Data Analyst Positions in Large U.S. Cities

Iyanu Adebisi, Laurie Cripe, Sadiyah Hotakey, Terrence Liu, Richard Tarbell

Introduction

Pursuing a new job can be intimidating, and everyone wants to make sure a new position can support (or improve) their current lifestyle, especially in a different city. In this report we explore over 2,200 job postings for Data Analyst positions all over the United States. While these postings include location, company, and a salary estimate, we wanted to create more meaningful comparisons by integrating the cost of living for the given location. Using a Cost of Living (CoL) index, we decided to look at the top 89 U.S. cities and see how these salaries and positions compare to the CoL for each city.

Data Sources

The two data sources we used were both adequately documented and taken from public records and datasets. As we joined the two datasets together, we completed an inner join and only allowed job postings that appeared within those top 89 cities to remain.

Our job posting dataset was scraped and made available by a user on Kaggle in July 2020. While these positions are over a year old, we feel that they are an accurate representation of positions in the current job market. The CoL indexes were scraped from Numbeo.com and are available under their public use license.

Problem

With a wide range of salaries and CoL throughout the U.S., our approach was to use a reference city with a CoL index in the bottom 25% of our list as a baseline. We selected San Antonio, Texas, which has a CoL ranked 84th out of the 89 cities included in the index. We used the following problem statement for our analysis:

As a data analyst in San Antonio, is it worth it to look for data analyst positions in other big cities? (i.e., Given the average salary and cost of living, how would your standard of living compare to your current situation?)

Through various comparisons of required skills, employer industries, and most importantly salaries, we hope to help potential job seekers make the best career decision compared to their current living situations.

Data Cleaning & Validation

Joining Tables and Addressing Missing Values

To focus on the job postings that provided the most meaningful comparisons, we joined the tables to exclude job postings in cities that were not listed in the CoL index. To exclude incomplete job postings, we excluded positions where the job title, salary estimate, company name, or industry were missing. After joining and cleaning the data, we used over 1,100 unique listings in our analysis. We verified the completeness of the joined data with a summary statistics report, and we used summary tables to compare the original and cleaned data sets.

Conversion of Salary Estimate Ranges to Numeric Columns

Within the Glassdoor data, the salary range values were formatted as "$35K-42K (Glassdoor est.)". The provided ranges were not consistent, with large overlapping ranges between estimates. Using

the TRANWRD function in SAS, we created two calculated columns containing numeric values for the upper and lower estimates: Lower_Salary_Estimate ($35,000) and Upper_Salary_Estimate ($42,000).

For the summary tables and analysis, we created a calculated column for the average salary estimate ($38,500). In order to present the charts and summary tables in a readable format, we created a user-defined format to group the average salary estimates by tens of thousands.

Classification of Job Levels

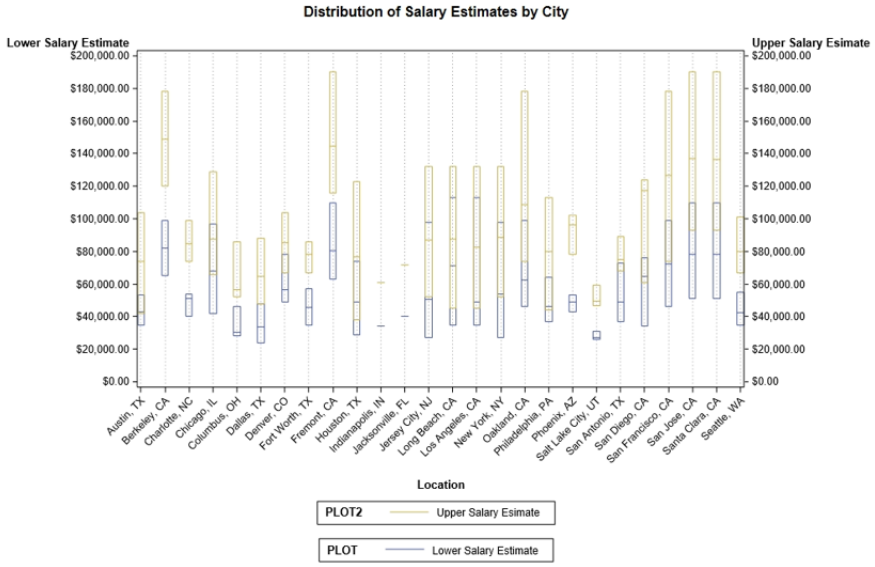
In order to compare the positions more efficiently, we created a recoded Job_Title column. We used keywords to categorize the job postings into internships, junior, mid-level and senior positions. For example, a job title listed as "Senior Forensic Data Analyst" would be recoded as "Senior Data Analyst" using the keyword "Senior". The keywords and their corresponding labels are shown in the table below.

Original Job Title	Recoded Job Title
I/II/III	Data Analyst
IV/Sr/Senior	Senior Data Analyst
Manager/Mgmt	Senior Data Analyst
President /VP	Senior Data Analyst
Principal	Senior Data Analyst
Jr/Junior	Junior Data Analyst
Summer/Intern	Data Analyst Intern

Analysis

Comparison of Salary Estimate Ranges by City

To provide a visual representation of the original salary estimates, we created a box plot which displays the original lower and upper salary estimates by city.

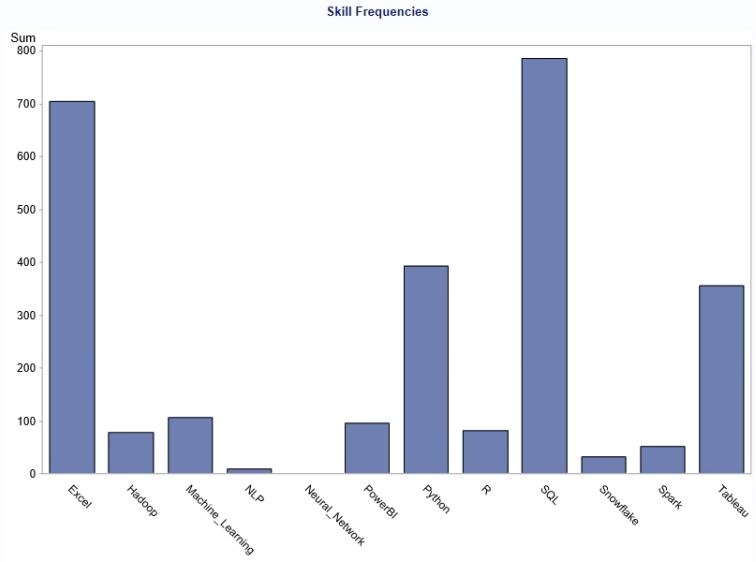


The initial estimates span a wide range and overlap to a large extent. In Berkeley, California, the posting with the lowest estimate range spans from approximately $60,000 to approximately $120,000, while the posting with the highest estimate range spans from approximately $100,000 to $180,000. In cases such as this, the actual salaries may not be accurately represented if we used the lower or upper estimates in our analysis.

Although the average salary estimates and CoL index are necessary to make more meaningful comparisons between cities, these averages are a representation of very large estimate ranges. While the lowest average salary estimate for Berkeley is approximately $90,000, the actual salary for this job posting may be as low as $60,000. Potential job seekers using this analysis should take the initial range of estimates into account before reaching final conclusions based on the average comparisons.

Analysis of Listed Job Skills

We created columns for relevant skills listed in the job descriptions, such as Python or SQL. We used one-hot encoding to create a count of the number of job postings with each listed skill. The chart below shows the frequency of each skill listed in the final job. These columns were then analyzed based on their independent frequencies as well as their combinations' frequencies in relation to the midpoint of the listed salary range.

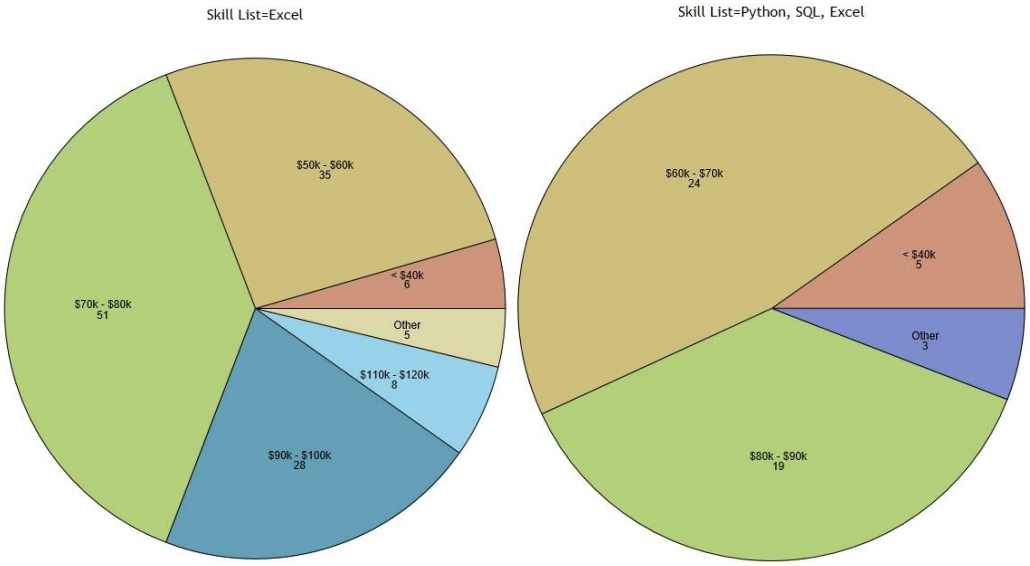


We used a combined frequency count to compare the most common skill combinations listed in the job postings.

	Skill List	SUM_of_Entry_Count
1	SQL, Excel	191
2	No Listed Skills	151
3	Excel	133
4	SQL	85
5	Python, SQL, Tableau, Excel	62
6	Python, SQL, Excel	51
7	SQL, Tableau, Excel	46
8	SQL, Tableau	33
9	Python, SQL	30
10	Tableau, Excel	19

If a potential job seeker is considering relocating and would like to earn additional certifications or prepare for interviews, these skill combinations may help to focus their efforts on the most valuable skill combinations. We created pie charts to compare the average salary for job postings for each skill

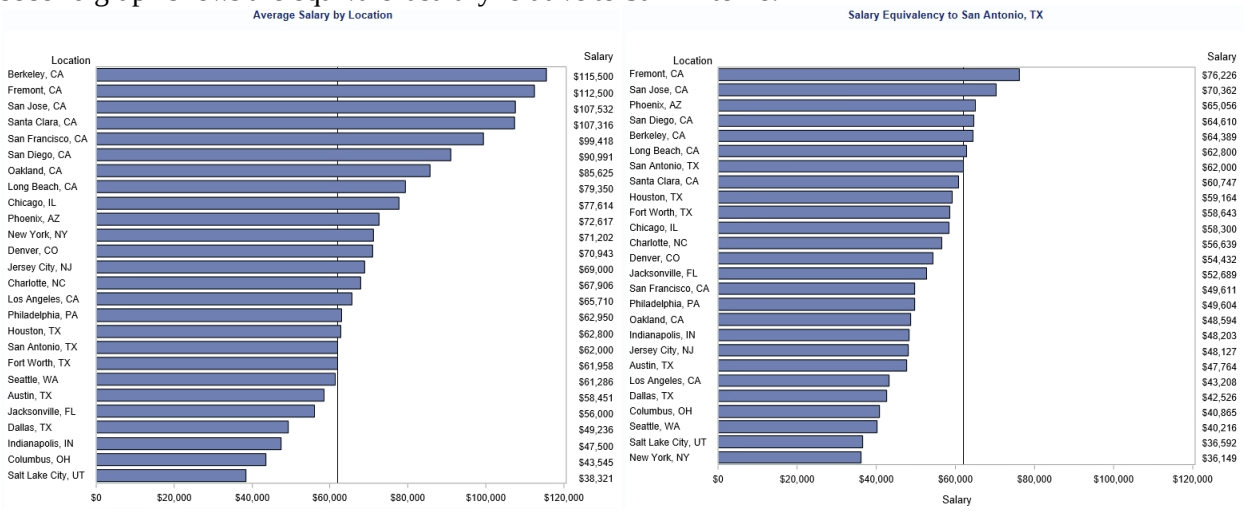
combination. The charts for the most frequent combinations in job postings with one listed skill and three listed skills (Excel; Python, SQL and Excel) are shown below.



Comparison of Indexed Salary Averages by City

Three calculated columns were created to compare salary and CoL across different cities, as well as relative to San Antonio, TX. The first column, Index Relative to SA, recalculated the original cost of living plus rent index relative to San Antonio. This took the cost of living plus rent index obtained from Numbeo data and divided it by San Antonio's CoL (50.77). The second column, Average Salary, is the average of the Lower Salary Estimate and Upper Salary Estimate, summarized by city. The third column, Salary Equivalency to SA, calculates the indexed average salary for each city relative to San Antonio. For example, the average salary for job postings in Fremont, California, is $112,500. This initially seems like a large increase relative to San Antonio's average salary of $62,000. However, due to the difference in the cost of living, this is equivalent to making $76,226 in San Antonio.

The difference between the average salary and indexed average salary is displayed in the following bar graphs. The first bar graph illustrates the average data analyst salary by location. The second graph shows the equivalent salary relative to San Antonio.



Comparison by Industry

To provide job seekers with information on sectors with the greatest level of opportunity, we created a frequency count for each industry. The top ten industries with the highest number of job postings listed in our final joined index are shown below. As shown from the table below IT services is the industry with the most data analyst positions with 232. While information technology is the most frequently referenced sector. With this being said, a data analyst seeking employment would find it beneficial to start searching in the IT services industry with a focus in information technology.

Number of Jobs	Industry	Sector
232	IT Services	Information Technology
192	Staffing & Outsourcing	Business Services
98	Healthcare Services & Hospitals	Healthcare
72	Computer Hardware & Services	Information Technology
66	Consulting	Business Services
52	Investment Banking & Asset Management	Finance
52	Internet	Information Technology
46	Enterprise Software & Network Solutions	Information Technology
34	Insurance Carriers	Insurance
34	Advertising & Marketing	Business Services

Conclusions and Recommendations

When considering job opportunities in different cities, this analysis provides potential job seekers with meaningful comparisons between skill sets that are in demand, industries and cities across the U.S. Each factor will play a significant role in the resulting salary and standard of living for data analysts considering positions in other large cities.

Data Sources

Cost of Living Index - https://www.numbeo.com/cost-of-living/country_result.jsp?country=United+States
Data Analyst Salaries - https://www.kaggle.com/andrewmvd/data-analyst-jobs