Analyzing Industry & Occupation Insights through the American Community Survey

BSAN 6300

Dr. Tang

May 6, 2021

Victoria Capobianco

John Lenhe

Karen Tafa

1. Executive Summary

This research project explores trends and insights regarding the current professional occupation & industry landscape through use of data cleaning and wrangling, correlation analysis and data visualizations. Research questions that guided the project include: Is a certain industry male or female dominated?”,” Are there certain education levels tied to specific industries?”, and, “What is the relationship between race and the professional landscape?” By answering these questions, our group found patterns related to social sentiments surrounding the gender wage gap, education accessibility and the racial and wealth inequality.

We began by collecting data from the American Community Survey regarding Educational Attainment, Income, and Industry as our main variables, along with demographic information such as gender and race. Our research was narrowed to California’s zip codes and county. The data was cleaned and wrangled using Excel functions such as Power Query and VLOOKUP.

Once the data was collected, we created visualizations using Tableau. Our visuals range from bar charts, maps, line charts, and bubble charts that illustrate the relationship Income by Race, Income by Percent of Population, Median Earning by Occupation/Industry/Gender, and Educational Attainment. Additionally, a correlation analysis was ran in Python between median earnings and educational attainment.

Our results depicted a concentrated pattern within the Bay Area and LA, Income differences due to education attainments, female earnings less than male, and discrepancy in income by race.

We hope this report helps other graduating students understand the professional landscape they are entering with more knowledge, as well as provide hiring managers with insights on how to make their workplace and industry a more equitable environment.

1. Introduction

As a research team, we are preparing to enter the real world post-graduation and pursue our professional goals outside of academia. The uncertainty of the professional world, combined with social sentiments surrounding income and gender inequality in corporate industries formed the basis for our research ideas.

Our aim is to answer the question of which industries are dominated by certain demographics of people? To fully explore this question we will ask ourselves a series of sub-questions that can allow us to dig deeper into certain topics. Questions like: “Is a certain industry male or female dominated?”,” Are there certain education levels tied to specific industries?”, and, “What is the relationship between race and the professional landscape?” tailors our research to how demographic information plays a role in the hiring process. Through answering these questions, we seek to provide stronger insights into an evolving work environment and measure the degree of inclusivity and equality in this modern world. The key piece of relevance for answering this question is to further understand societies push for more equal hiring practices and locating where change still needs to take place. Finally, answering this research question will aid us as students seeking careers after graduation, as we will be better equipped knowing demographically the type job market we will be stepping into.

The data used to answer our research questions comes from the American Community Survey (ACS). ACS is a detailed, annual survey organized by the US Census Bureau which covers topics such as educational attainment, employment status, income, household information, technology ownership, and much more. The purpose of the ACS is for local and public officials to learn more about changes taking place in their community to better understand and plan for the present and future.

1. Data

Due to the large amount of topics covered by the ACS, we decided to begin our search with ACS’s 2019 Subject Definitions page to understand what kinds of variables and features were available to us through this database. Because our research questions narrowly focus on analyzing the current industry and occupational landscape, we found the appropriate variables to study in the Population Variables section. We chose variables such as: educational attainment, income, industry, occupation, race, and sex. Educational attainment was asked of all respondents and the data is tabulated for individuals ages 18 and over. This variable records the highest degree or highest level of school completed by the individual. Income was asked of respondents ages 16 and older and refers to the sum of all wages and salary. Industry data was asked of respondents age 16 years and older through a series of write-in and check-box questions. This variable illustrates the kind of business conducted by a person’s employment organization. Occupation was asked of all people age 16 years and older and refers to the kind of work the respondent does on the job. Race data was collected from all respondents through participants’ self-identification of race in a social lense (as opposed to a biological or anthropological lense). ACS recognizes that race data includes identification with racial or national origin or sociocultural groups. Respondents were prompted to choose more than one race to indicate their racial mixture. Finally, sex was asked of all respondents to mark either male or female as an indicator of biological sex. Once we solidified our variables we began to look for table shells and lists that included these variables so we could begin to derive our data. Among each variable, several tables were created in the ACS database; therefore, in order to narrow our search even further we decided to focus on three key variables: income, educational attainment, and industry data and use the remaining variables to describe the demographic makeup of each of these categories.

For our first key variable, income, we found several tables reporting median, mean, household and per capita income. We settled on set B19301: Per Capita Income in the Past 12 Months in order to get a snapshot of income at the individual level within a particular year. This dataset contained individual tables for each year and racial group including White, Black, Hispanic, Asian, Hawaiian Native/Pacific Islander, American Indian/Alaskan Native, and Other. After evaluating the contents of the tables, we decided it would be best for our analysis to compare the change in income over time among these different racial groups. Therefore, data tables from the years 2015-2019 were downloaded for each race among each county in California so we could assess any changes over the 5 year period.

Once all the data tables were downloaded, the challenge was to combine all 35 individual tables together. Each table had the same four columns upon initial download: county id, county name, per capita income over the past 12 months, and margin of error. We removed county id and margin of error from the first table (Per Capita Income in 2015 White Only) because they were not necessary for our analysis and used this table as our base table to join the remaining 34 tables on to. Using the VLOOKUP function in Excel, we joined each table on county name and created a column for Income by each race within each year (2015-2019). This posed a challenge because immediately we noticed discrepancies in the number of counties reported. The tables reporting White Income only contained 42 California counties, while the others reported all 58 counties. We had to compare this table with our others to determine which counties were missing. From the database, it is unclear as to why the White population does not list every county in California. This led to several blank values within the White Income columns.

Continuing with tables displaying data regarding income, we also made sure to look at income from a gender and industry perspective. Another table acquired from the ACS website was a file that showcased median earnings by industry and by gender. This table was broken down by zip code in California, allowing us to compare earnings data by various regions in the state. The primary elements of cleaning that needed to be done in order to make the table ideal for analysis was to decide the necessary columns, clean up headers and label names, as well as understand how null values should be treated in the table. The columns used in this table are median earnings per industry, median earnings per industry for men, median earnings per industry for women, and women’s earnings as a percentage of men’s earnings per industry. Other columns such as the margin of errors and columns that were used as separators for the sections were removed in order to keep the data clean and only showcase pertinent information. From there we cleaned the labels and names so that they would be more easily used in our analysis as well as simpler to understand for the everyday user. Finally, we were confronted with many null values throughout the dataset. This is likely due to not enough people from a zip code responding to a certain category of the community survey. We decided to leave these values as null because changing them would be misleading. Having a null value become zero would imply that the zip code has earnings of zero and would therefore throw off our analysis and potentially showcase false information. Overall, these were the most significant pieces of data cleaning that needed to be done for the median earnings by industry and gender dataset adding onto some smaller cleaning elements like ensuring data types were correct and removing extra characters, like the letters that came before the numeric values of the zip codes, that did not belong in data fields.

The next table that was utilized in our analysis from the ACS website was a table that explores the industry and occupation metrics of various California zip codes. The table first outlines the number of people in each zip code by the industry they work in. Not everyone in the state of California is represented in the population as this data is limited to those who responded to the ACS study. After the population by industry is outlined, the table then breaks down the percentage of the population that works in certain occupations in those industries. Some of these occupations include management, business, science, arts, service occupations, sales occupations, construction, maintenance, agriculture, production, transportation, and hauling. Through looking at this data we can now see which zip codes and communities in California will hold certain occupations in certain industries. The data cleaning process for this table echoed a similar set of steps that needed to be done for the datasets explained earlier in this report. The margin of error and other unnecessary columns were removed, column names and data labels were changed for easier use, and other various cleaning elements like ensuring data types were correct as well as removing extra characters from numeric values was also done. Once the data was thoroughly cleaned we were left with a very usable data set that would allow us to analyze the industry and occupation metrics of various zip codes in the state of California.

The last table we needed to attain was one that would explain the relationship between education attainment and key demographic variables such as race and age in different zip codes around California. Although this may have seemed simple enough, we actually really battled with getting the right data. We found a lot of tables that had data that had already been aggregated and summarized. There were also a lot of websites that had created visualizations from that data that we would need, but it was nowhere to be found. After countless hours of downloading and trying to figure out how we were going to get that data, we finally decided to make use of the filter feature in the census.gov website. Although this seemed easy enough, the zip code was not directly available to us so we had to use the ZIP Code Tabulation Areas (ZCTAs) column then later use excel to extract just the zip codes from this column and leave out the string characters. We also removed the null values here.

Once we had the data, we had to make sure it was clean. This involved first making sure we had the right columns needed. There were a lot of unnecessary columns that included calculations that we would not be making use of, such as “margin of error” for each age group and each race. There were also 769 columns and this just made it clear that a lot of them needed to be assessed and gotten rid of if needed. The reason for these being so plentiful was because they were very repetitive for every race, which made it slightly easier to clean when creating code in Python. We did most of the cleaning in Python, we first dropped the unnecessary columns and that was pretty easy. We then cleaned up the column names using the replace method on the dataframe which made everything so much easier to read and understand. Now the data was ready to go.

1. Method

Once all of our data had been collected and wrangled, we loaded the data into Tableau to create visualizations and ran a correlation analysis against two of our variables to gain an in-depth understanding of patterns present within the data. Utilizing Tableau was the best method for our purpose because it allowed us to organize our large amount of data in a succinct manner and produce interactive visualizations that would be easily interpreted by stakeholders such as local officials, hiring managers, and government officials.

For the Income by Race data, the goal was to analyze any significant trends (either increases or decreases in income) over the past 5 years. The data was analyzed both at the county level, and as a view of California overall to produce eight different visualizations. The first visual is a snapshot of Average Income by Race based on all of California over the past 5 years in a simple and readable line chart. Each race has an overall upward trend. However, White Income rapidly decreased with a quick increase between 2016-2018. Following the line chart, a bar chart was created for each race that could be filtered at the county level. These in-depth bar charts are more specific as they allow the viewer to view side-by-side each population’s income year after year. Finally, after the bar charts, a bubble chart displays the Percent of Population in California for each racial group and a color to represent their average income in 2019. This is another high-level overview to see if income is distributed relative to the percent of the population. Altogether, these eight graphs combine more granular and specific charts with less granular and general overviews of income data by racial populations throughout 2015-2019.

With the understanding of how race plays a role in how income impacts the population of various counties in the state of California, we then looked to break down how different occupations and industries were distributed across the state. The data cleaned from the industry and occupations table, as mentioned earlier in this report, contained data that showcased the population distribution of zip codes across California. Creating a map with this data allowed us to fully represent the distribution of the working population, ages 16 and over, in these zip codes and fully understand where large quantities of the working population resided. Based on our results we could confirm that the vast majority of the working population resides in the Los Angeles and the San Francisco areas. Other hotspots for large population densities with high working populations were some pockets in central California. After mapping the working population, the next visualization to create from this data centered around the occupations and industries that the California population worked in. Again, through using Tableau we could map out the zip codes where high populations of people worked in certain occupations and layer on top of the map the density of people working those occupations in specific industries. Using the dashboard feature in Tableau, we mapped the percentage of the population that worked in each of the industries covered in the ACS. The map would show zip codes in green representing a high percentage of the population working in those industries or red signifying a low percentage of the population. A second layer was added to the map so as to showcase the relationship between the specific industries and the occupations within those industries. The industries were represented by using a filtered drop down menu in order to select the desired industry to look at. Once the industry is selected the percentage of the population working that occupation in the specific industry will appear as a blue circle, so as to stand out from the blue green base layer of the map. The circle will grow or shrink based on the magnitude of the population working in that industry. Through these visualizations we are able to point out trends with occupations and their respective industries across the state of California, picking out where one industry and occupation may be more dominant or even if there are industries and occupations that are widely worked in across the state.

The next set of data we wanted to analyze in order to gain a full picture of how demographics impact the work environment, was to look at the median earnings for different genders across various industries. To accomplish this analysis, we created Tableau visualizations that mapped out the distribution of earnings for each gender across the state of California, similar to what was done to visualize the industry and occupation levels. Looking at the final visualizations one can see the difference between the median earnings level of men versus the median earnings of women. Aside from some small pockets in the Los Angeles and San Francisco/Bay Area, male median earnings are vastly greater than those of female earnings across the state. Furthermore, this message is developed on the fourth map on this dashboard that showcases women’s earnings as a percentage of men’s. This graph showcases how a vast majority of women’s earnings across all industries are less than 100 percent, meaning women are being paid a fraction of the amount men are. Though the pay gap between women and men has been explored and brought to light in recent times, these visualizations highlight very clearly where that gap resides and how far equal pay practices still need to go. To continue our understanding of earnings in the population of California, we then decided to look at education levels and see if we could draw a correlation between education level and median earnings. To do this, we used Python to join two of our tables together using zip codes as the key and thereby allowing for the direct comparison of education level and median earnings. From there we appropriately cleaned the joined table so as to allow for a concise and organized output, and then ran our correlation matrix directly in the coding program. Our findings were as we would have expected, as one’s education level rises the correlation between education level and median earnings becomes more positively correlated. Overall, through looking at these maps and the correlation matrix, we can fully understand the relationship and impact that median earnings has with demographic features such as gender, industry, and education level within the state of California.

Finally, we wanted to look at education attainment by age and race. First, we had to assess the total number of people with a bachelor’s degree and how they were distributed around California. We created a map visualization in Tableau in order to do this. The map displayed the zip codes around California and how people of all ages with a Bachelor’s degree were distributed. In order to successfully achieve this without created maps for each age group, we had to create a calculated field that would combine all these groups together to get an overall picture. Although it was not completely clear, you could see that most of the people with a degree are situated in San Francisco and the bay area in general, and around Los Angeles. There were some people that were spread out in the middle of California too. We could assume that this makes sense for obvious reasons, for example, big cities attract a lot of people. They also attract a lot of talented individuals so this might be the reason why people with a high education end up in these big cities. The jobs are also often better and higher-paying.

We then created a similar map to visualize how people with less than a high school diploma were distributed. These people appeared to be situated more in the middle of the state and we assumed that this might be because of how expensive it is to live in the big cities. Or the fact that the people that are without a high school diploma can be without a home, therefore they may not be included in the statistics.

Next, we wanted to see the top ZIP Codes with a bachelor’s degree. Here we chose to see the top five, which could be easily changed by changing the choose value filter on the worksheet. We made use of the calculated field and created a treemap to display these zip codes. In these top five ZIP Codes, we found out that the median income was ranging from US$122,000 to US$145,000. Once again this makes sense as it falls into the assumption that more educated people tend to get paid higher. Furthermore, areas with these people will have a higher median income. However, we were very interested to see how race plays into this topic. So, we looked into the distribution of races in these ZIP Codes. We created a bar chart and we noticed that the white and asian populations were dominating these areas, while the black, native, and hispanic populations were extremely low here. It is important to note that this was based on the sum of each population, so to get a more clear picture, we created new calculated fields to show the proportion of each race in comparison to their population in each zip code. And now we saw a different picture. Now that we had accounted for the fact that the others tend to be minorities in the country, we could see that there is a higher black, native, and hispanic population in these areas as percentage-wise.

We ran the same analysis for high school graduates. These are people whose highest education level of education is high school. Once again, we looked up the median income range and it was between 68,000 US$ and 112,000 US$. As you can see this is significantly lower than those with a bachelor’s degree, although still good. We looked at the race again, and this time the hispanic population and the white population are dominating, and the asian population seems to be much lower as compared to the white and hispanic. Again, the black and native populations remained really low. Once we accounted for the different country populations again, we saw a much more equal picture that shows that in general, all races in these areas have received at least a high school education, which made more sense.

Lastly, we did the same exact process, for the least educated people. These are people with less than a high school diploma. We looked at the medium income range once again, and this time it was ranging from US$40,000-US$65,000. This is a big difference from the previous one. At this point, One might’ve already noticed that education attainment makes a big difference just by looking at the different income levels. Another interesting observation was that in these areas where they are the least educated people, the asian community almost completely vanished. It is now on the same level if not much lower than the black and native population. However, the hispanic population was leading in all the zip codes. The white population was also very present in these zip codes still, which was a very interesting thing to see. We weren't able to create the same proportion graph as the previous two as the data did not have data that showed races with less than a high school degree.

1. Results & Implications

Comparing the patterns across our three categories of Education, Income, and Industry, it is evident that several trends reflect social sentiments in today's climate such as the gender wage gap, lack of educational resources, and racial wealth inequality.

Beginning with the Income by Race data, it was not surprising that White individuals have the highest reported income above all others. Despite the dip in White income from 2016-2017, the Average Income Across 5 Years line chart illustrates this group as having the highest income at an average $43K, followed by Asian at $40K, and the remaining populations between $20-30K by 2019. Across all groups, it appears wealth is centered around the Bay Area in counties including San Francisco, Alameda, San Mateo, Marin, and Santa Clara. However, each racial group had its own individual outliers. For example, Black income skyrockets in Siskiyou County in 2018 and 2019. Siskiyou County is home to Mt. Shasta and several federally owned parks. A large amount of government-funded projects and parks in this area could be a reason for an influx in income. Another outlier is the American Indian Alaskan Native group has income (around $60K) in the early years and then significantly drops by more than half in 2019. Sierra County is near Lake Tahoe and only has one official city so most of the land remains untouched. This could potentially explain why indigenous people thrive here as they honor their ancestors and heritage on untouched land. Additionally, as seen in the Percent of a Population bubble chart, it is evident that income is not distributed relative to the percent of the population by racial group. Looking at the chart, the Hispanic population is the second-largest in California, yet, their per capita income is less than half of the Asian population and $9,000 less than the Black population, both of which have a smaller percentage of the population compared to Hispanic people. Overall, the insights derived from these charts are a great jumping-off point for a further investigation surrounding racial and wealth inequality.

This wealth inequality is further explored in the median earnings and industry visualizations where one can clearly delineate the gap in male and female earnings across industries. Looking at the maps created in Tableau, one can see how large swaths of California are greener in the male earnings map compared to the female earnings map. Showcasing how median male earnings in the state are much greater than those of female earnings. To further explore this gap in earnings, one can look towards the women’s earnings as a percentage of men’s map. There we can draw the conclusion that for a majority of the state of California women are only earning a fraction of what men are. The good news is that for much of the state women’s earnings are getting closer to men’s, but much progress still needs to be made in this area. When overlaying specific industries onto these maps, the trend continues with a larger number of large circles appearing on the male map with fewer large circles appearing on the female map. One again showcasing the differential in earnings between the two genders. Overall, through analyzing these maps we can affirm the conclusion that men in the state of California are in fact earning more than women, despite the growing awareness of this issue in recent times. We hope that should this analysis be run several years in the future that the gap between men’s and women’s earnings shown in our analysis will be significantly less than it is now and much closer to true equality.

Adding the education attainment on top of all that has been mentioned above, we can see that in general, the higher the education level, the higher incomes. However, one must start to wonder if there are other factors that may prevent certain people from attaining this higher level of education. We also noticed that zip codes with highly educated people had higher incomes and were limited to certain races. This could possibly encourage more research into ways of making these zip codes more evenly distributed in terms of race. This might include figuring out if there is any way to improve education attainment among these races.

All in all, analyzing these datasets and exploring this topic was highly beneficial to us as upcoming college graduates and we believe that many people like us could benefit from it. These insights could also be used to provide solutions to inequality in certain areas in California.

1. References

American Community Survey Table Links:

<https://data.census.gov/cedsci/table?q=United%20States&t=Business%20and%20Economy%3ARace%20and%20Ethnicity&g=0400000US06&tid=ABSCS2018.AB1800CSA02&hidePreview=false>

<https://data.census.gov/cedsci/table?t=Occupation%3ARace%20and%20Ethnicity&g=0400000US06.860000&tid=ACSDT5Y2019.C24010A>

<https://data.census.gov/cedsci/table?t=Occupation&g=0400000US06.860000&tid=ACSST5Y2019.S2405&hidePreview=false>

<https://data.census.gov/cedsci/table?q=California&t=Age%20and%20Sex%3AClass%20of%20Worker%3AEducational%20Attainment%3AEmployment%3AIncome%20and%20Poverty%3AIndustry%3AOccupation%3ARace%20and%20Ethnicity&g=0400000US06.050000&y=2019&tid=ACSSPP1Y2019.S0201&moe=false&tp=false&hidePreview=true>