DS Average Salary for the State of FL

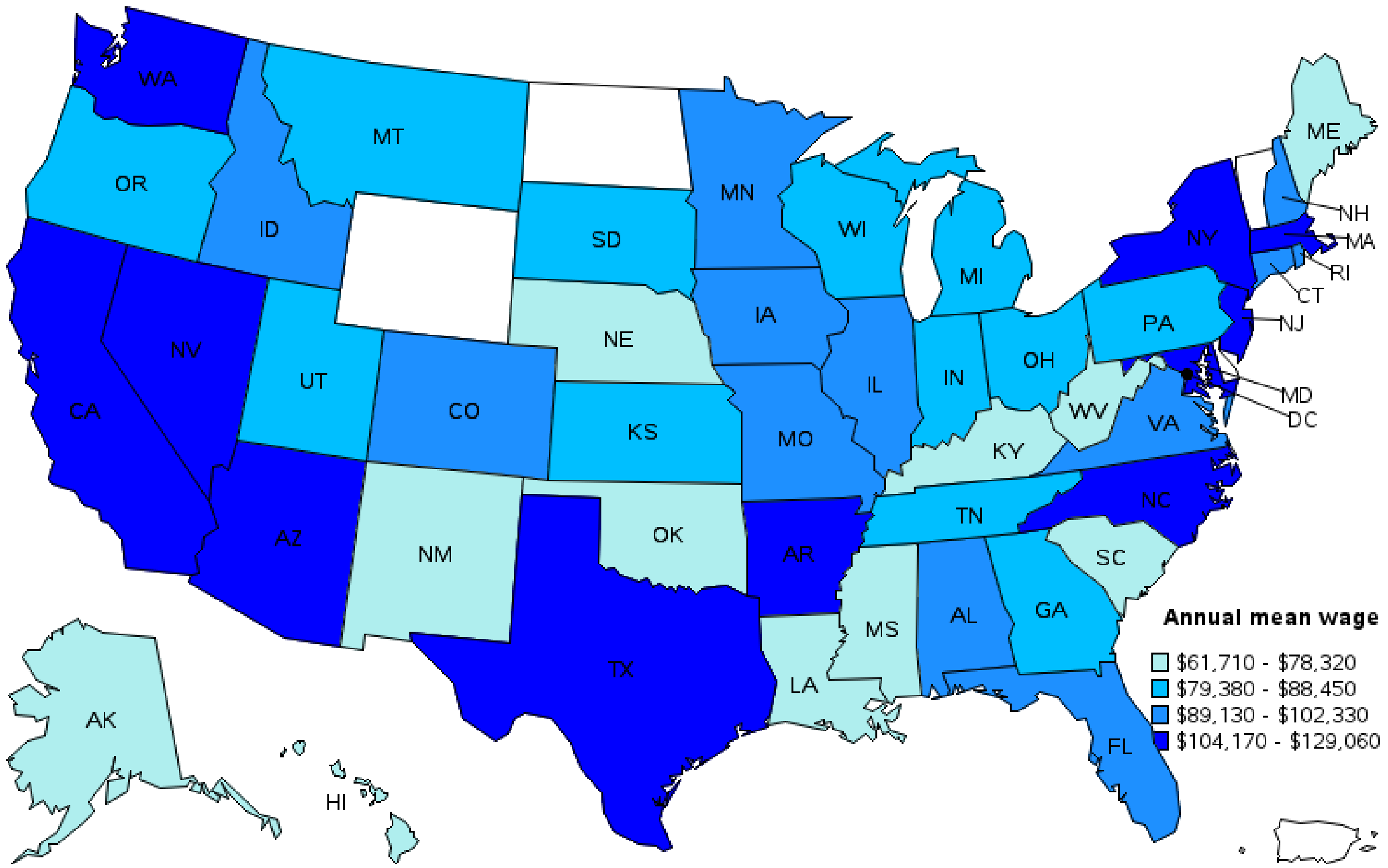
1. Introduction

a) Problem statement
The data scientist job market has current variability due to conditions arising from the Covid 19 pandemic
The transfer of companies from other states to Florida could influence the behavior of this DS job market.

Our Goal: We want to predict the salary behavior of data scientists in the state of Florida based on the opening of new positions and the skills of the applicants.

From U.S. BUREAU OF LABOR STATISTICS
https://www.bls.gov/oes/current/oes152098.htm

Annual mean wage of data scientists and mathematical science occupations, all other by state, May 2020



Blank areas indicate data not available.

b) Background

In this 2021, the labor market in general suffers contractions at the beginning of the year. In recent months due to the improvement of the situation with the covid, a recovery of jobs is beginning and specifically in the state of Florida, it is observed that a small increase in Data Scientist positions may occur and we expect an increase higher at the end of the year.

The causes in general can be different, the movement of companies from the North and West to the East could be one of them, which would lead to an increase in jobs, but we also know that there are other factors such as personal skills that can influence increase in average wages.

c) Goal

We want to make a prediction of the Florida average salary from the nation's average wages and the different factors that could influence their increase or decrease.
We also want to see the states and cities that currently have the most demand for DS jobs.

2. Datasets

I used Kaggle website source with 3 different files.

a) Data Scientist information from GlassDoor_milan
The 28 attributes of this dataset 1_glassdoor1.csv can be categorized in the following way.
Categorical:
- ●Job_Title,Job_Description,Company Name,Location, Headquarters , Type of ownership,Industry, Sector,Competitors,   employer_provided,company_txt, job_state,same_state, python_yn,R_yn, spark,aws,excel

Numerical:
- ●Salary_Estimate,Rating,Size,Rounded,Revenue,hourly,min_salary,max_salary, aveg_salary,age

After cleaning I eliminated redundant information or information not relevant for my purpose.

b) 900DS_jobs_Mirek
In the second file 2_900_DS_jobs_US_raw.csv I did not change any attribute

c) Data Scientist Jobs_Larxel
In the third File 3_DataScientist.csv, I eliminated redundant information or information not relevant for my purpose like:
- ● Unnamed: 0 ,index,Headquarters, Competitors,Easy Apply

All 3 Files finally have the same fields
Final csv has this attributes :

1 Job Title
2 Salary Estimate
3 Job Description
4 Rating

5   Company Name
6   Location
7   Size
8   Founded
9   Type of ownership
10 Industry
11 Sector
12 Revenue

With the same attributes we concatenated and obtained one DataFrame ready for cleaning and wrangling

3. Data Cleaning and Data Wrangling

The raw data set after concatenation has 5561 rows and 12 columns. We need to find unique rows and columns. Total rows decrease to 3909 rows by elimination of duplicate rows.

On the other hand, since I eliminated the redundant columns, I must now create new columns with the information I need for the analysis. For example, from the column "Salary Estimate" I can get the minimum, maximum and average salary.

From the Revenue column that has a range of values, I obtain the maximum Revenue of the company.
From the Size column which is range values as well I obtain the maximum Size of the company.
From the Founded column we can calculate how many years of foundation the company has.
From the Location column we have the city and the state of the company, we only have to separate it into two independent columns.

Rating less than zero I filled it in with zero because I consider that people do not want to evaluate the company at all, but honestly average value could be a better option I guess.

The company was founded before 1500. I consider that it is not a real number and the current year is given, but a more recent year could well be used, although the average calculation was 32 years, which tells us that there are no significant outliers .

The name of the company has Rating and the end. I removed this part of the company name as it was totally unnecessary, since we have a well defined Rating column.

"Revenue" for being a text field you have to find the numerical values inside it, also convert and multiply the words million, billion or trillion by their numerical value

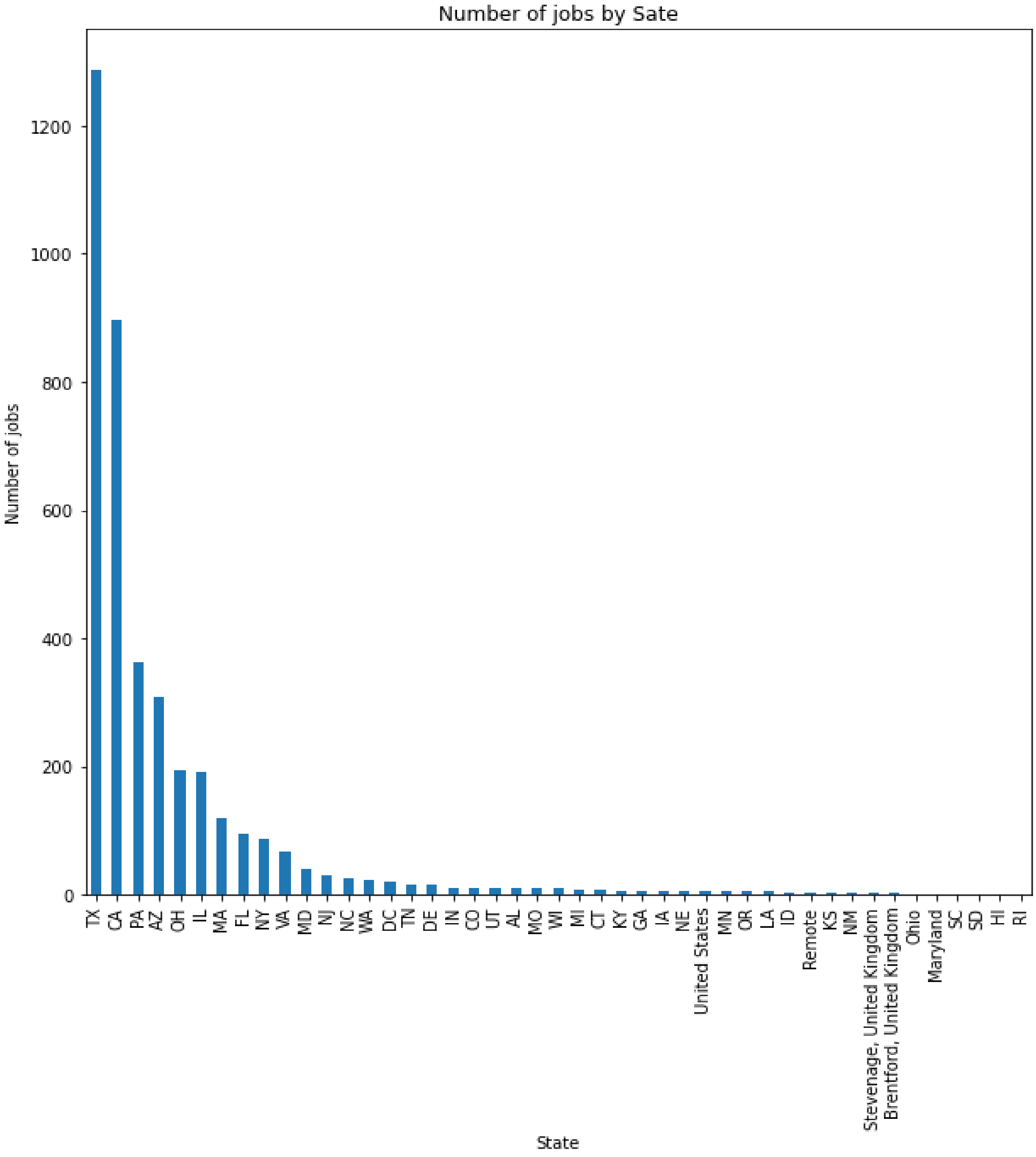
As a general rule, I avoided dropping the NaN values and consequently quite adequate values were given.

New columns are created from the Job description with interesting values such as knowledge of Python, SQL, PowerBI, DataBase, Math, Mathlab, ETL, GitHub, or Work Remote

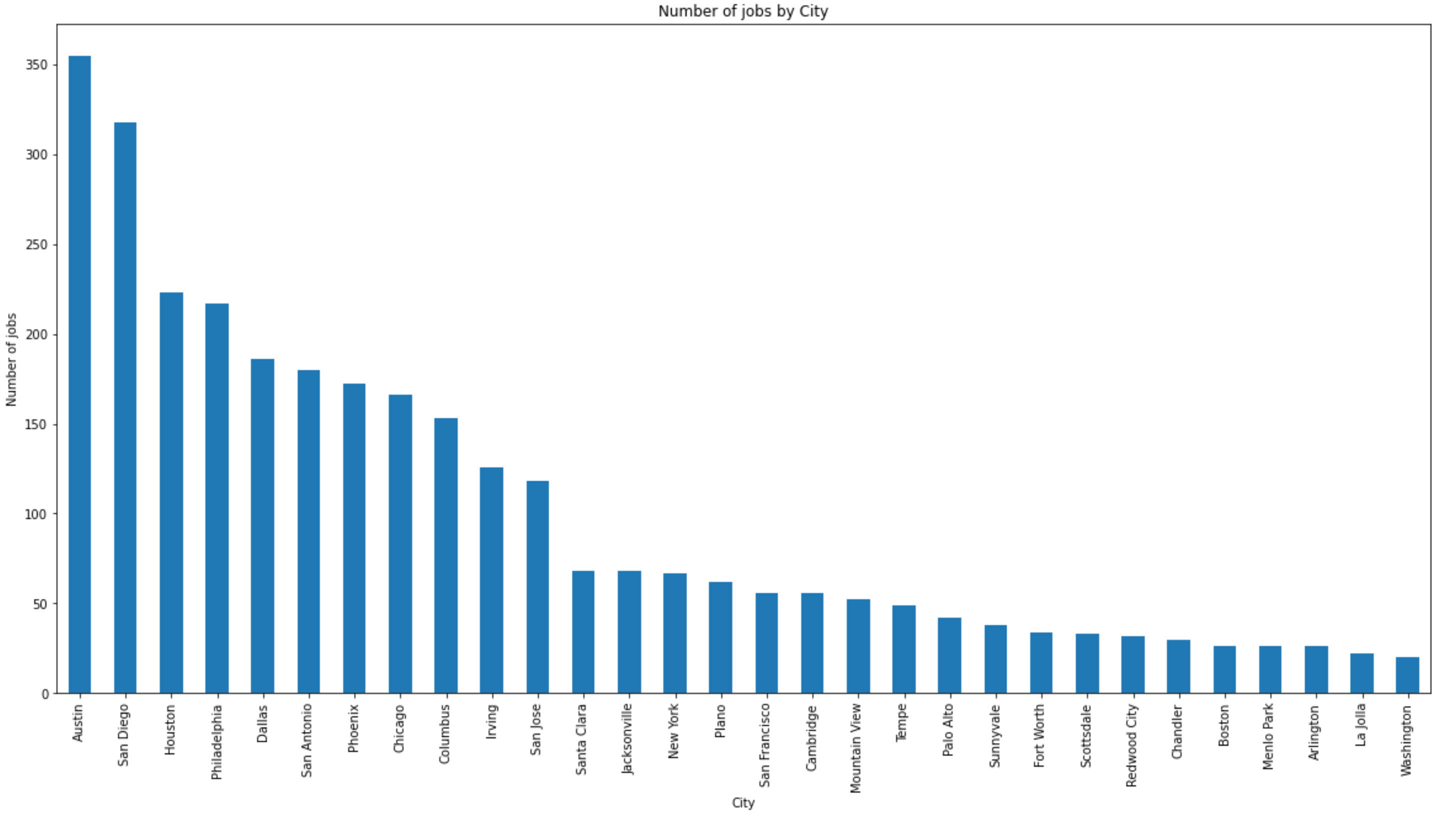
4. Exploratory Data Analysis and Initial Findings

The average salary by State and by Cities gives us an idea of the demand, as well as the number of job options available in each State.

A) Jobs by State



B) Jobs by City



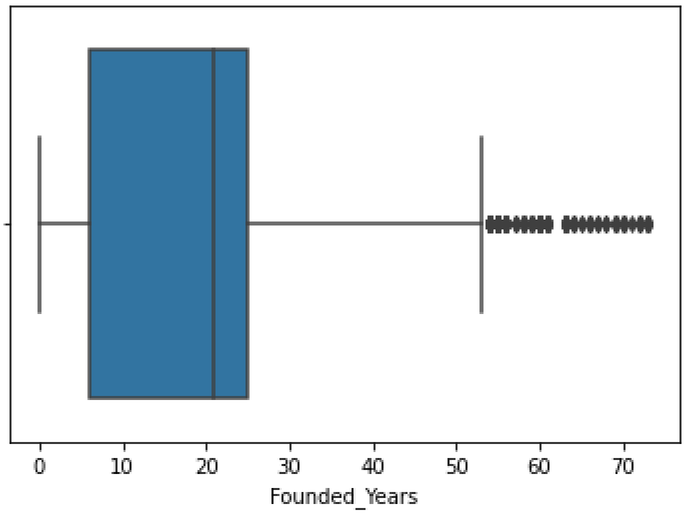
It's interesting the amount of jobs in Texas state and for the same reason the most cities of Texas have high demand for DS jobs. Example: Austin,Houston,Dallas, San Antonio,Irving,Plano,Fort Worth

At first I did a typical division of numerical and categorical columns. I did corresponding analyzes to see the relationship between them, but it was more interesting for me to work these categories as numerical due to the possibility of finding correlations between them and being able to apply ML to them.

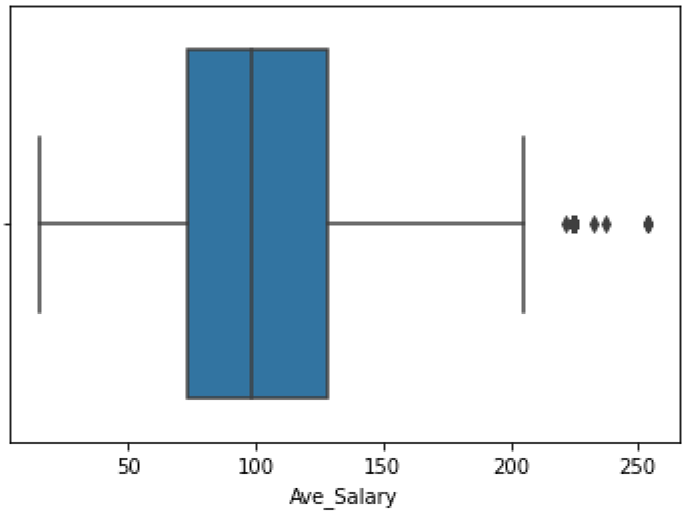
Detecting Outliers. Using IQR score technique.
To put an example of detecting Outliers for Founded Years of a Company

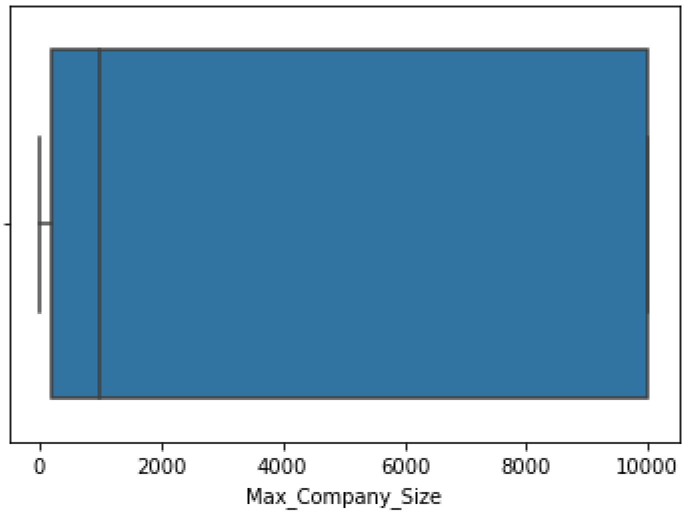
```
Q1s=result['Founded_Years'].quantile(0.25)
Q3s=result['Founded_Years'].quantile(0.75)
IQRs=Q3s-Q1s
Lower_Whiskers = abs(Q1s-1.5*IQRs)
Upper_Whiskers = Q3s+1.5*IQRs
medians = float(result['Founded_Years'].median())
result['Founded_Years']=np.where(result['Founded_Years']>Upper_Whiskers,medians,result['Founded_Years'])

print(Q1s,Q3s,IQRs,Lower_Whiskers,Upper_Whiskers)
print(medians)
sns.boxplot(x=result['Founded_Years'])
```

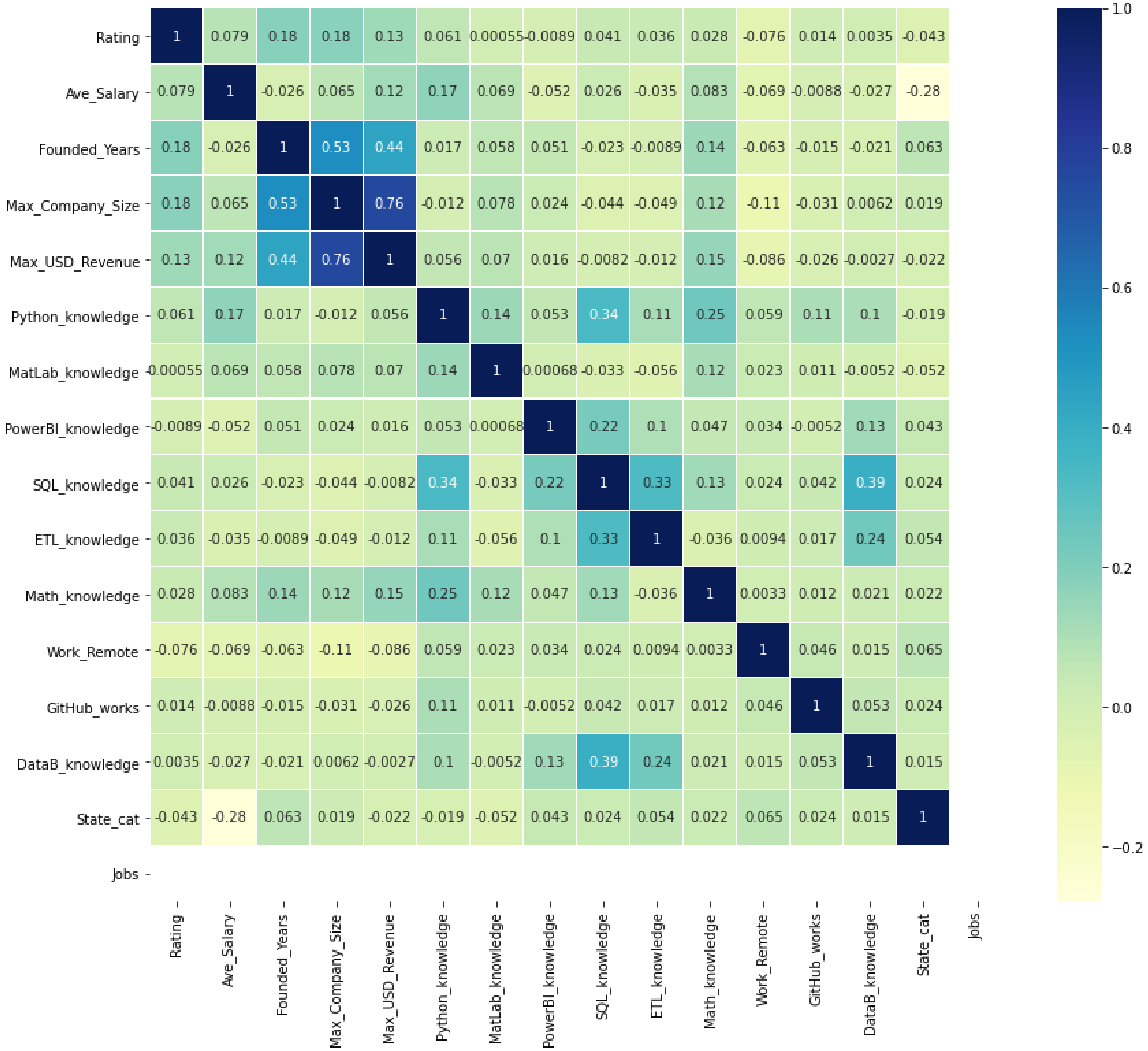


I used the same procedure to ‘Max_Company_Size’ and ’Max_USD_Revenue’





Trying to find any relation between features



	Rating	Ave_Salary	Founded_Years	Max_Company_Size	Max_USD_Revenue	Python_knowledge	MatLab_knowledge	PowerBI_knowledge	SQL_knowledge	ETL_knowledge	Math_knowledge	Work_Remote	GitHub_works	DataB_knowledge	State_cat	Jobs
Rating	1	0.079	0.18	0.18	0.13	0.061	0.00055	-0.0089	0.041	0.036	0.028	-0.076	0.014	0.0035	-0.043	
Ave_Salary	0.079	1	-0.026	0.065	0.12	0.17	0.069	-0.052	0.026	-0.035	0.083	-0.069	-0.0088	-0.027	-0.28	
Founded_Years	0.18	-0.026	1	0.53	0.44	0.017	0.058	0.051	-0.023	-0.0089	0.14	-0.063	-0.015	-0.021	0.063	
Max_Company_Size	0.18	0.065	0.53	1	0.76	-0.012	0.078	0.024	-0.044	-0.049	0.12	-0.11	-0.031	0.0062	0.019	
Max_USD_Revenue	0.13	0.12	0.44	0.76	1	0.056	0.07	0.016	-0.0082	-0.012	0.15	-0.086	-0.026	-0.0027	-0.022	
Python_knowledge	0.061	0.17	0.017	-0.012	0.056	1	0.14	0.053	0.34	0.11	0.25	0.059	0.11	0.1	-0.019	
MatLab_knowledge	0.00055	0.069	0.058	0.078	0.07	0.14	1	0.00068	-0.033	-0.056	0.12	0.023	0.011	-0.0052	-0.052	
PowerBI_knowledge	-0.0089	-0.052	0.051	0.024	0.016	0.053	0.00068	1	0.22	0.1	0.047	0.034	-0.0052	0.13	0.043	
SQL_knowledge	0.041	0.026	-0.023	-0.044	-0.0082	0.34	-0.033	0.22	1	0.33	0.13	0.024	0.042	0.39	0.024	
ETL_knowledge	0.036	-0.035	-0.0089	-0.049	-0.012	0.11	-0.056	0.1	0.33	1	-0.036	0.0094	0.017	0.24	0.054	
Math_knowledge	0.028	0.083	0.14	0.12	0.15	0.25	0.12	0.047	0.13	-0.036	1	0.0033	0.012	0.021	0.022	
Work_Remote	-0.076	-0.069	-0.063	-0.11	-0.086	0.059	0.023	0.034	0.024	0.0094	0.0033	1	0.046	0.015	0.065	
GitHub_works	0.014	-0.0088	-0.015	-0.031	-0.026	0.11	0.011	-0.0052	0.042	0.017	0.012	0.046	1	0.053	0.024	
DataB_knowledge	0.0035	-0.027	-0.021	0.0062	-0.0027	0.1	-0.0052	0.13	0.39	0.24	0.021	0.015	0.053	1	0.015	
State_cat	-0.043	-0.28	0.063	0.019	-0.022	-0.019	-0.052	0.043	0.024	0.054	0.022	0.065	0.024	0.015	1	
Jobs																

Apparently I don’t see any high relation between the features. I considered high over 0.65 and I have 0.44, 0.39, some 0.53 and 0.76 the highest

State-wide summary data

```
state_summary=numer_data.groupby('State').
    agg(jobs_per_state=pd.NamedAgg(column='Jobs',aggfunc='sum'),
    state_ave_salary=pd.NamedAgg(column='Ave_Salary', aggfunc='mean'),
    state_total_Python=pd.NamedAgg(column='Python_knowledge', aggfunc='sum'),
    state_total_Sql=pd.NamedAgg(column='SQL_knowledge', aggfunc='sum'),
    state_total_DataB=pd.NamedAgg(column='DataB_knowledge', aggfunc='sum'),
    state_total_Math=pd.NamedAgg(column='Math_knowledge', aggfunc='sum')
    ).reset_index()
state_summary.head()
```

Explore The Data

Top States By Order Of Each Of The Summary Statistics

Visualizing High Dimensional Data

I started digging more.I used PCA and Scaled the data

The basic steps in this process are:

1) scale the data
2) fit the PCA transformation (learn the transformation from the data)
3) apply the transformation to the data to create the derived features
4) (optionally) use the derived features to look for patterns in the data and explore the coefficients

Scale the data

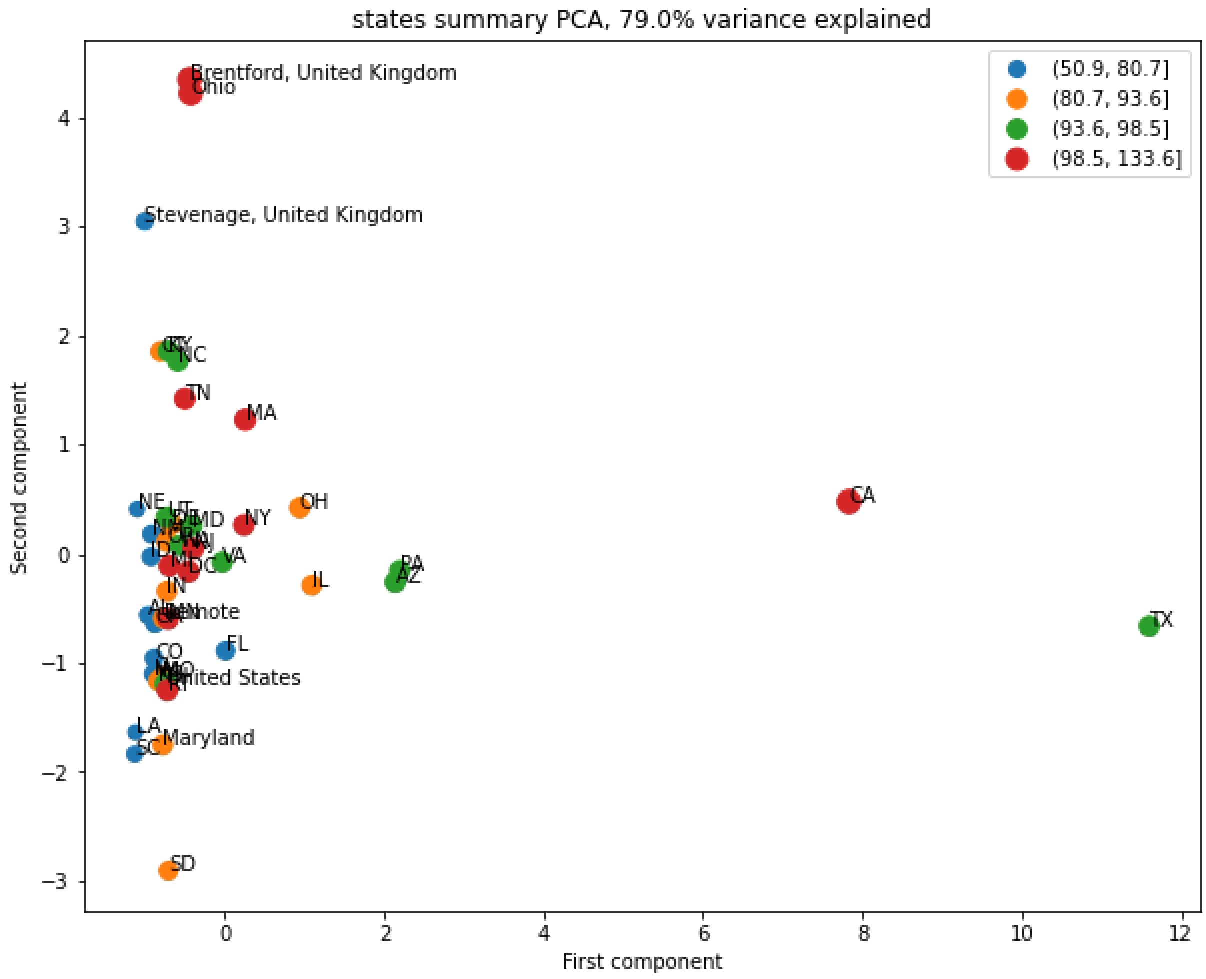
Verifying the scaling

```
state_summary_scale = scale(state_summary_scale)
```

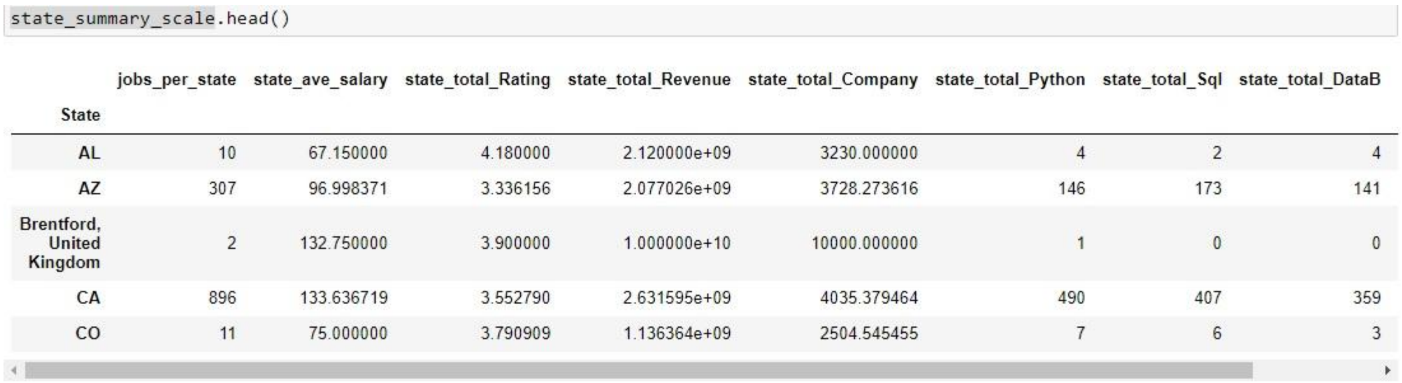
Calculate the PCA transformation

```
state_pca = PCA().fit(state_summary_scale)
```

Average salary by state



Feature engineering



```
state_summary_scale.head()
```

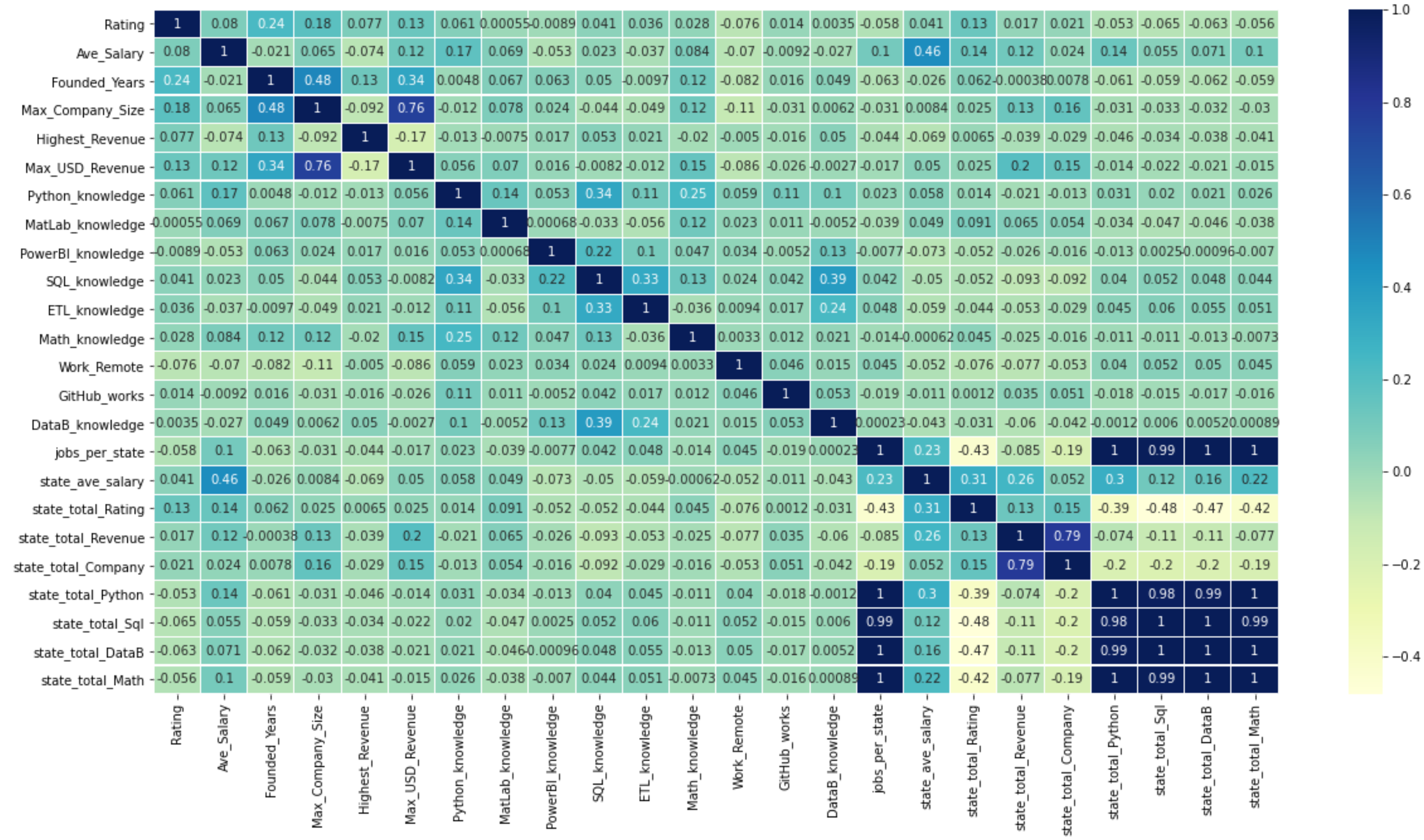
	jobs_per_state	state_ave_salary	state_total_Rating	state_total_Revenue	state_total_Company	state_total_Python	state_total_Sql	state_total_DataB
State								
AL	10	67.150000	4.180000	2.120000e+09	3230.000000	4	2	4
AZ	307	96.998371	3.336156	2.077026e+09	3728.273616	146	173	141
Brentford, United Kingdom	2	132.750000	3.900000	1.000000e+10	10000.000000	1	0	0
CA	896	133.636719	3.552790	2.631595e+09	4035.379464	490	407	359
CO	11	75.000000	3.790909	1.136364e+09	2504.545455	7	6	3

Feature correlation heatmap

I see a little correlation between Jobs per State and Average Salary per State ,that tells us that states that have more jobs also generally increase wages as well

We can see a total correlation between the amount of Jobs per State and amount of Python, SQL ,DataBase and Math knowledge requests which is evident. DataBase knowledge is correlated with SQL and ETL knowledge which is also evident.

On the other hand, the correlation between the size of the company, the years it was founded and the maximum revenue is maintained.



Scatterplots of numeric features against average salary

Summary

I can see different kinds of correlations which are not unusual for me and are not the purpose of the search. For example: SQL_knowledge vs Python_knoledge. SQL_knowledge vs DataB_knowledge. Math_knowledge vs Python_knowledge. DataB_knowledge vs ETL_knowledge.

5 Pre-Processing and Training Data

Fit the dummy regressor on the training data

```
dumb_reg = DummyRegressor(strategy='mean')
dumb_reg.fit(X_train, y_train)
dumb_reg.constant_
```

Metrics

sklearn metrics

We tried R-squared (median_r2), Mean Absolute Error (median_mae) and Mean Squared Error(median_mse)

Initial Models
Pre-Processing and Training Data

Extract FI State Data
Train/Test Split
Initial Not-Even-A-Model
Metrics
R-squared, or coefficient of determination
Mean Absolute Error
Mean Squared Error
Sklearn metrics
R-squared
Mean absolute error
Mean squared error

Refining The Linear Model

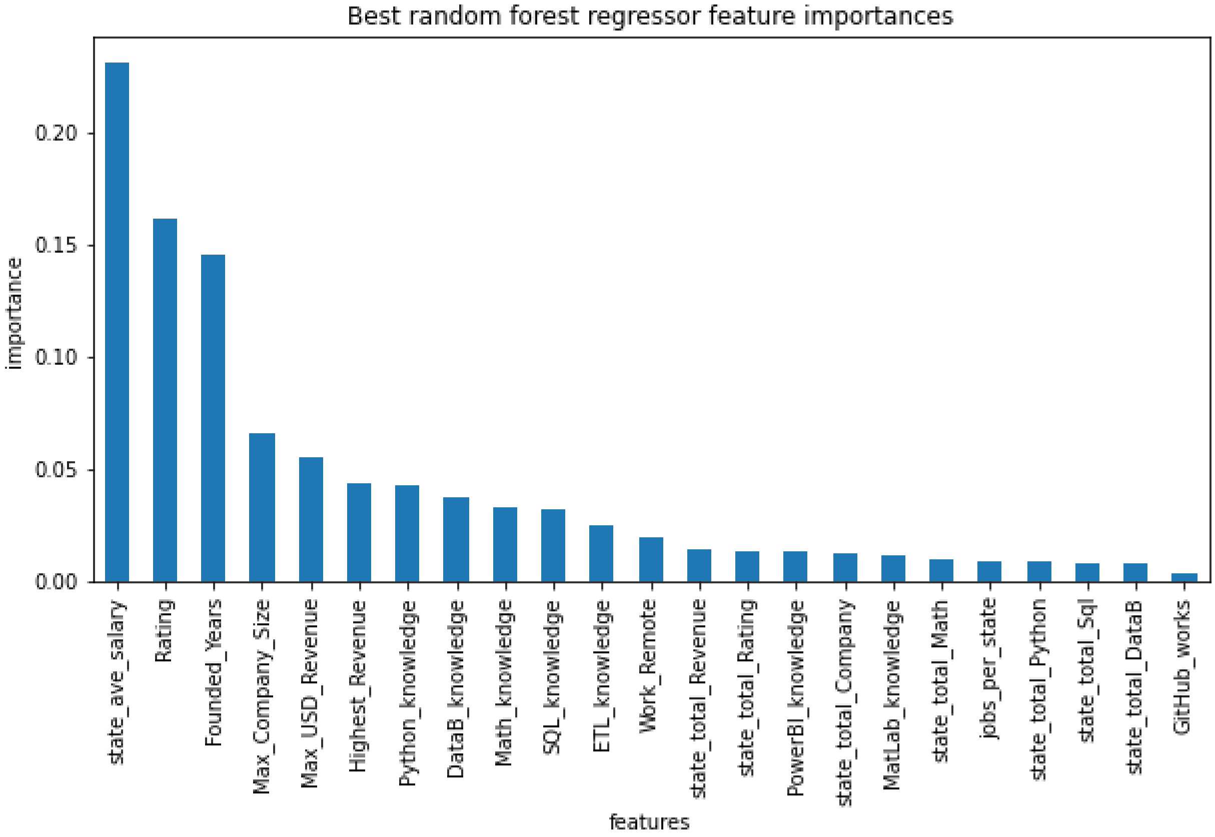
Define the pipeline
Fit the pipeline
Assess performance on the train and test set
Define a new pipeline to select a different number of features
Fit the pipeline
Assess performance on train and test data
Assessing performance using cross-validation
Hyperparameter search using GridSearchCV
Pulling the above together, we have:
a pipeline that

- imputes missing values
- scales the data
- selects the k best features
- trains a linear regression model

a technique (cross-validation) for estimating model performance

Random Forest Model

Define the pipeline
Fit and assess performance using cross-validation
Hyperparameter search using GridSearchCV



Final Model Selection

Linear regression model performance
Random forest regression model performance
Conclusion: The random forest model has a lower cross-validation mean absolute

6 Modeling

Refit Model On All Available Data (excluding FL State)

Calculate Expected Fl Average Salary From The Model

Average Salary = 79.962766

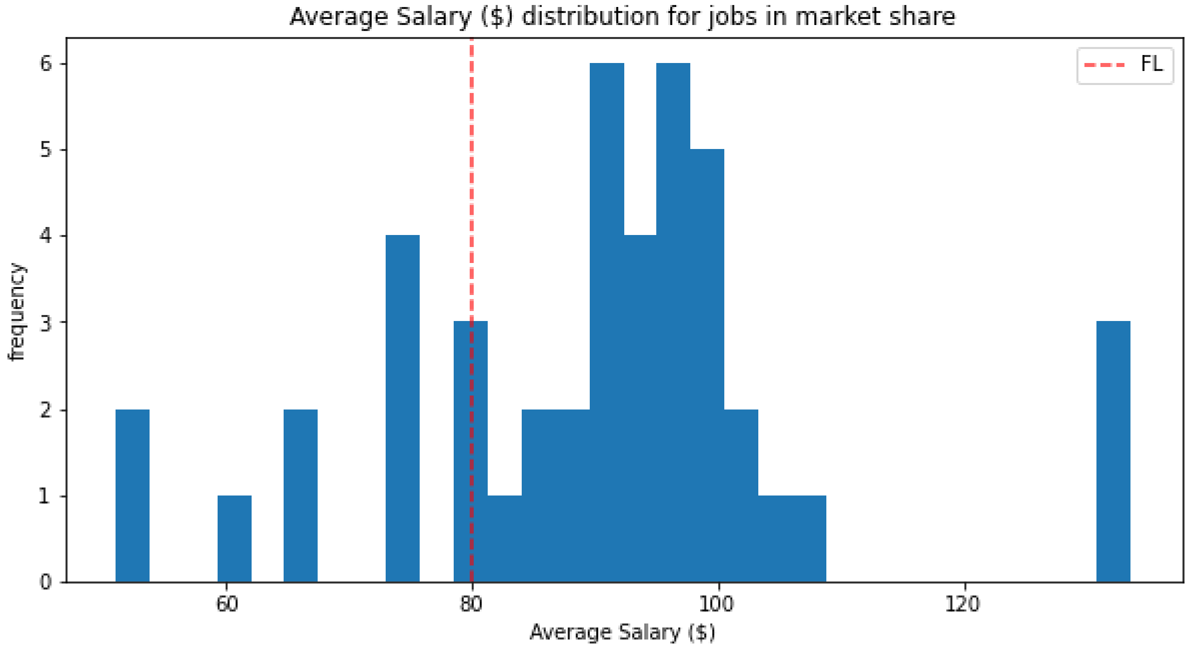
Florida Jobs In Market Context

We know from the beginning that there is a certain relationship between jobs by state and average salary by state, but we need to know other features or factors that can mainly influence the average salary.

Features that came up as important in the modeling included:

Founded_Years
Rating
Highest_USD_Revenue
Max_Company_Size
Python_knowledge
DataB_knowledge
SQL_knowledge
Math_knowledge

Average Salary



We also analyze and did also histogram for: Founded Years of a Company,Rating of Company, Highest Revenue possible for a Company, Maximum possible Company Size, Python knowledge, DataBase knowledge, SQL knowledge and Math knowledge

Summary

FL State currently has a Salary of 80K dollars per job. A modelling suggest 78K

If I want to change another variable to see the influence in my model for example:Rating and Max_Company_Size We don't see any real impact. The same occurs with jobs_per_state.

Conclusions:

1) Certain States(Like Texas, California,PA) and certain cities(Like Austin,Houston, San Diego,Dallas, San Antonio) have more job options as a Data Scientist

2) The maximum size of the company, the highest income of the companies, the rating of the companies and the years of foundation have some directly relation to the average salary paid

3) Certain states have the highest salary than others without a high relationship to knowledge of Python, math, SQL, or knowledge of databases. It could be because they have a higher cost of living, being areas of technological development such as Silicon Valley or simply because that state has greater growth

4) The amount of jobs offered are directly related to the knowledge of Python, SQL, Math, and Database and it is logical because that general knowledge is currently needed in these areas, the interesting thing is that the relationship is not so high with the average salary and this it may be because all jobs ask for it

5) The rating of companies by states fluctuates with a mean between 3 and 4.4 with a median of 3.58, which could be considered low or unsatisfied aspirations by data scientists

6) Power Bi, GitHub, ETL and work remotely have no relation to the average salary

Highlight any deficiencies in the data that hampered or limited this work.

The only salary data in our dataset was average salary.
We can check minimum and maximum salary as well
We can analyze others features in EDA like I'll recommend checking other features, maybe "Experience" I guess has
more relation with Average Salary. "Machine Learning Models"

Assuming the business leaders felt this model was useful, how would the business make use of it?

They can use it to do job marketing related to the state of Florida.

To know and predict how adjusted the average salary is by State