Property Investment Case Study

Jingyi Ran

1. Business Understanding:

We are interested in targeting the best areas for property investment by applying data science. There are two business requirements:

1. build a model to target the top 10 zip codes for property investment
2. explore the characteristics of close zip codes to figure out how geographically close zip codes impact each other

2. Data Understanding:

The dataset includes demographic information for each zip code from 2012 to 2017, by doing exploratory data analysis, we can summarize the variables into 6 categories:

1. Population (Total population, avg household size, etc.)
2. Age
3. Rent (Gross rent, Gross rent as % household income, etc.)
4. Housing units (Total housing units, housing units with the mortgage, etc.)
5. Occupation (Number of tech workers, Number of service workers, etc.)
6. Home Value

3. Business Metrics Selection

To target top zip codes for property investment, firstly we need to define an evaluation metric for property investment. There are many possible metrics, and we choose Median Estimated Home Value as the metric to evaluate if it's worth investing in the properties in the zip code area.

Selected Metric

```
data.head()
```

	id	Tract_number	zip_code	Year	state	Unweighted Sample Count of the population	Avg household size of occupied housing units	Total population in occupied housing units	Median Estimated Home Value(owner occupied units)	Total Population	...	Renter households	Housing units with mortgage	Total housing Units	Number of Sales and office workers	r
0	1	1001020100	36067.0	2012	Alabama	195.0	2	1764	121500.0	1812	...	203	357	724	289	
1	2	1001020200	36067.0	2012	Alabama	222.0	2	2074	130500.0	2218	...	722	311	893	180	
2	3	1001020300	36067.0	2012	Alabama	261.0	2	3103	118700.0	3155	...	643	551	1361	334	
3	4	1001020400	36067.0	2012	Alabama	764.0	2	4329	133500.0	4337	...	915	848	1851	547	
4	5	1001020500	36066.0	2012	Alabama	540.0	2	10431	174500.0	10498	...	3499	1815	4114	1409	

There are other possible metrics, but because of the time limit, I didn't choose other metrics or find the trade-off among different metrics, however, I'd like to go over my thoughts and show why I didn't choose those metrics.

1. Population: although in general, the larger the population, the more prosperous the development of the region, the population cannot be directly linked to the value of the property. The slum is a good example. It's risky to use population or population increase rate to evaluate if it's worth investing in the properties
2. Gross Rent: after comparison, I think the home value would be better because renters are only a small proportion of the total population, so revenue from renting the house is not closely relevant to how worthy it is to invest the properties in this area

3. Median Estimated Home Value ×Total housing units: in some regions, the total housing units might be very high, but the median estimated home value is pretty low, and slum is an example. The median estimated home value is better than the "total" estimated home value in this scenario.

Because there is no variable related to the cost of investment and only a few variables related to revenue, it's hard to calculate the profit and Return on Investment rate. If we have relevant data, I think profit (Revenue-cost) and Return on Investment would be better metrics than the median estimated home value.

4. Approach Scope

My approach is to forecast the Median Estimated Home Value of each zip code in the next 10 years based on the data from 2012 to 2017 and further calculate the compound annual growth rate (CAGR)

$$\mathrm{CAGR} = \left(\frac{V_{\text{final}}}{V_{\text{begin}}} \right)^{1/t} - 1$$

V final is the predicted median home value at the final year (2027)
V begin is the predicted median home value at the first year (2018)
t=10 because we are forecasting the home value in next 10 years

After implementing the model, we can get the CAGR of each zip code. The top 10 zip codes with the highest CAGR are the target zip codes.

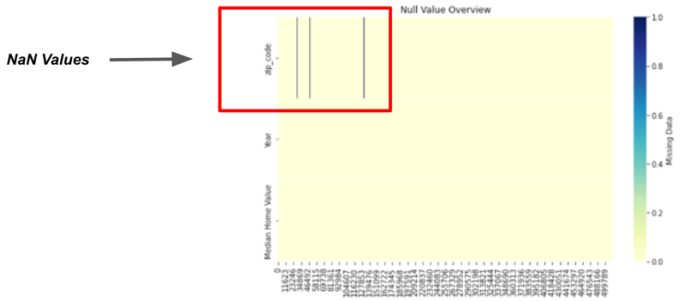
5. Data Cleansing

a. Select relevant variables:

According to the business understanding and approach scope, I choose 'zip code', 'state', 'year' and 'Median Estimated Home Value' as variables for modeling

b. Deal with NaN Values:

After a quick check, we found there are 511392 rows in the dataset and in the "zip code" columns, there are 3035 rows missing and in the "Median Home Value" column, there are 345 rows missing



For "zip_code" column, since we cannot do any analysis without zip code information, we choose to delete rows with missing zip codes
For "Median Home Value" column, we choose to impute the NaN values with median values at the same zip code and in the same year.

c. ***Deal with outliers:***
We found there are 9628 rows with zero median home value (Median Home Value=0), but this value cannot be 0. Considering 9628 out of 511392 is still a small proportion, we choose to delete those 9628 rows

d. ***Exclude zip codes with missing years***
The goal of the model is to forecast the median estimated home value of each zip code in the next 10 years based on the data from 2012 to 2017 (6 years), but if we don't have all six years' data points for a zip code (for example, we only have in 2012 but not from 2013-2017), the model would lose some predictive power. There are 21953 zip codes in the dataset after deleting NaN values and outliers, and we found 102 zip codes don't have all 6 years' data points, since 102 out of 21953 is a very small proportion, we exclude those 102 zip codes from the dataset.

6. Time Series Forecasting

We choose to apply the ARIMA time series forecasting algorithm to forecast the median home value of each zip code in the next 10 years starting from 2018 and further calculate the compound annual growth rate (CAGR)
Each zip code has one time series. For a specific zip code, we use the median estimated home value from 2012 to 2017 (6 data points) to predict the median estimated home value from 2018 to 2027 (10 data points). The logic of ARIMA in this scenario is using data from 2012 to 2017 to predict the value in 2018 and using data from 2013 to 2018 to predict the value in 2017, etc. (Autoregressive Integrated Moving Average principle)
Finally, we have 21912 time series because we have 21912 different zip codes after data cleansing.
After generating the forecasting median home value for the next 10 years, we calculate the CAGR in the next ten years of each zip code and get the top 10 zip codes with the highest CAGR.

7. Model Evaluation

Evaluation Metric: Mean Absolute Error
Since time series forecasting is a special type of regression problem, we need to choose metrics for evaluating the regression model, so I choose Mean Absolute Error

For this model, the average of the mean absolute error of all time series is 8783 (MAE=8783), since the average of the 'Median Home Value" column is 230886, and 8783 is much smaller than the mean value of the target variable, so we can assume the model has a great performance.

8. Business Suggestion

The top 10 zip codes to target: (top 10 zip codes with highest CAGR in the next 10 years)

	zip_code	state
0	92115.0	California
1	32772.0	Florida
2	70375.0	Louisiana
3	68874.0	Nebraska
4	58830.0	North Dakota
5	75751.0	Texas
6	77003.0	Texas
7	76701.0	Texas
8	77358.0	Texas
9	79605.0	Texas

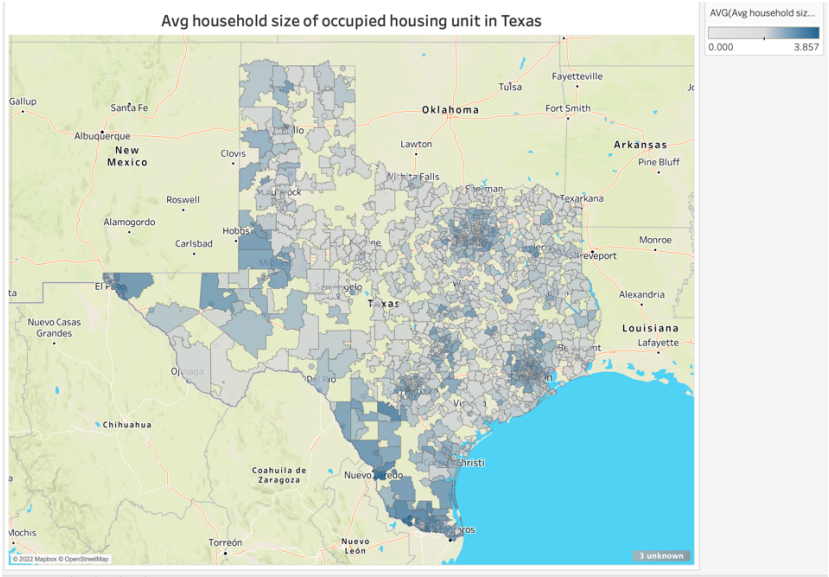
In addition, we can see among those 10 zip codes, 5 zip codes are in Texas. Maybe Texas is one of the best states to invest in properties

9. Data Visualization

(to answer "How do geographically close zip codes impact each other?")
we predicted the top 10 zip codes for property investment and we found 5 zip codes among those 10 are in Texas. So, we choose to analyze the demographic characteristics of close zip codes in Texas using Tableau.

There are two findings:

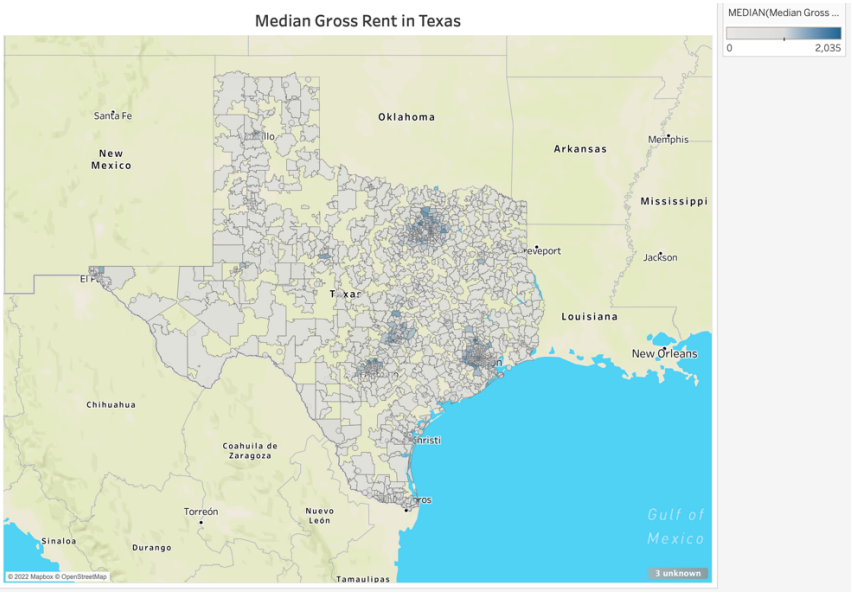
a. Close zip codes tend to have similar average household size of occupied housing unit in Texas



The darker the color, the higher the average household size of occupied housing unit. Similar colors represent similarities in average household size.

From the visualization, we can see close zip codes tend to have similar average household sizes. In addition, families in the south of Texas, the northwest of Texas, and downtown Texas tend to have more households in a housing unit compared with other regions in Texas.

b. Close zip codes may have similar rent.



The darker the color, the higher the gross rent. Similar colors represent similar gross rent.

From the visualization, we can clear clusters with darker colors, which means the rent in those regions is higher than rent in other regions.

In addition, the development of one area might boost the development of its neighbors or impede the development of its neighbors. For future work, we can choose metrics as a proxy for “development” and create visualizations to see if there are relationships between demographic similarity and economic development.

10. Future Work

a. **Dealing with outliers:** Because of time limit, we only consider the case of "zero median home values". For future work, we need to do more work in identifying and dealing with outliers because outliers have non-negligible negative impact on the predicting results. Common approaches include visualize data using histograms and box plots and then dealing with highly skewed distribution and excluding values outside 3 standard deviation from the mean.

b. **Exploratory Data Analysis:** Usually, before modeling, we will do some exploratory data analysis to know the data better. But because of the time limit and considering

the EDA doesn't help a lot for building the forecasting model especially we have decided to use ARIMA and we have a clear goal for model building, we choose to make it "Future Work".

Lots of analysis can be done in the EDA part. For example, we can have map visualization to see in which states or in which regions, the properties have relatively higher value and higher rent prices. In addition, it's also interesting to pick some regions (eg: New York and California) and visualize the trend of housing prices from 2012 to 2017. Also, age is another factor that we could explore, maybe we can make sorted bar plots to compare the age of the population in different regions, from which we can combine the characteristics of different age groups and the property information to find investment opportunities.

c. **Model Improvement:** Firstly, we need to tune the hyperparameters of the ARIMA model. In this case study, I just use the default parameters to do the forecasting, which is risky. In addition, the model training and model evaluation should be an iterative process.

11. Assumptions:

a. Ignore the impact of pandemic and other issues like financial crisis
b. One thing that I'm unclear is that there are multiple rows for the same zip code at the same year. Since my goal is to forecast the median home value of each zip code, I choose to aggregate the data. The outcome of aggregating data is one record for each zip code at each year