of experience at the same company have an effect on the salary increase?

**NOW
HIRING**
DATA ANALYST

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 NEXT MIAMI



Miami Tech Scene 'Growing Exponentially' & Will Eventually Surpass San Francisco, Investors Say

February 1, 2022 · 122 Comments

The Economist

Menu Weekly edition Search

Special report | Miami's makeover

The bid to make Florida's most famous city a tech hub

Forbes

CAREERS

Forget About San Francisco And Silicon Valley—Miami Is Planning On Becoming The Next Great Tech Hub

Jack Kelly Senior Contributor @
I write actionable interview, career and salary advice.

Jan 26, 2021, 11:18am EST

Follow



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

kaggle

Create

Home

Competitions

Datasets

Code

Discussions

Courses

More

Your Work

RECENTLY VIEWED

Data Science and STE...

How to read csv files i...

Simple Linear Regressi...

The Movies Dataset

View Active Events

Search

Data Science and STEM Salaries

Data

Code (18)

Discussion (1)

Metadata

88

New Notebook

Download (3 MiB)

Detail

Compact

Column

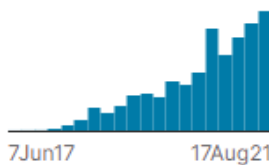
10 of 29 columns

About this file

Dataset containing records from levels.fyi

timestamp

When the data was recorded.



company

Company.

Amazon	13%
Microsoft	8%
Other (49300)	79%

level

What level the observation is at.

L4	8%
L5	8%
Other (52757)	84%

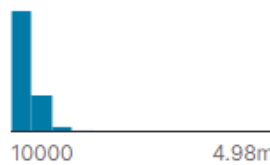
title

Role title.

Software Engineer	66%
Product Manager	7%
Other (16738)	27%

totalyearlycompe...

Total yearly compensation.



Job

Job

Sea

San

Oth

Red

San

Sea

Sun

Mou

Sea

Red

Sea

Red

6/7/2017 11:33:27

Oracle

L3

Product Manager

127000

Red

6/10/2017 17:11:29

eBay

SE 2

Software Engineer

100000

San

6/11/2017 14:53:57

Amazon

L7

Product Manager

310000

Sea

6/17/2017 0:23:14

Apple

M1

Software Engineering Manager

372000

Sun

6/20/2017 10:58:51

Microsoft

60

Software Engineer

157000

Mou

6/21/2017 17:27:47

Microsoft

63

Software Engineer

208000

Sea

6/22/2017 12:37:51

Microsoft

65

Software Engineering Manager

300000

Red

6/22/2017 13:55:26

Microsoft

62

Software Engineer

156000

Sea

6/22/2017 23:08:16

Microsoft

59

Software Engineer

120000

Red

Levels_Fyi_Salary_Data.csv

Summary

1 file

29 columns

```

'Gender' column has 68% of missing values
salaries_df.drop(['gender'], axis = 1, inplace=True)

'cityid' column is not relevant
salaries_df.drop(['cityid'], axis = 1, inplace=True)

'rowNumber' column is not relevant
salaries_df.drop(['rowNumber'], axis = 1, inplace=True)

'dmaid' column is not relevant
salaries_df.drop(['dmaid'], axis = 1, inplace=True)

drop all columns related to race, as it is not relevant
salaries_df.drop(['Race_Asian'], axis = 1, inplace=True)
salaries_df.drop(['Race_White'], axis = 1, inplace=True)
salaries_df.drop(['Race_Two_Or_More'], axis = 1, inplace=True)
salaries_df.drop(['Race_Black'], axis = 1, inplace=True)
salaries_df.drop(['Race_Hispanic'], axis = 1, inplace=True)
salaries_df.drop(['Race'], axis = 1, inplace=True)

```

Delete Columns

```

1 salaries_df.head()

```

	timestamp	company	level	title
0	6/7/2017 11:33	Oracle	L3	Product Manager
1	6/10/2017 17:11	eBay	SE 2	Software Engineer
2	6/11/2017 14:53	Amazon	L7	Product Manager
3	6/17/2017 0:23	Apple	M1	Software Engineer Manager
4	6/20/2017 10:58	Microsoft	60	Software Engineer

rows x 28 columns

Standardize

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

```

2
3 from sqlalchemy import create_engine
4 from config import db_password

```

```

1 # Connect to SQL Database
2
3 db_string = f"postgresql://postgres:{db_password}@localhost:5432/postgres"
4 engine = create_engine(db_string)

```

```

1 # Create a pandas df from the SQL table, with the following columns:
2
3 ba_salaries_df = pd.read_sql_table(
4     "ba_level1_salaries",
5     con=engine)

```

```

1 ba_salaries_df.head()

```

	year_column	company	city	state1	title	salary
0	2020	UBS	Krakow	MA	Business Analyst	100000
1	2021	HCA Healthcare	Nashville	TN	Business Analyst	100000
2	2021	Qualtrics	Provo	UT	Business Analyst	100000
3	2021	Clearwater	Boise	ID	Business Analyst	100000

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)
```

0.0303030303030304

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

56574.97037318953

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

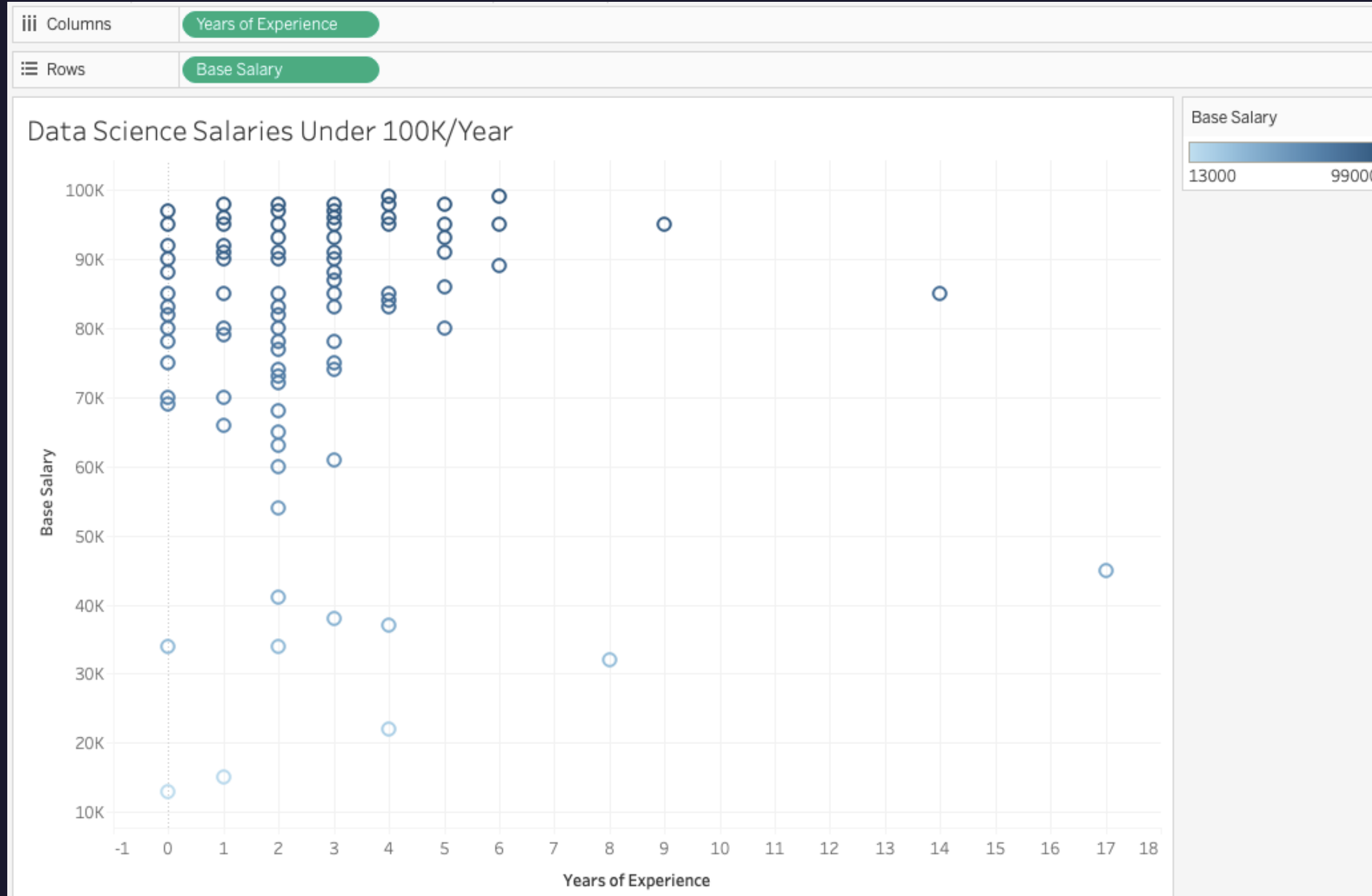
-0.03185276782647617



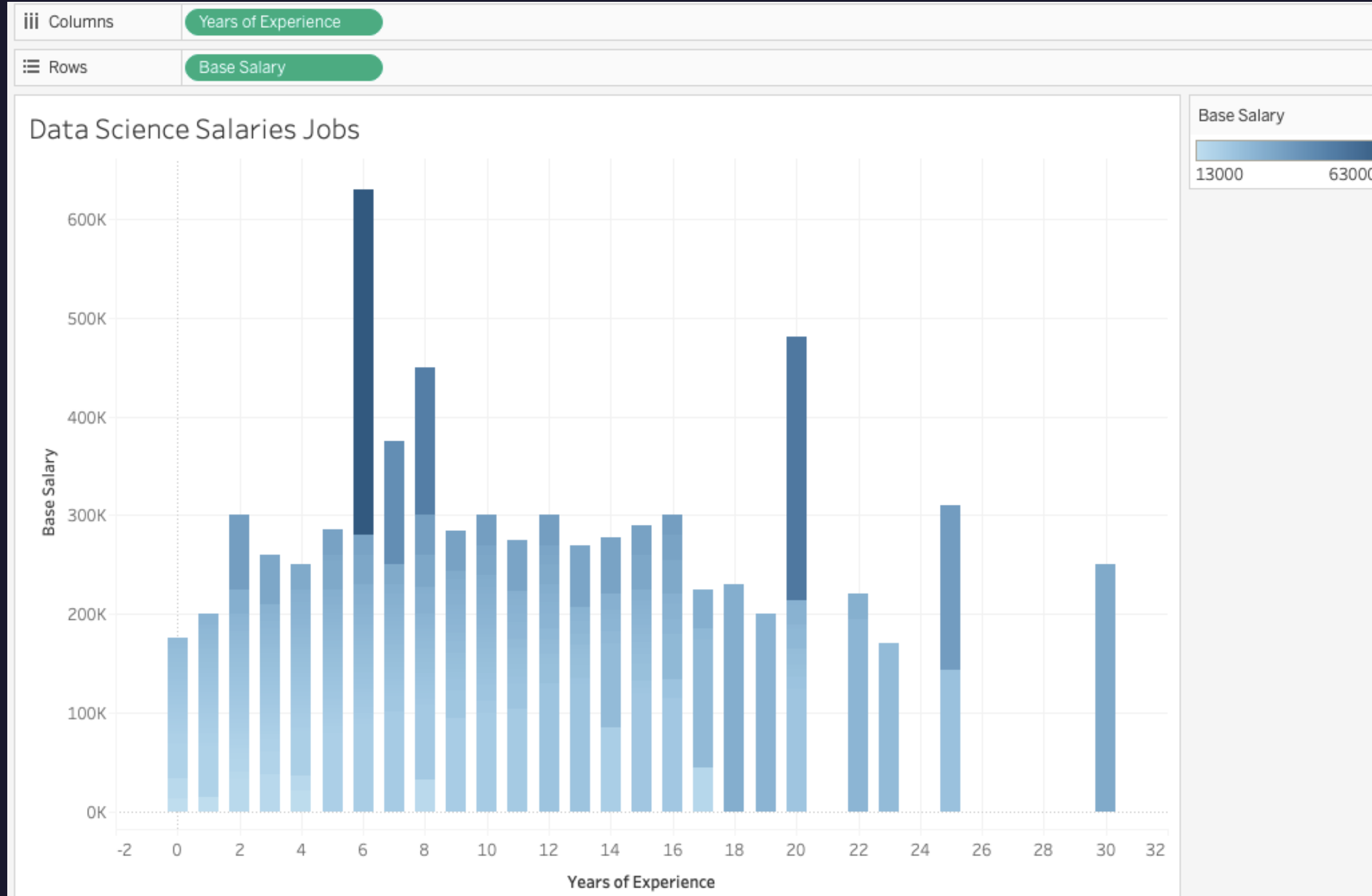
MACHINE LEARNING

Random Forest Model on Business
Analyst positions

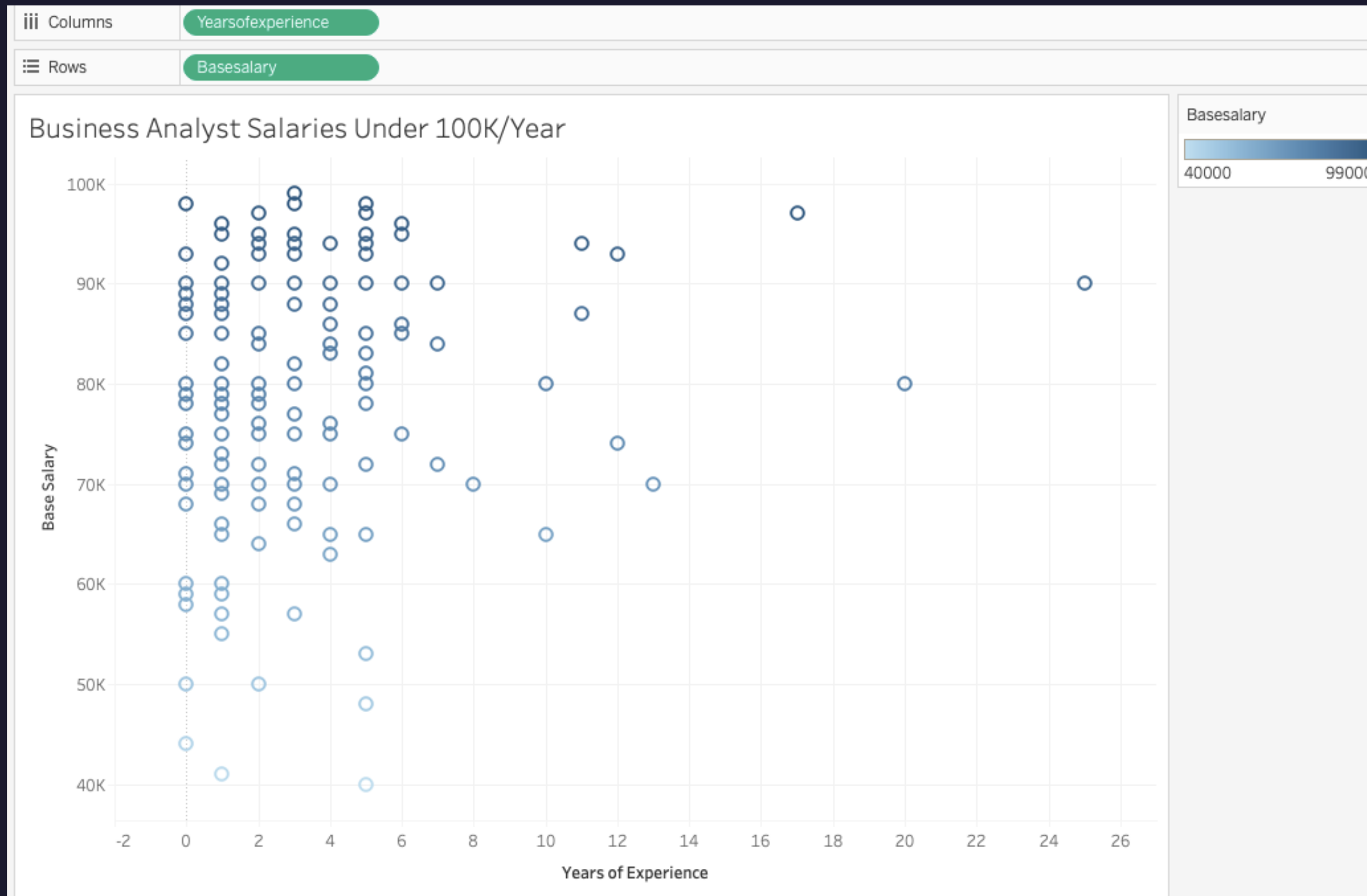
University of Miami



Tableau



Tableau



Tableau



Tableau



THANK YOU!

