Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021

Project Part I: Descriptive Analytics Data Mining

1. Executive Summary

The data selection process sought to answer the following question: “Where are auditors and accountants paid the most? Due to the nature of the data, the group supposed it could be available via a government organization such as the IRS, Census Bureau, or other similar agencies. The supposition proved true although, locating a robust easily accessible data set proved challenging. Ultimately the easiest and cleanest data set was retrieved from the Department of Labor’s Bureau of Labor Statistics (BLS), which stores employment data across a wide spectrum of occupations and geographic regions. The first step in gathering the data was to study the information available through each prospective source. While some had more robust information via formal Freedom of Information Act Requests they proved too time-consuming to pursue. Conversely, information via the Census Bureau was easily and readily accessible but not robust enough to differentiate household income by occupation across large geographic regions. The BLS provided a data set of over 500 distinct geographic regions which could be further filtered by occupation and filtered once more to view salary.

Given this information, the dataset is comprised of observations identified by metropolitan statistical areas, counties, cities, and rural areas. The variables of interest are mean annual wage and average household price for a 3-bedroom home in the respective area. The values represent fixed determinations from May 2020. Additional variables were created or omitted to provide further information such as income-to-price ratios and income quartiles.

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021

2. Preprocessing and Cleaning the Data

As previously mentioned after identifying the preferred source of information a deeper review of the source was carried out. Data pages, definitions, and publications provided through the BLS were analyzed to determine the best method for gathering data. Ease of access was a determining factor as the BLS provided a native data query page from which to extract curated information. Wage information was extracted along with corresponding geographic areas ranging from rural border towns to large metropolitan areas comprised of multi-city regions. Due to the variety of geographic areas, it became necessary to identify an additional variable as a benchmark of sorts for the wages associated with a given region. Housing data from Zillow was extracted via their site which provided dollar value historical cost for approximately 95% of the observations pulled from the BLS. Using search, filter, and other excel tools we matched BLS regions with their respective housing data extracted from Zillow. During this process, it became apparent that two issues existed in the data. First, Alaska and Hawaii had very inconsistent data due to their remoteness and size, respectively. Thus, non-metropolitan areas of Alaska often had Null or Naan values due to insufficient data. Likewise, non-metropolitan wage information for Hawaii was heavily skewed even though these areas remained within single-digit miles of the metropolitan areas. Thus, only data for the contiguous United States is provided in the data. Additionally, similar issues presented themselves with extremely remote areas such as in Montana or Texas where at times no home sales data existed. These extremely remote areas often comprised larger rural areas where some data was available. Thus, those sub-level areas were often omitted to prevent skewing. Every effort was made to create a representative mean housing value of a region by incorporating as many subordinate mean housing values as possible into the

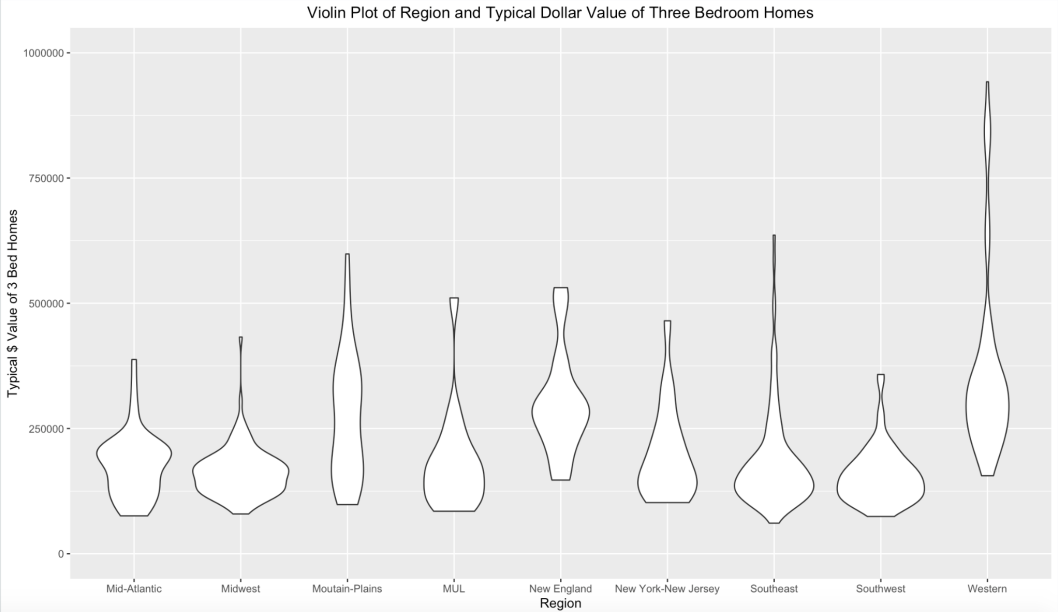
higher-level mean calculation. Additionally, minor formatting issues were addressed to improve importability to R studio such as formatting characters to numbers when applicable, removing unnecessary cells, deleting observations due to lack of data, and omitting but retaining certain variables.

As a result of these efforts, the final form of the data included variables related to wage, the number of individual observations within an observed area, concentration, housing price, and ratios related to housing and wage. This data was representative of the contiguous 48 states of the U.S. across a variety of areas such that large cities and rural regions are equally observed and compared by way of ratios.

3. Exploratory Analysis:

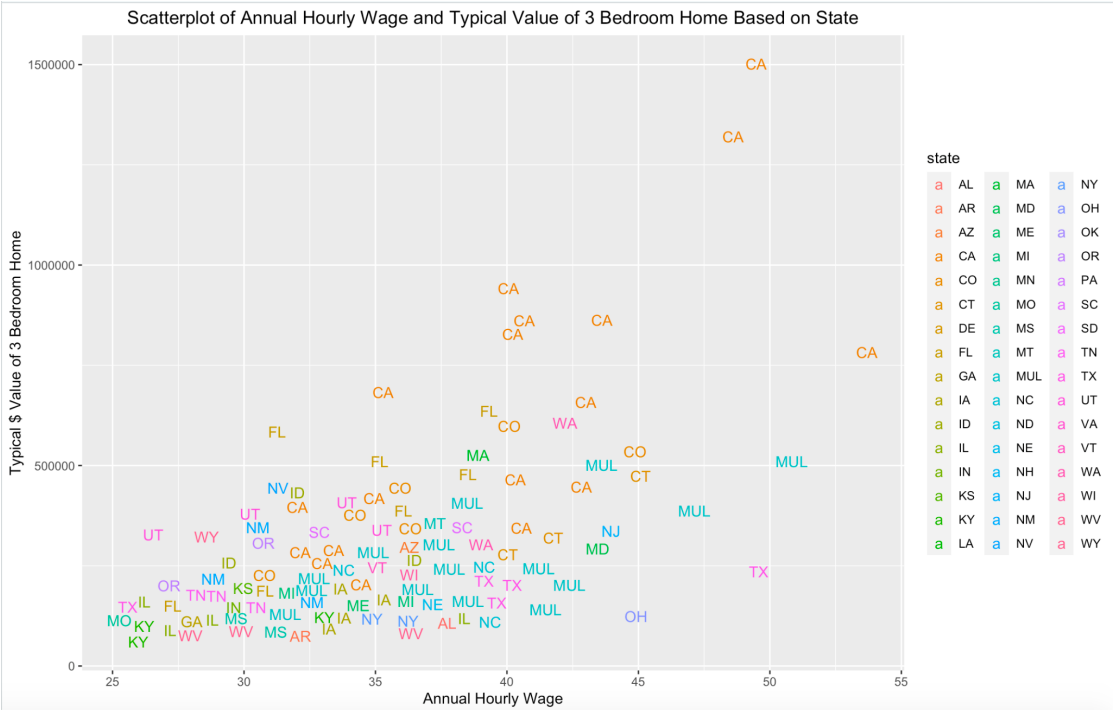
The first graphic we created was a scatterplot of hourly mean wage and employment per 1,000 jobs. We created this scatter plot to see the relationship between hourly wages and the number of accounting/auditing jobs out of 1,000 jobs in the area. We created this graphic to see if there was

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021
a relationship between the proportion of jobs in an area that were accounting jobs and what the hourly mean wage was in the area. We added a smooth geom to make it easier to see the relationship between the two variables. As is clear in the graph, there appears to be a positive relationship between these two variables - when the number of accounting jobs out of 1,000 jobs in the area increases, the average hourly wage for accountants also increases. Thus showing us that the areas with a greater proportion of accounting jobs tend to have a greater average hourly salary for accountants.



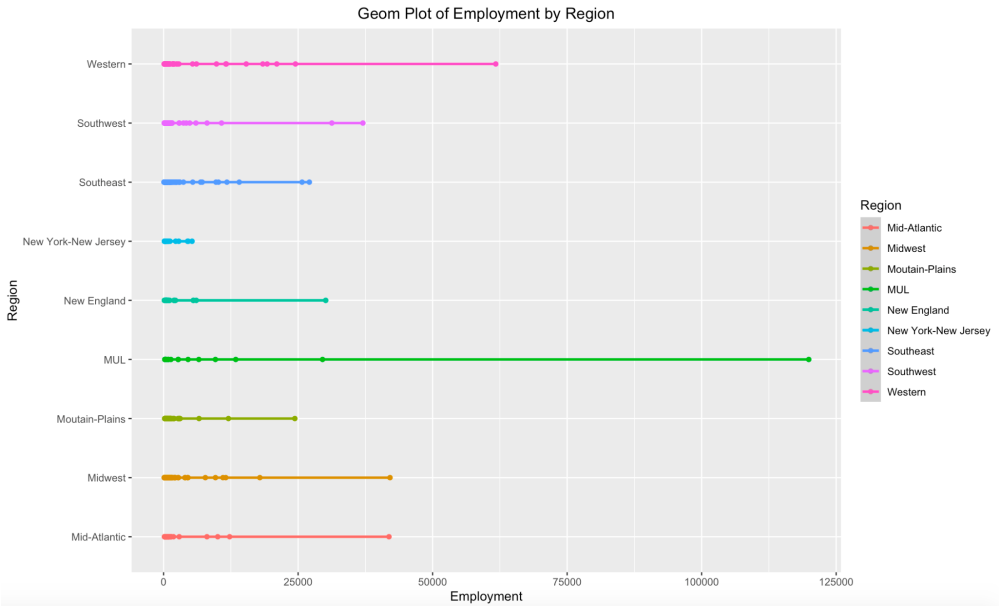
The next plot we created was a violin plot of the typical dollar value of three-bedroom homes by region. We created a violin plot to be able to visualize the number of modes per region and to understand the skew of home prices based on region. One thing to note about this visualization is that the western typical home value line continued further up into the graphic to about 1,500,000 because the maximum home value in the western region was significantly higher than the other regions. We chose to set the y axis max at 1,000,000 to be able to better visualize the mode frequency and distribution of the other regions, which were hard to see when

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021
the max was at 1,500,000. This graph shows us that the minimum home value is similar for all the regions other than New England and the Western region which have higher minimum home values. The Mid-Atlantic region seems to have one clear mode below 250,000, with a greater frequency of homes costing less than the mode, and fewer homes costing more. The midwest region has a smaller mode than the mid-Atlantic region and also has a thicker lower tail than upper tail meaning that there is a higher frequency of cheaper homes than expensive homes in comparison to the midwest home value mode.



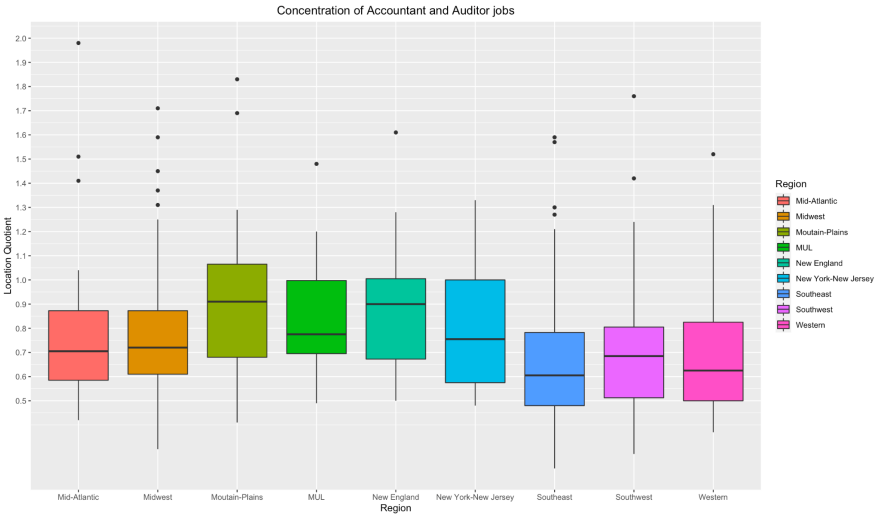
The third graphic we created was a scatterplot of annual hourly wage and the typical value of 3 bedroom home based on states. This scatterplot plots each observation and allows us to see the relationship between the annual hourly wage and the typical dollar value of three-bedroom homes. As we can see there appears to be a slight positive relationship between the two variables, meaning that areas with higher typical home values of three-bedroom houses

have higher annual wages for accountants and auditors. By coloring by state, it allows us to see that the outlier data points with higher typical home values are all California data plots. These areas also tend to all have annual hourly wages of greater than 40 dollars. All the areas with an annual hourly wage of under 30 dollars an hour also have the typical home value of under 500,000 dollars as well. The areas with an annual wage of over 50 dollars an hour all have typical home values of over 500,000 dollars.



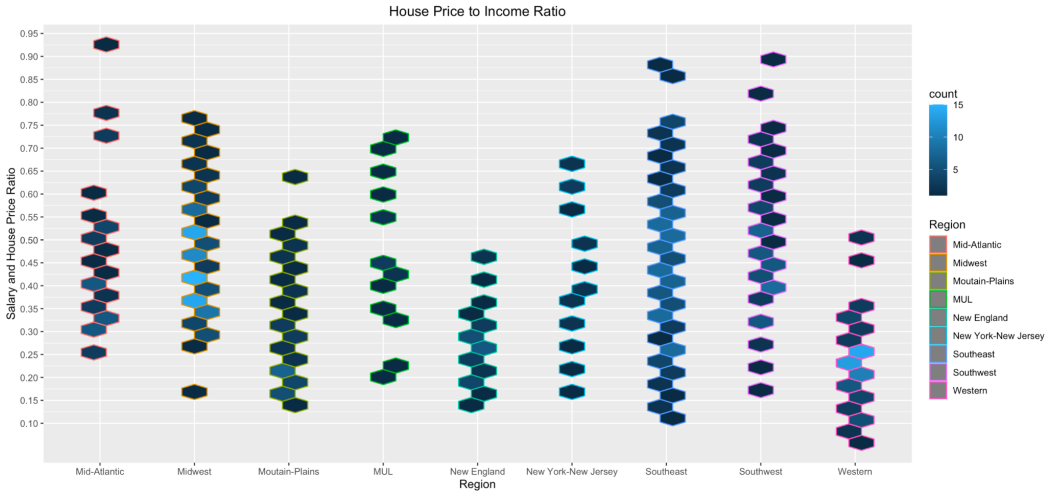
In this point range plot, we decided to see how much employment there is per region. Each dot represents an area within each region and the x-axis displays the amount of employment. The line connects all the dots so that it can display better how much employment of accountants/auditors each region has. A relationship we noticed was that each region has a great **concentration** of areas that in total have roughly 6,000 jobs in total. It was no surprise that the Western region had the greatest amount of areas due to the fact that California is a huge state. I would say that the relationships are pretty strong because the data is distributed fairly well (not counting “MUL”) among the regions.

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021



Our fifth visual is a boxplot. We decided to compare the location quotient which quantifies the concentration of accountants/auditors per region compared to the United States. With the location quotient, we decided to compare it among regions. Each boxplot displays the median, lower and upper quartiles, upper and lower extreme values, and the outliers. A relationship we can see is that all of the regions' median is between 0.6 and 0.9. A finding I noticed is that the two largest concentrations of accountants/auditors are in the Mountain-Plains and New York-New Jersey regions. This may be because of the numerous industries there are and so there may be high concentrations of accountants/auditors with the regions. The relationship between regions is not as strong due to the numerous outliers that can throw off the data.

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021



Our last visual is a hexagon chart. We used it to compare the salary to house price ratio between each region. Each hexagon is shaded with the number of areas that fall within each ratio. The chart shows the ratio of the average US house price to the average annual income of an accountant/auditor. The obvious relationship I see between the regions is that they do not offer many places where accountants/auditors can afford to purchase a home (most are shaded dark blue). It is not a surprise that the Western and New England regions are the ones with the smaller range of ratios due to their overwhelming house prices. Unlike the southeast region, there are more options. I would say that the relationship looks pretty strong because the data is well distributed among the regions. There seem to be more ranges in ratios in regions where it is less expensive compared to the regions where house prices tend to be higher.

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021

4. Conclusion

From this analysis, we learned about the relationships between location, home value, and both hourly and annual mean wages for accountants/auditors. Based on our above visualizations we can conclude that hourly mean wage and employment per 1,000 people are positively related, the mode and spread of home values vary by region, annual hourly wage and typical home value are slightly positively related, regions have similar employment concentrations with around 6,000 jobs, all regions have a median location quotient within the range of 0.6 to 0.9, and there are greater ranges in salary and home price rations in areas where it is less expensive.

Some lessons we learned along the way are that you must choose wisely what two variables to compare and which table suits them because you will get an error in your code if you don't take into account the type of variable you are analyzing. Drawing out the plots before writing them into RStudio is a helpful way to visualize what you are trying to depict with your data to make sure your variables can help you achieve such visualizations. We also learned of the immense quantity of helpful coding information online. Through the use of online tools, we were able to learn about new ggplot2 plots that we didn't learn in class which allowed us to expand our knowledge of the ggplot2 package and have a greater understanding of how to create unique visualizations. We also found it helpful to comment out our code using hashtags to organize our different visualizations and keep track of what the goal of each visualization was. It is much easier for an external viewer to understand our code if it is properly commented out, thus allowing anyone to understand the goals of our different visualizations as opposed to having to figure it out themselves. We also learned the benefits of using color in our visualizations. Using

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021
color allowed us to better understand the visuals and learn more complex things about the different data points. Creating the many different visualizations we experimented with in class allowed us to familiarize ourselves with the dataset and better understand the different relationships present between variables that we otherwise wouldn't have been able to understand without the help of visualizations.