Analysis of Population Age and Residential Rental Prices

Dustin M. Kraft

Western Governors University

**Table of Contents**

[Project Overview 3](#_Toc114756717)

[A. Project Highlights 3](#_Toc114756718)

[Project Plan 3](#_Toc114756719)

[B. Project Execution 3](#_Toc114756720)

[Methodology 4](#_Toc114756721)

[C. Data Collection Process 4](#_Toc114756722)

[C1. Advantages and Limitations of Data Set 4](#_Toc114756723)

[D. Data Extraction and Preparation Processes 5](#_Toc114756724)

[E. Data Analysis Process 5](#_Toc114756725)

[E1. Data Analysis Methods 5](#_Toc114756726)

[E2. Advantages and Limitations of Tools/Techniques 5](#_Toc114756727)

[E3. Application of Analytical Methods 5](#_Toc114756728)

[Results 6](#_Toc114756729)

[F. Project Success 6](#_Toc114756730)

[F1. Statistical Significance 6](#_Toc114756731)

[F2. Practical Significance 7](#_Toc114756732)

[F3. Overall Success 7](#_Toc114756733)

[G. Key Takeaways 7](#_Toc114756734)

[G1. Summary of Conclusions 7](#_Toc114756735)

[G2. Effective Storytelling 7](#_Toc114756736)

[G3. Findings-based Recommendations 8](#_Toc114756737)

[H. Panopto Presentation 8](#_Toc114756738)

[Appendices 8](#_Toc114756739)

[I. Evidence of Completion 8](#_Toc114756740)

[Sources 8](#_Toc114756741)

# Project Overview

# Project Highlights

**Research Question**

The question that this analysis intended to examine is whether population age could be used to determine higher residential rent in relation to market assessments for the real estate investment industry. Real estate investors use various metrics to identify geographic areas in which to purchase free-cash-flowing rental properties. If available demographic data can be used to pinpoint an additional metric to improve odds of profitability, it may provide a competitive edge in the rental market.

**Project Scope**

The scope of this project included the determination of whether or not the gathered dataset suggests a relationship between the target variables can be used in an initial assessment of a real estate market. This is only one element of a full-scale market analysis and, even if a correlation is found, cannot fully determine profitability. Additionally, this analysis did not examine individual properties or types of properties.

**Solution Overview - Tools and Methodologies**

The datasets were selected from the American Community Survey at data.census.gov and broken down by Zip Code Tabulation Area (ZCTA), the US Census Bureau’s classification for geographic areas by zip code. Data was downloaded in .csv format. Because of the manageable size, and simple tasks needed, the data was cleaned in Excel. The analysis application was coded in Python and relies on libraries related to data, such as Pandas and NumPy. The application reads the .csv files, filters records with null values or high margin of error values, aggregates the required ratios. The complete dataset is exported in .csv format for future use. The linear regression is then executed, outputting the graphs and calculated statistical measures.

# Project Plan

# Project Execution

**Project Plan**

The original plan for this analysis relied on the use of the percentage of population between the ages of 20 and 34 as well as using median rent and median home value data to calculate the “price-to-rent ratio” as described in the article “*8 Must-Have Numbers for Evaluating a Real Estate Investment* (Lerner, 2022).” After completing the analysis with these measures, it was determined that, because of the nature of the price-to-rent ratio, it does not provide the intended measure. Generally, areas with a price-to-rent ratio of 15% or less are targeted by investors. In looking for a positive correlation between this ratio and the age variable, the analysis would actually be targeting areas with lower rents. In its place, a rent-to-income ratio was created to measure the portion of income the average person can expect to pay in rent each year. This new ratio creates a standardized measure of whether rent is considered high or low relative to other ZCTAs.

**Project Planning Methodology**

The KDD project management methodology was used, as outlined in Task 2.

1. Goal-Setting and Application Understanding
   1. A goal was established to perform an analysis of data regarding population and contract rents.
   2. A testable hypothesis was formed as follows: “A higher percentage of adult residents between the ages of 20 and 34 will correlate with higher rent-to-income ratio in that area.”
2. Data Selection and Integration
   1. Used data.census.gov to download tables containing population age, median rent, and median income broken down by ZTCAs.
3. Data Cleaning and Preprocessing.
   1. Several unnecessary columns were removed and Null values added to fields with missing or unusable values, such as “\*\*\*\*\*”, or “-”.
4. Data Transformation
   1. Python was used to combine tables into one dataframe, remove empty records, limit the dataset to records below a certain threshold for margin of error and calculate the values for percentage of population in the target age group and median rent-to-income ratio.
5. Data Mining
   1. Performed regression analysis on the explanatory and response variables.
6. Pattern Evaluation/Interpretation
   1. Generated scatterplot and plot of regression line.
   2. Calculated and returned statistics necessary for hypothesis test.
7. Knowledge Discovery and Use
   1. Evaluated target statistics to complete test of the hypothesis.

**Project Timeline and milestones**

The original data extraction, cleaning and analysis were completed on the scheduled dates and within the planned duration. The change in project plan required additional data collection and analysis which was performed on the next available time slot of 9/19/2022 to 9/20/2022. The completion date of the final analysis document is 9/21/2022.

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Start Date** | **Completed Date** | **Duration (days/hours)** |
| Establish requirements | Completed in Task 2 | | 3 hours |
| Locate and download appropriate data tables | 9/15/2022 | 9/15/2022 | 2 hours |
| Clean data | 9/15/2022 | 9/15/2022 | 2 hours |
| Perform analysis | 9/16/2022 | 9/16/2022 | 2 hours |
| Additional Data Collection and Analysis | 9/19/2022 | 9/20/2022 | 2 hours |
| Generate analysis results document | 9/21/2022 | 9/21/2022 | 3 hours |

# Methodology

# Data Collection Process

**Actual data selection vs. planned collection process**

Data on median household income was collected to replace the data on median home value. Data was still collected from the U.S. Census Bureau American Community Survey, via data.census.gov.

**Obstacles to data collection**

The U.S. Census Bureau data tables contain several data points for each area, broken down by age, race, household characteristics, etc. The data collection and cleaning process required narrowing large datasets down to a small number of attributes before use within the analysis.

**Unplanned data governance handling**

The data taken from the American Community Survey are publicly available and the datasets used contain no personally identifiable or proprietary information. No data governance issues were encountered.

# C1. Advantages and Limitations of Data Set

The American Community Survey (ACS) dataset contains data for every community in the US, making it one of the most extensive surveys of the American population. This allows for the collection of data on topics such as the age and real estate elements used in this analysis in a common format, rather than combining data from multiple sources. However, the ACS is an average of samples over five years. Since it is an average of multiple surveys, performed at multiple times, the samples may be less accurate than a single point survey. Margin of error (MOE) is included for all data points, but many MOE values were far too large to be considered usable. Larger MOE values are often associated with smaller population sizes. Limiting the use of datapoints with large MOE values, may be skewing the data toward more densely populated areas.

# Data Extraction and Preparation Processes

The raw data was downloaded in .csv format using the U.S. Census Bureau’s data platform at data.census.gov. Because of the manageable size of the .csv data, it was most efficient to clean in Excel before importing, combining and aggregating in Python.

# Data Analysis Process

# E1. Data Analysis Methods

To complete this analysis, the inferential statistical methods of regression and hypothesis testing were used. Hypothesis testing is the accepted method used to address a specific business question through data. Linear regression was chosen because the hypothesis is relational in nature and relies on two variables, both of which are continuous.

# E2. Advantages and Limitations of Tools/Techniques

Excel allowed for the quick deletion of the large number of unneeded columns and replacement of non-numeric values with null values. The limitation of this approach is that it would not scale to larger datasets. The use of Python and .csv files allowed for data collection and transformation without the construction of a database. The preparation for regression analysis required the suppression of datapoints with a high margin of error. This may have limited data from smaller communities.

# E3. Application of Analytical Methods

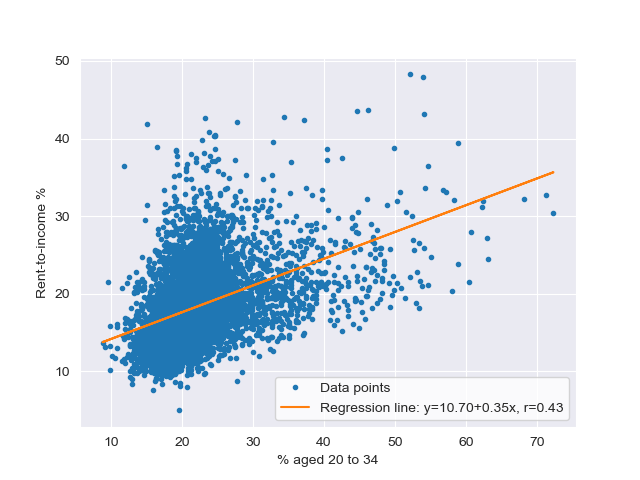
For the regression analysis, the percentage of the population between 20 and 34 was used as the explanatory variable (x). The median rent-to-income ratio was used as the response variable (y). The hypothesis posed stated that a higher percentage of adult residents between the ages of 20 and 34 would show no statistically significant relationship with a higher rent-to-income ratio in the same area. I then attempted to disprove that hypothesis through linear regression, thereby proving that a higher percentage of residents in the target demographic shows a statistically significant relationship with higher rent-to-income ratios.

# Results

# Project Success

# F1. Statistical Significance

The explanatory variable (x) used in the regression was the percentage of population between 20 and 34 by individual ZCTA. The response variable (y) used was the rent-to-income ratio aggregated from the median rent and median income for each ZCTA. The relationship between the two variables was measured by the regression coefficient (the measure of the slope of the regression line) at 0.35. This figure shows a positive correlation between x and y. The p-value was used as the measure of significance and was calculated to be 1.9301994559238115e-247, which was rounded to 0.00. This result shows that the positive relationship of 0.35 is statistically significant when compared to a 95% confidence interval and the null hypothesis is rejected.



**Results of Regression Analysis:**

slope: 0.34509386854740903

intercept: 10.703969717610681

r: 0.4276449048752109

p-value: 1.9301994559238115e-247

The correlation coefficient, shown by the r-value, was used to measure the strength of the relationship between the variables. The calculated value of 0.43 failed to meet the chosen cut score for success. While the selection of an r-value threshold is often arbitrary and varies from one application to another, the value lower than 0.50, combined with the wide distribution shown in the scatterplot suggests that, although there is statistically significant correlation between x and y, it is likely not a reliable predictor on its own.

# F2. Practical Significance

In practical terms, what was found in this analysis is not a smoking gun. While a correlation was found and the null hypothesis was rejected, if business decisions are made solely on this relationship, many profitable markets such as city centers or retirement communities, could be eliminated unnecessarily. However, the analysis does suggest that the portion of a population aged 20 to 34, combined with other factors such as employment and marriage rates, might be useful as a more nuanced method of assessment for new markets.

# F3. Overall Success

The project was a success in that a correlation was proven and the null hypothesis was rejected. This shows that there is a relationship between the amount of young people in an area and an upward pressure on residential rents. However, the project failed in its goal to establish a time-saving metric to quickly target, or eliminate markets on its own.

# Key Takeaways

# G1. Summary of Conclusions

The main goals for this project were to find a positive correlation between the number of young adults in a population and higher rents, determine its significance and decide whether the relationship is useful in decisions over whether or not to target an area for investment. The statistical results are summarized in the following chart:

|  |  |  |  |
| --- | --- | --- | --- |
| **Criterion/Metric** | **Required Data** | **Cut Score for Success** | **Results** |
| Regression Coefficient | Slope of the regression line | > 0 | 0.35 |
| Measure of Significance | p-value | < 0.05 | 0.00 |
| Correlation Coefficient | r-value | > 0.5 | 0.43 |

As the results show, a positive correlation was found and proven to be statistically significant. However, the strength of the relationship failed to meet the level that should be considered reliable on its own to make decisions. The last goal of the project was to create a cleaned dataset that can be used for further study of the topic. The complete dataset is part of the included files.

# G2. Effective Storytelling

A scatterplot was used as it works well in illustrating the relationship and distribution with two variables. The line from the linear regression analysis was then plotted over a scatterplot of the same data to show the relationship.

# G3. Findings-based Recommendations

The question that this analysis intended to examine is whether population age could be used to determine higher residential rents in relation to market assessments for the real estate investment industry. Since a correlation was found, but not found to be strong enough to predict on its own, one recommendation is to investigate other relationships that may correlate with higher rent to combine into a more detailed model for prediction. Additionally, it is recommended to begin performing targeted surveys in zip codes where data is lacking or contains a high margin of error within the current dataset. Over time these two assets can add significant value to the company’s success and profitability.

# Panopto Presentation

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=238c0352-57e5-445f-9f8e-af1a01248326>

# Appendices

# Evidence of Completion

* Panopto presentation
* Python Analysis
* Results PowerPoint
* Full dataset in .csv format

# Sources

Lerner, M. (2022, July 13). *8 Must-Have Numbers for Evaluating a Real Estate Investment*. Investopedia. Retrieved September 6, 2022, from https://www.investopedia.com/financial-edge/0511/8-must-have-numbers-for-evaluating-a-real-estate-investment.aspx