Text

Description automatically generated

**Course:** 2023-1004:16645 IST-718 Big Data Analytics

**Name:** Jacob N. Conard

**NetID:** jconard

**Assignment:** Laboratory Exercise - 2

**Date due:** 12 November 2023

**Date submitted:** 08 November 2023

**Number of Pages:** 20

**Introduction**

In the dynamic landscape of real estate investment, data science has emerged as an indispensable tool for making informed decisions and maximizing returns. As the demand for reliable investment opportunities continues to grow, real estate investment trusts (REITs) play a pivotal role in channeling capital into potential markets. This paper embarks on a compelling journey into the realm of real estate data science, with a focused exploration into the Syracuse Real Estate Investment Trust's (SREIT) quest to predict the most lucrative investment opportunities. By leveraging the power of data acquisition, preprocessing, exploratory analysis, predictive modeling, and insightful interpretation, this study aims to unlock the latent patterns and trends within the United States real estate market. It seeks to empower SREIT with the knowledge and tools needed to strategically select the three most promising zip codes for investment. In this endeavor, the paper combines the rigor of data science methodologies with the practicality of real-world investment decisions, ultimately bridging the gap between data-driven insights and the pursuit of sustainable real estate investment success.

**Data and Data Cleaning**

The data utilized for this lab is from Zillow and includes the median housing price for a specific zip code around the United States of America. An shape of the dataset can be seen below:

A screenshot of a computer

Description automatically generated

The first thing that we need to do to clean the data is to ensure that all of the date columns are formatted in a YYYY-MM format and remove the day of the month associated with each column:

A screen shot of a computer program

Description automatically generated

The next thing that we need to do to clean the data is to only capture the dates that are covered in the lab. For this specific lab we are only concerned with data from 1996 – 2017:

A computer screen shot of text

Description automatically generated

Below is an example of the final filtered dataset with the data clean and the string data in the rows changed to lowercase:

A screenshot of a computer

Description automatically generated

**Arkansas and Arizona Exploratory Data Analysis**

Once the data has been cleaned the next step is to conduct exploratory data analysis on the data. We will do this by answering some key questions related to the states of Arkansas and Arizona. Those questions can be found below:

1. Develop time series plots for the following Arkansas metro areas:
   1. Hot Springs, Little Rock, Fayetteville, Searcy
   2. Present all values from 1997 to present
   3. Average at the metro area level
2. Develop time series plots for the following Arizona metro areas:
   1. Phoenix, Tucson, Payson, Yuma
   2. Present all values from 1997 to present
   3. Average at the metro area level

In order to answer the aforementioned questions we need to do some data transformations and slice the dataframe in order to answer the questions posed above.

The code below looks for the specific metro areas mentioned above that we need develop a time series plot for in Arkansas:

A screen shot of a computer code

Description automatically generated

Below is a function that will take those new data sets and create a time series plot for the average home prices in the specific metro areas mentioned above in Arkansas:

A computer screen shot of text

Description automatically generated

Now that we have the data in subsets and a function to graph the average home prices for those areas we can start generating time series plots for Searcy, Fayetteville, Hot Springs, and Little Rock.

**Hot Springs, AR**

A graph showing a line

Description automatically generated

**Little Rock, AR**

A graph showing a line

Description automatically generated

**Fayetteville, AR**

A graph showing a line

Description automatically generated

**Searcy, AR**

A graph showing a line

Description automatically generated

**All Metro Areas in Arkansas**

In order to merge all of these plots together, we need to merge the dataframes into a larger dataframe that includes all the metro series and plot that larger dataframe:

A screen shot of a computer program

Description automatically generated

We need to do a little more data wrangling to ensure that we can show the average income by year, or month, or day:

A screen shot of a computer program

Description automatically generated

A graph showing the average home values

Description automatically generated

**Arkansas Conclusion**

By plotting this data in a time series plot we can visually see that until 2017 Little Rock, AR had the highest average home value according to the Zillow dataset. Additionally, in 2017 Fayetteville, AR surpassed Little Rock, AR in the average home value by city in Arkansas. A couple of other data points that you can visualize in this data is that in 2008 the average home price in all areas took a fairly significant dip, but Searcy, AR did not seem to take the large dip that the other areas did. In 2008 there was a significant housing bubble burst that dropped average home prices around the United States and that can visually identified by plotting the data in a time series plot.

**Arizona Exploratory Data Analysis**

We took a very similar approach with the Arizona exploratory data analysis. First, we sliced the dataframe to only include the metro areas that we are doing time series analysis on (Yuma, Phoenix, Payson, Tucson).

A screen shot of a computer code

Description automatically generated

We have already created the functions to plot the data from these smaller dataframes and we can reuse those functions from the Arkansas analysis.

**Phoenix, AZ**

A graph showing a line

Description automatically generated

**Tucson, AZ**

A graph showing a line

Description automatically generated

**Payson, AZ**

A graph showing a line

Description automatically generated

**Yuma, AZ**

A graph showing the growth of a stock market

Description automatically generated

**All Metro Areas in Arizona**

In order to merge all of these plots together, we need to merge the dataframes into a larger dataframe that includes all the metro series and plot that larger dataframe:

**A screen shot of a computer program

Description automatically generated**

A graph showing the average home values

Description automatically generated

**Arizona Conclusion**

By plotting this data in a time series plot we can visually see that since 1997 Phoenix, AZ had the highest average home value according to the Zillow dataset. Payson, AZ was the cheapest place to live out of these four metro areas until 2007, and in 2007 Payson was overtaken by Yuma as the lowest average home value out of the four metro areas in this study. A couple of other data points that you can visualize in this data is that in 2008 the average home price in all areas took a significant dip. In 2008 there was a significant housing bubble burst that dropped average home prices around the United States and that can visually be identified by plotting the data in a time series plot.

**Best Investment Opportunities**

The key question we are trying to answer in this lab is what three zip codes provide the best investment opportunities for the Syracuse Real Estate Investment Trust (SREIT). In order to do this we need to split the data into a training set and a testing set and get some visualizations of the machine learning models to determine how well they are performing. We are going to split the data from 1997 – 2016 for testing and try to predict the home prices in 2017 to determine the best zip code investment opportunities. Below how we split the data into training and testing columns:

A screenshot of a computer program

Description automatically generated

Below is time series plot of the training data where you can easily see the changes in average home values throughout the years.

A graph showing a line

Description automatically generated

Multiplicative Decomposition is a fundamental technique in time series analysis used to disentangle the various underlying components that contribute to the observed fluctuations in time series data. This method aims to break down a time series into three primary components: trend, seasonal, and residual (error), with the key assumption that these components interact multiplicatively. By understanding how these components interact and influence one another, analysts gain valuable insights into the patterns, trends, and seasonal variations within the data, which can be instrumental in making informed forecasts, identifying anomalies, and uncovering underlying dynamics within the time series. Multiplicative Decomposition is a powerful tool for time series decomposition, facilitating in-depth exploration and analysis of complex temporal data sets. Below you will find a Multiplicative Decomposition chart for the real estate training dataset.

A graph of a number of different types of graphs

Description automatically generated with medium confidence

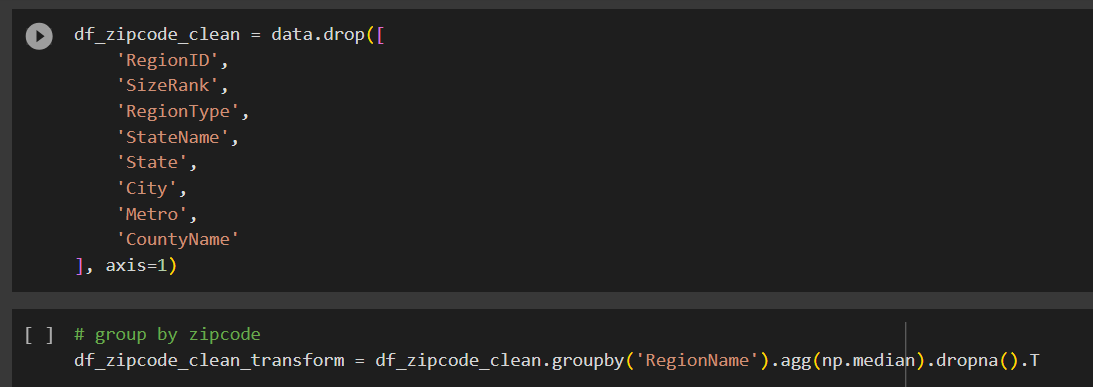
Next, we will conduct a Augmented Dickey-Fuller (ADF) test on a time series, aiming to assess the stationarity of the data. The null hypothesis of the test assumes that the time series is non-stationary, and if the resulting p-value is less than 5%, the null hypothesis is rejected, indicating that the time series is stationary.

**A screen shot of a computer program

Description automatically generated**

The ADF test results suggest that the time series data provided has an ADF statistic of approximately -1.615 and a p-value of approximately 0.476. The critical values at the 1%, 5%, and 10% significance levels are approximately -3.460, -2.875, and -2.574, respectively. Since the p-value (0.476) is greater than the commonly used significance level of 0.05 (5%), the null hypothesis of non-stationarity cannot be rejected. This indicates that the time series is likely non-stationary, meaning it exhibits trends or seasonality that affect its statistical properties, and further analysis may be required to make it stationary for more reliable time series modeling and forecasting.

In order to estimate the top three best zip codes to invest in for SREIT, I chose to utilize a Random Forrest Regressor model. Random Forest is an ensemble machine learning technique that combines the predictions of multiple decision trees to make accurate predictions. Before we utilize the Random Forrest model we need to clean up the data a little bit and group on the zip code column. First we will remove all non-numeric columns besides the zip code.



Next, we can view all of the zip codes by home value and date:

A graph of different colored lines

Description automatically generated

The code below will utilize the cleaned data to again split the data into testing and training data and then use the Random Forrest Regressor to make predictions on the best zip codes to invest in using their mean square error values:

A screenshot of a computer program

Description automatically generated

A screenshot of a phone

Description automatically generated

A blue and white graph

Description automatically generated

Based on the Random Forrest Regressor model that we utilized on the Zillow dataset the top three zip codes that should be invested in are Youngstown, Ohio; Muskogee Oklahoma; and Forrest City, Arkansas. We did down sample the data to only include data from 1997 – 2017 for this exercise due to performance. Even with down selecting the data the Random Forrest model still took over 15 minutes to run and predict the best zip codes to invest in.

**Conclusion**

In the ever-evolving landscape of real estate investment, the utilization of data science has emerged as an essential resource for making well-informed decisions and optimizing investment returns. The demand for dependable investment opportunities continues to rise, and real estate investment trusts (REITs) are pivotal in directing capital into promising markets. This paper has embarked on an intriguing journey into the realm of real estate data science, with a particular focus on the Syracuse Real Estate Investment Trust (SREIT) and its ambitious endeavor to forecast the most financially rewarding investment prospects. Through the effective utilization of data acquisition, preprocessing, exploratory analysis, predictive modeling, and insightful interpretation, this study was conducted to unveil latent patterns and trends within the United States real estate market. It aims to equip SREIT with the knowledge and tools essential for strategically selecting the three most promising zip codes for investment. By combining the rigor of data science methodologies with the practicality of real-world investment decisions, this paper successfully bridges the gap between data-driven insights and the pursuit of sustainable real estate investment success. The identification of Youngstown, Ohio, Muskogee, OK, and Forrest City, AR as the top three zip codes for investment represents a significant step toward achieving this objective, offering a solid foundation for future investment decisions.

**Bonus Content**

Top 30 Median Home Values by State

A computer screen shot of a code

Description automatically generated

A graph of a number of home values

Description automatically generated

US Map based on Median Home Values

A screen shot of a computer program

Description automatically generated

A map of the united states

Description automatically generated

Note to Professor:

Did not utilize another dataset outside of the Zillow dataset. I had a hard time with this lab and did not have time to add in another dataset.