Johnson Simon

ISM6419.001S22

Data Visualization

May 3, 2022

US Housing Market

INTRODUCTION

Description of the background of the project

As a new student moving to the US, I found it very difficult to find a place, and the reason was that the real-estate prices in Tampa are skyrocketing. Tampa has become a hotspot (A place of more than usual activity). Zillow estimated Tampa's home values will grow by 24.6% this year. Nationally, Zillow forecasts a 14.3% growth in national home value. I wanted to see if I can evaluate properties of my own using real estate and data. We know that Real Estate data is rich and interesting. However, it is in a silo between many different areas and so it is important to understand the landscape of the data before making decisions.

The Ambitiousness of the project

The Ambition of the project was to look at the various factors and outcomes involved in real estate prices, and understand the correlation between the parameters and the rates.

Some of the questions and ideas I wanted to explore are:

- Did values go up last year?
- Is Tampa really the Hottest Market?
- What type of market are we in? Buyers or Sellers?
- How is the age of the construction or the proportion of owner-occupied units established related to the price of real estate?

And knowing the Key Performance Indicators, if any.

Research questions

Theory **1**: If lower interest rates drive home buying, then inventory (listing of a property for sale) will decline.

Theory **2**: What are the Housing Market hotspots in the US? Is Tampa the hottest market for 2022 / 2023 as said in news?

Theory **3**: If high property value spurs housing, the Income generated will be high.

Methodology

Description of the data sources used

Some of the things I got to know when looking for data sources related to the project are that the US housing data is centralized using Multiple Listing Service (MLS) for free. And there are multiple parties like Zillow and Redfin which act as aggregators for data. We also get to access census data using API's.

Zillow is a leading Real-Estate marketplace company based out of Seattle. They help customers with housing solutions by providing buyers and sellers' market. They have a dedicated research team that does all of the housing market analysis and provides end-users with consumable information like Home Value Index, Home Value Forecast, and Observed Rent Index. All of which are useful in determining the market trends.

The Economic Research data Realtor.com listings database offers an extensive inventory of for-sale and rental listings. The data provided by them is down to zip-code level.

Federal Reserve Economic Data, FRED is an online database consisting of hundreds of thousands of economic data time series from scores of national, international, public, and private sources.

Apify is the tool used for web scraping. There are developers that write code to integrate the data available from multiple websites and sources. The data can be obtained in different file formats by providing search texts as input in pre-configured format. I found a few data sets but could not integrate them with the existing ones.

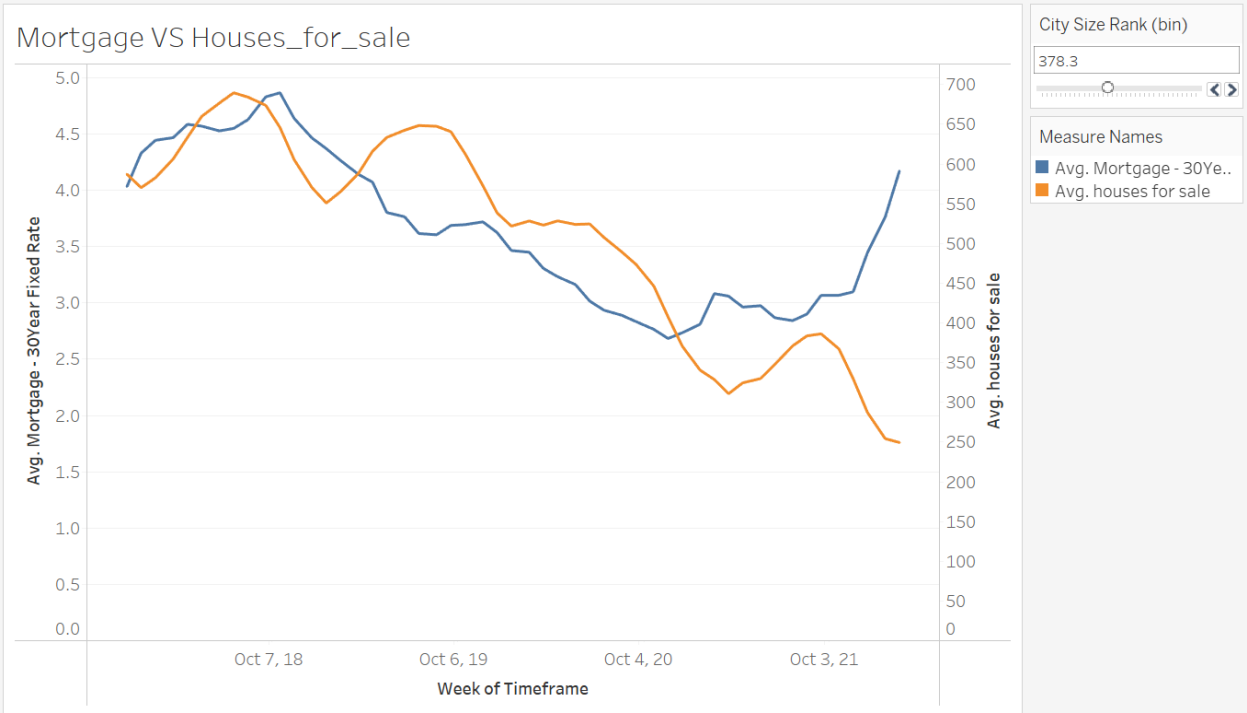
Analysis

A problem I faced with Prediction of Hotspots in the US data set, that was found on Zillow was noteworthy. The usual data sets I had been working with, were mostly structured with fewer columns and high row count. One of the datasets on Zillow had timeline as column headers (about 1000 columns) and states as row list. the data sets. Tableau would not accept over 700 columns. I had to pivot the data present in columns and as I tried to map the data, I ended up with a lot of data anomalies. Finally, I could tweak data and map it based on State Name, City Name and Date formatted in Month/Year format. This resulted in over 5.6 million rows.

One Hot Encoding is a method of converting data as input to an algorithm where we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. A data set that was available in Kaggle had the following encoding done as it was used to feed python and draw insights out of it. I had to reverse engineer the encoding and group the columns together as categorical variables.

Research Questions:

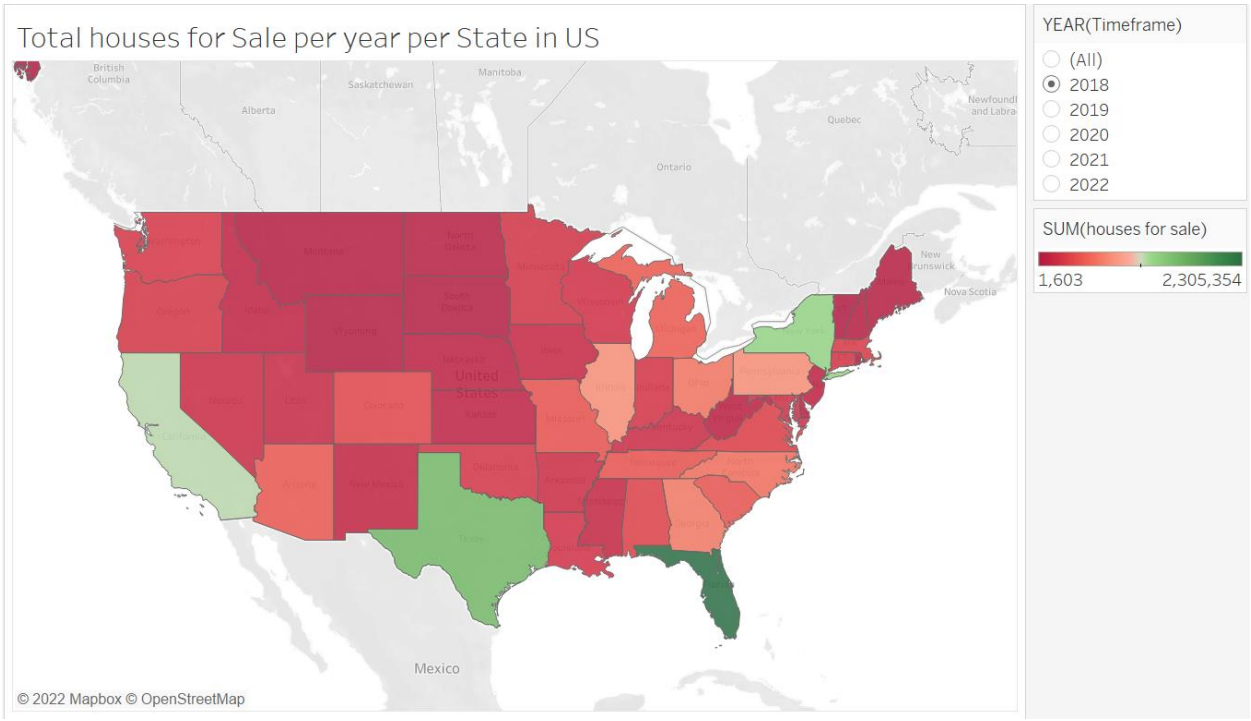
Theory 1:



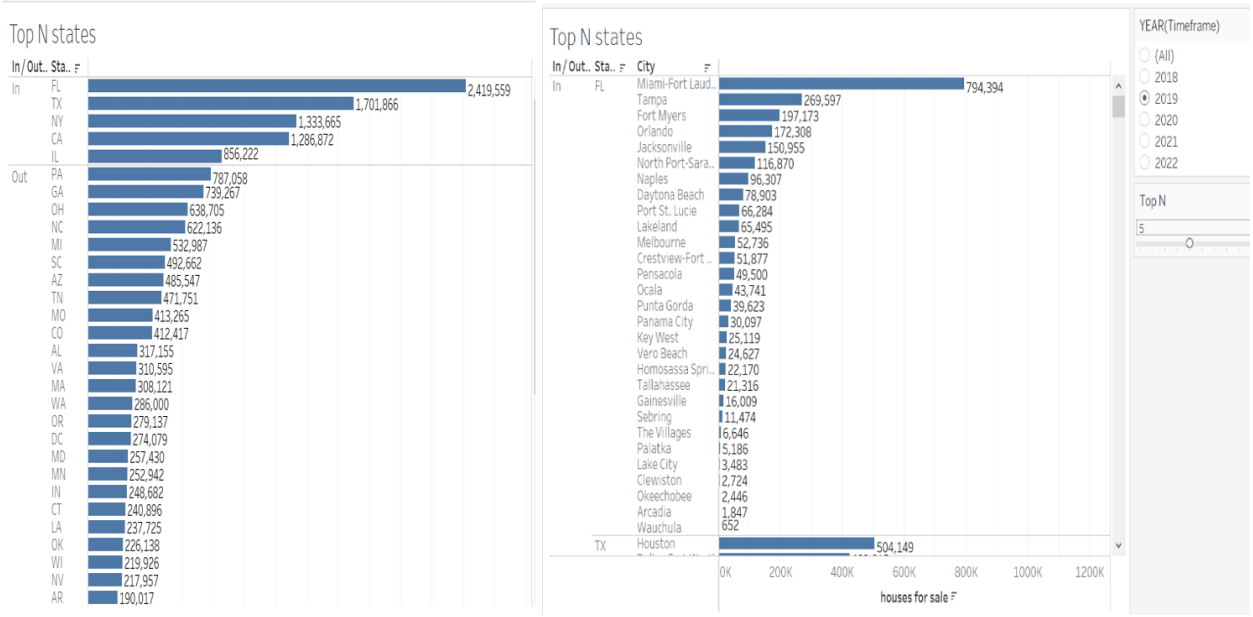
The visualization draws a trend with blue line on the average of Mortgage values for 30 Year Fixed rate over the timeframe January 2018 to present. The other line represents the average number of houses that were listed or made available for sale over the same time line. The data shown for house listing is cummulative over that particular month. The line graphs show a similar pattern over time showing us a correlation between the two(Which was covid in this case). In genaral, as there is a change in the mortgage rate, there has to be a proportionate change in the listings of house that are available each month.

We see a strange pattern in between the time period Apr 2019 to Dec 2019. As there was an inverse relationship. Another instance during which we see this similar economic trend is over the months Oct 2021 to present. The Mortgage values have started going up once again, but the houses that are up for listing are still declining. These trends could be an implication of work from home culture slowly disappearing as the covid cases have dropped and people are getting vaccinated.

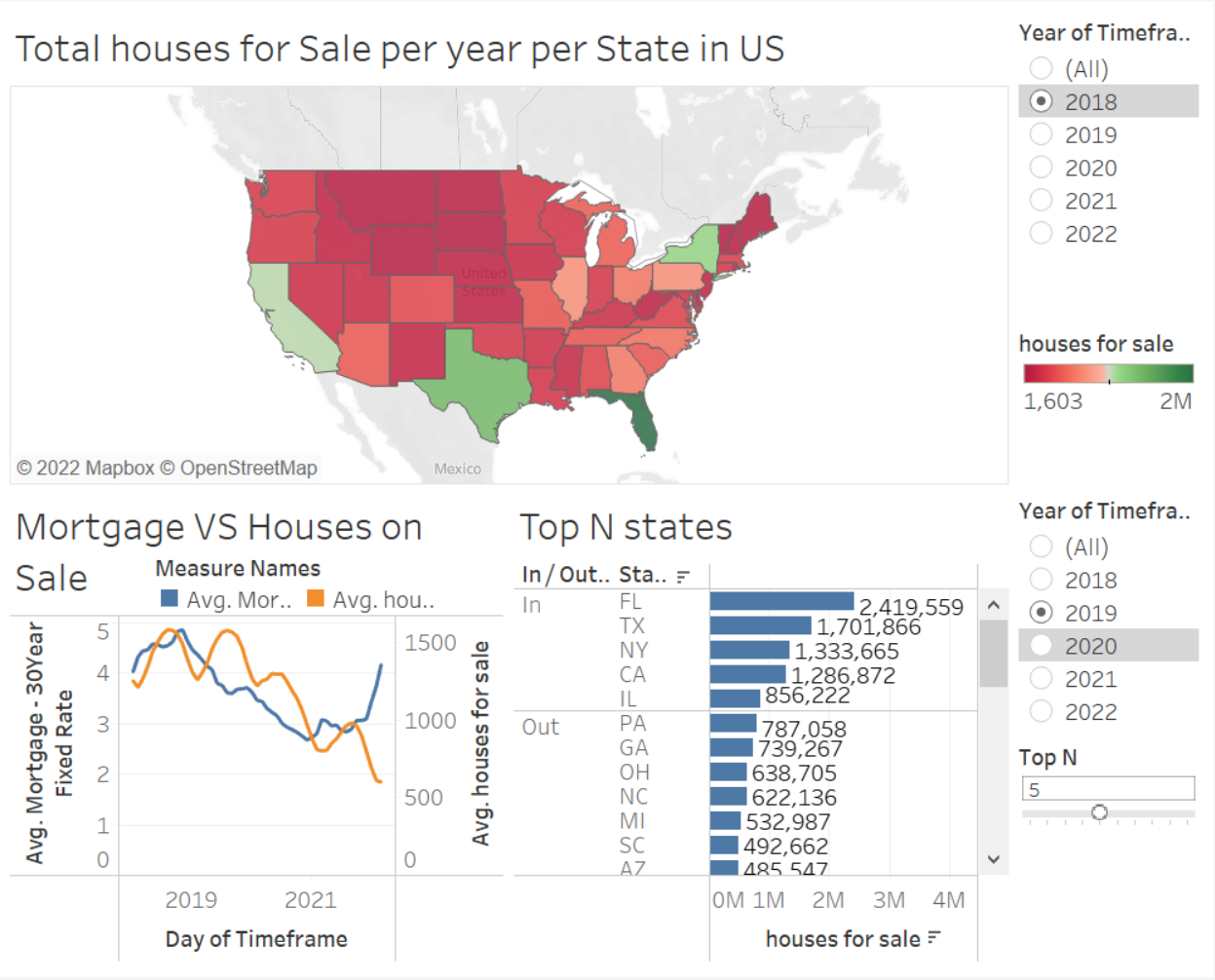
The graph has a bin set to aggregate cities that have similar city size rankings. An animation for the same is done over the monthly timeframe with show trails option enabled to compare the current values with historic trend.



The Map represents the total houses that are up for sale in US with states as the category. A filter card to show the change in aggregate total of houses for sale over the years has been added. The colors transition from red to green because of sale count is shown. We observe that the states that are to the south have a better housing market when compared to those in north, with New York state being the exception. Also the states that are in the center geographically do not seem to perform well in the housing market. A possible interpretation for this could be that people like states with beaches/bay area over others.

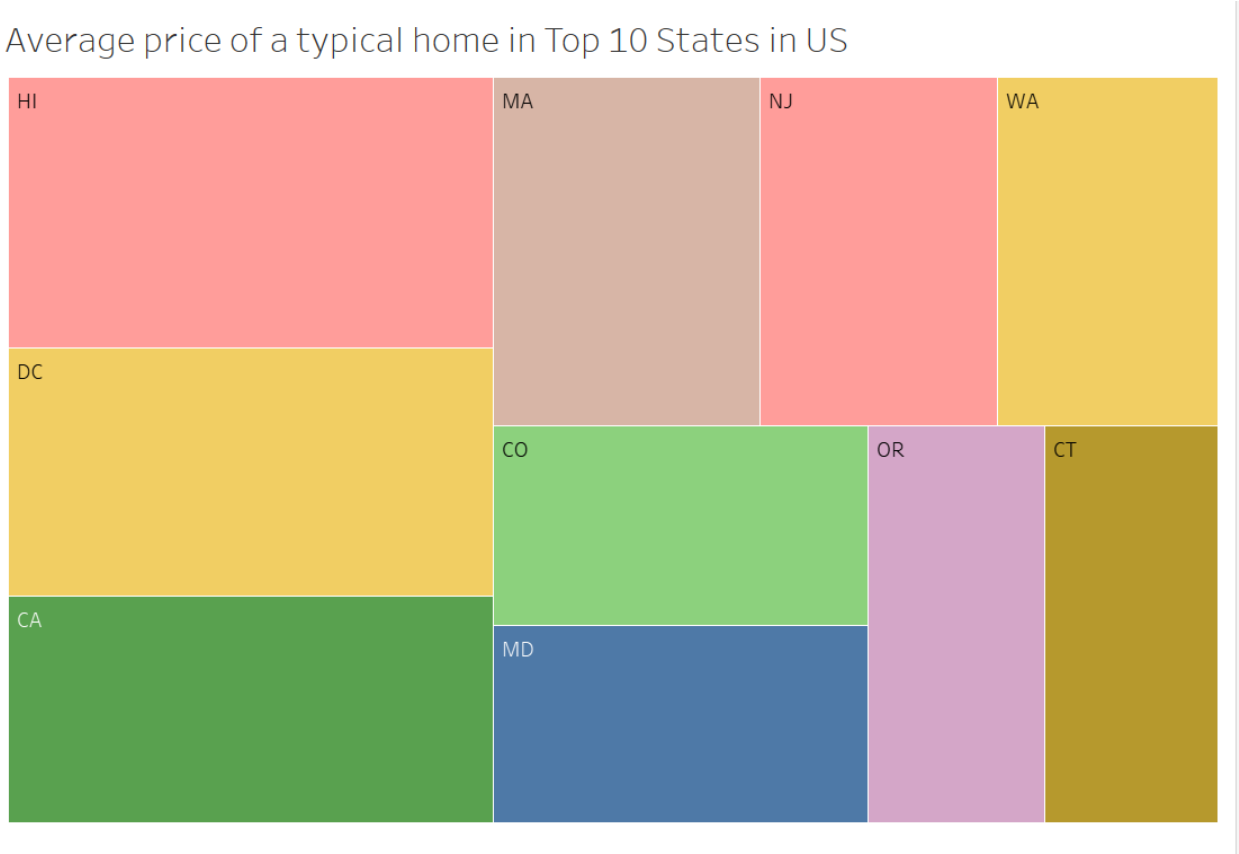


Top N is a tabular visualization with stacked bars representing the magnitude of the listings per state/city for sale that is available for that year. The graph has made use of hierarchy for state and city. I have added a Set to group values of a dimension(state) into two categories, namely in and out. And I have also added the Top N parameter by linking it with the Set created. This is to filter out the values in 'in & out' categories. It provides us with the flexibility to filter out the top N states with the greatest number of sales using a slider. This worksheet shows the year wise aggregation of count of sales. Since the order of operations in tableau says that set filters are performed before dimensional filters, we have added the dimensional filter- which is date(year) to context i.e., we have made use of context filter.

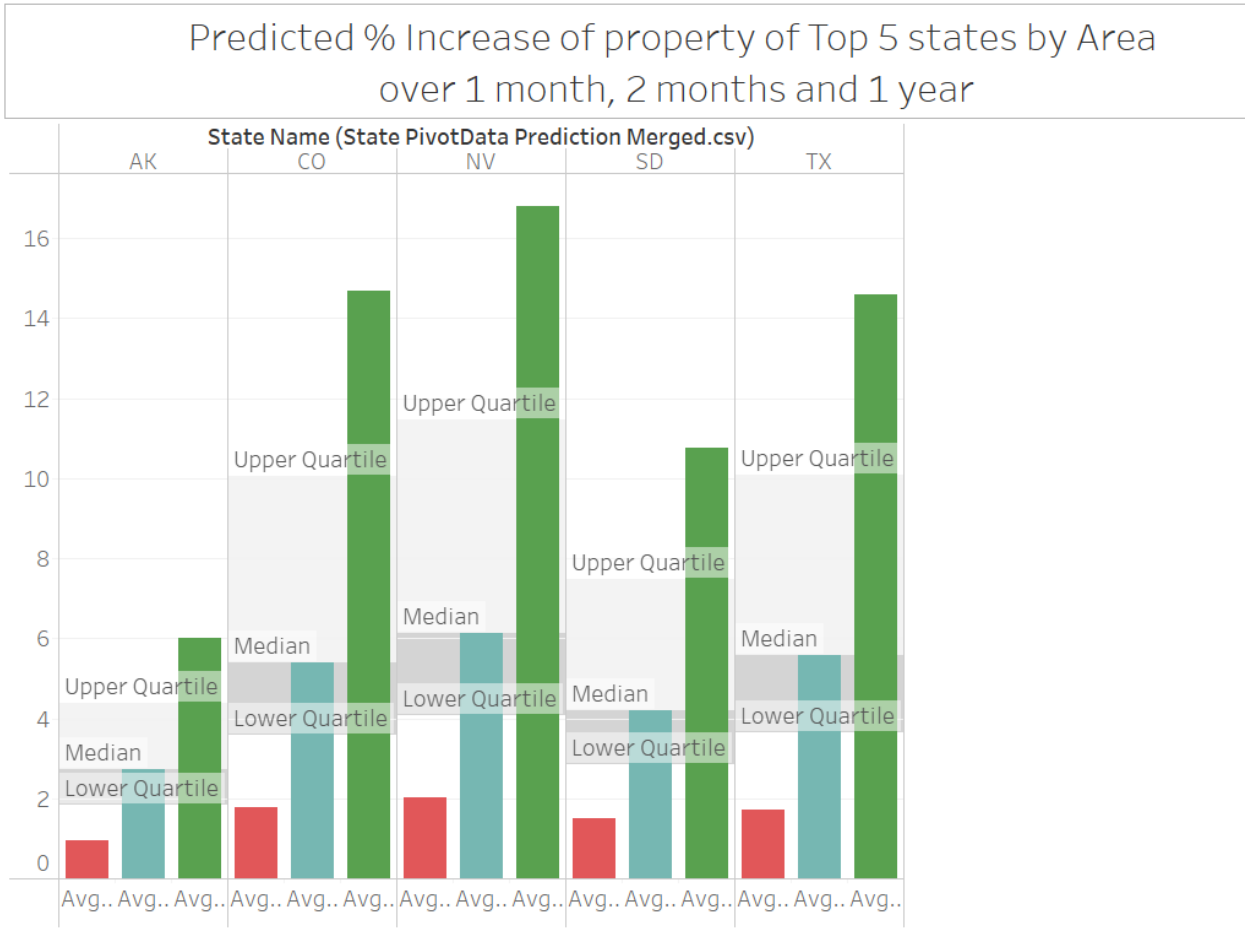


The dashboard provides us with all the information discussed above to helps us view all data at once, that is spread across various workbooks and draw meaningful insights out of it.

Theory 2:

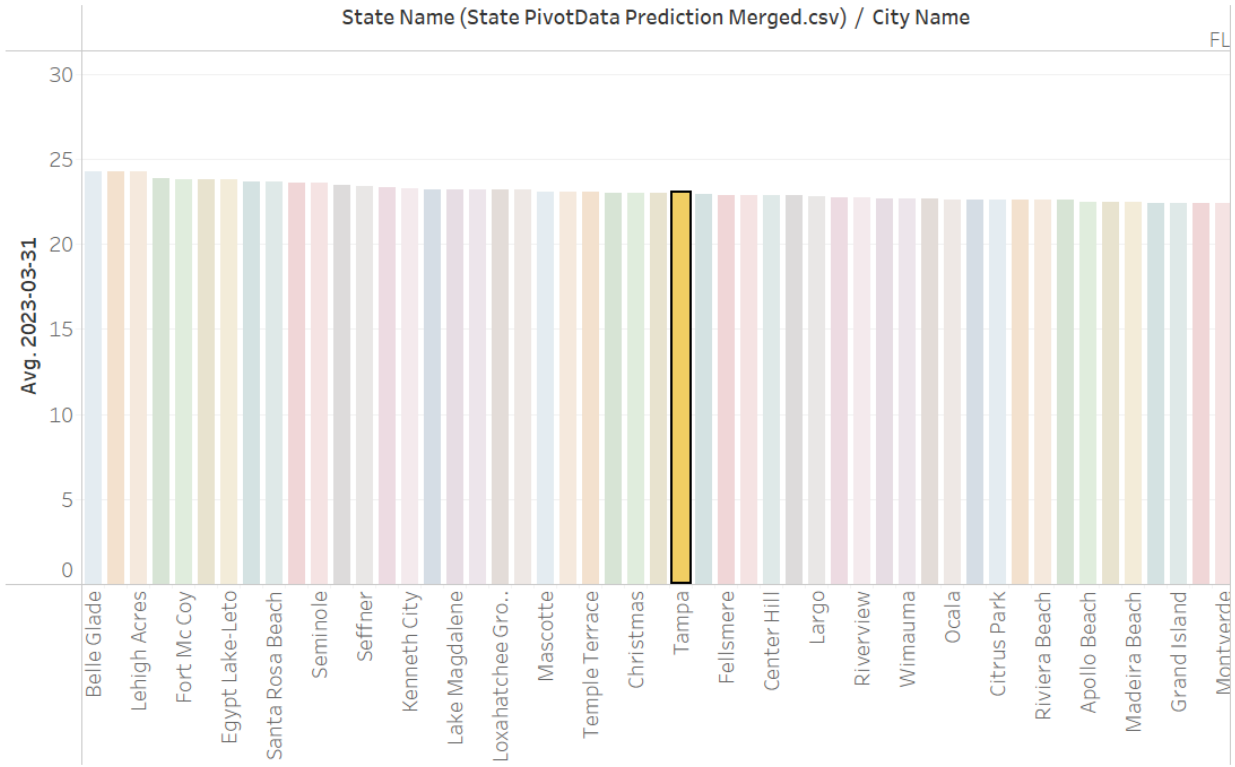


We are trying to draw insights on the Housing Market all over US by considering the typical price of a home over all states. We observe that Hawaii has the most expensive properties. Something to note is that Florida does not rank in the Top 10.

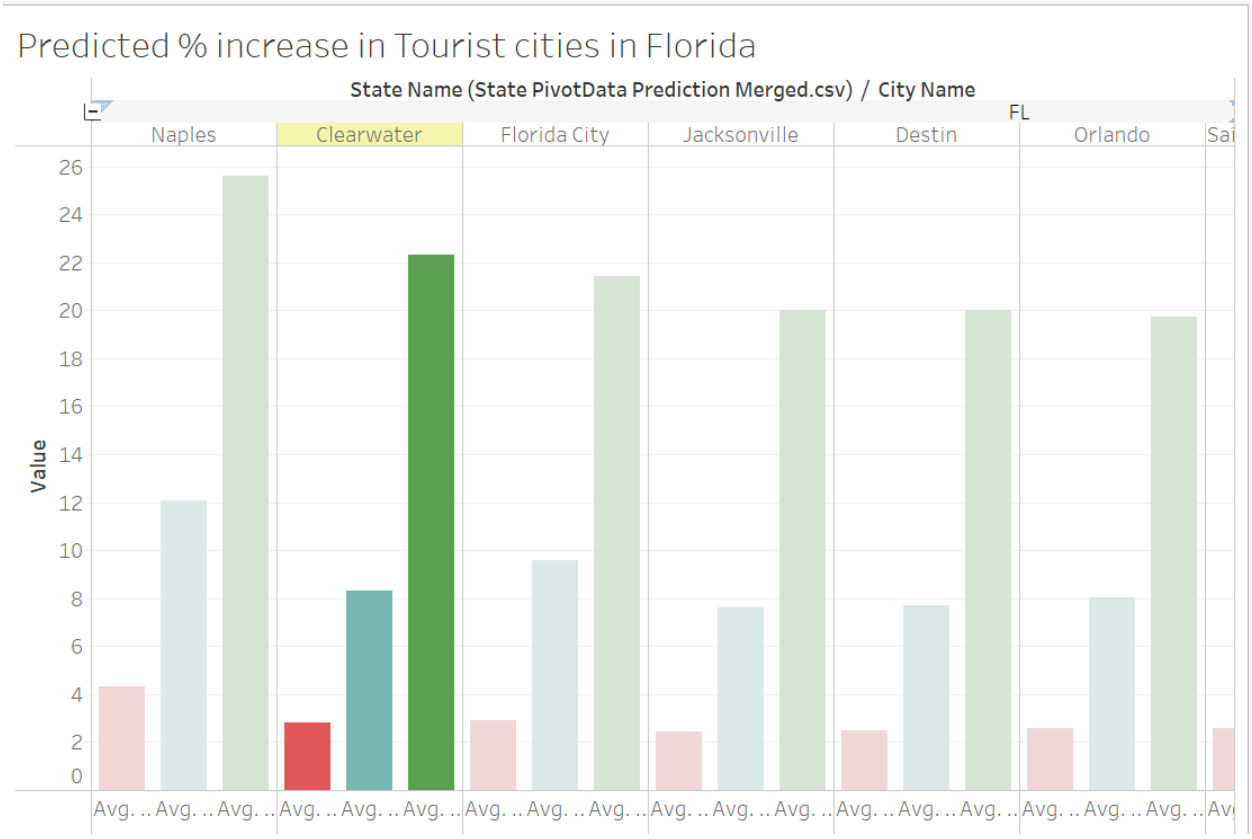


This graph draws the insights on prediction of increase in Top 5 states in US by Area. Observe that Nevada has the best percentage increase in property prices by area for both 2022 and 2023 predictions. The graph also estimates the Median and Quartile ranges on the percentage increase which is typically seen in a Boxplot. Observe that Florida does not fall under this category either.

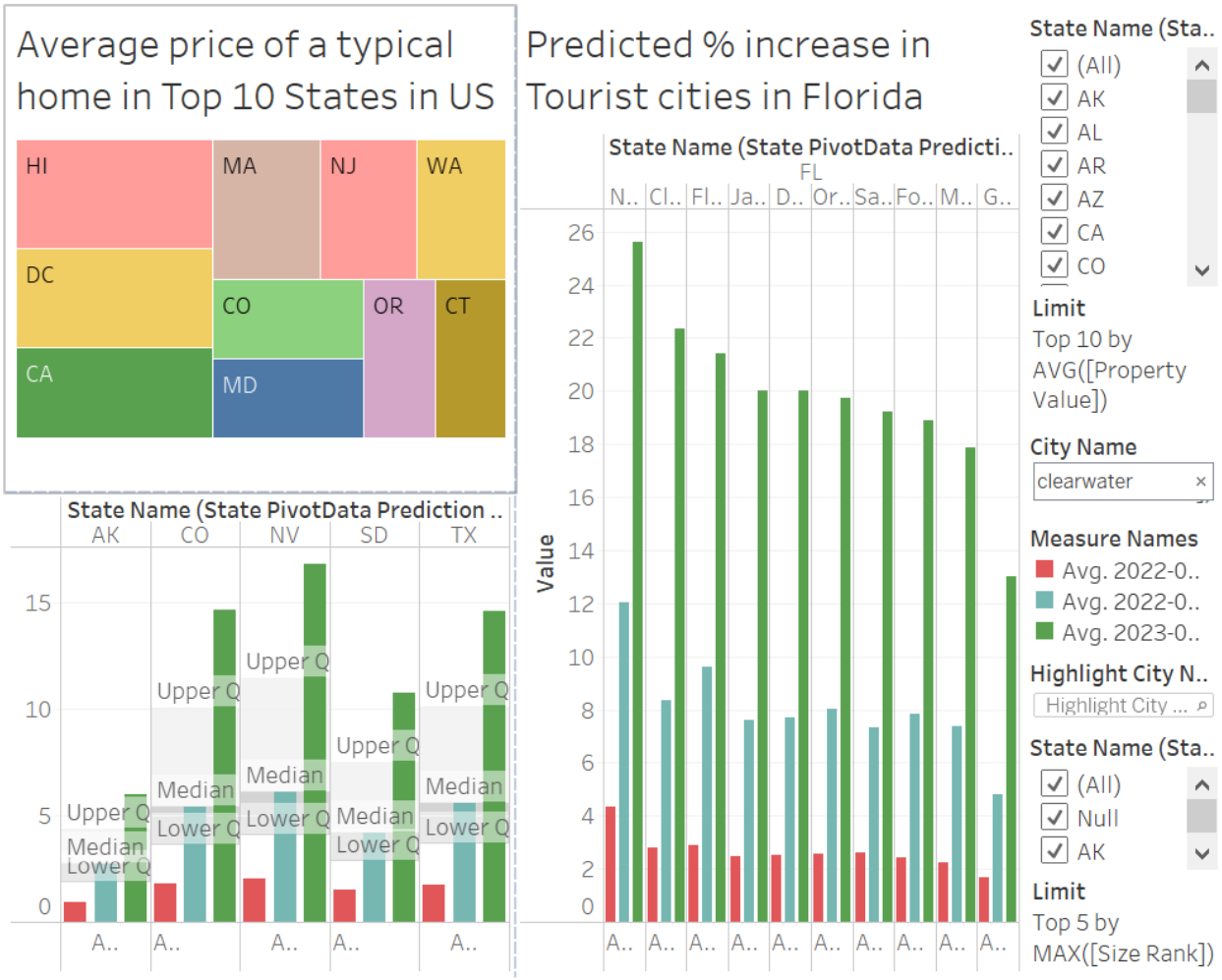
Predicted % increase within Florida over a year



We now consider all the cities and metropolis available around cities to draw a graph on the percentage increase in property prices with Florida state. Observe that Tampa is surpassed by various cities, but is still has close to the 22% increase. Holmes beach is predicted to rank first by having about 30% increase in the year 2023.



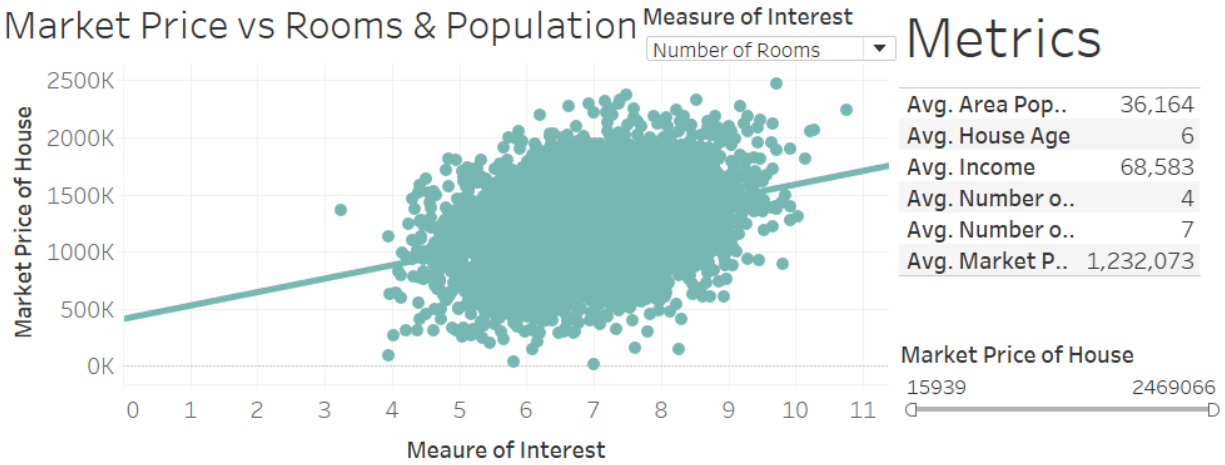
Now we try to further drill-down on the popular tourist cities and metropolis around cities in Florida. We observe that Naples ranks first, followed by Clearwater with about 22% increase in property price within a year. Notice that Clearwater is a location close to Tampa city.

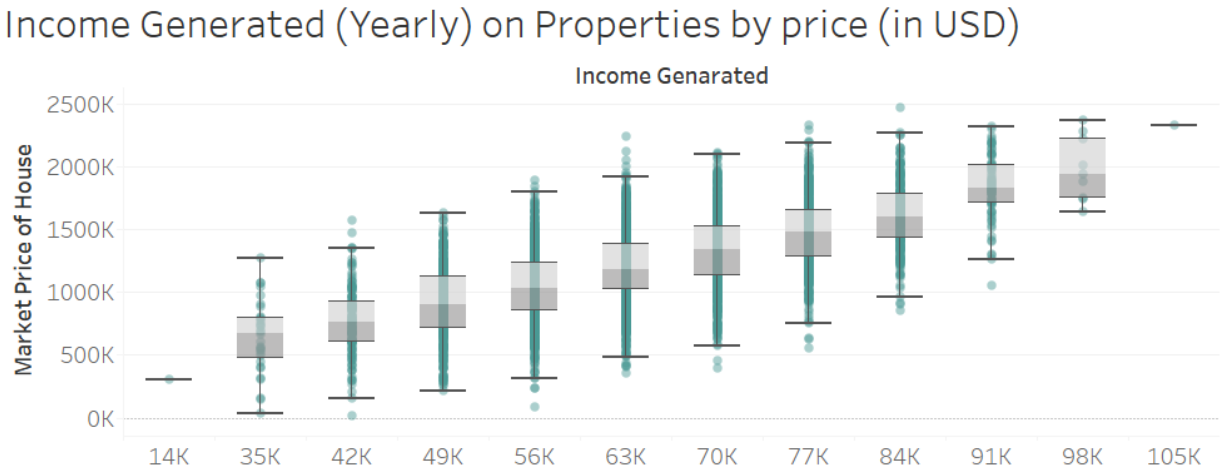


This Dashboard helps us answer the question on weather Tampa is the hottest market in US for 2022/2023. From the data and visualizations at hand, I could conclude that Florida/ Tampa might not have the properties of highest value in the US. But Clearwater, a city which is a part of Tampa bay area is definitely leading in the Housing market within Florida and has a good position within US too.

A limitation for the data set would be the consideration of a Typical number as price of property by considering typical value for homes in the 35th to 65th percentile range. Also, Something we must consider when answering the questions is the data we have at hand. Which is the On-Market prices. We have to take into account the deviation caused by Off-Market deals.

Theory 3:





The Tableau workbook1 was used to determine the relationship between the price of a property and income generated on that property. The text table shows the average values of some of the metrics available like the average area population and average age of the houses. The visualization shows information on income made by leasing out a wide range of (2-6.5) bed houses with about 2-10 rooms. We see a not so evident picture of number of rooms in a house being directly related to the price of the house. The appropriateness of the graph was checked by using regression analysis on the graph and result had a line with a p-value < 0.05. The low p-value indicates that the graph prediction is reliable. The dashboard generated is interactive with a drop-down to change the Measure of Interest value to Number of Bedrooms or average Area of Population.

The other graph we observe in the dashboard is the box and whisker plot. This box and whisker plot works represents a box around the first and third quantile of points and highlights the median. It also shows us data points that could be potential outliers for the data set. The plot here visualizes relationship stated between the price of the house and the income it generates over a year. We observe that there is a linear relationship and there is shift in median values. An observation drawn is, for a house to generate over 90,000$ a year, the house should be worth at least a Million dollars.

The appropriateness of this theory is that, about five thousand data points are considered for analysis after data cleaning. Certain rows with Null and Empty fields were dropped. The data is reliable as I did not need much processing of source data set and all fields have appropriate values. And the trends in the US Housing Market using the visualization did not show anything in contrast to logical assumption.

Conclusion for Theories

1. From the data at hand, we could conclude that it is a Buyer's market now, all over the US. Reason being the number of properties being listed for sale is going down.
2. The visualizations made on the price of a property all over US concludes that Florida does not rank among the top 10 in terms of property value. But Tampa Bay area certainly has cities which forms one of the Hottest market in Florida(US).
3. The data on Housing market concludes that there is proportionate revenue generated by a property based on value of the property.
   The question for research drawn out is, would the same property making low income currently, make more money, if additional investment was made?
   The plan of action would be to get information on the list of housings with their market price and leasing price before and after renovation. And drawing metrics on, if that additional renovation cost was worth it?

Conclusion

By having the right kind of data, we can reap the benefits of Data Driven decision making. Instead of simply having to look at a house before making an investment on it, we should consider the factors that influence us on making the decision in that direction. Is the house cash flowing? Is it a good deal based on the parameters like number of bedrooms, bathrooms, tiles, and square footage? By using data to make decisions, we not only have the power to make decisions relative to the market. But it also helps us reduce costs, increase speed, spending on continuous improvement, collaborative decisions and ultimately planning.

Identification of Research questions

Theory: If we can find sellers before they put their properties on the Multiple Listing Services (MLS), we can buy properties at a discount.

One of the ideas for solving this would be, using methodologies to find quality seller leads.

A useful website found to start with the idea is, ListSource. It is a listing service of Homeowners that want to sell their homes off market. Typically, data listed here is also on the internet but does not grab as much attention and popularity as it should. And It has data available that can be filtered out at multiple levels based on search criteria.

Some more research questions I would like to explore would be:

- How does pollution affect the price of real estate?
- Is there any relationship between the proportion of lower financial class and price drop in real estate?

Resource

Kaggle: https://www.kaggle.com/datasets/aariyan101/usa-housingcsv

Zillow Economics: https://www.zillow.com/research/data/

Realtor Data: https://www.realtor.com/research/data

FRED Economic Data: https://fred.stlouisfed.org/series/MORTGAGE30US

ListSource: https://www.listsource.com/list/ui/search.htm?search=searchAdvanced

PropStream: //paid service

Web Scraping: Using Apify on Zillow, County Page, Facebook and Craigs List

YouTube for Project questions on Housing Market analysis.