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

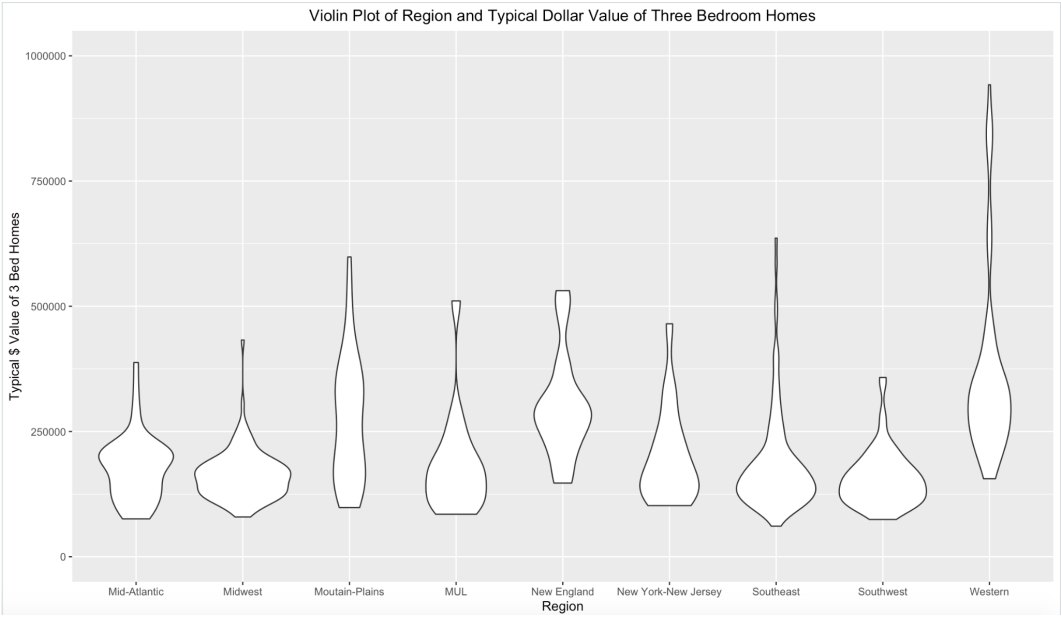
higher-level mean calculation. Additionally, minor formatting issues were addressed to improve importability to R studio such as formatting characters to numbers when applicable, removing unnecessary cells, deleting observations due to lack of data, and omitting but retaining certain variables.

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

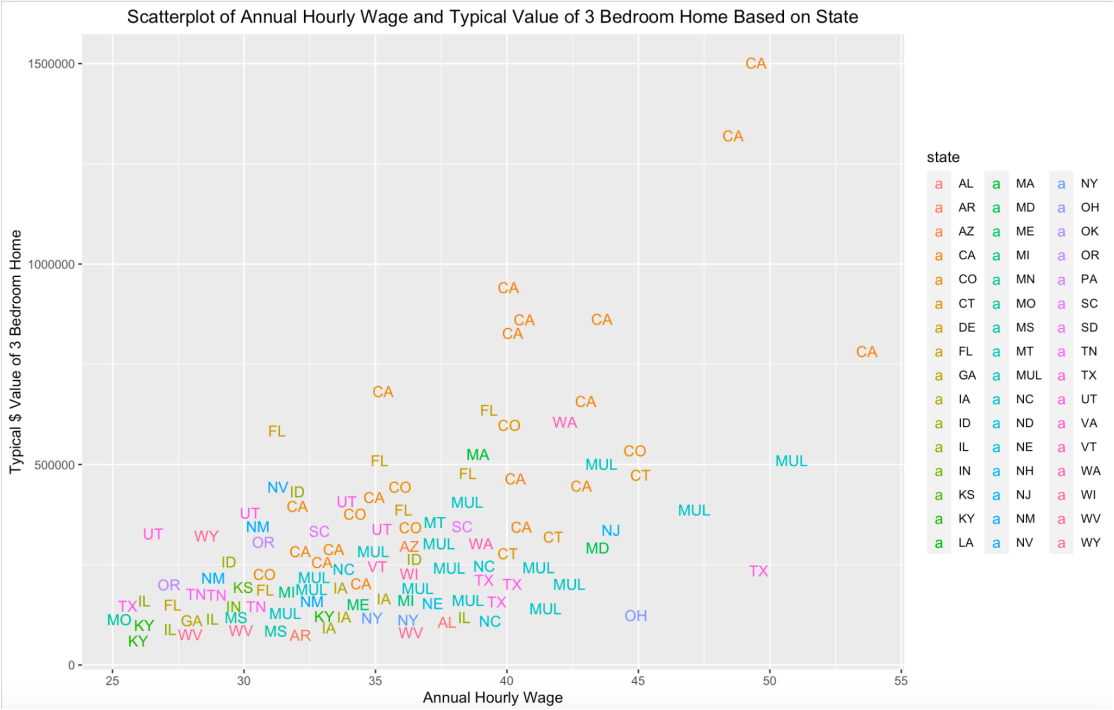
Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021
a relationship between the proportion of jobs in an area that were accounting jobs and what the hourly mean wage was in the area. We added a smooth geom to make it easier to see the relationship between 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. We ran a correlation on these two variables and found the correlation to be .4164707 meaning there is a slight positive relationship between these two variables. 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

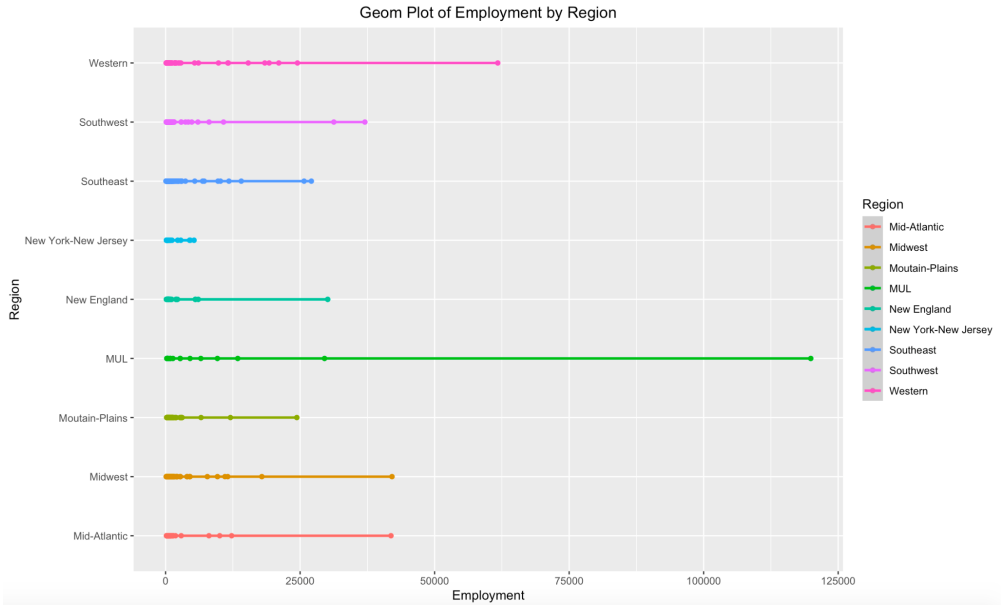
Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021

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 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.



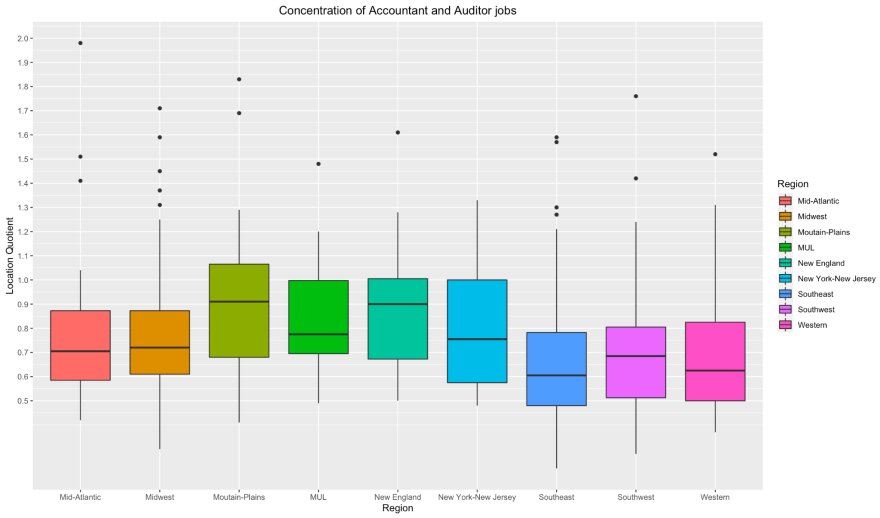
The third graphic we created was a scatterplot of annual hourly wage and the typical value of 3 bedroom home based on states. This scatterplot plots each observation and allows us to see the relationship between the annual hourly wage and the typical dollar value of

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021
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. The correlation between these two variables is 0.5276043 which makes it a positive correlation.



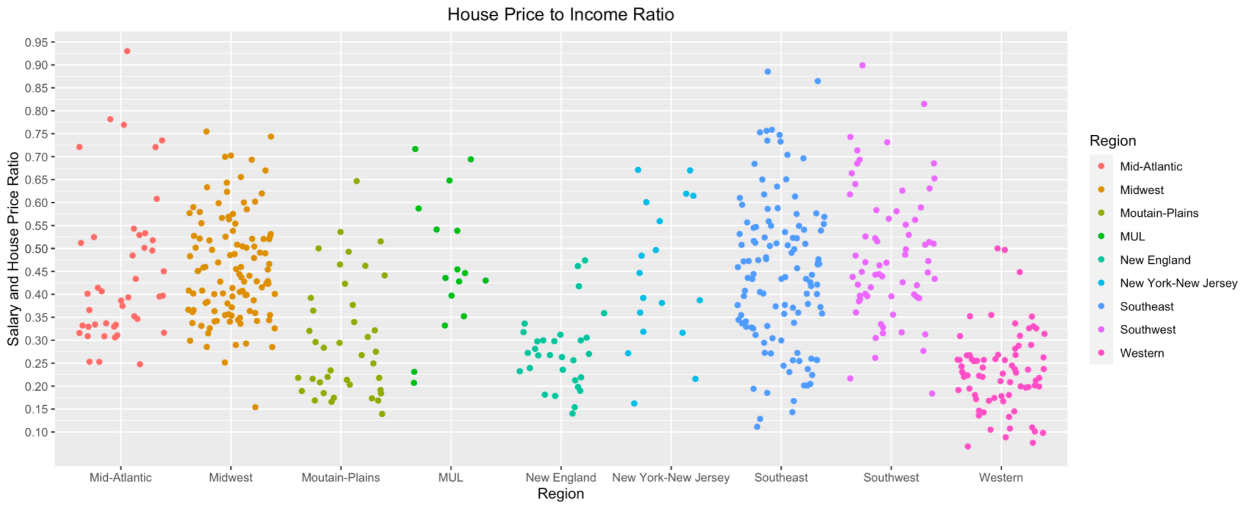
In this point range plot, we decided to see how much employment there is per region. Each dot represents an area within each region and the x-axis displays the amount of employment. The line connects all the dots so that it can display better how much employment of accountants/auditors each region has. 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

Alejandra Sotelo, Flavio Garcia, and Christine Partington
ECON 494 - 02R
Levkoff
April 18th, 2021
Western region had the greatest amount of areas since 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.



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 jitter plot. We used it to compare the salary to house price ratio between each region. Each data point is 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 (regions with fewer data points). 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/longer ranges. I would say that the relationship looks pretty strong because the data is well distributed among the regions. There seem to be longer ranges 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.

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

Project Part 2: Predictive Modeling and Testing:

1. Executive Summary:

For the second part of our project, we decided to dive into the variables that had a positive correlation from the first part of our project. In our first plot, we compared the hourly median wage and the employment per 1,000 jobs. We ran the correlation function in R and found that it was .4164707, which is great because it is a somewhat strong positive correlation. We did the same for our third plot which compares the hourly wage and home value. We found that the correlation was 0.5276043, which is a stronger positive correlation relationship. Because of the stronger positive correlation, we decided to choose home value and hourly wage as our primary variables of focus. For our multiple linear regression model, we decided to choose variables that had a positive correlation with hourly wage and home value.

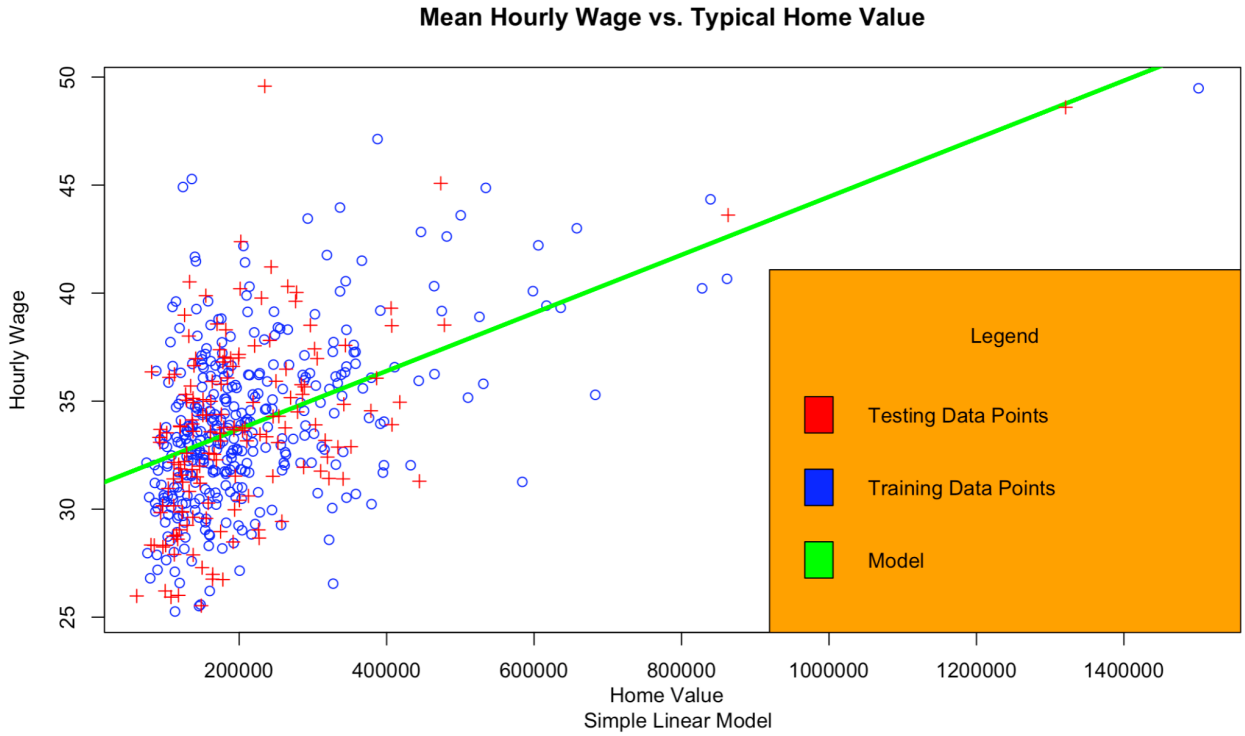
```
> cor(df$homeval, df$Employment)
[1] 0.3075275
> cor(df$LocationQuotient,df$homeval)
[1] 0.2782024
> cor(df$Hourlywage, df$LocationQuotient)
[1] 0.4164431
> cor(df$Hourlywage, df$Employment)
[1] 0.4068923
```

Above are the correlations between our primary variables of focus (Hourly Wage and Home Value) and the variables we chose for our multiple linear regression models, which were; Employment and Location Quotient. As you can see, they are positively correlated. We also decided to work hard in getting our in-sample and out-of-sample error for our multiple linear regression model as low as possible. So, we decided to add every state in our data to one of our multiple linear regression models.

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

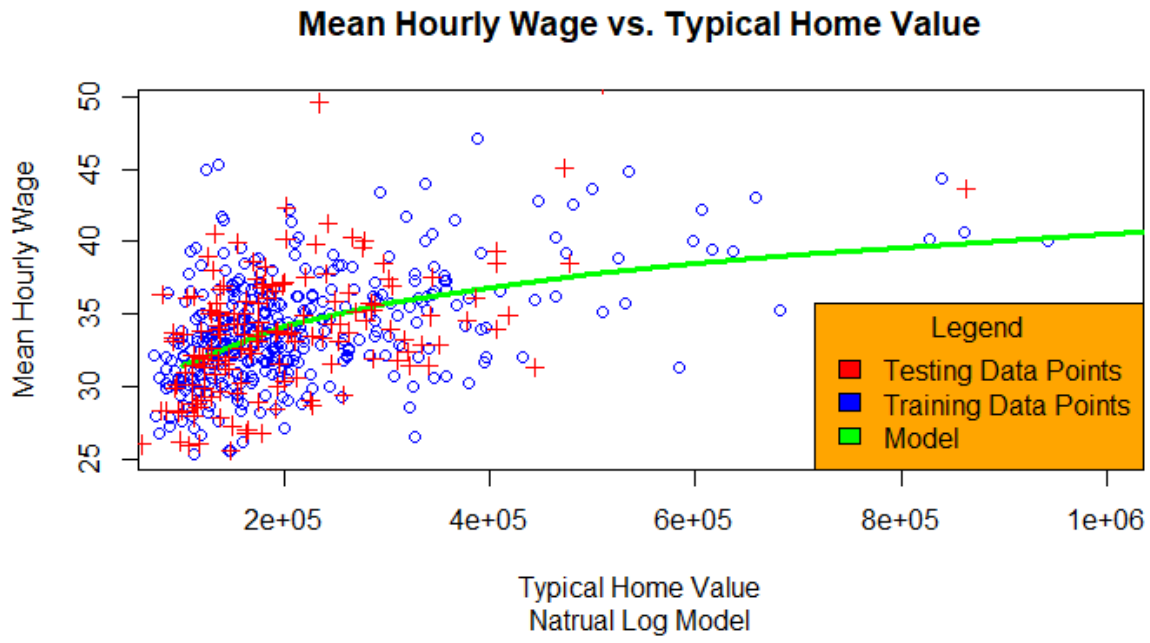
2. Single Variable Linear Regression Models:

The first model we created was a simple single variable linear regression: hourly wage= home value or predicted hourly wage= 3.102+ 1.344x . This regression model shows that a unit (measured in the thousands) increase in home value will increase the hourly wage by $1.344. This finding was not surprising because for the average professional to purchase a home, they must be able to afford it. So, it only makes sense that as the home value increases so should the hourly wage. My residuals were -8.8682 for the min, -0.1227 for the median, 12.4380 for the max. This tells me that there are residuals that are not normally distributed due to the extreme positive and negative values, and such a low negative median. My model does not suffer from multicollinearity because this is a single variable regression model. My R-Squared is 0.2569. This means that the regression model only explains 25.69% variation in Y (goodness of fit). This variable is significant at 1%. My in-sample error for this model is 3.327.



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

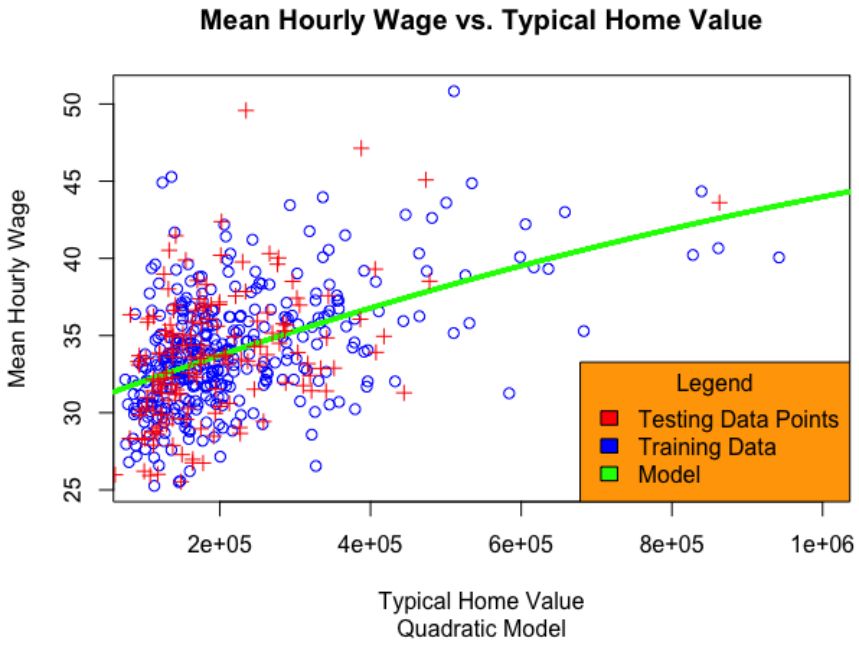
The second model we developed was a logarithmic model: Hourly wage= home value + ln(home value) or Predicted hourly wage = -13.98 + (3.9418*x). The natural supposition is that as home values increase so too should the mean hourly wage due to purchasing power considerations. So, when the beta is transformed through the natural log the positive correlation remains. Furthermore, plots were generated to test for homoscedasticity, linearity, and normality of residuals; which found the model satisfied all assumptions. With this in mind statistical analysis could be carried out with confidence to compare overall fit in and out of sample. This analysis produced an in-sample RMSE of 3.342023 and out of sample RMSE of 3.934744. Ultimately the group decided that comparing the in-sample and out of sample RMSE would be the optimal way to judge which model performed better. Although this model did not perform as well as the others we felt all models were competitive given their minor differences in outputs.



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

The third linear model we created was a quadratic equation. The model looked as follows: Hourly wage = home value + (home value)^2. When we input our coefficients we get the following regression equation: Hourly wage = 3.032e+01 + 1.782e-05(home value) - 4.15e-12(home value^2). What this model is essentially showing is that hourly wage is a function of homevalue to the second power. Looking at the residuals we see that the median is -.0736 and the 1Q is -2.0989 and the 3Q is 1.7151. We can see that these residuals are not normally distributed. The F-statistic tells us the overall significance of the model. The f statistic is 68.48 with a p-value of 2.2e-16 meaning that the model is overall significant at the 99% confidence level. To look at the significance of individual independent variables in the model we look at the t-tests. The intercept and homeval betas are significant at the 99% confidence level with p-values of 2e-16 and 3.29e-10 respectively. The homeval^2 variable is not significant at the 90% level of significance with a p-value of .116. The adjusted r-squared of the model is 27.27% meaning that 27.27% of the variance in hourly wage can be explained by this model. To measure how well the model performs in the training data in-sample we calculated the root mean squared error (RMSE) of the training data to be 3.323516.

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



To see which model is best in-sample we compared the root mean squared errors on the training data for each model. We chose to use the RMSE because it is easy to compare across models because the units of measurement are the same. Model 1 had an in-sample RMSE of 3.327, model 2 had an in-sample RMSE of 3.34, and model 3 had an in-sample RMSE of 3.323516. When comparing these in-sample errors they are very close, but the third model has the lowest in-sample error meaning that it performs the best in-sample.

3. Testing our Models Out of Sample:

To measure which model performs best out-of-sample we calculated the RMSE on the testing data for each model. We chose to use the RMSE because of the consistency of the units which would make it easy to compare our models against one another. The out-of-sample errors calculated were as follows: model 1 had an out-of-sample error of 3.90, model 2 had an out-of-sample error of 3.93, and model 3 had an out-of-sample error of 3.866603. These error

terms are similar out-of-sample as they were in-sample. That being said, model 3 has the lowest out-of-sample error and therefore performs the best out-of-sample. Model 3 also performed the best in-sample, and therefore we had the same model perform best both in-sample and out-of-sample.

4. **Multiple Linear Regression Models:**

The first multiple linear regression model we created is the following: Hourly Wage = employment^4 + employment^3 + employment^2 + employment + homeval^2 + homeval. With the coefficients added in our regression equations looks as follows: Hourly wage = 2.702e+01 - (1.191e-04)employment^4 + (6.498e-03)employment^3 - (1.055e-01)employment^2 + (1.022e+00)employment - (3.434e-12)homeval^2 + (1.496e-05)homeval Looking at the residuals we see that the median is 0.0174 and the 1Q is -2.0492 and the 3Q is 1.5601. From this, we can see that the residuals are not perfectly normally distributed. The F-statistic tells us the overall significance of the model. The f statistic is 30.27 with a p-value of 2.2e-16 meaning that the model is overall significant at the 99% confidence level. To look at the significance of individual independent variables in the model we look at the t-tests. The intercept and homeval betas are significant at the 99% confidence level with p-values of 9.49e-08 and 8.71e-08 respectively. The rest of the independent variables are not significant at the 90% confidence level. The adjusted r-squared of the model is 32.79% meaning that 32.79% of the variance in hourly wage can be explained by this model. To measure how well the model performs in the training data in-sample we calculated the root mean squared error (RMSE) of the training data to be 3.176993.

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

For our second multiple linear regression we decided to incorporate the variable employment into our regression equation: predicted hourly wage= home value + employment or predicted hourly wage= 3.177 + 1.109x + 1.827x. This multiple regression model reveals that an increase in home value and employment raises hourly wage by 1.109 and 1.827. This finding is surprising because if employment is rising, which therefore the job market may be getting bigger for accountants/auditors, the increase of home value does not affect hourly wage as much as it did with the simple linear regression model. But, by incorporating the employment variable and noticing that it affects the hourly wage, it seems that there could be more factors that may influence hourly wage as well. My residuals were -11.4075 for the min, -0.0003 for the median, 12.5631 for the max. They do not seem to be normally distributed because the same case is here where the median is centered around a very small negative value and there are extreme negative and positive values. My regressors seem to be positively correlated with each other. My model does suffer from multicollinearity due to the R-squared of 0.3144 being much higher than my simple single variable linear regression. So, this multiple regression model only explains 31.44% variation in Y. Both of these variables are significant at 1%. The in-sample error for this model is 3.19648.

By far the largest takeaway from creating the models was that by incorporating categorical variables by way of dummy variables we were able to exponentially increase our prediction accuracy. The model in question was:

Hourly wage = homeval + LocationQuotient + emp + AL + AR + AZ + CA + CO + CT + DE + FL + GA + IA + ID + IL + IN + KS + KY + LA + MA + MD + ME + MI + MN + MO +

MS + MT + NC + ND + NE + NH + NJ + NM + NV + NY + OH + OK + OR + PA + SC + SD + TN + TX + UT + VA + VT + WA + WI + WV + WY

Although overfitting was a possible consequence of the model's complexity we found that we were able to maintain an out of sample RMSE of 3.473065, while producing an in sample RMSE of 2.675078. Although some trade offs were made in the model, the in sample and out of sample performance remained better than the other multivariate models. The model's integrity was further substantiated by analyses to evaluate linearity, homoscedasticity, and normality of residuals. Given the model's complexity special care was taken to evaluate explanatory variables for collinearity. All of the assumptions were found to be true and therefore the model could be used in further regression analysis.

To see which model is best in-sample we compared the root mean squared errors on the training data for each model. We chose to use the RMSE because it is easy to compare across models because the units of measurement are the same. Model 1 had an in-sample RMSE of 3.176993, model 2 had an in-sample RMSE of 3.19648, and model 3 had an in-sample RMSE of 2.675078. When comparing these in-sample errors, the third model has the lowest in-sample error meaning that it performs the best in-sample.

5. **Testing our Models Out of Sample:**

To measure which multiple regression model performs best out-of-sample we calculated the RMSE on the testing data for each model. We chose to use the RMSE for the same reasons we explained before. The out-of-sample errors calculated were as follows: model 1 had an out-of-sample error of 3.587377, model 2 had an out-of-sample error of 3.692579 and model 3

had an out-of-sample error of 3.473065. Model 3 has the lowest out-of-sample error and therefore performs the best out-of-sample. Model 3 also performed the best in-sample, and therefore we had the same model perform best both in-sample and out-of-sample. The incorporation of dummy variables exponentially increased the predictive power of the multivariate model 3. This process allowed the model to take into account the "State" variable by way of several variables denoting each state. In this manner, the model's ability to account for effects on the predicted y value of each independent variable improved significantly. A major observation of this improvement was that when incorporating the state variable the model could account for approximately 52% of the variation in y.

6. Conclusion and Proposal:

The most important observation the group extracted from the overall analysis and modeling was the major predictive improvements that came about when incorporating dummy variables. Compare for example the adjusted R squared values of model 3 (dummy variables included) as opposed to model 4 (dummy variables not included). Model 3 produced an R squared value of 0.4424 while model 4 produced a value of 0.3128. Thus the conclusion is that the incorporation of dummy variables produced a 12.96% increase in the model's ability to explain the variation in y due to the x's. Ultimately the best model as dictated by our measure RMSE in and out of sample is the multivariate model 3.

Feedback from Part One of our project:

In our part one of the project, we had to fix two things, replace our hexagon plot and add a correlation analysis to our fitted scatter plot. Our hexagon plot was not the best plot to use due to the overcrowding of points in our data. It was really difficult to see the overcrowding in each

region. To fix this problem, we decided to use a jitter plot. It was a much better visual to see how much overcrowding there was in each region. To add a correlation analysis to our fitted scatter plot, we ran the correlation between hourly wage and employment in R so it could assist us in finding relationships that could help us in our second part of the project.