Salary Predictions of Data Science Roles

Why Data Science Salaries?

- We selected this topic because it was one that was of interest to all of us including our classmates. As we graduate from this course and enter the job market we have a frame of reference when negotiating offers.
- Our goal throughout this analysis was to better educate us on the salary levels of different data-related jobs.
- We wanted to see how factors such as company, location and experience affect the expected salary.

Our Data

- The dataset we selected to use is:
- Kaggle: Salary and more-Data Scientist, Analyst, Engineer, Retrieved 3/16/2022 from https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries.
- Some of the data this dataset includes are job title, salary, company, years of experience, gender, education level and location
- During our analysis, we hoped this dataset would help us answer the question of how do different factors such as location, years of experience, gender, and company affect the anticipated salary?
- **Limitations to Dataset**
  - Accuracy of the data because it was self-reported by humans
  - Data did not include job descriptions, and all jobs were the same title
  - There were several imbalances in the dataset such as company size, location, years of experience
  - Total Yearly Compensation means different things to different people when self reporting (are stock options included, benefit packages, etc)

Exploration of Data

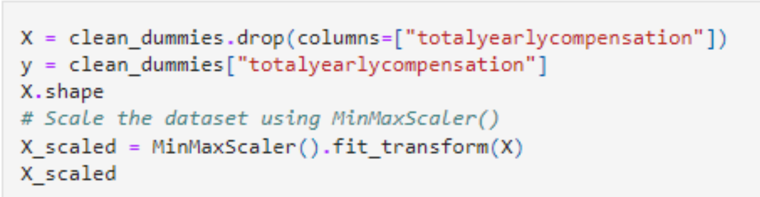
- Imported data in csv form into a Jupyter notebook and cleaned the data using pandas:
  - Dropping unnecessary columns in Data Frame, such as
  - Split the location into States, Cities and Countries
  - Use Drop Na function to remove the empty columns
  - Bin Companies, years of experience, states and by data count size
  - OneHotEncoder all the object column
- Description of Database component
  - Connected local Postgres to AWS server using pgAdmin with pyspark in Google colab.

Machine Learning Model Selection

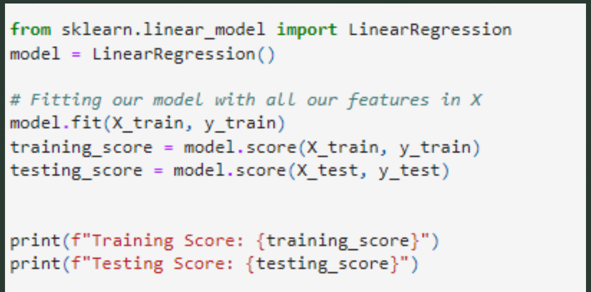
- Once the data was clean we did visualizations of the data to decide which models may lend itself better to the data
- We initially decided upon a linear model because what we were trying to find was a correlation rather than a classification
- Once we tired to linear regression model we decided to explore and build upon it with the Random Forest Regressor
- To see if we could improve the model and employ skills learned in class we also tested with a Neural Network

Analysis of Data – Using Linear Regression Machine Learning Model

- How was data split into training and testing sets:
  - We used both the default settings (75/25) and 80/20



```python
X = clean_dummies.drop(columns=["totalyearlycompensation"])
y = clean_dummies["totalyearlycompensation"]
X.shape
# Scale the dataset using MinMaxScaler()
X_scaled = MinMaxScaler().fit_transform(X)
X_scaled
```

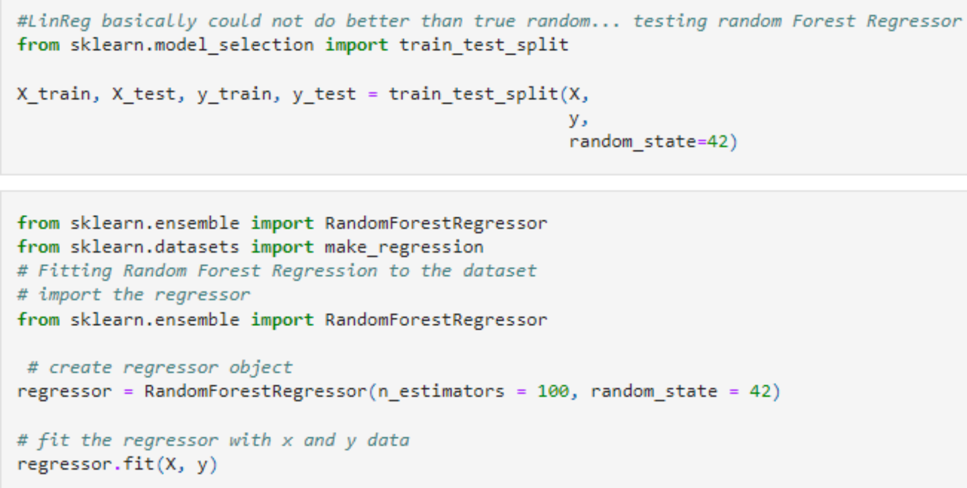


```python
from sklearn.linear_model import LinearRegression
model = LinearRegression()

# Fitting our model with all our features in X
model.fit(X_train, y_train)
training_score = model.score(X_train, y_train)
testing_score = model.score(X_test, y_test)


print(f"Training Score: {training_score}")
print(f"Testing Score: {testing_score}")
```

- Explanation of our model choice
  - Benefits: Find the relationship between variables (Total Annual Salary vs other factors), to construct a linear equation to the observed data.
  - Limitations: Some features of the dataset may not be linear correlated with the Salary

Analysis of Data – Using Random Forest Regressor Machine Learning Model

- How was data split into training and testing sets



```
#LinReg basically could not do better than true random... testing random Forest Regressor
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y,
                                                    random_state=42)
```

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.datasets import make_regression
# Fitting Random Forest Regression to the dataset
# import the regressor
from sklearn.ensemble import RandomForestRegressor

 # create regressor object
regressor = RandomForestRegressor(n_estimators = 100, random_state = 42)

# fit the regressor with x and y data
regressor.fit(X, y)
```

- Explanation of our model choice
  - Benefits: can solve both regression and classification problems.
    - This model didn’t require scaling the data
  - Limitations: The ability to explain this model to stakeholders is a bit more “black box” than the Linear Regression

Cross Validation– Using Neural Network Machine Learning Model

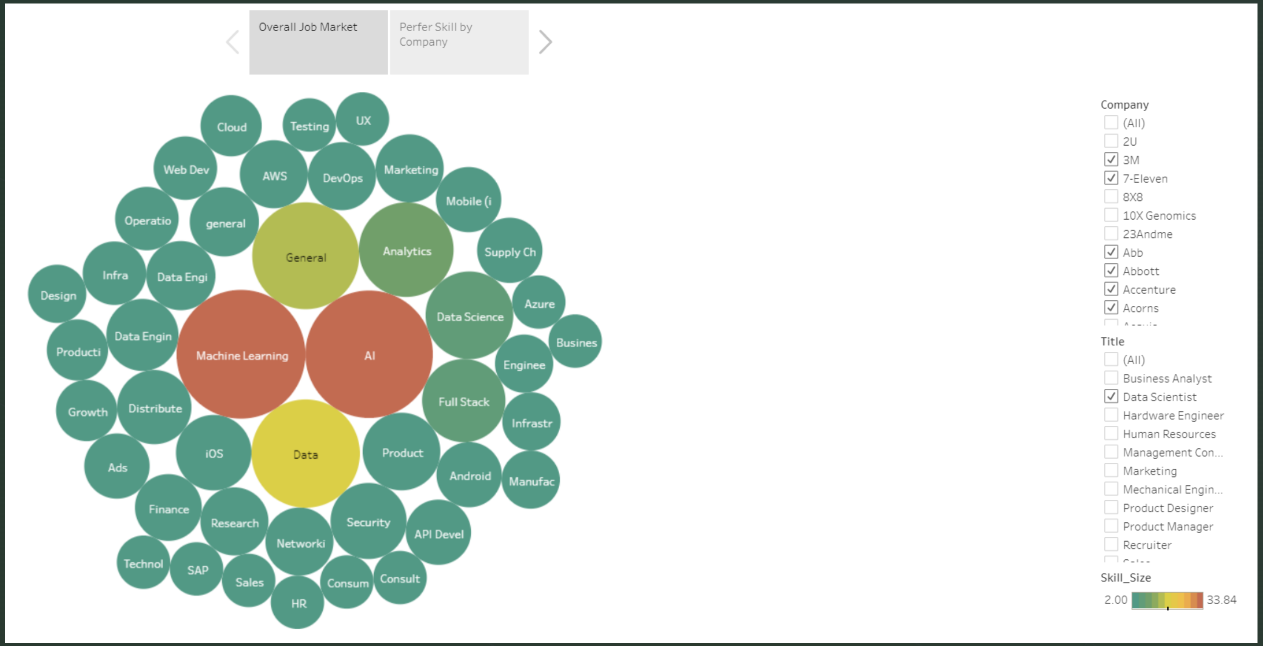
- How was data split into training and testing sets:
  - We used the standard train test split
- Explanation of our model choice
  - Benefits:
    - It is a more sophisticated model, however our amount of data did not really lend itself to this type of model
  - Limitations:
    - The amount of data we had, the model would perform better with more data

Interactive Dashboard

- We will create a website that displays our findings
- The tools that will be used to create our dashboard are:
  - Tools we will use:
    - Bootstrap
    - Html
    - Github Pages to host
    - Tableau to create interactive visuals for the user

Storyboard:
Prefer Skills amount market and different companies

User can filter the titles and companies to see what skills are required or highly acquire by the current employees.



Company

- □ (All)
- □ 2U
- ☑ 3M
- ☑ 7-Eleven
- □ 8X8
- □ 10X Genomics
- □ 23Andme
- ☑ Abb
- ☑ Abbott
- ☑ Accenture
- ☑ Acorns

Title

- □ (All)
- □ Business Analyst
- ☑ Data Scientist
- □ Hardware Engineer
- □ Human Resources
- □ Management Con...
- □ Marketing
- □ Mechanical Engin...
- □ Product Designer
- □ Product Manager
- □ Recruiter

Skill_Size

2.00 33.84

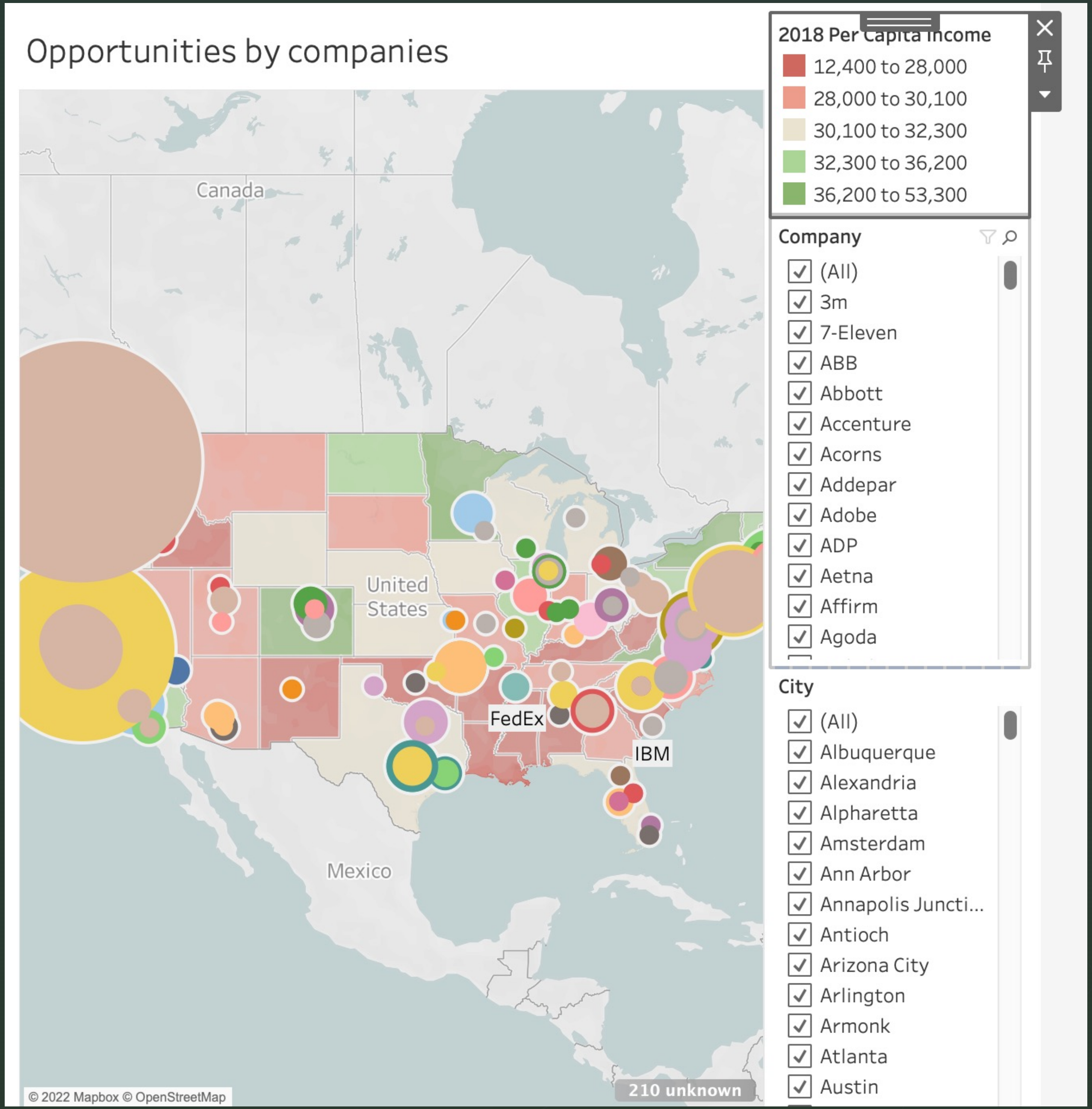
https://public.tableau.com/app/profile/sirius.liao/viz/Preferskillsfordatascientist/Story1

Storyboard:
External factors that affect the average salary

- We will have a Tableau dashboard that is interactive for the user. They can filter on company name and city. It will show the number of reported data science employees for that company.

  There is also a layer added to the map that has the per capita income of the state reported by Tableau.

  https://public.tableau.com/app/profile/mackenzie.coushay.richter/viz/SalaryAnalysis-Summary_16482343926450/location_dashboard?publish=yes



Opportunities by companies

2018 Per Capita Income

- 🟥 12,400 to 28,000
- 🟥 28,000 to 30,100
- ⬜ 30,100 to 32,300
- 🟩 32,300 to 36,200
- 🟩 36,200 to 53,300

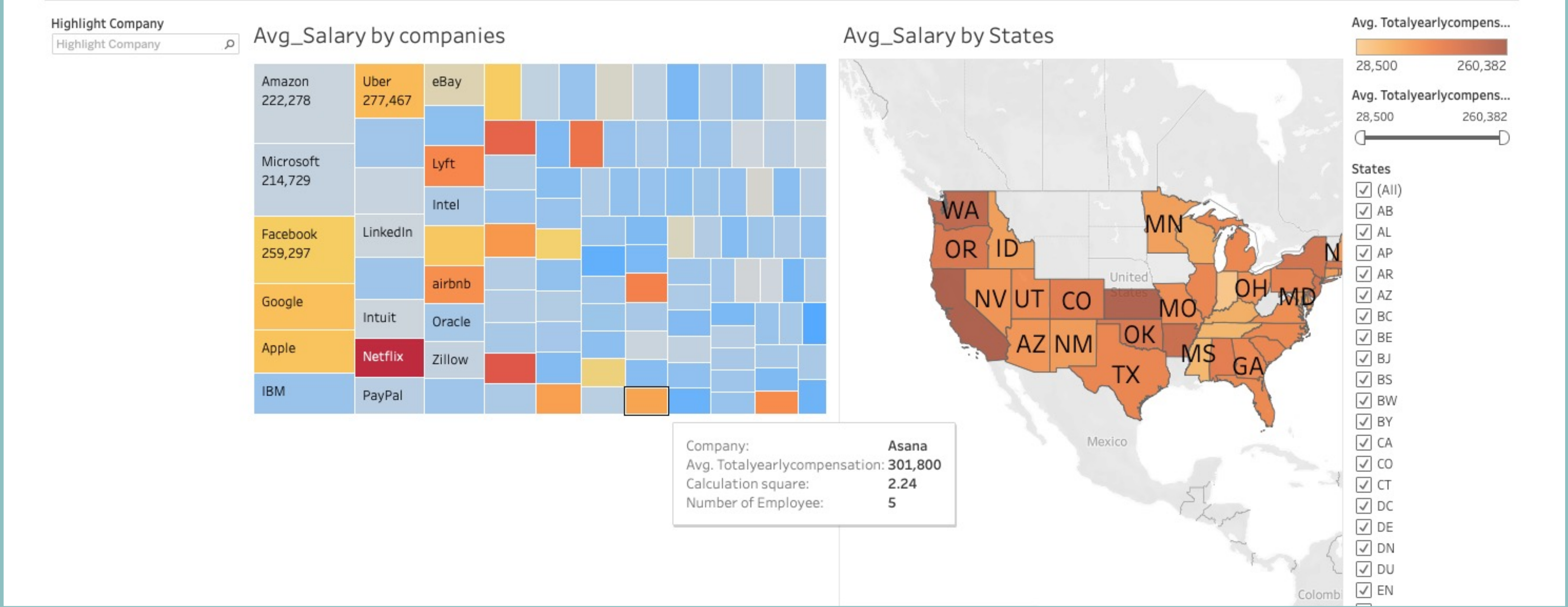
Company

- ☑ (All)
- ☑ 3m
- ☑ 7-Eleven
- ☑ ABB
- ☑ Abbott
- ☑ Accenture
- ☑ Acorns
- ☑ Addepar
- ☑ Adobe
- ☑ ADP
- ☑ Aetna
- ☑ Affirm
- ☑ Agoda

City

- ☑ (All)
- ☑ Albuquerque
- ☑ Alexandria
- ☑ Alpharetta
- ☑ Amsterdam
- ☑ Ann Arbor
- ☑ Annapolis Juncti...
- ☑ Antioch
- ☑ Arizona City
- ☑ Arlington
- ☑ Armonk
- ☑ Atlanta
- ☑ Austin

Storyboard: Salary by company and state



Company: Asana
Avg. Totalyearlycompensation: 301,800
Calculation square: 2.24
Number of Employee: 5

States

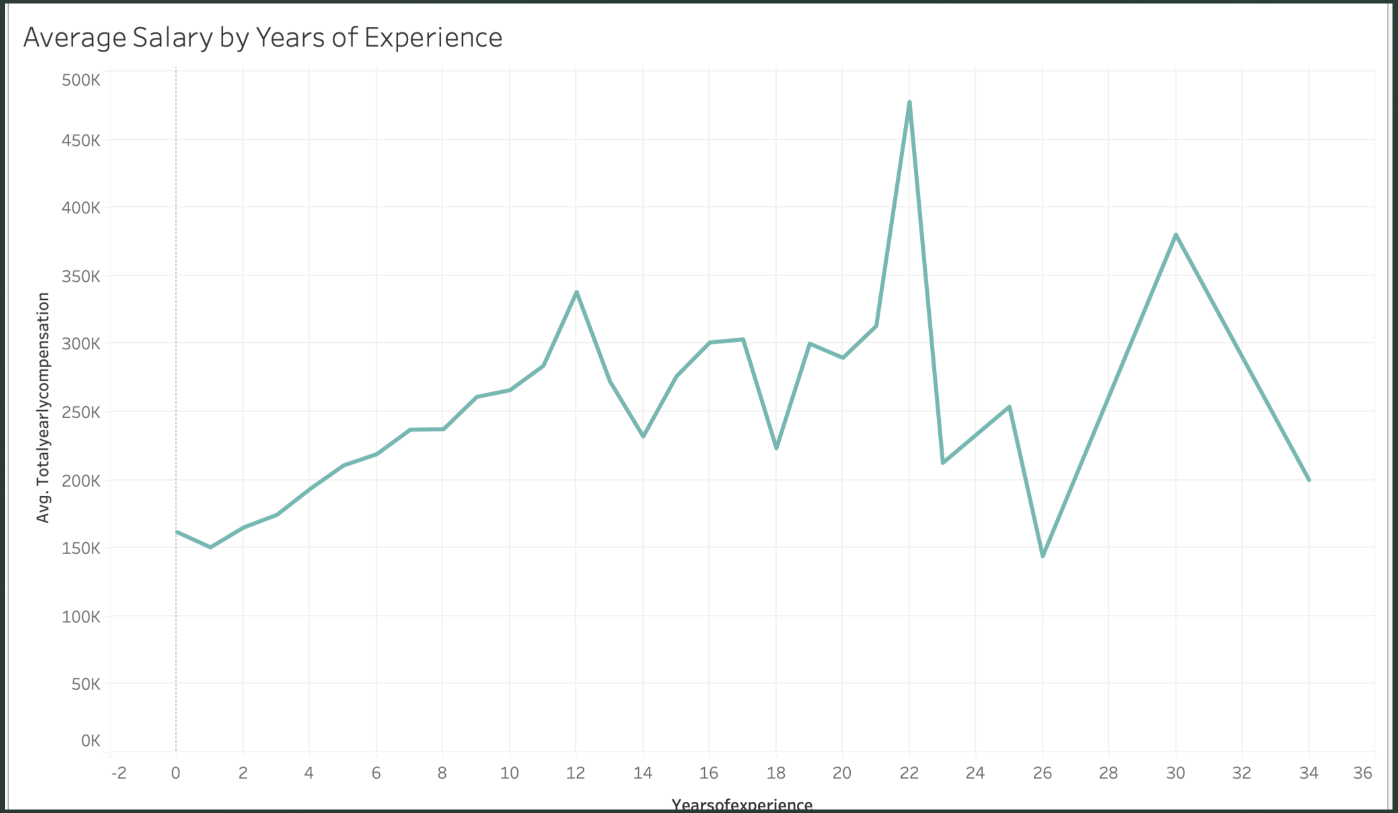
☑ (All)
☑ AB
☑ AL
☑ AP
☑ AR
☑ AZ
☑ BC
☑ BE
☑ BJ
☑ BS
☑ BW
☑ BY
☑ CA
☑ CO
☑ CT
☑ DC
☑ DE
☑ DN
☑ DU
☑ EN

https://public.tableau.com/app/profile/mackenzie.coushay.richter/viz/SalaryAnalysis-Summary_16482343926450/AverageSalarybyLocationCompany?publish=yes

This Interactive Tableau Dashboard visual will show the average salary by company. It is interactive for the user because they can highlight the company they want to see either type in the company name or select from a dropdown

This Interactive Tableau Dashboard will show the average salary by state. It is interactive for the user because they can filter by both state and desired average Salary.

Storyboard: Average Salary By Years of Experience



https://public.tableau.com/app/profile/mackenzie.coushay.richter/viz/SalaryAnalysis-Summary_16482343926450/Salarybyexp_years?publish=yes

This visual created with Tableau shows average salary by Years of experience collected from our dataset