Group 6

Kelsey Brantner, Laura Jordan, Jiamin Li, Xiaolin Ruan, Diane Guzman

Background

The world of data science is evolving so fast and data science was named the fastest-growing job in 2017 by LinkedIn. A recent study by PriceWaterhouseCoopers said, “the best jobs right now in America include titles like data scientist, data engineer, and business analyst.” As a result of this growing trend, our team will be analyzing the salaries of the data science profession based on years of experience, type of work you’re open to, your location and the size of the company you’re joining.

Motivation

- Interest to learn how well data science could make
- To understand the salary range of data related jobs
- To understand recent layoff trends and information

Requirements for Project:

1. Data and data delivery
2. Back end (ETL)
3. Visualiations
4. Group Presentation
5. Slide deck

Meet these requirements by using machine learning and we plan to use the following:

- Pandas
- Tableau
- HTML/CSS
- Matplotlib, Plotly or Tableau

Questions to explore

1. Map of where the jobs are
2. Who are the top salaried companies for 2017 to 2019 and compare with the percentage of layoffs.
3. Map of where the layoffs are
4. What was the stage of the company when the layoffs occurred (post-IPO, Series B, etc)?
    a. Correlation between layoff amounts and stage
5. How frequently has the tech industry had layoffs between 2020-2023?
    a. Do comparison with other industries

Machine learning: Create app where the model is used to predict salary based on input features

Potentially look at:

1. Graph if there are findings here: *Are there differences in salary based on gender or race (visualizations)*
2. Graph if there are findings here: *Education level, years of experience, job title, years at company etc*
3. Size of company and pay range
4. Will run statistical analysis on some results
5. Comparison of year over year data

Tools/Modules to use

Tableau (for map) or Plotly map

Data sets to use

- Salary: https://www.kaggle.com/datasets/jackogozaly/data-science-and-stem-salaries
- Layoff Info: https://layoffs.fyi/
- Layoff Data: https://www.kaggle.com/datasets/swaptr/layoffs-2022

Tasks Breakdown

- Preprocessing/Clean Data (Diane)
    - Get long/lat for cities
    - Extract the month and year
    - Convert the currency
    - Determine how to handle blank fields/misclassified fields
    - Use value counts to help determine relevancy of columns
    - Just use total yearly compensation
    - Drop level column
- Create the Machine Learning model (test multiple models, clusters, tf, linear) (Kelsey)
    - Target value is the salary
- Create graphs based on questions (Laura)
    - Who are the top salaried companies for 2017 to 2019 (from salary data)
    - Compare with the percentage of layoffs (from layoff data)
    - Top companies with layoffs

    - Layoffs by industry
    - Layoffs by location
    - What was the stage the company was in when the layoffs occurred (post-IPO, Series B, etc)?
    - Trend of layoffs by Quarter
- Run statistical analysis on key questions (Jiamin) correlation
    - Salary vs gender, education, race, location, experiences
    - Laid off vs stage, location, industry
- Create a location map of positions (Maybe Tableau, Plotly or Leaflet?) (Jiamin)
    - Map of where jobs are
    - Map of where the layoffs are
- Create app/HTML to run model (Xiaolin)
    - Drop down list to display the models/results based on selection

Tasks and Timeline

Date	Task	Notes	Check-In
5/15/203	Group Project Proposal is Due	Proposal is Due	Class
5/16/2023	Review proposal, work through individual pieces, clean data	Group GitHub site set up	Class
5/17/2023	Review proposal, work through individual pieces, clean data		6:30pm PST
5/18/2023 Project 4, Work Day 1	Data is cleaned, all working off cleaned dataset	Data is cleaned and ready for team to use	Class
5/19/2023	Individual work day, troubleshoot issues anyone comes up with		7pm PST
5/20/2023	Individual work day, troubleshoot issues anyone comes up with	Machine Learning piece complete Maps complete Graphs complete	10:30am PST
5/21/2023	Individual work day, troubleshoot issues anyone comes up		Individual work / work with team member(s) as

	with		necessary. Send update to team
5/22/2023 Project 4, Work Day 2	Finalize all parts of project, be prepared for mock presentation on 5/23/2023	Xiaolin to complete HTML/CSS Storyline creation	Class
5/23/2023 Project 4, Work Day 3	Project deliveries complete, talk track complete, entire group is ready for mock presentation	Final Group Work and Mock Presentations	Class
5/24/2023	Mock Presentations	Mock Presentations	7pm PST
5/25/2023 Project 4, Presentations	Presentation Day	Group presentations	

Presentation

Presentation Flow & Timing:

Opening: Diane (~30 seconds)
Salary Exploratory Info: Diane (~2 min)
- Top Paying Companies
- YoY Salary Changes by Job Title

Salary Statistical Analysis: Jiamin (~2 min)
Transition to Layoff Data and Layoff Analysis: Xiaolin (~2 min)
- Total Layoff Counts
- Trends by Quarter

Deeper Dive into Layoff Data: Laura (~2 min)
- Layoff Map, integrate top companies and US cities seeing layoffs
- Industry view

Machine Learning: Kelsey (~2 min)
- 2 Machine Learning pieces

Closing: Kelsey (~30 sec)

Team Background for Presentation: