Scuola universitaria professionale
della Svizzera italiana

SUPSI

DATA VISUALIZATION – DS&AI2

DATA ENGINEER JOBS – PROJECT DOCUMENTATION

SUBMITTED TO PROFESSOR GIOVANNI PROFETA

BY GEORGIY FARINA, NOMIN ENKH-OYU AND OLEG LASTOCICHIN

Abstract

This documentation was written for our final project of the course Data Visualization. The domain that was explored in this project is Data Engineer Jobs in different states from the US with the data collected from Kaggle. We were able to put into practice all that we have learned during the Data Visualization lectures using Python Notebook and the web development languages such as HTML, CSS and JavaScript. We would like to, especially thank Professor Giovanni Profeta for his guidance, reviews, and recommendations during the development of the project.

Table of Contents

Introduction

The purpose of this documentation is to provide a comprehensive overview of the project we made for the module Data Visualization, second-year Data Science and Artificial Intelligence. As we were instructed by the professor to select a domain of our interest, so that we could work on it efficiently, one common interest that we found between us was Data Engineer Jobs. Our primary interest as Data Science students and potential Data Scientists is based on the future expectations surrounding various technical careers and therefore the wages. Our main goal is to provide and highlight what others in similar industries, or even students like ourselves, could likewise expect by providing information and insights on the top paying states, industries and sectors, as well as the most common sector overall.

Data Sources

The Data Engineer Jobs dataset that we used for this project was taken from Kaggle, an online community platform that allows users to find and publish data sets. Moreover, it has been collected from Glassdoor which is an American website where current or ex-employees evaluate anonymously different companies. We also looked for a dataset that contains the latitude/longitude of the US states needed for plotting the visualizations.

Data Pre-Processing

The dataset has various information about engineering jobs in the US including their location, average salary, rating and many more. It contains 2528 samples and the following 15 features (from which 15 categorical variables and 3 numerical):

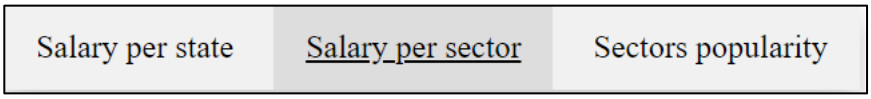
- Job Title: the title of the job
- Salary Estimate: range of the salary
- Job Description: description of the job
- Rating: rating of the job in float format
- Company Name: the name of the company
- Location: the location of where the job is located
- Headquarters: the headquarters of where the job is located
- Size: a range of how many employees are in the company
- Founded: the year when the company has been founded
- Type of ownership: the type of ownership (private or public)
- Industry: the industry of the job
- Sector: the sector of the job
- Revenue: the range of the revenues of the company
- Competitors: names of the competitor(s) of the company
- Easy Apply: if it is easy to apply or not (Boolean)
- City: the city where the job is located
- State: the acronym of the state where the job is located

Initial formatting of the data was "human readable," but later changes made it easier for us to aggregate and modify the data for the visualizations. We began the pre-processing of this data by identifying any missing values, outliers, and duplicate rows, 24 of which were eliminated. We saw that we had values like "-1" or "Unknown" representing missing values which we replaced them with “NaN” values so that we could impute them using KNN imputation. Using the percentiles method, we checked for outliers and removed any that were present because, given the size of our data, their exclusion wouldn't have had a significant impact on our research.

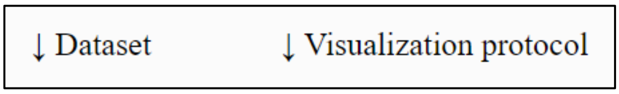
Interface design

With the help of the template that Professor Profeta has provided us, we customized it to meet our requirements. A header describes briefly what the webpage is about. After the header, follows a centered designed body of the page containing various sections, each of which describe something specific regarding our findings and visualizations.

The first section contains a choice of three major visualizations that we created using buttons that, when clicked, alter the image to the user wants to see. The three main images are describing respectively the following findings:



Below this image, there are present two links that allow you to download both the dataset and the visualization protocol:



At the end, before diving into the detailed description of every visualization, there is present a short abstract explaining our reasons for researching this topic and what we are expecting to find. Following the latest, we have a section for every visualization, with a description of what we obtained and some comments on the results. There are present three sections, in the same order of the visualization choices in the beginning of the page, where the first and last are simple description-visualization-conclusions combination, whereas the second one has also the selection of the visualization to display.

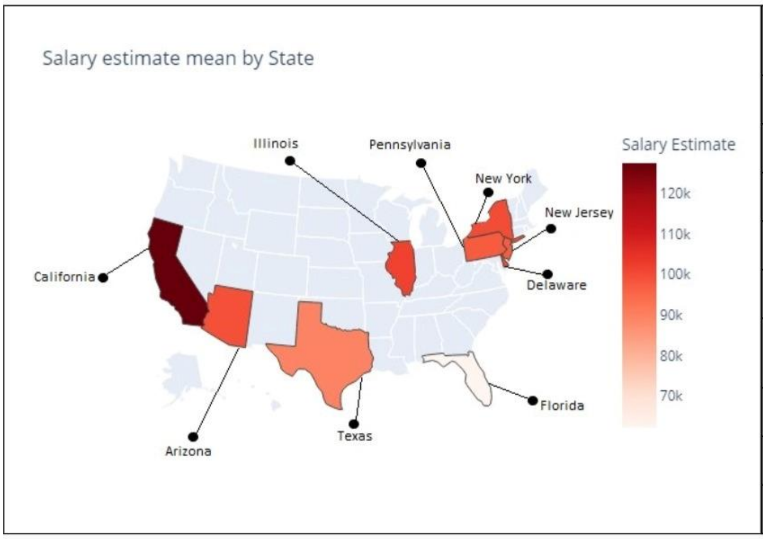
After the visualizations part, the last two sections are respectively the overall conclusions and the footer containing general information regarding this project.

Data Visualizations

The majority of the visualizations come in a map format except for the last one being the sectors distribution, represented as a pie chart. All the visualizations have been generated through the usage of a Python notebook.

Salary by state:

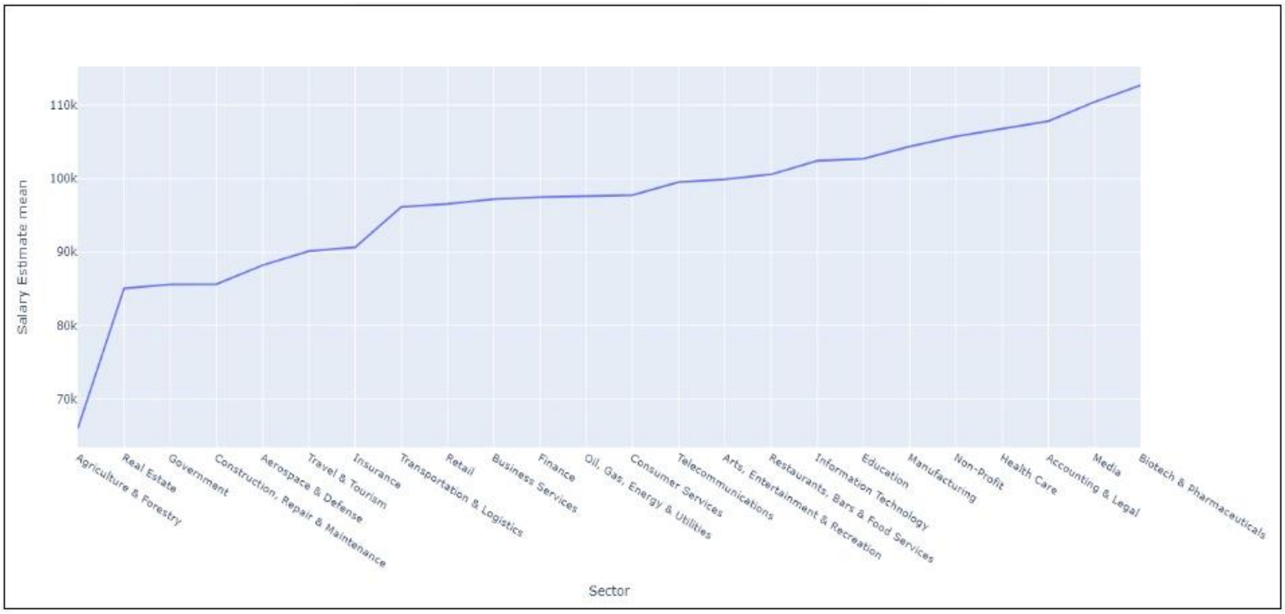
This visualization is a choropleth map using color scales to help the reader understand which the average salary estimates are for each state. We have used a sequential and unclassed color palette with the colors shading the respective geographic area from a lighter shade of red to a darker one, correlated to different average salaries. Shortly said, the darker is the color, the higher is the average salary. The visualization's right side features a table with three columns: "State," which contains the name and acronym of a particular state, "Average salary estimate ($)," which, as the name suggests, represents the average salary estimated in USD dollars, and "Amount of Data," which indicates the number of workers employed in a given state.



State	Average salary estimate ($)	Amount of data
California (CA)	127'000	530
Illinois (IL)	101'800	182
Delaware (DE)	101'000	12
New York (NY)	100'000	160
Arizona (AZ)	99'000	163
Pennsylvania (PA)	97'400	158
New Jersey (NJ)	97'000	23
Texas (TX)	90'000	901
Florida (FL)	62'000	118

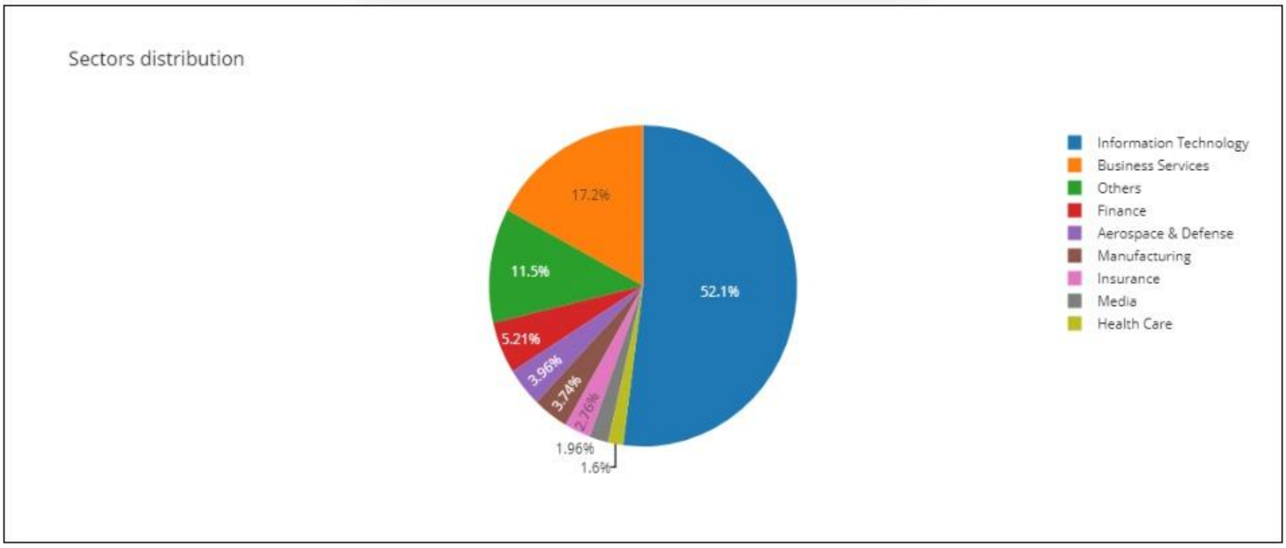
Salary estimates by sector:

This line plot provides insight into the industries that pay more on average. We selected the industries based on average salary to create an ascending line that highlights the disparity between some industries and others.



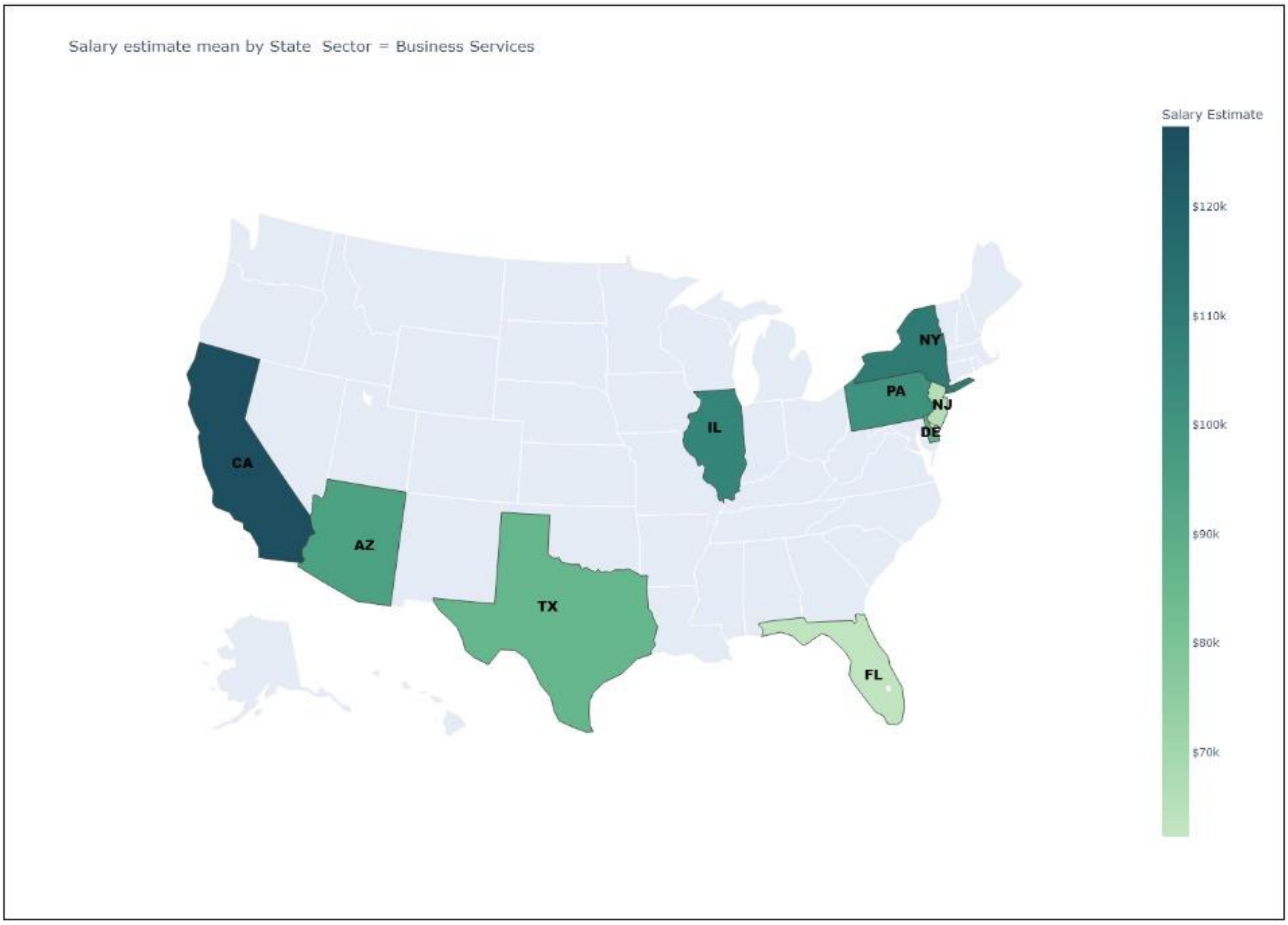
Sectors distribution:

The graph that we wanted to show the distribution of the sectors with is the pie chart. We represented it using this graph because we thought it being the best one to visualize the disparage of the sectors and the dominance of certain ones with respect to others.



Salary per state and sector:

As the first one, this is a choropleth map using color scales to help the reader understand which the average salary estimates are for each state. The difference with respect to the first one is that every one of these plots represents the salary estimate for every one of the sectors present in the analysis. Depending on the selection from the website, there is going to be a separate visualization representing the salary estimate for the chosen sector. The color in this case is different from the one used previously because we wanted to associate a color for every different kind of visualization: the first color (red) was referring to an overall estimate for every state, whereas this one (green) is related to a visualization for a specific sector.



Results

Salary by state

The highest paying state out of this list is California, which is also the second one in terms of estimations collected, so we can confirm the reliability of such salary. It was a result that we could've expected, having of course the famous Silicon Valley in that state. At the bottom of the list we can find the state of Florida, with a massive gap in salary with respect to the second to last, being Texas. Except for these two extremes, the other states have an average variation of salary between each other, where 6 of the states (from Illinois to New Jersey) vary in a range of roughly 5'000, and Texas is a bit further from this range, placed at the 8th place with a 90k expected salary.

Salary by sector

Out of the 24 sectors analysed, the top three in terms of yearly salary are:

1. Biotech & Pharmaceuticals: $ 112'700
2. Media: $ 110'000
3. Accounting & Legal: $107'800

Except for these three leaders, the other sectors are evenly distributed between 80'000 and 105'000, except for Agriculture & Forestry being the "poorest" sector with an estimated salary of just 66'000 dollars.

Sectors distribution

There is an evident dominance of the Information & Technology field with more than a half out of the ~2500 Data Engineering positions analysed. The top three is then completed with the Business and Finance companies. Following these three, there is also an "Others" slice in the chart, being all the sectors that contribute less than to 1.6% of the total amount of positions analysed. This means that a total of 16 sectors make up to 11.5%, 1/5 of how much the Information & Technology does.

Next steps

Overall, we are content with the project's results. Having it concluded, we had an overview of what has been done and we thought about what could be done next.

First of all, the dataset we found contains scraped data about just 8 states out of the total 50. An improvement could be done by scraping the data of all the other remaining states to obtain a better insight on what both the salaries per state and per sector could look like in the whole USA.

Another one that we thought would be an interesting future upgrade is to add data from other nations, in order not to have only USA data, by scraping data from Glassdoor and obtaining data from many different countries. With this there could be also done an interactive map, where the user could traverse and look for countries they are interested in. This would be a better alternative to the current static images, and would definitely improve the user's experience.

References

- Kaggle website - https://www.kaggle.com/datasets/andrewmvd/data-engineer-jobs
- Aishwarya Ramakrishnan, *Data Storytelling with Maps,* (accessed: 20.01.2023)