Capstone Project Technical Report

Han Luong, Jed Dryer & Phil Carbino

Introduction:

For our capstone project, we set out to explore home affordability and attempted to predict the near-term future affordability of houses for each county in the state of New Jersey. For the purposes of our project, a house is considered “affordable” if the mortgage payment for said house is no more than 25% of the household’s income. So, for a county to be considered “affordable”, a household making the median income for the area should be able to buy a median priced house, and their mortgage payment should not exceed 25% of their income. While attempting to build a model for our predictions, we were also interested in examining what effect, if any, amount of new housing construction projects had on whether or not housing was affordable.

The factors that we are taking into consideration for our prediction are:

- Median household income in the area
- Amount of new housing construction permits in the area
- Average mortgage interest rates
- Median house prices in the area

We also set out to take a look at some statistics surrounding these factors to see if we might learn anything interesting. In going through our data, we wanted to find the answers to the following questions:

1. What counties have the highest median income on average?
2. In what locations is home ownership most and least affordable?
3. During which time periods has home ownership been most and least affordable?
4. How has the median household income changed over the years when adjusted for inflation?
5. What age group is best situated to buy a home in different counties based solely on income? How has this changed over the years?

Data Sources:

The data sets that we used for our project came from three different sources:

- House prices were obtained from a public dataset available from Zillow:
    - (https://www.zillow.com/research/data/)
- Mortgage rate data was obtained from a Freddie Mac website:
    - (https://www.freddiemac.com/pmms/pmms30)
- Median income data was acquired from a US Census dataset:
    - (https://www.census.gov/data/developers/data-sets/acs-1year.html)
- New construction data was also obtained from the US Census:
    - (https://www2.census.gov/econ/bps/County/)

Data Overview:

After our ETL, the data we have for each county in New Jersey are:

- Median household income for 2005-2019
- Number of new units and new buildings construction permit requested for 2000-2022
- Average mortgage interest rates and average points for 2000-2022
- Median house prices for 2000-2022

Before diving into our initial questions, we performed a cursory exploratory data analysis. With the correlation matrix heatmap, we realized there was no strong correlation between our variables. Over time, as year increases the average rates were generally decreasing but the trend was not consistent. There were a small number of missing data points in our data, which we dropped for the visualization and machine learning process. Although there were no noteworthy findings with the initial data exploration, we found interesting points during our questions exploratory phase that will be addressed in each part.

Affordability Calculations:

We used a formula provided by EDUBA to calculate the monthly mortgage payment, based on the median house price, the mortgage rate. The downpayment in our calculation is 12%, which is the average down payment in the US, and we factored in an average New Jersey property tax of %1.89 across the board. We included property tax because that is included in a standard home affordability calculation. We did not include mortgage insurance since that varies by lender and does not add too much additional cost. Below is the formula for the monthly mortgage calculation, link in references.

$$\text{Fixed Periodic Payment} = P * [(r/n) * (1 + r/n)^{n*t}] / [(1 + r/n)^{n*t} - 1]$$

- **P** = Outstanding Loan Amount
- **r** = Rate of interest (Annual)
- **t** = Tenure of Loan in Years
- **n** = Number of Periodic Payments Per Year

Based on our definition above, if the monthly median mortgage payment is more than 25% of the monthly median income, we labeled it as unaffordable. We understand that each lender has a different threshold for affordability, however we decided to go by the National Association of Realtors (NAR) equation, which is a measurement of true affordability.

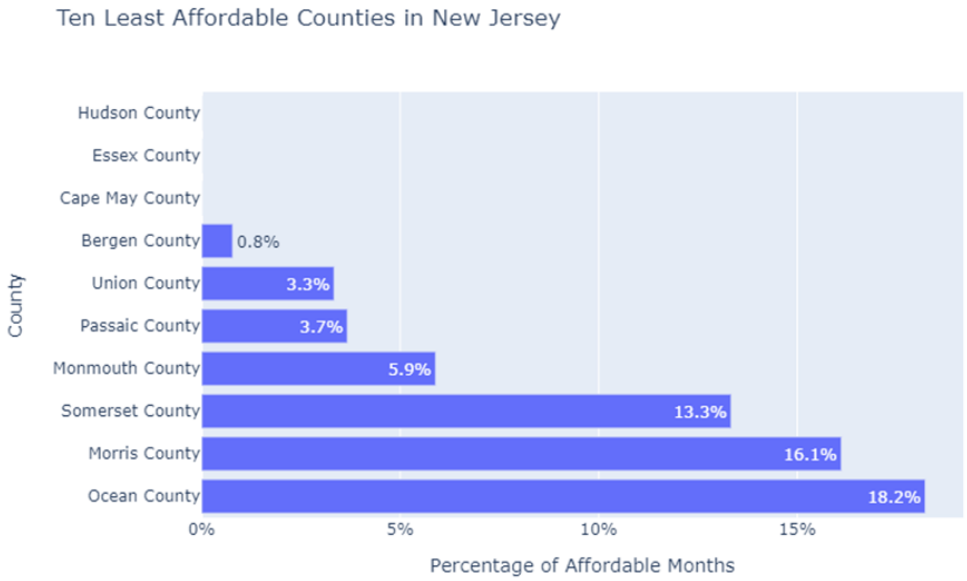
Question Exploration:

(Note that any data and information using median income for year 2020-2022 are predictions from our ARIMA machine learning, which will be explained in our later section.)

Q1. The first question we sought to answer was regarding which counties had the highest median income on average:

As we can see here, the 5 top counties for overall average income are Hunterdon County, Morris County, Somerset County, Bergen County, and Monmouth County, all of which are located in the northern part of the state in the New York Metropolitan area. While we might imagine at first glance that this could be an indicator of whether or not these counties would be considered affordable, this is obviously going to depend on house prices in the area, which are also very likely to be higher than in other counties with lower average median incomes.

Q2. Using the median income, median home price, and mortgage rate data we acquired, we could then calculate how affordable it would be to purchase a home in each of the counties we had data for. For our second question, we wanted to gain some insight as to which counties were considered most and least affordable according to our affordability criteria:

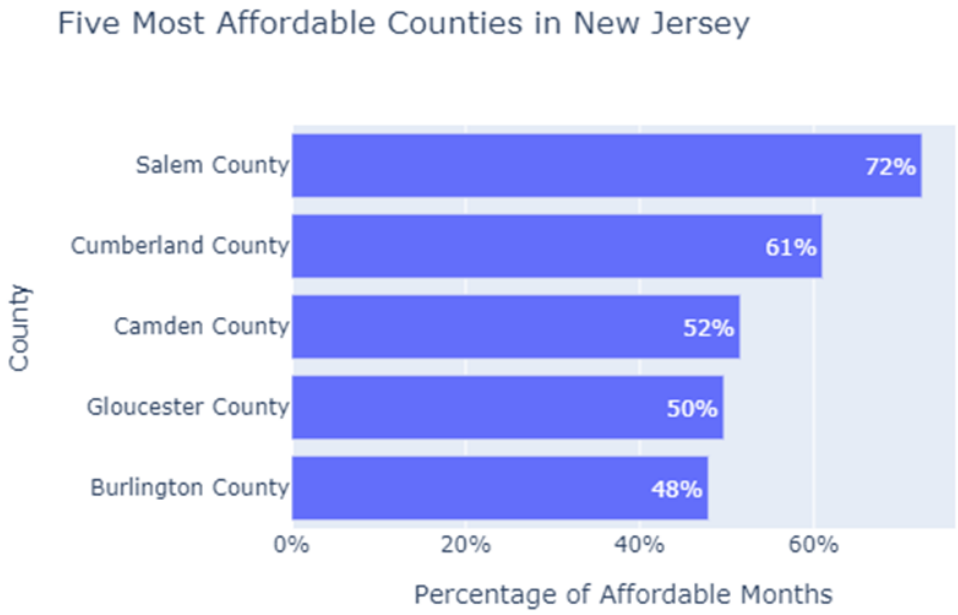


Above, we see the top 10 least affordable counties on average. Interestingly enough, three of the counties (Hudson, Essex and Cape May) are generally considered completely unaffordable by our standards. We also see that 4 of the top 5 counties for median income

discussed above (Bergen, Monmouth, Somerset and Morris) make an appearance on this list as well.

Being that Hudson and Essex counties are directly across the river from New York City, it's likely that these homes remain out of reach for most people because the areas are effectively an extension of New York and thus command comparable home prices. Bergen, Union, Passaic, Somerset and Morris counties surround Hudson and Essex to the north, south and west, which likely explains why homes in the area are still relatively unaffordable, albeit to a slightly lesser extent.

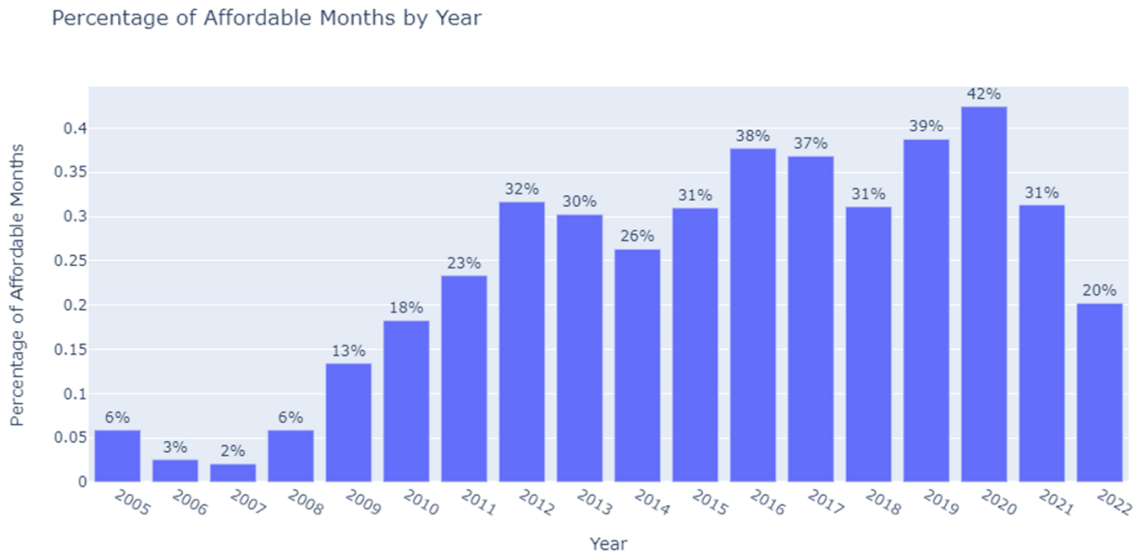
Monmouth County has the unique distinction of not only being relatively close to New York, but also being on the eastern coastline. This latter factor being something that the remaining counties, Cape May and Ocean Counties, have in common.



Moving on to the next portion of this exploratory question, we see that the 5 most affordable counties in New Jersey on average are, in order from most to least affordable, Salem County, Cumberland County, Camden County, Gloucester County, and Burlington County. All 5 of these counties are located in the south western part of the state, all of them substantially further away from New York City and the Atlantic coast than the least affordable counties.

Judging from the above discoveries, it appears that proximity to New York City or the Atlantic Coastline are quite significant in predicting the affordability of an area. The closer an area is to either of these, the less likely someone making the median income for the area would be able to afford a home.

Q3. The third question we asked was regarding what years were best and worst for home affordability:



According to the chart above, home affordability rates were extremely low from 2005 to 2008. Given that this was the peak of the housing bubble and the run up to the subprime mortgage crisis of 2008, that's entirely expected. After 2008, we see home affordability increase rather dramatically, having an initial peak in 2012, a second in 2016, and ultimately an all time high in 2020. However, the Covid-19 pandemic sent the world economy into somewhat of a tailspin. Supply chains were disrupted resulting in rising building costs during a time when demand for housing was rising as well. According to usafacts.org, "Between April 2020 and April 2021, the price of an average home increased 16%, the largest single-year increase since 1992," which is relatively consistent with the trends we discovered in our data.

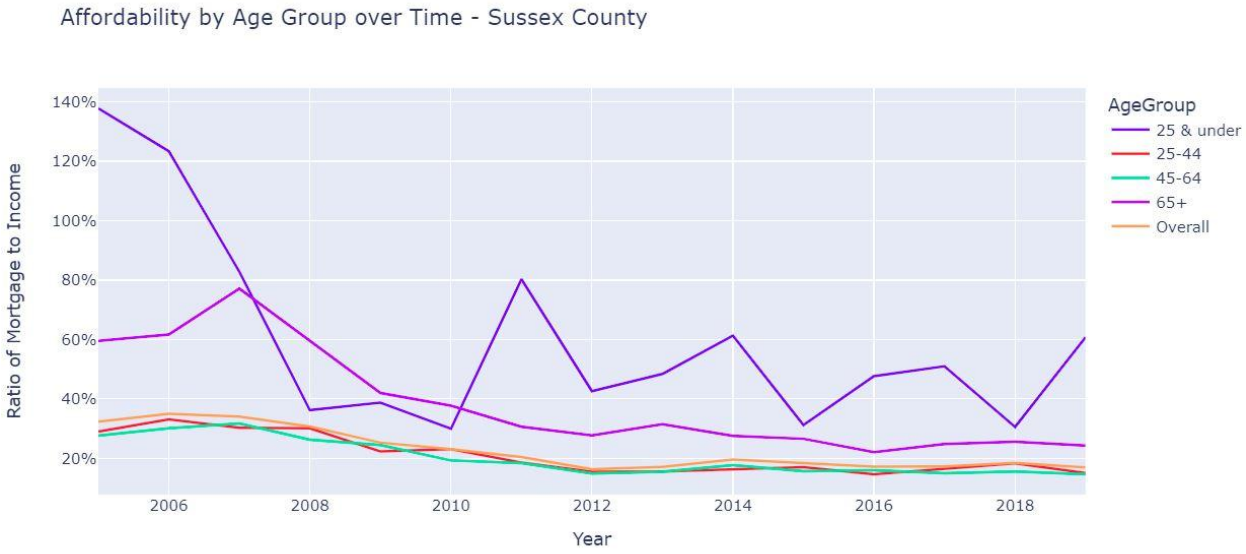
Q4. Our fourth question regards how has average median income changed over the years:

We can see that income has steadily risen over the years for all age groups. The age group 45-64 lead with the highest income every year followed by the age group 25-44. The lowest income age group, not surprisingly, is age group 25 and under. What is surprising is that the age group 65 plus follows very closely, since this is income and not savings, it makes more sense because that might be a retired group with savings instead of steady income.

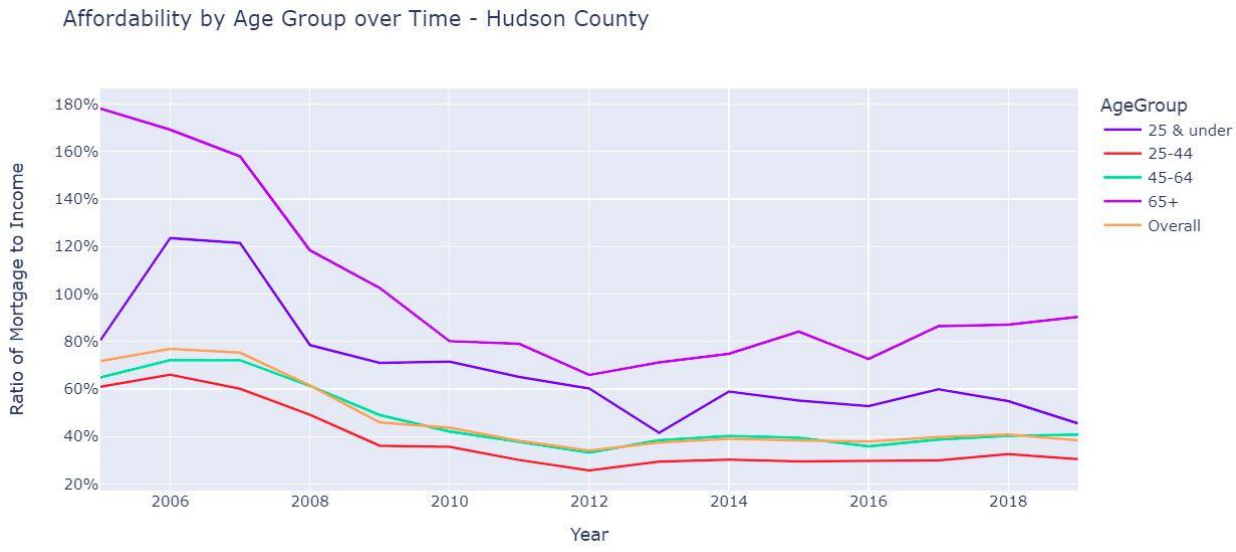
Q5. We then wanted to get an idea of what age demographic has been best situated to afford a home over time and whether or not the trends are similar for different counties:

As we can see, while the amount of the statewide average mortgage to income ratio for each age group has differed significantly over time, the relative positions of each group on the above graph does not change. Generally speaking, people 25 and under have historically been unable to comfortably afford a home being that their mortgage to income ratio has not dipped below 60% for any appreciable amount of time. People 65 and older are in the next best position, however their mortgage to income ratio is quite high from 2005 until 2012, when it begins to level off at around 40%, which is still significantly higher than the 25% threshold we've set in order for home ownership to be considered affordable. The demographics most and second most able to afford a home are people ages 45-64 and 25-44 respectively. The two groups' mortgage to income ratios are quite close and correlate very strongly across the entire span of time examined, but overall, people 45-64 years old are consistently making slightly more money than people 25-44.

We looked into whether or not these trends changed from county to county, however, in all counties except for Hudson, we discovered no significant deviation. In some counties, the data might show a period where people in the under 25 demographic tend to have a higher income than people in the over 65 demographic, which differs from the line graphs of the averages above, however, as you can see in a representative example chart below for Sussex County, these differences are usually quite short lived and return to the patterns above:



Hudson County was the only area where we found any major difference in these trends:



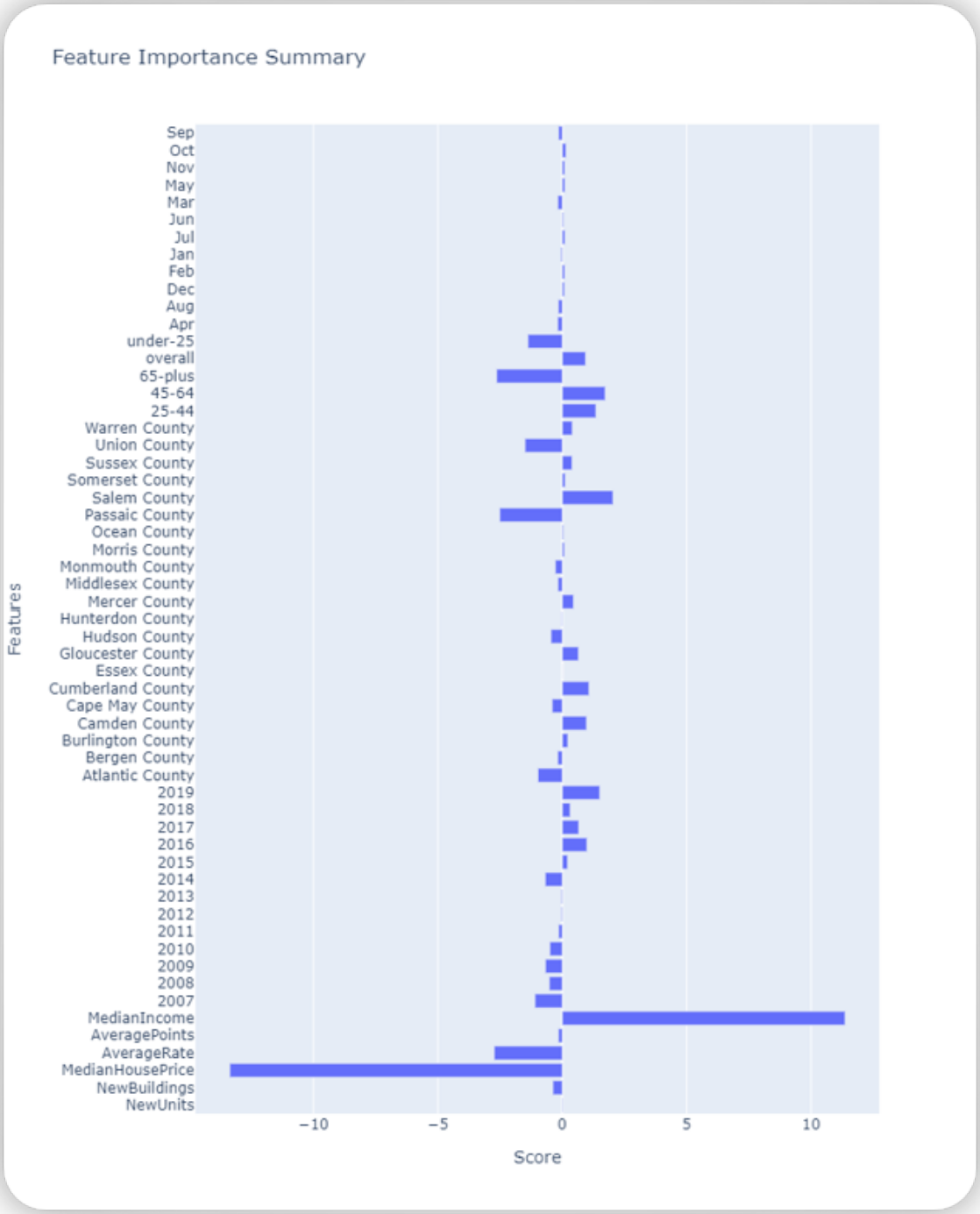
As we can see here, people in the under 25 demographic are consistently in a better position to buy a home than people in the 65 and up demographic, due to higher median incomes. However, given that Hudson County tops our list of most unaffordable, they are still not in a position to buy a home affordably, by our criteria.

Machine Learning Process:

Logistic Regression

As stated, affordability is something we are calculating with Median House Price, Median Income, and Mortgage Rates. Outside of this calculation, we have new building permit construction information. We wanted to see if this relates to home affordability - if it plays any role in determining home affordability. We decided on a simple logistic regression model to find out. The coefficient weights of the building permit information ('NewUnits' and 'NewBuildings')

indicate that building permits are not important in determining the probability of a yes or no in whether a home is affordable for that month with the Median Income, Median House Price, and Mortgage Rates. We then lagged building permits year by 2 years since building construction might not have an effect until the constructions are built. Even with the lagged data, building permits do not play a major role. Given the accuracy score of 99%, the model is pretty accurate at determining affordability. Building permit is close to zero in its weight meaning it does not play a big role in determining the probability of home affordability.



Linear/Polynomial Regression

In order to eventually predict home affordability, we needed to first make some predictions about median income, because we were lacking data for it beyond the year 2019. Initially, we considered using a standard linear or polynomial regression model to forecast, however, due to the fact that the data we were using was a time series, we decided that this likely wasn't an appropriate use case. We then considered using exponential smoothing, which is intended for use on time series data, but, ultimately we elected to use an ARIMA model.

ARIMA

ARIMA, which stands for "Autoregressive Integrated Moving Average", is a type of multilinear regression model that uses a set of previously observed target values to form a linear regression model to predict what the next value will be in the time series. The ARIMA model also uses the dependency between an observation and a residual error from a moving average model that is applied to lagged observations.

Essentially the model functions in this way; once one target value is predicted, it then applies the same formula to another set, dropping the oldest value from the previous set, and replacing it with the newly predicted value. It then makes a prediction about what the next target value will be, and repeats the process. While this algorithm is useful for predicting short term future values, its predictions lose accuracy the more it relies solely on predicted rather than observed values.

In order to appropriately use a time-series model, it must be stationary data. Stationarity is mandatory for time-series data, and this occurs when the properties of a time series do not change over time. We check stationarity by using the Augmented Dickey-Fuller test on the data which tests the null-hypothesis that a unit root is present in the sample. Stationarity is achieved through a process called differencing in which the differences between observations are calculated.

There are three hyper-parameters that are used in the ARIMA model:

- **p** : The number of lagged observations included in the model
- **d** : The number of times that the raw observations are differenced.
- **q** : The size of the moving average window, which is also called the order of moving average

We used ARIMA to predict median income values for each county and age group in New Jersey. We then used it along with the rest of the data we collected to determine which counties have previously been considered affordable by our standards, and ultimately make predictions about home affordability beyond the scope of the data we started with.

Conclusion and Recommendations

We came into the research wondering how home affordability has changed over the years, what it might look like in the future, how different age groups fared, and whether new building construction has an impact. Surprisingly, we found it was getting more affordable after the 2008 crash based on dropping house prices and median income trending the same. With other factors involved such as unemployment rate, that might look differently. The relationship between income and age groups made sense - as age increases, income increases until age 65 which may be because of retirement. New building constructions seem to trend alongside house prices but ultimately have no effect on home affordability, even when lagged and examined on its own. As house prices rise to levels of pre-2008 crash while income is not nearly increasing with the same trend, we see home affordability dropping quite drastically post pandemic.

In addition to the research on home affordability, we learned that time series data are more appropriate for non-independent variables. With more time, we would like to use ARIMA or other time series machine learning models to predict the other variables such as house prices, mortgage rates, and building constructions. Our recommendations are to incorporate other data that might further help to explain home affordability such as cost of rent, savings in different income groups for a varied down payment, unemployment, and potentially include other states within the US. Tweaking the formula to include other numbers for property tax rates, mortgage to income ratio, and down payment based on industry practices will also help complete the picture of home affordability.

References

Bonaros, B. (2021, April 28). *Arima model in Python*. Medium. Retrieved May 20, 2022, from https://towardsdatascience.com/arima-model-in-python-7bfc7fb792f9

Mortgage formula: Examples with Excel template. EDUCBA. (2021, October 13). Retrieved May 13, 2022, from https://www.educba.com/mortgage-formula/

Brownlee, J. (2020, December 9). *How to create an Arima model for time series forecasting in Python*. Machine Learning Mastery. Retrieved May 19, 2022, from https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/