Laboratory Exercise – 2

IST 718 Big Data Analytics

Dan Caley

Table of Contents

[Disclaimer 2](#_Toc88401113)

[Obtain 3](#_Toc88401114)

[Zillow Data 3](#_Toc88401115)

[Census Data 3](#_Toc88401116)

[Scrub 6](#_Toc88401117)

[Zillow Data 6](#_Toc88401118)

[Census Data 7](#_Toc88401119)

[Merging Zillow and Census Data 8](#_Toc88401120)

[Explore 10](#_Toc88401121)

[Arkansas Metro Time Series Plot 10](#_Toc88401122)

[Arkansas Metro Percentage Return (1997 – 2019) 11](#_Toc88401123)

[Arkansas Metro Results (2010 – 2019) 12](#_Toc88401124)

[Arkansas Metro Population and Household Median Income (2010 – 2019) 12](#_Toc88401125)

[Bonus Geographic Visualization –Median Housing Price 13](#_Toc88401126)

[Bonus Geographic Visualization – Population 14](#_Toc88401127)

[Bonus Geographic Visualization - Household Median Income 15](#_Toc88401128)

[Modeling 16](#_Toc88401129)

[Historic Risk and Return - USA 16](#_Toc88401130)

[Historic Risk and Return - Arkansas 17](#_Toc88401131)

[Forecasting Arkansas Return - Scrubing 18](#_Toc88401132)

[Forecasting Arkansas Return - Results 21](#_Toc88401133)

# Disclaimer

The bonus section was completed which included not just the Median house price mapped across the United States but also included Census population and household median income both at the County and State level.

To see all the analysis please look the Jupyter Notebook.

# Obtain

There were 4 datasets used in performing this analysis:

1. Zillow Static Data set found at <https://files.zillowstatic.com/research/public/Zip/Zip_Zhvi_SingleFamilyResidence.csv>
2. Zip Code Tabulation Area (ZCTA) Household Median Income and Population level data from the Census Bureau.
3. County Household Median Income and Population level data from the Census Bureau.
4. State Household Median Income and Population level data from the Census Bureau.
5. State code data mapping to plot on a geographic map.

## Zillow Data

To obtain the Zillow data set the Pandas read csv function was used by inserting the above url.

Table

Description automatically generated

## Census Data

To obtain the census data, the following code lines were used:

1. Ping the census bureau api for Household Median Income and Population by year
2. Append the data to a dataframe
3. Loop through 2011 – 2020, this is everything that the census bureau has.
4. Rename the headers to Median Income and Population
5. Create a Pandas Dataframe

A picture containing text

Description automatically generated

Table

Description automatically generated

This data frame shows the House Hold Median Income by year. We can clearly see that there is some data to be cleaned with every min is -66666, representing 0. Due to this the average gets thrown off. The median Household income can be seen at the 50%. By 2019 this increased by $8k. If accurate this track with inflation. Meaning inflation on average is 2%, over the course of 10 years that comes out to roughly a little above 20%

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

This process was repeated for both the County and State level.

The mapping for State Name and State abbreviation for to map the geographic visualizations were manually inserted.

Text

Description automatically generated

# Scrub

## Zillow Data

For the Zillow Data the following was needed to be scrubbed:

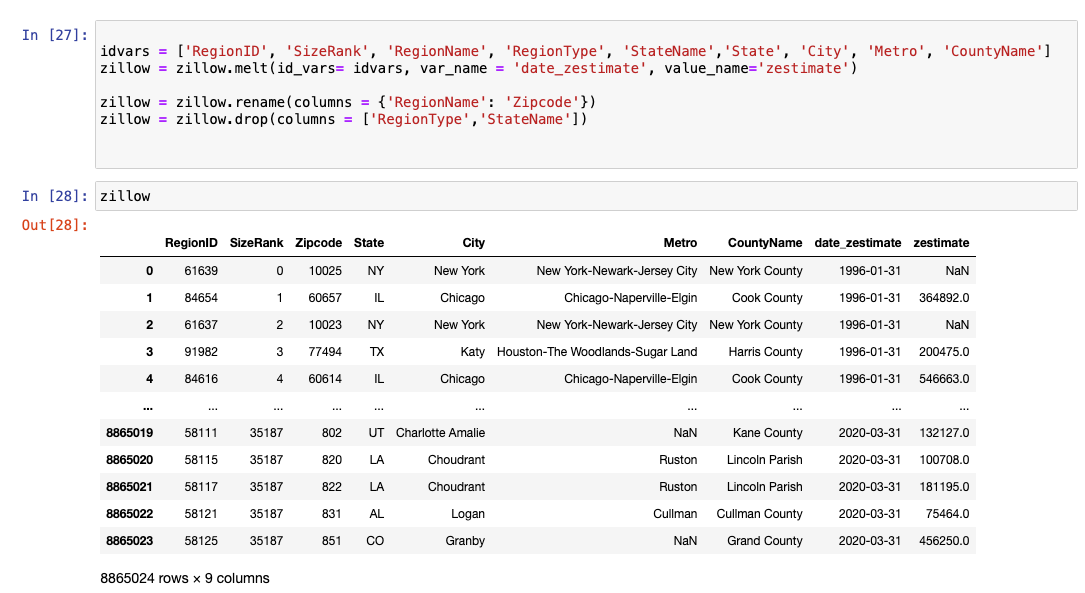
1. Zip code needed to have leading 0’s. Meaning a Zip Code is 5 digits with 0 at the front in some instances
2. The dataframe was melted to have the dates as rows instead of columns
3. The date name and the values were then called date\_zestimate and zestimate.
4. NaN were dropped from the dataset completely. Due to having so much data losing about 25% wasn’t a huge hit like normal datasets.

Before

Text, table

Description automatically generated with medium confidence

After



Dropped Data Results:

Graphical user interface, text, application

Description automatically generated

## Census Data

For the census data the following was needed to scrubbed:

1. The ZCTA data was pretty clean after obtaining the data.
2. The county data needed to be reformatted to create FIPS codes which is just the state code concatenated with the county code.

Table

Description automatically generated

## Merging Zillow and Census Data

Merged the Zillow data with the census data. This is joining Census onto Zillow. The Zip Code in Zillow is a copy right of the postal service and the ZCTA is a copyright of the census bureau. These two Zip Codes are different, so this is not a perfect match. There is a mapping file but for this analysis combining on the Zillow Zip Code will be sufficient due to only 2% of the Zip Codes were unable to be mapped. This is opportunity to improve the precision of the analysis.

Graphical user interface, text, application, email

Description automatically generated

An example of a Zip Code 85203

Graphical user interface, text, email

Description automatically generated

# Explore

## Arkansas Metro Time Series Plot

Chart, line chart

Description automatically generated

The graph above shows the 4 Metro areas Average Zestimate. In terms of growth Hot Springs appears to be the best with Little Rock being second best. This is hard to tell looking pearly out this graph. Also there is a lot of risk or volatility in all the Metro’s besides Searcy. Using some finance techniques lets look at the overall Return, Risk, and Return over Risk also known as Sharpe Ratio.

## Arkansas Metro Percentage Return (1997 – 2019)

Chart

Description automatically generated

This graph is a good visualization in volatility. Hot springs from 1996 until current has more volatility than the group with Searcy having the second most. Confirming our suspicions. Although Searcy appears to have done better after 2008.

Overall, from 1996 until the end of 2019 Overall Returns were as followed:

Text

Description automatically generated

The Higher the Sharpe Ratio is better. Meaning that for every Return an investor received they took on smaller risk compared to other investments. Hot Springs return overall was 71% but an investor had to take on seen in the below chart compared to Little Rock where very risk was needed.

Text

Description automatically generated

## Arkansas Metro Results (2010 – 2019)

Text

Description automatically generated

Before Hot Spring overall was the better choice from a historical perspective with the Return and Risk balanced very well. Looking to just this past decade Searcy has better return and a much higher Sharpe Ratio. The Riskiest area is actually Hot Springs now with Little over much smaller risk. Looking at more recent data will be important in the modeling stage of the analysis.

Looking at 2010 until 2019 the results are different. For the remaining of the analysis.

## Arkansas Metro Population and Household Median Income (2010 – 2019)

Chart, bar chart

Description automatically generated

The population represented hear is just the zip codes that Zillow provided. This might not be a full representation of the population growth. There appears to be a disconnect with Searcy and Zillow data. Meaning that housing prices are going up, yet population increased the first 4 years and then is on the decline. Either this population is now becoming homeowners or houses are disappearing from the market. Census data is not always accurate and with the 2020 census data coming next year Searcy might see an adjusted population growth.

Chart, bar chart

Description automatically generated

Confirming the same data issue with Searcy shows that household median income isn’t increases although housing prices and population are. A big disconnect. The other graphs track with steady growth.

## Bonus Geographic Visualization –Median Housing Price

A picture containing text

Description automatically generated

## Bonus Geographic Visualization – Population

Map

Description automatically generated

A picture containing text

Description automatically generated

## Bonus Geographic Visualization - Household Median Income

Map

Description automatically generated

# Modeling

## Historic Risk and Return - USA

When looking to perform the modeling historic returns should be considered by Zip code. The chart below shows all Zip Codes color coded by State. The objective in the analysis is too have the most return with lower risk. Meaning If a zip code achieves 12% return and a Standard Deviation (Risk) of 2% and another zip code achieves 12% return with a Risk percentage of 1% then taking the ladder zip code is the most optimal solution. This can also be described as a Sharpe Ratio where return is divided by risk. The higher the Sharpe Ratio means a more balance Return over Risk solution.

Chart, scatter chart

Description automatically generated

This graph shows every state from 2010 – 2019. The size is represented the Sharpe Ratio. As can be seen that California has a higher return with lower risk than many of the zip codes across the nation. Let’s focus though on Arkansas.

## Historic Risk and Return - Arkansas

Chart, scatter chart

Description automatically generated

Historically Speaking zip code 72447 has the highest return at 5.5% with some of the highest risk. Depending on the risk tolerance of an investor will depend if investing in this zip makes sense.

The next 3 highest zip codes have similar returns at 5% but have different risk levels:

* 71740
* 72675
* 72645

The better of the 3 and of even the highest would be 72645 achieving 4.7% return with much lower risk than the highest.

Chart, scatter chart

Description automatically generated

KS = Blue

AR = Red

OK = Yellow

TN = Orange

When looking at Neighboring states, excluding Texas There is better returns over risk then Arkansas. Meaning any different color bubble than red and is above red means that there is more return for the same amount risk.

## Forecasting Arkansas Return - Scrubing

When forecasting for Arkansas in 2020 and finding the 3 highest Zip codes Facebook Prophet was used in modeling these results. Before running the prophet additional scrubbing had to be done:

1. Filtered for just Zestimates from 2010 – 2019.
2. Only looked at Zipcodes with historic returns greater than 3% over the past 4 years.
3. Why looking over 3% is because inflation on average is 2% and investors minimum need to be compensated for taking on housing risk by a factor of 1%.
4. Then looked at just Arkansas which comes out to be 420 data points.
5. Changed names so that the Prophet would ingest the data.
6. Date\_zestimate changed to ds and zestimate to y
7. Looped the data through prophet by zipcode and appened to an empty dataframe including the zip code as a column.
8. Found the annualized returns predicted, annual risk predicted, and the sharpe ratio predicted.
9. Please see the below code for these steps.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

## Forecasting Arkansas Return - Results

The below graphs shows that the 3 highest return for 2020:

1. 72630 – Diamond City with a return of 17% and risk of 0.9%
2. 72431 – Grubbs with a return of 16% and risk of 0.8%
3. 71935 – Caddo Gap with a return of 13% and risk of 0.5%

Of the 3 that are the highest in graph 3 – 5 the predictive power illustrates the error bands in this forecast. Zip code 71935 appears to have much better predictive power of the 3 where the other’s do not.

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

Zip code: 72630

Chart, line chart, scatter chart

Description automatically generated

Zip code: 72431

Chart, line chart

Description automatically generated

Zip code: 71935

Chart, line chart

Description automatically generated

Table

Description automatically generated