Final Project

Lanqing Zhao

Spring 2023

Contents

1 Introduction and Motivation

This project is to investigate 2 main problems:

- In the tech industries specifically in Software Engineering and Development (SWE), is the salary in the United States Mountain states higher than that in Western European countries such as the United Kingdom and Germany?
- Within the US mountains states and / or within the western European countries such as the UK and Germany, is there a difference between salary in SWE positions?

The motivation of this project is from the facts that I live in Denver and wish to know the reality of benefits and salaries of tech industries in the Mountain states, which is the fastest growing region in the US. Additionally, there are debates online about the qualities of life between European countries and the US, and I wish to use this as an example to discuss it statistically.
The project will apply the web scraping to gather data directly from grassdoor.com, which is a job search and sharing website, and process data of salary and geographic locations to compare the regions listed above and determine statistically the answers to the two main questions. The project will also display visualizations of data to depict the comparison visually aside from statistical tests.
The conclusion I get from the project is that the SWD salaries in Greater Denver area and Greater Salt Lake City area are statistically significantly higher than that of Berlin and Greater London area. Additionally, the salary in Denver is slightly higher than that in Salt Lake City, but the salary of Greater London area is not statistically significantly higher than that in Berlin. We will also discuss the limits of this project.
The report is only a presentation of main results, and python notebook is the comprehensive workflow of the entire project.

2 Literature Review

The literature review will focus on the preexisting data on population shifts in the United States based on census 2020 and the general economic data found online of macro-economic between the United States, the UK, and Germany. Since we are studying the salary of those different countries, it is essential to understand why I would pick Denver and Salt Lake City besides

personal reasons, and it is also important to note that besides salary in tech industry how well the macro-economic performs in general in those countries.

The figure 1 just outlines that Mountain States in the United States are the fastest growing region in the entire country, and especially in Utah the population grows the fastest in the country. Colorado, on the other hand, has the population growth faster than the west coast. Therefore, using Denver and Salt Lake City can be valid enough to represent the population status quo of the United States.

The other set of data useful to be used for general picture of this project is

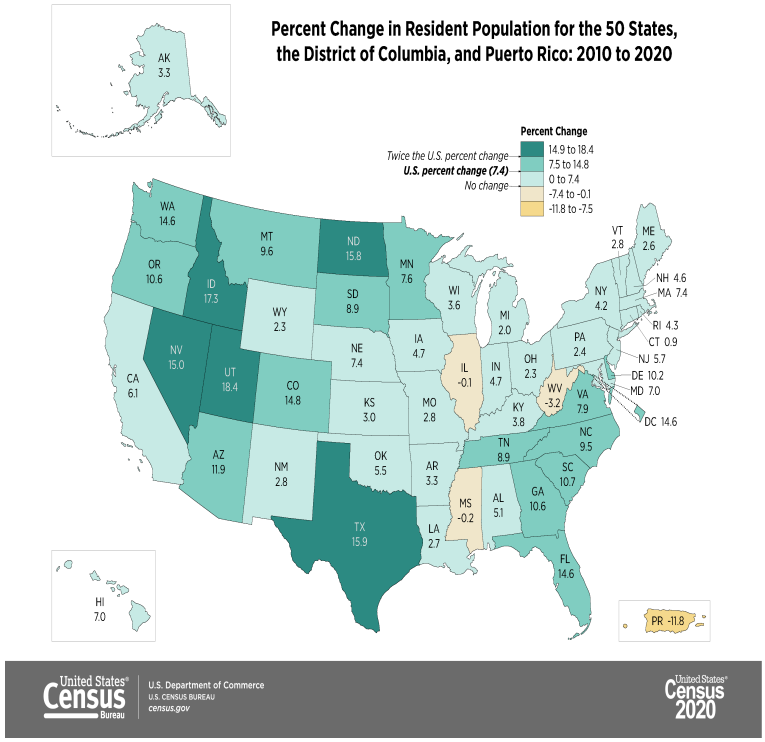

Figure 1: The population growth by percentage of the US states and territories in US census 2020

the overall economics of those countries. The good measurement is nominal GDP per capita as well as the median household income. The data are as the figure 2 and 3. It clearly indicates that the US performs better than the other two countries in those two categories. It is worth mentioning

the household income is before tax, and for Germany it is average income. Therefore. the real median household income is going to be lower than it is for Germany.

Figure 2: The median/average household income of three countries

Figure 3: The GDP per capita of three countries

3 Data Gathering and Cleaning

Data are gathered directly using web scraping from grassdoor.com. The webs that contain those data are found through the search tool of grassdoor. I basically set the keyword to software engineer, and I set the location of job post to be within 50 miles radius of Denver, Salt Lake City, London, and Berlin. The reason for choosing a large area than the city limit is that we all know the nature of sub-urbanization of cities in the US as well as in London or Berlin. Additionally, the job posts are indeed ads posted by the companies that seek to hire new employees, so in this sense the data can reflect the status quos of job market. We will only look for information of locations, company names, the range of salary, and the review of the companies on grassdoor.

After gathering data, we have data set in the form of .csv for two datasets. One is for the US, and the other is for Europe. The fields are

- name-of-company: the name of company
- review: the review by 1-5 stars

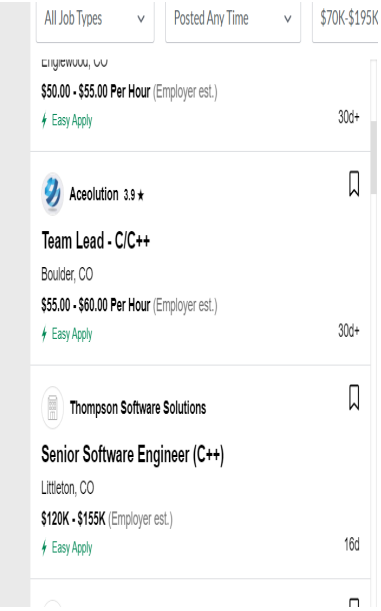

Figure 4: The web interfaces used for web scraping

- name-of-job: the name of the positions
- location: the location of jobs
- salary: a string of range of salary in either annual or hourly rate

If we take a look at dataset, we realize that there are a couple of problems we have to address in data cleaning. First, we have salary in string, so we must find a method to convert it to numerical value(s). Another issue with salary field is that it is in different currency, and therefore we need to convert them to USD to unify the currency. The other issue is that we have a lot of missing data, and some rows do not have review or even salary. We cannot substitute them by our common methods such as fill in with average because we are using data to compare average. The last problem is that we have too many duplicates, and it is because it is an automated posting system that may generate duplicate posts to boost visibility,
The solution are going to be the following. I use regular expression and algorithms to find numerical values and ending phrase in the string to convert the salary in small new columns as high, low, and median values of annual salary. If the salary is in hourly form, it will be converted to annual salary by multiplying 40*52, with the assumption that an employee working 40

hours a week and 52 weeks a year. With additional method attached, the Euro for German data and GBP for the UK data would be converted to USD. The exchange rate would be 1 EUR to 1,07 USD and 1 GBP to 1.24 USD. Then, for missing data and duplicate rows, I drop all of rows with missing data and drop all identical rows. After this, we get 20% of rows left. Thus, we still have about 200+ for the US, 100+ for UK, but Germany data drops to double digits , which is a problem. However, this data cleaning does help us get our data clean and even generates some more features with extra feature engineering.

4 Feature Engineering

As we have already done in the last section, we engineer more features from salary strings. The salary will have three tiers: low, median, and high. For some rows that only have a single value, the three columns will use this single value. This process generates three new columns useful for the analysis.
Another set of feature engineering is related to location information. The location will be encoded in two ways. The first way is to divide all data within 2 data sets into 5 regions: Denver, Salt Lake City, London, Germany, and Non-London, with number from 1 to 5. Another way is to encode country code with US, GB, and DE to represent the data of the US, the UK, and Germany respectively. Those are used for our analysis. Thus, we have generated more columns from our limited data sets. The details are in notebook.

5 Analysis

The analysis has two parts: visual analysis and statistical tests. In visual analysis, we generally display the visualization of statistics or results directly, and in statistical test we conduct one way ANOVA.

The details are in the notebook file, and we will just outline the most important ones in this report. The comparison will be between five regions we designated and encoded in last part and be between three countries for the answer of the first question in sec 1. The additional question from question 1 is that if there is any correlation between review and salary level. For the purpose of good representation, we will primarily use median salary as our source for comparison. Figure 5 displays the boxplot, and it shows that Salt Lake City has a lot of outliers at higher end. More than that, it shows

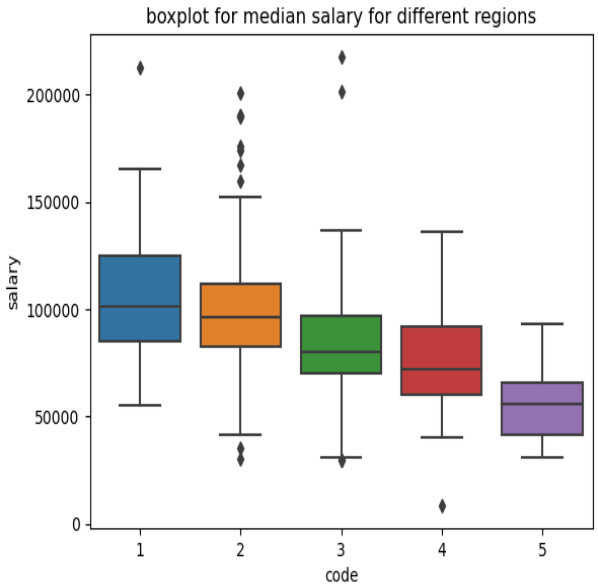

Figure 5: The boxplot of median salary from five regions

that Denver as well as Salt Lake City has higher salary level than Germany and the UK. London has higher salary than Non London. Figure 6 shows the average median salary of 5 regions. It is obviously showing that Denver has the highest, and non London area has the lowest. We can observe the order easily.
There is also a comparison between average review, and we display here as well. From figure we can see that the reviews do not have a lot of difference.However, non-London is the highest.

Another comparisons are between countries. We have three countries in total, and they have been encoded in the last part. The boxplot are displayed here to show that US is in whole higher than the other two countries. However, we do observe some outliers, so the 75 percentile , which is used to reduce the effects of outliers, is being used to display the comparison. We observe that the US data shows a higher 75 percentile, and the UK and German data do not have a lot of differences.
After visualization, I also conducted statistical tests between countries and within countries. The details are in notebook. The one-way ANOVA is used to compare the populations of salary between 3 different countries as well as between Denver and Salt Lake City. At the end, we do have the correlation

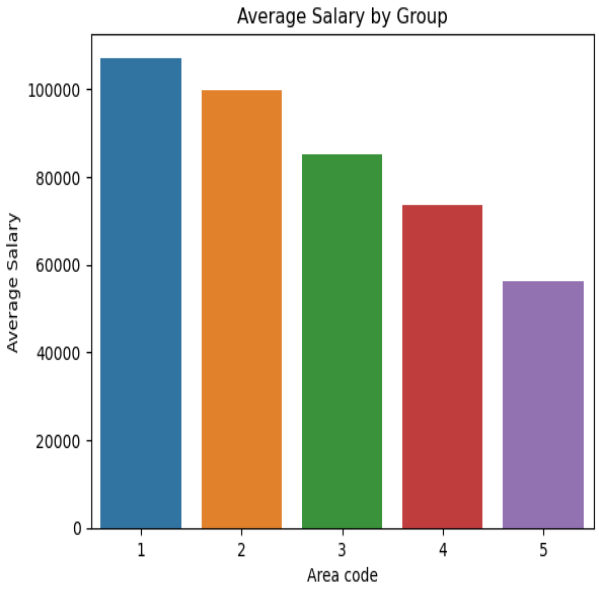

Figure 6: The bar chart of average median salary in five regions

between review and salary level for the 3 countries. The correlation is 0.115, and it means that there is not a strong correlation

6 Discussion and Conclusion

Based on statistical test, we can conclude that there are a statistically difference between salary levels for software engineer jobs in the US and the UK and a statistically difference between salary levels for software engineer jobs in the US and Germany. However, between the UK and Germany, there is not a significant difference. Within the US, Denver is slightly different from Salt Lake City as the p value for F test is around 0.06. The review is not strongly related to the salary level.

However, we do have some issues with our studies. First, we do not have a lot of data for Germany, and it is because grassdoor.com is an English speaking website. The second issue is that we also do not have non-London data, and it is because we select Greater London area as the region. The last is that Berlin may not be a good representation of Germany as it is for a long time behind the average level of economic development in Germany

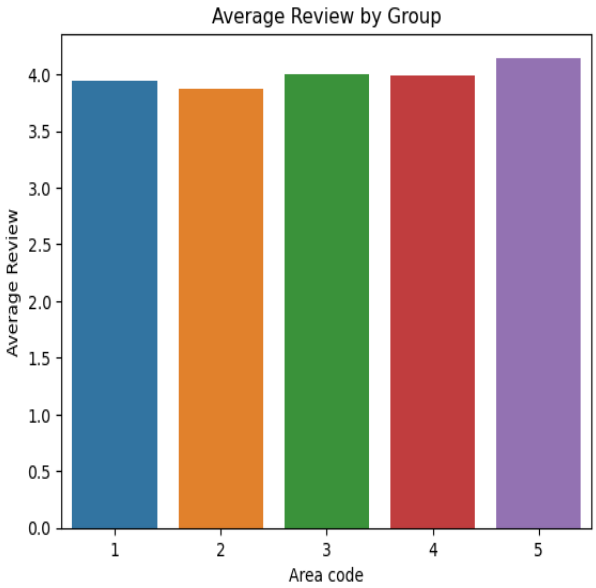

Figure 7: The bar chart of average review in five regions

due to historical reasons. Berlin was divided into 2 during the Cold War, so it is a enclave within the eastern part of Germany, which is economically more struggled than the western Germany. Thus, when we apply the data of Berlin, it may be biased.

7 Reference

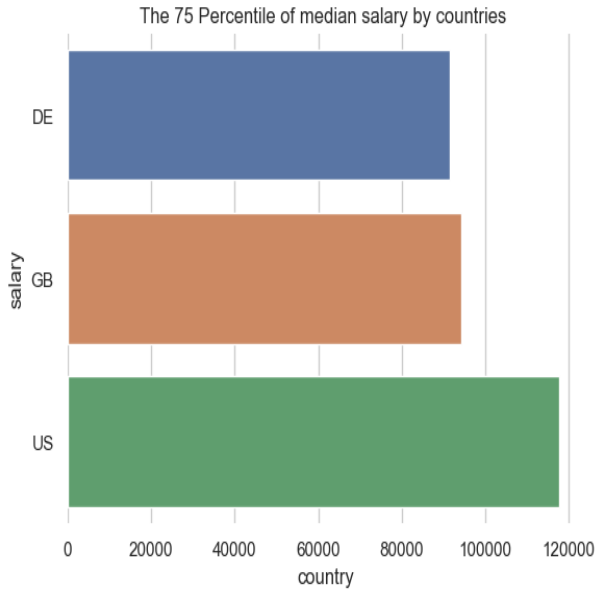

Figure 8: The bar chart of 75 percentile salary in 3 countries

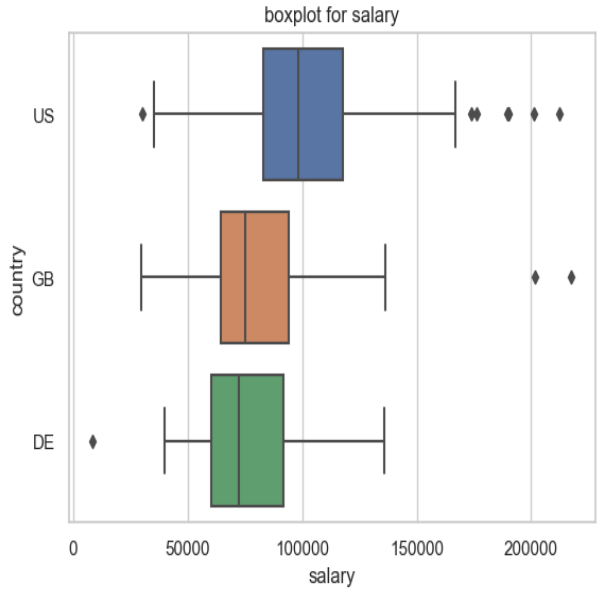

Figure 9: The boxplot of salary in 3 countries

https://www.scrapingdog.com/blog/scrape-glassdoor/

https://www.census.gov/library/visualizations/2021/dec/2020-percent-change-map.html

https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=US

https://data.worldbank.org/country/united-kingdom?view=chart

https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=DE

https://www.census.gov/library/publications/2022/demo/p60-276.html

https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/householddisposableincomeandinequality/financialyearending2022#:~:text=Median%20income%20for%20non%2Dretired,FYE%202013%20to%20FYE%202022).

https://www.destatis.de/EN/Themes/Society-Environment/Income-Consumption-Living-Conditions/Income-Receipts-Expenditure/Tables/income-expenditure-territory-lwr.html

Figure 10: The links of references