CS2410 Final Report

Scott Chang
California State Polytechnic University, Pomona
Pomona, CA
swchang@cpp.edu

Marc Cruz
California State Polytechnic University, Pomona
Pomona, CA
marccruz@cpp.edu

Andrew Benavides
California State Polytechnic University, Pomona
Pomona, CA
aibenavides@cpp.edu

***Abstract*—The report is an effort to combine data on salaries and cost of living to determine if there is a connection between local purchasing power for a given job role. This is relevant to CS majors, for which this report is intended, and is an attempt to provide information that will allow individuals to determine what cities could allow the most benefit with their intended occupation. To do so, information on the cost of living and salaries for varying job titles were collected, cleaned, and enriched. We use this data to categorize all occupations' potential purchase power based on city. To keep the project to a manageable scale, only jobs that fall into the 'Computer and Mathematical Occupations' job taxonomic family were analyzed. Although other data categories within the Computer and Mathematical Occupations are available for analysis, we focus specifically on salary pay and work cities. We normalize pay data in an effort to understand the potential local purchasing power, and the data set we used is from 2021. Other features available for further future analysis include Age, Industry, State, Gender, Race, Years of Professional Work Experience, Years of Professional Work Experience in the Field, and Highest Education Level.**

***Index Terms*—horizontal boxplot, purchasing power, housing cost index, salary**

I. Introduction

Starting in 2021, as the economy began recovering after the start of COVID, the US experienced rising inflation, well above the 2% mark targeted by the U.S. Federal Reserve for price stability [1]. Because different parts of the country have different wage scales and living costs, the question arises: **in which city or area would a person have the greatest purchasing power for a given job role?** The objective of this project is to conduct an analysis and provide findings on the purchasing power based on particular job occupations in different US cities, in order to give individuals more insight into the purchasing power that their occupational salary provides. To achieve this, various features in the data set such as Age, Industry, State, Work City, Gender, Race, Years of Professional Work Experience, Years of Professional Work Experience in the Field, Highest Education Level, and Job Title were analyzed. The focus of this project is centered on the Work City although other features can be used for additional insight.

II. Data Source

To conduct this research, we required data on the salaries, the type of job role held, and the city in which they work. We also required data on the Cost of Living (CoL) index of cities across the United States. We carefully look at the data set sources we used to see the features and information contained, to determine that it is usable for the scope of this project as well as any additional data for additional enriched analysis.

A. Salary and City Data

Our source for salary and city data came from the "Ask a Manager Salary Survey" [2] for 2021. Ask a Manager is a website focused on professional management themes with a variety of topics, one of which is to conduct a survey to get a glimpse of salaries across the US. This data set is publicly available, and while it was for 2021, the survey is still open and actively collecting data. It is feature-rich in the survey questions asked and contains the key salary, job role (as determined via Job Title), and city location data required for this project. At the time of download, it carried over 24,000 entries and in addition to the required information, it also contained other attributes such as Age, State, Gender, Race, Industry, and Education. The data distribution of entries is depicted in Figure 1.

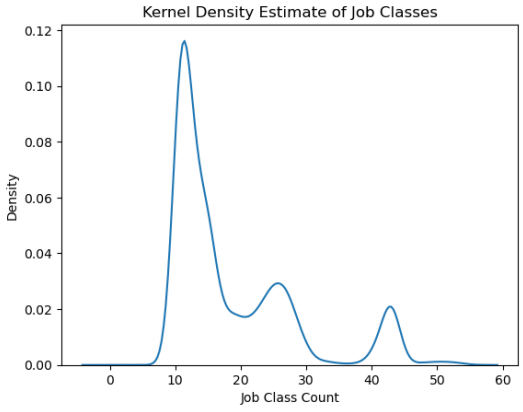

Fig. 1. Kernel density of entries categorized job classes/occupations

B. Cost of Living Index

We then utilized the AdvisorSmith Cost of Living Index, to extract the CoL index based on US cities. AdvisorSmith's CoL index is publicly available and was constructed using reliable and publicly reliable sources such as the US Bureau of Labor Statistics, the US Department of Housing, Zillow's Home Value Index, the US Department of Energy's Natural Gas data, and the Bureau of Economic Analysis. This source contained index information on 510 city locations across the United States. AdvisorSmith is a private company that uses its own research in developing its business, consulting, and insurance products. The combination of a business motive and the use of reliable sources lends credence to this as an overall trustworthy source [3]. The data distribution of the Cost of Living Index data is depicted in Figure 2.

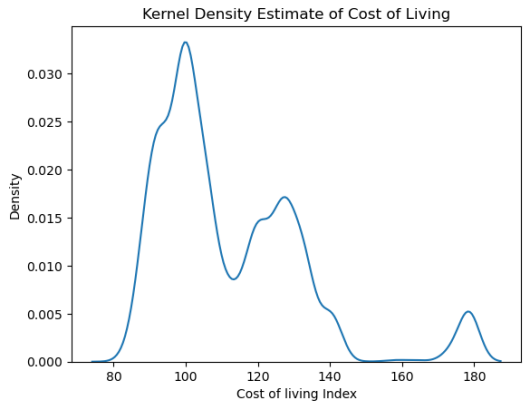

Fig. 2. Kernel density of entries Cost of Living Index

It should be noted that early in the project, we used another Cost of Living data source from Numbeo but transitioned in favor of AdvisorSmith data. The Numbeo source offered only 94 US cities in the reporting year of 2021. To expand the number of cities to allow better analysis, we looked at nearby years. Numbeo offers 6-month intervals in data years (eg "2021 mid-year") so we looked at 2020 mid-year, and 2021 mid-year reporting periods to check for additional city locations reported. We also looked at the 2022 beginning year for additional cities but understood since inflation was already rising by then it may affect our CoL analysis. In total, with Numbeo data, there was only a total of 109 available city locations available. Further, because Numbeo is publicly crowdsourced, it was deemed less reliable than AdvisorSmith data.

III. Methods

A. Data Preparation

1) Salary Data: The first step in data preparation for this project involved removing rows that did not meet the criteria set for the study, specifically, those not compatible with the project's objectives. Afterward, the team identified all unique job titles and industries in the Ask a Manager data set, using the "Job Title Context" column as a supplementary source of information to make informed guesses on ambiguous job titles. Since the same job title can have different meanings across different industries, standardizing job roles was necessary. The team assigned a BLS Taxonomic code to each job title to identify the job family/type and used O-Net keyword searches to help clarify any unclear roles. When no information was available, informed guesses were made based on the nature of the work.

In total, the team identified 1 4,086 d istinct, self-reported job titles/industry combinations. After joining the table back to the original data set, the team had a final t able o f 17,025 usable records. Since the process of standardizing job titles with their taxonomy codes took an exorbitant amount of time, to improve productivity we took advantage of Excel's "Power Query" to quickly filter, transform, and clean the extracted raw data. We made values uniform, trimmed white spaces, and put Names and Titles in the proper case. Additionally, the team selectively chose data from the year 2021 only and excluded rows not in that time period. We also required entries with salaries given in USD currency and required respondents to reside and report their domestic US city work location. Entries that explicitly indicated remote work were excluded. Finally, the team renamed some cities that are suburbs of a larger US city to be considered part of the target city if the target city was identified in the CoL index (i.e. Pomona is a suburb of Los Angeles, Tempe is a suburb of Phoenix, and Bellevue is a suburb of Seattle). Similarly, some cities are "dual cities", i.e. large cities that are in close proximity to each other, are re-identified based on the listing on the CoL index (e.g., Dallas/Ft Worth as Dallas, Arlington/VA as Washington DC, and Minneapolis/St. Paul (Twin Cities) as Minneapolis).

2) Cost of Living Index Rescaling: According to AdvisorSmith, they scale their cost of living index to 100. With 100 being the average cost of living for living in the United States. If a city has an index of 120, it indicates that the city has a CoL that is 20% higher than the average American city. However, within the 510 locations reported in the index, we found that the overall median value was 91.9. We, therefore, conclude that most city locations listed are in fact living "below average", with the implied assumption that most city locations may not have high earners. Because payscale is relative, and we believe it is more meaningful to compare positions based on the number of those earning more and those earning less, comparing the median values of reported data points was more meaningful.

To recalibrate the median value of 91.9 as the new center point of comparison, we rescaled by taking the distance of each city's index to the median and dividing it by the median. As an example, if a city was previously CoL = 100, rescaled against the median value, the city is now Relative CoL = 108.8. The calculation is 1 + [(100 - 91.9) / 91.9] = 1.088. To reach the final form, we multiply by 100. The rescaled Cost of Living Index is then joined to the "Ask a Manager Survey". Where the CoL index value was available for a city listed in the Ask a Manager survey, the CoL index was matched.

B. Data Occupation Labeling

Originally, the "Ask a Manager" data set was large. This data set contains 27,935 rows and 18 columns of data. Among the data, respondents were allowed to self-report their job title and industry and provide optional context to the nature of their roles. It was quickly discovered that job titles come in many variations. There were variations of titles that did the same type of job role. Some job titles also mean different roles in different industries. We termed this phenomenon "Job Role Ambiguity".

For example, there is a record of a job role "Analyst" working in "Organized Labor". The context provided was "I work with data", but we're not sure what type of data. If "working with data" meant business intelligence, it would be a job class code of "13", but if it was related to database management or other digital record keeping, it would be a code of "15".

Another example is the job title of "Knowledge Analyst". There were two records, each in a different industry, however, one provided the context "Analyze business and project requirements and develop procedural documentation" which would carry a code of "13" for business and finance roles, whereas the other provided the context "I'm basically a librarian", which is assigned to code "25" for library occupations.

In a third example, and perhaps most commonly encountered, generic titles like "Specialist" can have a wide range of meanings and roles. Frequently, a Specialist can be an assistant or support worker of some type (e.g. part-time general office worker is a code #43), whereas other roles define a much more specialized role (e.g. a legal specialist requires years of education and training, and as a legal occupation assigned code #23).

We cross-referenced reported job titles with the BLS taxonomic code list and made use of an ONET online occupational keyword search to help us narrow down and identify the type of job role. Where roles remained ambiguous even after cross-referencing, we made an informed best guess and assigned the most likely job code.

The list of occupation categories is depicted in Figure 3. Generally, any title that is at least a Manager falls under code #11, so all the VPs, Heads, Directors, Chiefs, and managers are included in code 11. The supervisors below are then categorized by their respective job categories from the list in Figure 3. All teachers are categorized in code #25. A title that contains "Administrator" can vary, they can be considered code #11, or they can be code 43, depending on the level and the type of work. Administrative "Assistants" and "Specialists" get code #43, but other types get the code of their job category (i.e. Dental #31 or Teaching Assistants #25). For Business Analysts, as it is too broadly used in different occupations we simplified their categorization into two codes: Code #13 for anything finance/business intelligence related or code #15 for Math/Computing occupations. Librarians, Library Technicians, and Library Specialists are code #25 (same as Education), but a Library Assistant is code #43 for Office Support and technician roles. Engineering Repair Technicians could be code #17 or code #49, depending on the kind of engineering or machinery they are repairing.

SOC Code	Description
0	All Occupations
11	Management Occupations
13	Business and Financial Operations Occupations
15	Computer and Mathematical Occupations
17	Architecture and Engineering Occupations
19	Life, Physical, and Social Science Occupations
21	Community and Social Service Occupations
23	Legal Occupations
25	Educational Instruction and Library Occupations
27	Arts, Design, Entertainment, Sports, and Media Occupations
29	Healthcare Practitioners and Technical Occupations
31	Healthcare Support Occupations
33	Protective Service Occupations
35	Food Preparation and Serving Related Occupations
37	Building and Grounds Cleaning and Maintenance Occupations
39	Personal Care and Service Occupations
41	Sales and Related Occupations
43	Office and Administrative Support Occupations
45	Farming, Fishing, and Forestry Occupations
47	Construction and Extraction Occupations
49	Installation, Maintenance, and Repair Occupations
51	Production Occupations
53	Transportation and Material Moving Occupations

Fig. 3. List of Job Occupations and its assigned code value

The final data table created to be used for analysis has the following characteristics:

- Data was from the year 2021
- The salary was given in USD currency
- Must reside in the United States
- Must have reported a domestic US city work location
- Entries that explicitly indicate remote work was excluded
- Job Titles are classified by job family/taxonomic code as defined by the Bureau of Labor Statistics
- Informed best guesses used when roles remained ambiguous
- Suburbs of metro areas are considered to be part of the larger target city and grouped and listed with their target city.
- CoL data rescaled against the median value of CoL index data
- Where CoL listing was available, the CoL index was matched to their respective city.

C. Coding/Visualization

1) horizontal boxplot: To create the visualizations depicted in Figure 4, several functions were created:

- true_pay()
- normalized_pay()
- raw_pay()
- color_palette()

- label_formatter()

The function true_pay() is the main code and takes input arguments to specify job class, visualization category, # of minimum observations of the visualization category needed before inclusion, and optional search terms. The main code calculates the different variables needed to visualize and passes on those values to the support functions. The code is flexible and allows many types of analysis. For the purpose of this report, we are interested in analyzing job code 15 (Computer and Mathematical Occupations), the work cities that are reported, and more than 9 reported instances of the work city for visualization. We used a minimum of more than 9 observations because we felt that instances, where the cities were reported less, would not offer a reliable indication of common pay range.

The functions normalized_pay() and raw_pay() are support functions that create the 2 plots. It receives the data frame and additional arguments from the main function to use in the graph. color_palette() and label_formatter() are two more helper functions that take arguments passed by the support functions to further enhance readability and color coding.

We color code the data in the following manner. Color coding conveys where the median of the city (on a nominal basis) is first located compared to the overall median and where the median of the city location ends (on an adjusted normalized basis).

- Blue: From Above, Stayed Above Median
- Green: From Below to Above Median
- Red: From Above to Below Median
- Black: From Below, Stayed Below Median

To further supplement the color coding, we created a metric called "Magnitude of Relative Movement (%)" ("MRM") to convey the distance of the movement of the median. Taken together, it allows the user to understand how the data point moved and by how much.

IV. Findings/Suggestions

We are able to produce two charts, one showing the plot of normalized median and pay scale ranges by city, and another chart showing the raw, reported, unadjusted median and pay scale range. In the normalized pay chart, we see the data sorted by median value in descending order. Cities with a higher normalized median value are at the top and lower median values are at the bottom. Each bar also indicates the relative distance to the overall median of the plot data. In the nominal, unadjusted pay chart, we see how the pay ranges are first reported in the data. Comparing the two charts lets the user understand the reported pay values and how it compares after adjusting for the Cost of Living.

Our findings suggest different ways to answer the problem statement "**For a given job role, in which city does your income provide the most purchasing power?**". The answer that is most relevant to the user depends on the context from which they are answering the problem statement. For this project, we use the inputs: true_pay(15, 3, 9, ''). In terms of solely considering median values, the city with the most relative purchasing power is the first city listed. In our chart, Pittsburg, PA is the city with the best purchasing power after adjusting for CoL with a high MRM. However, cities that are color-coded green but have a large positive magnitude of movement could be a very satisfactory result if the user is considering an unconventional location with salaries that could offer very comfortable living. In this context, St. Louis, MO could offer the best value as it has the highest MRM. From our data, both of these cities offer seemingly average nominal pay but offer excellent purchasing power after CoL adjustment.

Additionally, if the user is searching for a city that has high nominal pay that remains MRM stable after adjustment of CoL, Boulder, CO, and Seattle, WA both have very little MRM movement. We interpret this to mean that pay levels are very nicely balanced with CoL. Boulder has nearly "0%" movement, and Seattle, as a location with a tech reputation, has a very low MRM movement.

In general, when viewing the normalized pay chart results, it is assumed that any city location that has an adjusted median value above the overall chart median would be desirable (green and blue bar cities). However, on a pay scale basis, within this grouping, cities higher on the scale are seemingly more attractive. Finally, it is important to note that the results above are specific to the input arguments entered. Our code is flexible, and a different output may be returned if the user specifies different starting inputs. Results of the charts can be depicted in Figure 5 and Figure 6.

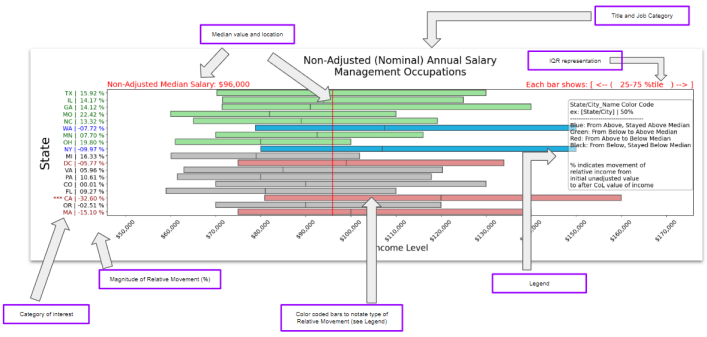

Fig. 4. Modified horizontal boxplot explanation

V. Future Work

For future work, there are several potential improvements that could be made to increase the accuracy and scope of our findings. One limitation of our data source, the Ask a Manager Survey, is that it may not have reached a representative sample of the workforce. In our data exploration, we found certain occupational job classes to be over-represented. To address this, a larger and more reliable data set could be obtained from large market share companies or entities that have access to payroll information, such as ADP, Paychex, and Paylocity. Companies and entities like them are likely to have more representative payroll information, as well as information on

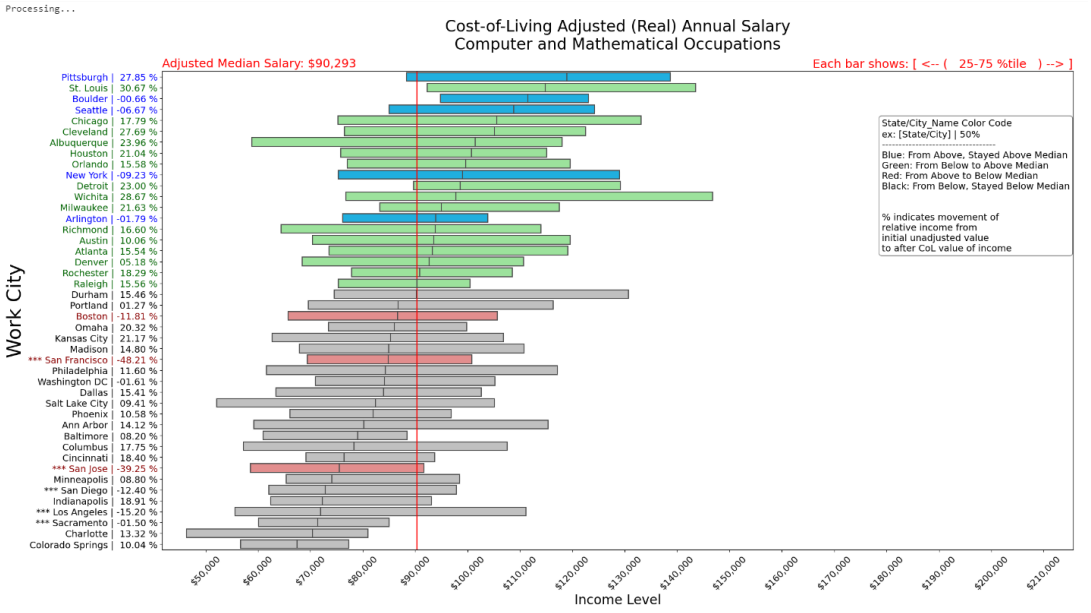

Fig. 5. Horizontal Boxplot of Cost of Living Adjusted Real Annual Salary for Computer and Mathematical Occupations

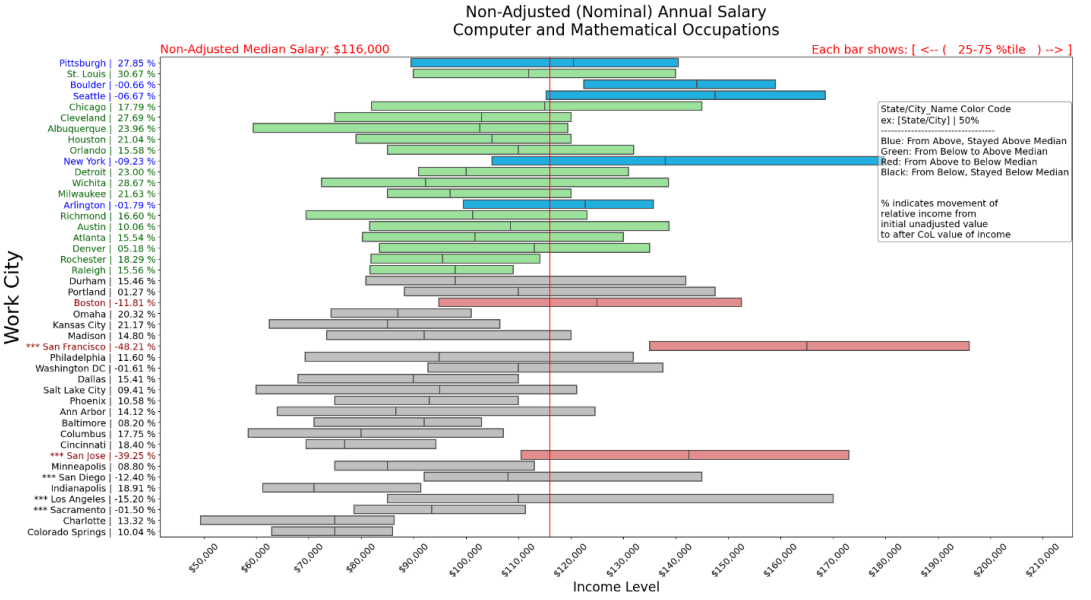

Fig. 6. Horizontal Boxplot of Non-Adjusted (Nominal) Annual Salary for Computer and Mathematical Occupations

their occupational category. This could improve the accuracy of determining their job role.

Another potential improvement is to use the ACCRA Cost of Living Index, which is considered the gold standard and most comprehensive in Cost of Living studies, done by the Council for Community and Economic Research (C2ER). Although this data is expensive to access, additional funding could allow for a more accurate analysis. Because they collect data year by year, if the ACCRA CoL index was used, a possible avenue of investigation is a comparative study on purchasing power of different cities over time. Finally, including compensation factors such as stocks, benefits, and other amenities, could provide a better perspective on purchasing power.

VI. Conclusion

In conclusion, the group project provided a guideline for finding the US city with the most purchasing power specifically for Computer and Mathematical Occupations. Doing so by going through the process of adjusting and rescaling salaries with correlations to salaries and U.S. cities with relevant suburbs included. The results are mainly represented through horizontal boxplot IQR ranges of real and nominal

annual salaries to show the movement of the 25th percentile to the 75th percentile of each city's annual salaries. Comparing the change in annual salaries between the two graphs to determine the movement and with the movement provides a measurement of how much a city's purchasing power has positively or negatively increased. The measurements were further color-coded and represented in percentage form to allow efficient identification of potential higher purchasing power. Lastly, we want to emphasize that the charts are merely a guideline. Our project is limited in funding and the Ask a Manager Salary survey has a limited reach of respondents. It is over-represented in certain occupational categories. Larger data sets will allow for larger, more reliable analysis and could better paint the picture of the state of CoL-adjusted annual salaries.

REFERENCES

[1] The Fed, "The Fed - Why does the Federal Reserve aim for 2 percent inflation over time?," Board of Governors of the Federal Reserve System, 2015. https://www.federalreserve.gov/faqs/economy_14400.htm

[2] A. a Manager, "look at 24,000 people's real-life salaries and sort by industry, job, and location," Ask a Manager, May 05, 2021. https://www.askamanager.org/2021/05/look-at-24000-peoples-real-life-salaries.html (accessed February 20, 2023).

[3] "U.S. Cost of Living Index by City: Downloadable Data – Advisor-Smith," advisorsmith.com. https://advisorsmith.com/data/coli/

[4] "Northern America: Cost of Living Index by City 2021," www.numbeo.com. https://www.numbeo.com/cost-of-living/region_rankings.jsp?title=2021&region=021 (accessed May 11, 2023).