Data 670 Data Analytics – Capstone Project

Geovanna Karina Meier

Professor Dr. Steve Knode

August 6, 2018

Executive Summary

Through exploratory data analysis and three separate predictive models, this report examines the relationship between employment and median housing prices in metropolitan areas in the United States, and how employment by industry influences housing prices. First, the report considers the historic trends of median housing prices and how they changed in different regions from 2012 to 2017. Then, it examines the employment trends by industry and worker's age and how they differ across industries. And finally, it compares three different predictive models of median housing prices in metropolitan areas, using total employment counts by industry as predictors. The metropolitan areas have been segmented into eight economic regions as defined by the Bureau of Economic Analysis in order to compare median housing prices within regions with similar "economic activity and growth" (Bureau of Economic Analysis, n.d.).

The data came from Zillow Research and the U.S. Census Bureau. The housing data, which came from Zillow Research, provided historical monthly median housing prices and monthly rent prices in metropolitan areas from 2012 to 2017. The employment data, which came from the U.S. Census Bureau, provided quarterly employment statistics by industry in metropolitan areas from 2012 to 2017.

The three models to predict median housing prices were multiple linear regression, decision tree and random forest. Following are the main findings and conclusions from the data analysis and the predictive models:

- Median housing prices have seen a rapid increase across metropolitan areas in the U.S. between 2013 and 2017, particular in the North East and West Regions, where

the technology sector (west region) and finance industry (northeast region) are predominant.

- The Retail Trade and Health Care and Social Assistance industries have both (1) a wide spread presence across the country and (2) consistent employment counts in coastal cities.
- The Health Care and Social Assistance industry has seen steady hiring of a young workforce and job creation continues to get stronger in coastal metropolitan areas, especially among the 25 to 34 age group.
- The Retail Trade industry saw an increase in employment numbers across all metropolitan areas between 2012 and 2017, particularly in five metropolitan areas, which also saw a surge in median home prices in 2017.
- Based on the predictive models, total employment counts by industry impact median housing prices differently in different economic regions and not all industries have a positive impact on housing prices. The same industry can be both positive and negative correlated to median housing prices in different regions. For instance, while the Finance and Insurance sector is associated with an increase in housing prices in the Rocky Mountains region, it is linked to lower median housing prices in the New England and in the Plains region.
- Total employment counts in the construction industry is positively correlated to median housing prices in five of the eight economic regions.
- Lastly, based on a combination of accuracy and simplicity, the Multiple Linear Regression Models provided a ‘better’ predictive model of median home prices, with high Adjusted R-Squares and low Root Mean Squared Errors.

Table of Contents

Project Scope

Problem Description

As industries gain and lose traction across metropolitan areas in the United States, new employment trends arise, bringing changes to the housing market, which becomes a key indicator of the region's development. Hence, identifying and understanding employment trends and housing affordability is not only important for employers and job seekers, but the entire population, as these inevitably shape the area's cultural, academic, and socio-economic landscape

Thus, the aim of this project is to analyze employment flows (job creation, earnings, employment by industry, etc.,) and housing values across metropolitan areas in the United States to find associations between:

- workforce changes by industry and evolving housing markets,
- job creation and job losses by industry in different metropolitan areas,
- average earnings by industry across metropolitan areas, and
- housing affordability across different metropolitan areas.

The main objective is to develop a predictive model of median housing prices (target attribute) in metropolitan areas in the United States using total employment counts by industry and housing data from 2012 to 2017. The idea is to understand if there is a real correlation between industries, employment, and housing prices and identify the main attributes that influence housing prices. This will be accomplished by creating a predictive model that uses employment and housing historical data as inputs and helps predict price surges or declines.

As the debate continues on whether Amazon's second headquarters (HQ2) will impact housing prices, as recently pointed out by several articles that insist that HQ2 will "bring soaring housing prices" (Garfield, 2018) and that "having [HQ2] as a neighbor will result in very different consequences for homeowners and renters" (Passy, 2018), the idea is to consider if industries and employment can have a real effect on housing prices that extend beyond Amazon's HQ1 example in Seattle, where housing prices' have increased at an "average annual rate of 10.3 percent from 2012 to 2017" (Schuetz, 2018). Or even California, which hosts some of the most prominent tech companies in the country and it's experiencing "exorbitant housing prices and transportation failings, fueled by the growing ranks of tech companies" (Salinas, 2018).

Other objectives addressed throughout the analysis are:

- Identifying the industries with the highest traction and employment rate in each metropolitan area.
- Identifying metropolitan areas with the highest median rent and housing prices to highlight the most and least affordable micropolitan areas in the United States.
- Identifying the industries with the highest payed workforce by metropolitan area.

The datasets used for the analysis will be retrieved from three different sources:

- **Employment data** was retrieved from the U.S. Census Bureau, Center for Economic Studies (U.S. Census Bureau, Center for Economic Studies, n.d.) by year, quarter, and state via a combination of the LED Extraction Tool and the Census API. Six datasets were retrieved by year for 48 states and Washington D.C. using the Census API, and two separate datasets were retrieved for WY and WI, for all the years using the LED Extraction tool. The 8 datasets were later

combined to have a single employment dataset.

The dataset for each state provides information on year, reference quarter, industry, state code, metropolitan statistical area/micropolitan statistical code (CBSA), workforce age by industry, and employment statistics (beginning of quarter counts, full quarter counts, new hires, separations, stable hires, turnover rate, job gains at firms, job lost at firms, firm job net change in employment, average monthly earnings of employees with stable jobs, average monthly earnings for new hires, total quarterly payroll). Employment indicators also include accompanying variables for each attribute with information on why items are missing.

- The housing data was retrieved from the Zillow Research Data page (Zillow Group, 2016). Three different datasets are being used:
    - **Median Rent Values (ZRI dataset)**, which provides monthly data from 2010 to 2018 on median rent prices by metropolitan area, including regionID, regionName, sizeRank and monthly columns with median rent prices (Zillow Research, n.d.).
    - **Median Home Values** (**ZHVI dataset**) with monthly data from 1996 to 2018 on median home prices by metropolitan area, including regionID, regionName, SizeRank and separate columns with monthly median housing prices, which will become the target variable once they are combined into a single feature. (Zillow Research, n.d.).

    - **Conventional 30-Year Fixed Mortgage Rates** with the "average mortgage rate quoted on Zillow Mortgages for a 30-year" (Zillow Research, n.d.) in 15 minutes increment from 2011-2018.

- Zip to CBSA code crosswalk files from Zillow Research. Initially, these files were retrieved from the Office of Policy Development and Research, U.S. Department of Housing and Urban Development, however significant issues arose during the exploration and cleaning process and they had to be replaced by Zillow's crosswalk files. These files are used to translate ZIP codes to CBSA code in order to combine the employment data with the housing data.

The main tools used for this project are Python, R, Tableau, and SAS Enterprise Miner[1]. Python and R are used to retrieve and combine the individual datasets from each data source, prepare them and later to merge them by metropolitan area. Tableau will be used for visualization purposes, progress reports and presentations. Since Tableau allows to change data types with "geographic roles" (Tableau, n.d.), data can be easily visualized analyzed by region and time period. Python, R and SAS Enterprise Miner will be used for the predictive models.

Business Understanding

The two main domains for this analysis are the labor and housing sectors. Although this type of analysis is not new and there are reports that examine housing costs and employment trends in the United States, either separately or on the impact of one on the other, this project will specifically analyze employment statistics by industry,

[1] Use of SAS software is dependent on usability and will vary or may be limited due to restrictions or accessibility in Germany. Watson Analytics may be used instead

workforce demographics, and housing costs on metropolitan areas in the United States from 2014 to 2016, along with the relationship between various employment statistics and housing affordability.

The labor market analysis uses employment statistics by industry to give insights on employment trends, workforce flow, regions with most stable hiring, on-demand skills based on the local industries, and the overall socio-economic outlook of a region. The housing market analysis uses mortgage rates, local home values, and rental values across metropolitan areas to gain a deeper understanding on how affordability can facilitate or hinder workforce migration. Hence, the combined analysis done in this project can show which areas of the country are seeing increase hiring, workforce migration, and industry grow.

Thus, following the CRISP-DM guidelines (IBM, n.d.) the main questions that would be answered by this analysis are:

- What industries have the highest hiring rates?
- What metropolitan areas have the highest housing prices?
- What are the metropolitan areas with the highest average earnings?
- Which industries have the highest payrolls and highest average wages?
- What are the main employment measures that impact and help predict housing values?

The answers to these questions will give employers and job seekers, insights on local labor trends (i.e. workforce flow by industry) and housing costs.

Organization

Since this analysis focuses on the convergence of the housing market and employment, the business purpose is to provide information on key socio-economic aspects to the general public. Employment statistics, which are obtained from employers, provide information on workforce demographics, jobs created, and jobs lost, which gives a general overview on the strength of the labor market in a geographic area.

The housing market statistics provide an overview of local real estate and migration dynamics, which along with the employment data can be used by economists, employers, workers, and job seekers to make informed decision on their social and financial future.

Thus, as depicted in the diagram below, the data gathered from businesses and real estate, would provide insights that fuel socio-economic decisions by different stakeholders at multiple levels.

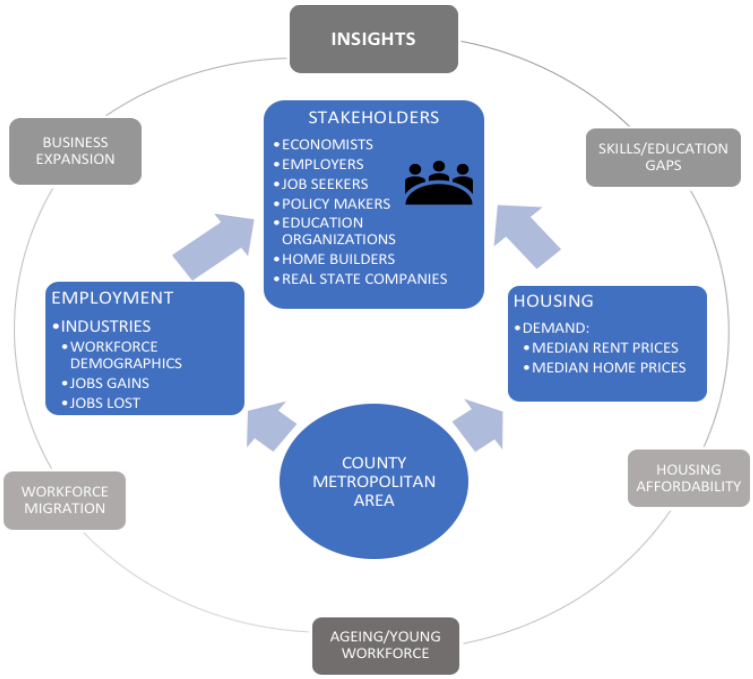


Figure 1. Diagram showing data gathering and usability by stakeholders

Stakeholders

The insights of this analysis can be used by a wide array of stakeholders. From employers and jobseekers, to economists, home builders, education institutions looking to tailor their curriculum to match the local labor environment, and policy makers looking to make their states more attractive, this analysis can provide useful information on the socio-economic outlook of the state. The findings of this project, however, will focus particularly on employers, jobseekers and policy makers.

Employers can use the insights to make informed decisions on business expansions, job creation, and on-the-job training programs in specific geographic areas. Jobseekers can use the insights to compare locations that offer similar employment opportunities but differ on housing affordability or workforce demographics. Policy makers can use the insights to better serve the demographics of their workforce, and to decide whether or when to implement housing assistance programs when housing affordability becomes an issue, especially for low-income residents.

Define Business Area

As previously stated, this analysis focuses on the convergence of the housing market and the labor market. More precisely, whether and how housing prices vary as a result of local industries and workforce demographics. The state of the local housing market is a key factor for both employers, who need to offer wages that are in line with the cost of living so they can retain workers, and for jobseekers who may be discourage to relocate if housing costs will take most of their salary.

An example is Amazon's planned HQ2. Seattle, home of Amazon's primary HQ, saw a 10.3 percent year-over-year increase in average home prices between 2012-2017

(Schuetz, 2018). If the tech giant's presence and economic impact on the local economy is anything like it is in Seattle, the metropolitan area that wins the bid for HQ2, could see not only "50,000 future employees and eight million square feet of office space over the next several years" (Schuetz, 2018), but also a rapid growth in the housing market value and inventory.

Another example is Salt Lake City, which has seen a steady job creation growth of "3.6 percent year-over-year… and 3% unemployment" (Vivas, 2018), but also a steady growth in the housing market with median housing prices coming over 15% above the national median housing price in 2017 (Vivas, 2018). This rapid growth is making it increasingly harder to afford a home, especially for millennials moving in, whose salaries "are only slightly higher than the national average" (Olick, 2018).

Hence, using the output of the predictive model (median housing prices), companies considering regional expansions or job seekers considering moving, could consider estimated home values to decide whether they could afford to move in or even remain in their city.

More generally, this project will look into:

1. The impact of employment and workforce dynamics (i.e. job creation and job losses) on median housing prices.
2. The local economic impact of industry and employment dynamics (average earnings and total payroll by industry) in different metropolitan areas across the U.S. when it comes to rent and housing prices.

Business Objectives

1. Identify industries with the highest hiring rates across metropolitan areas to determine which industries favor particular regions. Since the predictive model will use employment statistics by industry to predict median housing prices, the output of the model could help state legislators use this information to determine if and how to invest in infrastructure and housing projects to attract employers of these specific industries.

Additionally, this information could help job seekers with skills on these industries and interested in relocating, narrow their search to consider locations that are actively hiring and where the median housing prices are still 'affordable" but are *expected* to increase and appreciate over time. For instance, "home price appreciation in San Francisco, [CA] between 2012 and 2017 was 84 %… [mainly] fueled by the combination of a strong local economy steadily adding tech jobs, rising stock prices that benefit tech professionals, and confident buyers" (Hansen, 2018).

This shows how creating a predictive model that considers employment trends by industry can help understand if the impact of industries on housing values is broad and can help predict surging housing prices in other metropolitan areas.

2. Identify industries whose workforce is consistently younger across metropolitan areas and the average earnings per industry. Determining whether the industries attracting young population have wages that are inline or higher than the national average, can help identify regions expected to see changes in housing market due to higher demand. Since young adults "aged 25 to 34 make up 13.6% of the U.S. population, but 30% of the current population of existing-home buyers" (Realtor.com cited by Sullivan, 2018), understanding which areas provide employment and good

paying jobs for these older millennials, can shed some light on where the next generation of home buyers is expected to be.

This information, couple with the predictions of median housing prices can be used by state legislators and construction industry to estimate their regions likelihood and ability to host a wave of housing demand.

3. Create a predictive model that helps identify whether external factors are correlated to housing prices and how they influence housing affordability. The idea is to identify if there is a clear positive (close to 1) or negative (close to -1) relationship between housing values and external attributes such as location, workforce flow, and employment indicators. The end goal is to provide stakeholders with information that directly affect their socio-economic outlook.

- Employers can use the predictive model to adjust recruitment practices (increase remote jobs) and hiring benefits (transportation or housing assistance) in regions that are *expected* to see housing increases that represent more than 28 percent of household's income, based on housing expense ratio (Quicken loans, n.d.).
- Policy makers can use the model's output to understand what factors may affect housing affordability in their districts and devise plans that help protect low-income families from being left behind.
- Job seekers can use the model to estimate the housing costs of relocating to particular regions based on the model's key predictors of housing prices.

Business Success Criteria

- One criterion for the successful outcome of this project is to recognize workflow and employment trends by industry, clearly defining the geographic locations

where specific industries dominate. For instance, an initial assumption is that tech-related jobs are primarily in costal locations (i.e. Washington State, California, Boston), and most manufacturing jobs are locating in the Rust Belt. The analysis will help determine if these assumptions are true and whether these trends are growing or decreasing over the years.

The use and benefits of these results are subjective and will be different for each stakeholder (Chapman et al., n.d). For instance, employers could use this information along with the results from the predictive model to make decisions on business expansions based on places where their industry dominates and hiring is proving successful, which shows the presence of skilled workers, along with the outlook of the region's housing market.

- A second criterion is to find if there is the correlation between employment indicators, workforce flow, geographical location and housing prices. A clear correlation can help identify which attributes are the most likely predictors of housing prices and could help predict geographic areas that could see a decrease in housing affordability. For instance, identify whether there is a positive correlation, closer to 1, between wages and median house value. As with the first criterion, these insights would be used differently by each stakeholder, and just like with the last objective, they would be useful for employers considering relocating as they are "today more than ever are locating new offices in lower cost of living" (Chamberlain as cited in Mejia, 2017).

Background

The progressive rise in housing prices across the Unites States (Olick, 2018), particularly in coastal cities, is making big corporations rethink expansion plans or even relocating and forcing some residents to move to more affordable cities. This is a prime example of the situation in California (Daniels, 2018). As pointed out by Redfin CEO Glenn Kelman, “the technology companies, the Wall Street companies, they're chasing the talent, [and] the talent is chasing affordable housing… The "mass migration" to the center of the country is populating cities that were “economically dead five or 10 years ago” (as cited in Salinas, 2018). Hence, as more industries move in, employment trends change, and housing becomes a key competitive factor for areas looking to attract employers and a new workforce.

However, the rising housing costs impacts more than local employers and job seekers wanting to relocate. They fuel housing instability among current residents and contribute to “regional housing disparities [that] interfere with the recruitment of workers and a firm’s performance” (Kneebone, Snyderman, Murray, 2018). This, in the end, has a ripple impact on the social and economic development of a region.

As an example, a 2017 research from Freddie Mac showed how rental affordability across different metropolitan areas is worsening for very low-income (VLI) families – those who make less than 50 percent of the median regional income (Freddie Mac Multifamily, 2017). The analysis, which considered nine states where Freddie Mac Multifamily financed loans at least twice between 2010 and 2016, shows how the number of “rental units qualified as VLI” fell at an alarming pace. For instance, “in Colorado, the

number fell from 32.4 percent to just 7.5 percent of the units" (Freddie Mac Multifamily, 2017) between the first and second financing.

Thus, this project intends to analyze how employment trends and workforce demographics impact housing affordability beyond coastal cities and develop a predictive model for housing prices across metropolitan areas. The idea is to find the key predictors of housing prices and help determine which areas can see a new wave of housing crunch or development opportunity. Since as pointed out by a business leader who is actively involved in the Metropolitan Planning Council in Chicago, "economically competitive regions *need* housing [emphasis added] that connects their workers to jobs, schools, and transportation" (Harris, 2018) and state legislators need to find solutions to the increasing affordability housing of their constituents in large metropolitan areas.

Research

As previously mentioned, there have been previous reports that examine housing costs and employment trends in the United States, but most of them have done it separately or focus on predicting housing values based on internal features. In this analysis, the focus will be on the geographic and socio-economic features where the homes are located.

Recent analyses have examined the relationships between *job posting* and housing costs in the largest metro areas, as seen in the latest indeed.com article titled "The Jobs Priced out of Expensive Metros" (Kolko, 2018). The report examines *job title's postings* and housing costs in the ten most expensive metro areas in the U.S., however it doesn't take into consideration other elements such as workforce demographics or hiring rates.

Another example of research done in this area is LinkedIn's monthly "Workforce report" using LinkedIn data, which looks "into hiring, skills gaps, and migration trends across the country [at the national level], and … insights into localized employment trends in 20 of the largest U.S. metro areas [at the city level]" (2017). These reports provide useful insights on industry hiring and the cities with the widest skill gaps, but there is no analysis on housing prices.

In 2016, the U.S. Bureau of Labor Statistics presented a report on the "Major industries with highest employment, by state, 1990-2015" (2016), using the Quarterly Census of Employment and Wages. The report shows how five major industries: Manufacturing, retail, health care, accommodation and food services, and professional and technical services, have move across the country and in some cases, like the manufacturing industry, hiring has seen a steady decline. Just like previous reports, the analysis is centered on employment and not the relationship with housing prices.

Gaps in this Problem Resolution

The gap in research is the relationship between local employment trends (hiring by industry, and industry/workforce migration) and housing prices. That is, how local industries and workforce characteristic may influence housing affordability. In general, most research projects tend to focus on one aspect or the other, but not whether and how they interact. Additionally, projects done to predict housing prices, primarily use internal factors to explain home values.

What this analysis expects to accomplish is to show that there are other factors that may weight more on home prices than the internal characteristics of the home. Mainly, the industries, employment rates, and workforce located in the area. A case in

point is housing affordability in California. Currently, California is "facing exorbitant housing prices and transportation failings, fueled by the growing ranks of tech companies" (Salinas, 2018) and is not uncommon to have exorbitant new listings just because they are located near major tech companies. For instance, a recent listing showed a "dilapidated and boarded-up house [that] sits on a 5,800 square foot lot in San Francisco's Bay Area" and burned out two years ago being listed for $800.000 (Turak, 2018). This shows that the type of employment and workforce of a geographical area may have a big impact on housing affordability.

Proposed Project

One of the reasons I chose this analysis came from a personal desire to identify the most prominent industries in different metropolitan areas in the U.S. As a soon to be graduate student, identifying the most adequate place to relocate, means considering two key factors: job prospects and housing options. However, there is a gap in analysis, which is why with this project I want to consider workforce demographics, local demographics, and the likelihood that areas with similar workforce and employment factors, see a different degree of housing affordability.

Another reason is that as shown by the research conducted by Freddie Mac Multifamily in 2017, housing affordability is increasingly – and dangerously affecting low income families. Between 2010 and 2016, in the U.S. "11.2 percent of the total number of rental units were affordable to very low-income (VLI) households …[and] by the second financing, rents had increased so significantly relative to income that just 4.3 percent of the same units were affordable" (Freddie Mac Multifamily, 2017). Thus, using the model's output policy makers could identify and understand which factors are most

likely to impact housing values, in order take preemptive measures in their districts that help alleviate the situation for current and future residents.

Additionally, the output of the predictive model could help employers, as some of the biggest industries are seeing the 'unintended' consequences of their own success as seen in California. In a 2017 research and survey report by Raphael Bostic, a USC Price School of Public Policy Professor, showed that "60% of employers surveyed [in California] cite the region's high cost of living as impacting employee retention, and 75% cite housing costs specifically as an area of concern" (Bostic as cited by Bedrosian Center, 2017). Thus, the intention of this analysis and the predictive model is to identify whether the expansion of certain industries have a significant impact on housing prices beyond coastal cities.

Lastly, as stated in a 2017 report by Zillow "the decision on the right time to buy [or rent] a home generally revolves around two key factors: Landing a steady job and affording the monthly mortgage [or rent] payments on a home" (Zillow Research, 2017). Hence, doing an analysis on the two key factors that drive one of the most important choices people make: buying a home, makes perfect sense as it provides useful insights to a large number of people.

In general, the output of this analysis could be used as input in other models, as this is a subject that can produce further analysis and answer additional questions, such as: Do historically leaning blue states and red states attract different type of industries? Do they differ substantially when it comes to housing affordability? Are some congressional districts more or less prone to implement housing assistance programs for

low income families or workers moving in? Do the primary industries in a region impact in gentrification. If so, which industries? And which ones contribute to more inclusion?

Key Performance Indicators

1. Deliverables for the project are thorough and on time. By July 8, 100% of the restructuring has been accomplished for all datasets: age and industry attributes of quarterly workforce indicators are joined to the indicators, and the median housing and median rent datasets are in quarterly long format.
2. 25 percent of the visualizations are done in week 8. The goal is to build five different visualizations in total. One data visualization in Tableau using mapping and showing median housing prices by metropolitan area developed by Jul 9. The visualization will show five bins with ranges for median housing prices for metropolitan areas using the year attribute as a filter to see how prices have evolved over time. The quality of the EDA can help discover additional insights and aid data preparation for the development of the predictive models. It will also help ensure there is direct engagement with the data before developing the models.
3. By July 7, a correlation matrix has been developed to find clear correlations (close to 1 or -1), or lack of (0), between employment indicators. This allows to remove highly correlated variables before modeling
4. By July 14, 50 percent of the datasets merging has been accomplished. All the housing related datasets, median housing prices, crosswalk files, median rent prices and mortgage datasets have been merged. Since the datasets come from different sources, ensuring data is combined and merged using the same metrics (year, quarter and CBSA).

5. By July 17, 100 percent of the merges are done after completing the imputation process. The maximum overdue time for mergers is four days.
6. The predictive models explain at least 50 percent of the variation in housing prices using geographic and employment indicators as inputs.
7. Train at least two different predictive models and evaluate their performance using the Root Mean Square Error and/or Adjusted R squared to obtain over 50% predictive accuracy.

Project Insights of the Data Analysis

Before starting the analysis, I anticipate there is a strong relationship between the industries that predominate in a region and the median housing prices. One assumption is that manufacturing and construction are predominant in the center of the country and that housing tends to be more affordable in these areas than in coastal cities. Another assumption is that hiring in the health care industry has increased over the years and it tends to be located in the major metropolitan areas, with housing prices progressively higher. Hence, the analysis will help verify whether these assumptions are correct.

I also expect the analysis would find a positive relationship between earnings and median housing prices. As the average earning goes up, so will housing prices. Even though there is clear data showing that housing prices are high in areas with major presence of technology companies, such as Seattle, San Jose, and San Francisco (Daniels, 2018), I anticipate the relationship extends to other parts of the country and it's growing particularly in the west and south (Austin, Denver, parts of Florida).

Regarding workforce demographics, I expect the analysis to show that there is a clear difference in age among industries. For instance, before the analysis, I assume that

industries such as health care, food services and accommodation, along with professional, scientific, and technical Services tend to have a younger workforce than manufacturing and construction. The project may also identify other patterns on workforce demographics and industries.

Project Milestones

1. Data Selection (*June 17*): The datasets will be retrieved from the different sources.
    - Employment datasets will be retrieved individually by state from 2014-2016 for all metropolitan areas and then combined as a single dataset before merging with the housing data.
    - Housing data will be retrieved from Zillow Research.
2. Data Preparation (*July 8*):
    - Detailed description of datasets and variables will be provided.
    - The monthly columns in the housing data are binned into quarters to match the employment dataset' time frame and being able to merge.
    - From the employment data set, redundant features, attributes that don't provide new information or highly correlated attributes are removed.
    - After combining and restructuring all employment datasets, additional years (2012, 2013, and 2017) had to be included as the number of rows was dramatically reduced to just over 10,000. Using the Census API, this time data was retrieved by year for all states and later combined, which helped with processing speed and capacity.
    - Employment and housing datasets will be merged by metropolitan area.
    - New dataset with all the information will be reformatted if necessary.

- EDA will be performed to find outliers, missing data, or problems with the data.
- Initial visualizations will be performed with Python and then with Tableau to find additional insights on the data.

3. Data Analysis Report (July 15)

- Summary of what has been done thus far.
- Report with current steps taken to clean, transform, and prepare the data.
- Visualization done for the EDA will be presented in Python and Tableau.
- Software used to create the predictive model has been determined.
- Predictive models being considered will be explained.

4. Analysis Evaluation (July 29)

- A summary of what has been done thus far will be done.
- Analysis of the patterns uncovered in the EDA and the predictive model will be explained.
- Steps done to create the predictive model will be explained, including variables used, transformations done and what tuning parameters were used.
- Findings, review of success, and recommendations for future analysis will be explained.

5. Final Report (August 12)

1. Summary of all the steps done thus far will be provided.
2. Presentation and reports will be submitted.

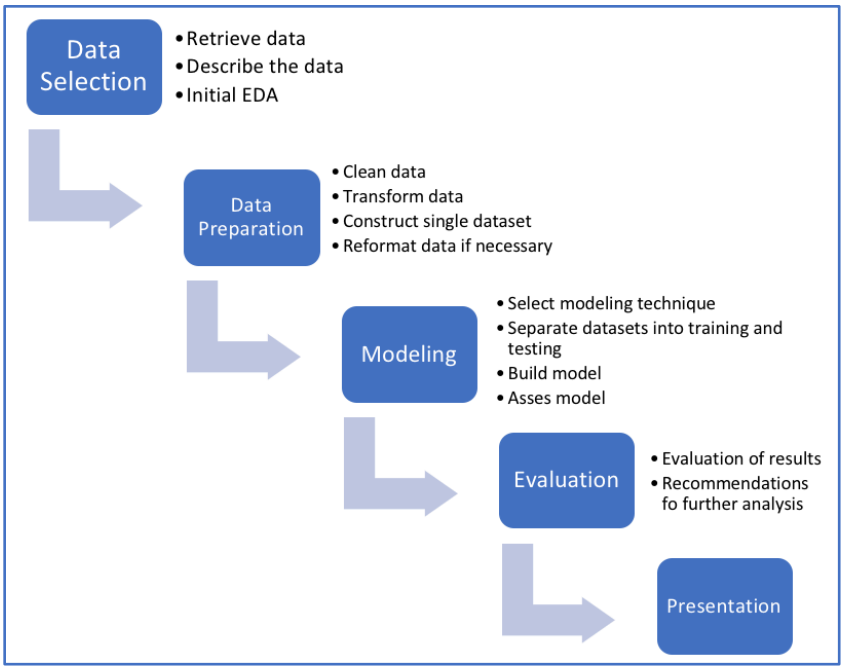


Figure 2. Diagram showing outline of project milestones based on CRIPS-DM (Chapman et al., n.d., p.12)

Completion History

Week 1 (5/21-5/26)	Defined project approached and researched suitable topics for project. • Identify several topics worth exploring: Gun violence and employment, Housing affordability and job announcements, Employment and housing prices. • Searched datasets suitable for selected project that complied with minimum requirements for project.
Week 2 (5/28-6/2)	Defined project subject and objectives and identify appropriate sources. 1. Set project's objective: Analyze relationship between employment and housing prices and create a model to predict median housing prices based on employment indicators and location.

	2. Identify data sources: Housing data from Zillow and employment data from U.S. Census Bureau
Week 3 (6/3-6/9	Selected datasets, reviewed their structure and how they could be connected: • datasets will be merged by year, quarter and zip code • identified alternative and additional datasets. Crosswalk files are needed to match zip codes to CBSA codes and being able to merge housing and employment data.
Week 4 (6/10-6/17)	Started downloading and inspecting datasets.
Week 5 (6/18-6/23)	Reviewed individual employment and housing datasets and possible ways to restructure the data: change wide format to long format, change frequency, etc. Preliminary review of housing distribution in Tableau.
Week 6 (6/25-6/30)	• Identified possible outliers, input errors, and missing values. • Identified issues with median housing prices distribution using the zip codes datasets and merging conflicts with crosswalk files. • Different data subsets for employment and housing were retrieved as data quality and quantity issues arose with both housing and employment data.
Week 7 (7/1-7/7)	Employment data is restructured to connect age groups and industry with workforce indicators Housing data by metropolitan area is restructured in long format from 2012 to 2017, missing values are imputed and frequency is changed from monthly to quarterly using the median.
Week 8	Visual examination of employment and housing data is done in Tableau. Data preparation process continues in both R and Python. Different predictive models are researched and evaluated.
Week 9	Continue with data cleaning and preparation for the model. Data is segmented in regions to develop eight different models to predict the (logarithm) median price of home in different regions.
Week 10	Summarize project purpose, analysis and models' outputs

Lessons Learned

Week 1	Lessons learned: There were a lot of different project and ideas that were worth examining and analyzing. Researching topics was certainly interesting and I found other topics that will analyze after this class.
Week 2	Finding and selecting 'good' datasets that suit the research topic and would fit the project requirements was not easy. There were interesting datasets, but they didn't have enough cases or attributes.
Week 3	• Downloading and merging datasets from different sources, can be quite time consuming, particularly when they are timeseries and they have to be downloaded manually. • Initially, I selected employment datasets from all states and metropolitan areas from 2014 to 2017 and statistics were split both by age group (7 in total) and sex (male, female, and both). However, soon I realized that the final dataset will all metropolitan areas had over 2 million rows, which seemed highly ambitious for the timeline of this project. Hence, I decided to work with a more manageable dataset including only 2014 to 2016, four age groups and all genders.
Week 4	Coding in both R and Python has been one of the most interesting and rewarding aspects of the project thus far. Although I'm not an expert, it has been very valuable to gain an intuitive understanding of the research process from finding and loading to inspecting the data.

Week 5	Setting up the data in the right format can take significant time and effort. My target variable comes from the housing dataset, which recorded data monthly since 1996 and had a different variable for each year. Hence, the data had to be restructure in a long format to have the target in a single variable and the frequency had to be changed to quarterly to match the quarterly workforce indicators datasets.
Week 6	• Using the same geographic measure to compare and merge datasets is vital to obtain unbiased results. During the EDA of the median housing prices dataset (target variable), it was discovered there was a wide spread of prices inside the same CBSA and some zip codes were unincorporated statistical areas (Missouri Census Data Center, n.d.). Hence, a new data subset for median housing prices was retrieved by metropolitan area. • More is better. After selecting the data from 2014-2017 and restructuring it to have the workforce indicators tied to each industry and age group for the predictive model, the dataset went from 545,323 rows to just over 10,000 and over 950 variables. Hence, additional years needed to be added.
Week 7	• Restructuring can take a significant amount of time, especially when datasets come from different sources.
Week 8	• Data preparation is not linear. It's a process that is done multiple times.

Week 9	• Coding is more reliable than using commercial tools that need to be accessed through the company's website. Using commercial software that needs to be connected to the company's site can be challenging at times. Although SAS EM is a very powerful platform, my sessions timed out several times while in the middle of building a model and a few times the models had to be built from scratch.

Data Set Description

The datasets used for this project come from two different sources. The employment data come from the Quarterly Workforce Indicators (QWI) and it was retrieved from the U.S. Census Bureau. The housing data, which has the predictive variable, median housing prices, come from Zillow Research.

Initial EDA Findings and Rationale for Data subsets inclusion and exclusion

Initial findings from the EDA led to a reconsideration of data subsets for the analysis and predictive model. Originally, the median housing prices dataset, which represents the target variable, was retrieved by zip code. However, as it was pointed out in the previous report, one of the potential issues of using the housing data by zip code and employment data by cored based statistical area (CBSA), was a possible wide spread in the median housing prices within the same metro area, which was confirmed by the initial EDA done in Tableau.

A number of states showed that the same CBSAs have zip codes with median housing values that ranged from 200 thousand to over a million dollars. Thus, comparing and creating a predictive model for median housing prices by zip codes using employment statistics by CBSA codes would lead to a biased model and analysis, as they are providing measures for different geographical areas. For this reason, a new subset of

housing data from Zillow Research was retrieved by metropolitan areas to match the statistics provided by the Quarterly Workforce Indicators (QWI) datasets.

Additionally, almost 30% of the zip codes (132,825), once the dataset was in long format, had a 99999 CBSA code of indicating that they were unincorporated statistical areas (Missouri Census Data Center, n.d.). Hence, these data would have had to be removed as they wouldn't had been able to be merged with the employment data.

The following visualizations show a sample of the range of median housing prices by zip code in the same CBSA for 2015 in California, New York and Texas. For instance, in CA, the 21260 CBSA includes zip codes with median housing prices that go from $150,000 to over a million dollars

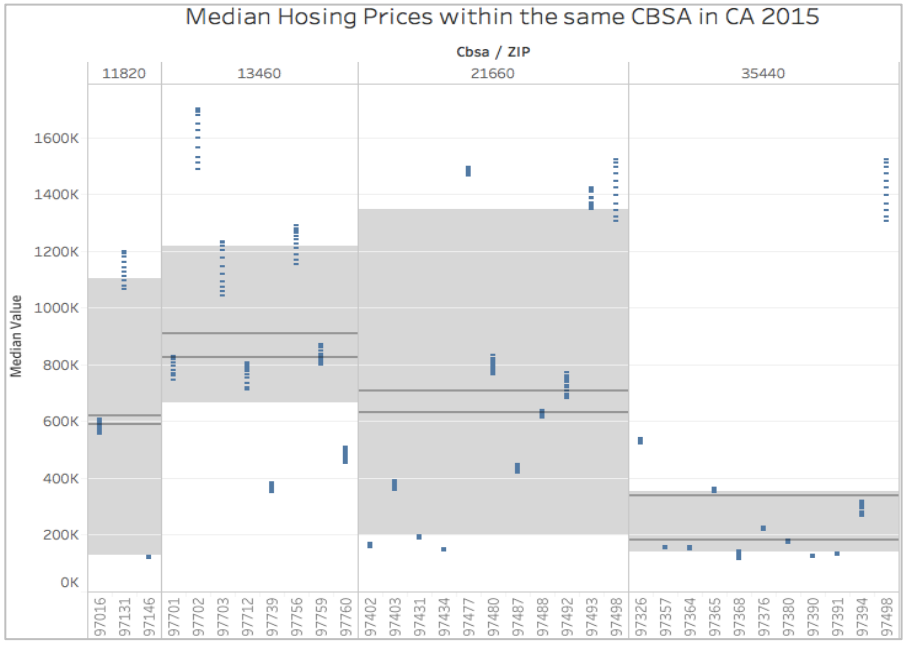


Figure 3. Median housing prices in CA by Zip code for three CBSA Codes, showing the wide range of values

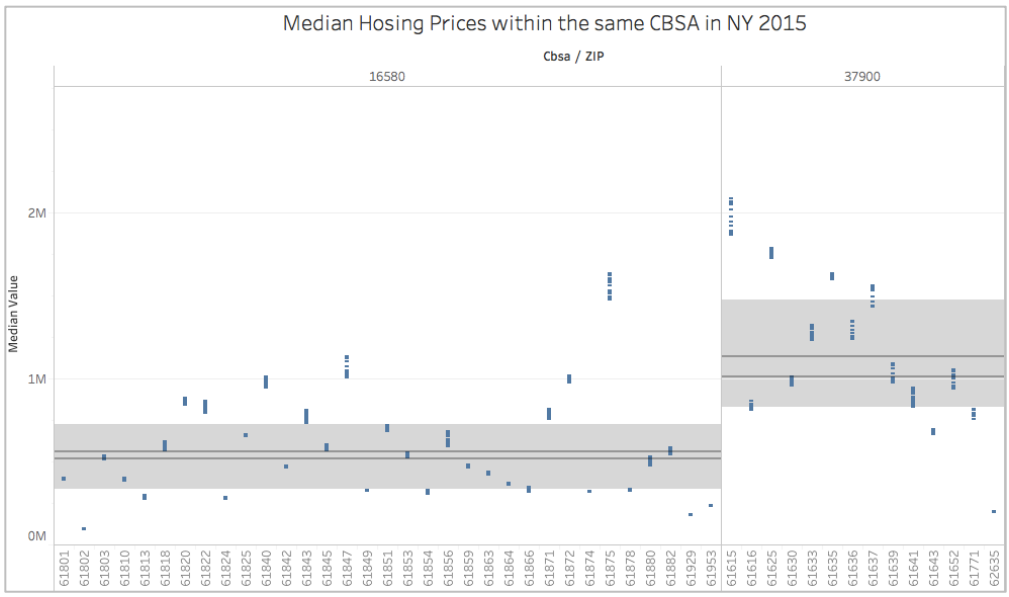


Figure 4. Median housing prices in NY by Zip code for two CBSA Codes, showing the wide range of values

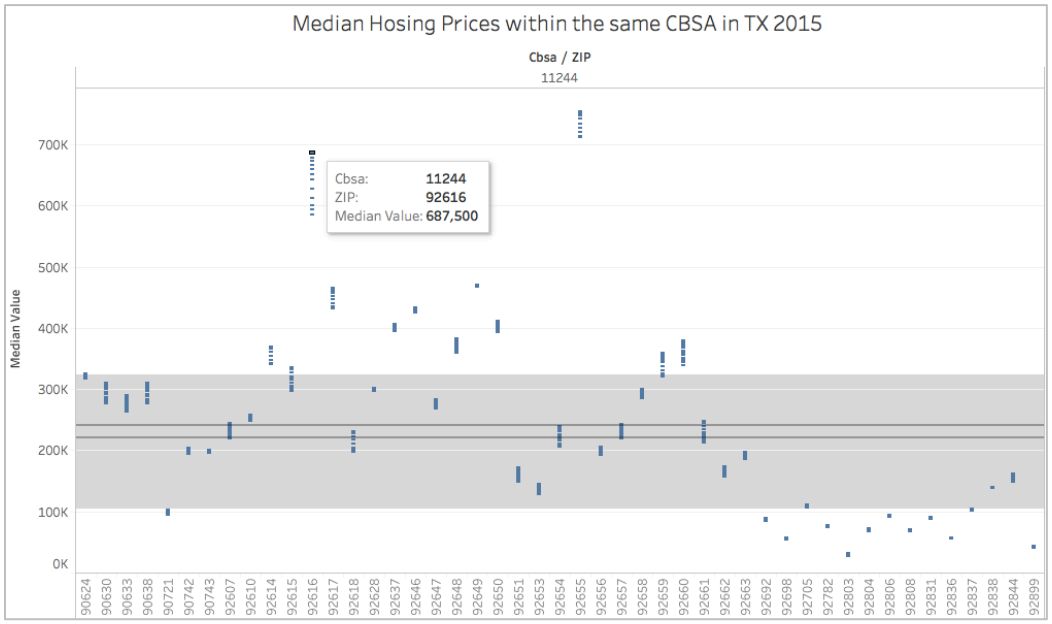


Figure 5. Median housing prices in TX by Zip code for two CBSA Codes, showing the wide range of values: from $29,400 in zip code 92803 to $755,000 in zip code 92655.

For the employment data, after completing a preliminary data restructuring seeking to transpose the age groups and the industry categories and join them with the employment indicator, the size of the dataset went from over 500,000 rows to 10,888 rows and from 30 attributes to 1,156 attributes. Since the number of records was

significantly reduced, additional years were added to have at least 20,000 records for the predictive model. Thus, the employment dataset will include employment statistics from 2012 to the third quarter of 2017.

	state	metropolitan statistical area/micropolitan statistical area	year	quarter	23 A03 Emp	23 A03 sEmp	23 A03 EmpS	23 A03 sEmpS	23 A03 EarnS	23 A03 sEarnS	...	72 A05 TurnOvrS	72 A06 TurnOvrS	81 A03 TurnOvrS	81 A04 TurnOvrS	81 A05 TurnOvrS
0	1	10700	2014	1	51.0	1.0	40.0	1.0	1808.0	1.0	...	0.104	0.081	0.173	0.130	0.055
1	1	11500	2014	1	39.0	1.0	32.0	1.0	2188.0	1.0	...	0.128	0.091	NaN	0.057	0.070
2	1	12220	2014	1	83.0	1.0	64.0	1.0	2064.0	1.0	...	0.132	0.099	0.271	0.100	0.090
3	1	13820	2014	1	1179.0	1.0	937.0	1.0	2584.0	1.0	...	0.125	0.086	0.169	0.111	0.079
4	1	17980	2014	1	30.0	1.0	25.0	1.0	1443.0	1.0	...	0.158	0.083	NaN	0.119	0.122

5 rows × 1156 columns

```
df_reset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 10888 entries, (1, 10700) to (9, 49340)
Columns: 1154 entries, year to 92
```

Figure 6. Screenshot showing number of records obtained after preliminary restructuring of employment data.

Datasets Overview

The Employment data come from the U.S. Census Bureau, Center for Economic Studies (U.S. Census Bureau, Center for Economic Studies, n.d.) and it's a collection of 51 Quarterly Workforce Indicators (QWI) time series datasets. One for each state and Washington D.C. The QWIs are retrieved by year, quarter, state, and metropolitan area and they are combined into a single dataset with employment indicators for metropolitan areas between 2012 and 2017. The datasets were retrieved via a combination of the LED Extraction Tool and the Census API. Although all the datasets have the same workforce attributes, the datasets retrieved with the Census API had different names and structure than the two datasets retrieved using the LED Extraction tool. Hence, additional formatting had to be done before being able to join all datasets vertically in RStudio.

The purpose of these datasets is to provide statistics on the following employment indicators by industry and metropolitan area: beginning of quarter counts, full quarter

counts, new hires, separations, stable hires, turnover rate, job gains at firms, job lost at firms, firm job net change in employment, average monthly earnings of employees with stable jobs, average monthly earnings for new hires, and total quarterly payroll. Hence, these datasets provide a variety of local workforce flow and employment indicators needed to analyze the relationship between employment and housing.

The main reason for using the QWI datasets is that their data come from the Longitudinal Employer-Household Dynamics employee microdata, which covers over 95% of U.S. private sectors jobs (U.S. Census Bureau, 2017). This ensures that the employment statistics used in this project are representative of the U.S. labor market.

The Housing Data come from Zillow Research and they are retrieved by metropolitan area. There are three time series datasets being used for analysis:

1. The **Zillow Home Value Index (ZHVI)** for all homes [(SFR, Condo, Co-op)], which contains the target variable, median housing prices by metropolitan area.
2. The **Zillow Rent Index (ZRI)** for all homes [(SFR, Condo, Co-op)], which has median rent prices by metropolitan area.
3. The **Conventional 30-year Fixed Mortgage Rates**.

The **ZHVI** dataset is published – and updated – monthly and it provides "seasonally adjusted measure of the median estimated home value" (Zillow, n.d) by state and metropolitan area between 1996-2018. The reason for using the ZHVI dataset is that median home values are based on Zillow's estimate value (Zestimate) of over 110 million homes in the U.S., which considers "special features, location, and market conditions" (Zillow, n.d.) of all homes, and not just the ones being sold, as well as inventory volatility. Thus, "the distribution of actual sale prices for homes sold in a given time

period looks very similar to the distribution of estimated sale prices for this same set of homes" (Zillow, 2014), making it easier to compare across time periods.

The **ZRI** datasets are published monthly and provide information on the monthly median rent price for all homes in a region from 2010 to 2018. For this project, the metropolitan area dataset is being used. The goal is to determine if there is a direct correlation between rent prices and housing prices and whether rent prices can affect home values. Similar to the ZHVI, the ZRI is a based on the estimated rental prices (Rent Zestimate) on all homes and not just the one being rented. The reason for using the ZRI for this project, is that just like the ZHVI, it is "unaffected by the mix of homes for rent at any particular time… and makes it easier to compare [with the ZHVI] since they are based on a similar set of homes" (Zillow, 2012).

Lastly, the **Conventional 30-year Fixed Mortgage Rates** is also published by Zillow Research and it provides information on mortgage rates *quoted on Zillow Mortgages* (Zillow, n.d.) for people with a 720-credit-score or better and are published on 15-minute increments between 2011-2018. The idea of using this dataset is to determine if mortgage rates could have an impact on housing prices. The assumption is that low-mortgage rates would encourage more people to buy a home and take advantage of the low rates, which would increase demand and possibly raise values.

Before merging the ZHVI and the ZRI datasets, they are reshaped from wide format to long format. This is done for two reasons. First, so the years and the monthly values become distinct attributes with each row representing the median housing and median rent price of a specific metropolitan area in a year-month. Second, to have the target variable, median housing prices, in a single attribute. This restructuring is done

using Python's pandas.melt() function. The median was used to combine monthly values into quarters. Then, the datasets are combined by location, quarter, and year using an inner join. The mortgage dataset is also reshaped to have a quarterly time series using the median and then it's merged with the rent and housing dataset on quarter.

The **Crosswalk files** are used to merge Zillow housing data with the U.S. Census Bureau employment data and they are retrieved directly from Zillow Research. Since the data from Zillow uses different region codes from the U.S. Census Bureau, these files provide a crosswalk between the Core Based Statistical Areas (CBSA) codes using in the employment data and Region codes used in housing data. Originally these files were retrieved from the U.S. Department of Housing and Urban Development when the housing data was retrieved by zip codes. However, when the housing data, both for median housing and median rent prices was changed, the crosswalk files provided by Zillow had the necessary information and were used instead. The change was also done because when attempting to do the inner merge with the median housing prices (ZHVI) dataset by zip codes, over 8,000 records were lost, as some of the RegionID codes used by Zillow were not recognized.

Since the crosswalk files are the direct connection between employment and housing data, they would be merged first with the housing data and then this single dataset will be merged with the housing data by zip code and year.

High-Level Data Diagram

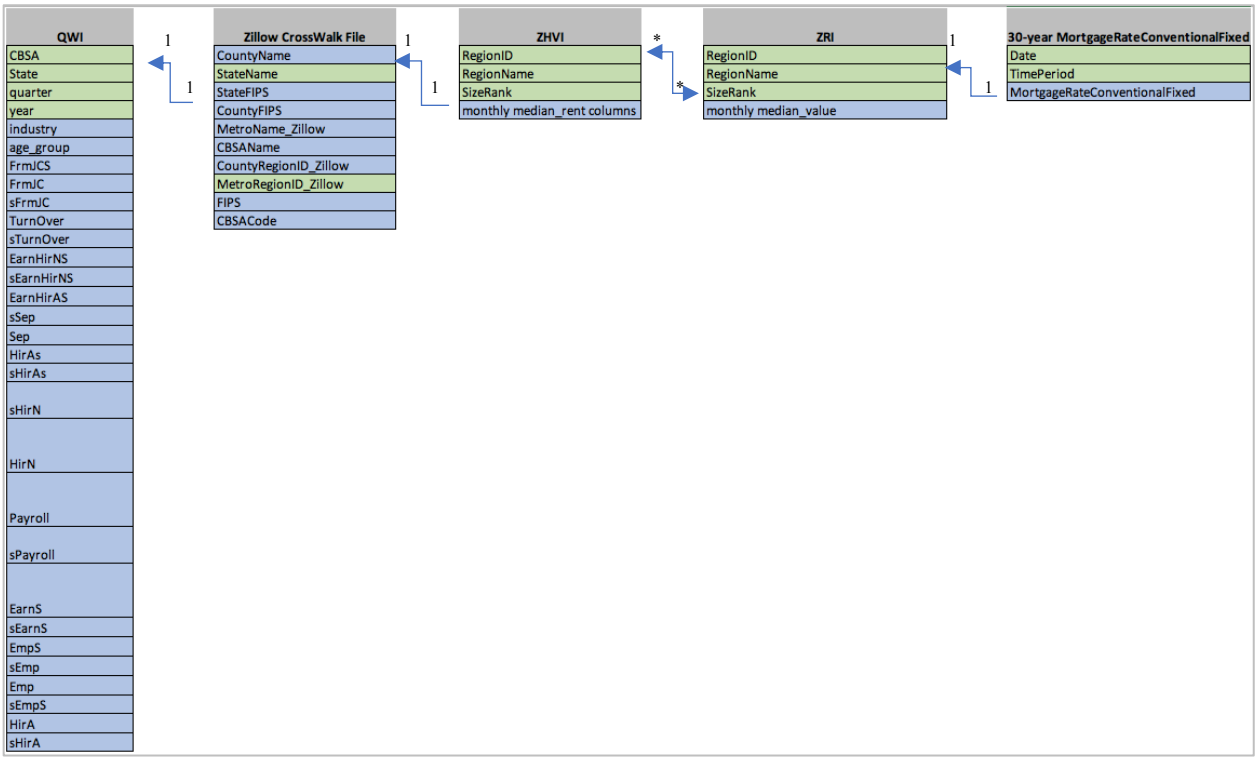


Figure 7. Relationship between datasets. Left to right: employment data (QWI), Zillow Crosswalk files, median housing prices dataset (ZHVI), median rent housing dataset (ZRI), mortgage dataset.

The housing values (ZHVI) and rent values (ZRI) datasets have four distinct attributes in common with the same name convention, data frequency and structure, so they can be easily joined. The conventional 30-year Fixed Mortgage Rates has a many to many relationships with ZHVI and ZRI. However, since the date and timePeriod attributes from the mortgage rate dataset have different frequencies than the attributes from the ZHVI and ZRI datasets, they need to be reshaped to have the same frequencies, so they can be merged. The crosswalk dataset with the MetroRegionID_Zillow and CBSA codes have a one to one relationship with the employment data (QWI): CBSA to

CBSA and one to one relationship with the housing data (ZHVI and ZRI): MetroRegionID_Zillow to RegionID.

Data Definition/Data Profile

Employment Data

The **employment data** is a collection of 51 Quarterly Workforce Indicators (QWI) time series datasets retrieved by state and metropolitan area from 2012 to 21017. Each dataset provides information on year, reference quarter, industry, state code, metropolitan statistical area/micropolitan statistical code (CBSA), workforce age by industry, and employment statistics by industry.

Each attribute has an accompanying variable, called flag status, with information on why values are missing or important information on the values released. This variable is helpful in identifying and inspecting missing and suspicious values and how to handle them. For instance, it will describe if a record is missing because a feature is not available for that period or is not applicable, hence, the record could be replaced by a 0, denoting no information (see Table 1 in Appendix for full description of all variables).

The combined datasets have 1,084,699 rows and 44 attributes. The workforce attributes provide information on two main employment measures:

1. Employment at the employee and at the company level. This measure includes count of jobs, hiring, separations, job creations, and job losses (Hayward & Tibbets, 2016).
2. Earnings. This measure includes payroll and average wages by industry and age group attributes (Hayward & Tibbets, 2016).

	emp	emps	earns	payroll	hira	hirn	hiras	sep	earnhiras	earnhirns	...	earnseps	frmjbgns	frmjblss	hiraendrepl	year	quarter	agegrp	industry	state	cbsa
0	51.0	39.0	1780.0	255345.0	9.0	9.0	8.0	15.0	1949.0	2219.0	...	1748.0	6.0	6.0	2.0	2012	1	A03	23	1	10700
1	237.0	208.0	2206.0	1609028.0	42.0	37.0	21.0	44.0	1719.0	1492.0	...	1768.0	14.0	11.0	10.0	2012	1	A04	23	1	10700
2	203.0	177.0	2426.0	1453073.0	27.0	22.0	18.0	30.0	1735.0	1751.0	...	2396.0	15.0	18.0	8.0	2012	1	A05	23	1	10700

Figure 8. Screenshot of sample view of employment dataset structure

The following table shows the summary statistics of the workforce indicators without the flag status variables. As shown by the skewness and kurtosis values, all variables have extreme values that are positively skewing the distribution of the data.

variable	n	mean	sd	min	max	range	se	Skewness	kurtosis	missing values
emp	1084512	1599.51	6533.99	0	265916	265916	6.27	12.881843	265.71	187
emps	1035257	1439.13	5926.36	0	249399	249399	5.82	13.034608	273.57	49442
earns	1035579	3207.8	2827.27	68	1323112	1323044	2.78	278.053334	115073.1	49120
payroll	1084699	21302881.55	124700616.2	0	15512062530	15512062530	119733.03	30.039181	2123.91	0
hira	1027676	256.05	1047.51	0	47441	47441	1.03	13.46645	284.93	57023
hirn	1016350	222.94	911.87	0	41620	41620	0.9	13.298597	279.27	68349
hiras	937971	137.67	550.93	0	31106	31106	0.57	13.303377	296.47	146728
sep	978392	242.89	995.72	0	45122	45122	1.01	13.432589	279.5	106307
earnhiras	1000234	2457.42	2605.51	0	1213685	1213685	2.61	237.885987	90340.28	84465
earnhirns	992744	2490.75	2422.33	0	1213685	1213685	2.43	246.081676	104711.5	91955
frmjbc	1036633	12.3	172.25	-10555	16178	26733	0.17	5.734077	598.05	48066
frmjbcs	1035338	8.71	155.28	-9558	14009	23567	0.15	10.854796	779.07	49361
emptotal	1084157	1842.83	7450.85	0	307170	307170	7.16	12.698985	256.13	542
hirr	870051	41.62	194.13	0	20738	20738	0.21	22.102925	900.57	214648
hirns	924418	121.9	489.09	0	28715	28715	0.51	13.267967	290.91	160281
seps	930732	129.03	516.41	0	29300	29300	0.54	13.096412	276.15	153967
earnseps	950813	2437.87	2907.55	1	1947294	1947293	2.98	368.369224	224362.2	133886
frmjbgns	1035338	74.1	306.13	0	20485	20485	0.3	13.763845	326.8	49361
frmjblss	1035338	65.39	270.17	0	16433	16433	0.27	13.477744	294.8	49361
hiraendrepl	1034847	79.5	361.75	-558	22350	22908	0.36	15.000568	356.07	49852
year	1084699	NA	NA	2012	2017	5	NA	NA	NA	0
quarter	1084699	2.43	1.1	1	4	3	NA	NA	NA	0
agegrp	1084699	NA	NA	NA	NA	NA	NA	NA	NA	0
industry	1084699	NA	NA	NA	NA	NA	NA	NA	NA	0
state	1084699	NA	NA	1	56	NA	NA	NA	NA	0
cbsa	1084699	NA	NA	10100	49820	NA	NA	NA	NA	0

Figure 9. Descriptive statistics for workforce employment indicators

The high kurtosis values show there is a large number of outliers pushing the tail of the distribution farther to the right, particularly in the earnings variables, which makes sense, since there is a larger number of employees whose monthly average earnings are less than $4000 and fewer ones that make more. One approach to deal with the skewness

and kurtosis is to apply a log transform to all the variables in SAS EM to make the data nearly normal before modeling.

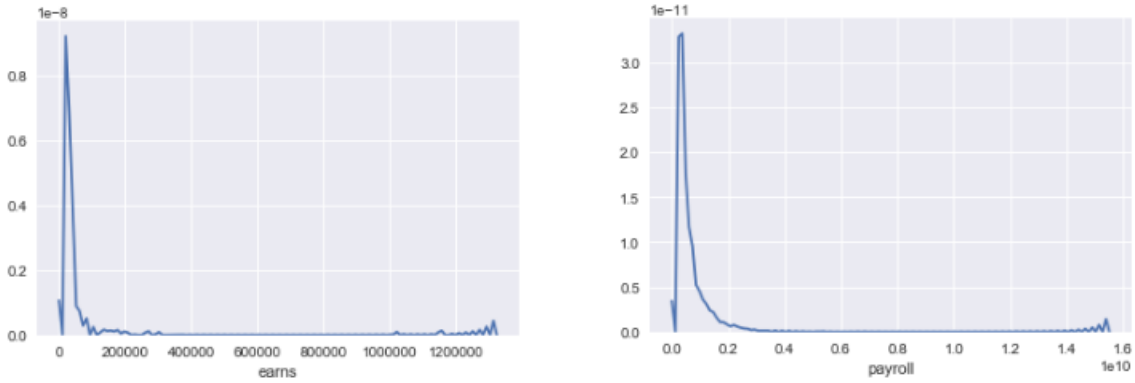


Figure 10. Density distributions showing the longer tails in the distributions for earnings (left) and payroll (right).

After looking at the summary statistics and plotting earnings and employment in Tableau, it was discovered that there were four observations that had average quarterly earnings of over 1M. Since they had a 9-flag status, saying that "Data significantly distorted, distorted value released" and they were all in state 48: Texas, they were set aside and omitted from the analysis.

The missing values will be imputed using the Amelia II package in R, which allows to do multiple imputation based on location and time trends, since this is a multivariate time series dataset and values need to be imputed accordingly.

Housing Data

The **housing data** includes three datasets: **Median Home Values (ZHVI dataset)**, **Median Rent Values (ZRI dataset),** and the **Conventional 30-Year Fixed Mortgage Rates**.

The Median Home Values (ZHVI) dataset, which includes the target variable, provides the monthly median home values for metropolitan area from 1996 to 2018. In its original wide format, each column represents the monthly median price. Hence, in order

to conduct the analysis and build the predictive model, all the value columns will be melted and transposed into a single column to become the target variable – median home values.

The dataset in long format has 208,278 rows representing U.S. metropolitan areas and 5 columns providing information on RegionID (Zillow's metropolitan area code), metropolitan area name, size rank, year and month of the median home value, and the median home value. In the original wide format, the dataset had 783 rows and 269 columns. However, as it was previously pointed out, in order to have the target variable in a single column, the data needed to be restructured.

	RegionID	RegionName	SizeRank	variable	value
0	102001	United States	0	1996-04	100600.0
1	394913	New York, NY	1	1996-04	165000.0
2	753899	Los Angeles-Long Beach-Anaheim, CA	2	1996-04	170600.0

Figure 11. Screenshot of sample view of the data structure in long format with median values (Target variable) in one column

	RegionID	RegionName	SizeRank	1996-04	1996-05	1996-06	1996-07	1996-08	1996-09	1996-10
0	102001	United States	0	100600.0	100600.0	100600.0	100700.0	100800.0	101000.0	101200.0
1	394913	New York, NY	1	165000.0	164800.0	164600.0	164300.0	164100.0	163900.0	163800.0
2	753899	Los Angeles-Long Beach-Anaheim, CA	2	170600.0	170400.0	170100.0	169800.0	169500.0	169300.0	169200.0

Figure 12. Screenshot showing data in original wide format with median values in different columns.

Since only housing values from 2012 to 2017 will be used to match the employment data time period, prior values are removed during the data preparation process and to provide summary statistics, leaving 113574 rows. The description of the variables is as follows:

Name	Description
RegionId	Zillow code for metropolitan area

RegionName	Name of metropolitan area
SizeRank	Rank size of population for metropolitan area
Variable	Variable indicating the month and year when the median housing value was reported
Value (Target variable)	Median housing value for each metropolitan area, measured monthly

The summary statistics for all median housing values for all states, show that the skewness is outside the normal range and the high kurtosis indicates there are several outliers that are pulling the tail of the distribution to the right.

vars	mean	sd	median	min	max	range	skew	kurtosis	se	missing	N values dataset
value	154267	87344	131500	35200	1069400	1034200	2.92	13.9	259.51	294	113574

Figure 13. Summary statistics of target variable: median housing prices

Since the skewness and kurtosis for the entire target variable is outside the acceptable range. A log transform will be performed before modeling.

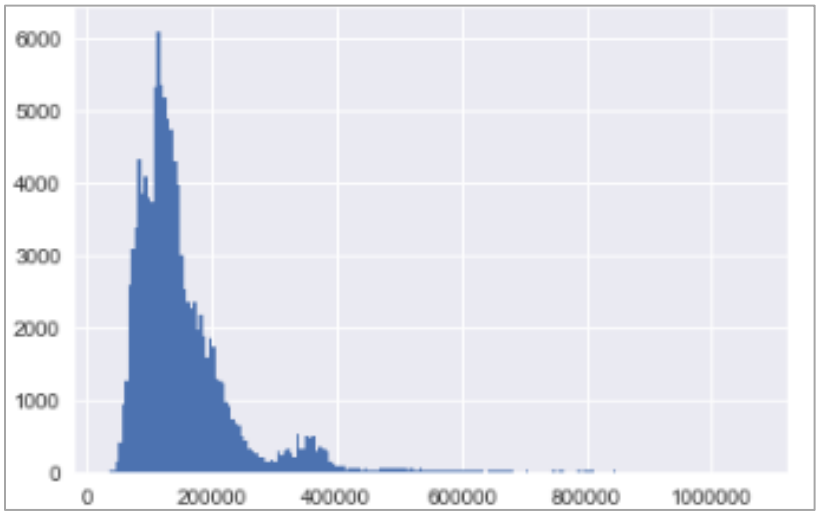


Figure 14. Distribution of median housing prices (Target variable)

Further examination by state, show that when looking at the skewness of median housing values by state, four states have the highest skewness: WY, FL, UT, and WV.

StateName	Skewness median home value
South Dakota	-0.673175
North Dakota	-0.565496
District of Columbia	-0.49245
New Jersey	-0.445234
Louisiana	-0.419347
Vermont	-0.382908
Hawaii	-0.335185
Missouri	-0.312118
Kentucky	-0.298934
Delaware	-0.201803
Alaska	0.015948
Arkansas	0.025261
Oklahoma	0.065738

StateName	Skewness median home value
Maryland	0.088071
Minnesota	0.110975
Maine	0.114704
Nebraska	0.168991
Iowa	0.190392
Georgia	0.295407
Wisconsin	0.34949
Ohio	0.360625
Kansas	0.361245
Indiana	0.455538
North Carolina	0.46949
Nevada	0.493302
Illinois	0.576013

StateName	Skewness median home value
Mississippi	0.610068
New Mexico	0.627942
Rhode Island	0.661107
Colorado	0.6772
Oregon	0.686437
Alabama	0.712766
Virginia	0.735212
Tennessee	0.756671
Texas	0.79579
California	0.90985
New Hampshire	0.928091
New York	0.949536
Washington	1.008098

StateName	Skewness median home value
South Carolina	1.039741
Arizona	1.054717
Michigan	1.060427
Massachusetts	1.283635
Montana	1.314661
Connecticut	1.708472
Pennsylvania	1.709969
Idaho	2.061925
Wyoming	2.299842
Florida	2.500632
Utah	2.503834
West Virginia	3.001554

The Kurtosis by states, also show there were four states with high kurtosis indicating the presence of a few outliers in PA, UT, WV and FL, with the last one having the biggest indication of outliers, which was confirmed by the histograms by states.

StateName	Kurtosis median home value
Maine	-1.743964
New Jersey	-1.412398
Alaska	-1.307271
Oklahoma	-1.207345
Kentucky	-0.864368
Missouri	-0.857483
Ohio	-0.806538
Vermont	-0.781816
Georgia	-0.776115
North Dakota	-0.770899
District of Colum	-0.739391
Hawaii	-0.735365
Nebraska	-0.727955

StateName	Kurtosis median home value
Louisiana	-0.726928
New York	-0.695181
Minnesota	-0.634808
Iowa	-0.610008
Rhode Island	-0.563828
Wisconsin	-0.557156
Arkansas	-0.501247
Delaware	-0.489734
Kansas	-0.433271
Maryland	-0.358487
Illinois	-0.30462
Oregon	-0.277116
North Carolina	-0.260113

StateName	Kurtosis median home value
Virginia	-0.226011
New Hampshire	-0.063922
California	-0.040741
Nevada	0.016991
South Dakota	0.09329
New Mexico	0.336711
Tennessee	0.363453
Indiana	0.529795
Alabama	0.573533
Arizona	0.57537
Washington	0.777898
Texas	0.800986
Colorado	0.811921

StateName	Kurtosis median home value
South Carolina	1.124964
Mississippi	1.521589
Michigan	1.58895
Massachusetts	1.660753
Connecticut	1.812901
Montana	2.365278
Wyoming	3.731506
Idaho	4.4639
Pennsylvania	5.866527
Utah	6.35631
West Virginia	9.422285
Florida	10.703563

Since there were 294 missing values, they were investigated by state. Using Tableau, it was discovered that the same metropolitan areas in three counties had missing values for the same periods of time: 2012, 2013 and 2014. Since there is so much missing data, they were removed. The rest of the missing values will be imputed using the average housing value for the same cities in the following quarters.

The second dataset used for housing is the **Median Rent Values (ZRI dataset)**. This dataset provides monthly data on median rent prices by metropolitan area from 2010 to 2018. Missing values will be imputed using the median rent value for the previous

quarter of the same metropolitan area. Highly skewed values will be evaluated by state to determine if they should be removed and analyzed separately. The median rent dataset in the long format has 208,278 rows and 5 columns.

	RegionID	RegionName	SizeRank	variable	value
0	102001	United States	0	1996-04	100600.0
1	394913	New York, NY	1	1996-04	165000.0
2	753899	Los Angeles-Long Beach-Anaheim, CA	2	1996-04	170600.0
3	394463	Chicago, IL	3	1996-04	138300.0

Figure 15. Screenshot of sample dataset structure in long format with rent values in one column

In the original wide format, the dataset had 659 rows and 94 columns. Each column representing the median rent value from 2010 to 2018. However, for the analysis in order to have all values in the same column, the data needed to be restructured and transposed.

	RegionID	RegionName	SizeRank	2010-11	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05
0	102001	United States	0	1254.0	1255.0	1254.0	1248.0	1245.0	1243.0	1245.0
1	394913	New York, NY	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 16. Screenshot of sample dataset structure in wide format.

The following table gives a description of the variables in the median rent dataset in long format:

Name	Description
RegionId	Zillow code for metropolitan area
RegionName	Name of metropolitan area
SizeRank	Rank size of population for metropolitan area
Variable	Variable indicating the month and year when the median housing value was reported
Value	Median rent value

Since only housing values from 2012 to 2017 will be used to match the employment data time period, prior values are removed during the data preparation process and to provide summary statistics, leaving 45470 rows.

	complete values	mean	sd	median	min	max	range	skew	kurtosis	se	missing vales	records in dataset
value	45166	1145	324.25	1086	569	3682	3113	2.55	10.52	1.53	304	45470

The skew and kurtosis values indicate that the distribution of the median rent is positively skewed and that there are several outliers that are pulling the right tail farther. In order to get a nearly normal distribution for the predictive model, this variable will be log transformed.

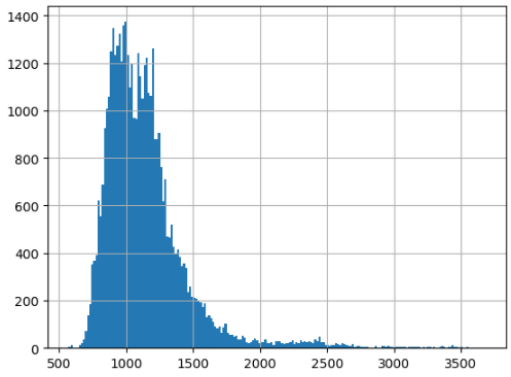


Figure 17. Distribution of median rent values.

Data visualization in Tableau showed that the same metropolitan areas had missing values for the same period of times and that one region in particular, Sikeston, MO, had missing values for 2012 – 2015. Hence, this record was removed as there was too many missing rent values.

The Conventional 30-Year Fixed Mortgage Rates dataset provides mortgage rates *quoted on Zillow Mortgages* (Zillow, n.d.) for people with a 720-credit-score or better and are published on 15-minute increments between 2011-2018. The description of the dataset is as follows:

- Total number of rows: 87, 627 rows representing "the average mortgage rate quoted on Zillow Mortgages for a 30-year, fixed-rate mortgage in 15-minute increments during business hours, 6:00 AM to 5:00 PM Pacific" (Zillow, n.d.).
- Total number of columns: 3 columns giving the date in yyyy-mm-dd format, the time in 15 min increments and the average mortgage rate.
- Attribute names and descriptions:

Variable Name	Description	Data Type
Date	Date of mortgage rate quote in yyyy-mm-dd format	date/factor
TimePeriod	Time of mortgage quote in 15-min increment	time/factor
MortgageRateConventionalFixed	mortgage rate	integer

Date	TimePeriod	MortgageRateConventionalFixed
2011-06-01	06:00	4.37
2011-06-01	06:15	NA
2011-06-01	06:30	NA
2011-06-01	06:45	4.38

Figure 18. Screenshot of a dataset sample view.

The descriptive statistics below show that the distribution of the data is nearly normal with a light left tail, by shown by the negative value of the kurtosis. The missing values are replaced by the average rate once the frequency of the data is changed from 15-min increment to quarterly.

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
MortgageRateConventionalFixed	23	3.84	0.28	3.83	3.84	0.26	3.38	4.42	1.04	0.29	-0.75	0.06

The **Crosswalk File** will be used to combine the employment data with the housing data, as they provide the crosswalk from Zillow's metropolitan codes to the CBSA codes used by employment data. The file was retrieved from Zillow research and it has 3144 rows and 10 columns. Since this file is only used to merge the employment and housing data and does not provide continuous variables for the model, no summary statistics were created. The County, StateName, MetroRegionID and CBSA codes will be

used in the final dataset to have a sense of the counties present in the same CBSA areas and compare housing prices and employment statistics by metro areas and counties.

	CountyName	StateName	StateFIPS	CountyFIPS	MetroName_Zillow	CBSAName	CountyRegionID_Zillow	MetroRegionID_Zillow	FIPS	CBSACode
0	Hudson	New Jersey	34	17	New York, NY	New York-Newark-Jersey City, NY-NJ-PA	1106	394913	34017	35620
1	Morris	New Jersey	34	27	New York, NY	New York-Newark-Jersey City, NY-NJ-PA	1241	394913	34027	35620

Figure 20. Screenshot with sample of dataset structure of crosswalk file

Variable Name	Description
CountyName	County name
StateName	State's full name
StateFIPS	Federal Information Processing Standard Publication (FIPS) numeric code
CountyFIPS	FIPS numeric county code
MetroName_Zillow	Zillow's full name of metropolitan area
CBSAName	Name of metropolitan area
CountyRegionID_Zillow	Zillow's county region ID code
MetroRegionID_Zillow	Zillow's metro region ID code
FIPS	Combination of state and county FIPS codes
CBSACode	U.S. Census bureau Core Based Statistical Area Code

Potential Quality Issues

One of the potential issues that was considered in assignment two was the integration of zip codes and metropolitan codes and how that could affect the median housing values or the relationship between housing and employment. After reviewing the median home values by zip code and employment data, there was in fact a wide spread in the median within the metro area. Hence, a new housing data subset was retrieved by metropolitan data also from Zillow Research.

Another issue has been character encoding, particularly when attempting merges in Tableau and exporting the files as .CSV files. Tableau uses a 'utf-16' encoding and not a 'utf-8' encoding, which was causing problems when reading the data back into R and Python, as some of the characters were not properly recognized by R or Python when

reading them. One way this issue has been solved is by resaving the file in ‘utf-8’ and enforcing the encoding when reading it.

A potential quality issue is missing employment data for some he states. Depending on the reasoning for the missing values (using the accompanying flag status data) these could be imputed using the median values of the previous quarter.

Data Preparation

There are multiple steps that need to be accomplished with each dataset before they are ready for analysis.

For the **median housing values dataset**, which has the target variable, the main task is to convert all the columns that have the monthly median housing prices into a single column. This is done to have the target variable, median housing prices, as a single attribute. Before this task can be accomplished, the first step is to remove the years that are not included in the analysis, so the data only contains median housing prices for all metropolitan areas from 2012-01 to 2017-09, since this is the time period for the employment data.

Then, the data is changed from wide format to long format in order to do EDA, visualizations, discover missing values, outliers, and obtain descriptive statistics for the median housing variable, which is the target variable. This step is done using Pandas’ melt function to combine all the columns with the median housing prices into one column and the time period into another column.

RegionID	RegionName	SizeRank	1996-04	1996-05	1996-06
102001	United States	0	100600.0	100600.0	100600.0
394913	New York, NY	1	165000.0	164800.0	164600.0
753899	Los Angeles-Long Beach-Anaheim, CA	2	170600.0	170400.0	170100.0

RegionID	RegionName	SizeRank	variable	value
102001	United States	0	1996-04	100600.0
394913	New York, NY	1	1996-04	165000.0
753899	Los Angeles-Long Beach-Anaheim, CA	2	1996-04	170600.0

Figure 20. Sample of dataset in original wide format (left) and dataset in long format after melting (right) with the value attribute containing the target variable.

After the EDA is done, the data will be again changed to wide format to impute outliers. This is done because it's easier to have all values in the same row to obtain the average median housing value for the same metropolitan area and being able to impute missing values with the average of previous or future months.

Once the imputation is done, the dataset is merged with the crosswalk files by 'RegionID' on the median housing dataset and 'MetroRegionID_Zillow' in the crosswalk file. The StateName, CountyName and CBSACode are left after the inner merge.

After the merge is complete and with the dataset still in wide format (with all the values in different columns), the time frequency is changed from monthly to quarterly using the median quarterly value to match the frequency of the quarterly workforce indicators. This process is done with pandas by first grouping the columns using the groupby and PeriodIndex functions and then melting the columns with the median prices once again to have the quarterly median values in a single column.

The same process is done with the **median rent housing dataset**. First, the years that are not considered in the analysis are removed and only rent prices for all metropolitan areas from 2012-01 to 2017-09 are left. The dataset is changed to long format in order to do EDA, obtain summary statistics, detect outliers, and missing values. Values are imputed with the dataset in long format and then the dataset is changed back to wide format to adjust the frequency of the monthly rent values to quarterly and melt the value columns, so they are in a single column.

For the **mortgage rates dataset**, the first step is to remove the dates that are not considered in the analysis, as it was done with the median housing and the median rent

datasets. The next step is to resample the data to turn the 15-min increment rates into quarterly rates using the average. This is done in pandas by first using the resample function with the mean.

For the **quarterly workforce indicators datasets,** after retrieving them using the census API and the LED extraction tool from the U.S. Census Bureau, the first step is to combine them horizontally in order to have a single dataset. Although, the same attributes were selected using the API and the LED extraction tool, the two states that were retrieved using the LED tool (WY and WI) needed to be renamed and rearranged as the order and name of the attributes did not match the states. This was done in RStudio. Then, all the datasets were combined.

Since it was necessary to obtain additional years (2012, 2013, and 2017). The retrieval was done by year for all states to improve retrieval time using the API. Hence, there were eight total datasets combined. The next step was to look at the correlation between variables to decide which variables were highly correlated or didn't provide new information, so they could be removed from the analysis.

The next step is to transpose the age and industry categories, so they are tied to each employment indicator and become input variables. This is done in JypyterLab using pandas groupby, reindex, stack and unstack.

Data Cleansing

The data cleansing will be done separately for each dataset. For the median housing dataset, the first step is to impute the missing values using the average median housing prices for the same metropolitan area in different quarters, this ensures that the imputed values are based on time trends for the same area. Based on visualizations done

in Tableau, it was discovered that the same three counties in the same metropolitan areas had the 30 missing values for the same period of time 2012, 2013 and 2014: 'Marion', 'Manitowoc', ‘Grant'. Since there were three years in a row with missing data and they appeared to be missing completely at random, they were removed from the analysis. The imputation of the rest of the 204 missing values is done in RStudio.

The same process is used for the median rent values dataset. Using Tableau, it was discovered that Sikeston, MO had 46 missing values in total for 2012 to 2015. Since there were three years in a row with missing values and they appeared to be missing completely at random, they were removed from the analysis. The imputation of the remaining 258 missing values was done using the average of the median rents for the same metropolitan area in different quarters to ensure the imputation was based on the same geographic location and followed times trends. The imputation was done in RStudio.

For the mortgage rate, the missing values were replaced using the average rate. There were no unusual values and the imputation, which also helped restructured the dataset into quarters, was done using Pandas.

There was no need to do any pre-processing to the crosswalk files as they only provide names and codes to merge the employment and housing datasets.

For the employment dataset, using Tableau, it was discovered that Texas had some observations with average monthly earnings over 1 million dollars. Since the flag status indicated the “Data [was] significantly distorted, distorted value released”, these values were omitted from the main analysis. For the missing values, the Amelia II package in R will be used for imputation as it will allow to do the multiple imputation

using time trends for the same locations and all the available variables. Hence, the imputation will be done after the age and industry categories had been merged with the other workforce indicators. One of the reasons for using the Amelia package in R is that it allows to do the imputation before the transformation is done, by accepting "a natural logarithm transformation of [the variable with missing values as an argument] in order to normalize its distribution" (Honaker, King, & Blackwell, 2018, p. 19).

Another important cleaning step was to examine correlations and removed variables that are highly correlated. Since there is only one numeric variable in the median housing dataset, which is the target variable, a correlation matrix was run on the quarterly workforce indicators. The matrix showed there were several correlation coefficients close to 1, showing a strong positive linear correlation between the variables. Some of the correlations were expected. For instance, the correlation coefficient between beginning of quarter employment and stable employment is 0.998, which makes sense, as the last measure is counting jobs that are present the previous and the current quarter. Hence, emps is removed.

	emp	emps	earns	payroll	hira	hirn	hiras	sep	earnhiras	earnhirns
emp	1.000000	0.998799	0.106346	0.821811	0.880947	0.866943	0.923457	0.880278	0.076883	0.081419
emps	0.998799	1.000000	0.111442	0.828996	0.859254	0.843120	0.908163	0.857003	0.080269	0.085132
earns	0.106346	0.111442	1.000000	0.181246	0.047442	0.042034	0.059973	0.046025	0.678985	0.733271
payroll	0.821811	0.828996	0.181246	1.000000	0.637155	0.613559	0.681457	0.633206	0.137149	0.138846
hira	0.880947	0.859254	0.047442	0.637155	1.000000	0.993469	0.951258	0.985974	0.039780	0.041818
hirn	0.866943	0.843120	0.042034	0.613559	0.993469	1.000000	0.950602	0.983683	0.036753	0.038206
hiras	0.923457	0.908163	0.059973	0.681457	0.951258	0.950602	1.000000	0.944690	0.052316	0.052984
sep	0.880278	0.857003	0.046025	0.633206	0.985974	0.983683	0.944690	1.000000	0.037100	0.038862
earnhiras	0.076883	0.080269	0.678985	0.137149	0.039780	0.036753	0.052316	0.037100	1.000000	0.898943
earnhirns	0.081419	0.085132	0.733271	0.138846	0.041818	0.038206	0.052984	0.038862	0.898943	1.000000
frmjbc	0.243030	0.249222	0.020835	0.187021	0.353330	0.327040	0.298182	0.192814	0.026001	0.028405
frmjbcs	0.199887	0.208563	0.017955	0.147434	0.190452	0.183531	0.362764	0.115466	0.028352	0.024248
emptotal	0.997900	0.993970	0.099850	0.808238	0.909792	0.896622	0.940464	0.907277	0.072880	0.077126
hirr	0.758974	0.751571	0.118853	0.615841	0.814676	0.742687	0.745680	0.781697	0.070493	0.089062
hirns	0.913533	0.896957	0.057572	0.669543	0.949163	0.954433	0.992382	0.943222	0.050100	0.052270
seps	0.925427	0.906285	0.065726	0.682945	0.958076	0.959570	0.955121	0.974971	0.049547	0.052190
earnseps	0.077299	0.080962	0.661161	0.141218	0.035411	0.032400	0.044612	0.034624	0.482249	0.529618
frmjbgns	0.895016	0.883226	0.066750	0.676013	0.907896	0.902940	0.978891	0.892275	0.059470	0.059300
frmjblss	0.899247	0.880901	0.065312	0.681243	0.919104	0.917437	0.900001	0.944591	0.051070	0.053235
hiraendrepl	0.870233	0.847756	0.035362	0.598800	0.967771	0.975164	0.945250	0.967770	0.028409	0.029074

Figure 20. Partial output of correlation matrix for quarterly workforce indicators dataset.

Data Transformation

There were two additional features that needed be created for the mortgage dataset, the median housing dataset, and the median rent dataset in order to merge them with the employment data for accuracy and to match not only with the CBSA code, but the corresponding year and quarter. These were year and quarter.

For the Median Housing Price dataset, this was accomplished after the values were in long format and in a quarterly basis, using pandas. The yr_qtr column was converted to string and then it was parse into two columns: year_str and quarter, as shown below:

MetroRegionID_Zillow	FIPS	CBSACode	yr_qtr	median	year_str	quarter
394913	34017	35620	2012Q1	336800.0	2012	1

Figure 21. Screenshot of Median Housing Prices Dataset after adding a year and quarter features.

Just like with the Median housing prices dataset, the year and quarter features for the Median Rent prices, were done using pandas.

RegionID	RegionName	SizeRank	yr_qtr	median	year_str	quarter
394913	New York, NY	1	2012Q1	2022.0	2012	1
753899	Los Angeles-Long Beach-Anaheim, CA	2	2012Q1	2217.0	2012	1

Figure 23. Screenshot of Median Rent prices dataset after adding a year and a quarter feature.

For the Mortgage Rate this was accomplished using Pandas, by converting the date column into a string and splitting the string at the "-", which separated the year and the month. Then a function was created to map the last month of the quarter with the quarter number, which went to a different column. Hence, the final dataset looks like this:

	Date	MortgageRateConventionalFixed	year_str	month	quarter
0	2012-03-31	3.854417	2012	03	1
1	2012-06-30	3.738913	2012	06	2

Figure 24. Screenshot of Mortgage Rate Dataset after adding a year, month and quarter features.

The transformation of the quarterly workforce indicators was more involved as it required converting each category of the age and industry features into individual attributes that could be combined to each employment indicator. The idea was to have the workforce indicator statistics (earns, hires, separations, etc.) for each age group and industry in a single column to use them as input attributes in the predictive model. Thus, using pandas, the dataset went go from having one age group and industry feature, as shown below:

earnseps	frmjbgns	frmjblss	hiraendrepl	year	quarter	agegrp	industry	state	cbsa
1748.0	6.0	6.0	2.0	2012	1	A03	23	1	10700
1768.0	14.0	11.0	10.0	2012	1	A04	23	1	10700
2396.0	15.0	18.0	8.0	2012	1	A05	23	1	10700
2236.0	14.0	9.0	4.0	2012	1	A06	23	1	10700
1825.0	181.0	12.0	76.0	2012	1	A03	31-33	1	10700

Figure 25. Quarterly workforce indicators dataset showing age and industries as independent variables.

To having each industry and age group tied to each indicator:

year	quarter	state	cbsa	23A03emp	23A03earns	23A03payroll	23A03hira	23A03hirn	23A03hiras
2012	1	1	10700	51.0	1780.0	255345.0	9.0	9.0	8.0
2012	1	1	11500	37.0	1852.0	208296.0	14.0	12.0	6.0
2012	1	1	12220	118.0	2065.0	710765.0	38.0	37.0	7.0
2012	1	1	13820	1256.0	2473.0	9441351.0	464.0	420.0	177.0
2012	1	1	17980	36.0	1392.0	153130.0	16.0	10.0	6.0

Figure 26. Same dataset showing age and industries as part of the employment indicators.

The new employment dataset went from having 1,035,575 rows and 18 to having 21,736 rows and 292. Hence, as it was considered in the previous report, in order to reduce the number of attributes and simplify the model, the age groups were added by industry and indicator. Hence, instead of having four attributes for each indicator providing the statistics for each age group, a single attribute was created. Thus, all the

indicators were now by industry and all age groups, reducing the size of the data set to 21, 736 rows and 64 columns.

	year	quarter	state	cbsa	Emp23Total	Ears23Total	Payroll23Total	jblossTotal	jbgainTotal	Emp31_33Total
0	2012	1	1	10700	676.0	9380.0	5027993.0	44.0	49.0	8261.0
1	2012	1	1	11500	626.0	11390.0	5775010.0	54.0	45.0	4614.0

Figure 27. Same dataset showing all ages combined for industries and employment indicators.

Once the employment data was ready, the employment and the housing datasets were merged. At that point, eight different features were created to identify each metropolitan area's economic region as defined by the Bureau of Economic Analysis. Each indicator had a 1 if the metropolitan area was part of the region and a 0 otherwise.

Hence, the new features allowed to segment the data in eight different regions that shared similar "economic activity and growth" (Bureau of Economic Analysis, n.d.). Each region was then used to create different models to predict median housing prices within similar regions and identify which industries drive median home prices (target variable). The regions are New England, Mideast, Great Lakes, Plains, Southeast, Southwest, Rocky Mountain, and Southwest.

Data Analysis

As it has been previously mentioned the main tools used for this analysis and the predictive model are Pandas (Python), R in RStudio and Tableau, Tableau and SAS EM.

R in RStudio has been used for data retrieval, summary statistics and imputation of missing values for all the datasets. R in Tableau has been used to identify outliers and visual analytics that helped uncover the outliers and missing

values by state in the housing datasets and the outliers in the workforce indicators datasets.

The main reason for using R is that since it's a dedicated statistical language, it provides a vast number of libraries that make statistical analysis and EDA quick and simple. Even with the basic packages obtaining descriptive statistics and visualizations for the housing and employment data was simple, yet insightful.

One of the main tasks accomplished in R was retrieving the employment data using the census package, which uses the census API. A task that I could not accomplished in Python. With this package, additional years were added and variables that were identified as having a large number of missing values in the original subset, were retrieved. A simple function was used to read each year into a list and then combine all metropolitan areas for that year.

R is also being used for imputing missing values. As it was previously described, for the housing data, the average median housing value for the same cities from other quarters is being used. For the employment data, the Amelia II package is used for the imputation. The visualizations done in R have been mainly as part of the EDA to examine the distribution of the datasets and investigate outliers and missing values.

Python in JupyterLab has been the language and programing environment for data retrieval, manipulation, restructuring, and EDA as it allows to write code, generate output, write markup text, and make visualizations, all in a single setting. The main reason for using Python is that it's faster than R and data cleaning and

manipulation with Pandas is very easy even with large datasets. The main tasks accomplished with Pandas have been:

1. The restructuring of all the housing datasets to convert them into long format and change the time frequency.

2. The restructuring of the quarterly workforce indicators to transpose the age and industry categories to tied them to each employment indicator.

Python has also been used for data visualization with the Matplotlib and Seaborn libraries. The reason for using these libraries is that the visualizations are highly customizable and they offer a wide range of build-in statistical plots to examine the distribution of the data.

Tableau has been used for visualizations and EDA. The reasons for using Tableau are that it offers a very simplistic UI that allows to create quick visualizations and discover patterns with simple drag-drop features. And it provides mapping capabilities that allow data analysis using geographic and time filters. This is a very useful feature as it gives a better idea of how median housing prices and employment trends are distributed across the U.S. in different periods of times.

Several visualizations will be done for the median housing prices, median rent prices and workforce indicators. For the quarterly workforce indicators, visualizations are done to identify outliers and to examine employment trends by age and industry in different metropolitan areas. Using the mapping function, this last visualization will help identify if there are particular industries that thrive in some regions.

For the median housing prices, mapping will be done to show prices by metropolitan area by year. The values will be combined into five bins and the filter

feature will be used to show values by year to show how prices have changed and if there is concentration of higher prices in certain areas. A preliminary visualization was done before imputing values and it's shown below.

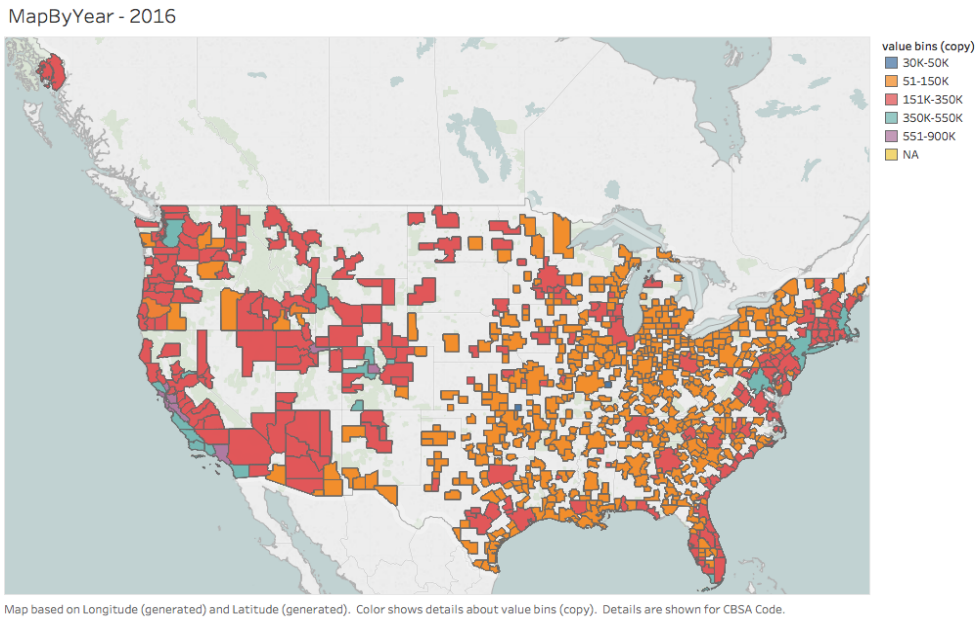


Figure 27. Map visualization of median home price across metropolitan areas in 2016

SAS EM will be used for transformations and predictive modeling. The transformation of the skewed variables for all datasets was done in SAS as well as building the predictive models. The reasons for building the model and doing the transformations here is that it has a very easy and intuitive interface and it offers multiple tuning parameter for both transformations and modeling, making it easier and faster to build and compare models. The creation of the models will begin in week 9 after all the transformation and imputation has been completed.

Data Visualizations

Data Visualization 1 – Median Housing Prices within CBSAs

Among the first visualizations created there were several used to compare the spread of median housing prices by zip codes within the same Cored Based Statistical

Area. The visualizations were done because a potential issue considered at the beginning of the analysis was a possible wide range of values within the same CBSA, which is the measured used for the employment statistics. As previously discussed, the spread was analyzed for different states and year.

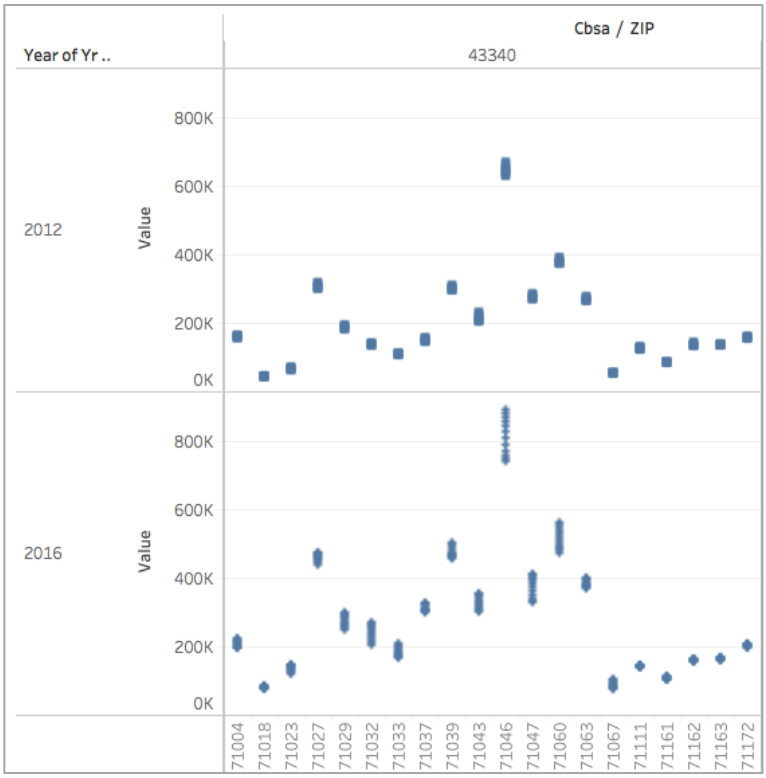


Figure 28. Visualization of median home prices spread in the same CBSA for GA in 2012 and 2016.

The above visualization shows the wide ranges in median home prices in 2012 and 2016 for two statistical areas in Georgia. The same statistical area has a zip code 71018, which median home price in 2012 was $46,000 and $87,500 in 2016. Along with a zip code, 71046, with a minimum median price of $629,900 in 2012 and $741,600 in 2016. Since the median homes values are provided by zip codes and the employment indicators are given by metropolitan areas, the analysis wouldn't be considering the same measurements as one takes the median price for home within each zip code and the other (employment) takes the measurement for the whole statistical area. Although these

findings didn't' alter the scope of the analysis, which is predicting median home prices using employment indicators, they led to reconsider the subset of housing data being used.

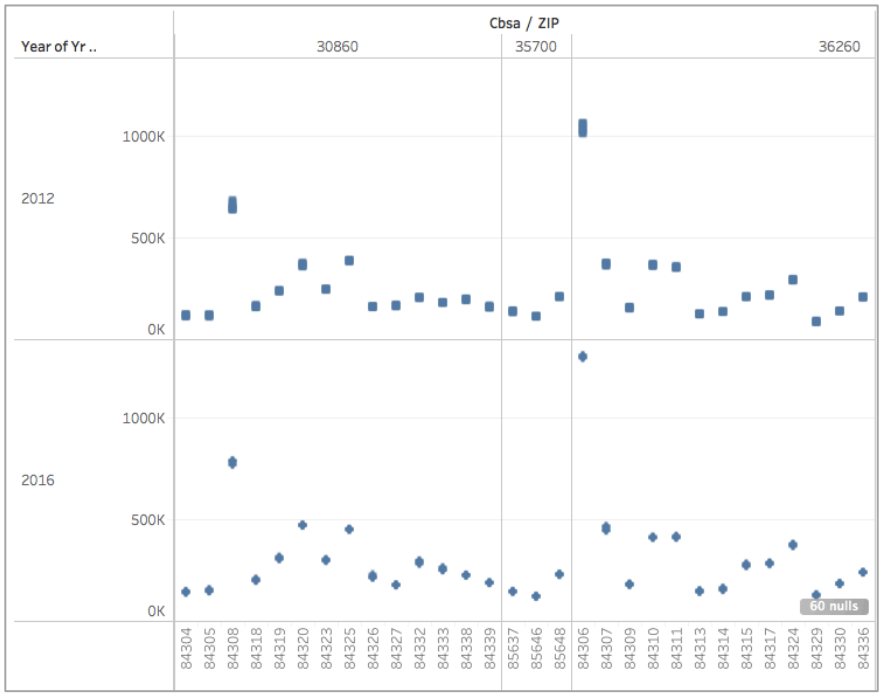


Figure 29. Visualization of median home prices spread in the same CBSA in IL in 2012 and 2016.

The same analysis was done for Illinois. The graph shows two statistical areas with prices that range from just over $100,000 to over 1 million dollars. These visualizations show that the same statistical area can have a very different socio-economic makeup and the quarterly employment indicators will not reflect or provide a reasonable measure to predict median home prices for metropolitan areas based on historical data for home values by zip code. Since Zillow provides median home values for metropolitan areas, this measure will be used instead.

Data Visualization 2

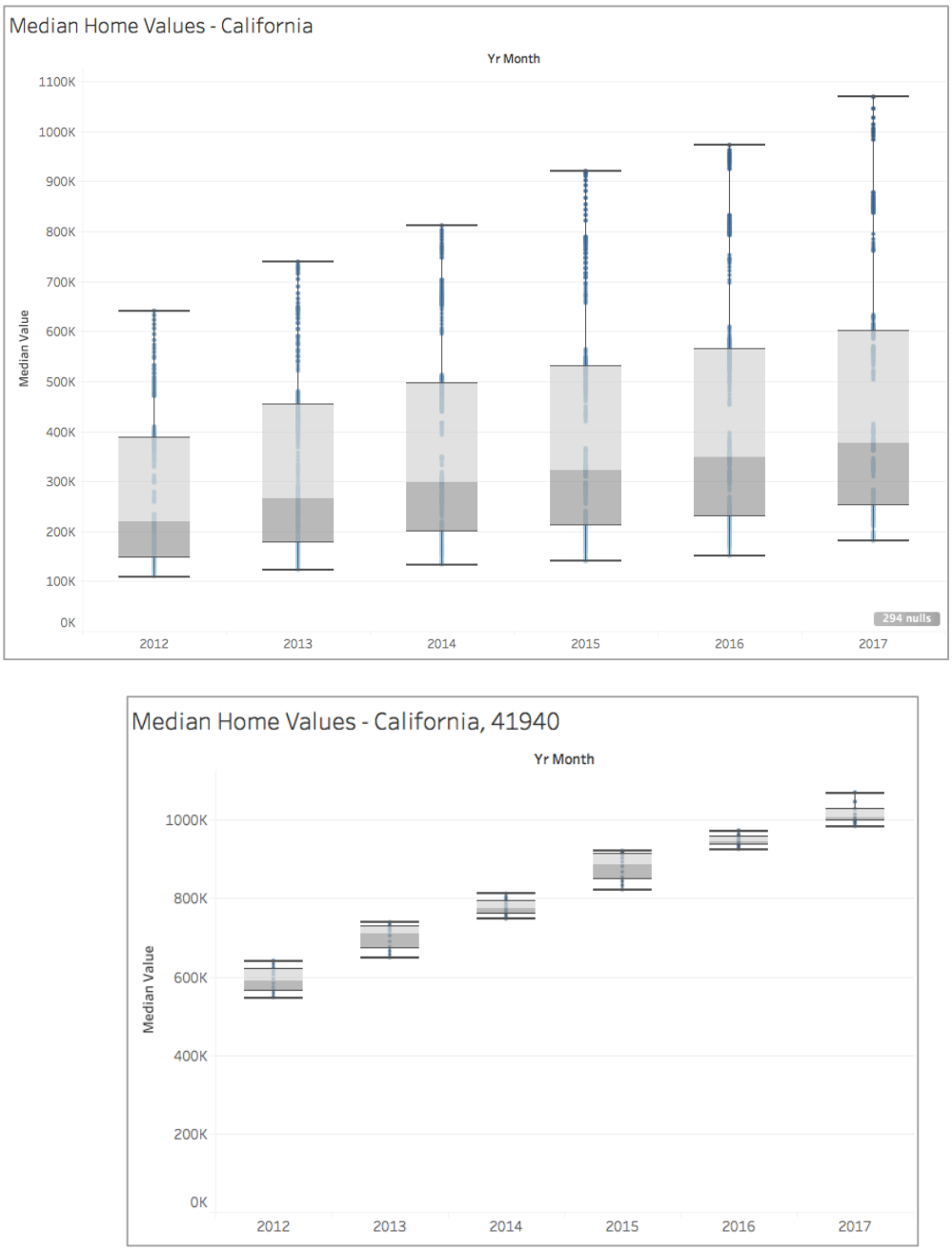
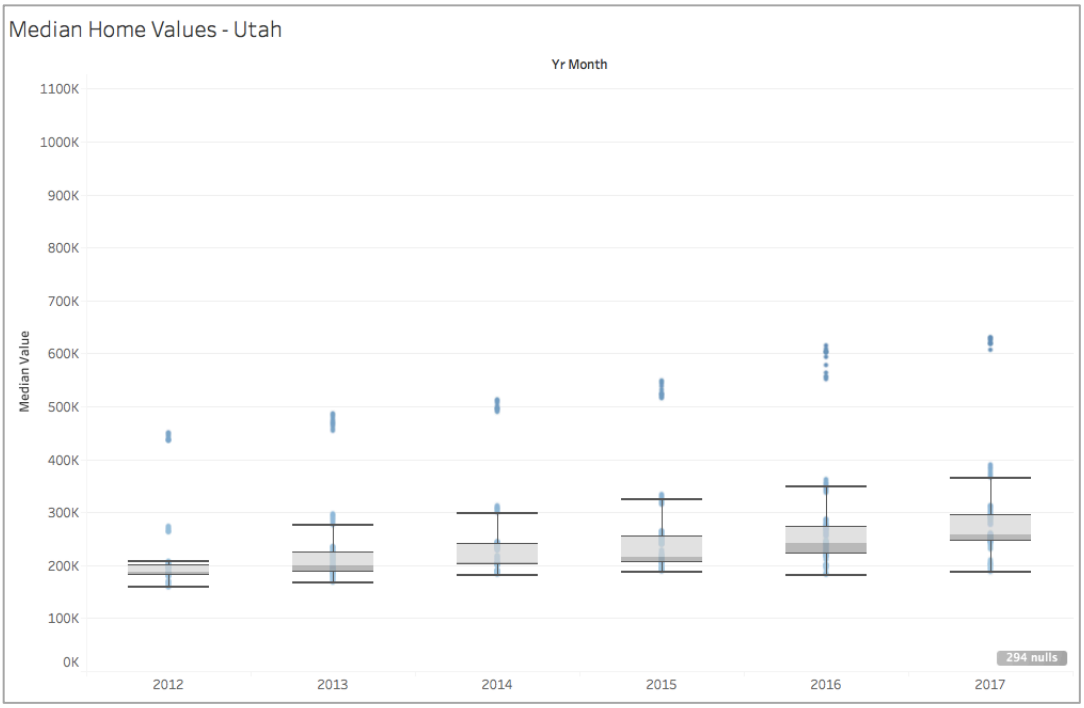


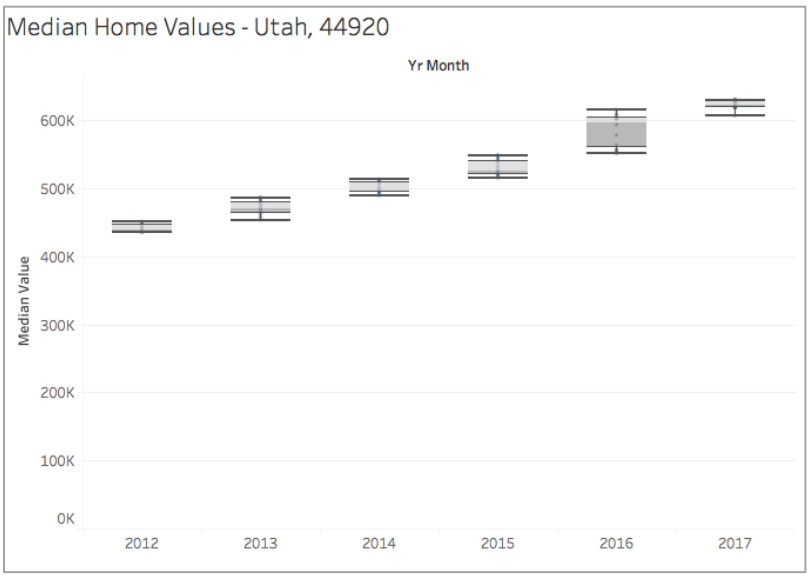
Figure 30. Top – Median housing prices by year in California. Bottom – Median home prices by year in San Jose-Sunnyvale-Santa Clara, CA 41940 statistical area

With the new median home prices dataset by metropolitan areas, one visualization was done in Tableau to help uncover outliers by year and state and to verify the distribution of the data. Looking at the distribution of median home prices in California,

it's clear that an upward spread of values has gained momentum since 2013. In 2012, 75% of the median home values fell between $146,000 and $386,000. In 2017, the spread was much greater, with 75% of the median values falling between $251,200 and $601,100. Median home prices overall have also accelerated, going from $267,000 in 2013 to over $370,000 in 2017. This is particularly visible in metropolitan areas like San Jose-Sunnyvale-Santa Clara, where the median values have gone from $773,000 in 2014 to over 1 million dollars in 2017. This shows how areas that have seen an increasing presence of technology companies, have also seen a significant surge in median home values.

The visualizations also show, that although the data is positively skewed, there are no extreme outliers, as seen by both the skewness and kurtosis coefficients obtained for the overall distribution from 2012 to 2017, which have a skewness of 0.909 and a Kurtosis of -0.040741.

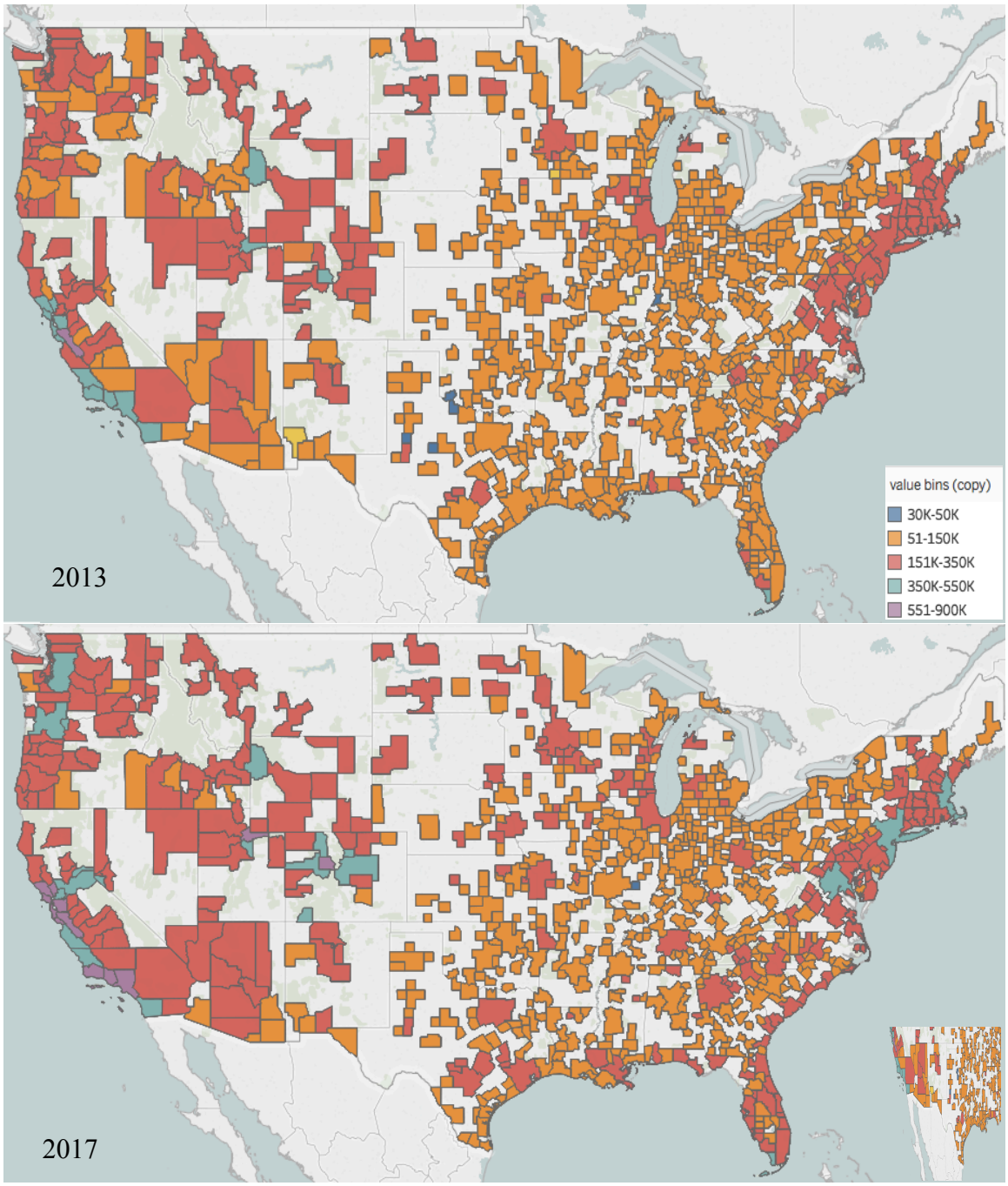




Other states like Utah, show less variance in the median home prices, but there is an increasing number of extreme values that are skewing the distribution. For instance, the 44920 CBSA code, which corresponds to Summit Park, UT, saw an increased in median home price from $525,00 in 2015 to $621,000 in 2017. This reflects a housing crisis that the state is currently living, as "the median sales price of a home in Utah's two largest metropolitan areas is already 20 percent higher than home prices in Boise, Las Vegas and Phoenix" (Gorell, 2018). These changes have been fueled in part by the migration of millennials who are drawn by strong economy and the increasing presence of big companies like Delta and eBay (Olick, 2018).

What these previous visualizations shows is a growing trend of rising median home prices, particularly in the west region. In order to investigate this further, another visualization was done, this time using the map functionality.

Data Visualization 3

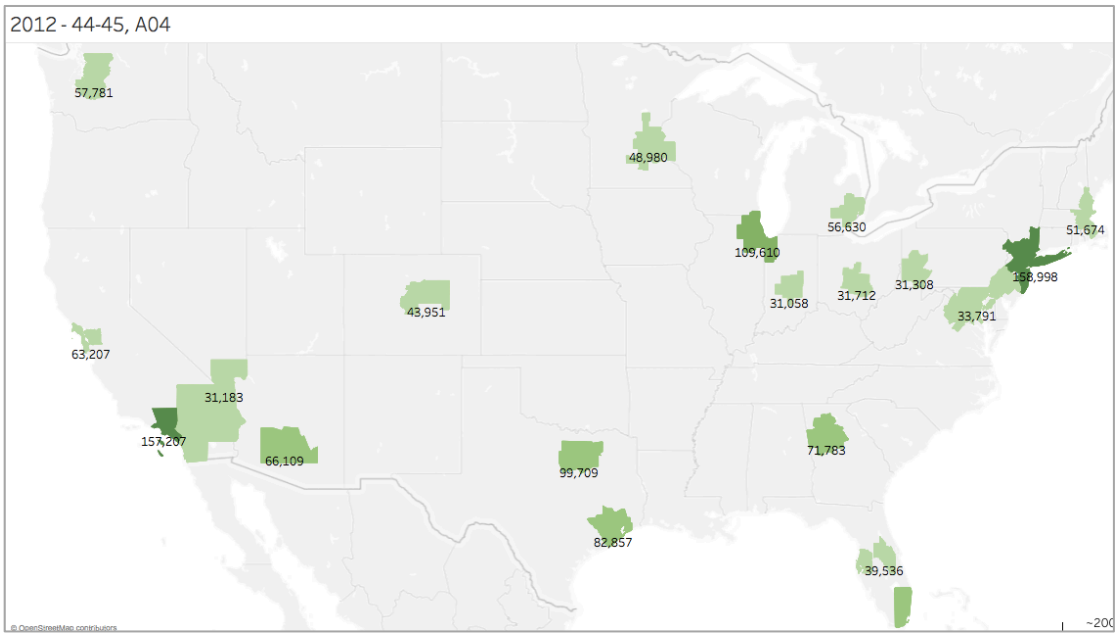


The maps above show the rapid change in median housing prices across metropolitan areas in the U.S. between 2013 and 2017, particular in the North East and West Regions, where over 80% of the median home prices are over $151,00 in 2017. The visualization shows a socio-economic change in the makeup of the regions, which could

be explained by a growing migration due to economic growth fueled by the two major industries that predominate in these areas: technology (west region) and finance (northeast region). In general, this shift can provide further insight on how increasingly, the "East Coast and West Coast disproportionately contribute to the total economic output of the US" (Investopedia, n.d.) and the dominant industries are helping fuel the surge in median housing prices.

Data Visualization 4

One more visualization was created using the quarterly workforce indicators to identify the industries with the largest number of employees by year and metropolitan area. There were two industries, Retail Trade (44-45) and Health Care and Social Assistance (62), that had both (1) a wide spread presence across the country and (2) a large number of employees in coastal cities. For illustration purposes in this report and to emphasize the areas with the largest number of employees, only metropolitan areas that had over 32,000 employees are shown.



The visualizations show how Retail Trade saw an increase in employment numbers across all metropolitan areas between 2012 and 2015. However, between 2015 and 2017 the number of total employees only increased in major metropolitan hubs, with a decrease in other parts of the countries. This was seen particularly in five metropolitan areas, which also saw a surge in median home prices in 2017:

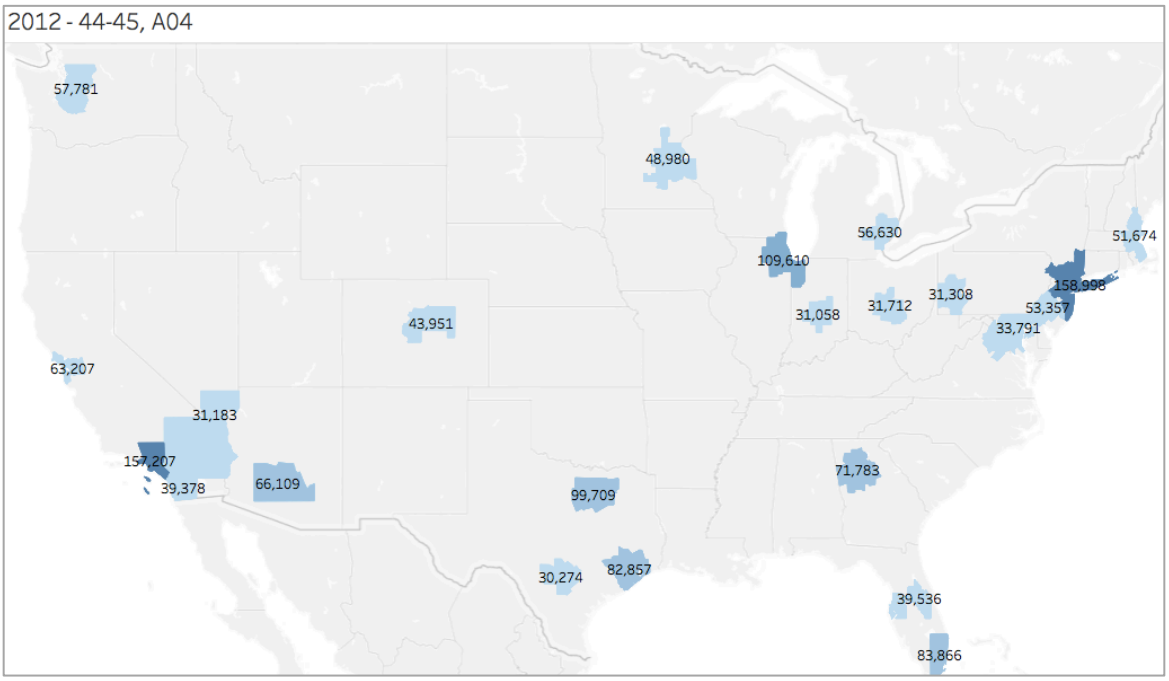
- Los Angeles-Long Beach-Anaheim, CA;
- New York-Newark-Jersey City, NY-NJ-PA;
- Chicago-Naperville-Elgin, IL-IN-WI;
- Dallas-Fort Worth-Arlington, TX;
- Atlanta-Sandy Springs-Roswell, GA.

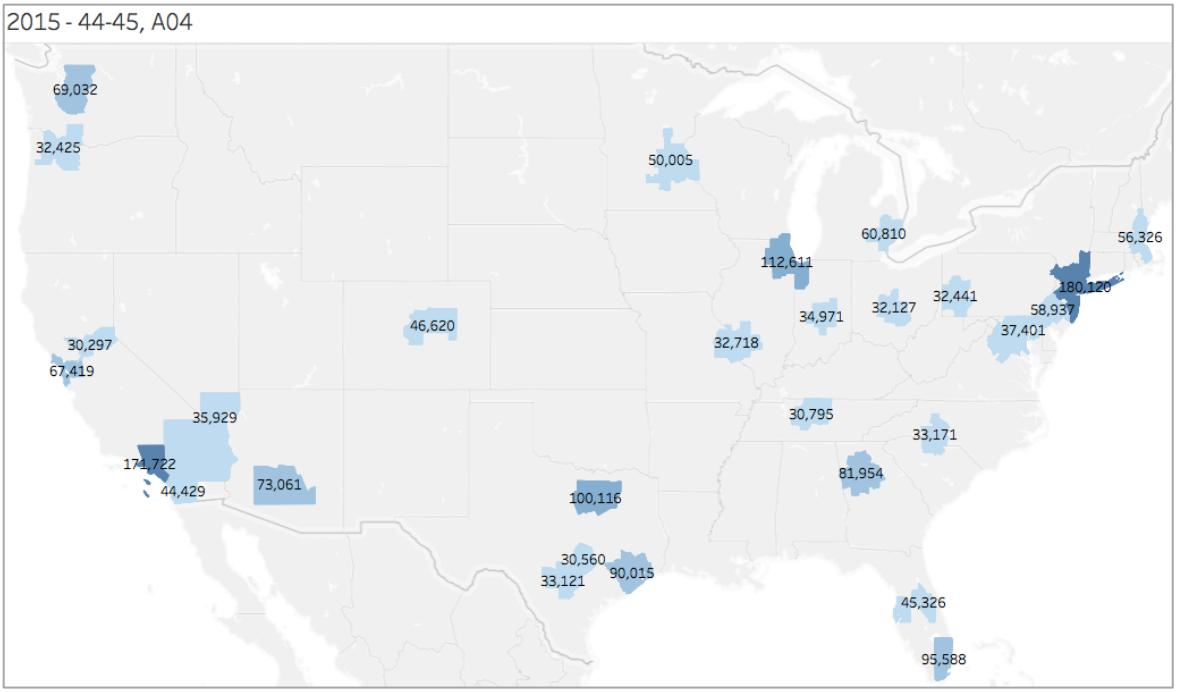
The employment numbers also showed that the greatest growth was seen for people between 25 to 34, showing retail has a steady young workforce. Although the employment indicators don't breakdown industries into subsectors to understand which the main drivers in the sector are, these areas share a common feature, they offer major airport hubs and container ports, which helps create an overall demand in the entire retail industry. This increase in employment numbers is in line with the $5.7 trillion retail sales record in 2017 reported by the U.S. census bureau (Amadeo, 2018). Just like the housing market, retail is a "key economic indicator and [retail sales] are considered a major driver of the economic health of a nation" (Investopedia, n.d.).

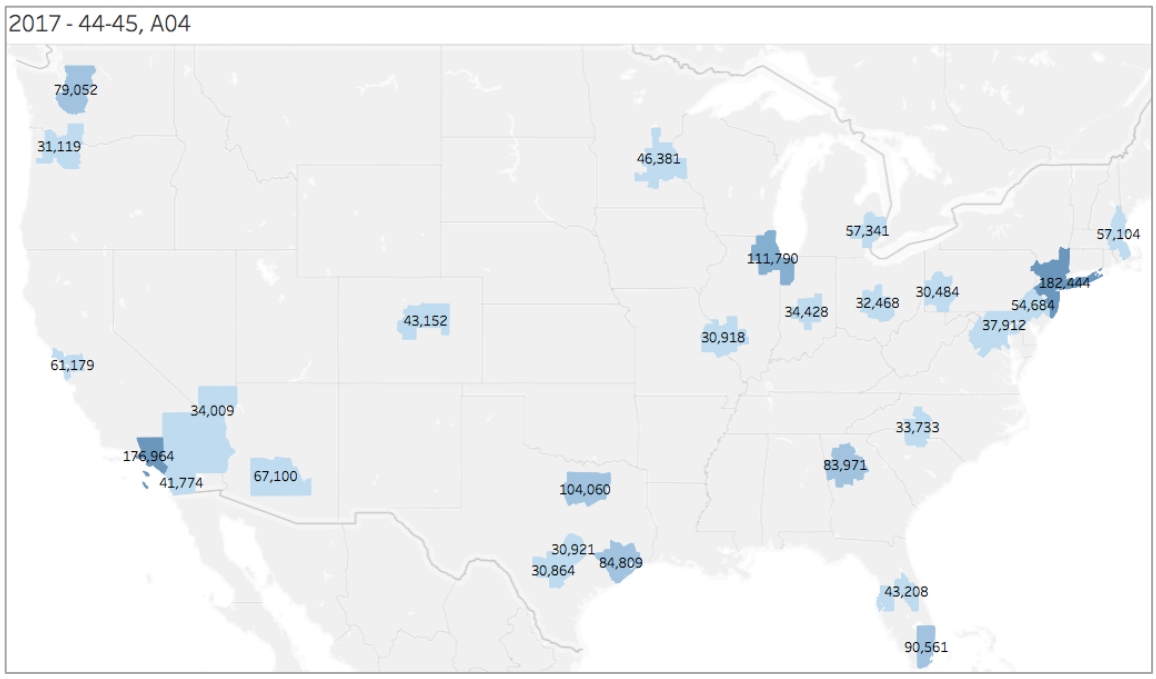
A similar trend was shown for the Health care and social assistance industry in the same metropolitan areas. In this case, an interesting change was seen in a different metropolitan area: Miami-Fort Lauderdale-West Palm Beach, FL, where there was a decrease from 2015 to 2017; from 95,558 to 90,561 in total employment count. The same metropolitan areas that saw an increase in Retail trade, also saw an increase in Health care.

These visualizations show that job creation continues to get stronger in these metropolitan areas, which drives migration by a young workforce. This migration waves

bring housing demand, which in fuels increasing home prices, which is the objective of this analysis.





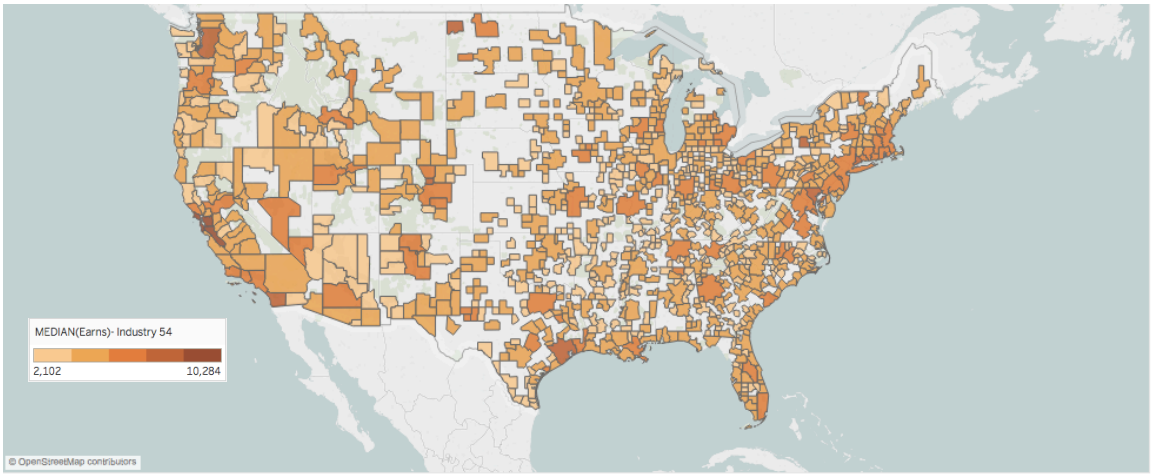


The visualizations done thus far have shown that the metropolitan areas that have seen surges in median home prices have also seen a steady increase in employment numbers, particularly for the Retail Trade (44-45) and Health Care and Social Assistance (62) industries. This shows, so far, that there is at least an empirical relationship between median housing prices and employment trends, which is the main purpose of this analysis, find if there is a real relationship between employment indicators and median home prices.

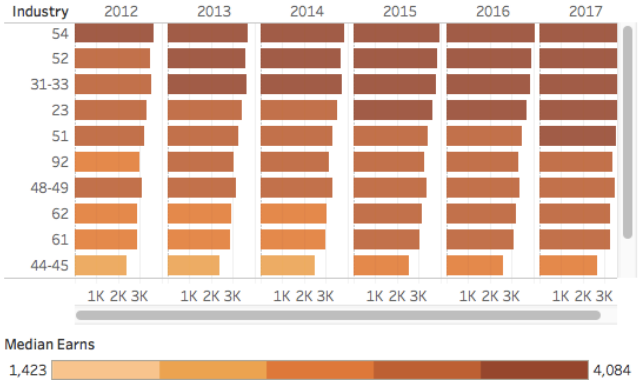
Data Visualization 5

After looking at the industries with the largest number of employees, a set of visualizations was created to examine how competitive they were when it came to wages. The visualizations show that even though Retail Trade (44-45) and Health Care and Social Assistance (62) had the highest numbers when it came to employment, Finance and Insurance (52) and Professional, Scientific, and Technical Services (54) had the

highest median monthly average wages overall, with the highest earners located in coastal cities. This shows that coastal cities are attracting both industries that are increasingly hiring and industries that are paying well.



	YearQuarter					
Industry	2012	2013	2014	2015	2016	2017
54	3,640	3,714	3,840	3,973	4,064	4,084
52	3,479	3,580	3,690	3,833	3,934	4,049
31-33	3,545	3,615	3,739	3,808	3,855	3,942
23	3,331	3,402	3,511	3,619	3,710	3,686
51	3,180	3,239	3,331	3,439	3,477	3,554
92	2,998	3,061	3,144	3,247	3,301	3,341
48-49	3,103	3,167	3,285	3,366	3,392	3,453
62	2,897	2,949	3,018	3,132	3,194	3,234
61	2,870	2,903	2,967	3,036	3,092	3,236
44-45	2,366	2,412	2,471	2,545	2,602	2,649
81	2,134	2,229	2,314	2,396	2,459	2,489
72	1,423	1,440	1,474	1,533	1,574	1,596



The above visualization shows the median average earnings by industry and year, showing that the Professional, Scientific, and Technical Services (54) sector had the highest median average from 2012 to 2017, closely followed by Finance and Insurance (52).

In general, the visualizations done thus far, show that employment (employment count and earnings) is closely related to median home prices, as the regions with higher median prices also have the industries with the highest employment and earning statistics.

Predictive Models

Since the data was segmented by economic regions in order to compare and predict median housing prices within similar regions, and simplify the model, eight different datasets were used to create three different predictive models of median housing prices using employment indicators for each region.

After removing all the highly correlated variables, each data set had 79 attributes. Since no transformation had been done to treat skewness, the individual datasets were imported into SAS EM and the log transform was applied to all the variables using the Transform Variables node, including the target variable: median housing prices in order to have nearly normal distributions. This was donesince as shown in the data profile section in figure 9, the skewness of all the employment variables was over 2 and because highly skewed variables could bias the model, especially when it comes to linear regression models as they are sensitive to outliers. Unless otherwise specified, all models for all regions were built using the log form of the input variables and the target variable.

For each dataset, three different predictive models were built: a linear regression model, a decision tree model, and a rain forest model. Following is a description of the predictive models created:

Predictive Models		
Multiple Linear Regression	Decision Trees	Random Forest
➨ Multiple Linear Regression can be very powerful and efficient. They predict the relationship between a set of predictors and the dependent variable.	➨ They split the predictor variables to identify the ones that help predict the target variable. Two of the main advantages: they help find no linear relations and	➨ An ensemble model of multiple decision tress and “the results are aggregated to improve accuracy” (Kabacoff, 2015).

	are not disturbed by outliers of missing values	➨ It works well with large datasets

Since there were 79 input variables in each dataset, the first step was attribute selection. Attributes like CBSA (metropolitan) codes and county name were removed from all datasets as they didn't provide significant information.

The seed was set to 12345 in order to reproduce results and the data was split into 60% to train the model, 20% to validate and tune the model, and 20% to assess the model. Since the objective is to identify which industries have an impact in median housing prices, only the 12 attributes with the total employment count by industry were used, along with median rent prices, and the mortgage rate variable. For all regression models, each predictor must have a 0.05 significance level to enter and stay in the model.

The total employment count for the following industries were considered as inputs for the predictive models:

23 Construction
31-33 Manufacturing
44-45 Retail Trade
48-49 Transportation and Warehousing
51 Information
52 Finance and Insurance
54 Professional, Scientific, and Technical Services
61 Educational Services
62 Health Care and Social Assistance
72 Accommodation and Food Services
81 Other Services (except Public Administration)
92 Public Administration

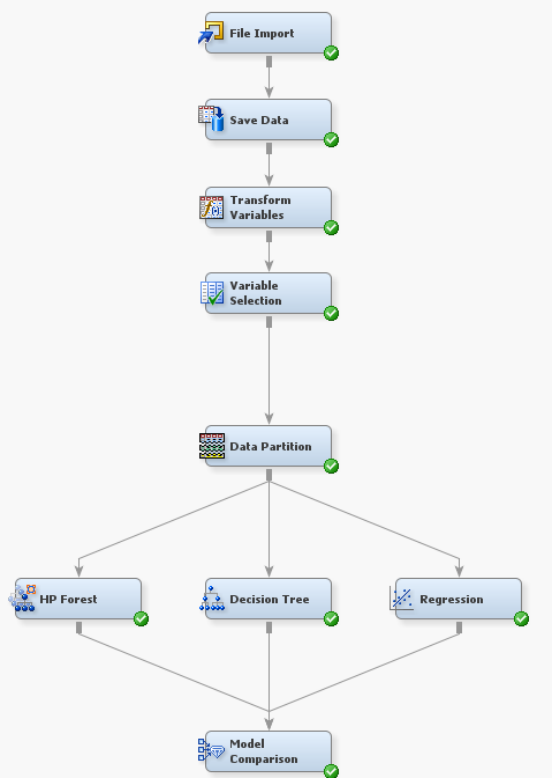
Additionally, for all the models, unless otherwise specified, the variable selection node, which uses the R-squared criterion to determine variable significance, was run for each dataset to determine the variables that were more significant to build the models. The R-squared criterion uses a "stepwise method of selecting variables that stops when the improvement in the r-square value is less than 0.0005" (SAS, n.d.). This means that variables whose contribution to the model is not statistically significant are rejected.

Since each region is inherently different, the variable selection for each dataset produced a different set of features to predict median housing prices.

Due to the large size and number of splits of all the Decision Trees developed, only the output of the algorithm is shown in the appendix and not the graphical representation of the trees. Thus, only a chart with the variable importance and the number of splits is shown.

The 'best' model for predicting (logarithm) median housing prices in each region was chosen using a combination of simplicity and the lowest Root Mean Squared Error (RMSE).

Predicting Median Housing Prices in the Farwest Region



The first set of predictive models were created to predict median home prices in the Farwest region, which includes the states of CA, HI, NV, AK, OR, and WA.

Multiple Linear Regression

The first predictive model was a multiple linear regression model using backwards selection and the following log transformed variables were used as inputs: Emp31_33, Emp48_49, Emp54, Emp72, Emp81, Emp92, median rent and states. The multiple linear regression model of the (logarithm) median housing price is given by the following equation:

$$\log\left(\widehat{median\ housing\ values}\right) = \alpha + \beta_1 \log(X_{i1}) + \beta_2 \log(X_{i2}) + \cdots + \beta_p \log(X_{ip})$$

After running the model, the prediction equation for the estimated (logarithm) median housing prices is:

$$\log\left(\widehat{\boldsymbol{median\ housing\ values}}\right) = 2.20 + 0.04 * \log(\text{construction}) + 0.029 * \log(\text{manufacturing}) - 0.081 * \log(\text{transportation/warehousing}) + 0.038 * \log(\text{professional-scientific-technical services}) + 0.079 * \log(\text{accommodation services}) - 0.07 * \log(\text{other services}) - 0.036 * \log(\text{public administration}) + 1.40 * \log(\text{median rent})$$

The symbols of the coefficients in the equation represent the positive (+) or negative (-) impact of the features in predicting median housing prices. The (+) sign before each coefficient indicates that a 1% increase in the variable, increases the median house prices by the coefficient in the model output. While the (-) symbol indicates that a 1% increase in the variable, reduces the median house prices by the coefficient with all other variables held constant.

Since the models were built using a log transformation for the target and input variables, in order to interpret the output, each coefficient is multiplied by 100 to show the percentage change in median housing prices. Thus, in this case, the model predicts that with all other predictors held constant, a 1% percent increase in the total employment

count for the Professional, Scientific and Technical service industry (54), is associated with a 0.038 % increase in housing prices, while a 1% increase in total employment count in the Transportation and warehousing industry is associated with a 0.081% decrease in housing prices (see Fig1 for complete output).

The model also estimates that a 1% increase in median rent prices along with a 1% increase in total employment count in construction, manufacturing, Professional-Scientific-Technical Services, and Accommodation and Food Services, is associated with an increase in median housing prices. Thus, as employment number increases in these industries, the model estimates that median housing prices would go up in this region.

Since this region includes states like CA, NV and HI, it makes sense that industries such as Transportation and Accommodation-services have a significant influence in housing prices, as these states are major commercial and touristic hubs with a large presence of these industries. Hence, as these industries grow, they bring more employees, which increases housing demand and prices.

The adjusted R-Squared for this model is 0.94, which means that everything held constant, together, these set of predictors account for 94% of the variance in median housing prices in this region, and the Root Mean Squared Error (RMSE) in the validation set is 0.116 and 0.108 in the test set, which means, that overall the typical prediction error is approx. 1% (see Fig2 for partial output of observed vs. predicted values).

Decision Tree

The second model built to predict median housing prices was a decision tree. The algorithm used a minimum of 10 observations per node and the average square as the

evaluation method. The decision tree split the input variables based on their chi-square to identify the variables that better help predict median home prices. Since the target variable is continuous, the predicted value is the ***average*** median housing price and the algorithm tries to minimize the mean square error (MSE) at every split. This is true for all subsequent decision tree models.

In this case, the model estimates that median rent prices, is the most influential attribute in predicting median home prices, followed by total employment count in Public Administration industry (92), Transportation-Warehousing (48-49), Accommodation-Food Services (72), professional/scientific/technical services, and other services. There were 22 splits in total, with median rent, being the most significant at the root node with 17 splits, which indicates that a large number of observations were associated with this node. The Root Mean Squared Error (RMSE) in the validation test was 0.12 and 0.13 in the test set, which means the overall prediction error rate is very low (see fig3 for output).

Just like the linear regression model, the decision tree found that total employment count in the Professional-Scientific-Technical industry and the Accommodation and Food Services industry are significant features in predicting housing prices.

Random Forest

The random forest is an ensemble model of multiple decision tress and "the results are aggregated to improve accuracy" (Kabacoff). Since random forest is a combination of trees, "instead of searching for the best feature to split a node, it searches for the best feature among a random subset of features" (Geron, 2017). The RMSE in the validation set was very low at only 0.052. Since the most relevant attributes in a Random

Forest "are likely to appear closer to the root of the tree" (Geron, 2017), the number of splits gives a good indication of the features the model considers to be significant in predicting median home prices in this region. Hence, based on the number of splits, median rent was considered to be the most important factor, followed by the total employment count in Public Administration, Accommodation-Food Services, Professional-Scientific-Technical Services, other services, Transportation-Warehousing, Manufacturing, and Construction.

Using the model comparison node, the three models were compared based on the Mean Squared Error of the validation set. The random forest model was chosen over the linear regression and the decision tree model, as it had a RMSE of 0.05 vs. 0.117 and 0.125 in the other models respectively.

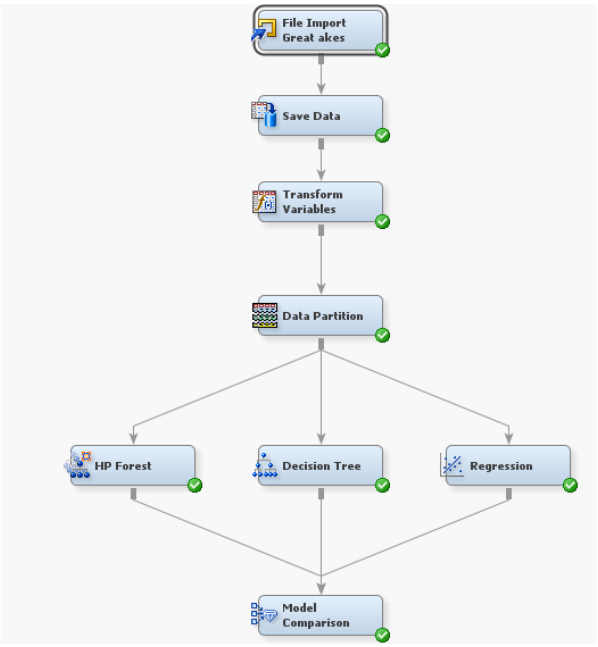
Since the RMSE of the three models is very close and choosing a predictive model is not only about finding the model with the highest predictive accuracy and the smallest RMSE, but also one that can be easily interpreted, and in this case, one that is able to clearly provide most significant variables, the linear regression model is the preferred model as it can not only explain over 90% of the variance in housing prices, but provides the coefficient of the most significant predictors to predict median housing prices in 2019.

RMSE results for all models and Adjusted R-squared for multiple linear regression:

Multiple Linear Regression	**Decision Tree**	**Random Forest**
0.12	0.12	0.05
Adj. R-squared: 0.94		

Predicting Median Housing Prices in the Great Lakes Region

The great lakes region is composed of IL, IN, MI, OH and WI. Just like the previous model, the 12 attributes with the total employment count by industry is being considered.



Linear Regression

The first predictive model was a multiple linear regression model using stepwise selection. Using this type of feature selection, the algorithm begins training the model by adding one predictor at a time with each iteration, “but may remove effects already in the model” (SAS EM, n.d.) as it iterates when the interaction between variables is not statistically significant.

The total count of employment for all industries, the median rent price, and mortgage rate were used as predictor variables to build the model. The multiple linear regression model of the (logarithm) median housing price is given by the following equation:

$$\log\left(\widehat{median\ housing\ values}\right) = \alpha + \beta_1 \log\left(X_{i1}\right) + \beta_2 \log\left(X_{i2}\right) + \cdots + \beta_p \log\left(X_{ip}\right)$$

Using all the employment count variables as inputs, the best model shows that construction (+), manufacturing (+), finance (+), Educational Services (+), Accommodation and Food Services (+), public administration (+), and median Rent (+), positively impact housing prices. This means that a 1% increase in these indicators results in an increase in housing prices by the coefficient of each feature. While Health Care and Social Assistance (-), transportation-warehousing (-), retail trade (-), and other services (-), have the opposite impact on median housing prices in this region. This means that in this region, a 1% increase in total employment in retail is associated with a 0.33% decrease in median housing prices or an increase in the total count of employment in health care and assistance service is associated with a 0.05% decrease in housing prices.

The adjusted R-squared of the model is 0.760, meaning that taken together and held constant, these predictors account for 76% of the variability in median housing prices in this region. The RMSE in the validation set is 0.15 and 0.14 in the test set. The prediction equation for the estimated (logarithm) median housing prices is:

$$\log\left(\widehat{\boldsymbol{median\ housing\ values}}\right) = 3.734 + 0.122*\log(\text{construction}) + 0.065*\log(\text{manufacturing}) - 0.329*\log(\text{retail}) - 0.046*\log(\text{transportation-warehousing}) - 0.051*\log(\text{information}) + 0.094*\log(\text{finance-insurance}) + 0.064*\log(\text{education}) - 0.058*\log(\text{health care-social services}) + 0.196*\log(\text{accommodation}) - 0.054*\log(\text{other services}) + 0.028*\log(\text{public administration}) + 1.10*\log(\text{median rent})$$

The model shows that the presence of more industries is an important predictor of median housing prices in the great lakes region. In this case, higher employment counts

in finance/insurance, education, accommodation and manufacturing are strong predictors of higher median housing prices, while public administration, health care-social services and the information industry indicate a slight decrease in median housing prices (see Fig4 for complete model output).

Decision Tree

The decision tree algorithm splits the variables based on their significance using the chi-square. In this case, the model identified median rent prices as the most influential attribute in predicting median home prices, followed by whether the home is in Illinois, total employment count in Accommodation and Food Services (72), whether the home is in Illinois, the total count of employment manufacturing, health care/assistance services, public administration, finance/insurance, information, professional services, and accommodation/food services. There were 23 splits in total, with median rent having 7 splits, which indicates that a large number of observations were associated with this leaf node. The RMSE in the validation and test set was 0.12, which means the overall prediction error rate is very low (see Fig5 for complete output).

Just like the linear regression model, the decision tree found that (logarithm)m median rent prices is a strong predictor of median housing values, followed by the total employment count in Accommodation and Food Services.

Random Forest

Using the same features as predictors, the random forest model showed that based on the number of splits, median rent followed by total count of employment in the public administration industry are the most significant features in predicting median housing

prices in the Great Lakes Region. The subsequent splits considered the total number of employees in the other services industry, followed by the employment count in accommodation, health care-social assistance, educational services, professional/scientific/technical services, finance/insurance, information, transportation manufacturing, retail, construction, and lastly the state the home is located. The RMSE in the validation and test sets was only 0.04 (see Fig6 for complete model output).

Using the comparison node, all tree models were compared to select the best model based on the MSE. The 'best' model was the Random Forest with an MSE of 0.04 in both the validation and the test set, followed by the Decision Tree with 0.12 RMSE in the validation set and 0.11 in the test set, and the regression model with 0.15 in both the validation and test sets. However, the plot of the distribution of predicted values, shows the random forest may be overfitting the data and it also provides a more complex structure to interpret than the other models (see Fig7).

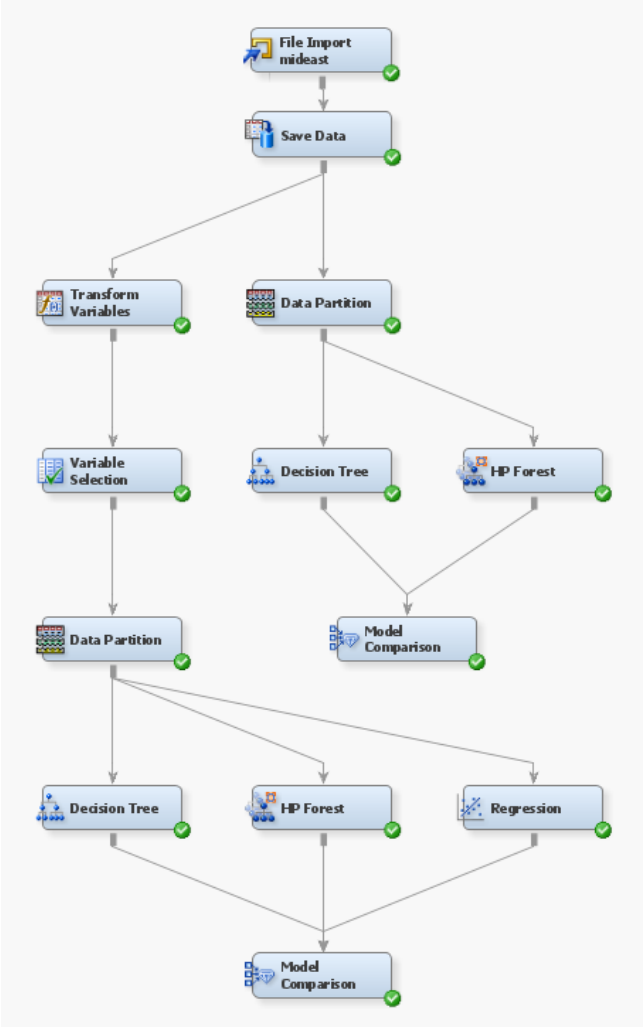
Thus, the linear regression model was chosen as the best model to predict median housing prices in the Great Lakes Region as it was not only easy to interpret, but it was able to explain 76% of the variance in housing prices in the region with a simple model.

The predictive models for this region show that a different set of industries influence median housing prices in the great lakes region compared to the industries that help predict prices in the Farwest region. While the finance and insurance industry along with mortgage rates play a role in predicting median housing prices in the great lakes region, these attributes had no statistical significance in predicting median housing prices in the Farwest region.

RMSE results for all models and Adjusted R-squared for multiple linear regression:

Multiple Linear Regression	Decision Tree	Random Forest
0.15	0.12	0.04
Adj. R-squared: 0.760		

Predicting Median Housing Prices in the Mideast Region



The states in the Mideast region are NJ, NY, DE, MD, PA and the District of Columbia. Just like with the previous regions, only the total employment counts for all industries, 12 in total, the median rent price and the mortgage rate were considered as inputs for the predictive models.

Since the decision trees and the random forest can handle skewed variables, an additional decision tree and one more random forest were build using untransformed predictors including the target variable: median housing prices to compare the RMSE.

Linear Regression

The multiple linear regression model of the (logarithm) median housing price is given by the following equation:

$$\log\left(\widehat{median\ housing\ values}\right) = \alpha + \beta_1 \log\left(X_{i1}\right) + \beta_2 \log\left(X_{i2}\right) + \cdots + \beta_p \log\left(X_{ip}\right)$$

Using stepwise selection and all total employment count for all industries, 12 in total, the log of median rent price, and log of mortgage rate as predictors, the best model shows that construction (+), transportation-warehousing (+), median rent (+), finance-insurance (+), accommodation and food services (+), and information (+) have a positive impact in median housing prices. While total employment count in manufacturing, other services, health care and social assistance (-), and public administration have the opposite effect. The adjusted R-squared of the model was 0.90, which means that with everything else held constant, together these predictors account for 90% of the variance in median home prices in the region.

The prediction equation for the estimated (logarithm) median housing prices in the Mideast region is:

$$\begin{aligned}\log\left(\widehat{median\ housing\ values}\right) &= 1.160 + 0.208 * \log(construction) - 0.095 * \log(manufacturing) \\ &+ 0.036 * \log(information) + 0.094 * \log(finance) - 0.143 \\ &* \log(health\ care) + 0.214 * \log(accomodation) - 0.260 \\ &* \log(other\ services) - 0.056 * \log(public\ administration) + 1.51 \\ &* \log\left(median\ rent\right)\end{aligned}$$

Looking at the formula for this model, we can see that the Mideast and Great lake regions share similar industries and similar predictors for median housing prices. However, in this case, the total count of employment in manufacturing has a negative impact in median housing prices, meaning that a 1% increase in the total number of jobs in manufacturing is associated with a 0.095% decrease in housing prices, while a 1% increase in employment in the finance/insurance industry is associated with a 0.094% increase in median housing prices (see Fig8 for model output).

Decision Tree

Just like the previous decision tree models for other regions, this model shows that median rent is a significant factor to predict median housing prices. This was the root node with 10 splits, followed by total employment count in accommodation/food services, finance/insurance, public administration, other services, manufacturing, information, health care and assistance services, and professional/scientific/technical services. The RMSE of the model in the validation set is 0.12 and 0.14 in the test set. See Fig9 for decision tree output.

Another decision tree was also tried in this region splitting the data 55% for the training set, 25% of the validation set and 20% for the test set, but without transforming the data to better interpret the results. The output show that the RMSE of the model was $16,005 in the validation set and $18,868 in the test set, which are good scores. There were 23 splits in total and tree considered the median rent as the root node, then state, followed by the total employment count in accommodation/food services, finance/insurance, mortgage rate manufacturing, total employment in other services, transportation/warehousing, and retail. See fig10 for model output.

Both decision trees performed very well with the transformed and untransformed data.

Random Forest

As previously mentioned, two different random forest were built for the Mideast region. One using the log transformed data and the other using the data without transformation. The datasets were randomly split using the same percentage used for the decision tree: 55%-25%-20% for the training, validation, and test sets respectively.

The random forest built using the log transformed data had a 0.03 RMSE in the validation set and test 0.04 in the test set. Based on the number of splits, median rent was the most significant attribute to split the observations, followed by total employment counts in public administration, information, accommodation/food services, finance/insurance, other services, and health care/assistance services. See fig11 for output.

The second random forest was built using untransformed data. The RMSE in the validation set was $6,270 and $6,101 in the test set, which once again, are good scores considering the range of median home prices. The most significant attribute to split the average median housing prices was median rent, followed by mortgage rate, total employment count in public administration, other services, accommodation/food services, health care/assistance services, manufacturing, professional/scientific/technical Services, information, finance/insurance, and educational services (see fig.12 for model's output).

Using the model comparison node, the models build using the transformed data were compared. Based on the RMSE, the random forest model had the best mean square error in the validation set (see fig13 for model comparison plot). The random forest and the decision tress models built using the untransformed data were also compared and the random forest also had the lowest RMSE.

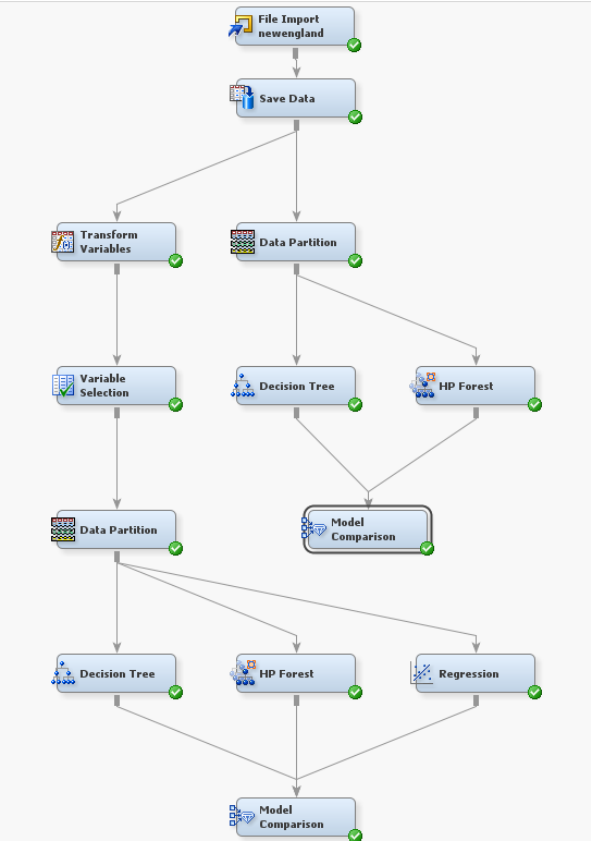
Although the random forest had the overall lowest score using both the transformed and the untransformed data, the accuracy was achieved by sacrificing simplicity. Hence, the multiple linear regression model was chosen as the 'best model' to predict the (logarithmic) median housing prices in the Mideast region, as it was able to explain 90% of the variation in median housing prices with a simpler model.

In general, the models built to predict median housing prices in the Mideast region show that median rent prices, total count of jobs in construction and manufacturing are significant indicators of median housing prices. Based on the results of the decision tree and the random forest, the state where the home is located, also plays a significant role, with NY and DC being the strongest predictors of higher housing prices.

RMSE results for all models and Adjusted R-squared for multiple linear regression:

Multiple Linear Regression	**Decision Tree**	**Random Forest**
0.14	0.12	0.03
Adj. R-squared: 0.90		

Predicting Median Housing Prices in the New England Region



The states in the New England region are CT, MA, ME, NH, and RI. Just like with the previous regions, only the total employment counts for all industries, 12 in total, the median rent price and the mortgage rate were considered as inputs for the predictive models.

Linear Regression

The multiple linear regression model of the (logarithm) median housing price is given by the following equation:

$$\log(\widehat{median\ housing\ values}) = \alpha + \beta_1 \log(X_{i1}) + \beta_2 \log(X_{i2}) + \cdots + \beta_p \log(X_{ip})$$

After running the model, the prediction equation for the estimated (logarithm) median housing prices in the New England region is:

$$\log(\widehat{median\ housing\ values})$$
$$= 4.201 + 0.1136 * \log(construction) + 1.0193 * \log(median\ rent)$$
$$+ 0.125 * \log(retail) - 0.1089 * \log(finance|insurance) - 0.084$$
$$* \log(information) + 0.1415$$
$$* \log(Professional|Scientific|Technical\ Services) - 0.1825$$
$$* \log(health\ care|social\ assistance)$$

The model estimates that a 1% increase in the total employment count in construction, retail, and professional/scientific and technical services is linked to an increase in median housing prices in the new England region. While a 1% increase in health care/social assistance employment, in the information industry and the finance/insurance industry is associated with a decrease in median housing prices. The output shows that the adjusted R-squared of the best model is 0.9175, which means that with everything else held constant, together these predictors account for 91.7 % of the variance in median home prices in the region. The RMSE of the model in the validation set was 0.081 and 0.096 in the test set (see Fig14 for model output).

Decision Tree

Two decision trees were built for the new England region, one using transform data and one with untransformed data. The decision tree with the transformed data had a 0.096 RMSE in the validation set and it considered median rent to be the most significant attribute to predict median housing prices, followed by total employment count in professional/scientific/technical services, whether the home is located in MA, and the total employment count in the transportation/warehousing industry. Unlike the linear

regression model, the decision tree considered total employment count in transportation/warehousing a statistically significant predictor of median housing prices (see fig15 for model output).

The second decision tree model using untransformed target and input variables showed that slightly different variables were significant in predicting median home prices. Median rent was still the most significant features, followed by employment count in information, public administration, professional/scientific/technical services, state, accommodation/food services and lastly total employment count in the transportation/warehousing industry. The RMSE was $21,725 in the validation set and $20,376 in the test set (see fig16 for model output).

Random Forest

The random forest built on the transformed data had a 0.033 and 0.030 RMSE in the validation and test sets respectively. The most important features in splitting order were median rent, followed by total employment count in public administration, other services, health care/social assistance, educational services, Professional/Scientific/Technical Services, finance/insurance, information, transportation, construction, and whether the home is MA, ME or NH (see fig17 for model output).

The RMSE of the random forest built using the untransformed input and target variables was $10,227 in the validation set and $8,751 in the test set. Once again, the most significant feature was median rent, followed by total employment count in other services, health care/social services, public administration, manufacturing, mortgage rate, total employment count in finance/insurance, information, transportation/warehousing,

professional/scientific/technical services, accommodation and services, retail and educational services.

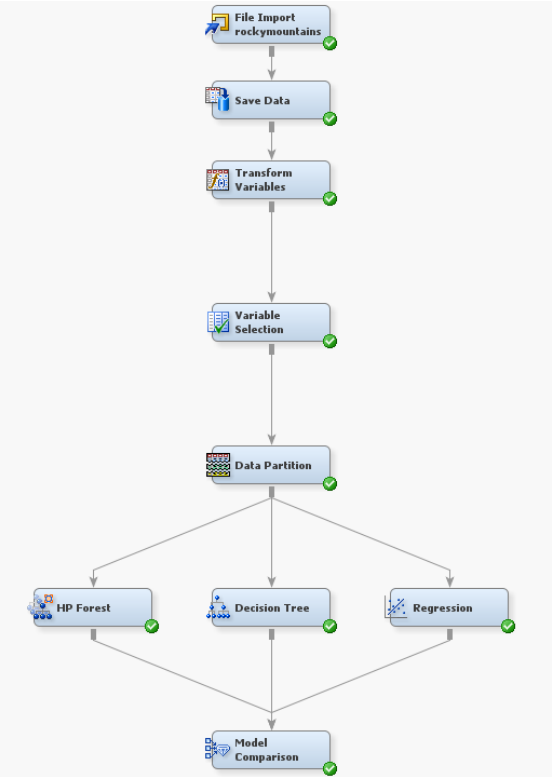
Although the random forest had a lower RMSE than the multiple linear regression model and the decision trees, the regression model was simpler and taken together, the predictors were able to account for 92 % of the variance in median home prices in the new England Region (see fig18 for models score distribution).

In general, all the models for the New England Region show that total employment counts in the health care/social assistance and manufacturing industries are significant predictors of median housing prices. This shows that higher percentage of stable hires in these industries have an overall impact on median housing prices.

RMSE results for all models and Adjusted R-squared for multiple linear regression:

Multiple Linear Regression	**Decision Tree**	**Random Forest**
0.08	0.096	0.03
Adj. R-squared: 0.92		

Predicting Median Housing Prices in the Rocky Mountains Region



Linear Regression

The multiple linear regression model of the (logarithm) median housing price is given by the following equation:

$$\log\left(median\ \widehat{housing}\ values\right) = \alpha + \beta_1 \log\left(X_{i1}\right) + \beta_2 \log\left(X_{i2}\right) + \cdots + \beta_p \log\left(X_{ip)}\right.$$

Using stepwise regression to select the most significant predictors and the log of total employment counts for all 12 industries, log median rent prices, and log median mortgage rate as inputs, the prediction equation for the estimated (logarithm) median housing prices in the Rocky Mountains region is:

$$\log(\widehat{median\ housing\ values})$$
$$= 6.0978 + 0.110 * \log(construction) - 0.0231 * \log(manufacturing)$$
$$+ 0.1137 * \log(finance|insurance) + 0.0631$$
$$* \log(professional, scientific, and\ technical\ services) - 0.1876$$
$$* \log(health\ care|social\ assistance) - 0.1366$$
$$* \log(public\ administration) + 0.9390 * \log(median\ rent)$$

The results of the best linear regression model show that among the most significant features in predicting median housing prices Rocky Mountains region are construction (+), finance/insurance (+), health care/social assistance, professional, scientific, and technical services (+), and median rent. According to the model, a 1% increase in the presence of the finance/insurance industry, the construction industry and the professional/scientific/technical services are associated with an increase in median housing prices, while a larger presence of the public administration and the health care/social assistance industry are estimated to be associated with lower median housing prices. More specifically a 1% increase in employment count of the finance industry is associated with a 0.1137% increase in median housing prices (see fig19 for model output).

The adjusted R-squared of the model was 0.8740, which means that using these set of predictors the model explains 87.4% of the variability in median housing prices in the region.

Decision Tree

The decision tree model to predict median home prices in the Rocky Mountains region estimates that the most influential features in predicting median home prices are

median rent, total employment count in the retail industry, whether the home is located in Utah or Colorado, and the total employment count in the manufacturing industry. The RMSE for the model in the validation set was 0.107 and 0.109 in the test set. See fig20 for model details.

Contrary to the logistic regression, the decision tree model did not find the presence of the finance/insurance industry or the health care/service assistance industry important features in predicting median housing prices.

Random Forest

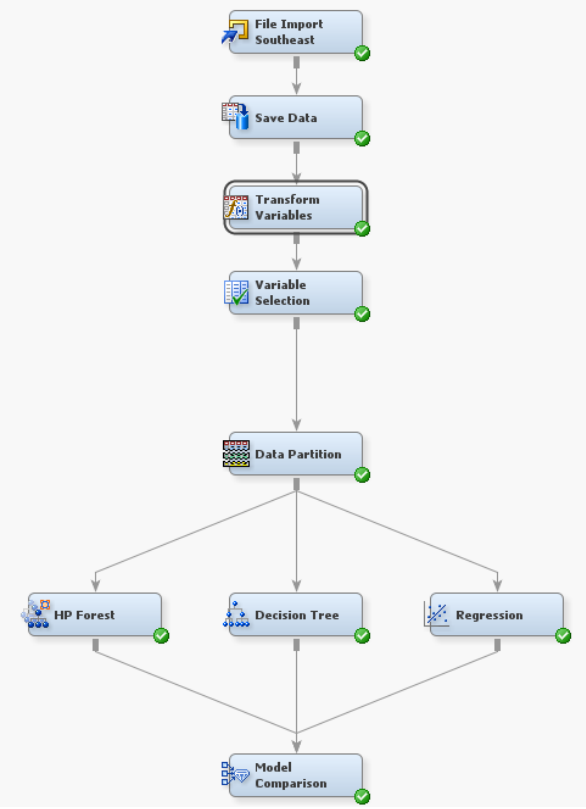
The random forest model used a greater number of features to predict median housing prices compared to the decision tree. Just like the linear regression model, the Random forest model estimates that the construction, finance/insurance, professional/scientific/technical services, and health care/assistance services industries are important features in predicting median housing prices. The RMSE of the random forest model was 0.053 in the validation set and 0.046 in the test set (see Fig 21 for output models).

Using the comparison node, all three models were evaluated based on the lowest MSE. The result shows that between the linear regression, the decision tree, and the random forest model, the random forest had the smallest MSE and it was chosen as the best model. However, the random forest was the most complex model with a large number of nodes and splits. Hence, in this case, the decision tree is preferred as it's easier to interpret and the RMSE is also relatively small in both the validation and test sets.

RMSE results for all models and Adjusted R-squared for multiple linear regression:

Multiple Linear Regression	Decision Tree	Random Forest
0.11	0.11	0.05
Adj. R-squared: 0.874		

Predicting Median Housing Prices in the Southeast Region



The states in the Southeast region are AL, AR, FL, GA KY, LA, MS, NC, SC, TN, VA, and WV.

Linear Regression

The multiple linear regression model of the (logarithm) median housing price is given by the following equation:

$$\log\left(\widehat{median\ housing\ values}\right) = \alpha + \beta_1 \log\left(X_{i1}\right) + \beta_2 \log\left(X_{i2}\right) + \cdots \beta_p \log\left(X_{ip}\right)$$

Using stepwise regression to select the most significant predictors and the log of

total employment counts for all 12 industries, log median rent prices, and log median mortgage rate as inputs, the prediction equation for the estimated (logarithm) median housing prices in the Southeast region is:

$$\log\left(\widehat{median\ housing\ values}\right)$$
$$= 3.839 + 1.118 * \log(median\ rent) + 0.296 * \log(construction)$$
$$- 0.1777 * \log(retail) - 0.0485 * \log(transpotation|warehousing)$$
$$+ 0.032\ * \log(professional|scientific|technical\ ) + 0.1016$$
$$* \log(educational\ services) - 0.1091$$
$$* \log(health\ care - assistance\ services) + 0.1691$$
$$* \log(accomodation) - 0.055 * \log(public\ administration) + 0.080$$
$$* \log\ (other\ services)$$

The output of the model to predict median housing prices in the Southeast region showed that median rent (+), total employment in construction (+), professional/scientific/technical services (+), accommodation (+), and educational services (+) have a positive impact on median housing prices, while retail (-), health care and assistance services, transportation (-), and public admiration have a negative impact. This means that a 1% increase in the total employment counts in construction is associated with a 0.296 % increase in median home prices, while a 1% increase in total employment counts in retail is associated with a 0.177% decrease in median housing prices.

The adjusted R-squared of the model was 0.7591, which means that together, these predictors account for 75.91% of the variance in median home prices in the region. The RMSE of model was 0.15 in the validation set and 0.14 in the test set (see Fig23 for

model output).

Decision Tree

The decision tree used the 12 log-transformed total employment counts, log median rent prices and log mortgage rate as predictors. The output of the decision tree showed that median rent prices, total employment count in educational services, accommodation/food services, transportation/warehousing, retail, whether a home is located in NC, VA or WV, total employment in construction, professional services, other services, if the home is located in FL, total employment in health care/assistance services, and lastly total count of employment in public administration are important features in estimating median housing prices. The decision tree model had a RMSE of 0.13 in both the test and validation sets, which is comparable with the RMSE of the logistic regression model, but achieved with greater complexity (see fig24 for model output).

As previously mentioned, the decision tree algorithm splits the input variables based on their chi-square to identify the variables that better help predict median home prices. In this case, the decision tree had to do multiple splits in order to minimize the MSE, which made the model increasingly complex.

Random Forest

The random forest model also used the total employment count for the 12 industries, log median rent prices and log mortgage rates as predictors. The model estimates that the most influential features in predicting median housing prices are median rent, total count of employment in public administration, other services, accommodation/food services, health care/social assistance, educational services,

professional/scientific/technical services, transportation/warehousing, construction and retail trade. The RMSE of the model was 0.03 for both validation and test sets. See fig25 for model output.

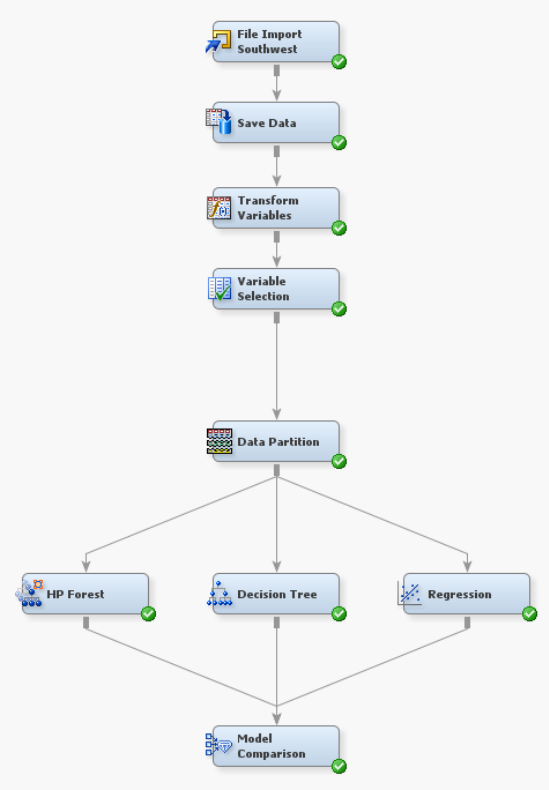
Just like with previous predictive models for the different regions, all the predictive models show that there is not a single industry that drives median housing prices, but a combination of predictors that taken together explain the variability in median housing prices.

In this case, even though the random forest had the lowest RMSE, 0.03 vs. 0.15 for the linear regression model and 0.13 for the decision tree, the regression model offered a simpler and more straightforward predictive model. Hence, even though there is a tradeoff of simplicity vs. accuracy, the linear regression model is the best model, as it was able to estimate over 75% of the variance in median housing prices with a 0.15 RMSE.

RMSE results for all models and Adjusted R-squared for multiple linear regression:

Multiple Linear Regression	**Decision Tree**	**Random Forest**
0.15	0.13	0.03
Adj. R-squared: 0.76		

Predicting Median Home Prices in the Southwest Region



Linear Regression

Just like the predictive models of the previous regions, the first predictive model was a multiple linear regression model using stepwise selection. The 12 features representing the log of the total employment count by industry, the log median rent price, and the log mortgage rate were used as input variables. The multiple linear regression model of the (logarithm) median housing price using log transformed input variables is given by the following equation:

$$\log\left(median\ \widehat{housing}\ values\right) = \alpha\ +\ \beta_1 \log\left(X_{i1}\right) +\ \beta_2 \log\left(X_{i2}\right) + \cdots +\ \beta_p \log\left(X_{ip)}\right)$$

The model estimates that the prediction equation of median housing prices in the southwest region is :

$$\log(\widehat{median\ housing\ values})$$
$$= 2.4309 + 1.278 * \log(median\ rent) - 0.0649$$
$$* \log(manufacturing) + 0.127 * \log(retail) - 0.09$$
$$* \log(transportation|warehousing) + 0.0327 * \log(information)$$
$$+ 0.08 * \log(professional|scientific|technical\ services) - 0.064$$
$$* \log(educational\ services) - 0.073$$
$$* \log(health\ care|assistance\ services) + 0.127$$
$$* \log(accomodation) - 0.058 * \log(other\ services)$$

The model estimates that the most significant predictors of median housing prices in the southwest region are median rent(+), total count of employment in manufacturing (-), retail (+), transportation/warehousing (-), information (+), professional services(+), educational services (-), healthcare and assistance services (-), accommodation (+) and other services (-). According to the model, a 1% increase in employment counts in retail is associated with a 0.12% increase in median home prices. The adjusted R-squared of the model was 0.8940, which means these set of predictors account for 89.4% of the variance in median home prices in the southwest region. The RMSE was 0.11 in both test and validation sets (see fig26 for model output).

Decision Tree

The decision tree used the 12 features representing the total employment counts by industry, the log median rent prices and the log mortgage rate as predictors. There were 22 different splits and the output of the model showed that median rent, total count of employment in retail, whether a home is located in TX, OK or NM, total employment in accommodation, manufacturing and health care/ assistance services, and transportation

are the most significant features to predict median housing prices. The RMSE of the model was 0.13 in the validation and test sets (see fig27 for model output).

Random Forest

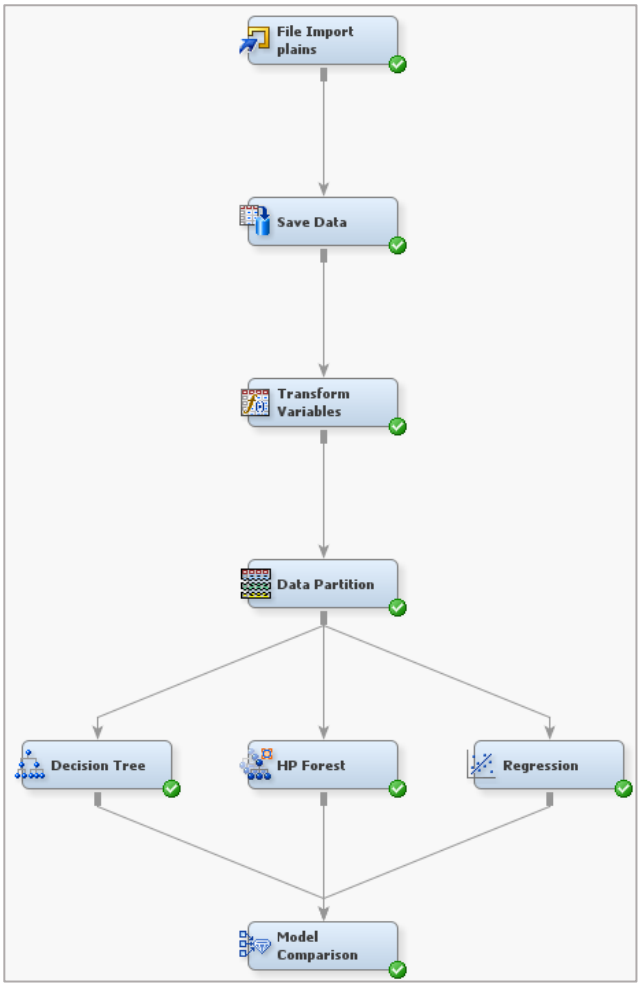
The random forest used the same input features and the model estimated that the most influential factors in predicting median housing prices are median rent, total employment count in accommodation/food services, other services, health care/social assistance, educational services, professional services, transportation/warehousing, information, retail, manufacturing and whether the home is located in TX, OK or NM. The RMSE of the model was 0.04 in both the validation and the test sets.

Using the comparison node and the average squared error to assess the models, the random forest was considered to be the best model. However, since the multiple linear regression model performed well on both the test and validation sets and it was able to explain over 89% of the variability of median home prices, the linear regression model was chosen over the random forest and the decision tree, as it was simpler and the RMSE was 0.11 (see Fig28 for output comparison).

RMSE results for all models and Adjusted R-squared for multiple linear regression:

Multiple Linear Regression	**Decision Tree**	**Random Forest**
0.11	0.13	0.04
Adj. R-squared: 0.894		

Predicting Median Housing Prices in the Plains Region



The states in the Plains region are MN, IA, MO, KS, ND, and NE. Just like the previous models, the 12 log-transformed features for the total employment counts by industry, the log median rent, and the log mortgage rate are considered for all the models.

Linear Regression

The first predictive model was a multiple linear regression model using stepwise selection. The multiple linear regression model of the (logarithm) median housing price is given by the following equation:

$$\log\left(\widehat{median\ housing\ values}\right) = \alpha + \beta_1 \log\left(X_{i1}\right) + \beta_2 \log\left(X_{i2}\right) + \cdots + \beta_p \log\left(X_{ip}\right)$$

Using all the employment count variables as inputs, the best model shows that total employment in manufacturing, retail, transportation/warehousing, finance/insurance, educational services, accommodation/food services, median rent and public service are strong predictors of median housing values in the Plains region.

The adjusted R-squared of the model is 0.7942, meaning that taken together and held constant, these predictors account for 79.42% of the variability in median housing prices in this region. The RMSE in the validation set and the test set is 0.11. The prediction equation for the estimated (logarithm) median housing prices is:

$$\log(\widehat{median\ housing\ values}) = 4.4892 - 0.0388*\log(\text{manufacturing}) - 0.122*\log(\text{retail}) + 0.014*\log(\text{transportation-warehousing}) - 0.0167*\log(\text{finance-insurance}) + 0.0568*\log(\text{education}) + 0.084*\log(\text{accommodation}) + 0.0431*\log(\text{other services}) + 1.0382*\log(\text{median rent})$$

The model estimates that higher employment counts in manufacturing, retail, and finance are associated with a decrease in median housing prices, while an increase in transportation, education, accommodation and other services is linked to an increase in median housing prices. Thus, according to the model, a 1% increase in manufacturing is associated with a 0.038% decrease in housing, while a 1% increase in the total employment count in the accommodation industry is associated with a 0.084% decrease in housing (see Fig29 for model output).

Decision Tree

The decision tree had a total of 24 splits and it found that employment by a larger number of industries is an important predictor of median housing prices in the Plains

region. The algorithm estimates that median rent is the most influential predictor of median housing prices, followed by total count of employment in accommodation/food services, finance/insurance, whether the home is in KS, total employment count in health care/assistance services, educational services, whether the home is in NE, and lastly the total employment count in transportation and warehousing. Unlike the linear regression model, the decision tree did not consider manufacturing as a significant feature in predicting housing prices.

The RMSE in the validation set was 0.094 and 0.093 in the test set. Based on the splits from the decision tree, total count of employment in accommodation/food services, and finance/insurance is more significant in predicting prices than the total count of employment in the transportation and warehousing industry (see Fig30 for model output).

Random Forest

The random forest for this model showed that based on the number of splits, median rent is the most significant feature in predicting median housing prices and it's the root node. The other variables considered in the model in order of importance were: total employment count in the public industry, accommodation/food services, other services, health care/assistance service, professional/scientific/technical services, retail, information, construction, if the home was located in KS, manufacturing, finance/insurance, and transportation. The RMSE in the validation set was 0.037 and 0.040 in the test sets (see Fig31 for complete model output).

Even though the RMSE is smaller than the one obtained by the regression and the decision tree models, the random forest is more complex, considering a larger number of inputs, which resulted in greater number of splits.

Using the comparison node, all tree models were compared to select the best model based on the MSE. The results show that the random forest had the lowest MSE in the validation set. However, due to the complexity of the model and considering that the RMSE of the random forest was comparable to the RMSE of the multiple linear regression model, the multiple linear regression model is selected as the 'best model', as it was able to explain over 79% of the variability of the (logarithm) median home prices in the Plains region with a 0.11 RMSE.

In general, the three predictive models for this region show that employment by a greater number of industries influence median housing prices.

RMSE results for all models and Adjusted R-squared for multiple linear regression:

Multiple Linear Regression	**Decision Tree**	**Random Forest**
0.11	0.094	0.037
Adj. R-squared: 0.794		

Overall Predictive Model Review

In general, all the predictive models confirmed the initial assumption of this project that different industries influence housing prices in different ways across regions in the U.S. Although some industries have a positive impact in the median housing prices in a region, that same industry can have a negative impact in a different region.

Additionally, the models showed that a large presence of an industry in a region is not always associated with increase housing prices, and in fact, as mentioned above, there are industries that in some regions tend to be linked to a decrease in housing prices. For instance, while finance and insurance were associated with increase in housing prices in

the Rocky Mountains region, the same industry was associated with lower median housing prices in the New England Region and in the Plains region.

Even though the Random Forest models outperformed the multiple regression models and decision trees in predicting median housing prices across all regions based on the RMSE, the complexity of the models' output and the large number of nodes and splits made them less likely to be chosen as the champion model. As previously noted a 'good' model not only provides increased accuracy and lower error rate, but it's also simple and easy to read and interpret. For that reason, the multiple linear regression models were predominately chosen, as they provided not only high adjusted R-squared values that indicated that the multiple linear regression models were able to explain, in average, 80% of the variability of median housing prices in all regions, but they are easy to implement, interpret, and offer an equation to easily predict future median housing prices using the same set of predictors.

Final Results

Analysis Justification

The previous analysis was done to examine the relationship between employment and median housing prices in the United States for two reasons: First, an exorbitant rise in housing prices in multiple states, is being linked to the presence and employment of fast growing industries like high-tech, causing housing instability, which triggers other social and economic gaps, particularly among very low-income families who can't afford their homes. Second, since the real estate market is a key indicator of a regions' economic growth, understating the industries that influence housing prices can provide local officials information on the estimated valuation of their real estate market to make

informed decision when it comes to infrastructure investment, building regulations, and affordable housing initiatives.

Thus, the main objective was to understand how employment by industry is corelated to median housing prices. As mentioned above, the results of the analysis and predictive models can provide local officials with information necessary to examine initiatives on affordable housing and review current building policies, such as "zoning regulations that limit the housing supply" (Scott, 2015) that can further fuel housing prices.

Even though the housing and employment data were retrieved by metropolitan area for the fifty states and Washington D.C., since not all metropolitan areas are alike and their geographic and economic characteristics mat not be able to be replicated in other metropolitan areas across the country, the data was segmented into eight economic regions in order to create predictive models of median housing prices in areas that share similar characteristics.

The segmentation, allowed to consider for each predictive model, the special geographic and economic characteristics that attract a diverse set of industries, employers, and workforce, which in the end confirmed the initial assumption of this project that different industries influence housing prices in different ways across regions in the U.S.

The total workforce counts by industry were the only employment variables used in the predictive models as they provide information not only on the industries that are active within the region, but also the industries that are having the largest impact in the socio-economic makeup of the region. As part of the housing data, the median rent data

were used to determine if they could affect home values. The assumption was that median rent prices could be a clear indicator of a surge in housing demand. As rent prices increase, potential home owners may feel encourage to invest in homeownership instead of paying high rents.

Findings

The following are the findings and conclusions from the data analysis and the predictive models

- Median housing prices have seen a rapid increase across metropolitan areas in the U.S. between 2013 and 2017, particular in the North East and West Regions, where the technology industry (west region) and the finance sector (northeast region) are predominant.
- The Retail Trade and Health Care and Social Assistance industries have both (1) a wide spread presence across the country and (2) a large number of employees in coastal cities.
- The Health Care and Social Assistance industry have shown steady hiring of a young workforce (25 to 34 years old) and job creation continues to get stronger in coastal metropolitan areas.
- The Retail Trade industry saw an increase in employment numbers across all metropolitan areas between 2012 and 2017, particularly in five metropolitan areas, which also saw a surge in median home prices in 2017.
- Based on the predictive models, total employment counts by industry influence median housing prices differently depending on the economic region. The same industry can have a positive impact on housing prices in one region and negative

impact in a different region. For instance, while the finance and insurance industry was associated with increase in housing prices in the Rocky Mountains region, the same industry was associated with lower median housing prices in the New England Region and in the Plains region.

- Total employment counts in the construction industry is positively correlated to median housing prices in five of the eight economic regions.
- Median rent prices and median home prices are positively correlated. Increases in median rent prices may encourage homeownership as higher rents may persuade motivated buyers to consider investing in a home instead of paying high rents, which increases housing demand, increasing median home prices.
- Based on a combination of accuracy and simplicity, the Multiple Linear Regression Models provided a 'better' predictive model of median home prices, with high Adjusted R-Squares and low Root Mean Squared Errors.

Review of Success

Through the analysis of housing and employment data and the development of three different predictive models for estimating median housing prices in eight different economic regions, this report was able to:

1. Identify the industries with the highest hiring rates across metropolitan areas.

2. Confirm the initial assumption that employment by industry influences housing prices and confirm that different industries have a different impact in median housing prices depending on the economic region they are located

3. Mete the key performance indicator of finding a model that could estimate over 51% of the variability of median home prices.

First, the analysis showed that the Retail Trade industry along with the Health care and Social Assistance sector have had the broadest presence across the country since 2012 with the highest average employment rate among the 12 industries being considered in the analysis. Additionally, the data showed that the Health care and social assistance sector has seen a steady hiring of a young workforce (24 to 35).

Second, the analysis showed that median housing prices have shown a steady increase in coastal cities since 2013.

Third, although the predictive models confirmed the initial assumption that median housing prices are influenced by employment counts in particular industries, it also showed that higher employment counts are not necessarily positively correlated with housing prices. For instance, higher employment counts in industries like health care and assistance care were associated with decrease in median housing prices in multiple regions.

Fourth, since the predictive model uses total employment count by industry to predict median housing prices, the output of the model could help state officials (1) determine if and how to invest in infrastructure and housing projects to attract employers of specific industries, (2) review construction policies and regulations to allow more flexible constructions, such as additional buildings in areas with limited housing inventory.

Recommendations for Future Analysis

Future analysis can consider aside from geographic segmentation, quarterly segmentation, to provide insights into seasonal employment trends by industry that could

influence median housing prices at different periods of time. For instance, quarter analysis could show that different industries have a variable impact on housing prices based on the quarter, as their employment counts could substantially differ from quarter to quarter. Thus, multiple linear regression models and random forest models can be run on quarterly datasets to compare results. Since the data would be further segmented, additional years may be needed to have robust datasets for the models.

Additionally, although the total employment counts by industry provided low RMSEs and high adjusted R-squared in all the predictive models, indicating that together the employment counts by industry could explain on average over 75% of the total variability in (logarithm) median housing prices, a combination of a diverse set of economic predictors could be considered to see if they have a greater impact in predicting housing prices. Thus, average quarterly income data from the U.S. Census Bureau, gross domestic product data by region from the Bureau of Economic Analysis, building permit counts by quarter and state from the U.S. Census Bureau, and data on quarterly new residential construction starts and completions from the U.S. Census Bureau, could be considered in future models to see if they can improve the accuracy of the models.

References

Amadeo, K. (2018, June 14). *U.S. retail sales report, current statistics, and recent trends*. The balance. Retrieved from https://www.thebalance.com/u-s-retail-sales-statistics-and-trends-3305717

Bureau of Economic Analysis. (n.d.). Regional economic accounts: About regional. U.S. Department of Commerce. Retrieved from https://www.bea.gov/regional/docs/regions.cfm

Bostic, R. (2017, March). *The affordable housing crisis in Los Angeles* (Research and survey report). Retrieved from University of Southern California website: https://bedrosian.usc.edu/featured/major-l-a-employers-raise-concerns-that-high-cost-of-living-is-a-barrier-to-attracting-and-retaining-top-talent/

Bun, Y. (2012, March 12). *Zillow rent index methodology*. Zillow. Retrieved from https://www.zillow.com/research/zillow-rent-index-methodology-2393/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (n.d.), *CRISP-DM 1.0.* Document posted in University of Maryland University College DATA 670 9041 online classroom, retrieved from: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

Daniels, J. (2018, March 19). Californians fed up with housing costs and taxes are fleeing state in big numbers. *CNBC*. Retrieved from https://www.cnbc.com/2018/03/19/californians-fed-up-with-housing-costs-and-taxes-are-fleeing-state.html

Freddie Mac Multifamily. (2017, November 02). Rental affordability is worsening. Freddie Mac Multifamily. Retrieved from https://mf.freddiemac.com/research/insight/20171103-rental-affordability.html

Garfield, L. (2018, April 11). Amazon's HQ2 is expected to bring soaring housing prices — here are the cities that could be hit hardest. *Business Insider Deutschland*. Retrieved from https://www.businessinsider.de/amazon-hq2-cities-rent-and-home-prices-effect-2018-4?r=US&IR=T

Geron, A. (2017). Hands-on machine learning with Scikit-learn & Tenserflow. Sebastopol, CA: O'Reilly Media.

Gorell, M. (2018, March 28). Fast rises housing prices. A problem for those with lower income and could hurt Utah's competitiveness according to a new study. Retrieved from https://www.sltrib.com/news/business/2018/03/21/fast-rising-housing-prices-a-problem-for-those-with-lower-income-and-could-hurt-utahs-competitiveness-according-to-new-u-study/

Gudell, S. (2016, October 20). *Q3 2016 Market Report: High demand continues to put pressure on home values*. *Zillow*. Retrieved from https://www.zillow.com/research/september-2016-market-report-13641/

Hansen, L. (2018, February 28). Bay Area homes deliver record-breaking returns. *The Mercury News*. Retrieved from https://www.mercurynews.com/2018/02/28/bay-area-home-prices-continue-to-rise-in-record-breaking-streak/

IBM. (n.d.). *CRISP-DM*. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm?pos=2

Investopedia. (n.d.). Retail sales. Retrieved from https://www.investopedia.com/terms/r/retail-sales.asp

Harris, K. (2018, February 8). *A business leader's perspective on meeting regional housing needs*. Brookings. https://www.brookings.edu/blog/the-avenue/2018/02/08/a-business-leaders-perspective-on-meeting-regional-housing-needs/

Honaker, J., King, G., Blackwell, M. (2018, May 7). AMELIA II: A program for missing data. Retrieved from https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf

Kabacoff, R. (2015). *R in Action*. Shelter Island, NY: Manning publications.

Kneebone, E., Snyderman, R., Murray, C. (2017, October 17). *Advancing regional solutions to address America's housing affordability crisis*. Brookings. Retrieved from https://www.brookings.edu/blog/the-avenue/2017/10/19/advancing-regional-solutions-to-address-americas-housing-affordability-crisis/

Kolko, J. (2018, May 17). The Jobs priced out of expensive metros. Retrieved from https://www.hiringlab.org/2018/05/17/jobs-priced-expensive-metros/

LinkedIn. (2017, December 7). *LinkedIn workforce report – United States – December 2017*. Retrieved from https://economicgraph.linkedin.com/resources/linkedin-workforce-report-december-2017

Mejia, Z. (2017, July 27). The 10 best cities for getting a job in tech beyond Silicon Valley. *CNBC*. Retrieved from https://www.cnbc.com/2017/07/27/tech-jobs-silicon-valley.html

Missouri Census Data Center (n.d.). Metadata for dataset. Retrieved from http://mcdc.missouri.edu/data/georef/zcta_master.Metadata.html

Olick, D. (2018, May 29). Run-up in home prices is 'not sustainable': Realtors' chief economist. *CNBC*. Retrieved from https://www.cnbc.com/2018/05/29/run-up-in-home-prices-is-not-sustainable-realtors-chief-economist.html

Olick, D. (2018, April 26). These are the 5 worst housing markets for millennials. *CNBC*. Retrieved from https://www.cnbc.com/2018/04/26/these-are-the-5-worst-housing-markets-for-millennials.html

Passy, J. (2018, January 19). What Amazon's HQ2 means for homeowners, home buyers and renters in the chosen city. *MarketWatch.* Retrieved from https://www.marketwatch.com/story/what-amazons-hq2-means-for-homeowners-home-buyers-and-renters-in-the-chosen-city-2017-10-20

Quicken Loans (n.d.). Mortgage glossary: housing expense ratio. Retrieved from https://www.quickenloans.com/mortgage-glossary/housing-expense-ratio

SAS. (n.d.). Using the Variable Selection Node. *SAS*. Retrieved from http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#p19a2aoz75ivw7n15rh1tju5agwx.htm

Salinas, S. (2017, December 27). Silicon Valley will soon see a 'mass migration' of tech companies and talent, says Redfin CEO. *CNBC*. Retrieved from https://www.cnbc.com/2017/12/27/silicon-valley-will-soon-see-mass-migration-redfin-ceo-says.html

Salinas, S. (May 31, 2018). Facebook says it needs to address high-priced housing 'if we're going to remain a company in Silicon Valley. *CNBC*. Retrieved from

https://www.cnbc.com/2018/05/31/facebook-on-silicon-valley-housing-this-is-an-existential-issue.html

Schuetz, J. (2018, January 19). *Which metros have enough housing capacity to absorb Amazon's HQ2?* Brookings. Retrieved from https://www.brookings.edu/blog/the-avenue/2018/01/19/which-metros-have-enough-housing-capacity-to-absorb-amazons-hq2/?utm_campaign=Brookings%20Brief&utm_source=hs_email&utm_medium=email&utm_content=60103947

Sullivan, B. (2017, October 30). Why are millennials moving to these small towns? *MarketWatch*. Retrieved from https://www.marketwatch.com/story/why-are-millennials-moving-to-these-small-towns-2017-10-30

Turak, N. (2018, 12 April). A Silicon Valley house that burned out two years ago is now on the market for $800,000. *CNBC*. Retrieved from https://www.cnbc.com/2018/04/12/a-burned-out-silicon-valley-house-is-now-on-the-market-for-800000.html

U.S. Census Bureau, Center for Economic Studies. (n.d.). *Quarterly Workforce Indicators (QWI)* [Datasets in CSV format by state]. Retrieved from https://ledextract.ces.census.gov/static/data.html

U.S. Census Bureau, U.S. Department of Commerce. (n.d.). *American Community Survey (ACS) demographic and housing estimates* [Datasets in CSV format]. Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YR_DP05&src=pt

U.S. Bureau of Labor Statistics. (2016, August 5). *Major industries with highest employment, by state, 1990-2015*. Retrieved from https://www.bls.gov/opub/ted/2016/major-industries-with-highest-employment-by-state.htm

U.S. Department of Housing and Urban Development. HUD USPS zip code crosswalk files [Datasets in .xlsx format and variable descriptions]. Office of Policy Development and Research. Retrieved from https://www.huduser.gov/portal/datasets/usps_crosswalk.html#codebook

Vivas, J. (2018, February, 23). *Is Salt Lake City, the new Denver?* Realtor.com. Retrieved from https://www.realtor.com/research/salt-lake-city-the-new-denver/

Zillow. (n.d.) *Zestimate*. Retrieved from https://www.zillow.com/zestimate/

Zillow Research. (n.d.). County and Metro Data [Datasets in CSV format]. Retrieved from https://www.zillow.com/research/data/

Zillow Research. (2014, January 3). *Zillow home value index: methodology*. Retrieved from https://www.zillow.com/research/zhvi-methodology-6032/

Zillow Research. (2017, June 28). *Why is home buying demand so high? partly because buying conditions (in theory) are so good*. Retrieved from https ://www.zillow.com/research/why-is-home-buying-demand-high-15781/

Zucchi, K. (n.d.). West coast vs. east coast economy. *Investopedia.* Retrieved from https://www.investopedia.com/articles/investing/052715/west-coast-vs-east-coast-economy.asp

Appendix

Table 1. Name and description of Variables for Quarterly Workforce Indicators Dataset

Variable Name	Description	Data Type	Notes
CBSA	Core Based Statistical Areas codes	categorical	
State	state code. 01-56	categorical	
quarter	Numerical - quarter being reported 1 to 4	categorical	
year	year being reported 2012 to 2017	categorical	
industry	industry code	categorical	23 Construction 31-33 Manufacturing 44-45 Retail Trade 48-49 Transportation and Warehousing 51 Information 52 Finance and Insurance 54 Professional, Scientific, and Technical Services 61 Educational Services 62 Health Care and Social Assistance 72 Accommodation and Food Services 81 Other Services (except Public Administration) 92 Public Administration
age_group	age group codes. This variable provides information on the workforce demographics by industry.	categorical	A03 for 22-24 A04 for 25-34 A05 for 35-44 A06 for 45-54
FrmJbC	Difference between jobs gain and jobs lost at companies	numeric	
sFrmJbC	Flag status for FrmJbC	categorical	Accompanying code with information on why items are missing in this variable
Emps	Counts for full quarter stable employment	numeric	
sEmps	Flag status for Emps	categorical	accompanying code with information on why items are missing in this variable

FrmJbCS	Net growth in jobs that lasts a full quarter	numeric	
EmpTotal	Count of people employed by a company at any time during the quarter	numeric	
sEmpTotal	Flag status for EmpTotal	categorical	Accompanying code with information on attribute values
sSep	Flag for Separations: Counts	categorical	accompanying code with information on why items are missing in this variable
Sep	Separations: Counts	integer	
HirAs	Hires All (Stable): Counts (Flows into Full-QuarterEmployment)	integer	
sHirAs	Flag for Hires All (Stable): Counts (Flows into Full-QuarterEmployment)	categorical	accompanying code with information on why items are missing in this variable
sHirN	Flag for Hires New: Counts	categorical	accompanying code with information on why items are missing in this variable
HirN	Hires New: Counts	integer	
Payroll	Total Quarterly Payroll: Sum	integer	
sPayroll	Flag for Total Quarterly Payroll: Sum	categorical	accompanying code with information on why items are missing in this variable
HirR	Estimated number of workers who returned to the same employer	numeric	
sHirR	Flag status for HirR	categorical	accompanying code with information on why items are missing in this variable
HirNS	Estimated number of workers who started a new job	numeric	
sHirNS	Flag status for HirNS	categorical	accompanying code with information on why items are missing in this variable
EarnS	Full Quarter Employment (Stable): Average Monthly Earnings	integer	

sEarnS	Flag for Full Quarter Employment (Stable): Average Monthly Earnings	categorical	accompanying code with information on why items are missing in this variable
Emp	Counts: Beginning of quarter employments.	integer	
sEmp	Flag for Beginning-of-Quarter Employment: Counts	categorical	accompanying code with information on why items are missing in this variable
EmpS	Full-Quarter Employment (Stable): Counts	integer	
sEmpS	Flag for Full-Quarter Employment (Stable): Counts	categorical	accompanying code with information on why items are missing in this variable
HirA	Hires All: Counts (Accessions)	integer	
sHirA	Flag for Hires All: Counts (Accessions)	categorical	accompanying code with information on why items are missing in this variable
EarnHiraS	Average monthly earnings all hires	integer	
sEarnHiraS	Flag status for EarnHiraS	categorical	accompanying code with information on why items are missing in this variable
EarnHirnS	Average monthly earnings for new hires	integer	
EarnSeps	Average monthly earnings of separations	integer	
sEarnSeps	Flag status for EarnSeps	categorical	accompanying code with information on why items are missing in this variable
FrmJbGnS	Estimated number of jobs gained at firms	integer	
sFrmJbGnS	Flag status for FrmJbGnS	categorical	accompanying code with information on why items are missing in this variable
FrmJblSs	Estimated number of jobs lost at firms	integer	
sFrmJblSs	Flag status for FrmJblSs	categorical	accompanying code with information on why items are missing in this variable

Hiraendrepl	Count of replacement hires.	integer	
shiraendrepl	Flag status for Hiraendrepl	categorical	accompanying code with information on why items are missing in this variable

Table 1. Variable names and description were retrieved from the Quarterly Workforce Indicators 101 online documentation, retrieved from the U.S. Census Bureau, Center for Economic Studies, LEHD.

Figure 1. Output Linear Regression Model Farwest region

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	2.2041	0.1188	18.55	<.0001	1.9712	2.4369
LOG_Emp23Total		1	0.0408	0.0120	3.41	0.0007	0.0173	0.0642
LOG_Emp31_33Total		1	0.0292	0.00512	5.71	<.0001	0.0192	0.0392
LOG_Emp48_49Total		1	-0.0814	0.00656	-12.40	<.0001	-0.0943	-0.0685
LOG_Emp54Total		1	0.0385	0.00855	4.50	<.0001	0.0218	0.0553
LOG_Emp72Total		1	0.0795	0.00871	9.12	<.0001	0.0624	0.0966
LOG_Emp81total		1	-0.0770	0.0109	-7.05	<.0001	-0.0984	-0.0556
LOG_Emp92		1	-0.0359	0.00609	-5.90	<.0001	-0.0478	-0.0240
LOG_medianRent		1	1.4020	0.0171	82.20	<.0001	1.3686	1.4354
TI_stateName3	0	1	-0.0812	0.0128	-6.34	<.0001	-0.1063	-0.0561

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	-5482.71	.	.
ASE	Average Squared Error	0.01	0.014	0.015
AVERR	Average Error Function	0.01	0.014	0.015
DFE	Degrees of Freedom for Error	1246.00	.	.
DFM	Model Degrees of Freedom	10.00	.	.
DFT	Total Degrees of Freedom	1256.00	.	.
DIV	Divisor for ASE	1256.00	419.000	419.000
ERR	Error Function	15.71	5.910	6.293
FPE	Final Prediction Error	0.01	.	.
MAX	Maximum Absolute Error	0.48	0.448	0.439
MSE	Mean Square Error	0.01	0.014	0.015
NOBS	Sum of Frequencies	1256.00	419.000	419.000
_NW_	Number of Estimate Weights	10.00	.	.
RASE	Root Average Sum of Squares	0.11	0.119	0.123
RFPE	Root Final Prediction Error	0.11	.	.
RMSE	Root Mean Squared Error	0.11	0.119	0.123
SBC	Schwarz's Bayesian Criterion	-5431.35	.	.
SSE	Sum of Squared Errors	15.71	5.910	6.293
SUMW	Sum of Case Weights Times Freq	1256.00	419.000	419.000

Figure 1. Output Linear Regression Model Farwest region.

Figure 2. Sample of observed vs. predicted data Linear Regression Model Farwest region

Obs #	Obse...	CountyName	cbsa	yrQtr	medianValue	Transformed medianValue	Predicted: LOG_me...
1	2	Los Angeles	31...	01Jan1...	381200	12.85108	12.94099
2	6	San Francisco...	41...	01Jan1...	472000	13.06474	13.05748
3	8	Riverside	40...	01Jan1...	186500	12.13619	12.37909
4	9	San Bernardin...	40...	01Jan1...	186500	12.13619	12.37909
5	10	Pierce	42...	01Jan1...	254900	12.44863	12.49847
6	25	Washington	38...	01Jan1...	215700	12.28165	12.30499
7	31	Yolo	40...	01Jan1...	210100	12.25534	12.30372
8	39	San Joaquin	44...	01Jan1...	156900	11.96337	12.14333
9	41	Stevens	44...	01Jan1...	145800	11.89	11.90467
10	42	Pend Oreille	44...	01Jan1...	145800	11.89	11.90467
11	46	Storey	39...	01Jan1...	155700	11.95569	12.2123
12	47	Washoe	39...	01Jan1...	155700	11.95569	12.2123
13	64	Whatcom	13...	01Jan1...	238400	12.38171	12.19588
14	66	Shasta	39...	01Jan1...	160700	11.9873	12.0394
15	68	Sutter	49...	01Jan1...	140700	11.85439	12.1293

Figure 2. Sample of observed vs. predicted values.

Figure 3. Fit Statistics of Decision Tree Farwest region

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error
LOG_medianRent	8467	0.110889	0.170216	0.111626	0.160424	0.103970	0.160117
LOG_Emp92	6796	0.005989	0.023727	0.003626	0.011956	0.003937	0.012755
LOG_Emp72Total	4000	0.029416	0.044568	0.026571	0.035522	0.024019	0.035961
LOG_Emp54Total	3763	0.036597	0.049251	0.033151	0.040276	0.026015	0.034571
LOG_Emp81total	3573	0.004134	0.013654	0.002683	0.007152	0.002749	0.007814
LOG_Emp48_49Total	2506	0.008471	0.021089	0.006621	0.015180	0.004938	0.012107
LOG_Emp31_33Total	2387	0.021389	0.035057	0.019220	0.027342	0.015270	0.024786
LOG_Emp23Total	1260	0.006661	0.013475	0.006037	0.009999	0.005465	0.009484
TI_stateName3	11	0.000165	0.000235	0.000196	0.000293	0.000238	0.000353

Input variables

LOG_Emp23Total	8	Input	Interval	Numeric
LOG_Emp31_33Total	8	Input	Interval	Numeric
LOG_Emp48_49Total	8	Input	Interval	Numeric
LOG_Emp54Total	8	Input	Interval	Numeric
LOG_Emp72Total	8	Input	Interval	Numeric
LOG_Emp81total	8	Input	Interval	Numeric
LOG_Emp92	8	Input	Interval	Numeric
LOG_medianRent	8	Input	Interval	Numeric
TI_stateName3	8	Input	Class	Numeric

Data Role=VALIDATE Target Variable=LOG_medianValue Target Label=Transformed medianValue

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
13.553 - 13.653	13.5951	13.5931	8	13.6032
13.454 - 13.553	13.5240	13.5254	1	13.5037
13.354 - 13.454	13.4008	13.3963	3	13.4042
13.255 - 13.354	13.3185	13.3115	4	13.3047
13.155 - 13.255	13.1964	13.1932	12	13.2052
13.056 - 13.155	13.1097	13.1056	15	13.1056
12.956 - 13.056	12.9687	13.0010	8	13.0061
12.857 - 12.956	12.8793	12.9057	8	12.9066
12.757 - 12.857	12.7949	12.7807	12	12.8071
12.658 - 12.757	12.6998	12.7005	25	12.7076
12.558 - 12.658	12.6306	12.6185	24	12.6081
12.459 - 12.558	12.5289	12.5108	21	12.5086
12.359 - 12.459	12.4186	12.4101	39	12.4091
12.260 - 12.359	12.3255	12.3080	49	12.3096
12.160 - 12.260	12.2200	12.2144	34	12.2101
12.061 - 12.160	12.1246	12.1125	39	12.1106
11.961 - 12.061	12.0023	12.0225	37	12.0111
11.862 - 11.961	11.8895	11.9109	40	11.9116
11.762 - 11.862	11.8022	11.8184	23	11.8121
11.663 - 11.762	11.6977	11.7195	17	11.7126

Figure 3. Output of Decision Tree model. Top to Bottom: Variable importance, Fit statistics and observed vs. predicted value range.

Figure 4. Output of Linear regression model of Great Lakes Region.

Model Fit Statistics

R-Square	0.7622	Adj R-Sq	0.7608
AIC	-7891.2624	BIC	-7889.0779
SBC	-7818.0853	C(p)	11.5017

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	3.7342	0.1536	24.31	<.0001	3.4331	4.0353
LOG_Emp23Total	1	0.1222	0.0129	9.51	<.0001	0.0970	0.1474
LOG_Emp31_33Total	1	0.0658	0.00810	8.12	<.0001	0.0499	0.0817
LOG_Emp44_45Total	1	-0.3292	0.0301	-10.93	<.0001	-0.3882	-0.2702
LOG_Emp48_49Total	1	-0.0464	0.00708	-6.55	<.0001	-0.0602	-0.0325
LOG_Emp51total	1	-0.0514	0.00829	-6.20	<.0001	-0.0676	-0.0351
LOG_Emp52Total	1	0.0945	0.0103	9.21	<.0001	0.0744	0.1147
LOG_Emp61Total	1	0.0645	0.0105	6.12	<.0001	0.0439	0.0852
LOG_Emp62Total	1	-0.0588	0.0160	-3.67	0.0002	-0.0903	-0.0274
LOG_Emp72Total	1	0.1966	0.0180	10.90	<.0001	0.1613	0.2320
LOG_Emp81total	1	-0.0548	0.0180	-3.04	0.0024	-0.0901	-0.0195
LOG_Emp92	1	0.0281	0.00977	2.87	0.0041	0.00893	0.0472
LOG_medianRent	1	1.1016	0.0216	51.02	<.0001	1.0592	1.1439

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	-7891.26	.	.
ASE	Average Squared Error	0.02	0.024	0.021
AVERR	Average Error Function	0.02	0.024	0.021
DFE	Degrees of Freedom for Error	2044.00	.	.
DFM	Model Degrees of Freedom	13.00	.	.
DFT	Total Degrees of Freedom	2057.00	.	.
DIV	Divisor for ASE	2057.00	935.000	748.000
ERR	Error Function	43.82	22.389	15.809
FPE	Final Prediction Error	0.02	.	.
MAX	Maximum Absolute Error	0.66	0.702	0.646
MSE	Mean Square Error	0.02	0.024	0.021
NOBS	Sum of Frequencies	2057.00	935.000	748.000
_NW_	Number of Estimate Weights	13.00	.	.
RASE	Root Average Sum of Squares	0.15	0.155	0.145
RFPE	Root Final Prediction Error	0.15	.	.
RMSE	Root Mean Squared Error	0.15	0.155	0.145
SBC	Schwarz's Bayesian Criterion	-7818.09	.	.
SSE	Sum of Squared Errors	43.82	22.389	15.809
SUMW	Sum of Case Weights Times Freq	2057.00	935.000	748.000

Figure 4. Top output of fit statistics linear regression model Grate Lakes Region

Figure 5. Output of decision tree for Great Lakes region

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_medianRent	Transformed medianRent	10	1.0000	1.0000	1.0000
LOG_Emp72Total	Transformed Emp72Total	2	0.1978	0.1968	0.9947
TI_stateName1	stateName:IL	1	0.1720	0.1545	0.8987
LOG_Emp44_45Total	Transformed Emp44_45Total	1	0.1462	0.1459	0.9983
LOG_Emp31_33Total	Transformed Emp31_33Total	1	0.1436	0.1346	0.9377
LOG_Emp92	Transformed Emp92	3	0.1199	0.1682	1.4032
LOG_Emp81total	Transformed Emp81total	1	0.0877	0.0772	0.8799
TI_stateName4	stateName:OH	2	0.0721	0.0834	1.1567
TI_stateName5	stateName:WI	1	0.0708	0.0609	0.8611
LOG_Emp52Total	Transformed Emp52Total	2	0.0701	0.0668	0.9532
LOG_mortgageRate	Transformed mortgageRate	1	0.0063	0.0052	0.8366

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	2871.00	1305.00	1044.00
MAX	Maximum Absolute Error	0.51	0.57	0.54
SSE	Sum of Squared Errors	29.93	14.44	12.10
ASE	Average Squared Error	0.01	0.01	0.01
RASE	Root Average Squared Error	0.10	0.11	0.11
DIV	Divisor for ASE	2871.00	1305.00	1044.00
DFT	Total Degrees of Freedom	2871.00	.	.

Data Role=VALIDATE Target Variable=LOG_medianValue Target Label=Transformed medianValue

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
12.174 - 12.237	12.1867	12.1953	89	12.2055
12.112 - 12.174	12.1228	12.1240	39	12.1434
12.050 - 12.112	12.0850	12.0845	56	12.0813
11.988 - 12.050	11.9917	11.9884	76	12.0192
11.926 - 11.988	11.9661	11.9706	43	11.9571
11.864 - 11.926	11.8652	11.8745	53	11.8950
11.740 - 11.802	11.7458	11.7801	37	11.7708

Figure 5. Variable importance, fit statistics and range of predicted values of decision tree model to predict the (logarithm) price of median house prices.

Figure 6. Output of Random Forest Great Lakes Region

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

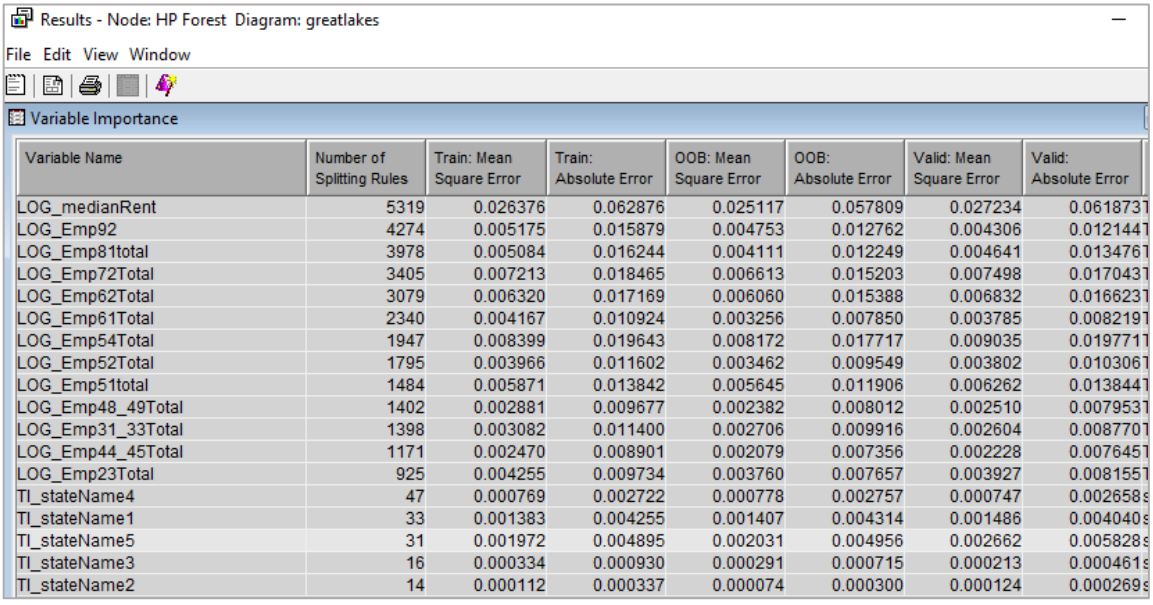
Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.00	0.00	0.00
DIV	Divisor for ASE	2871.00	1305.00	1044.00
MAX	Maximum Absolute Error	0.12	0.24	0.27
NOBS	Sum of Frequencies	2871.00	1305.00	1044.00
RASE	Root Average Squared Error	0.01	0.03	0.03
SSE	Sum of Squared Errors	0.51	1.14	0.97

Data Role=VALIDATE Target Variable=LOG_medianValue Target Label=Transformed medianValue

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
12.238 - 12.309	12.3269	12.2831	3	12.2737
12.168 - 12.238	12.1967	12.1954	57	12.2031
12.097 - 12.168	12.1381	12.1380	79	12.1325
12.027 - 12.097	12.0657	12.0639	53	12.0619
11.956 - 12.027	11.9939	11.9917	90	11.9914
11.885 - 11.956	11.9277	11.9263	37	11.9208
11.815 - 11.885	11.8460	11.8479	86	11.8502
11.744 - 11.815	11.7696	11.7686	153	11.7796
11.674 - 11.744	11.7213	11.7179	129	11.7090
11.603 - 11.674	11.6513	11.6419	160	11.6385
11.533 - 11.603	11.5749	11.5710	71	11.5679
11.462 - 11.533	11.4941	11.4892	63	11.4973
11.391 - 11.462	11.4250	11.4297	57	11.4267
11.321 - 11.391	11.3525	11.3644	66	11.3561
11.250 - 11.321	11.2585	11.2819	57	11.2856
11.180 - 11.250	11.1955	11.2184	76	11.2150
11.109 - 11.180	11.1493	11.1525	49	11.1444
11.039 - 11.109	11.0809	11.0931	10	11.0738
10.968 - 11.039	10.9909	11.0285	4	11.0032
10.897 - 10.968	10.9041	10.9143	5	10.9326

Figure 6. Output of Random Forest for Great Lakes Region showing Fit statistics and Mean target values vs. mean predicted values.

Figure 7. Output of Random Forest Great Lakes Region



Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error
LOG_medianRent	5319	0.026376	0.062876	0.025117	0.057809	0.027234	0.0618731
LOG_Emp92	4274	0.005175	0.015879	0.004753	0.012762	0.004306	0.0121441
LOG_Emp81total	3978	0.005084	0.016244	0.004111	0.012249	0.004641	0.0134761
LOG_Emp72Total	3405	0.007213	0.018465	0.006613	0.015203	0.007498	0.0170431
LOG_Emp62Total	3079	0.006320	0.017169	0.006060	0.015388	0.006832	0.0166231
LOG_Emp61Total	2340	0.004167	0.010924	0.003256	0.007850	0.003785	0.0082191
LOG_Emp54Total	1947	0.008399	0.019643	0.008172	0.017717	0.009035	0.0197711
LOG_Emp52Total	1795	0.003966	0.011602	0.003462	0.009549	0.003802	0.0103061
LOG_Emp51total	1484	0.005871	0.013842	0.005645	0.011906	0.006262	0.0138441
LOG_Emp48_49Total	1402	0.002881	0.009677	0.002382	0.008012	0.002510	0.0079531
LOG_Emp31_33Total	1398	0.003082	0.011400	0.002706	0.009916	0.002604	0.0087701
LOG_Emp44_45Total	1171	0.002470	0.008901	0.002079	0.007356	0.002228	0.0076451
LOG_Emp23Total	925	0.004255	0.009734	0.003760	0.007657	0.003927	0.0081551
TI_stateName4	47	0.000769	0.002722	0.000778	0.002757	0.000747	0.002658s
TI_stateName1	33	0.001383	0.004255	0.001407	0.004314	0.001486	0.004040s
TI_stateName5	31	0.001972	0.004895	0.002031	0.004956	0.002662	0.005828s
TI_stateName3	16	0.000334	0.000930	0.000291	0.000715	0.000213	0.000461s
TI_stateName2	14	0.000112	0.000337	0.000074	0.000300	0.000124	0.000269s

Figure 8. Output of Linear Regression Model Mideast Region

Model Fit Statistics

R-Square	0.9026	Adj R-Sq	0.9020
AIC	-5780.7497	BIC	-5778.6338
SBC	-5721.8588	C(p)	13.8115

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
LOG_Emp23Total	1	4.7903	195.29	<.0001
LOG_Emp31_33Total	1	3.0566	124.61	<.0001
LOG_Emp51total	1	0.1863	7.59	0.0059
LOG_Emp52Total	1	1.1763	47.95	<.0001
LOG_Emp62Total	1	1.1582	47.22	<.0001
LOG_Emp72Total	1	4.1336	168.52	<.0001
LOG_Emp81total	1	4.5791	186.68	<.0001
LOG_Emp92	1	0.7218	29.42	<.0001
LOG_medianRent	1	79.1084	3225.04	<.0001
TI_stateName6	1	1.7839	72.72	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	1.1601	0.1963	5.91	<.0001	0.7753	1.5449
LOG_Emp23Total		1	0.2082	0.0149	13.97	<.0001	0.1790	0.2375
LOG_Emp31_33Total		1	-0.0954	0.00855	-11.16	<.0001	-0.1122	-0.0787
LOG_Emp51total		1	0.0367	0.0133	2.76	0.0059	0.0106	0.0629
LOG_Emp52Total		1	0.0944	0.0136	6.92	<.0001	0.0677	0.1211
LOG_Emp62Total		1	-0.1436	0.0209	-6.87	<.0001	-0.1846	-0.1026
LOG_Emp72Total		1	0.2145	0.0165	12.98	<.0001	0.1821	0.2469
LOG_Emp81total		1	-0.2602	0.0190	-13.66	<.0001	-0.2976	-0.2229
LOG_Emp92		1	-0.0565	0.0104	-5.42	<.0001	-0.0770	-0.0361
LOG_medianRent		1	1.5146	0.0267	56.79	<.0001	1.4623	1.5669
TI_stateName6	0	1	-0.0627	0.00735	-8.53	<.0001	-0.0771	-0.0483

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	-5780.75	.	.
ASE	Average Squared Error	0.02	0.022	0.023
AVERR	Average Error Function	0.02	0.022	0.023
DFE	Degrees of Freedom for Error	1551.00	.	.
DFM	Model Degrees of Freedom	11.00	.	.
DFT	Total Degrees of Freedom	1562.00	.	.
DIV	Divisor for ASE	1562.00	710.000	568.000
ERR	Error Function	38.05	15.533	13.132
FPE	Final Prediction Error	0.02	.	.
MAX	Maximum Absolute Error	0.66	0.607	0.638
MSE	Mean Square Error	0.02	0.022	0.023
NOBS	Sum of Frequencies	1562.00	710.000	568.000
_NW_	Number of Estimate Weights	11.00	.	.
RASE	Root Average Sum of Squares	0.16	0.148	0.152
RFPE	Root Final Prediction Error	0.16	.	.
RMSE	Root Mean Squared Error	0.16	0.148	0.152
SBC	Schwarz's Bayesian Criterion	-5721.86	.	.
SSE	Sum of Squared Errors	38.05	15.533	13.132
SUMW	Sum of Case Weights Times Freq	1562.00	710.000	568.000

Figure 8. Fit statistics and linear regression coefficients of Mideast Region model.

Figure 9. Output of Decision Tree Model Mideast Region

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
LOG_medianValue	Transformed medianVal...	_NOBS_	Sum of Frequencies	1562	710	568
LOG_medianValue	Transformed medianVal...	_MAX_	Maximum Absolute E...	0.837789	0.792887	0.841294
LOG_medianValue	Transformed medianVal...	_SSE_	Sum of Squared Erro...	24.34747	11.69599	11.93639
LOG_medianValue	Transformed medianVal...	_ASE_	Average Squared Error	0.015587	0.016473	0.021015
LOG_medianValue	Transformed medianVal...	_RASE_	Root Average Squar...	0.124849	0.128348	0.144965
LOG_medianValue	Transformed medianVal...	_DIV_	Divisor for ASE	1562	710	568
LOG_medianValue	Transformed medianVal...	_DFT_	Total Degrees of Fre...	1562	.	.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_medianRent	Transforme...	10	1.0000	1.0000	1.0000
LOG_Emp72Total	Transforme...	1	0.2169	0.2150	0.9916
LOG_Emp52Total	Transforme...	2	0.1138	0.1443	1.2674
LOG_Emp92	Transforme...	2	0.0991	0.0848	0.8550
LOG_Emp81total	Transforme...	1	0.0826	0.0791	0.9578
LOG_Emp31_33Total	Transforme...	3	0.0731	0.0621	0.8493
LOG_Emp51total	Transforme...	1	0.0391	0.0192	0.4922
LOG_Emp62Total	Transforme...	1	0.0038	0.0046	1.1986
LOG_Emp54Total	Transforme...	0	0.0000	0.0000	.
LOG_Emp23Total	Transforme...	0	0.0000	0.0000	.
TI_stateName1	stateName:...	0	0.0000	0.0000	.
TI_stateName6	stateName:...	0	0.0000	0.0000	.

Figure 9. Variable importance, and fit statistics.

Figure 10. Output Decision Tree Mideast Region without Log Transformation of Variables

Target=medianValue Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	1562.00	710.00	568.00
MAX	Maximum Absolute Error	88478.95	88878.95	139278.95
SSE	Sum of Squared Errors	416676827878.34	181885775194.96	202225166066.89
ASE	Average Squared Error	266758532.57	256177148.16	356030221.95
RASE	Root Average Squared Error	16332.74	16005.53	18868.76
DIV	Divisor for ASE	1562.00	710.00	568.00
DFT	Total Degrees of Freedom	1562.00	.	.

File Edit View Window

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
medianRent		11	1.0000	1.0000	1.0000
stateName		2	0.1709	0.1610	0.9419
Emp72Total		1	0.0925	0.0937	1.0131
Emp52Total		2	0.0731	0.0890	1.2172
mortgageRatePercentage		3	0.0287	0.0279	0.9713
Emp31_33Total		1	0.0152	0.0140	0.9254
Emp81total		1	0.0122	0.0148	1.2121
Emp48_49Total		1	0.0058	0.0046	0.7964
Emp44_45Total		1	0.0014	0.0009	0.6158
Emp23Total		0	0.0000	0.0000	.
Emp61Total		0	0.0000	0.0000	.
Emp62Total		0	0.0000	0.0000	.
Emp92		0	0.0000	0.0000	.
Emp51total		0	0.0000	0.0000	.
Emp54Total		0	0.0000	0.0000	.

Figure 10. Fit statistics and variable importance.

Figure 11. Output of Random Forest using transformed data.

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.00	0.001	0.002
DIV	Divisor for ASE	1562.00	710.000	568.000
MAX	Maximum Absolute Error	0.22	0.222	0.466
NOBS	Sum of Frequencies	1562.00	710.000	568.000
RASE	Root Average Squared Error	0.02	0.033	0.041
SSE	Sum of Squared Errors	0.48	0.765	0.956

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error	Label
LOG_medianRent	4325	0.081918	0.136921	0.079024	0.129933	0.084837	0.138888	Transformed m...
LOG_Emp92	3475	0.010863	0.028125	0.010029	0.025528	0.010048	0.026161	Transformed E...
LOG_Emp51total	1945	0.019989	0.030748	0.020353	0.029908	0.021793	0.032480	Transformed E...
LOG_Emp72Total	1945	0.013238	0.023897	0.012339	0.021440	0.012381	0.021531	Transformed E...
LOG_Emp52Total	1898	0.009887	0.019845	0.009396	0.018467	0.010032	0.019395	Transformed E...
LOG_Emp81total	1894	0.014102	0.026414	0.013457	0.024699	0.014693	0.026267	Transformed E...
LOG_Emp54Total	1831	0.055578	0.078829	0.052882	0.074910	0.059014	0.083352	Transformed E...
LOG_Emp62Total	1806	0.008076	0.017509	0.007687	0.016312	0.007556	0.015492	Transformed E...
LOG_Emp31_33To...	1493	0.010142	0.022261	0.010075	0.021235	0.010171	0.020787	Transformed E...
LOG_Emp23Total	1154	0.023194	0.032083	0.021317	0.028857	0.023956	0.031910	Transformed E...
TI_stateName6	31	0.002929	0.006842	0.002850	0.006628	0.002890	0.007085	stateName:PA
TI_stateName1	12	0.000938	0.001893	0.001208	0.002200	0.001392	0.002506	stateName:DC

Data Role=VALIDATE Target Variable=LOG_medianValue Target Label=Transformed medianValue

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
12.769 - 12.863	12.8062	12.8045	164	12.8159
12.676 - 12.769	12.7325	12.7335	78	12.7226
12.583 - 12.676	12.6568	12.6560	17	12.6293
12.489 - 12.583	12.5062	12.5145	1	12.5360
12.396 - 12.489	12.4494	12.4448	12	12.4427
12.303 - 12.396	12.3595	12.3595	20	12.3494
12.209 - 12.303	12.2511	12.2507	24	12.2561
12.116 - 12.209	12.1682	12.1644	80	12.1628
12.023 - 12.116	12.0828	12.0697	41	12.0695
11.930 - 12.023	11.9773	11.9649	22	11.9762
11.836 - 11.930	11.8999	11.9002	30	11.8829
11.743 - 11.836	11.7420	11.7763	25	11.7896
11.650 - 11.743	11.7031	11.7034	79	11.6963
11.556 - 11.650	11.5946	11.6082	41	11.6030
11.463 - 11.556	11.5150	11.5268	39	11.5097
11.370 - 11.463	11.3488	11.4087	10	11.4164
11.276 - 11.370	11.2486	11.3224	8	11.3231
11.183 - 11.276	11.1927	11.2202	9	11.2298
11.090 - 11.183	11.0250	11.1319	4	11.1365
10.997 - 11.090	11.0068	11.0264	6	11.0432

Figure 11. Fit statistics, variable importance and predictive ranges for median housing prices in the Mideast region.

Figure 12. Output of Random Forest model for Mideast Region with untransformed data.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
medianValue		_ASE_	ared Err...	25804283	39318892	37230114
medianValue		_DIV_	Divisor for ASE	1562	710	568
medianValue		_MAX_	Maximum Absolute ...	47183.33	47837.48	59483.88
medianValue		_NOBS_	Sum of Frequencies	1562	710	568
medianValue		_RASE_	Root Average Squar...	5079.792	6270.478	6101.648
medianValue		_SSE_	Sum of Squared Err...	4.031E10	2.792E10	2.115E10

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error	Label
medianRent	1054	2.933E9	24551	2.9569E9	24291	3.1487E9	25548	
mortgageRatePercentage	777	6747663	494.847541	2369954	418.377575	4389188	470.605524	
Emp92	744	1.8154E8	2523.534492	1.9331E8	2461.373420	1.7997E8	2413.691130	
Emp81total	614	2.3695E8	2729.558008	2.1921E8	2547.929558	2.4913E8	2757.184713	
Emp72Total	564	5.5006E8	4742.592722	5.0989E8	4426.451772	5.5719E8	4775.159473	
Emp62Total	544	1.5772E8	2247.413899	1.3837E8	2018.694922	1.4415E8	2028.169080	
Emp31_33Total	469	4.3942E8	4299.965328	4.1596E8	4000.036456	4.3024E8	4163.652072	
Emp54Total	413	8.2047E8	6536.072553	8.2899E8	6588.524536	9.0357E8	7064.356691	
Emp51total	372	2.5601E8	2424.980255	2.2833E8	2212.043190	2.6672E8	2426.162627	
Emp52Total	364	65196004	1074.287304	55003794	1026.472058	62000815	1055.893849	
Emp48_49Total	353	1.2391E8	1496.024809	95917108	1296.396333	1.1783E8	1351.574425	
Emp44_45Total	275	1.2667E9	9040.331269	1.2446E9	8849.956658	1.3445E9	9451.783251	
Emp61Total	235	2.3254E8	2082.955317	2.2067E8	1978.267536	2.4157E8	2047.033842	
Emp23Total	234	2.3288E9	15627	2.3292E9	15711	2.5532E9	17008	
stateName	121	8.4066E8	8125.342408	7.9323E8	7793.955511	8.1923E8	7953.248560	

Figure 12. Variable importance and fit on untransformed data for the Mideast region.

Figure 13. Models Score Distribution for the Mideast region.

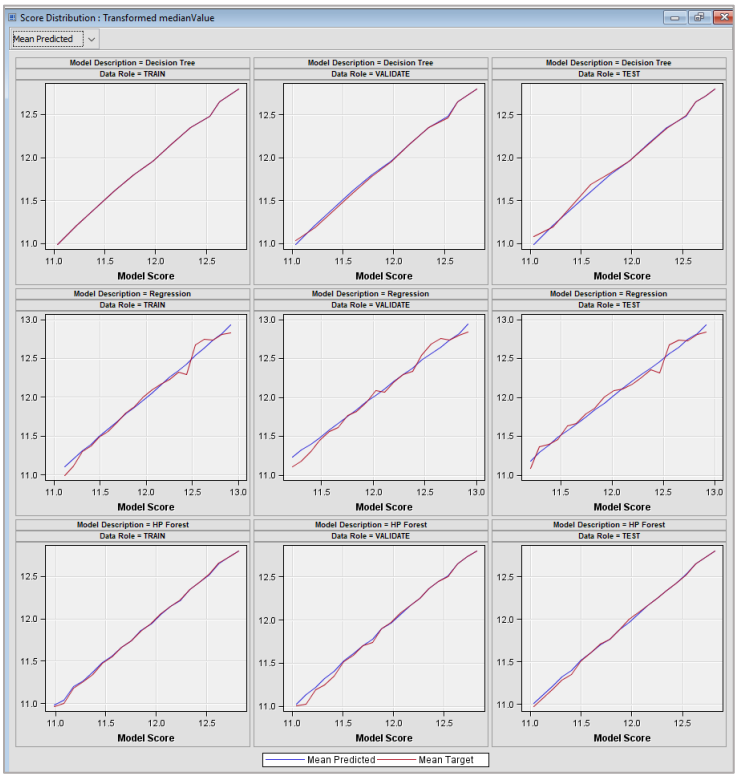


Figure 14. Output Linear Regression Model New England Region

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	-1870.52	.	.
ASE	Average Squared Error	0.01	0.007	0.009
AVERR	Average Error Function	0.01	0.007	0.009
DFE	Degrees of Freedom for Error	397.00	.	.
DFM	Model Degrees of Freedom	10.00	.	.
DFT	Total Degrees of Freedom	407.00	.	.
DIV	Divisor for ASE	407.00	185.000	148.000
ERR	Error Function	3.91	1.209	1.358
FPE	Final Prediction Error	0.01	.	.
MAX	Maximum Absolute Error	0.44	0.322	0.324
MSE	Mean Square Error	0.01	0.007	0.009
NOBS	Sum of Frequencies	407.00	185.000	148.000
_NW_	Number of Estimate Weights	10.00	.	.
RASE	Root Average Sum of Squares	0.10	0.081	0.096
RFPE	Root Final Prediction Error	0.10	.	.
RMSE	Root Mean Squared Error	0.10	0.081	0.096
SBC	Schwarz's Bayesian Criterion	-1830.44	.	.
SSE	Sum of Squared Errors	3.91	1.209	1.358
SUMW	Sum of Case Weights Times Freq	407.00	185.000	148.000

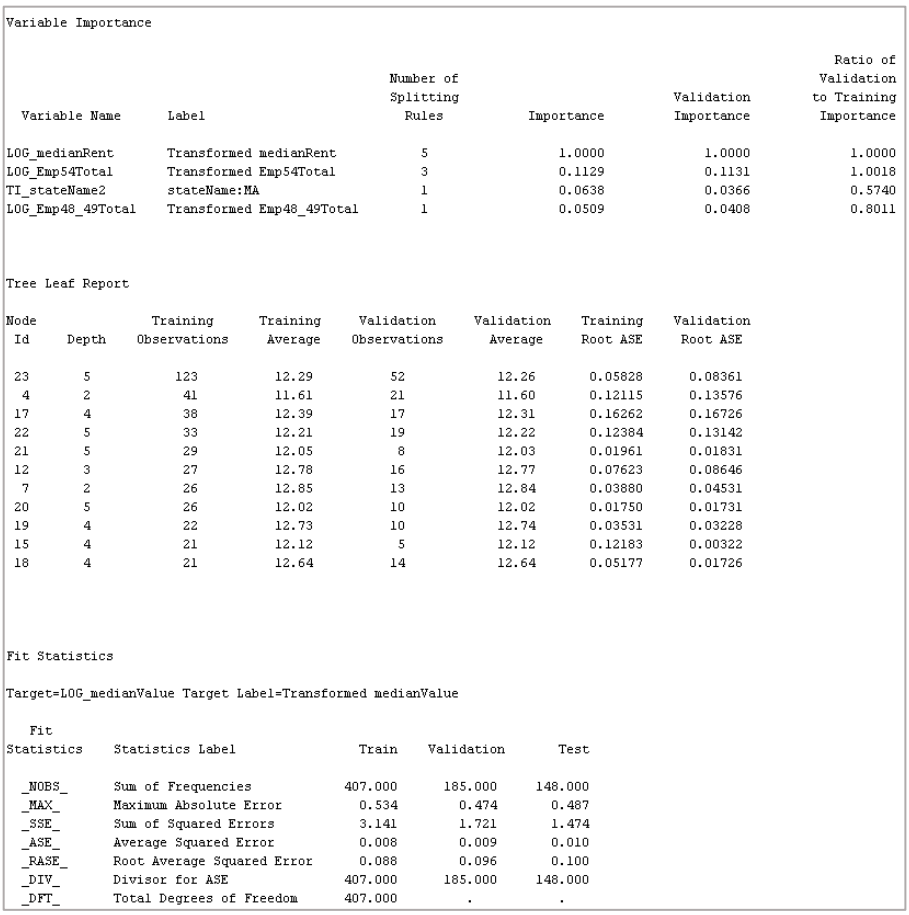
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	4.2016	0.3615	11.62	<.0001	3.4931	4.9102
LOG_Emp23Total		1	0.1136	0.0247	4.60	<.0001	0.0652	0.1619
LOG_Emp44_45Total		1	0.1252	0.0279	4.49	<.0001	0.0706	0.1798
LOG_Emp51total		1	-0.0840	0.0240	-3.50	0.0005	-0.1311	-0.0369
LOG_Emp52Total		1	-0.1089	0.0135	-8.05	<.0001	-0.1355	-0.0824
LOG_Emp54Total		1	0.1415	0.0212	6.69	<.0001	0.1000	0.1830
LOG_Emp62Total		1	-0.1825	0.0245	-7.44	<.0001	-0.2305	-0.1344
LOG_Emp81total		1	0.0689	0.0241	2.86	0.0044	0.0217	0.1161
LOG_medianRent		1	1.0193	0.0487	20.92	<.0001	0.9238	1.1148
TI_stateName3	0	1	0.0325	0.0152	2.14	0.0330	0.00273	0.0623

Model Fit Statistics

R-Square	0.9193	Adj R-Sq	0.9175
AIC	-1870.5237	BIC	-1868.1396
SBC	-1830.4356	C(p)	12.3153

Figure 14. Fit statistics and model coefficients for linear regression model.

Figure 15. Output Decision Tree for New England Region



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_medianRent	Transformed medianRent	5	1.0000	1.0000	1.0000
LOG_Emp54Total	Transformed Emp54Total	3	0.1129	0.1131	1.0018
TI_stateName2	stateName:MA	1	0.0638	0.0366	0.5740
LOG_Emp48_49Total	Transformed Emp48_49Total	1	0.0509	0.0408	0.8011

Tree Leaf Report

Node Id	Depth	Training Observations	Training Average	Validation Observations	Validation Average	Training Root ASE	Validation Root ASE
23	5	123	12.29	52	12.26	0.05828	0.08361
4	2	41	11.61	21	11.60	0.12115	0.13576
17	4	38	12.39	17	12.31	0.16262	0.16726
22	5	33	12.21	19	12.22	0.12384	0.13142
21	5	29	12.05	8	12.03	0.01961	0.01831
12	3	27	12.78	16	12.77	0.07623	0.08646
7	2	26	12.85	13	12.84	0.03880	0.04531
20	5	26	12.02	10	12.02	0.01750	0.01731
19	4	22	12.73	10	12.74	0.03531	0.03228
15	4	21	12.12	5	12.12	0.12183	0.00322
18	4	21	12.64	14	12.64	0.05177	0.01726

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	407.000	185.000	148.000
MAX	Maximum Absolute Error	0.534	0.474	0.487
SSE	Sum of Squared Errors	3.141	1.721	1.474
ASE	Average Squared Error	0.008	0.009	0.010
RASE	Root Average Squared Error	0.088	0.096	0.100
DIV	Divisor for ASE	407.000	185.000	148.000
DFT	Total Degrees of Freedom	407.000	.	.

Figure 16. Output Decision Tree Model for New England Region using Untransformed Data.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
medianRent		6	1.0000	1.0000	1.0000
Emp51total		1	0.1389	0.1354	0.9745
Emp92		1	0.0834	0.0734	0.8804
Emp54Total		1	0.0621	0.0633	1.0199
stateName		1	0.0590	0.0565	0.9574
Emp72Total		1	0.0524	0.0294	0.5604
Emp48_49Total		1	0.0434	0.0326	0.7507

Tree Leaf Report

Node Id	Depth	Training Observations	Training Average	Validation Observations	Validation Average	Training Root ASE	Validation Root ASE
18	4	101	209346.53	48	205652.08	20569.09	24718.12
11	3	38	243431.58	17	224617.65	42421.04	41128.53
24	5	33	217990.91	19	217142.11	5890.23	6169.86
20	4	32	316821.88	24	322000.00	16042.67	26171.07
23	5	29	170796.55	8	168375.00	3333.37	3113.40
22	5	26	166357.69	10	166080.00	2875.78	2859.15
7	2	24	384804.17	10	384810.00	13555.86	14227.61
25	5	22	233218.18	4	234750.00	3976.25	3987.35
14	4	21	98766.67	13	103492.31	3125.98	17987.52
17	4	21	185271.43	5	183700.00	20454.11	1653.30
12	3	20	359640.00	7	363285.71	30508.43	23691.73
15	4	20	123135.00	8	119775.00	7883.93	6866.33
21	4	20	343185.00	12	346591.67	9975.13	10397.65

Fit Statistics

Target=medianValue Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	407.00	185.00	148.00
MAX	Maximum Absolute Error	131953.47	112153.47	116353.47
SSE	Sum of Squared Errors	156630039303.75	87315769728.21	61451801220.59
ASE	Average Squared Error	384840391.41	471977133.67	415214873.11
RASE	Root Average Squared Error	19617.35	21725.03	20376.82
DIV	Divisor for ASE	407.00	185.00	148.00
DFT	Total Degrees of Freedom	407.00	.	.

Figure 16. Output of Decision tree for New England region showing fit statistics, variable importance and range of predicted values.

Figure 17. Random Forest model output for New England Region

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.000	0.001	0.001
DIV	Divisor for ASE	407.000	185.000	148.000
MAX	Maximum Absolute Error	0.066	0.231	0.209
NOBS	Sum of Frequencies	407.000	185.000	148.000
RASE	Root Average Squared Error	0.012	0.033	0.030
SSE	Sum of Squared Errors	0.055	0.203	0.134

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error
LOG_medianRent	1591	0.036563	0.078678	0.035561	0.073041	0.041858	0.084924
LO	1302	0.006448	0.025437	0.005168	0.020718	0.004689	0.019579
LOG_Emp81total	1088	0.017862	0.030506	0.015733	0.026261	0.019019	0.030788
LOG_Emp62Total	913	0.007867	0.026039	0.006401	0.021696	0.007172	0.023925
LOG_Emp61Total	654	0.005716	0.017711	0.006069	0.017099	0.006180	0.016695
LOG_Emp54Total	585	0.008585	0.013172	0.008689	0.012423	0.009804	0.013683
LOG_Emp52Total	393	0.006834	0.013304	0.007041	0.012368	0.007431	0.013451
LOG_Emp51total	375	0.007654	0.018939	0.006126	0.014666	0.007157	0.015639
LOG_Emp48_49Total	286	0.009192	0.014825	0.007331	0.011488	0.007873	0.011243
LOG_Emp23Total	250	0.008076	0.010734	0.009116	0.010586	0.009469	0.011516
LOG_Emp44_45Total	231	0.005221	0.009179	0.004462	0.006766	0.004782	0.007864
TI_stateName2	9	0.000845	0.002598	0.000705	0.002039	0.001091	0.002646
TI_stateName4	3	0.000127	0.000713	0.000183	0.000909	0.000093	0.000356
TI_stateName3	0	0.000000	0	0.000000	0	0.000000	0

Figure 18. Output of comparison of models for the New England Region.

Data Role=Valid

Statistics	HPDMForest	Reg	Tree
Valid: Average Squared Error	0.001	0.007	0.009
Valid: Average Error Function	.	0.007	.
Valid: Divisor for VASE	185.000	185.000	185.000
Valid: Error Function	.	1.209	.
Valid: Maximum Absolute Error	0.231	0.322	0.474
Valid: Mean Square Error	.	0.007	.
Valid: Sum of Frequencies	185.000	185.000	185.000
Valid: Root Average Squared Error	0.033	0.081	0.096
Valid: Root Mean Square Error	.	0.081	.
Valid: Sum of Square Errors	0.203	1.209	1.721
Valid: Sum of Case Weights Times Freq	.	185.000	.

Data Role=Test

Statistics	HPDMForest	Reg	Tree
Test: Lower 95% Conf. Limit for TASE	.	.	.
Test: Upper 95% Conf. Limit for TASE	.	.	.
Test: Average Squared Error	0.001	0.009	0.010
Test: Average Error Function	.	0.009	.
Test: Divisor for TASE	148.000	148.000	148.000
Test: Error Function	.	1.358	.
Test: Maximum Absolute Error	0.209	0.324	0.487
Test: Mean Square Error	.	0.009	.
Test: Sum of Frequencies	148.000	148.000	148.000
Test: Root Average Squared Error	0.030	0.096	0.100
Test: Root Mean Square Error	.	0.096	.
Test: Sum of Square Errors	0.134	1.358	1.474
Test: Sum of Case Weights Times Freq	.	148.000	148.000

Figure 19. Output Linear Regression Model Rocky Mountains Region.

Model Fit Statistics

R-Square	0.8757	Adj R-Sq	0.8740
AIC	-3131.4266	BIC	-3129.0686
SBC	-3080.9635	C(p)	10.3573

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	6.0978	0.1975	30.88	<.0001	5.7108	6.4848
LOG_Emp23Total	1	0.1310	0.0158	8.32	<.0001	0.1002	0.1619
LOG_Emp31_33Total	1	-0.0231	0.00728	-3.17	0.0016	-0.0374	-0.00883
LOG_Emp51total	1	0.0247	0.0111	2.23	0.0259	0.00302	0.0464
LOG_Emp52Total	1	0.1137	0.0162	7.03	<.0001	0.0820	0.1454
LOG_Emp54Total	1	0.0631	0.0120	5.27	<.0001	0.0396	0.0865
LOG_Emp61Total	1	0.0261	0.0111	2.36	0.0185	0.00444	0.0478
LOG_Emp62Total	1	-0.1876	0.0210	-8.94	<.0001	-0.2287	-0.1465
LOG_Emp81total	1	-0.0607	0.0225	-2.70	0.0072	-0.1048	-0.0166
LOG_Emp92	1	-0.1366	0.0120	-11.42	<.0001	-0.1600	-0.1132
LOG_medianRent	1	0.9390	0.0257	36.50	<.0001	0.8886	0.9895

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	-3131.43	.	.
ASE	Average Squared Error	0.01	0.012	0.015
AVERR	Average Error Function	0.01	0.012	0.015
DFE	Degrees of Freedom for Error	715.00	.	.
DFM	Model Degrees of Freedom	11.00	.	.
DFT	Total Degrees of Freedom	726.00	.	.
DIV	Divisor for ASE	726.00	330.000	264.000
ERR	Error Function	9.43	4.007	3.951
FPE	Final Prediction Error	0.01	.	.
MAX	Maximum Absolute Error	0.58	0.574	0.538
MSE	Mean Square Error	0.01	0.012	0.015
NOBS	Sum of Frequencies	726.00	330.000	264.000
_NW_	Number of Estimate Weights	11.00	.	.
RASE	Root Average Sum of Squares	0.11	0.110	0.122
RFPE	Root Final Prediction Error	0.12	.	.
RMSE	Root Mean Squared Error	0.11	0.110	0.122
SBC	Schwarz's Bayesian Criterion	-3080.96	.	.
SSE	Sum of Squared Errors	9.43	4.007	3.951
SUMW	Sum of Case Weights Times Freq	726.00	330.000	264.000

Figure 20. Output Decision Tree Rocky Mountains Region

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	726.000	330.000	264.000
MAX	Maximum Absolute Error	0.424	0.542	0.430
SSE	Sum of Squared Errors	6.831	3.809	3.151
ASE	Average Squared Error	0.009	0.012	0.012
RASE	Root Average Squared Error	0.097	0.107	0.109
DIV	Divisor for ASE	726.000	330.000	264.000
DFT	Total Degrees of Freedom	726.000	.	.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_medianRent	Transformed medianRent	9	1.0000	1.0000	1.0000
LOG_Emp44_45Total	Transformed Emp44_45Total	2	0.2498	0.1582	0.6331
TI_stateName4	stateName:UT	1	0.1367	0.1418	1.0378
LOG_Emp31_33Total	Transformed Emp31_33Total	2	0.1051	0.1294	1.2310
TI_stateName1	stateName:CO	2	0.0744	0.0408	0.5485
LOG_Emp52Total	Transformed Emp52Total	1	0.0404	0.0710	1.7559

Figure 21. Output of Random Forest Model

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.000	0.003	0.002
DIV	Divisor for ASE	726.000	330.000	264.000
MAX	Maximum Absolute Error	0.126	0.306	0.173
NOBS	Sum of Frequencies	726.000	330.000	264.000
RASE	Root Average Squared Error	0.022	0.053	0.046
SSE	Sum of Squared Errors	0.342	0.915	0.553

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error
mortgageRatePercentage	1738	0.000448	0.005358	-0.00052	0.000635	-0.00041	0.001219
LOG_medianRent	1731	0.028195	0.052885	0.02887	0.049865	0.02272	0.039787T
LOG_mortgageRate	1390	0.000300	0.003915	-0.00017	0.000600	-0.00013	0.000982T
LOG_Emp92	1191	0.002809	0.010359	0.00233	0.006327	0.00177	0.004980T
LOG_Emp81total	1098	0.006996	0.018281	0.00606	0.014428	0.00506	0.013015T
LOG_Emp72Total	979	0.012991	0.022495	0.01209	0.019091	0.00675	0.014045T
LOG_Emp62Total	919	0.009546	0.022576	0.00861	0.019931	0.00794	0.018059T
LOG_Emp61Total	723	0.003610	0.009271	0.00305	0.006929	0.00238	0.005264T
LOG_Emp54Total	632	0.006169	0.014230	0.00561	0.011649	0.00322	0.008556T
LOG_Emp52Total	576	0.003684	0.010132	0.00437	0.009777	0.00331	0.008738T
LOG_Emp51total	533	0.004770	0.011383	0.00391	0.008459	0.00230	0.005669T
LOG_Emp48_49Total	522	0.002747	0.009299	0.00212	0.006459	0.00162	0.005564T
LOG_Emp31_33Total	444	0.003987	0.009885	0.00323	0.007770	0.00338	0.007906T
LOG_Emp44_45Total	399	0.003146	0.007845	0.00304	0.006467	0.00239	0.005236T
LOG_Emp23Total	345	0.003746	0.010016	0.00261	0.007136	0.00183	0.006124T
TI_stateName2	28	0.008516	0.015171	0.00865	0.015039	0.00832	0.013493s
TI_stateName1	13	0.000934	0.001329	0.00077	0.001403	0.00051	0.001025s
TI_stateName4	10	0.000200	0.000099626	0.00007	0.000019794	0.00015	0.000086840s
TI_stateName3	8	0.000126	0.000368	0.00016	0.000414	0.00021	0.000595s
TI_stateName5	2	0.000056	0.000231	0.00003	0.000111	-0.00000	0.000063697s

Figure 22. Output of Comparison node showing statistics for validation and test sets

Data Role=Valid

Statistics	HPDMForest	Tree	Reg
Valid: Average Squared Error	0.003	0.012	0.012
Valid: Average Error Function	.	.	0.012
Valid: Divisor for VASE	330.000	330.000	330.000
Valid: Error Function	.	.	4.007
Valid: Maximum Absolute Error	0.306	0.542	0.574
Valid: Mean Square Error	.	.	0.012
Valid: Sum of Frequencies	330.000	330.000	330.000
Valid: Root Average Squared Error	0.053	0.107	0.110
Valid: Root Mean Square Error	.	.	0.110
Valid: Sum of Square Errors	0.915	3.809	4.007
Valid: Sum of Case Weights Times Freq	.	.	330.000

Data Role=Test

Statistics	HPDMForest	Tree	Reg
Test: Lower 95% Conf. Limit for TASE	.	.	0.001
Test: Upper 95% Conf. Limit for TASE	.	.	0.044
Test: Average Squared Error	0.002	0.012	0.015
Test: Average Error Function	.	.	0.015
Test: Divisor for TASE	264.000	264.000	264.000
Test: Error Function	.	.	3.951
Test: Maximum Absolute Error	0.173	0.430	0.538
Test: Mean Square Error	.	.	0.015
Test: Sum of Frequencies	264.000	264.000	264.000
Test: Root Average Squared Error	0.046	0.109	0.122
Test: Root Mean Square Error	.	.	0.122
Test: Sum of Square Errors	0.553	3.151	3.951
Test: Sum of Case Weights Times Freq	.	264.000	264.000

Figure 23. Output of linear regression model for the Southeast region.

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	-21749.79	.	.
ASE	Average Squared Error	0.02	0.02	0.02
AVERR	Average Error Function	0.02	0.02	0.02
DFE	Degrees of Freedom for Error	5687.00	.	.
DFM	Model Degrees of Freedom	11.00	.	.
DFT	Total Degrees of Freedom	5698.00	.	.
DIV	Divisor for ASE	5698.00	2590.00	2072.00
ERR	Error Function	124.83	54.62	43.44
FPE	Final Prediction Error	0.02	.	.
MAX	Maximum Absolute Error	0.92	0.75	0.81
MSE	Mean Square Error	0.02	0.02	0.02
NOBS	Sum of Frequencies	5698.00	2590.00	2072.00
_NW_	Number of Estimate Weights	11.00	.	.
RASE	Root Average Sum of Squares	0.15	0.15	0.14
RFPE	Root Final Prediction Error	0.15	.	.
RMSE	Root Mean Squared Error	0.15	0.15	0.14
SBC	Schwarz's Bayesian Criterion	-21676.66	.	.
SSE	Sum of Squared Errors	124.83	54.62	43.44
SUMW	Sum of Case Weights Times Freq	5698.00	2590.00	2072.00

Model Fit Statistics

R-Square	0.7596	Adj R-Sq	0.7591
AIC	-21749.7857	BIC	-21747.7431
SBC	-21676.6591	C(p)	11.0000

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	3.8693	0.1194	32.40	<.0001	3.6353	4.1034
LOG_Emp23Total	1	0.0296	0.00712	4.15	<.0001	0.0156	0.0435
LOG_Emp44_45Total	1	-0.1777	0.0143	-12.40	<.0001	-0.2057	-0.1496
LOG_Emp48_49Total	1	-0.0485	0.00348	-13.95	<.0001	-0.0553	-0.0417
LOG_Emp54Total	1	0.0321	0.00575	5.59	<.0001	0.0209	0.0434
LOG_Emp61Total	1	0.1016	0.00616	16.50	<.0001	0.0895	0.1137
LOG_Emp62Total	1	-0.1091	0.00788	-13.85	<.0001	-0.1245	-0.0937
LOG_Emp72Total	1	0.1691	0.00920	18.39	<.0001	0.1511	0.1871
LOG_Emp81total	1	0.0805	0.00880	9.15	<.0001	0.0632	0.0977
LOG_Emp92	1	-0.0553	0.00474	-11.67	<.0001	-0.0645	-0.0460
LOG_medianRent	1	1.1187	0.0164	68.09	<.0001	1.0865	1.1509

Figure 24. Output of Decision Tree Model to predict median housing prices in the Southeast region

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_medianRent	Transformed medianRent	6	1.0000	1.0000	1.0000
LOG_Emp61Total	Transformed Emp61Total	4	0.3329	0.3077	0.9243
LOG_Emp72Total	Transformed Emp72Total	3	0.2134	0.1925	0.9022
LOG_Emp48_49Total	Transformed Emp48_49Total	2	0.1830	0.1830	1.0002
LOG_Emp44_45Total	Transformed Emp44_45Total	1	0.1208	0.1012	0.8376
TI_stateName8	stateName:NC	2	0.1005	0.1044	1.0386
TI_stateName11	stateName:VA	2	0.0981	0.1082	1.1028
TI_stateName12	stateName:WV	1	0.0903	0.0823	0.9115
LOG_Emp23Total	Transformed Emp23Total	1	0.0765	0.0701	0.9163
LOG_Emp54Total	Transformed Emp54Total	2	0.0763	0.0887	1.1633
LOG_Emp81total	Transformed Emp81total	1	0.0742	0.0856	1.1533
TI_stateName3	stateName:FL	1	0.0601	0.0438	0.7285
LOG_Emp62Total	Transformed Emp62Total	1	0.0400	0.0438	1.0938
LOG_Emp92	Transformed Emp92	1	0.0328	0.0154	0.4695

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	5698.00	2590.00	2072.00
MAX	Maximum Absolute Error	1.01	0.96	0.95
SSE	Sum of Squared Errors	95.98	46.19	35.41
ASE	Average Squared Error	0.02	0.02	0.02
RASE	Root Average Squared Error	0.13	0.13	0.13
DIV	Divisor for ASE	5698.00	2590.00	2072.00
DFT	Total Degrees of Freedom	5698.00	.	.

Figure 25. Output of Random Forest model for the Southeast region

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error
LOG_medianRent	13730	0.033153	0.068795	0.032399	0.065018	0.031458	0.065043
LOG_Emp92	11739	0.003294	0.015059	0.002732	0.012503	0.002620	0.011845
LOG_Emp81total	10485	0.005166	0.016810	0.004752	0.014272	0.004093	0.013744
LOG_Emp72Total	9376	0.009563	0.022724	0.008734	0.019643	0.008561	0.020221
LOG_Emp62Total	7664	0.006263	0.018186	0.005431	0.015482	0.005755	0.015909
LOG_Emp61Total	5153	0.007582	0.019343	0.006649	0.016571	0.005890	0.015787
LOG_Emp54Total	4914	0.007074	0.018164	0.006576	0.016304	0.005914	0.015825
LOG_Emp48_49Total	4358	0.003619	0.012601	0.003117	0.010894	0.002652	0.010222
LOG_Emp23Total	3566	0.005106	0.014435	0.004471	0.012369	0.004188	0.012160
LOG_Emp44_45Total	3469	0.005168	0.014453	0.004717	0.012881	0.004186	0.012266
TI_stateName11	85	0.002679	0.006573	0.002563	0.006428	0.002438	0.006103
TI_stateName8	85	0.000577	0.001394	0.000677	0.001553	0.000650	0.001375
TI_stateName12	56	0.001063	0.003094	0.001069	0.003043	0.000757	0.002121
TI_stateName3	45	0.000257	0.000500	0.000176	0.000409	0.000273	0.000552
TI_stateName4	43	0.000085	0.000320	0.000082	0.000313	0.000066	0.000257
TI_stateName2	23	0.000106	0.000332	0.000100	0.000316	0.000101	0.000251

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.00	0.00	0.00
DIV	Divisor for ASE	5698.00	2590.00	2072.00
MAX	Maximum Absolute Error	0.22	0.50	0.50
NOBS	Sum of Frequencies	5698.00	2590.00	2072.00
RASE	Root Average Squared Error	0.02	0.03	0.03
SSE	Sum of Squared Errors	1.77	3.08	2.53

Figure 26. Output of Linear Regression Model for the Southwest Region

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	2.4309	0.1734	14.02	<.0001	2.0910	2.7709
LOG_Emp31_33Total	1	-0.0649	0.00493	-13.15	<.0001	-0.0745	-0.0552
LOG_Emp44_45Total	1	0.1278	0.0239	5.35	<.0001	0.0810	0.1746
LOG_Emp48_49Total	1	-0.0907	0.00541	-16.78	<.0001	-0.1013	-0.0801
LOG_Emp51total	1	0.0327	0.00747	4.38	<.0001	0.0181	0.0474
LOG_Emp54Total	1	0.0831	0.00670	12.41	<.0001	0.0700	0.0962
LOG_Emp61Total	1	-0.0664	0.00892	-7.44	<.0001	-0.0839	-0.0489
LOG_Emp62Total	1	-0.0733	0.0114	-6.45	<.0001	-0.0955	-0.0510
LOG_Emp72Total	1	0.1275	0.0206	6.18	<.0001	0.0871	0.1679
LOG_Emp81total	1	-0.0587	0.0112	-5.22	<.0001	-0.0807	-0.0367
LOG_medianRent	1	1.2780	0.0249	51.26	<.0001	1.2291	1.3268
TI_stateName2 0	1	0.0390	0.00773	5.05	<.0001	0.0239	0.0542
TI_stateName3 0	1	0.1149	0.00651	17.64	<.0001	0.1021	0.1277
TI_stateName4 0	1	0.1496	0.00633	23.62	<.0001	0.1372	0.1620

Model Fit Statistics

R-Square	0.8948	Adj R-Sq	0.8940
AIC	-7137.9609	BIC	-7135.7165
SBC	-7062.5243	C(p)	14.0000

Figure 27. Output of Decision Tree model for the Southwest Region

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_medianRent	Transformed medianRent	8	1.0000	1.0000	1.0000
LOG_Emp44_45Total	Transformed Emp44_45Total	1	0.4032	0.3561	0.8834
LOG_Emp54Total	Transformed Emp54Total	3	0.2396	0.2109	0.8801
TI_stateName4	stateName:TX	2	0.2270	0.2389	1.0523
TI_stateName3	stateName:OK	1	0.1400	0.1298	0.9273
LOG_Emp72Total	Transformed Emp72Total	2	0.1346	0.2171	1.6129
TI_stateName2	stateName:NM	1	0.1239	0.1042	0.8413
LOG_Emp31_33Total	Transformed Emp31_33Total	2	0.1204	0.1275	1.0588
LOG_Emp62Total	Transformed Emp62Total	1	0.1036	0.0841	0.8119
LOG_Emp48_49Total	Transformed Emp48_49Total	1	0.0775	0.0606	0.7829

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	1617.00	735.000	588.000
MAX	Maximum Absolute Error	0.52	0.610	0.558
SSE	Sum of Squared Errors	23.77	13.288	9.809
ASE	Average Squared Error	0.01	0.018	0.017
RASE	Root Average Squared Error	0.12	0.134	0.129
DIV	Divisor for ASE	1617.00	735.000	588.000
DFT	Total Degrees of Freedom	1617.00	.	.

Figure 28. Output of comparison node for predictive models of the Southwest region.

Data Role=Valid

Statistics	HPDMForest	Reg	Tree
Valid: Average Squared Error	0.002	0.014	0.018
Valid: Average Error Function	.	0.014	.
Valid: Divisor for VASE	735.000	735.000	735.000
Valid: Error Function	.	10.440	.
Valid: Maximum Absolute Error	0.358	0.653	0.610
Valid: Mean Square Error	.	0.014	.
Valid: Sum of Frequencies	735.000	735.000	735.000
Valid: Root Average Squared Error	0.047	0.119	0.134
Valid: Root Mean Square Error	.	0.119	.
Valid: Sum of Square Errors	1.658	10.440	13.288
Valid: Sum of Case Weights Times Freq	.	735.000	.

Data Role=Test

Statistics	HPDMForest	Reg	Tree
Test: Lower 95% Conf. Limit for TASE	.	0.005	.
Test: Upper 95% Conf. Limit for TASE	.	0.027	.
Test: Average Squared Error	0.002	0.014	0.017
Test: Average Error Function	.	0.014	.
Test: Divisor for TASE	588.000	588.000	588.000
Test: Error Function	.	8.122	.
Test: Maximum Absolute Error	0.285	0.534	0.558
Test: Mean Square Error	.	0.014	.
Test: Sum of Frequencies	588.000	588.000	588.000
Test: Root Average Squared Error	0.045	0.118	0.129
Test: Root Mean Square Error	.	0.118	.
Test: Sum of Square Errors	1.199	8.122	9.809
Test: Sum of Case Weights Times Freq	.	588.000	588.000

Figure 29. Output Multiple Linear Regression Model Plains Region

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	4.4897	0.1371	32.74	<.0001	4.2209	4.7585
LOG_Emp31_33Total		1	-0.0388	0.00439	-8.84	<.0001	-0.0474	-0.0302
LOG_Emp44_45Total		1	-0.1228	0.0366	-3.36	0.0008	-0.1944	-0.0511
LOG_Emp48_49Total		1	0.0143	0.00563	2.54	0.0110	0.00329	0.0253
LOG_Emp52Total		1	-0.0167	0.00594	-2.82	0.0049	-0.0284	-0.00508
LOG_Emp61Total		1	0.0568	0.00795	7.14	<.0001	0.0412	0.0724
LOG_Emp72Total		1	0.0844	0.0375	2.25	0.0243	0.0110	0.1579
LOG_Emp81total		1	0.0431	0.00958	4.50	<.0001	0.0244	0.0619
LOG_medianRent		1	1.0382	0.0200	51.88	<.0001	0.9990	1.0774
TI_stateName6	0	1	0.0456	0.00452	10.08	<.0001	0.0367	0.0544
TI_stateName7	0	1	-0.1097	0.00987	-11.12	<.0001	-0.1291	-0.0904

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	-7829.91	.	.
ASE	Average Squared Error	0.01	0.014	0.013
AVERR	Average Error Function	0.01	0.014	0.013
DFE	Degrees of Freedom for Error	1793.00	.	.
DFM	Model Degrees of Freedom	11.00	.	.
DFT	Total Degrees of Freedom	1804.00	.	.
DIV	Divisor for ASE	1804.00	820.000	656.000
ERR	Error Function	23.23	11.318	8.805
FPE	Final Prediction Error	0.01	.	.
MAX	Maximum Absolute Error	0.42	1.288	0.716
MSE	Mean Square Error	0.01	0.014	0.013
NOBS	Sum of Frequencies	1804.00	820.000	656.000
_NW_	Number of Estimate Weights	11.00	.	.
RASE	Root Average Sum of Squares	0.11	0.117	0.116
RFPE	Root Final Prediction Error	0.11	.	.
RMSE	Root Mean Squared Error	0.11	0.117	0.116
SBC	Schwarz's Bayesian Criterion	-7769.43	.	.
SSE	Sum of Squared Errors	23.23	11.318	8.805
SUMW	Sum of Case Weights Times Freq	1804.00	820.000	656.000

Figure 30. Output of Decision Tree Model for Plains Region

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_medianRent	Transformed medianRent	13	1.0000	1.0000	1.0000
LOG_Emp72Total	Transformed Emp72Total	1	0.2907	0.2386	0.8207
LOG_Emp54Total	Transformed Emp54Total	2	0.1887	0.1546	0.8192
TI_stateName2	stateName:KS	1	0.1665	0.1014	0.6090
LOG_Emp62Total	Transformed Emp62Total	1	0.1576	0.0911	0.5778
LOG_Emp31_33Total	Transformed Emp31_33Total	2	0.1450	0.0271	0.1870
LOG_Emp61Total	Transformed Emp61Total	1	0.1140	0.1331	1.1671
TI_stateName6	stateName:NE	2	0.0786	0.0738	0.9387
LOG_Emp48_49Total	Transformed Emp48_49Total	1	0.0261	0.0157	0.6005

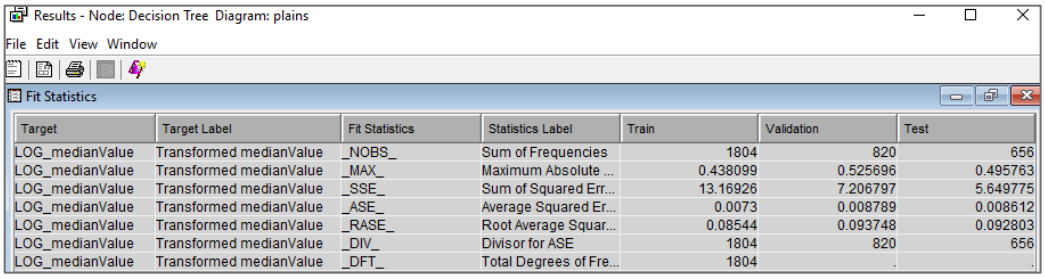


Figure 31. Output of Random Forest for Plains Region

Fit Statistics

Target=LOG_medianValue Target Label=Transformed medianValue

Fit Statistics	Statistics Label	Train	Validation	Test
ASE	Average Squared Error	0.00	0.001	0.002
DIV	Divisor for ASE	1804.00	820.000	656.000
MAX	Maximum Absolute Error	0.19	0.251	0.288
NOBS	Sum of Frequencies	1804.00	820.000	656.000
RASE	Root Average Squared Error	0.02	0.037	0.040
SSE	Sum of Squared Errors	0.52	1.106	1.060

Loss Reduction Variable Importance

Variable	Number of Rules	MSE	OOB MSE	Valid MSE	Absolute Error	OOB Absolute Error	Valid Absolute Error
LOG_medianRent	3963	0.014622	0.014878	0.014429	0.042302	0.039422	0.039199
LOG_Emp92	3180	0.007708	0.006742	0.007330	0.019610	0.016395	0.017460
LOG_Emp72Total	2465	0.005369	0.004907	0.004669	0.015405	0.013426	0.012925
LOG_Emp81total	2928	0.005126	0.004660	0.004670	0.015143	0.012349	0.012305
LOG_Emp62Total	2364	0.004930	0.004596	0.004937	0.015963	0.013965	0.014556
LOG_Emp61Total	1758	0.004531	0.003986	0.003585	0.010666	0.008746	0.007653
LOG_Emp54Total	1758	0.003508	0.002868	0.002887	0.011279	0.009048	0.009290
LOG_Emp44_45Total	1177	0.002785	0.002519	0.002440	0.008451	0.006839	0.006812
LOG_Emp51total	1434	0.002413	0.002228	0.002385	0.008287	0.006870	0.007231
LOG_Emp23Total	907	0.002260	0.001976	0.001676	0.006721	0.005415	0.004435
TI_stateName2	43	0.001881	0.001892	0.001508	0.005862	0.005987	0.004645
LOG_Emp31_33Total	1142	0.002072	0.001591	0.001073	0.008724	0.007057	0.006046
LOG_Emp52Total	1612	0.001612	0.001386	0.001286	0.007332	0.005962	0.005628
LOG_Emp48_49Total	1261	0.001655	0.001243	0.001228	0.006594	0.004965	0.004968
TI_stateName3	24	0.001198	0.000809	0.001166	0.001904	0.001274	0.002101
TI_stateName7	15	0.000384	0.000363	0.000267	0.001269	0.001182	0.000736
TI_stateName6	21	0.000253	0.000235	0.000274	0.000989	0.000901	0.001096
TI_stateName4	24	0.000165	0.000169	0.000165	0.000619	0.000643	0.000562
TI_stateName1	16	0.000111	0.000110	0.000099	0.000326	0.000333	0.000325
TI_stateName5	1	0.000038	0.000033	0.000031	0.000080140	0.000079492	0.000067217