Stats 140XP – Final Report

Exploring the Association Between States & Salary

Allen Chun, Nicole Lam, Neftali Lemus, Shirley Tang, Leo Tien

March 11 2022

Table of Contents

Abstract

There are several factors that can affect one's decisions when choosing a job: aspects such as location, compensation, and company reputation among others play different roles in the decision-making process. In particular, evaluating a job's salary can be a driving factor that influences one's career choice. This paper will focus on analyzing data from Indeed, a worldwide employment site that allows users to search and apply to jobs while having access to information like job requirements, average company ratings, company industries, and job locations among others.

Problem

Our group was interested in comparing two main questions focused on the effects of job listings and estimated salary:

- Do salaries across different states in the United States differ, and is this difference significant between the average estimated salary across the top earning states?
- How do the proportion of job listings based on job categories compare among the highest and lowest average estimated salary states?

Description of Dataset & Variables

The dataset from Indeed contains around 14.5 million unique observations and consists of information about 23 variables describing job postings from various countries and industries in the form of 9 numerical and 14 non-numerical attributes.

We worked with 3 variables in particular, two predictors: stateProvince (a categorical variable listing the name of the state or province of the job posting), normTitleCategory (a categorical variable describing the occupational category of the normalized job title), and a response variable, estimatedSalary (a numerical variable of the estimated annual salary).

In order to clean the data, we first filtered for US job postings and removed all missing and unknown values from the dataset for the variables of interest. We then took a random sample of about 5% of the filtered data, which was around 600,000 observations.

Exploratory Data Analysis

For the exploratory data analysis, we first grouped our data by state to gain a better understanding of how average estimated salary differed among states. After sorting the states by average mean salary, we found that Washington D.C, New York, Massachusetts, California, Virginia, Washington, Maryland, Connecticut, Delaware, and New Jersey were the top ten states with the highest average estimated salaries as seen in Table 1.

State	Average Estimated Salary
DC	66006.99
NY	55186.40
MA	52773.91
CA	52763.50
VA	52352.02
WA	51766.29
MD	50475.05
CT	49837.99
DE	49743.37
NJ	49318.15

Table 1. Top ten highest earning states in the United States

To visualize salary trends across the country, we created a heatmap (Figure 1) of the United States and a boxplot (Figure 2) showing the top ten highest average estimated salary states. From these two graphs, we observed that jobs in states along the east and west coasts, on average, seem to have a higher estimated salary than other states.

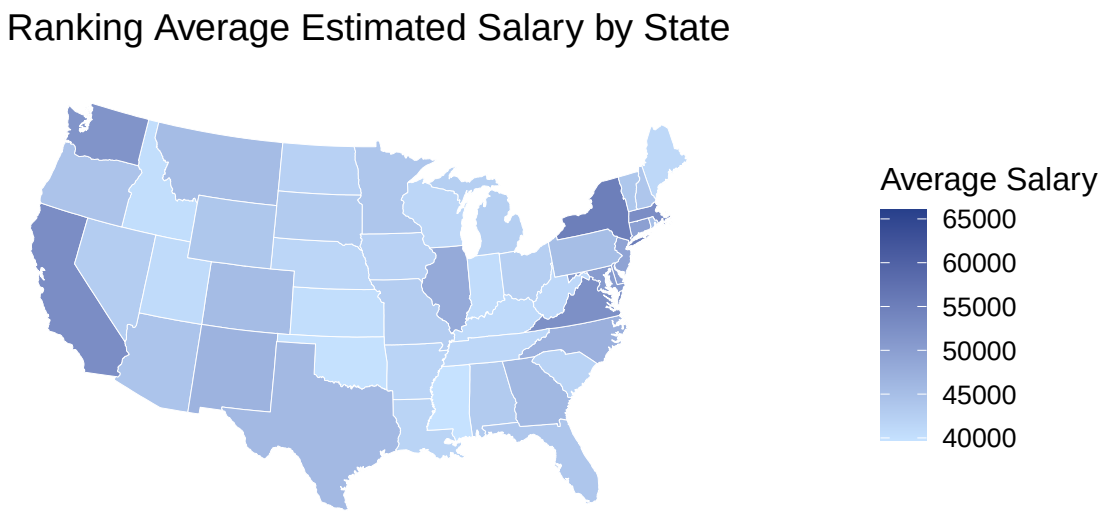

Figure 1: Heatmap of average estimated salaries

This figure shows the states with darker shades have, on average, higher estimated salaries. States such as California, Washington, New York, and Virginia have the highest average estimated salaries among all

states. Looking at the heatmap, coastal states seem to generally have a higher estimated salary than inland states.

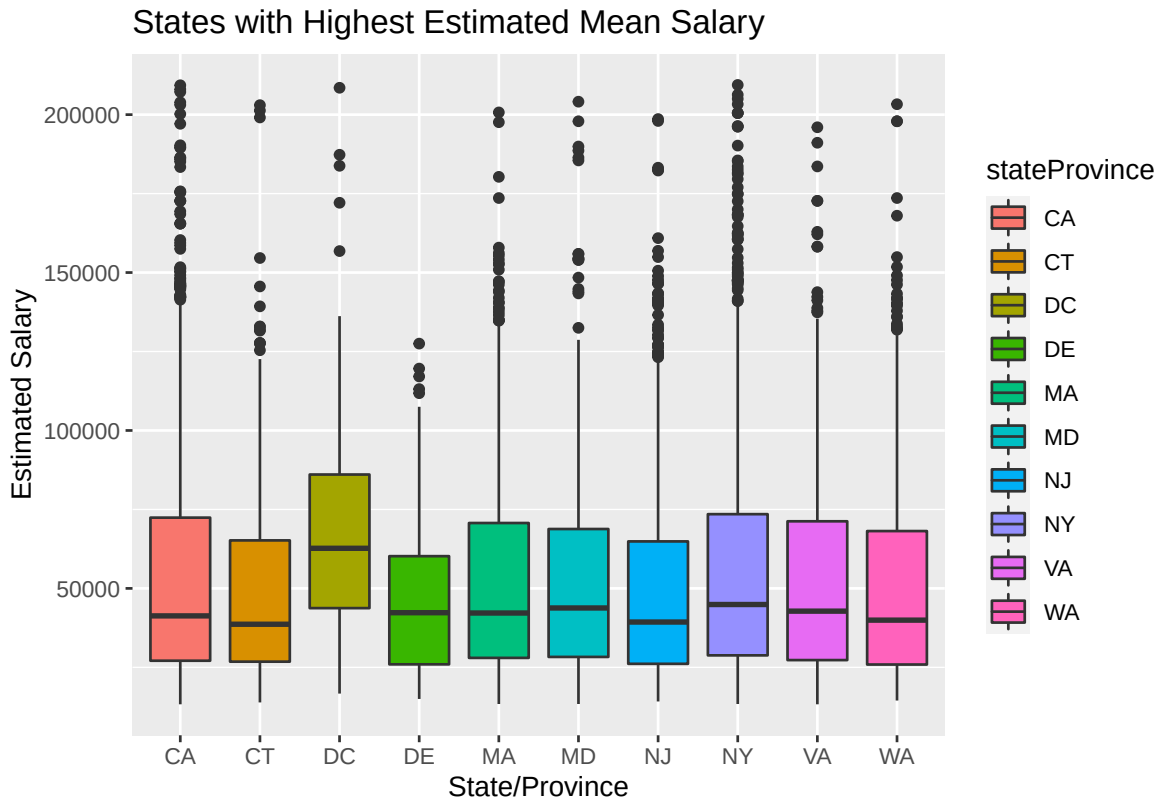

Figure 2: Boxplot of highest average salaries

This boxplot shows the top ten states with the highest average estimated salaries. In addition to the coastal states, Washington, D.C. has the highest average estimated salary among the states.

Statistical Methods Used

One-Way ANOVA

To address the first question of whether there are significant differences between the top ten highest paying states, we used a one-way ANOVA model to compare differences across multiple salary means. This model states the following hypotheses:

$$H_0 = \mu_{WashingtonD.C} = \mu_{NewYork} = \mu_{Massachusetts} = \mu_{California} = ... = \mu_{NewJersey}$$

$$H_a = \mu_i \neq \mu_j$$

In words, our null hypothesis states that there is no difference between the salary means of the top ten states, while our alternative hypothesis states that at least one of the means is different.

Tukey's HSD Test

After using the ANOVA model, we followed up with a post-hoc analysis using a TukeyHSD test to identify pairs of states where the difference in average estimated salary is significant.

For additional exploration, we sorted the job listings by category within each of the five highest earning and five lowest earning states to determine if there was a trend among which job categories appeared the most in high estimated salary states, versus low estimated salary states.

Results

After running the model, we found that the ANOVA model between estimatedSalary and stateProvince shows that stateProvince is a significant predictor for determining the estimated salary of a job, with a p-value of 2e-16. Given the small p-value at the $\alpha = 0.05$ level, we reject the null hypothesis that the average estimated salaries among states is the same.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
stateProvince	9	1.313e+11	1.459e+10	14.99	<2e-16
Residuals	20282	1.974e+13	9.733e+08		

We continued our analysis with Tukey's test for post-hoc analysis. Out of 53 comparisons, only 18 pairs turned out to have significant differences as seen in Figure 3. It is interesting to note that New York appeared to have a higher average estimated salary than many of the states.

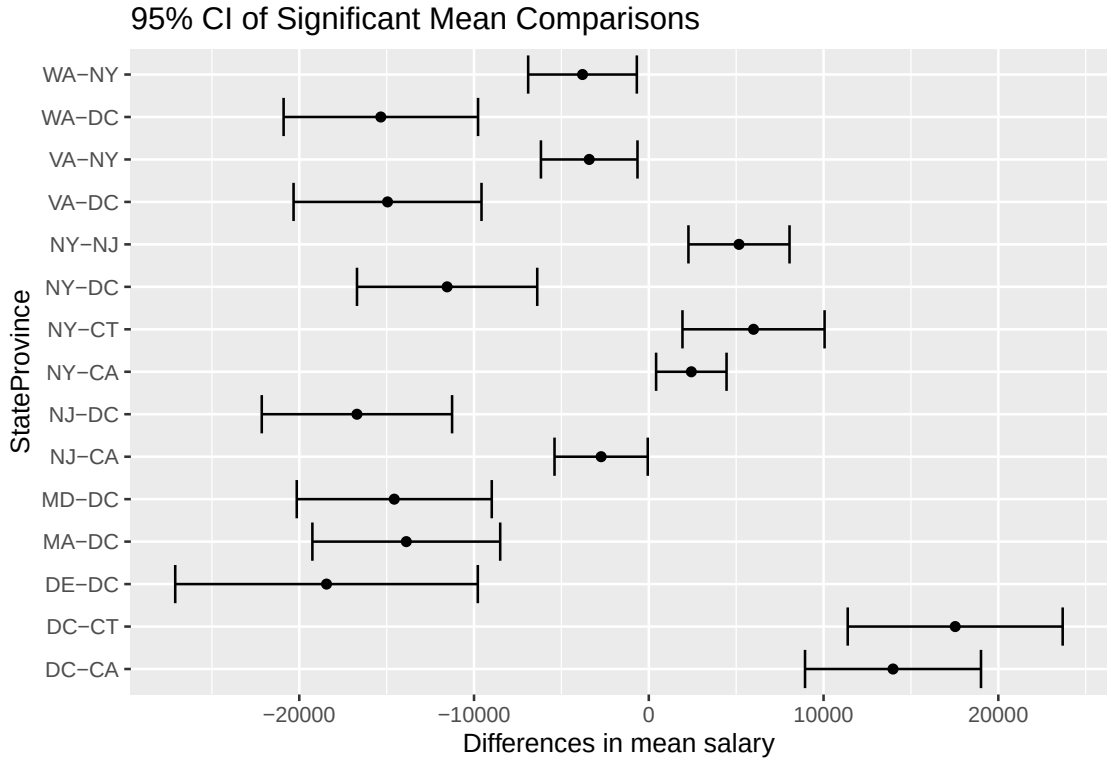

Figure 3: 95 percent confidence intervals of significant mean comparisons

We further analyzed the differences by examining the proportion of job listings based on job categories among the five highest and lowest average earning states, as shown below in Figure 4 and Figure 5.

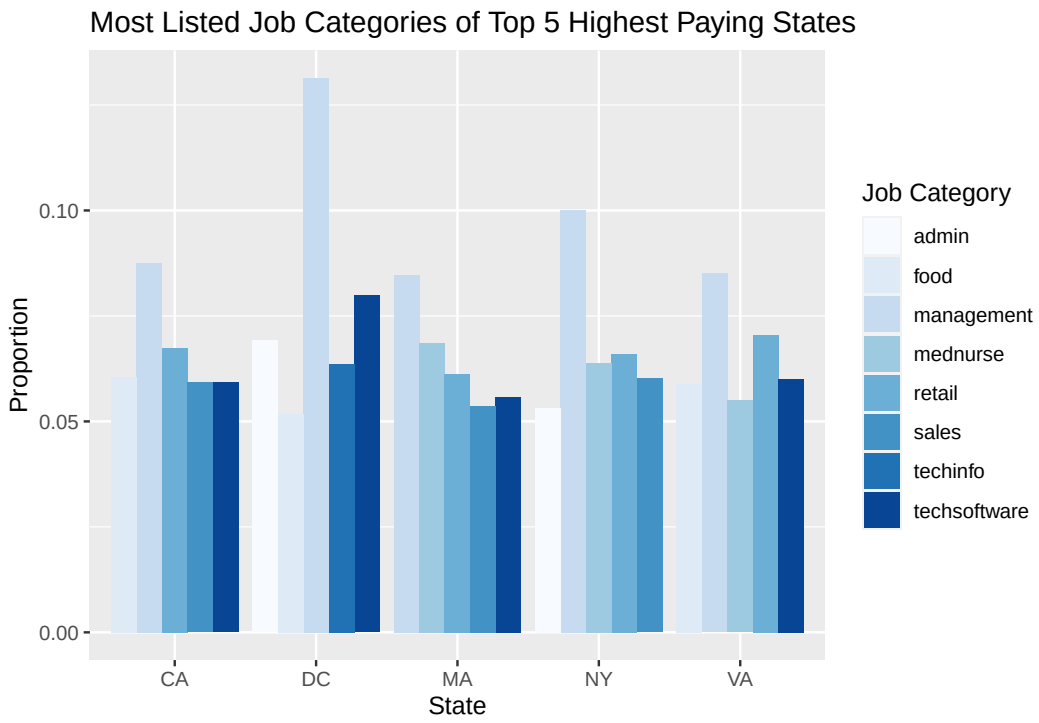

Figure 4: Highest proportion of job categories for top five highest paying states

Figure 5: Highest proportion of job categories for lowest five paying states

We found that while states with higher estimated salaries had more management and STEM-related roles, states with lower estimated salaries tended to have more jobs in the retail and service industries The management category consisted of jobs such as district manager, regional vice president, supervisor, and operations manager roles while the retail category was composed of titles like stock supervisor, cashier, and retail sales associate.

Conclusions

We concluded that the top ten states have the highest average salaries for several reasons. One reason being was that the top states tend to have more startups and opportunities in tech, government, and healthcare, whereas states in the middle of the country may be more agricultural based. As a result, the lower paying states tend to have more jobs in food and retail.

Another reason could be the cost of living of the top ten states. The majority of these states have big cities, and typically, the cost of living is higher in the big cities, resulting in a higher salary to afford the cost of living.

Limitations

Due to a large data set and limited laptop capacity, there were restrictions to the range of analyses we could conduct on the given data. Therefore, we had to take a 5% sample of the data to conduct our analysis, which affected our analysis as a whole.

It is also important to note that there were missing values in our variables of interest (estimated salary, job category), and the removal of these missing values could have affected our conclusions as well.

References

American Statistical Association (2018). *Indeed Data*.