An Exploratory Analysis
on Data-related roles
In the US Job Market

Prepared by The Trendsetters
Prepared for Norman Lo

December 12, 2024

Table of Contents

Executive Summary

The Salary dataset provides detailed information on the US data-related jobs market. Data-based insights help job searchers find opportunities based on desirable characteristics examined in the data. Using paid wage as the key indicator, this variable is examined across other variables such as State, Job Type, Education, Experience, and Prevailing Wage.

Question 1

Question 2 explores paid wages based on state data, job type, and cost of living. This provides valuable information for job searchers who are interested in desirable states and want to know what types of wages and jobs they can expect in these areas.

Question 3 explores offered salaries compared to the prevailing wage across job types and brands, and does the answer change when considering individuals' financial circumstances?

Question 4 takes the state analysis further and looks at the regional characteristics of wages and jobs.

Question 5

Question 6

Question 7

Ultimately, the data story examined provides key insights into the desirable characteristics associated with different data-related jobs on the US market, better informing job searchers and interested parties.

I. Background, Objective, & Goals

The dataset being explored provides salary information on data-related jobs in the US job market. Exploring this dataset helps examine the salaries of these jobs across qualitative and quantitative characteristics. This enables job seekers to be more informed when deciding which companies and positions to apply for in their job search.

The key performance indicator of the analysis is Paid Wage per Year. This is the comparable basis for the data and visualization found throughout the report. In addition to this KPI, other notable characteristics are: Employer Name, Education Level, College Major, Experience, Prevailing Wage, Work City/State, and Job Title Subgroup. These characteristics make up the focus data from the extracted dataset. Processing, cleaning, and manipulating the focus data within a software environment yields a more effective analysis by removing irrelevant data. This improves data integrity, accuracy, and visualization.

The objective of this exploratory analysis is to find insights on data-related jobs according to desirable characteristics such as Job Type, Location, Prevailing Wage, Company, Company Size, and Job Requirements.

The goal is to interpret this data through several high-level questions that look deeper into these characteristics and look for any disparities that may exist. Data from other datasets may be implemented into the analysis to accurately depict any disparities affecting the KPI or other characteristics.

Group members address the following questions:

- **Question 1** (Vinh) - Do specific sub-types of data-related jobs have higher or lower salaries than others?
- **Question 2** (Reese) - What states have the highest paying data-related salaries ?
- **Question 3** (Swikar) - How do offered salaries compare to the prevailing wage ?
- **Question 4** (Reese) - How do salaries differ per geographic region ? Are there specific data-related roles found in these regions ?
- **Question 5** (Vinh) - What specific subtypes require higher experience and education, is this correlated with salary ?
- **Question 6** (Kshitij) - What salaries are "normal" when looking at a distribution of salaries. Are there any trends or patterns found within this salary distribution?
- **Question 7** (Kshitij) - How does company size or industry influence salary levels for data-related roles

II. Question 1

III. Question 2

Question Two is composed of four subquestions:

- (1) - What states have the highest paying data-related salaries ?
- (2) - Differences between job subcategories ?
- (3) - Which companies have the highest salaries of those sub-types ?
- (4) - Will the answer change if I take the standard of living into account ?

Using R Studio, analysis will explore these questions helping provide data-based insights. This starts by loading in the necessary packages to help read the excel file, clean the data, process the data, and then organize it into meaningful information. I began constructing a data frame for the salary dataset by focusing on desired columns of the dataset, helping remove any irrelevant columns. Then a data cleaning was performed by mutating missing entries to show “NA” for string data and “0” for numerical data. The next step was to create a regional dictionary based on the states for more effective regional analysis. This is something that wasn’t present before so a new column is created for proper referencing. Additionally, salary bins were created to help group the salaries into bins of $20,000. This data preparation helps transition into the next part of analysis.

For the question “What states have the highest paying data-related salaries ?” I began filtering data by average salary per state, categorized them by region, and selected only the top ten highest average salary states. The resulting visualization depicts the top 10 states along the x-axis, the average salary for these states on the y-axis, and fills the area with colored bar charts indicative of their corresponding regions. The story in Visualization 2.1 depicts the top 10 states in the following order:

- (1) West Virginia - $109,426 (South)
- (2) California - $103,571 (West)
- (3) Washington - $102,176 (West)
- (4) New York - $91,601 (Northeast)
- (5) Arkansas - $90,270 (South)
- (6) Alabama- $87,326 (South)
- (7) Massachusetts - $86,610 (Northeast)
- (8) Pennsylvania - $83,889 (Northeast)
- (9) District of Columbia - $81,968 (South)
- (10) Mississippi - $81,950 (South)

This provides a great glimpse for job searchers to understand which states provide higher average salaries compared to others. Regionally, data analyst roles are paid quite high in most parts of the US except for the midwest. However, results must dig deeper into the job types and employers that compose the dataset to see how they influence salary numbers.

For the question “Differences between job subcategories ?” I constructed a histogram depicting the frequency of jobs by their salary bins. The columns are composed of the relative frequency of each specific job subtype. This process was accomplished by constructing a piece of code that put salary bins along the x-axis, frequency along the y-axis, and a fill function of the job subgroup. This depicts the normal distribution of salaries in the data-related jobs market.

Looking at Visualization 2.2 provides insight into salaries based on types of jobs. Some noticeable features include:

- Software Engineers make up a bulk of the market and have great salary potential, mostly ranging from $60,000 to $150,000. The distribution skew is telling of the high concentration of higher salary jobs. This depicts software engineering's high demand and value on the market.
- Business Analysts and Assistant Professors make up the next ranking concentration and are both fairly similar in the salary range of $50,000 to $100,000. These job types are the middle ground of the salary distribution reflecting high demand and average value.
- Teaching jobs make up the left side of the distribution of salaries mostly between $30,000 to $60,000 showing that teaching jobs are high demand yet low value.

To go more in depth, a histogram can be made to depict each individual job type salary distribution of lower demand jobs that aren’t represented in the overall graphic. Using the same data, Visualization 2.3 shows each job types distribution and its frequency providing the following additional insights:

- Management Consultants have great salary potential ranging mostly from $50,000 to $160,000 with a peak around $120,000
- Data Scientists have normal salary distributions spanning $50,000 to $150,000 peaking at $120,000.
- Data Analysts have a salary distribution spanning $50,000 to $100,000 with a peak at $60,000.
- Attorneys have a bimodal salary distribution spanning $50,000 to $200,000 and upward centered at $90,000 and $160,000.

For the question “Which companies have the highest salaries of those sub-types ?” I began constructing a data frame based on the average salary of those employers and grouped it by job subgroup, employer, and state. This was then arranged into descending order with the top 5 employers displayed. From this point, a set of dot plots can be constructed displaying the salary along the x-axis and companies & state on the y-axis per job group.

Visualization 2.4 helps interpret the highest paying average salaries per company and their locations. This provides insights on the top five companies in each category such as:

- Netflix providing higher average salaries for the categories of data scientist, data analysts, and management consulting
- High average salaries for attorneys are located in New York
- Highest Assistant professor salaries are a part of medical foundations
- Highest frequency states are: California, New York, New Jersey
- The top 5 companies of each subgroup typically pay around $160,000 and upward

For the question “Will the answer change if I take the standard of living into account ?” additional data is required before beginning analysis. Using Kaggle as an online dataset resource, I imported a zip file of Cost of Living information collected by Missouri Economic Research. This was then loaded into the R studio environment and merged with salary data across the state abbreviation category, successfully joining the two datasets. With this achieved, an adjusted salary was found by dividing the paid wage by the conversion factor and multiplying it by 100. From here, the average adjusted salaries were found for each state. A histogram can then be made depicting the top 10 states along the x-axis and the salary amount on the y-axis.

Visualization 2.5 tells a great story of the realistic salary expectations for the highest average salary paying states. The following insights can be made:

- The southern states of West Virginia, Alabama, Mississippi, Arkansas, and Kentucky have the highest average salary all hovering around $100,000 except for Kentucky which sits around $85,000
- The midwestern states of Kansas, Iowa, Indiana, and Missouri hover around $90,000
- Washington is the only west coast state with a high average salary accounting for the cost of living and sits around $90,000

IV. Question 3

Question 3 answers the following questions:

- “How do offered salaries compare to the prevailing wage ?”
- “Are there types of jobs that are paid too much or too little?”
- “Are there brands that tend to pay too much or too little?”
- “Will the answer change if I think about how much money people have?”

Answering the question ““Are there types of jobs that are paid too much or too little?”
The "Average Salary Difference by Job Subcategory (Offered vs. Prevailing)" table shows big differences between the average wage and the wages paid for different job categories. For example, lawyers and assistant professors are regularly paid $39,765 and $36,023, respectively, more than the prevailing wage. Software engineers and data scientists, on the other hand, are paid $8,437 and $17,475 more than the prevailing wage, but their pay goes down when cost-of-living changes (COLI) are taken into account. This means that when COLI is taken into account, high-paying jobs in places with high cost of living may not be as appealing. However, jobs like data analysts and business analysts are still competitive even though they are slightly overpaid by $7,277 and $6,550, respectively. This information is enough to tell which jobs are paid too much or too little, and it shows which groups have salaries that are in line with the industry standard, with and without COLI adjustments.

Answering the question “Are there brands that tend to pay too much or too little?” The "Top Companies with Highest Salaries for Each Data Job Subtype" table offers valuable insights into which employers pay exceptionally high salaries. For example, The University of Texas System Administration pays business analysts $677,508, significantly exceeding the industry average, while Netflix leads for data scientists at $220,000. Companies like Intuit and Knowledgent Group are top employers for data analysts, offering $433,161 and $185,000, respectively. While this data provides clarity on overpaying companies, there is no comparable table or metric for underpaying employers. If data on salary differences relative to prevailing wages by employer is available, it could help identify consistent underpayers. Without such data, analyzing underpayment trends across companies would require additional information. The available insights, however, are sufficient to identify which brands or organizations offer the most competitive compensation in their respective categories.

Answering the question “Will the answer change if I think about how much money people have?” The "Average Adjusted Salary Difference by Job Subcategory (Considering COLI)" table makes it clear how cost-of-living adjustments (COLI) affect how much people think they are paid for their work. For example, with adjusted differences of $34,734 and $23,134, respectively, assistant teachers and management experts are still paid way too much. When COLI is taken into account, on the other hand, software engineers and data scientists lose their pay benefits, with differences of -$2,797 and -$750, respectively. This shows how their low competitiveness is hurt by high living costs. This research is especially helpful for people who are looking at job offers in different places, since higher pay may not always mean more money in the bank. When you change the data for cost of living, you can see how the cost of living affects pay and see if salaries are in line with what workers are actually worth. This makes sure that choices based on numbers that have been adjusted for COLI are more accurate reflections of actual income potential.

V. Question 4

Question 4 seeks to answer the questions:

- (1) - How do salaries differ per geographic region ?
- (2) - Are there specific data-related roles found in these regions ?

These questions try to find the regional frequencies of job types on the basis of salary. Using the data frames from Question 2, a similar approach can be taken to conduct analysis. Each histogram has salary on the x-axis, frequency as the y-axis, and a column fill showing the relative frequency of each region spanning the different job groups.

Visualization 4.1 depicts the following insights:

- High relative frequency of jobs in lower salary roles across all regions for Assistant Professors beneath $120,000. There is a low concentration of jobs in the west for these job types compared to other regions.
- Management Consultants have an evenly distributed relative frequency mix above $40,000. There is a low concentration of jobs in the west for these job types compared to other regions.
- Evenly distributed frequency of jobs spanning lower salaries in all regions for Data Analysts & Business Analysts. There is a low relative frequency of jobs in the west for these job types compared to other regions.
- High relative frequency of Attorneys in the Northeast across most salary bins above $80,000.
- High relative frequency of Data Scientists and Software Engineers in the Western region spanning higher salaries peaking around $100,000 to $120,000 for both. Other regions peak at lower salaries around $60,000 to $80,000.
- High relative frequency of Teachers in the South that are under-paid peaking around $60,000.

‘

VI. Question 5

VII. Question 6

VIII. Question 7

IX. Conclusion

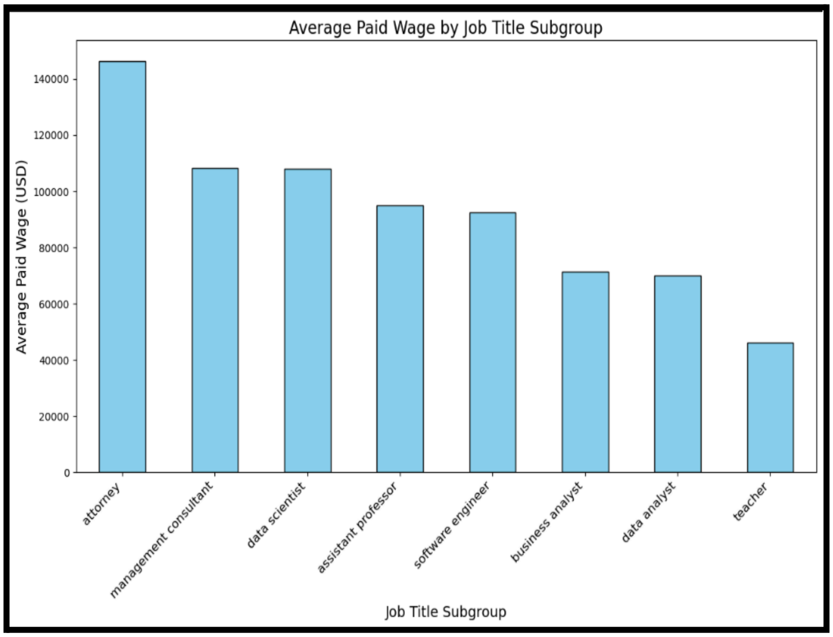
In conclusion, exploring the dataset constructs a transparent data story to be told by characteristics such as State, Job Type, Education, Experience, Prevailing Wage, and Adjusted Cost of Living. It depicts a diverse job economy across the US, where opportunity flourishes in unique combinations of all characteristics. Some areas flourish more than others, and disparities exhibited by the data expose unfortunate truths that exist in a modern world. However, with data analysis data-driven insights capitalize upon the existing information so then better decision-making can take place.

For job searchers, finding a job entails so much more than the job itself. The focus data for this project examines all influential factors determining the direction of one’s career. This refines the ability to make a more well-informed decision in the beginning of a journey or even in the middle of one. It provides a realistic view of what to expect.

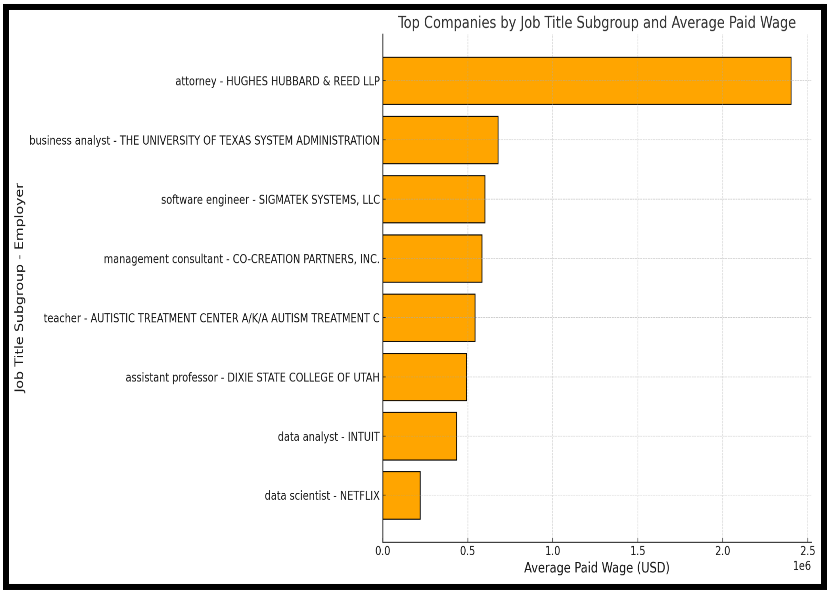
For interested parties, this data exploration provides an opportunity to shape areas that need development. Creating a well-balanced job ecosystem provides more available options for the workforce. Having large concentrations of certain characteristics isn’t necessarily negative. Once could invoke comparative advantages to regional job markets in the US. However, it is important to note that specialized job markets also have a range of economic implications as well, especially when it comes to salary differences. Therefore, getting more insight into the data allows for more questions to be asked, and further analysis to be conducted.

Technical Appendix

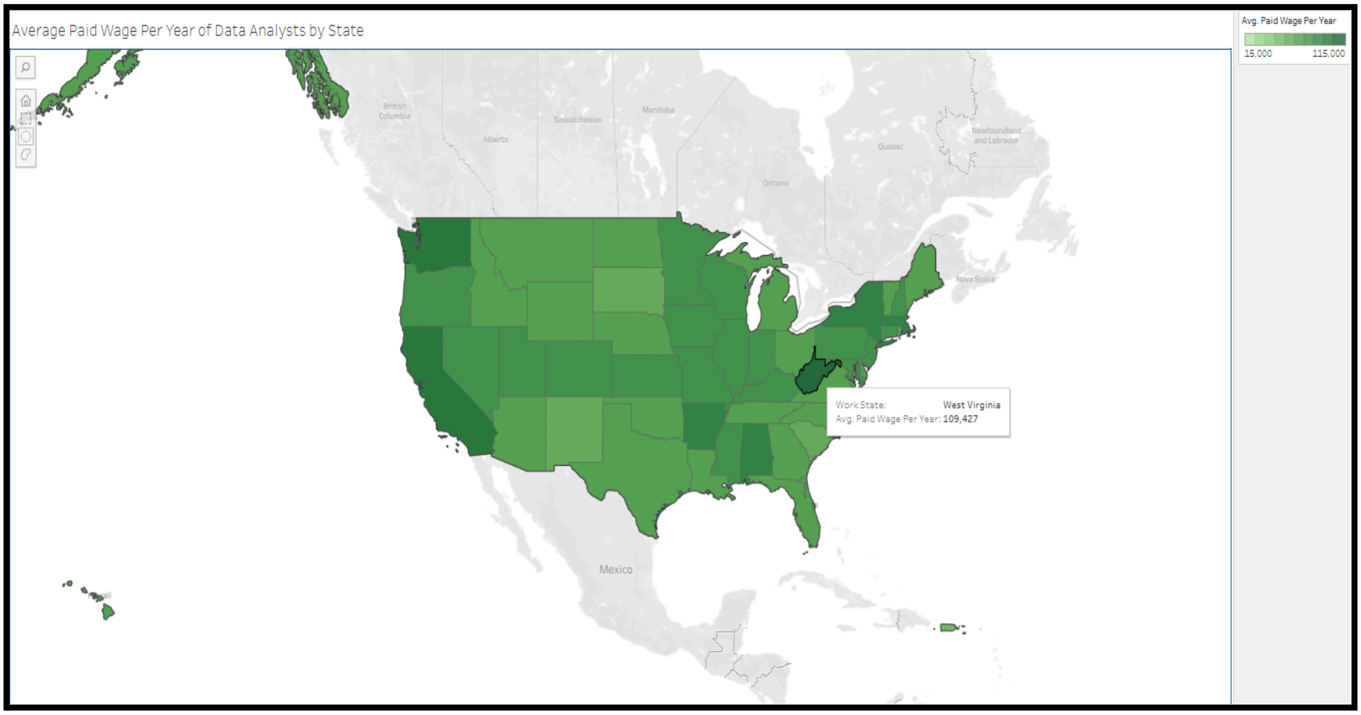
Visualization 1.1



Visualization 1.2



Visualization 1.3

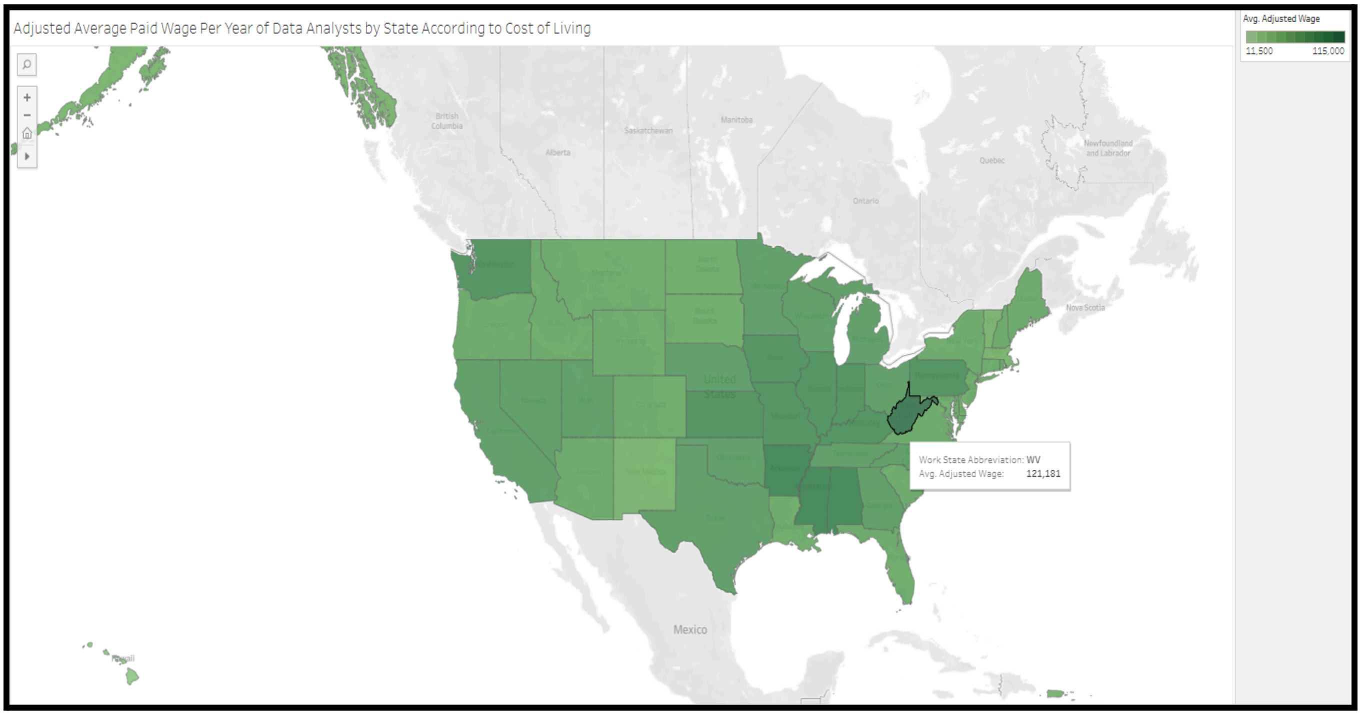


Average Paid Wage Per Year of Data Analysts by State

Avg. Paid Wage Per Year
15,000 115,000

Work State: West Virginia
Avg. Paid Wage Per Year: 109,427

Visualization 1.4

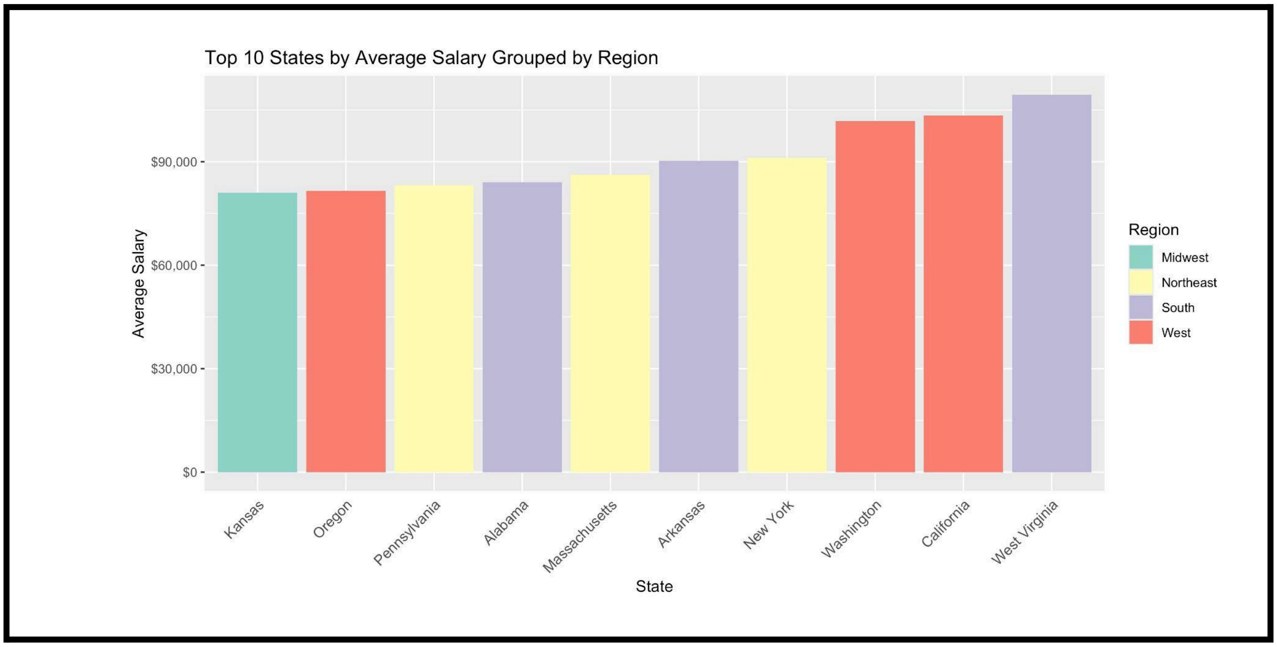


Adjusted Average Paid Wage Per Year of Data Analysts by State According to Cost of Living

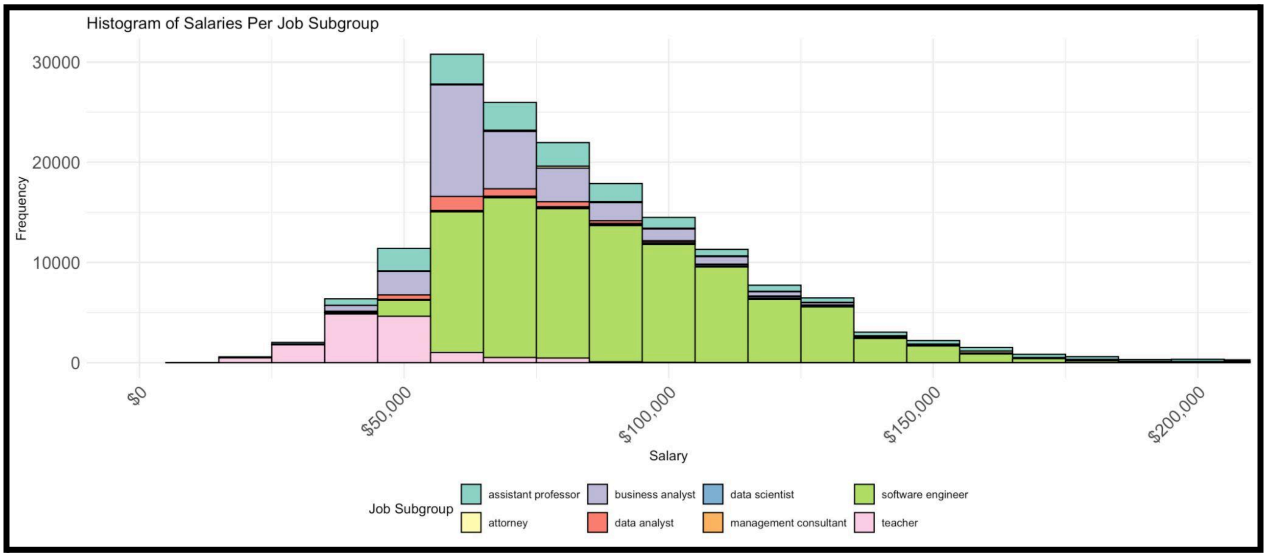
Avg. Adjusted Wage
11,500 115,000

Work State Abbreviation: WV
Avg. Adjusted Wage: 121,181

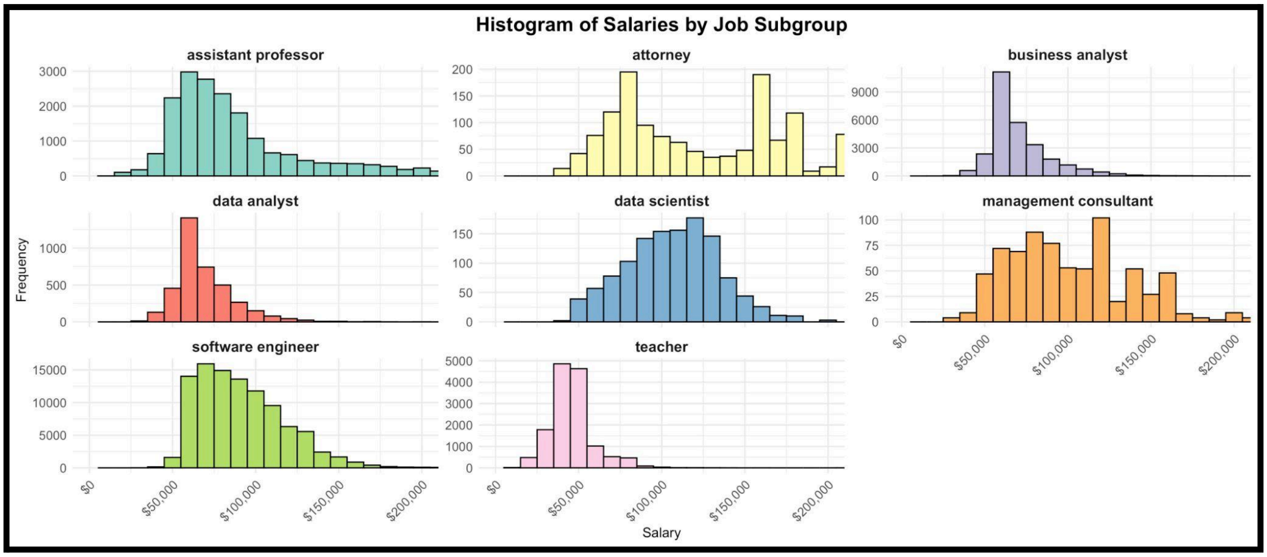
Visualization 2.1



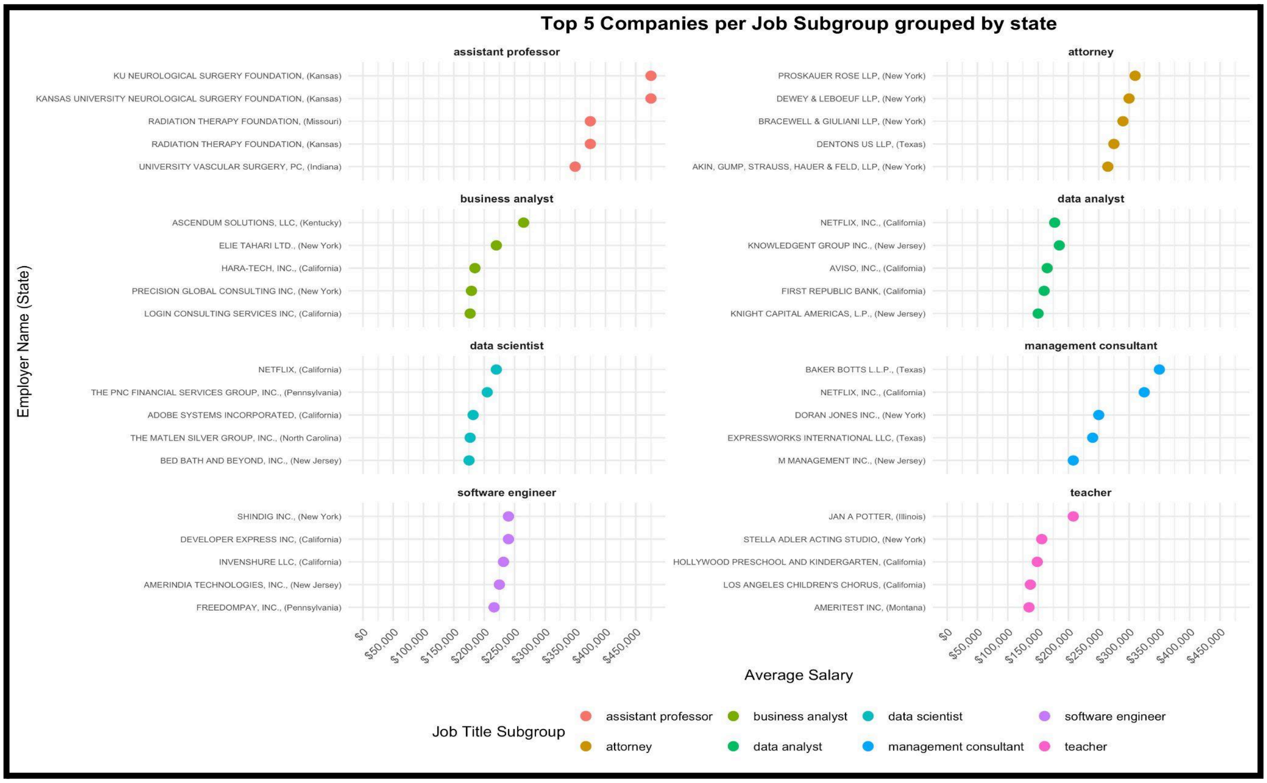
Visualization 2.2



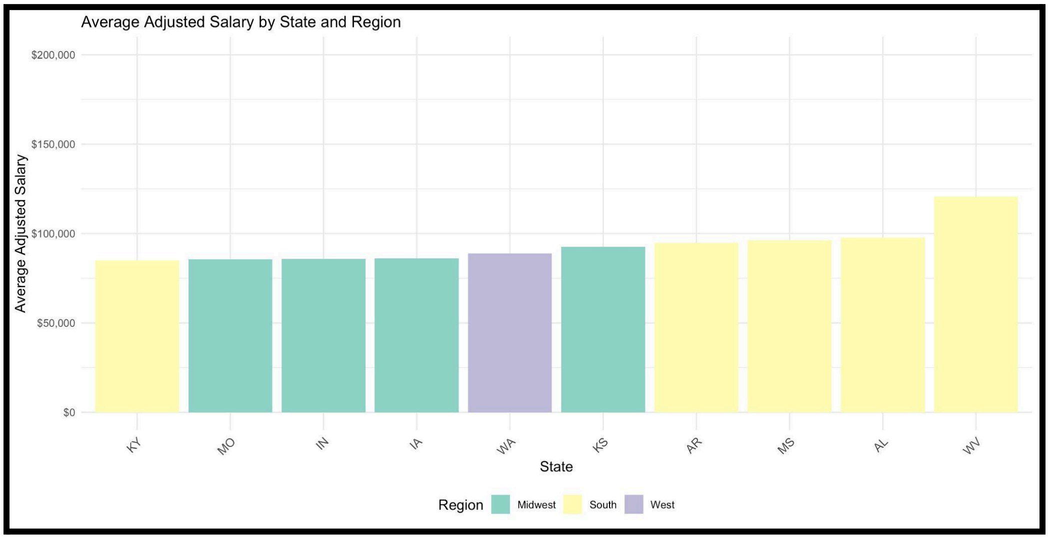
Visualization 2.3



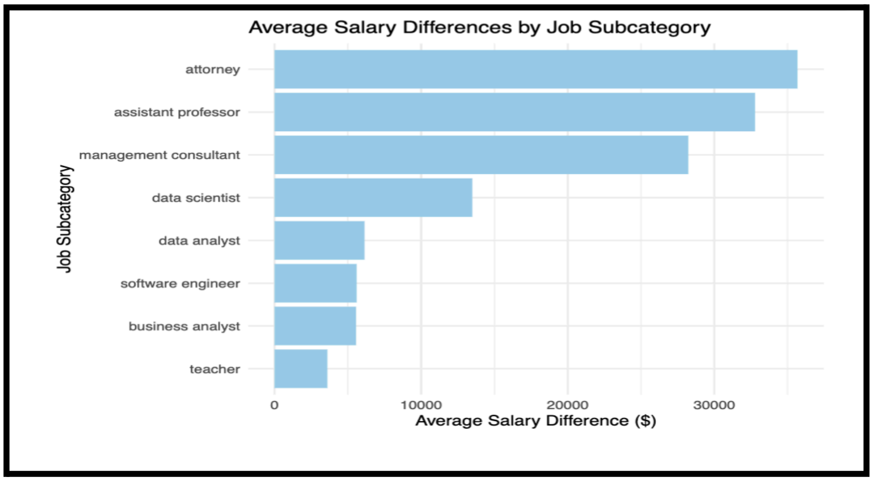
Visualization 2.4



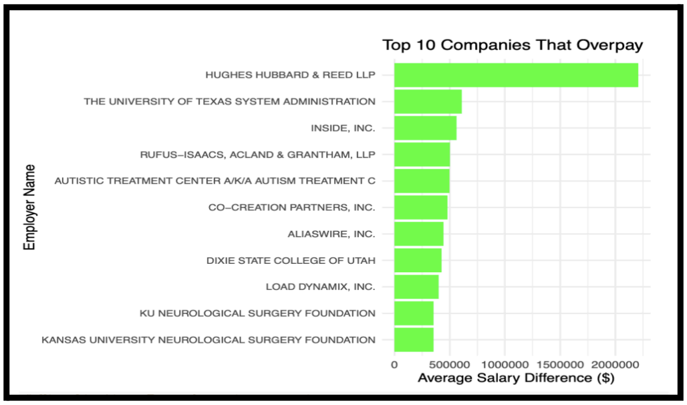
Visualization 2.5



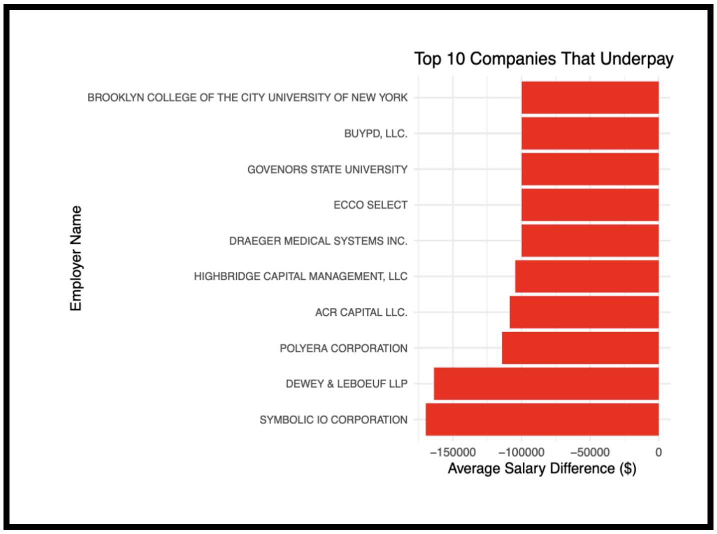
Visualization 3.1



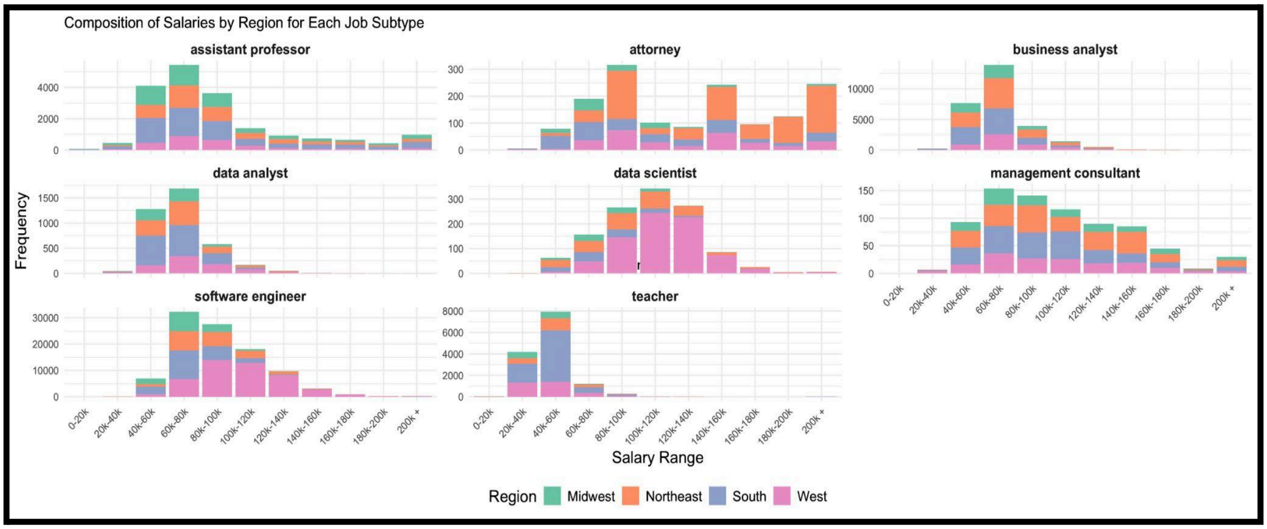
Visualization 3.2



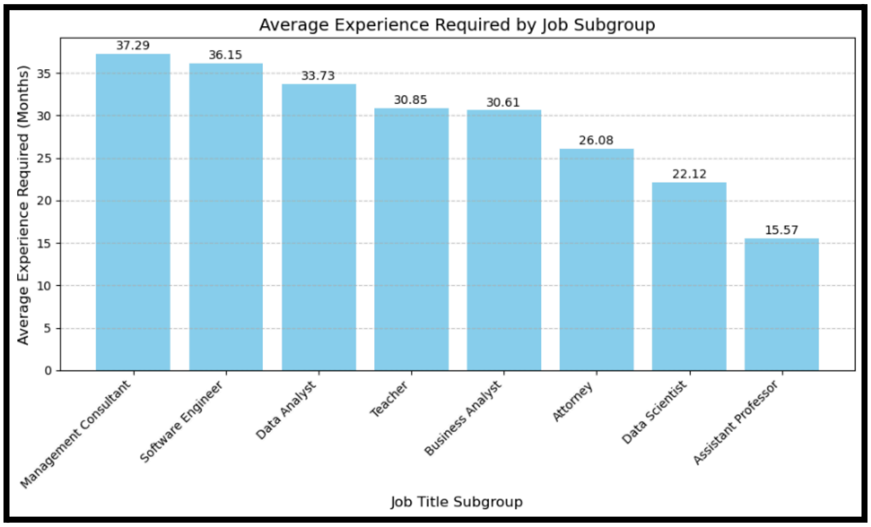
Visualization 3.3



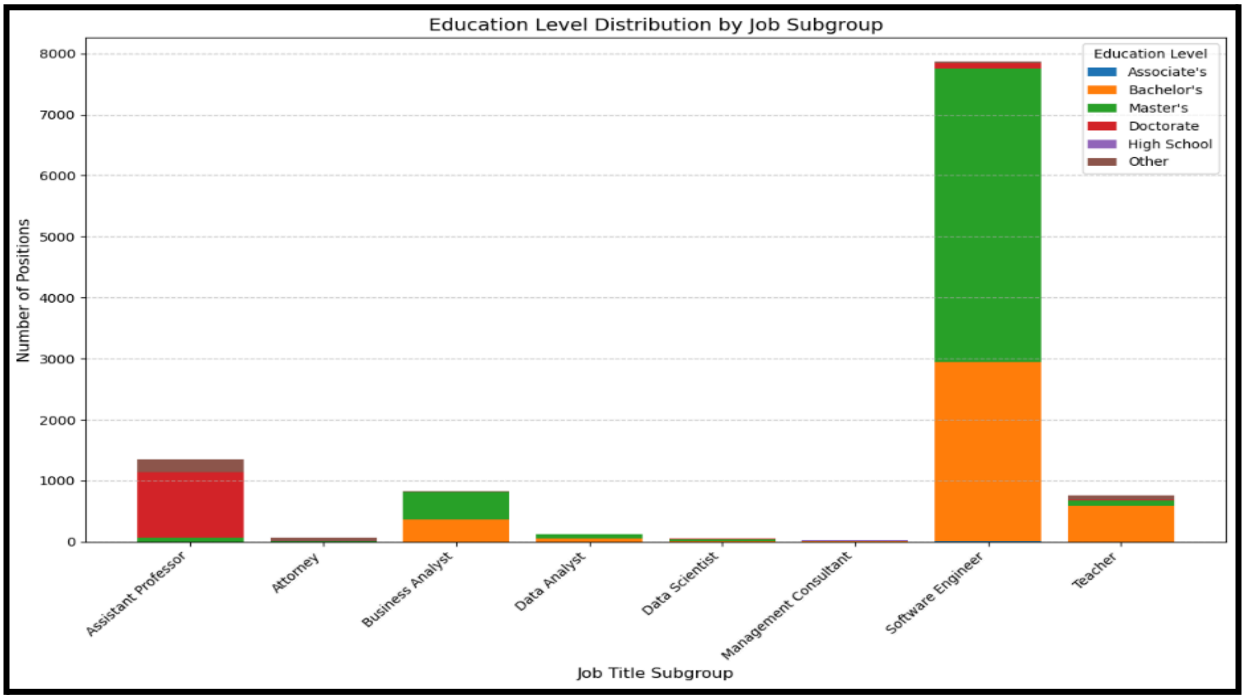
Visualization 4.1



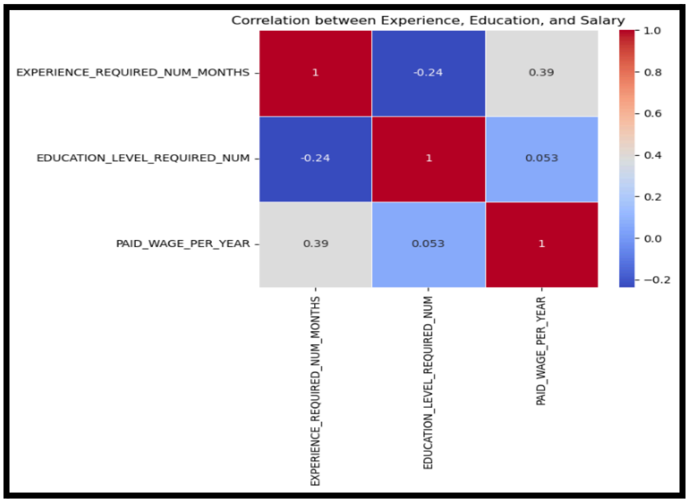
Visualization 5.1



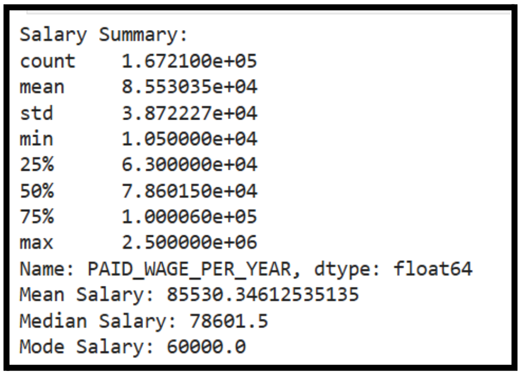
Visualization 5.2



Visualization 5.3

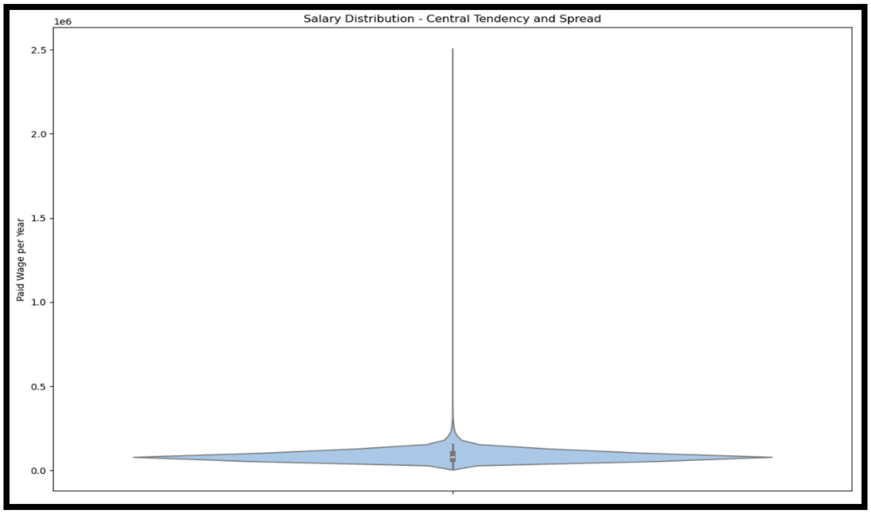


Visualization 6.1

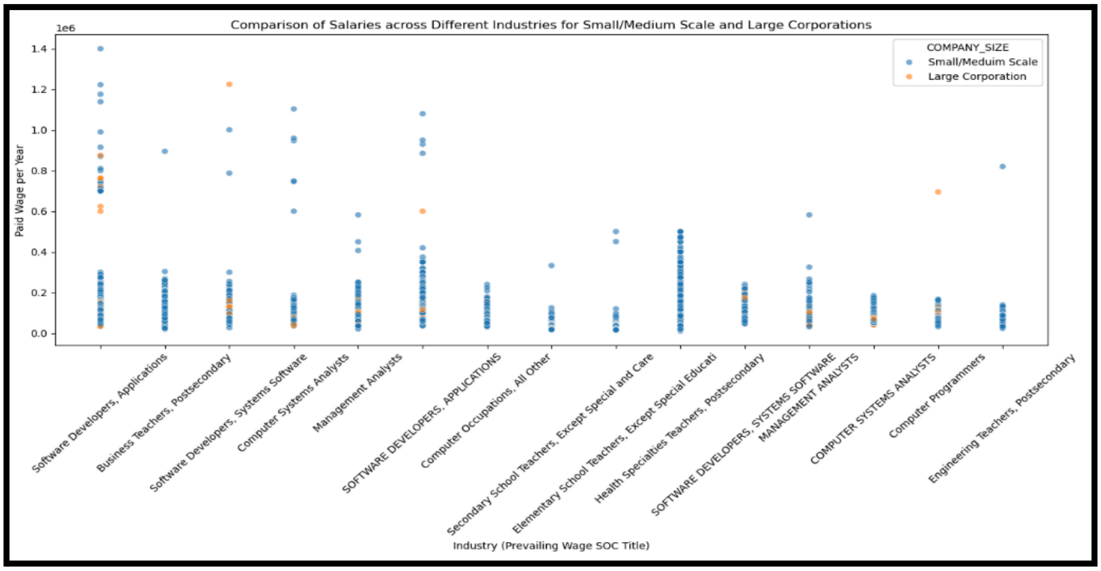


```
Salary Summary:
count    1.672100e+05
mean     8.553035e+04
std      3.872227e+04
min      1.050000e+04
25%      6.300000e+04
50%      7.860150e+04
75%      1.000060e+05
max      2.500000e+06
Name: PAID_WAGE_PER_YEAR, dtype: float64
Mean Salary: 85530.34612535135
Median Salary: 78601.5
Mode Salary: 60000.0
```

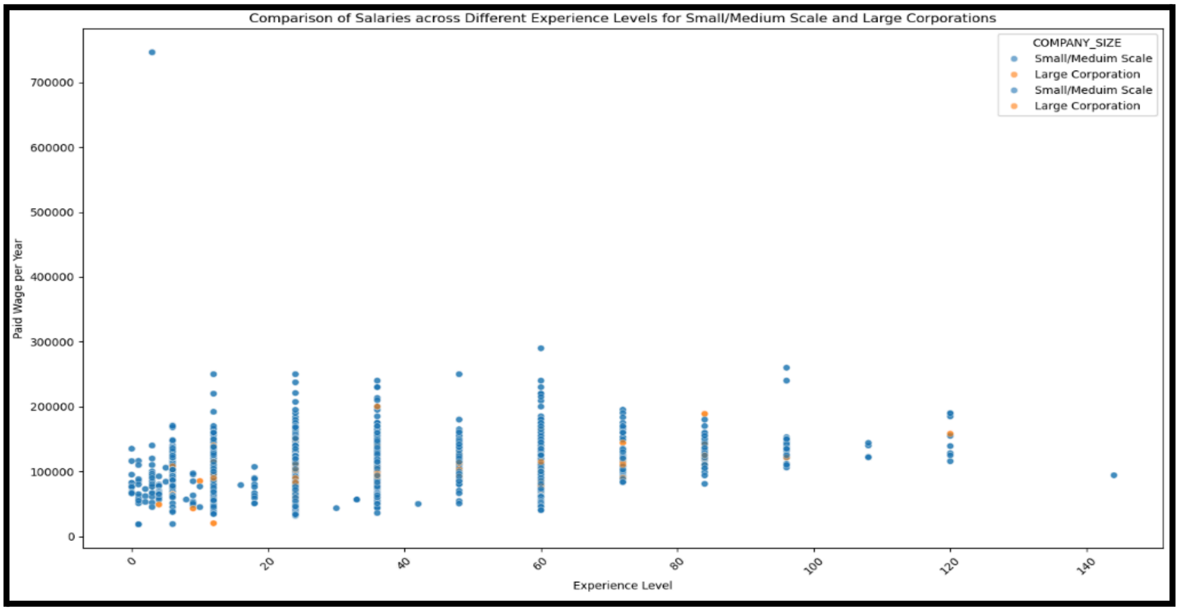
Visualization 6.2



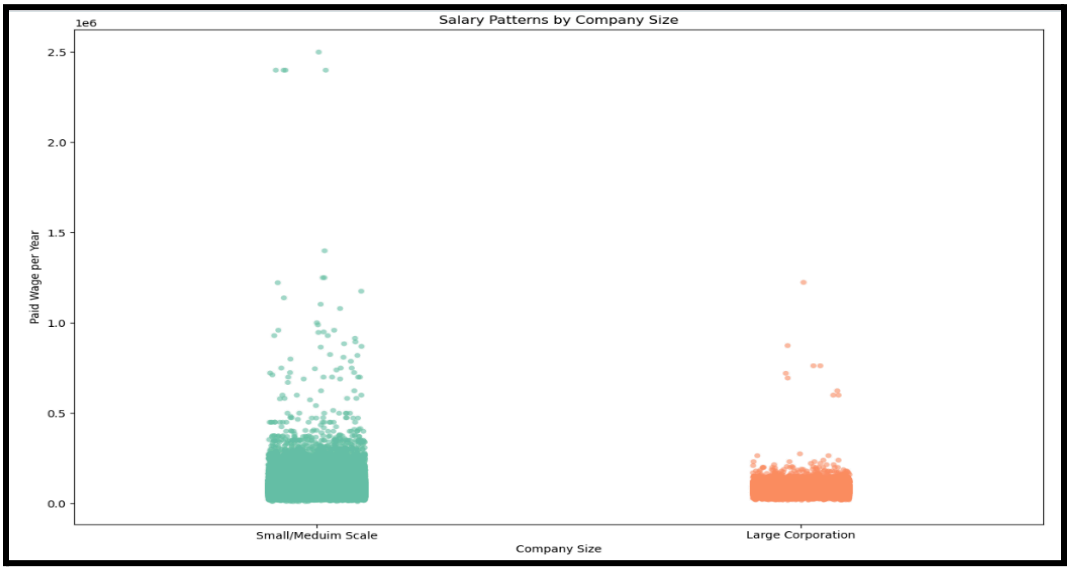
Visualization 6.3



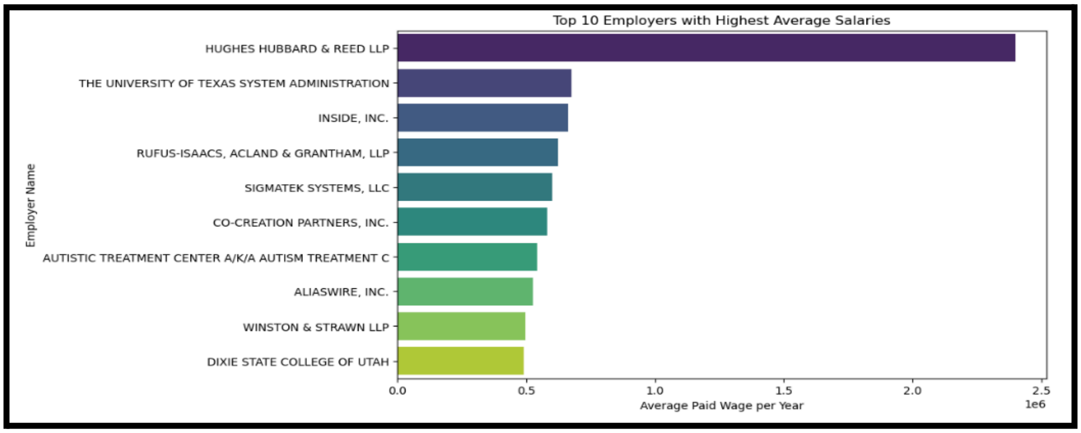
Visualization 6.4



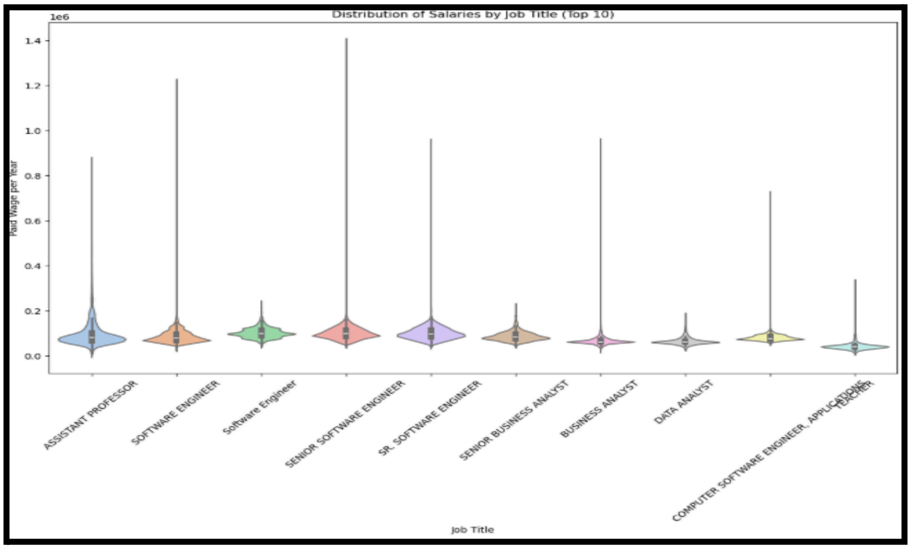
Visualization 6.5



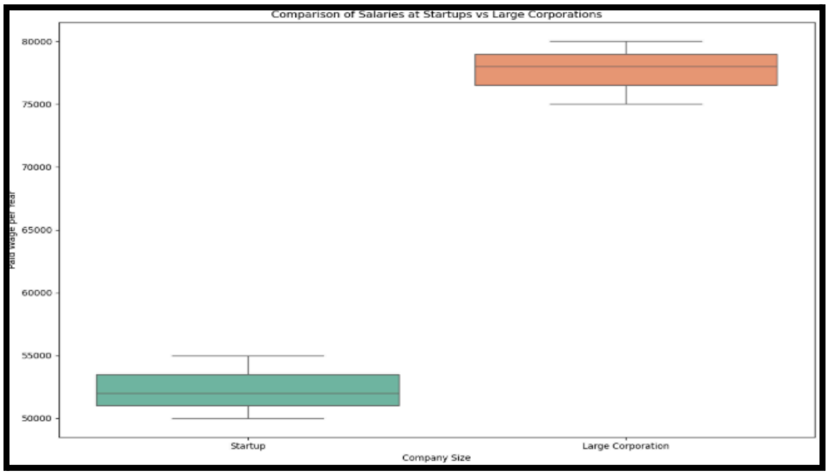
Visualization 7.1



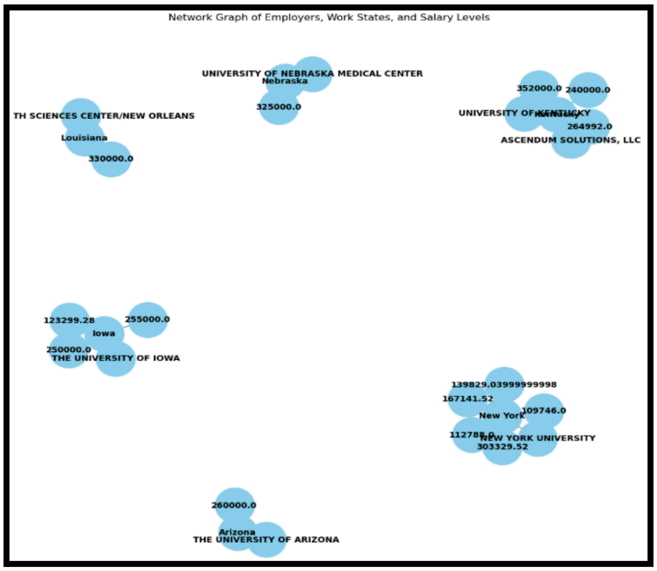
Visualization 7.2



Visualization 7.3



Visualization 7.4



References:

1 Lukkar Data. *Cost of Living Missouri Economic Research*. Kaggle, https://www.kaggle.com/datasets/lukkardata/cost-of-living-missouri-economic-research. Accessed 9 Dec. 2024.

2 U.S. Department of Labor. *Foreign Labor Certification Performance Data*. Foreign Labor Certification, http://www.foreignlaborcert.doleta.gov/performancedata.cfm. Accessed 9 Dec. 2024.