DATA101

Final Project Report - Group 8

Borro, Domingo, Paden, Roque

Target Audience

The visualization project aims to assist **job seekers**, such as students looking for part-time work, fresh graduates, underemployed, and unemployed individuals. This demographic consists of people who are either canvassing career paths/options or searching for job opportunities that fit their needs. These users do not necessarily belong to a specific type, but they are mostly part of the general public.

Through interactive data visualization, the project enables these users to explore salary ranges across various industries, skill sets, companies, and locations. However, the complexity of the dashboard and dataset could present challenges for visualizing such tasks effectively.

Given the large volume of data (124,000+ job postings from 2023 and 2024, along with various other variables), the dashboard's navigation and interactive features must be intuitive. Users should be able to easily understand, interact with, and extract insights.

A common issue with overly complex interactive dashboards is the ambiguity of task abstraction. Users may struggle to comprehend the interactive features (such as dropdowns, sliders, or clickable charts), particularly considering the varying levels of data literacy among the target audience.

Ultimately, the dashboard should be simple to use without sacrificing the ability to personalize the visualization through effective filtering. While it can be assumed that job seekers may generally be literate enough to gather insights from interactive data visualizations, it is still better to be as inclusive as possible — most especially with students who might not have prior experience being exposed to such tools.

Visualization Problem

LinkedIn is a common tool utilized by job seekers to look for employment opportunities. However, depending on the user's familiarity with LinkedIn's functions, they may find it difficult to ascertain if certain job postings are the best options available to them or if these postings are competitive by industry or location standards. They may struggle to find accurate and comprehensive salary insights when evaluating career opportunities as it requires them to manually search and compare job listings. It becomes difficult for them to analyze and compare salary trends effectively.

An interactive dashboard would address this issue as it will effectively visualize salary data, which can enable users to:

- Explore salary distribution across different industries, skill sets, companies, and locations
- Compare salaries by industry, skill set, company, and location

Dataset Description & Connections

The primary dataset that will be utilized for the visualization project is the LinkedIn Job Postings (2023 - 2024) dataset from Kaggle, which contains the locations, companies, industries, salaries, and other relevant information about job postings found on LinkedIn. Geospatial data of the United States of America State Boundaries taken from Opendatasoft will also be utilized, which will be used to visualize the data per state.

Collection Method

The job postings dataset was compiled using web scraping techniques to extract job postings from LinkedIn over a specified period. The data includes structured fields such as job title, industry, location, and salary estimates. The USA State Boundaries data, on the other hand, was provided by the U.S. Census Bureau.

Dataset Structure

The job postings data can mainly be divided into two (2) distinct datasets. The first dataset is a dataset of *companies*, while the second dataset is for *job postings*.

Companies Dataset: Each row in the companies dataset represents one company. It contains 4 separate csv files related to the hiring companies, namely:

- companies.csv
- company_industries.csv
- company_specialties.csv
- employee_counts.csv

These files are connected through the key identifier variable **company_id**, which is the unique identifier for each company. The following are the attributes of a company that can be found in the dataset:

Variable	Type	Description
name	string	Company name
description		Company description
company size	int	Company grouping based on the number of employees (0 smallest - 7 Largest)
state	string	US state where the company headquarters is located
country		Country where the company headquarters is located
city		City where the company headquarters is located
industry		Main industry of company

Job Postings Dataset: Each row in the job postings dataset represents one job posting. The dataset contains 5 separate csv files related to each posting:

- postings.csv
- benefits.csv
- job_industries.csv
- job_skills.csv
- salaries.csv

These csv files are connected through the **job_id** key identifier variable. Each job also has a **company_id** variable, which is the id of the company that posted the job and connects this dataset to the companies dataset. The following are relevant attributes that can be found in this dataset:

Variable	Type	Description
title	string	Job title
description		Job description
max_salary	float	Maximum salary offered in the job posting
med_salary		Median salary offered in the job posting
min_salary		Minimum salary offered in the job posting
pay_period	string	Pay period for salary (Hourly, Monthly, Yearly, etc.)
work_type		Type of work associated with the job (Fulltime, Parttime, Contract)
experience_level		Job experience level (entry, associate, executive, etc.)
skill_name		List of skills associated with the job
industry_name		Industry of the job
benefit_type		Type of benefit provided with the job (401k, Medical Insurance, etc.)

Initial Dataset Exploration and Cleaning

The dataset was partially explored and analyzed in the following Jupyter Notebook. This included the merging of files into a single dataframe, dropping unneeded columns, and mapping the correct categorical variables to specific features such as skills and industries. After merging the needed variables for the visualization, the initial dataframe looks like this:

	job_id	title	company_name	company_size	min_salary	med_salary	max_salary	pay_period	formatted_work_type	formatted_experience_level	city	state	industry_name	skill_name
0	921716	Marketing Coordinator	Corcoran Sawyer Smith	2.0	17.0	NaN	20.0	HOURLY	Full-time	NaN	Jersey City	NJ	[Real Estate]	[Marketing, Sales]
1	10998357	Assitant Restaurant Manager	The National Exemplar	1.0	45000.0	NaN	65000.0	YEARLY	Full-time	NaN	Mariemont	Ohio	[Restaurants]	[Management, Manufacturing]
2	23221523	Senior Elder Law / Trusts and Estates Associat...	Abrams Fensterman, LLP	2.0	140000.0	NaN	175000.0	YEARLY	Full-time	NaN	Lake Success	New York	[Law Practice]	[Other]
3	91700727	Economic Development and Planning Intern	Downtown Raleigh Alliance	1.0	14.0	NaN	20.0	HOURLY	Internship	NaN	Raleigh	North Carolina	[Non-profit Organization Management]	[Project Management]
4	103254301	Producer	Raw Cereal	NaN	60000.0	NaN	300000.0	YEARLY	Contract	NaN	Los Angeles	CA	[Design Services]	[Design, Art/Creative, Information Technology]

Figure 1. Initial Dataframe

The dataset was then further cleaned using this Jupyter Notebook. Firstly, the rows where the pay_period column was NaN were dropped, as standardizing the salary column would not be possible with null values. Next, values in the med_salary column were imputed by getting the average of the min_salary and max_salary. Moreover, rows with missing min_salary and max_salary values were dropped as imputing those from the given variables was not possible. Once the missing values were addressed, the pay_period column was used to standardize all the salaries to represent an annual amount. The work_type was also considered, ascertaining that the conversion for part-time work would be more accurate.

Looking at the distribution of salary, the group noticed that the histogram was incredibly skewed, which meant there were outliers — 231 in med_salary to be exact. These values were dropped to ensure that they would not affect the visualizations.

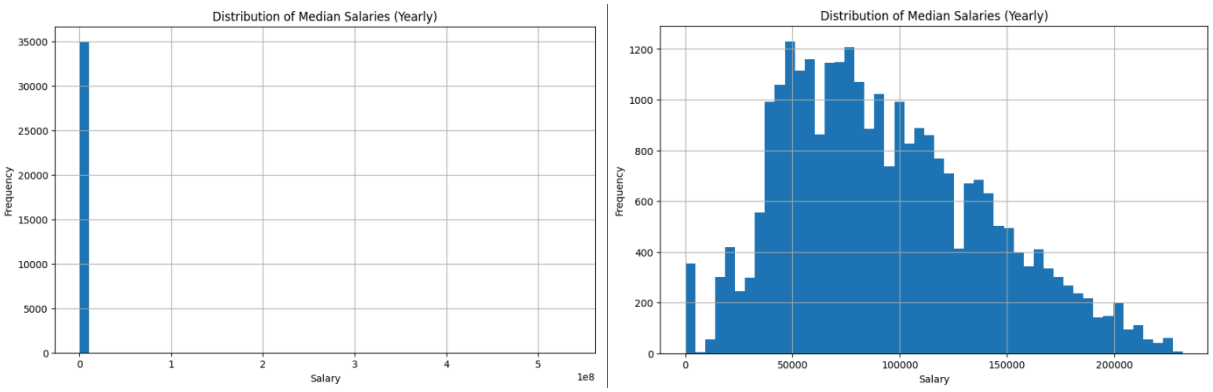

Figure 2. Histogram of Salaries Before (Left) and After (Right) Removal of Outliers

Lastly, the state column was cleaned. The initial dataset had the states listed in many different ways with no standardization, even including states that were not in America. To address this, a dictionary was made with all the unique values from the state column, where each value was then assigned to its corresponding state acronym. This dictionary was then used to correct and standardize the values in the dataset. Rows with states outside of America were removed.

Prototype Design/s

The group has decided to use Canva to design and sketch prototypes for the dashboard. The initial iteration contains a simple layout to give an idea of where each visualization will be located. This iteration is a simple one-page design with all the visualizations on one page. Colors, channels, and layout are all subject to change. The following variables will be used as metrics for the data visualizations:

- The minimum, median, and maximum salary of the job postings. This will be the variable that will connect all the visualizations.
- Industry, Skill, and Company are all categorical variables that will be used together with the salary variables in a bar chart to show a descending ranking.
- Location, specifically US states, will be used to filter the data.

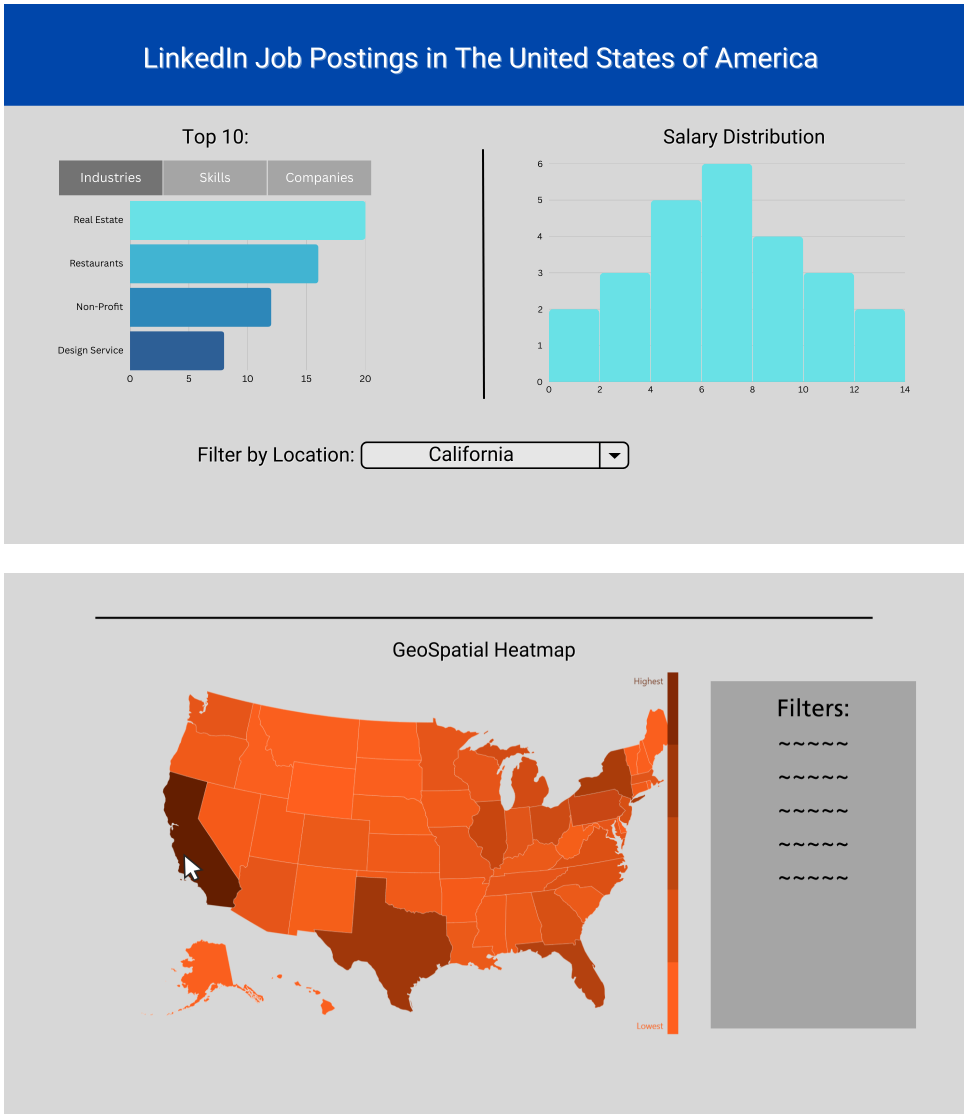

Figure 3. Initial Prototype of Project Dashboard

Visualization Choices

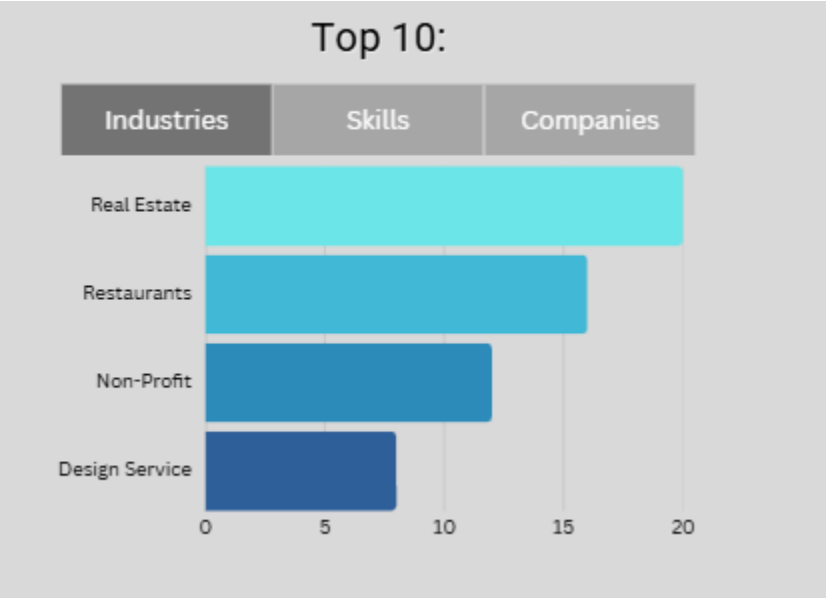

Figure 4. Bar Charts of the Top 10 Salaries per Industry/Skill Set/Company

Idiom	Bar Chart [Rankings - Descending]
Data	X-Axis: Min/Med/Max Salary Y-Axis: Industries/Skills/Companies
Channels	Blue Gradient
Task	Present Extremes

Figure 5. Histogram of the Salary Distribution

Idiom	Histogram
Data	X-Axis: Salary Amounts Y-Axis: Frequency
Channels	Blue
Task	Present Distribution

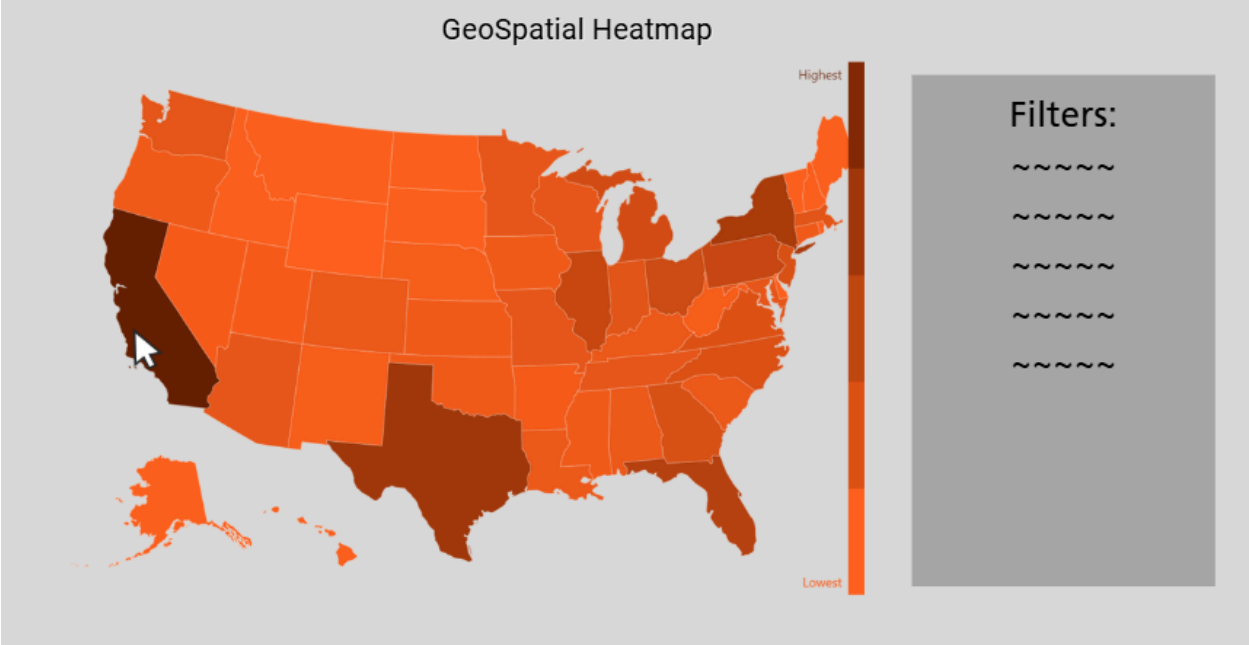

Figure 6. Choropleth of Min/Med/Max Salary per US State

Idiom	Choropleth
Data	Min/Med/Max Salary, US State Boundaries
Channels	Red Gradient Low Average Salary State: Pale/Light Beige Medium Average Salary State: Orange High Average Salary State: Red
Task	Lookup Features

Interactivity Techniques & Justifications

The bar chart, histogram, and choropleth would be connected with each other. These charts can be filtered by state via a dropdown box, which changes the values, and thus, the visualizations. This interaction was implemented to enable comparison of salary data between different states as well as allowing the bar charts and histograms to present more information effectively, as they only show what the user needs them to show, rather than visualizing salary information for all states at once.

In the same vein, the bar charts also have clickable options on top that change the values on the y-axis, depending on if the user wants to see salary information for industries, skills, or companies. This interaction was implemented to allow the dashboard to present more information to the user and also allows the bar chart to convey such information more effectively by only showing the specific information the user wants to see, rather than overwhelming them with all the information at once. It is salient to consider that there may be users as well who are not literate in terms of identifying clickable buttons. To reduce such possibilities, color distinctions or gradients would be used rather than having the buttons be of the same color.

Users can also hover over all three of the visualizations with their cursor to see more specific information. For example, hovering over the bins on the bar chart or the histogram will give the specific salary and count for that bin. This interaction was implemented to also allow the dashboard to present more information the user may want to see or need and helps these visualizations convey more detailed information about what they are showing.

Final Application

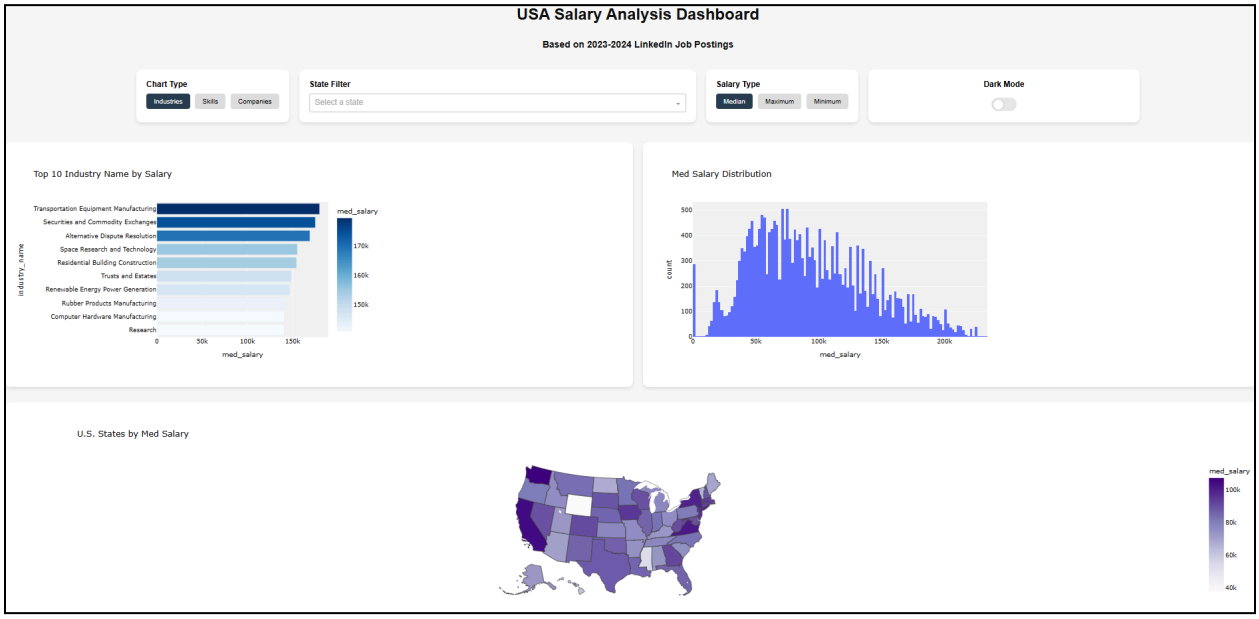

Figure 7. Final Dashboard Application in Light Mode

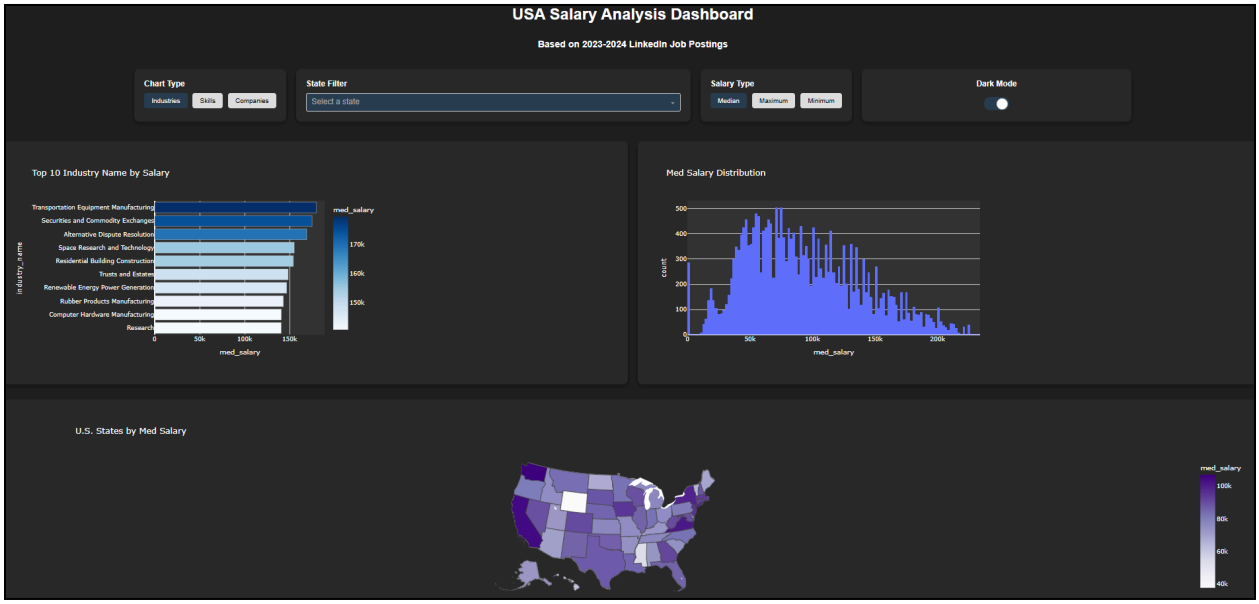

Figure 8. Final Dashboard Application in DarkMode

Link to Application: https://salary-analysis-dash-app.onrender.com/

Link to GitHub: https://github.com/DomDom727/DATA101-Dash-App.git