Exploration of the Common Factors Among the Highest Paying STEM Jobs and the Best Cities for STEM Careers

Crispin Corpuz
Connor Moorhous
Ken Drucker
Josh Reynoso
Alexis Walther

May 8, 2023

Abstract

In the job market, STEM fields are widely recognized as offering lucrative career opportunities, even at the entry level. However, it is important to understand which paths within STEM are most financially rewarding. This study investigates the common factors that significantly raise salaries in entry-level tech jobs. This study also examines the best cities in the United States for STEM jobs, taking into account factors such as salary, cost of living, and quality of life. By analyzing data from levels.fyi, the U.S. Bureau of Labor Statistics' "Occupational Employment and Wage Statistics," and teleport.org, we identified trends and patterns in the job market. We used various visualization techniques, including categorical distributions, time series analysis, and an interactive map dashboard, to explore the data. Our analysis revealed that managerial positions in tech jobs and individuals with graduate degrees generally came with higher salaries. Toward the second research question, we found that California, Washington, New York, and Massachusetts offer some of the highest mean and median salaries for tech jobs. Among the areas with high average salaries and good cost of living are Austin, Dallas, Raleigh, Detroit, and Phoenix.

1 Introduction

As technology continues to spread into countless parts of everyday life, it is easy to marvel at the fact that there are numerous people, teams, and organizations behind its development. While it is true that the jobs and organizations in the tech industry have become lucrative, it is also true that not all jobs are created equal in terms of salary. This raises the question about what factors contribute to the higher salaries in tech and STEM, in addition to what factors contribute to wage discrimination and quality of life for professionals. For context, some jobs in these fields could possibly pay higher salaries than others according to a multitude of factors that could include education level, years of experience, supply and demand for individuals qualified for these jobs, and many more.

The research in this project seeks to answer two questions. First, are there common factors between tech jobs that significantly raise their salary, and what are they? Finding categorical trends among the salaries of IT professionals that can provide deeper insight than the surface level stereotypes of moving upward through the corporate chain or small, startup company successes would be particularly valuable. Second, what are the best cities for STEM jobs in terms of salary, cost of living, and other quality of life factors? This question is in the same spirit as the first question in that building trends and profiles is valuable for quick information digestion. However, it differs in the sense that answering this question identifies specific

tech industry hotspots that can help young professionals discover feasible paths on which to start and build their career.

The answers and conclusions generated from this research and data analysis, as mentioned, will be a useful tool for a wide range of young individuals from high school students to college students and graduates to individuals looking to change careers into the tech industry. Interaction with this information can range from a detailed report to a simple web application in order to educate this demographic on the nuances behind jobs in the tech industry and STEM.

2 About the Data

The first data source for this report is levels.fyi[1], a website used for reporting and researching salaries. This site hosts crowdsourced salary data including information such as experience, company, location, and demographic information (gender, race, and education). This data aids the discovery of any common factors of the highest paid positions in STEM careers. Additionally, the location data on this site helps determine the best cities for STEM jobs. The data from this site can be obtained in two ways. The first is through a publicly accessible JSON object which stores over 62,000 entries from 2017 and 2021. The second way is manually scraping the site with a tool like Selenium.

The second source is the U.S. Bureau of Labor Statistics' "Occupational Employment and Wage Statistics."[2] This employment data from May 2021 contains occupational and wage information for over 1,000 various jobs. These data sets are broken down by geographic region into distinct categories such as: employment per thousand jobs, annual mean wage, and location quotient. This group of datasets will serve as a foundation for determining the relationship between geographic location, occupation, and salaries.

The final source is teleport.org[3], a site that calculates and reports quality of life metrics for cities across the globe. The site ranks cities in 17 categories including cost of living, commute, safety, and more. The data found on this site, which can be accessed using a public API, provides another perspective that benefits the evaluation of the best cities for STEM careers.

2.1 Data Scraping

The levels.fyi site organizes salary data by job title, location, and company. The most recent 1,000 entries are available for each combination of these categories. Therefore, to acquire a large amount of data, a list of job titles was compiled from the site and a list of cities was gathered from the JSON data set. However, there was a list of 20 titles and over 700 cities which, by executing only one script on one machine, would require weeks to scrape. Therefore, scraping was limited to only the list of 20 job titles.

To scrape the data from the website, a Python script[4] was written that utilizes Selenium, an automated testing tool, to navigate the site and record salary data. The script navigates to job title pages, accesses salary data in the form of strings stored in an HTML table, parses these strings into their individual components, and then adds the data into a data frame. Once a job title is fully scraped, a .csv file is generated with all of its entries. Once all titles were successfully scraped, another Python script was used to quickly compile these individual data sets into one.

While scraping ended up being successful, there were some challenges along the way. One of the first problems was addressing advertisements and hidden entries that interrupted the script. It was also quickly noticed that the levels.fyi site stopped showing salary entries after a certain number of web requests. Thus, the script had to be adjusted to detect when this happened, to wait for 5 minutes, and then to reload the page and continue scraping. Ultimately, all these challenges were overcome and over 13,000 entries were obtained.

[1] https://www.levels.fyi
[2] https://www.bls.gov/oes/current/oes_nat.htm
[3] https://teleport.org/cities/
[4] https://github.com/ConnorMoorhous/STAT4410-TeamUndergrad/blob/main/scrape/levels.fyi/scrape_titles.py

To acquire data from teleport.org, a Python script[5] was written that utilizes Teleport's developer API. The script takes a list of cities and searches for each one using the API. For each match, it grabs the city's scores and adds them to a data frame along with the city's name, first-level administrative division, and country. Once all cities have been searched, the data is exported as a .csv file.

3 Exploratory Data Analysis

3.1 Data Cleaning

3.1.1 Levels.fyi Data

The cleaning process for the scraped levels.fyi data started with dropping columns that are not relevant to the focus of this report. The kept columns were `date`, `company`, `title`, `total_annual_compensation`, `city`, `state`, `years_of_experience`, `years_at_company`, `gender`, `race`, and `education`. The next step was to convert the `date` column to the R standard date format. After that, the strings in `title` were reformatted by replacing '-' with the space character, and `total_annual_compensation` was converted to numeric after replacing '$' with the empty string.

Next, the columns `years_of_experience` and `years_at_company` needed to be converted to numeric. The strings 'years' and '+' were removed by substituting all lowercase letters and '+' with the empty string. Then, these columns were converted into numeric type. However, some entries still contained range values with hyphens (e.g., 1-5) which were automatically replaced with NA values after this conversion. Since there is no good way to convert a range to a numeric value, these rows were dropped. Finally, the `education` column was renamed to `education_level` to avoid merging issues with the teleport data set.

Similar to the scraped data, the cleaning process for the JSON levels.fyi data started with dropping irrelevant columns. Only the columns `timestamp`, `company`, `title`, `totalyearlycompensation`, `location`, `yearsofexperience`, `yearsatcompany`, `gender`, `Race`, and `Education` were kept. The next step was converting the `timestamp` column to the R standard date format and renaming the column to `date`. After that, the `location` column was separated by commas into individual `city`, `state`, and `country` columns. Since levels.fyi locations only excluded a country if they were in the United States, the NA values in `country` were replaced by 'United States' and then rows without 'United States' in country were removed. With only U.S. cities left in the data frame, the `country` column no longer served a purpose and was dropped.

The JSON data contained job titles outside the focus of this report. These titles ('Human Resources', 'Marketing', 'Sales', and 'Recruiter') were removed. An odd string, 'Title: Senior Software Engineer', was detected in the `gender` column and then replaced with 'Unknown'. Finally, to prepare this data set to be combined with the scraped data, some columns were renamed and the column named `type` with a default value of "Unknown" was added. The columns `totalyearlycompensation`, `yearsofexperience`, `yearsatcompany`, `Race`, and `Education` were renamed to `total_annual_compensation`, `years_of_experience`, `years_at_company`, `race`, and `education_level`, respectively.

With both the scraped and JSON levels.fyi data relatively clean, the next step was combining these data frames which was completed by appending the scraped data to the JSON data. After this, some columns still needed cleaning. The `company` column needed capitalization to be standardized. Therefore, all `company` strings were converted to title case. Additionally, entries with 'Software Engineer', a placeholder value on levels.fyi for entries with no company name listed, in the `company` column were replaced with 'Unknown'. The demographic columns `gender`, `race`, and `education` also needed cleaning. For each of these, the empty string, `NA`, and 'hidden' were replaced with 'Unknown' and strings meaning the same thing (e.g., 'Doctorate (PhD)' and 'PhD') were recoded to be the same. Finally, duplicate rows were removed and the categorical (`chr`) columns were converted to factors.

[5] https://github.com/ConnorMoorhous/STAT4410-TeamUndergrad/blob/main/scrape/teleport/teleport.py

3.1.2 BLS Occupational Employment and Wage Statistics Data

For the data from the U.S. Bureau of Labor Statistics (BLS) concerning "Occupational Employment and Wage Statistics," the cleaning was relatively straightforward for the three data sets—city, state, and national statistics—obtained for the year 2021. Extraneous columns for the purpose of answering the second research question and keeping data similar to other datasets involved the hourly wages per occupation, columns that were purely categorizers for the BLS's records, and columns that only held one value for every row entry (like "Cross-industry"). These included the columns named `NAICS`, `NAICS_TITLE`, `I_GROUP`, `PCT_TOTAL`, `PCT_RPT`, `H_MEAN`, `H_PCT10`, `H_PCT25`, `H_MEDIAN`, `H_PCT75`, `H_PCT90`, `ANNUAL`, `HOURLY`. Additionally, since the national statistics dataset did not differ in terms of `JOBS_1000` and `LOC_QUOTIENT`, these were removed from that set specifically.

When cleaning the rows for missing values, the first step was to convert all wage/monetary columns to numeric type (since importing them left them as characters). This way, any rows that are non-numeric would be forced as `NA` (missing value). There is no reason to impute these values with a mean or median because there is no analytical gain. As such, these rows were dropped from the dataset in addition to any rows that did not contain the keywords 'Software', 'Engineer', 'Product', 'Engineering', 'Architect', 'Solution', 'Data', and 'Mathematical'. With that, three clean data sets were ready to be explored.

3.1.3 Teleport.org Data

The raw teleport data started in long form with columns: `city`, `admin1`, `country`, `color`, `name`, and `score_out_of_10`. Upon loading the .csv file, duplicates were removed. Next, since Teleport's API reported NA scores as 0, 0 values were replaced with `NA` values which prevents these scores from being misinterpreted later on. Since this report's focus is on cities in the United States, entries without U.S. cities (i.e., rows without the United States in the `country` column) were removed. The next step was to drop the `country`(it no longer served any purpose) and `color` columns. After that, the `admin1` column was renamed to `state`, the `name` column was renamed to `category`, and the `score_out_of_10` column was renamed to `score` to better represent the columns. Then, the `state` values were converted to their abbreviations to better match the cleaned levels.fyi data.

To support mapping these cities, a subset of the US Cities Database from simplemaps.com – with selected columns: `city`, `state_id`, `lat`, `lng`, and `population` – was merged, based on city and state, with the teleport data. After this, the categorical columns `city` and `state` were converted into factors. Finally, the data set was cast into wide form with the `category` column as the variable names and the `score` column as the values.

3.2 Data Summaries

3.2.1 Levels.fyi Data

The cleaned levels.fyi dataset contains 12 columns with a total of 62,531 rows. As shown in Figure 1, the column `date` contains values from January 2018 to April 2023. The gap around early 2022 shows the divide between the scraped and JSON datasets.

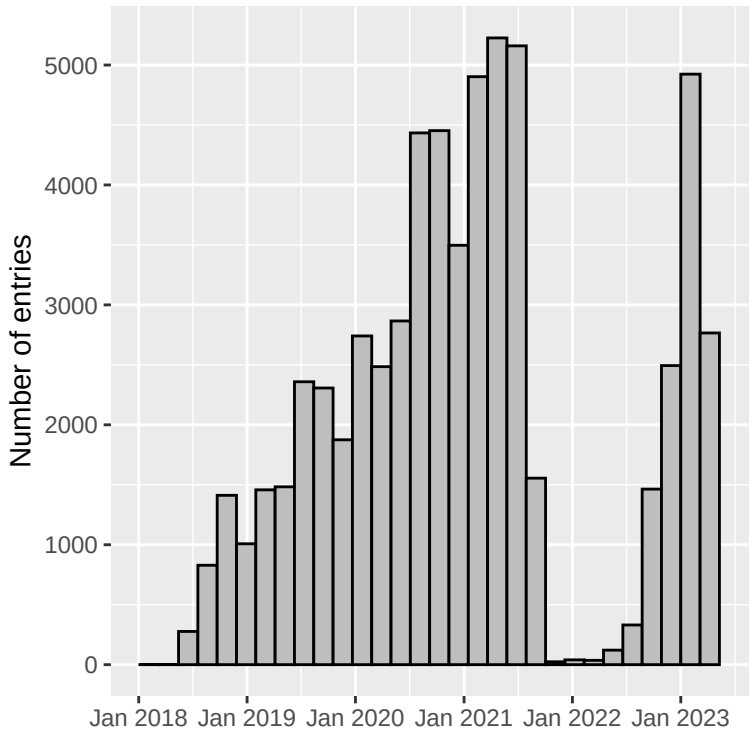

Figure 1: Distribution of levels.fyi entries by date.

The numerical columns in this dataset – `total_annual_compensation`, `years_of_experience`, `years_at_company` – are shown and summarized in Table 1.

Table 1: Numerical summary of levels.fyi data set.

	total_annual_compensation	years_of_experience	years_at_company
Min.	11000.00	0.00	0.00
Median	198000.00	6.00	2.00
Mean	229272.72	7.48	2.84
Max.	4980000.00	71.00	69.00
SD	137790.80	6.07	3.40

The other columns – `company`, `title`, `city`, `state`, `gender`, `race`, `education_level`, and `type` – are categorical factors. It is important to note that the distribution of entries in this dataset. The mast majority of jobs reported are software engineering positions, as shown in Figure 2. Additionally, most entries have unreported demographic data with over half missing race and education data. A large chunk (over 20,000) of entries are also missing gender data. However, male is the most reported gender.The distributions of these demographic variables and the rest of the categorical variables are represented as Figures 8-14 in Appendix A.

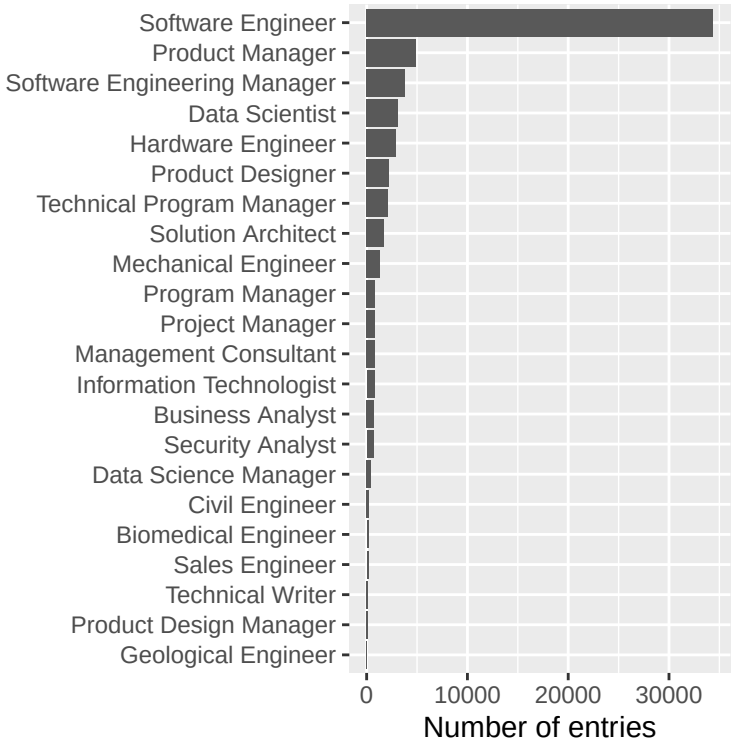

Figure 2: Distribution of levels.fyi entries by job title.

3.2.2 BLS Occupational Employment and Wage Statistics Data

The cleaned city BLS dataset contains 14 columns with a total of 7800 rows. In the `AREA_TITLE` column, there were 396 unique cities indicated with the most frequent one being the New York-Newark-Jersey City tri-city area at 258 counts. All states and major territories of the United States were represented in this dataset, since there were 52 unique states identified—the most frequent one was California with 5 counts. The last of the categorical factors involved occupation title, where there were 50 unique titles. The most frequent was the overarching category of Architecture and Engineering Operations at a count of 7. The remaining columns are numeric descriptors of salary data and are summarized below in Tables 2 and 3.

Table 2: City Numerical Summary 1

	TOT_EMP	EMP_PRSE	JOBS_1000	LOC_QUOTIENT	A_MEAN	MEAN_PRSE
Min.	30.00	0.00	0.01	0.05	17780.00	0.10
Median	160.00	10.40	0.84	0.86	84190.00	2.80
Mean	1289.00	12.45	3.08	1.26	82585.75	3.65
Max.	294770.00	49.40	123.39	100.55	170200.00	29.30
SD	7644.87	8.69	6.73	2.54	24868.03	2.96

Table 3: City Numerical Summary 2

	A_PCT10	A_PCT25	A_MEDIAN	A_PCT75	A_PCT90
Min.	15080.00	16510.00	16940.00	18260.00	18500.00
Median	48505.00	61470.00	78230.00	99340.00	123830.00

	A_PCT10	A_PCT25	A_MEDIAN	A_PCT75	A_PCT90
Mean	51170.12	62794.90	78778.43	98002.68	117762.89
Max.	130620.00	168370.00	197920.00	207960.00	207980.00
SD	16364.20	19836.67	24797.01	30735.28	36979.31

The cleaned state BLS dataset contains 14 columns with a total of 2059 rows. In the `AREA_TITLE` and `PRIM_STATE` columns, there were 54 unique states and territories identified. Among all states and major territories of the United States, the most frequent was Florida at 10 counts. The last of the categorical factors involved occupation title, where there were 51 unique titles. The most frequent was the overarching category of Architecture and Engineering Operations at a count of 8. The remaining columns are numeric descriptors of salary data and are summarized in Tables 9 and 10 in Appendix B.

The cleaned national BLS dataset contains 10 columns with a total of 78 rows. With no separate cities and states at the national level, there are no columns for such in the cleaned dataset. In terms of occupation title, there were 63 unique titles with the most frequent being Aerospace Engineers at a count of 2. The remaining columns are numeric descriptors of salary data and are summarized in Tables 11 and 12 in Appendix B.

3.2.3 Teleport.org Data

The cleaned teleport.org dataset contains 22 columns and 576 rows. The first four columns – `city`, `state`, `lat`, and `lng` – contain identifying city information (name and coordinates) for 576 U.S. cities. The next column named `population` stores each city's estimated population. The last 17 columns, shown with summary information in Table 4, represent categories that teleport.org uses to score and evaluate cities. Scores for these categories can range from 0 to 10.

Table 4: Summary information for teleport.org data.

	Min.	Median	Mean	Max.	SD
Business Freedom	5.50	8.67	8.63	8.67	0.32
Commute	0.93	4.52	4.23	6.14	1.10
Cost of Living	2.34	4.92	4.67	6.68	1.21
Economy	6.51	6.51	6.51	6.51	0.00
Education	3.62	5.82	6.08	8.62	1.77
Environmental Quality	4.05	6.69	6.53	9.95	1.22
Healthcare	7.52	8.60	8.62	9.02	0.20
Housing	0.50	4.21	3.93	7.26	2.21
Internet Access	3.44	6.17	6.21	9.05	1.18
Leisure & Culture	1.91	7.64	7.76	10.00	1.50
Outdoors	1.00	5.39	5.37	7.93	1.31
Safety	1.34	4.79	4.76	8.06	1.42
Startups	2.79	7.94	7.67	10.00	2.00
Taxation	3.78	4.06	4.25	4.77	0.34
Tolerance	3.68	6.55	6.40	9.67	1.38
Travel Connectivity	1.10	3.65	3.63	6.68	1.42
Venture Capital	1.00	6.12	6.55	10.00	2.67

3.3 Exploring the Data

3.3.1 Exploring Levels.fyi Data

To explore the first research question – that is, to find the common factors that contribute to higher paid tech salaries – different methods can be applied to the levels.fyi dataset based on classification of its variables. To

determine if there is a relationship between the numerical variables, a correlation plot can be used. Figure 3 is a correlation plot for the `total_annual_compensation`, `years_of_experience`, `years_at_company`, `population` (from teleport.org dataset), and `Cost of Living` variables. This plot reveals no strong correlation between any of the variables. The highest correlated are annual compensation and years of experience with r = 0.43 which suggests a weakly correlated relationship.

It might seem odd that `Cost of Living` is negatively correlated with `total_annual_compensation`. However, `Cost of Living` represents a score from 0 to 10, with 10 representing the better (more positive) score. That is, a lower `Cost of Living` score represents a higher cost of living. Thus, this relationship is in line with general expectations.

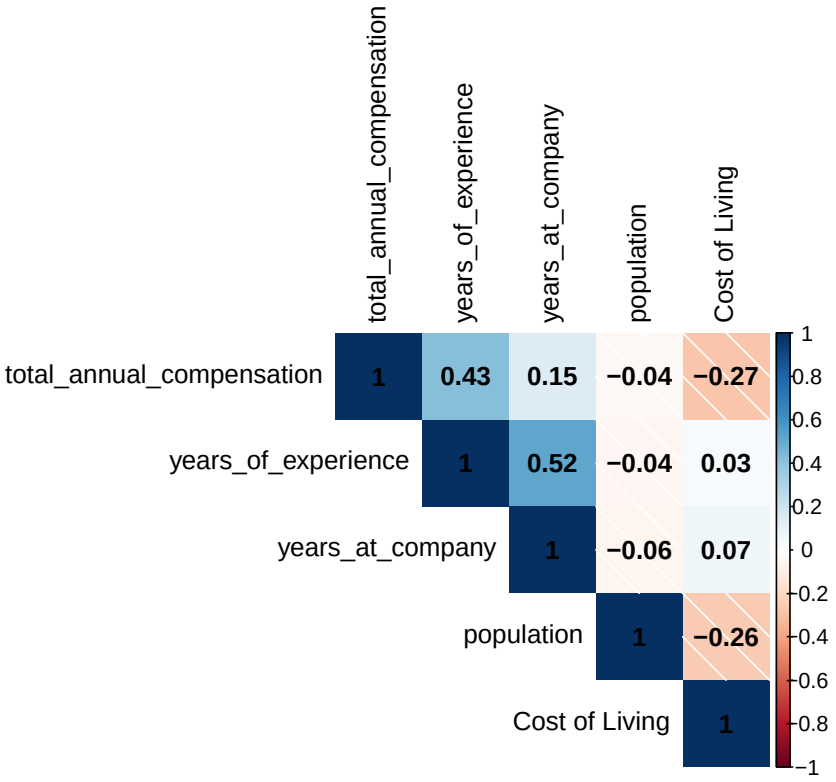

Figure 3: Correlation plot of numerical variables.

For the categorical variables, we can explore each one's relationship with median and average compensation. Figure 4 shows the average and median compensation by job title. Notably, this figure reveals that the top five most paid job titles are all manager positions. Additionally, hardware and software engineers are the next highest paid positions behind them.

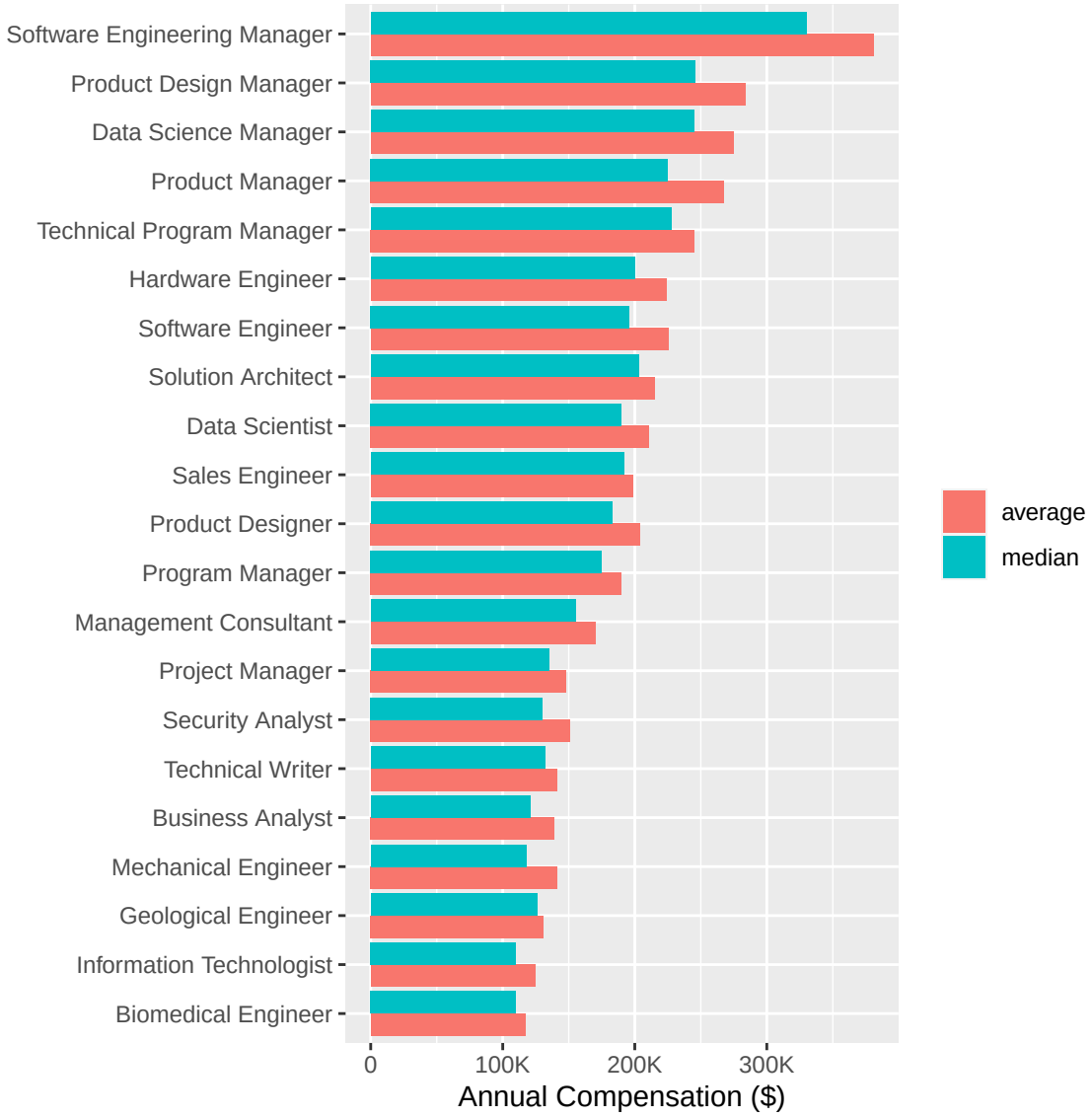

Figure 4: Annual compensation by job title.

Similarly, Figure 5 shows the top ten cities with at least 25 reported jobs (for better accuracy) and Figure 15 in Appendix C shows the top ten states with the highest average and median compensation. These figures provide some insight into the best locations for high paying jobs. California stands out the most in both of these figures as a location for high compensation. Notably cities around the San Francisco Bay Area dominate the top ten list.

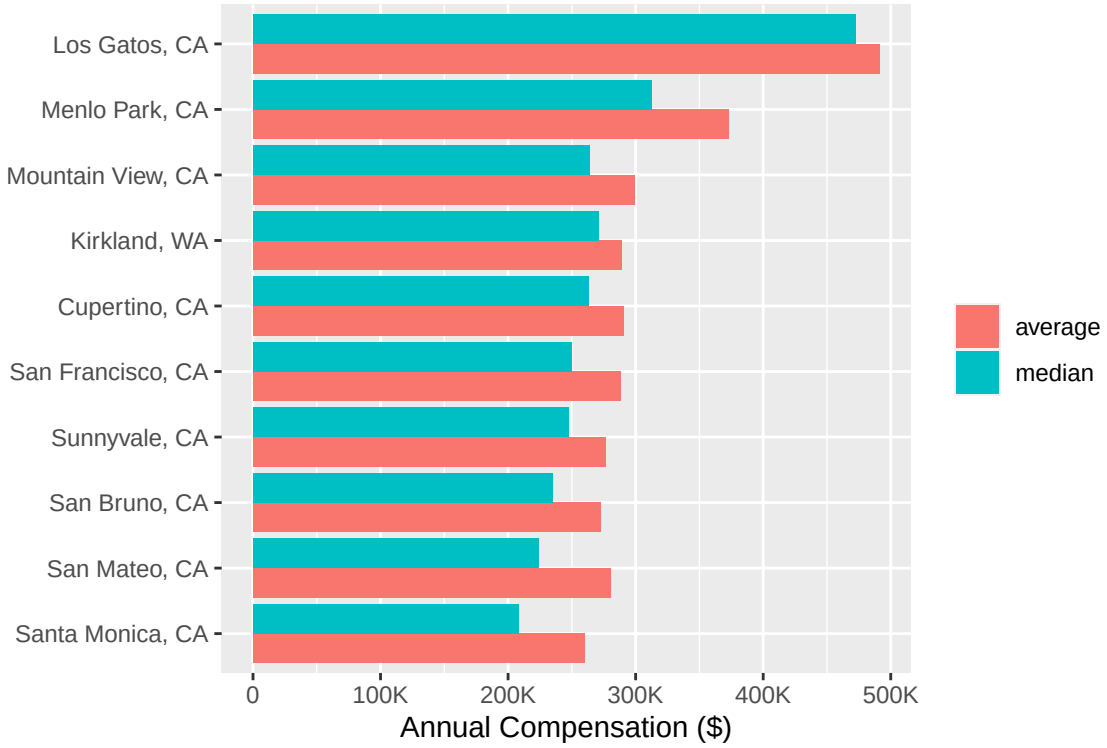

Figure 5: Cities with highest annual compensation (minimum 25 reported jobs).

Figure 6 shows the top average compensation by education level. The highest paid education levels, PhD and master's degree, are largely expected results. However, workers who completed high school or some college are performing better than those with associate's or bachelor's degrees, an unexpected result. This might be due to the low representation of workers with only high school or some college experience. Not including unknown entries, Figure 14 in Appendix A reveals that most individuals with a STEM job have at least a bachelor's degree.

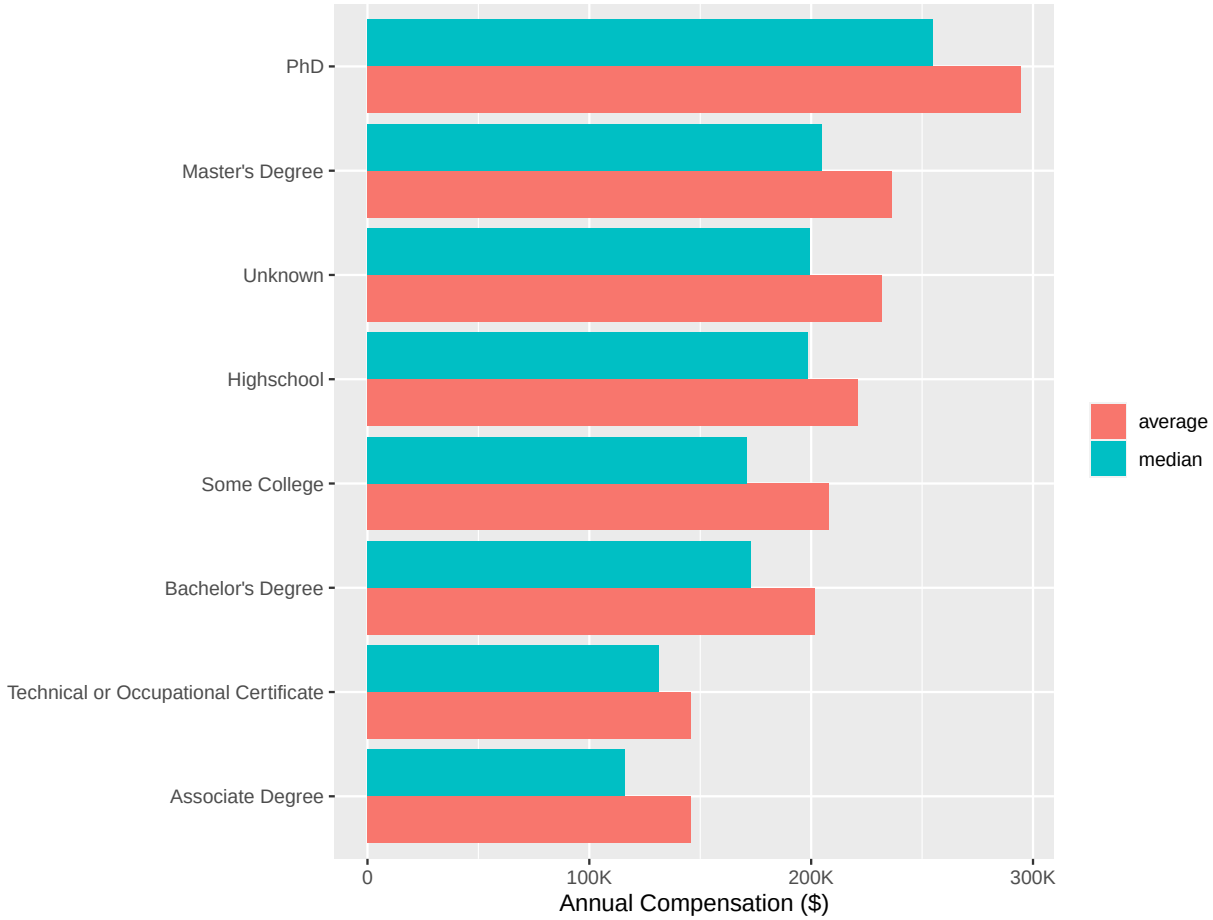

Figure 6: Annual compensation by education level.

With recent rise of remote work, an interesting area of study is the composition of jobs that are in person, remote, or hybrid. While the reported entries for this metric is quite low, Figure 16 in Appendix C shows that among the data reported, hybrid and in-person office jobs result in a slightly higher salary than remote ones.

While the levels.fyi site stores a vast amount of salary data over the past five years, this report's research is limited by the availability of the data. Time-series analysis is of particular interest, however, this limitation in the data makes it hard to reach any solid conclusions. Figure 7 is a time series plot showing the change in average and median compensation across the whole data set. The data appears to reveal a slight decrease in compensation from 2018 to 2021. However, the sudden decrease is due to the missing data and further analysis is unreliable. For another time series plot showing the change in compensation for the top 5 most reported job titles, see Figure 17 in Appendix C.

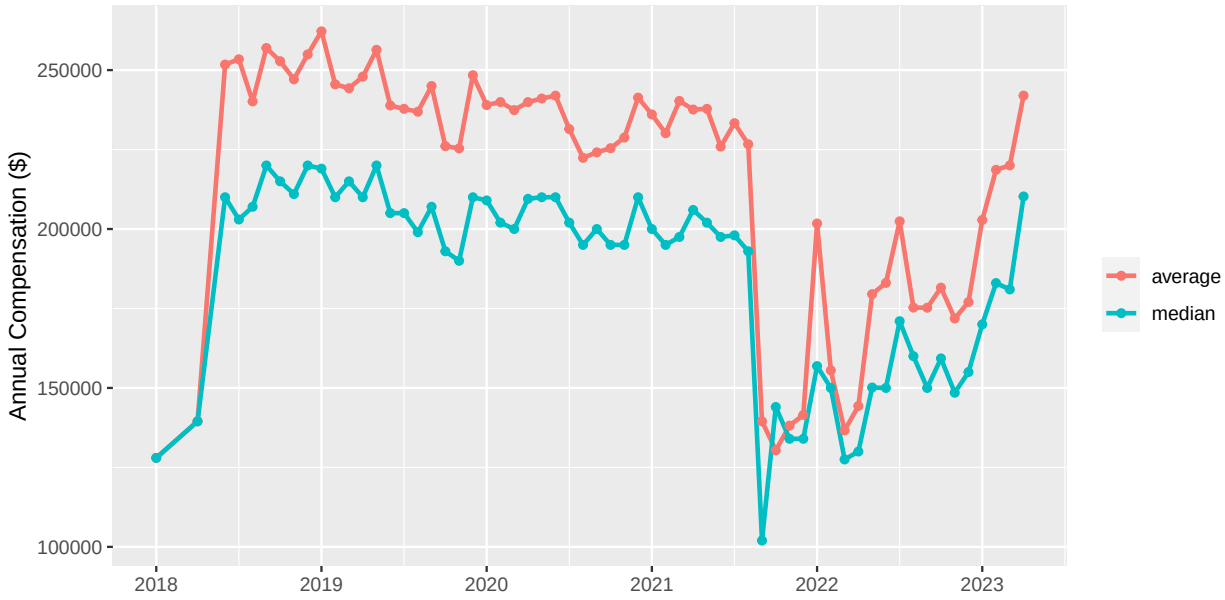

Figure 7: Annual compensation from 2018 to 2023.

3.3.2 Exploring Teleport.org Data

To explore the second question and find the best cities not only in terms of high salary but also other city metrics such as cost of living, a map can be used. Figure 8, a screenshot of an interactive Tableau map made with both the levels.fyi and Teleport data sets, shows cities based on both their average annual compensation and their cost of living. Cities with a higher score (a darker blue) have a better cost of living than cities with a lower score (and a darker orange). Additionally, cities with a larger circle have a higher average compensation. This map reveals that, while the San Francisco and New York areas have some of the highest salaries, they also have poor cost of living. Some cities in Texas (Austin and Dallas), North Carolina (Raleigh area), Michigan (Detroit area), and Arizona (Phoenix) might be a better choice for some trying to maximize both metrics.

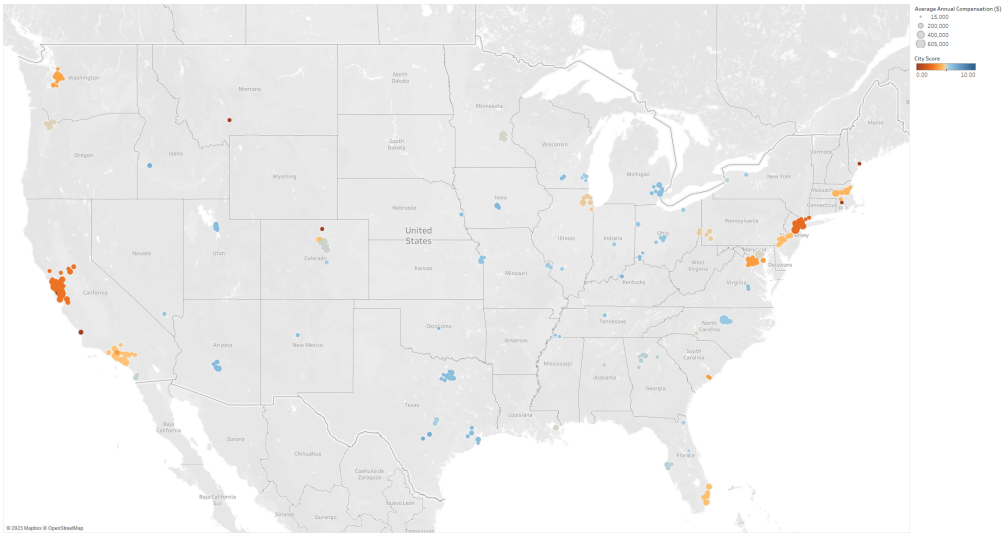

Figure 8: Map of cities by average annual compensation and cost of living.

3.3.3 Exploring BLS Data

Before continuing, it is important to preface with the fact of the BLS datasets being somewhat unreasonable to present visually by conventional means (distributions, boxplots, etc.). However, the simple visual format of a table containing the top ten cities and states to live in ordered by the top mean salaries accompanied by occupation title helps to answer the question of what cities are best to live in by salary with the nuance of occupation.

Table 5: Mean Salary Top-Ten by City

AREA_TITLE	OCCUPATION	MEAN
Norwich-New London-Westerly, CT-RI	Architectural and Engineering Managers	170200
Beckley, WV	Architectural and Engineering Managers	166150
Omaha-Council Bluffs, NE-IA	Aerospace Engineers	166060
Palm Bay-Melbourne-Titusville, FL	Architectural and Engineering Managers	161450
Orlando-Kissimmee-Sanford, FL	Architectural and Engineering Managers	160410
Providence-Warwick, RI-MA	Architectural and Engineering Managers	159580
Shreveport-Bossier City, LA	Architectural and Engineering Managers	159380
Syracuse, NY	Architectural and Engineering Managers	158790
Binghamton, NY	Architectural and Engineering Managers	158620
Savannah, GA	Architectural and Engineering Managers	158350

As such, Table 5 above indicates which cities are best to live in order of the mean salary. It is indicated that with the exception of the Omaha-Council Bluffs area, Architectural and Engineering managers (which is an overarching group of numerous professions in the categorization of the BLS) are the ones who will likely experience this in the listed areas. More exploration and summary needs to be done on an occupational basis to determine the distribution in the cities given in the BLS dataset. A map-like visualization would be ideal for exploring this data.

Table 6: Median Salary Top-Ten by City

AREA_TITLE	OCCUPATION	MEDIAN
Beckley, WV	Architectural and Engineering Managers	197920
Omaha-Council Bluffs, NE-IA	Aerospace Engineers	168380
Toledo, OH	Sales Engineers	166940
Seattle-Tacoma-Bellevue, WA	Database Architects	166480
Flagstaff, AZ	Bioengineers and Biomedical Engineers	165540
Rapid City, SD	Environmental Engineers	164410
Orlando-Kissimmee-Sanford, FL	Architectural and Engineering Managers	163650
Oklahoma City, OK	Mining and Geological Engineers, Including Mining Safety Engineers	162720
Bridgeport-Stamford-Norwalk, CT	Database Architects	162190
Norwich-New London-Westerly, CT-RI	Architectural and Engineering Managers	162150

When doing the same analysis as before on the median annual salary, however, more STEM fields are indicated as shown in Table 6. With how close the mean and median salaries are to each other, a more in-depth visualization of distributions for the most common occupations and least common occupations would be interesting.

Table 7: Highest Employment Areas by State

AREA_TITLE	OCCUPATION	TOT_EMP	LQ	MEAN
California	Computer and Mathematical Occupations	662110	1.21	122200
Texas	Computer and Mathematical Occupations	411390	1.02	93880
California	Architecture and Engineering Occupations	316980	1.11	108540
New York	Computer and Mathematical Occupations	263440	0.92	108470
Florida	Computer and Mathematical Occupations	229040	0.81	85540
Texas	Architecture and Engineering Occupations	224440	1.06	94030
Virginia	Computer and Mathematical Occupations	219330	1.77	110510
Washington	Computer and Mathematical Occupations	190160	1.80	124000
Illinois	Computer and Mathematical Occupations	175110	0.94	93580
Pennsylvania	Computer and Mathematical Occupations	163020	0.88	87530

Shown in Table 7, the high number of employment in California, Texas, New York, Florida, Virginia, Washington (state), Illinois, and Pennsylvania in the areas of computer and mathematical operations—with respect to data science, analysis, and management and also software engineering and management—indicate they are states where such skilled individuals are in high demand. With the location quotient also all being close to or above 1, these states (collectively) likely make up the majority of employment locations in these fields. Moreover, it is interesting to see that mean salary generally increases among this list of top ten employment areas as the location quotient also increases.

Table 8: Highest Annual Salaries Among Lowest Employment Totals

AREA_TITLE	OCCUPATION	TOT_EMP	LQ	MEAN
Rhode Island	Architectural and Engineering Managers	260	0.43	163460
District of Columbia	Architectural and Engineering Managers	1290	1.47	158100
Pennsylvania	Architectural and Engineering Managers	7300	0.99	150490
Washington	Database Architects	1860	1.62	150430
Nebraska	Aerospace Engineers	140	0.38	149860
Maryland	Architectural and Engineering Managers	4910	1.45	149780
Idaho	Architectural and Engineering Managers	990	0.98	148160
South Dakota	Architectural and Engineering Managers	190	0.35	147930
California	Nuclear Engineers	340	0.23	147060
Alabama	Architectural and Engineering Managers	2670	1.04	146600

This last exploration was interesting due to the widely varied location quotient. Because there is little trend between occupation title and location quotient (and the mean and median salaries were in the middle range around $150,000), understanding the discrepancy as to why some areas are more saturated than others could lead to a better understanding of why some states flourish in STEM and data science in comparison to others. For example, Architectural and Engineering Managers category varied widely in location quotient while staying in a relatively similar mean annual salary range.

3.3.4 Salary Explorer

The diverse nature of the levels.fyi salary data set makes it difficult to examine all angles of the data. The several categorical variables – including title, gender, education, etc. – combined with the geospatial and temporal data create a rich data set primed for exploration. To facilitate the exploration of this dataset, we created Salary Explorer, a web application for investigating the best cities for science, technology, engineering, and mathematics careers.

Salary Explorer[6] is an interactive application that presents an overview of different STEM job salaries for different locations. The app allows users to select a time range, one or more titles, and any combination of education levels, races, or genders to filter the levels.fyi data and present a more specialized overview.

The app's first page provides a national overview of the data. The page displays national summary data – median salary, average salary, and the number of jobs reported – for the user's selections. The page also displays an interactive time-series plot showing how this summary data has changed over time.

The app also hosts two other pages for the user to explore the data by state and city. The state page features an interactive table and map. The table displays median compensation, average compensation, and reported jobs for each state and can be filtered and sorted by each of these metrics. The map shows a coloring of the states based on median compensation and has a hover tooltip that includes the same information as the table. The user can select a state from the table or the map to focus on which displays the time series plot for the data.

The city page is similar to the state page, including both an interactive table and an interactive map displaying the same information. The cities can also be selected to display the time-series plot and a bar chart displaying the selected city's Teleport.org scores for different quality of life categories.

While Salary Explorer provides a great overview of salaries across the country, the application still has several areas for improvement. The maps could benefit from the ability to switch the type of information displayed and the Teleport.org bar chart could use more context explaining what the categories are and how their scores were calculated. Salary explorer would also benefit from a comparison feature that allows users to easily compare salaries for different titles and demographic information. Lastly, the biggest area in need of improvement is the data itself. The entire levels.fyi data set was unavailable for this project due to a paywall and limitations of scraping. Having access to this entire data set would likely make product more accurate and would provide essential context for time-series analysis.

4 Findings

4.1 Common Factors Behind High Paying Jobs

Regarding the question "Are there common factors between tech jobs that significantly raise their salary, and what are they?", there are indeed common factors that raise tech job salary. These include having a managerial role and an advanced degree. Additionally, there is a positive relationship between years of experience and total annual compensation. Specific jobs that commonly have high salaries include Architectural and Engineering Managers in addition to Aerospace Engineers based on mean and median salary by city.

4.2 Best Cities

Regarding the question "What are the best cities for STEM jobs in terms of salary and cost of living?", a few cities were identified in the San Francisco Bay Area, Seattle Area, and New York Area. In the San Francisco Bay Area, Los Gatos, Mountain View, Cupertino, San Francisco, and Sunnyvale were indicated as best by solely salary. Similarly, Kirkland and Seattle were identified for the same in Washington, while New York City was identified for the same in New York. Finally, when considering both salary and cost of living, Austin, Dallas, Raleigh, Detroit, and Phoenix were identified as cities with tech jobs that satisfy both categories sufficiently.

4.3 Interesting Findings

Most of the findings of this report are largely expected. However, there was one interesting finding in the data. As shown in Figure 6, those with only high school diplomas or some college had a higher median and

[6] https://connormoorhous.shinyapps.io/salaryexplorer/

average compensation than those with bachelor's or associate degrees. This might be due to a relatively small number of entries for high school, some college, and associate degrees, which was found in Figure ?. Also, it is important to note that the distribution of jobs among education levels suggests that a bachelor's degree is needed to get a STEM job in the first place. There are way more job entries for people with a bachelor's degree than those with only high school diplomas, some college, or an associate degree.

4.4 Challengs & Limitations

Some limitations of our analysis is the fact that there are skewed types of data collected from levels.fyi. A large amount of the data collected was from software engineers or software engineering managers. This is unfortunate because the tech job field is extremely diverse, and having the data saturated by one specific field of STEM is detrimental to coming to better conclusions. Secondly, a majority of the data from levels.fyi is embodied by the top five software engineering companies. (FAANG is an acronym that embodies the most popular large tech companies: Facebook, Amazon, Apple, Google/Alphabet, and Microsoft). Again, this limits the scope of conclusions and recommendations.

Moreover, there were a large amount of unknown values for work type, education level, and race that limited conclusions further. More data was missing from late 2021 to early 2023 due to not having access to the full dataset behind level.fyi's paywall and not being able to scrape more data due to the limitations of the scraping process and the data actually available on the website. This limited the time series analysis and overall accuracy of the research.

5 Conclusions

In conclusion, the research presented in this project sheds light on key factors that contribute to higher salaries in tech and STEM fields, in addition to best cities for professionals in these industries. The findings provide valuable insight for individuals at different stages of their careers, ranging from high school students to college students and those looking to change careers. The data used in this research came from various sources, including levels.fyi, U.S. Bureau of Labor Statistics' "Occupational Employment and Wage Statistics," and teleport.org. Data cleaning was performed to remove irrelevant columns and summarize the remaining data. Data exploration was conducted through distributions across categories, time series analysis, and an interactive dashboard of tech jobs and their salaries across the United States by city and state. Managerial roles and having an advanced degree among a few other factors contributed most to an increase in salary. Cities were identified in the San Francisco Bay Area, Seattle Area, and New York City Area as locations for positions with solely high salaries in tech jobs. Now, if both high salary and low cost of living are considered, then Austin, Dallas, Raleigh, Detroit, and Phoenix would be more appropriate.

Overall, this research provides valuable insights into the tech industry and STEM fields, and it offers a practical tool for individuals who are interested in pursuing careers in these fields. By understanding the nuances behind jobs in the tech industry and STEM, young professionals can better position themselves for successful careers in this rapidly growing field.

In the future, the primary improvement should largely involve finding data sets with more thorough and diverse STEM fields than solely Software Engineering. While the data collected hints at the popularity of this field and a collective willingness to share its lucrativeness, the scope of conclusions was severely limited as a result. Regarding the data product, an improvement to the breadth of data that the map can show would be impressive. An ability or set of features to show more time series data relevant to high school or college students regarding fields of interest would be extremely beneficial to the power of the tool in providing more holistic guidance to the user.

Appendix A

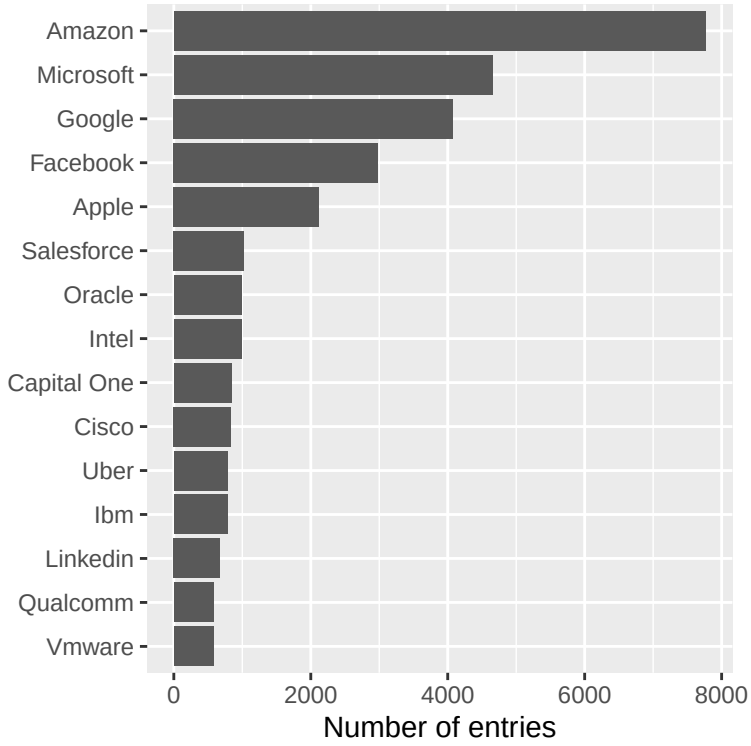

Figure 9: Distribution of top 15 companies.

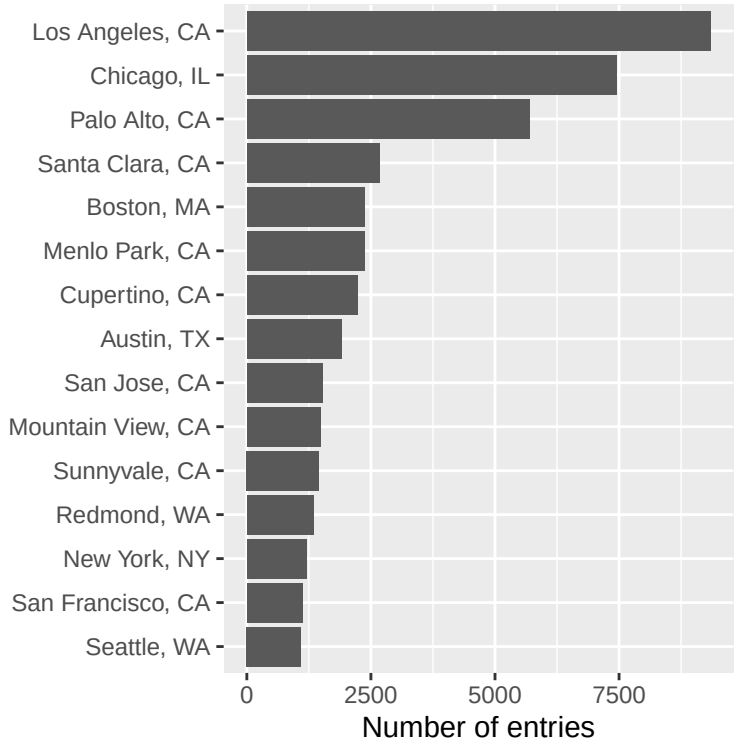

Figure 10: Distribution of top 15 cities with most reported jobs.

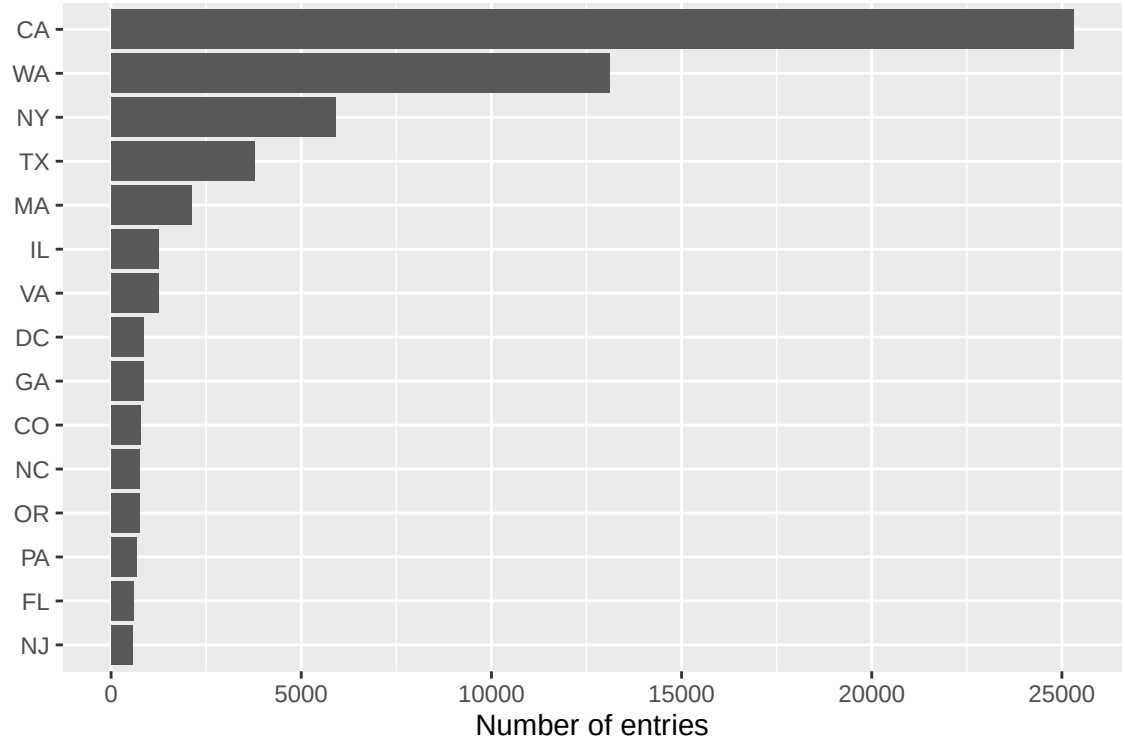

Figure 11: Distribution of top 15 states with most reported jobs.

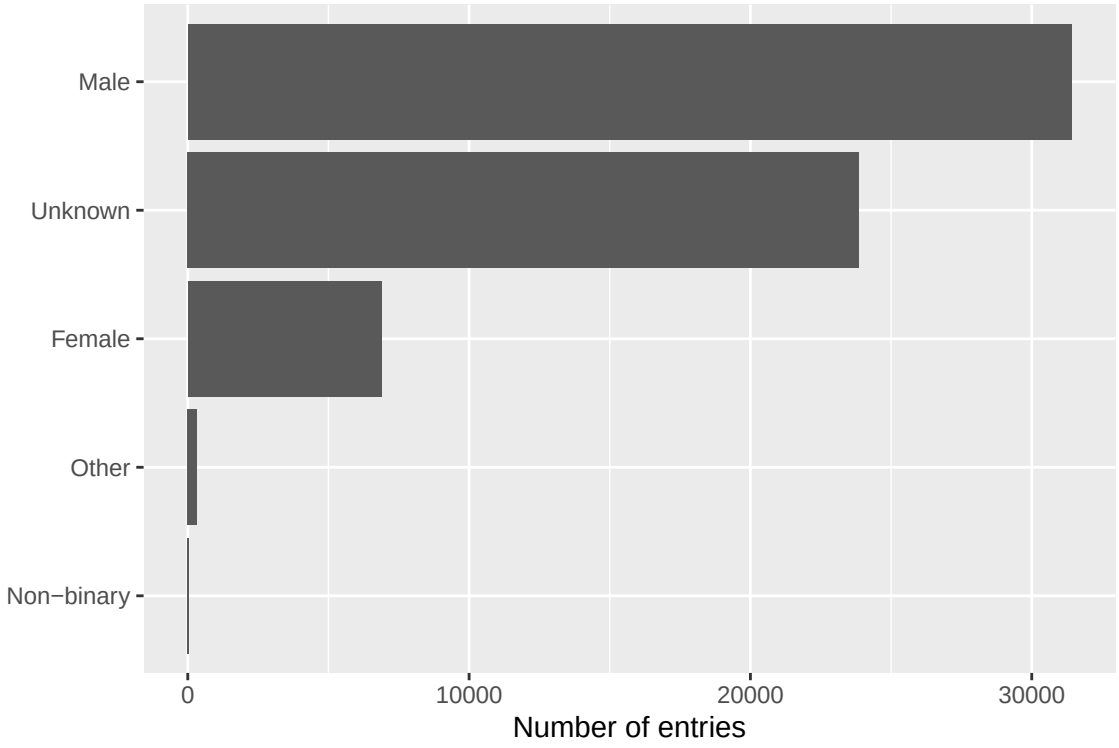

Figure 12: Distribution of levels.fyi entries by gender.

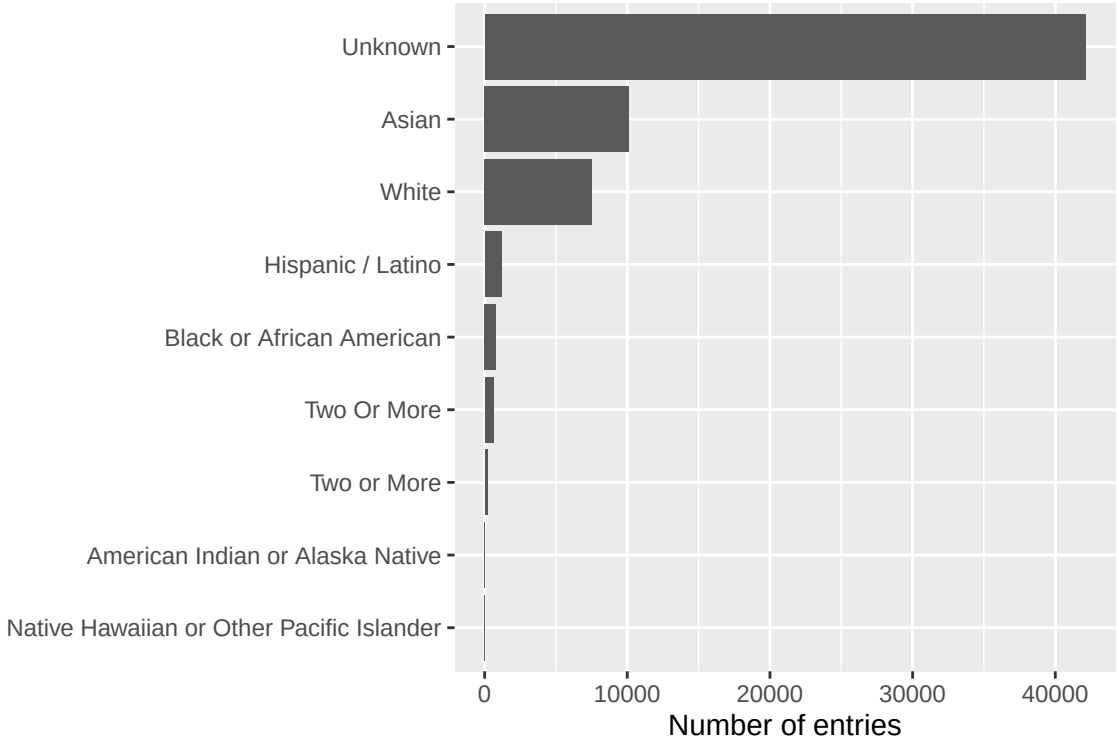

Figure 13: Distribution of levels.fyi entries by race.

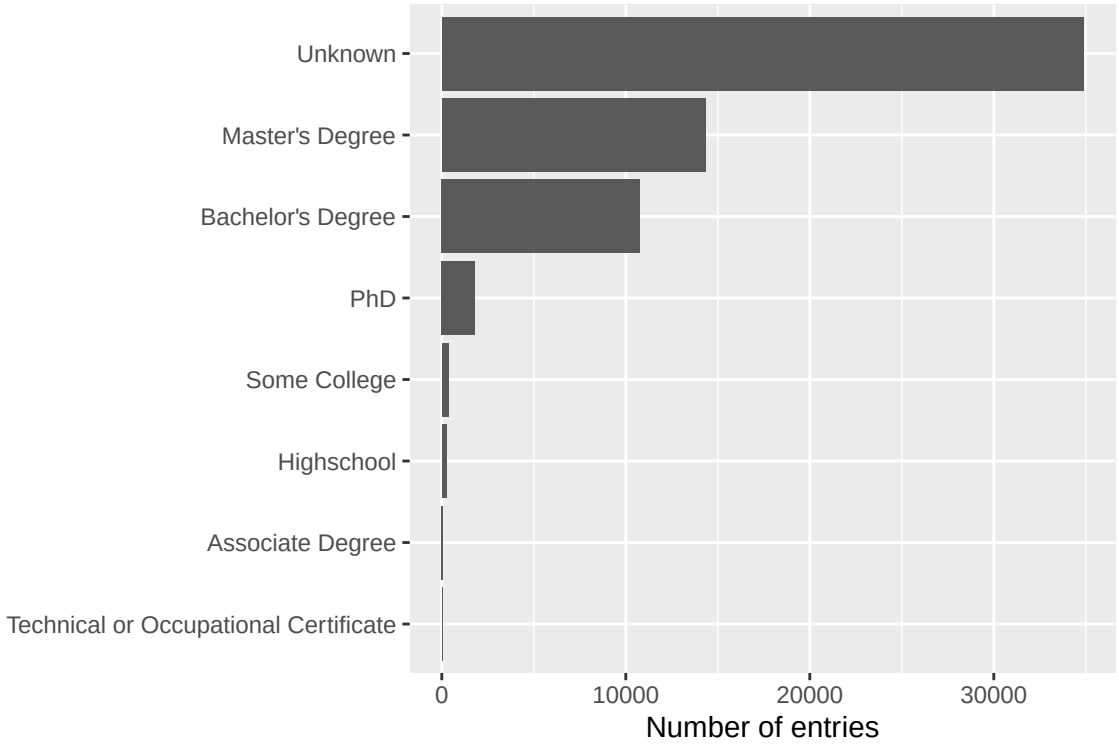

Figure 14: Distribution of levels.fyi entries by education level.

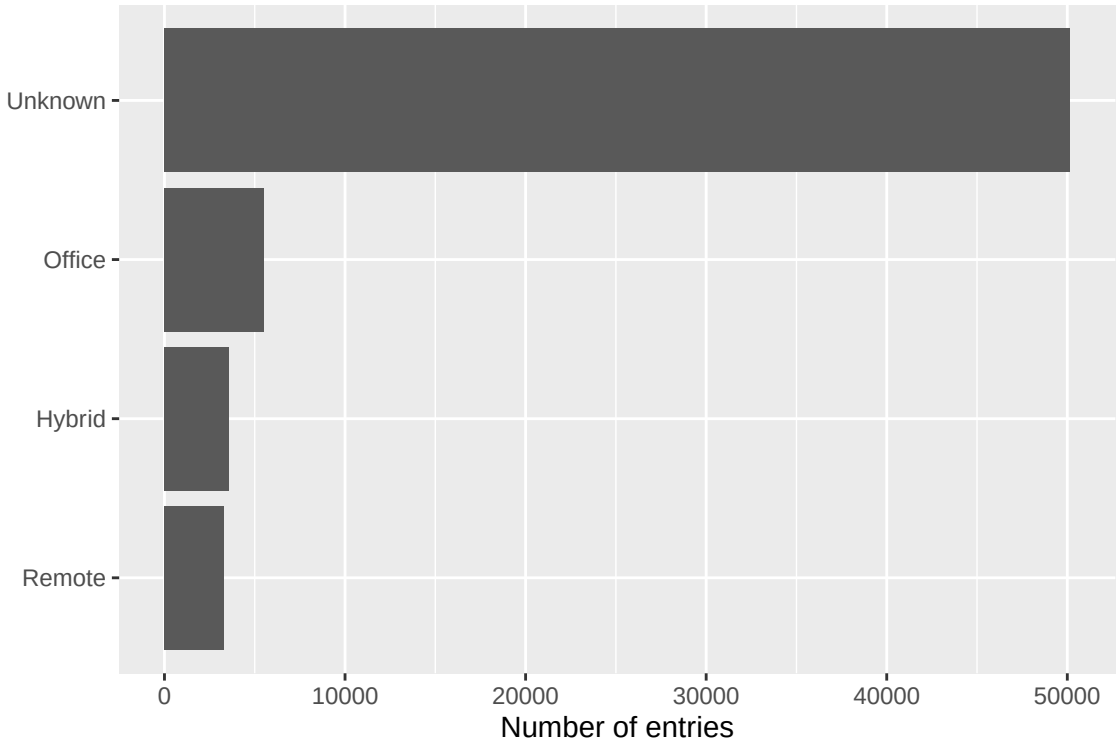

Figure 15: Distribution of levels.fyi entries by work type.

Appendix B

Table 9: State Numerical Summary 1

	TOT_EMP	EMP_PRSE	JOBS_1000	LOC_QUOTIENT	A_MEAN	MEAN_PRSE
Min.	30.00	0.00	0.00	0.07	19410.00	0.00
Median	740.00	9.70	0.44	0.87	87060.00	2.60
Mean	5740.72	12.48	2.04	1.08	84935.08	3.57
Max.	662110.00	49.90	74.02	18.84	163460.00	28.90
SD	26015.75	9.88	5.78	1.08	24986.79	3.24

Table 10: State Numerical Summary 2

	A_PCT10	A_PCT25	A_MEDIAN	A_PCT75	A_PCT90
Min.	16490.00	17060.00	17970.00	18430.00	23260.00
Median	48930.00	62360.00	79440.00	100880.00	127720.00
Mean	52019.10	64128.63	80924.96	100959.75	122155.04
Max.	130360.00	142940.00	164840.00	195410.00	207980.00
SD	16281.03	19960.15	24924.01	31415.18	37523.32

Table 11: National Numerical Summary 1

	TOT_EMP	EMP_PRSE	A_MEAN	MEAN_PRSE	A_PCT10
Min.	1120.00	0.30	35940.00	0.30	23790.00
Median	56640.00	1.95	97000.00	0.80	55810.00
Mean	253500.77	2.73	91317.69	1.23	51951.03
Max.	4654750.00	10.00	122970.00	5.60	77440.00
SD	652429.72	2.12	22855.55	1.14	12648.49

Table 12: National Numerical Summary 2

	A_PCT25	A_MEDIAN	A_PCT75	A_PCT90
Min.	29040.00	32350.00	39230.00	47700.00
Median	71815.00	95300.00	119395.00	151060.00
Mean	66041.79	86084.74	110062.95	136581.92
Max.	96130.00	123430.00	156140.00	187430.00
SD	16590.11	22318.38	29087.76	36075.06

Appendix C

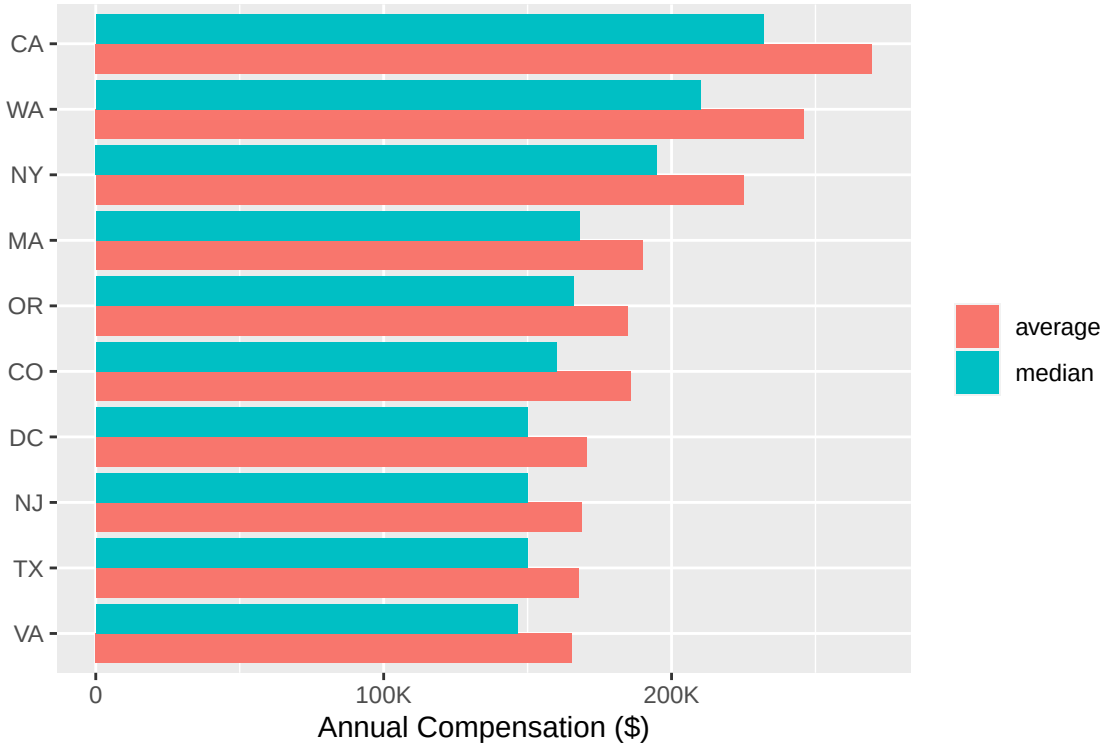

Figure 16: States with highest annual compensation (minimum 25 reported jobs).

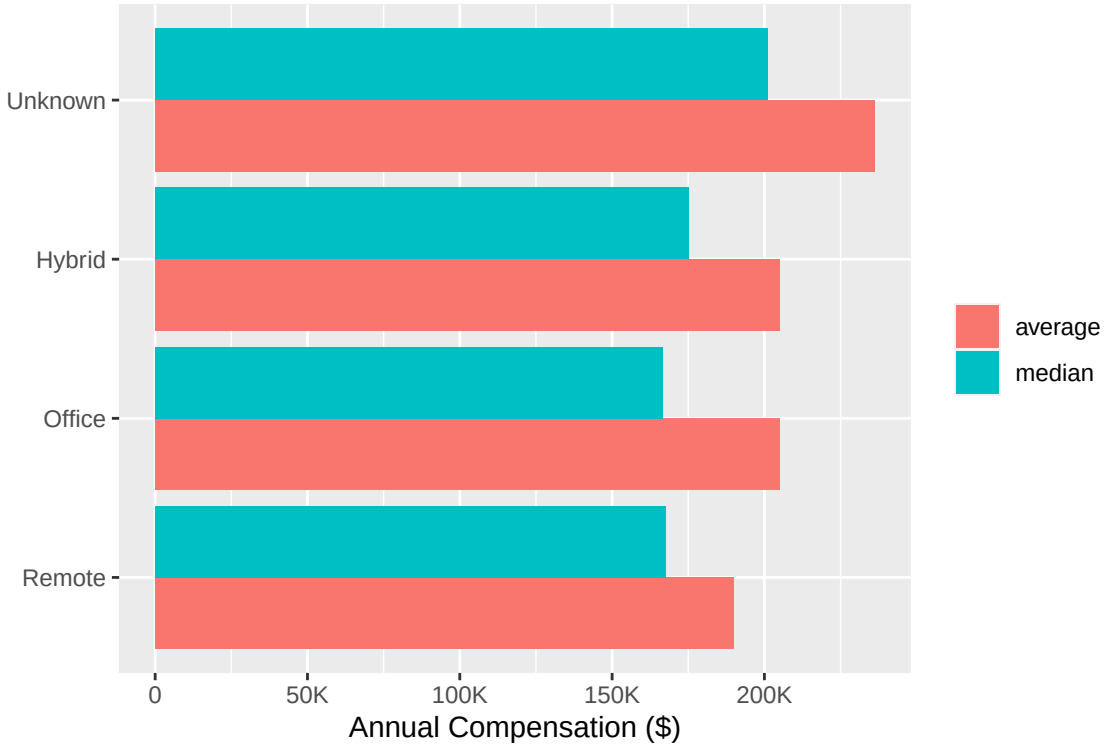

Figure 17: Average annual compensation by job location type.

Figure 18: Annual compensation of top reported titles over time.