ANALYSIS OF POPULATION GROWTH & HOUSING DEMAND

Randy Tilson

Table of Contents

Project Overview

A. Project Highlights

A1. Research Question

This project aspired to answer the following question: Where is population growth and housing demand the strongest, and what are the determinant factors that lie from within this growth and demand. As shown within the context of this project, the answer to this question is essential for survival within the dynamic environment of homebuilding.

A2. Project Scope

This project includes a Jupyter Notebook capable of accessing the Federal Reserve Database through API configuration. The notebook and the given Python script produces a CSV file output with the ranked results based on combined strength for each metropolitan area. The solution does not include automation at the level of automatic updates. Updates will need to be performed manually by removing or renaming the stored CSV file and running the script as deemed appropriate given the release dates of the Federal Reserve data series.

A3. Solution Overview

Within the tool, the individual variables of population, price parity, and personal income have been segmented by their mean value to create a high value and low value dataset. As an example, a segment of population created two separate data-frames, one that contains regions with population values that are above the overall mean value, and one that contains regions with population values that are below the overall mean value. From there, two variables were analyzed within each dataset, population, and home price index. The difference between the two isolated time periods was measured for the given variable within each metropolitan area and as such became the finalized metric to determine statistical significance.

Project Plan

B. Project Execution

B1. Project Plan

The primary goal of automating the task of accessing, combining, summarizing, and analyzing different U.S. metropolitan areas using a Python script from within a Jupyter Notebook was achieved without change to the original proposal planning. This was possible due to the large foundational knowledge of the data and its given limitations obtained within the project proposal phase of the project. As such, the objectives and deliverables shown below were achieved as projected within the proposal.

1) **Objective 1**: Collect and gather the required data.
    i) **Deliverable1**: A Python function capable of retrieving all available metropolitan area data for a given variable.
    ii) **Deliverable2**: Summarized and accurate data for each metropolitan area.
2) **Objective 2:** Identify patterns/trends within population growth and housing demand.
    i) Deliverable: The identification of at least two variable associations that show a measure of statistical significance.
3) **Objective 3:** Identify the strongest metropolitan areas within the predetermined statistically significant trend.
    i) Deliverable: Provide a ranked output of the results in CSV file format to be stored locally.

B2. Project Planning methodology

In completing this project, the KDD Process (Knowledge Discovery in Databases) was implemented. This process consists of data cleaning, data integration, data selection, data

transformation, data mining, pattern evaluation, and knowledge representation. What follows is a representation of how the proposed project progressed within each KDD phase.

1) **Data Cleaning**: Individual data sets were checked for null values. The results from within this were minimal, so mean column values were used as a replacement.
2) **Data Integration:** A function was created to easily call upon the Federal Reserve API.
3) **Data Selection**: The independent variables of Population, Gross Domestic Product, House Price Index, Labor Statistics, Regional Price Parities, and Personal Income for each metropolitan area were retrieved from the Federal Reserve Database.
4) **Data Transformation**: The independent variables for each metropolitan area were combined into a singular data-frame.
5) **Data Mining:** Each metropolitan area had its independent variables split across two individual timeframes with approximately 5 years of data within each time-series. The variables were calculated as a percentage of change from the first to the second time series.
6) **Pattern Evaluation:** Variable associations that show a measure of statistical significance towards population growth and housing demand were identified.
7) **Knowledge representation:** A CSV file was produced showing the strongest metropolitan areas using the identified patterns and combined values from within the variables.

B3. Project timeline and milestones

As mentioned earlier, the project plan was able to be followed without change due to the large foundational knowledge that was achieved within the proposal phase. As such, the project was delivered within the proposed timeline of 28 hours spaced over a period of 4 days. Of note though, 4 hours was dedicated within a milestone of creating a technical manual for interacting with the notebook. It was clear that sufficient markdown could be incorporated into the notebook to eliminate the need for this item. This time savings was beneficial, as the function needed for data retrieval took slightly longer than expected and was offset by the 4-hour time savings of not producing a technical manual.

Methodology

C. Data Collection Process

C1a. Actual data selection vs. planned collection process

Given the exploratory nature from within the project, a wide range of variables were chosen within the project proposal phase. These variables proved to be sufficient towards the identification of at least two variables that showed a measure of statistical significance within either population or home price index growth. Also, the collection of these variables was identical to the initial planning, in that a function was created that allowed for an API connection. Lastly, the independent variables were combined into a singular data frame as originally planned.

C1b. Data collection obstacles

The main obstacle encountered within the data collection phase of this project was the identification of similar data series sets for each individual metropolitan area. This was overcome within the primary function, where a list was populated with all similar series ids. This list was then able to be used and called upon within each individual metropolitan area.

C1c. Unplanned data governance

As documented within the proposal, the data from within this tool is of public domain and free to use internally in a commercial setting, however, to maintain regulatory and legal compliance, the data must not be distributed within a commercial fashion outside of the organization. Abiding to this with the use of public domain data, left the data governance aspect of this project without any unplanned issues.

C2. Advantages and Limitations of Data Set

The data set carries many advantages and limitations. One advantage that is also a limitation is The periodic release of new data within each series. This is advantageous with the aspect updating the provided tool; however, the tool is limited by the schedule of the releases. Data is released on differing schedules based on monthly, quarterly, and yearly releases. Given this, it

should be noted that the data will always be the most current available, but an annual release date can hinder the reactive speed of the tool. However, a wide variety of accurate and credible data from a singular source far outweighs the compromised effect of the release schedule.

D. Data Extraction and Preparation Processes

The data was collected through the API provided from within the FRED database service. The final dataset originated from multiple data series from within the FRED database, so the common unique id of region name was used within a final join to create a singular data frame of all variables as they relate to each metropolitan area.

E. Data Analysis Process

E1. Data Analysis Methods

Multivariate Descriptive analysis was used within this project. Data is then aggregated and segregated across two distinct time periods for comparison. The project aims to provide a descriptive summary for each metropolitan market and provide any correlated migratory patterns through a comparative analysis, as such a descriptive technique is most suitable. Splitting the data into two distinct time periods and aggregating the data from within allows for an easy comparison analysis between the two time periods to distinguish the patterns that lie from within.

E2. Advantages and Limitations of Tools/Techniques

Python was used within a Jupyter Notebook to perform all analyzations and calculations. This Provided an excellent method to both develop the required code and perform the necessary data analysis from within the results. It could be said that an R development solution may provide more built-in functionality in terms of statistical calculations and visualization, however Python is also widely used and has many built-in packages to overcome these limitations.

In reflecting on the techniques used, aggregating, and summarizing the data from each variable as they relate to an individual metropolitan area provided a simplified singular tabular result. This

single result containing each of the processed variables provided an efficient means of comparison. However, the singular result was dependent on a large amount of code and data processing which was time intensive within the production phase. Also, this approach was designed purely for a descriptive analyzation, and does not lend itself well towards building a machine learning model.

E3. Application of Analytical Methods

In this situation, descriptive methods were used. As such, multi-year data was aggregated and formatted as a single row, multiple field tabular result for each metropolitan region for easy comparison. The singular result consists of the most current measurement of a given variable such as population, and the variable percent change from the time series starting point. Also, time series data was split into two segments and treated as two separate columns within each data frame, with each time series being roughly 5 years in length. This was appropriately used to identify changing patterns or trends from one segment to the next in-between the selected features.

The project contains the assumption and requirement of identifying at least two variables that are of possible correlation within population or home price index growth. To verify this assumption and requirement, multiple independent t-tests were used to prove and provide a measure of statistical significance.

Results

F. Project Success

F1. Statistical Significance

Within the tool, the individual variables of population, price parity, and personal income are segmented by their mean value to create a high value and low value dataset. As an example, a segment of population creates two separate data-frames, one that contains regions with population

values that are above the overall mean value, and one that contains regions with population values that are below the overall mean value. From there, two variables are analyzed within each dataset, population, and home price index. The difference between the two isolated time periods is measured for the given variable within each metropolitan area and as such becomes the finalized metric to determine statistical significance. The null hypothesis was made that the difference between time periods amongst the high and low value datasets will not be significantly different. To reject the null hypothesis, and prove statistical significance, an independent t-test and P-value with an alpha value of .05 was used. Multiple measures of statistical significance using these methods were ultimately identified and are shown below:

Population growth:

- A comparison of smaller population vs larger population was made. The comparison produced a P-value of *0.0000296* with higher growth among smaller population areas.
- A comparison of lower cost of living vs higher cost of living was made. The comparison produced a P-value *0.0000284* with higher growth among lower cost of living areas.
- A comparison of lower income vs higher income areas was made. The comparison produced a P-value *0.0000000037* with higher growth among lower income areas.

These comparisons all show a P-Value that is significantly less than the benchmark of .05. It can then be said that there is measurable statistical significance pointing towards a population growth bias within areas of smaller population, lower cost of living, and lower income.

Home price growth:

- A comparison of lower income vs larger income areas was made. The comparison produced a P-value of *0.0000157* with higher growth among lower income areas.

This comparison also shows a P-Value that is significantly less than the benchmark of .05. It can also then be said that there is a measurable statistical significance pointing towards a bias within home price growth within areas of lower overall income.

F2. Practical Significance

To gain an understanding of the effect size, was important to place its evaluation in the hands of a mathematical model. With statistical significance determined by an independent t-test, a measure of the standardized distance between the two means was appropriate and is the nature of the Cohen's d calculation. Within this, a minimum factor of 0.4 was used as a baseline of acceptance. This ensures that at minimum, some meaningful strength exists within the relationship of the two variables and viability exists within the real world. Within the predetermined associations of statistical significance, Cohen's d values also showed a reflection of practical significance as measured by their values being higher than .40. For clarity, these values are shown below:

Population growth:

- Smaller population vs larger population produced a Cohen's d value of *0.49420.*
- Lower cost of living vs higher cost of living produced a Cohen's d value of *0.46855.*
- Lower income vs higher income produced a Cohen's d value of *0.66459.*

Home price growth:

- Lower income vs larger income produced a Cohen's d value of *0.42296.*

Applying this knowledge within the real world is the primary goal. In moving forward, this knowledge can be used to provide a deeper understanding as to why a given region is showing

strength over others. This deeper understanding will therefor lead to greater confidence within decision making. Lastly, a filtering layer based on these identified relationships can be placed within the final output of the proposed tool.

.

F3. Overall Success

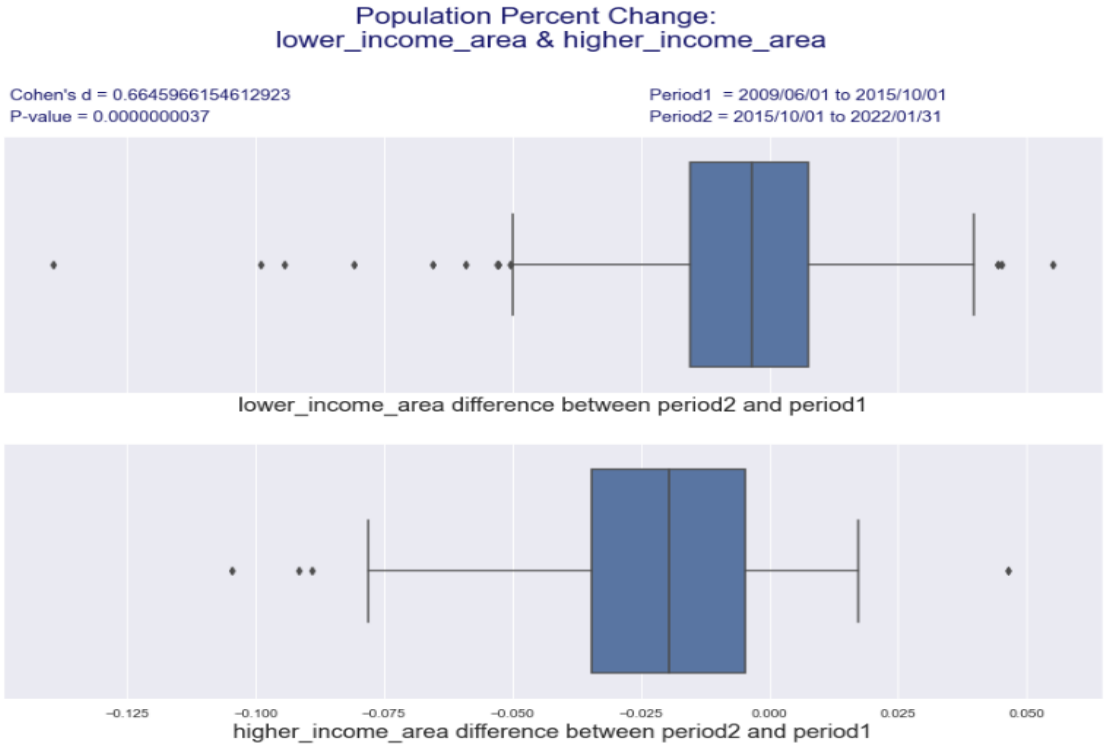
The project proposal provided three key metrics to be used as a measure of success. First was the identification of at least two variables that show a measure of statistical significance within either population or home price index growth. A total of four variables were identified, thus exceeding the original expectations set within this metric. The second metric is a locally stored CSV file containing a ranked list of metro areas based on the combined weight of all variables within a metropolitan area. This file is produced and updateable whenever needed. Lastly is providing easy interaction within the notebook using interactive text boxes to select data at an individual state level without the need for manual coding. This was also provided, and thus national rankings can be seen within granular state level. Given the usability and deliverance of the promised key metrics, this project is a success and viable business asset that can be used for all future decision making within the context of identifying where population growth and housing demand are the strongest, and the determinant factors that lie from within this growth and demand.

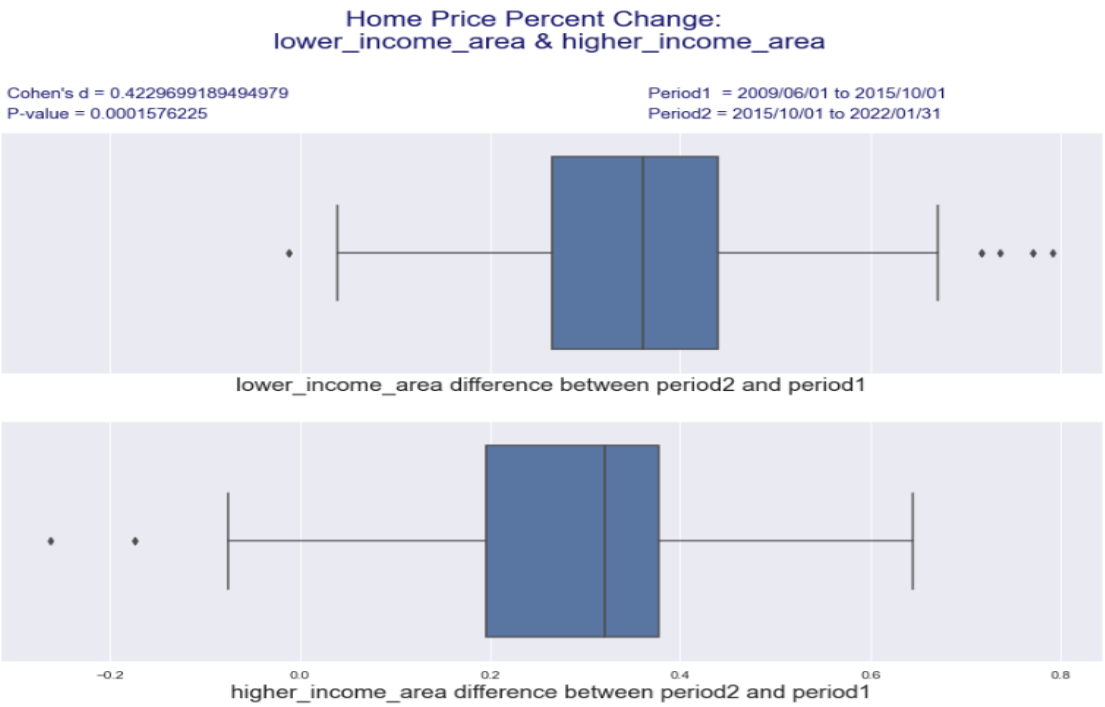
G. Key Takeaways

G1. Summary of Conclusions

The primary goal of this project was to automate the task of accessing, combining, summarizing, and analyzing different U.S. metropolitan areas. This was done to develop an easily updateable tool to assist within the decision-making process that pertains to the national expansion of a home-building enterprise. As trends change, this tool can and will update itself within the provided data to provide a filtering layer consistent with metrics that are proven to be both

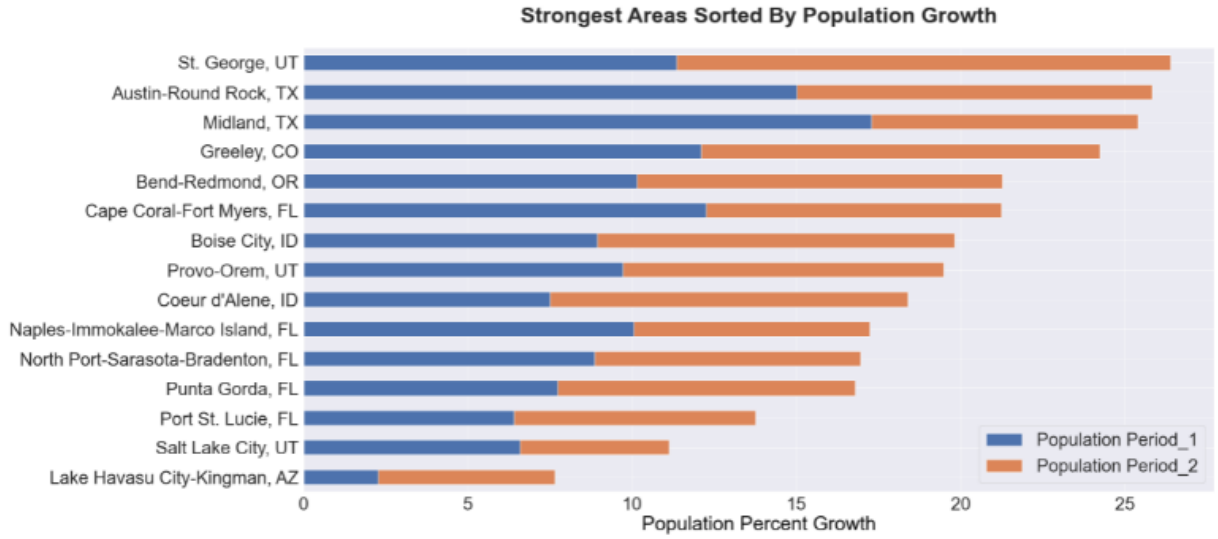
statistically and practically significant. As such, this tool can be incorporated as a reliable source of information within the operations of a home-building enterprise. A visual depiction of the measures of statistical significance can be seen below.

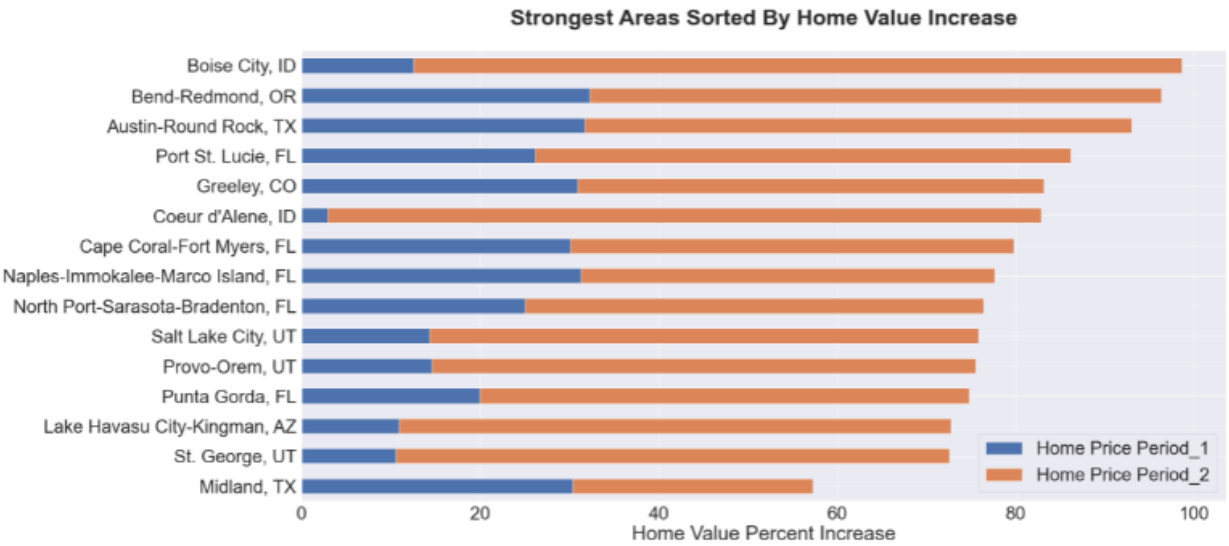




The boxplots shown above, are a representation of the distribution from within the values used to determine statistical significance. A bias towards home price and population increase within metropolitan areas of lower income is evident.

Within the filter of the identified statistical and practical significance, the strongest areas as measured by the combined value of their associated variables are shown below.





As shown above, population growth and home value increase are related, but not a 1 to 1 correlation. Exploring this within the context of identifying lower initial capital expenditures may be beneficial.

G3. Findings-based Recommendations

It should be noted that this tool is not a one size fits all. There are many possible other factors that can come into play when choosing an area of expansion, and all the results that are output within the CSV file are all strong and viable options. As an example, the strongest metropolitan area may present itself on the west coast while current operations are located on the east coast. In this scenario, it would be wise to explore some of the options that are geographically closer to the home base of operations.

The initial research question was, where is population growth and housing demand the strongest, and what are the determinant factors that lie from within this growth. Given the determinant factors found from within this tool, future homebuilding should focus primarily on metropolitan areas with smaller populations, a lower cost of living, and a lower personal income. In doing so, a final product will be met with the required homebuyer demand. Lastly, in narrowing down towards the demand results as measured by the combined variable values, it would be prudent to explore the idea of capitalizing on an area that has sustained high population growth and a more modest home price growth. In doing so, home values may have more room for sustained growth within the timeframe of the home-building project. Given the market structure of supply vs demand, it could also be assumed that these areas may have more land available to build, which can expedite the process.

Sources

Citations listed below are in the preferred provider format as provided by the FRED database at the origination point of the data download.

U.S. Census Bureau, Resident Population in Charleston-North Charleston, SC (MSA) [CRLPOP], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CRLPOP, January 4, 2022.

U.S. Bureau of Economic Analysis, Total Gross Domestic Product for Charleston-North Charleston, SC (MSA) [NGMP16700], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/NGMP16700, January 5, 2022.

U.S. Federal Housing Finance Agency, All-Transactions House Price Index for Charleston-North Charleston, SC (MSA) [ATNHPIUS16700Q], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/ATNHPIUS16700Q, January 5, 2022.

U.S. Bureau of Labor Statistics, All Employees: Mining, Logging, and Construction in Charleston-North Charleston, SC (MSA) [CHAR745NRMN], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CHAR745NRMN, January 6, 2022.

U.S. Bureau of Economic Analysis, Regional Price Parities: All Items for Charleston-North Charleston, SC (MSA) [RPPALL16700], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/RPPALL16700, January 5, 2022.

U.S. Bureau of Economic Analysis, Per Capita Personal Income in Charleston-North Charleston, SC (MSA) [CHAR745PCPI], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CHAR745PCPI, January 5, 2022.