Building a REIT Portfolio in Austin, Texas

Capstone Project for IBM, Applied Data Science Capstone

By Miguel Villanueva

I. Introduction

A. Overall Objective: The primary goal of the capstone project is to create a REIT portfolio compromising of real estate properties in Austin, Texas, USA.

B. What is an REIT?

1. An REIT or a **R**eal **E**state **I**nvest **T**rust is a company that manages investors' capital and invests it in income generating real estate properties. Income-generating meaning properties that have rent or mortgage such as shopping malls, groceries, hospitals, etc. A counter example would be non-income generating properties such public parks, public trails, public beaches, etc. An REIT can own and manage multiple property types and this collection of properties is what is called a portfolio, specifically in this case, a real estate investment portfolio.
2. Additionally, in order for a company to be considered an REIT by law, it has to return 90% of its net earnings back to the investors. This means that the total collection of rent and mortgage from properties, minus the cost of upkeep, is given back to the investors. This return is called the dividend.
3. The beauty of an REIT and why it was chosen as the focus of this capstone project, is that it has the option to be publicly-traded. Contrasting privately held REITs, publicly traded REITs are available to anyone who can afford the market price of the company's share. You don't have to be part of the billionaire or multi-millionaire club to participate.

C. Why build an REIT portfolio in Austin, Texas, USA?

1. The reason behind the interest in Austin real estate properties is due to the current trend of movement of technology companies from Silicon Valley in California to Austin. Examples of these large companies providing gainful employment are IBM, AMD, Intel, Tesla, Oracle and the like. Even F.A.A.N.G., the largest publicly-traded technology companies (with the exception of Netflix) is present in Austin—namely Facebook, Apple, Amazon, and Alphabet (parent company of Google). This has led many to dub Austin, Texas as "Silicon Hills" a play-on-words on its predecessor Silicon Valley.
2. These technology companies provide a good source of gainful employment and consequently a movement of labor and capital investment means a rise in demand for

goods in services in the area. Two of these demands are the rise in the desire for property ownership and the desire of business ownership. Speaking plainly, people want to own or rent real estate near their place of work (e.g. buying a home or renting an apartment), and people want to build businesses near strong population (customer) locations (e.g. opening a restaurant or a coffee shop). These are the two scenarios where an REIT can greatly benefit from.

D. Business Problem
    1. The business problem that this capstone project aims to solve can be subdivided into two questions. First, "Where are the neighborhoods that will benefit the most from the movement of Big Tech?" Second, "What are the types of properties within these neighborhoods?"
    2. To answer the first question, geospatial data will be visualized to find out which neighborhoods are in close proximity to high density clusters of these Big Tech companies. Furthermore, population size and growth will be accounted for to find places with the best chance of having renters or mortgage borrowers.
    3. To answer the second question, venues and venue category of selected neighborhoods will be acquired to find property types and frequency of venue category.
    4. Additionally, K-Means clustering will be used to find if there are any underlying or hidden trends between these neighborhoods with respect to venue categories as features.

E. Interested Parties
    1. Companies interested in building a REIT portfolio in Austin, TX.
    2. Companies who already have a REIT portfolio in Austin but wish to switch focus to properties accommodating the technology industry boom.
    3. Individual investors who are curious about property types surrounding these locations.

II. Data

A. Neighborhood data in Austin was obtained from the official data collection Austin governance. Specifically from the building and development webpage of data.austintexas.gov, the updated version on April 15, 2021. The direct link is: https://data.austintexas.gov/Building-and-Development/Neighborhoods/a7ap-j2yt

B. Big Tech data or the list of Big Tech companies was web-scrapped from the Wikipedia page of Silicon Hills. The specific website is: https://en.wikipedia.org/wiki/Silicon_Hills

C. Population size and population growth were taken from decennial data (10 year data) from 2000 and 2010. Specifically 10 years was chosen, since yearly trends might be too small to give a good predictive trend. The data file was provided by Kaggle, but was performed by the U.S. Census Bureau. The direct link: https://www.kaggle.com/census/us-population-by-zip-code

D. The geolocation (latitude and longitude) of both neighborhoods and Big Tech locations were found using the OSM (Open Street Map) database, since usage of it was free and allowed multiple queries (compared to Google API). The reference link: https://www.openstreetmap.org/#map=4/38.01/-95.84

E. Venue data and venue description (importantly venue category) about certain neighborhoods were found using Foursquare API. Venue category was used in K-means clustering as features. The direct link: https://developer.foursquare.com/

III. Methodology

A. Software Used:

1. Jupyter Notebook version 6.0.3 via anaconda navigator was used with Python 3.7.
2. The pandas package version 1.2.4 was used for data frame handling.
3. The geopy package version 2.1.0 with Nominatim was used to extract geolocation data.
4. The folium package version 0.12.1and matplotlib version 3.4.1 was used for mapping and plotting respectively.
5. The scikit-learn package version 0.24.1 was used for machine learning, specifically K-Means clustering.

B. Quick Overview of Flow of the Data Process/ Chronology of Steps Taken:

1. First, the neighborhood and Big Tech data was extracted and from the csv files. Then their geolocations (latitude, longitude, and zip code) were extracted from the OSM database.
2. Second, the neighborhood and Big Tech locations were then mapped with folium to visually search for significant overlaps of the two. Zip codes of neighborhoods with strong proximity to large clusters of Big Tech were chosen.

3. Third, population size and population growth were extracted from decennial data in 2000 and 2010. Their population size and population growth were summed by the zip codes present in Austin, TX. It was then plotted for visual inspection. The zip codes that were the best candidates were the ones that had good standing in both population size and population growth. These zip codes were chosen.
4. Fourth, from the gathered zip codes of those in close proximity to Big Tech, those with a good standing in population size and population growth, their respective neighborhoods were extracted from the previously cleaned neighborhood data.
5. Fifth, using Foursquare API, the venues of these locations were found, extracting the top 100 venues, and then turned numerical variable (via one-hot encoding) to be processed.
6. Sixth, using the one-hot encoded venue data, K-Means clustering was performed to find possible hidden trends between the neighborhoods with respect to venue categories as features.
7. Seventh, using the one-hot encoded venue data, frequency of the most common types of venue categories (the real estate types that will be invested in) was found.

C. 1st: Neighborhood Data and Big Tech Data Cleaning
1. The neighborhood names were extracted from official data provided by the Austin, TX database.
2. A *for loop* was used to go over the neighborhood names and associate latitude and longitude data from OSM database.
3. Using *regex* the zip code associated in the street address was extracted.
4. The Big tech data was web-scrapped from the "Silicon Hills" Wikipedia page. Unfortunately there was no collective directory for these companies, so their addresses had to be found manually.
5. A *for loop* was used to go over the Big Tech names and associate latitude and longitude data from OSM database.

D. 2nd: Mapping of the Overlap of Neighborhood and Big Tech Data Cleaning
1. Folium was used to map the neighborhood data and big tech data.
2. An overlap of the two were created and the zip codes of the neighborhoods with close proximity to high-density Big Tech clusters were selected.

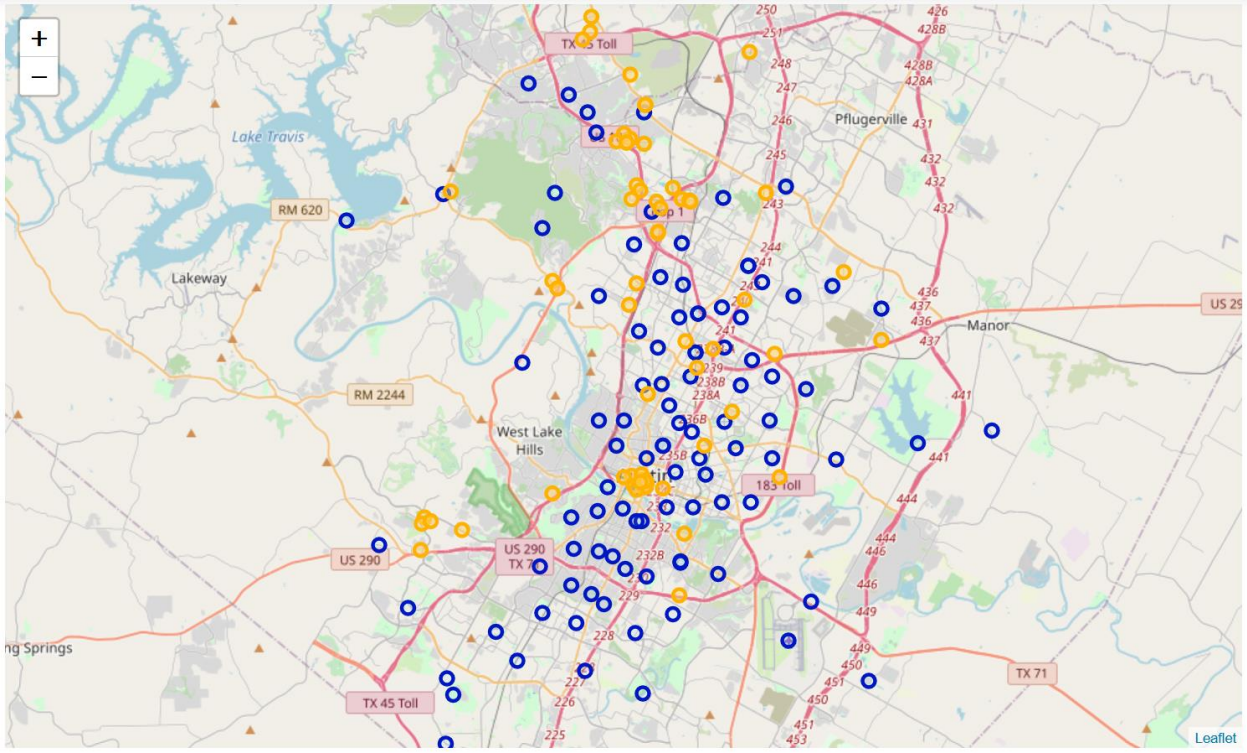

Fig 1. “Overlap of Neighborhoods Location and Big Tech Locations”. The orange circles represent the Big Tech companies, and the blue circles represent the neighborhood location centered on geolocation provided by the OSM database.

E. 3rd: Population Size and Population Growth Cleaning/Plotting

1. The population count for 2000 and 2010 were split into categorical grouping, specifically by age group and gender.
2. A list of unique zip codes in Austin was created from the cleaned neighborhood data, to use as a filter in the population data. The interest was only for zip codes in the Austin, TX area.
3. A *for loop* was used to sum populations by their respective zip codes. The difference of the 2010 population and 2000 population was used as a measure for growth.
4. Matplotlib was used to plot a bar graph of population size and a horizontal bar graph was used to plot population change between 2000 and 2010 numbers.
5. Zip codes that had a good standing, meaning both a high population size and a strong positive population growth, were selected.

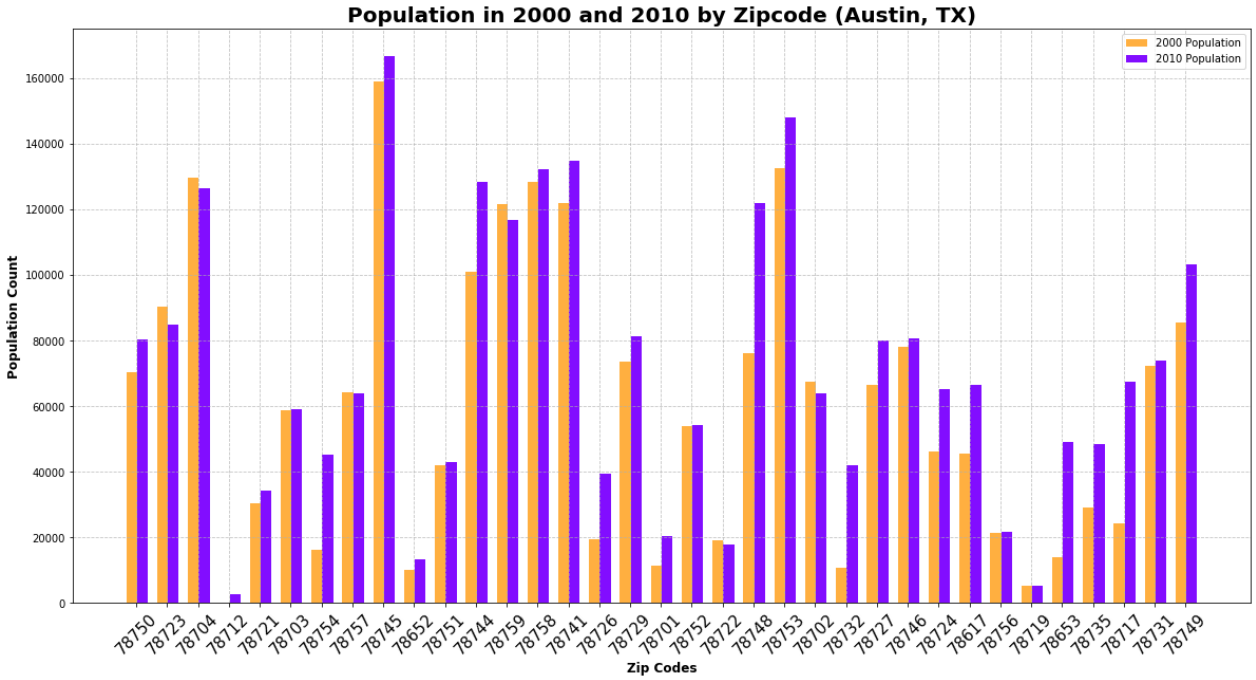

Fig 2. “Population in 2000 and 2010 by Zip Code”. The x-axis represents the zip codes and the y-axis represents the population count. The orange bars are for the year 2000 and the purple bars are for the year 2010.

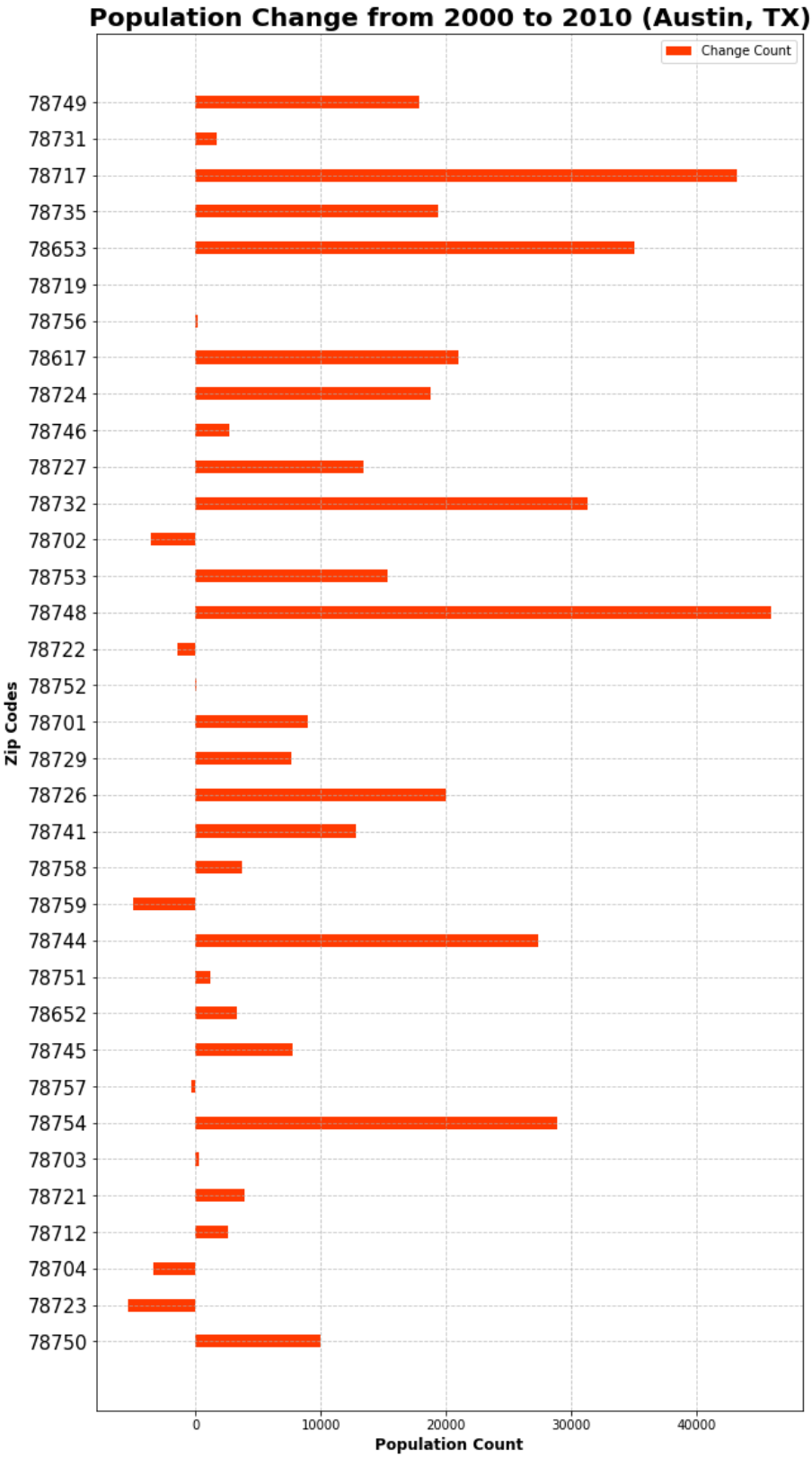

Fig 3. “Population Change from 2000 to 2010”. The x-axis represents population count and the y-axis represents the zip codes.

F. 4th: Neighborhoods in Selected Zip Codes

1. The selected zip codes fulfilling gathered were used to filter the neighborhood data and retrieve their corresponding neighborhoods.

	Neighborhood	Latitude	Longitude
0	Dittmar Crossing, Austin, TX, USA	30.183729	-97.813116
1	CHERRY CREEK, Austin, TX, USA	30.197113	-97.824812
2	SOUTH BRODIE, Austin, TX, USA	30.175622	-97.851637
3	SLAUGHTER CREEK, Austin, TX, USA	30.167554	-97.848329
4	MCKINNEY, Austin, TX, USA	30.205876	-97.728015
5	BLUFF SPRINGS, Austin, TX, USA	30.178720	-97.775929
6	ONION CREEK, Austin, TX, USA	30.168117	-97.744817
7	FRANKLIN PARK, Austin, TX, USA	30.196898	-97.748800
8	UT, Austin, TX, USA	30.279308	-97.742845
9	DOWNTOWN, Austin, TX, USA	30.268054	-97.744764
10	WEST AUSTIN NG, Austin, TX, USA	30.265587	-97.746996
11	EAST CESAR CHAVEZ, Austin, TX, USA	30.255896	-97.731707
12	WESTOVER HILLS, Austin, TX, USA	30.379933	-97.749599
13	GATEWAY, Austin, TX, USA	30.395185	-97.739643
14	SPICEWOOD, Austin, TX, USA	30.432946	-97.770378

Fig 4. “Neighborhoods of Interest”. The location of where to build the REIT portfolio from.

G. 5th: Top 100 Venues in these Neighborhoods

1. Foursquare API was used to extract the top 100 venues in these neighborhoods.
2. Venue categories, specifically were used as features in machine learning.
3. A data frame was created with the top 10 venue categories in each neighborhood and one-hot encoding was used so it can be processed in machine learning.

H. 6th: K-Means Clustering and the Elbow Method

1. The one-hot encoded data was normalized, and then the elbow method (an SSE plot or a sum of squares error plot) was used to find the best K.
2. The elbow method used was that of distortion and inertia comparisons to k-value or number of cluster. Euclidean distance was used for distortion and the inertia was found by extracting it from its attributes. Distortion, being the mean of the squared distance

from a feature to the cluster center. Inertia, being the sum of squared distances of features from the nearest cluster center.

3. The goal of either distortion or inertia measurements was to find the value that would minimize inside cluster distances and maximize outside cluster distances.
4. **Notably, no preferred k-value was found due to the plot having no sharp bend (will be discussed in the discussion area).**

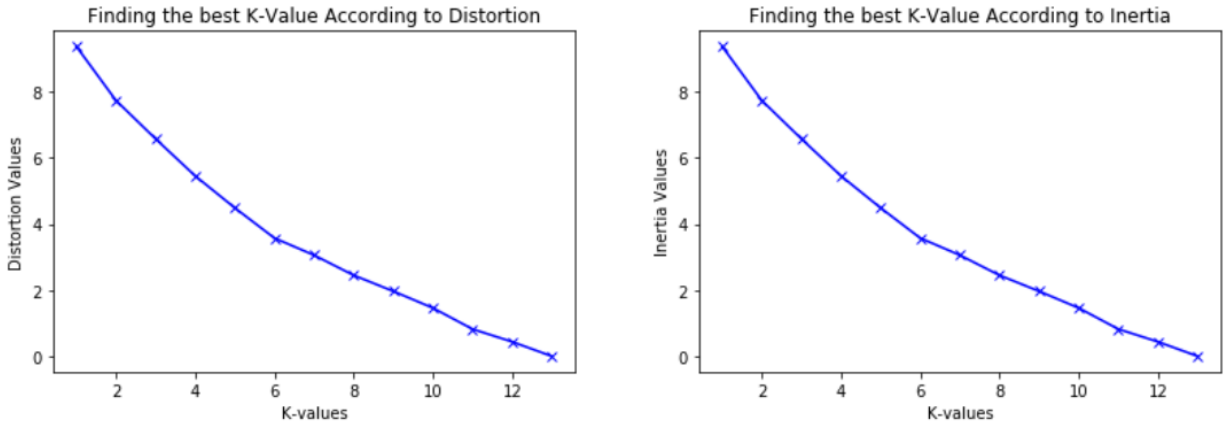

Fig 5. "Elbow Method via Distortion and Inertia". The left graph represents the elbow method using distortion values and the right graph represents the elbow method using inertia values.

I. 7th: Venue Category Frequency

1. The frequency of venue category by occurrence was then found for the top 5 most common venues.

Hotel	0.200000
Trail	0.133333
Video Store	0.133333
Food	0.066667
Soccer Field	0.066667
IT Services	0.066667
Dog Run	0.066667
Recreation Center	0.066667
Print Shop	0.066667
American Restaurant	0.066667
Restaurant	0.066667

Name: 1st Most Common Venue, dtype: float64

Food	0.200000
Coffee Shop	0.133333
Dog Run	0.133333
Gym	0.066667
Convenience Store	0.066667
Bakery	0.066667
Trail	0.066667
Gym / Fitness Center	0.066667
Pool	0.066667
Financial or Legal Service	0.066667
Mexican Restaurant	0.066667

Name: 2nd Most Common Venue, dtype: float64

Cocktail Bar	0.133333
Bus Stop	0.133333
Intersection	0.133333
Theme Park	0.066667
Bubble Tea Shop	0.066667
Yoga Studio	0.066667
Creperie	0.066667
Hotel	0.066667
Music Venue	0.066667
Dog Run	0.066667
Warehouse Store	0.066667
Mexican Restaurant	0.066667

Name: 3rd Most Common Venue, dtype: float64

Yoga Studio	0.266667
Mexican Restaurant	0.133333
Gay Bar	0.066667
French Restaurant	0.066667
Gym	0.066667
Soccer Field	0.066667
Bar	0.066667
Cycle Studio	0.066667
Liquor Store	0.066667
Indian Restaurant	0.066667
Athletics & Sports	0.066667

Name: 4th Most Common Venue, dtype: float64

Food	0.200000
Yoga Studio	0.133333
Financial or Legal Service	0.133333
New American Restaurant	0.066667
Steakhouse	0.066667
Food Truck	0.066667
Cosmetics Shop	0.066667
Pizza Place	0.066667
Speakeasy	0.066667
Scenic Lookout	0.066667
Dance Studio	0.066667

Name: 5th Most Common Venue, dtype: float64

Fig 6. "Venue Frequency." Shows the frequency count of each venue category. From top left to bottom middle: 1st most common venue, 2nd most common venue, 3rd most common venue, 4 most common venue, and 5th most common venue.

IV. Results

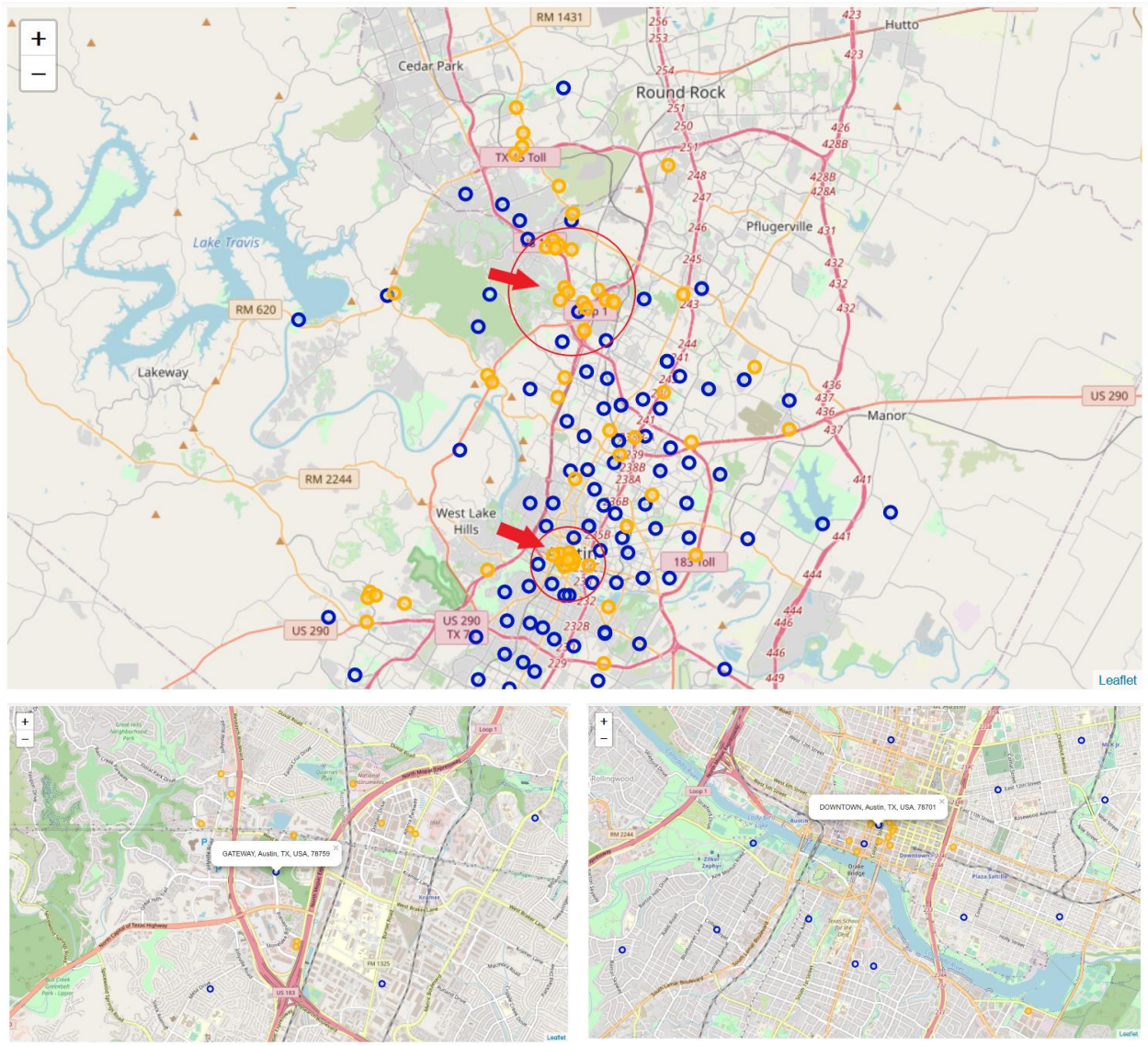

Fig 7. "Overlap Zoomed-In." The orange circle represent Big Tech locations and the blue circles represent neighborhood location. The top map is an observation of Fig.1 where the high density locations of Big Tech were noticeable, specifically two red circles identify these areas of interest. The bottom left map show the zip code of the top (larger) red circle. The bottom right map show the zip code of the bottom (smaller) red circle.

A. Selected Zip Codes of Neighborhoods that have high proximity to Big Tech clusters
    1. A look at fig.7 show that there are two main areas where Big Tech congregates. Specifically, north of Downtown Austin and Downtown Austin itself.
    2. Fig.7 also reveals there zip codes, namely 78759 and 78701.
    3. The neighborhoods in these two zip codes are: Westover Hills, Gateway, Spicewood, UT Austin, Downtown Austin, West Austin, and East Cesar Chavez.

B. Selected Zip Codes with High Population Size and Positive Growth

1. Zip code 78748 was selected because it ranked 7th in population size and 1st in population growth.
2. Zip code 78744 was selected because it ranked 5th in population size and 6th in population growth.
3. Neighborhoods in these zip codes are: Dittmar Crossing, Cherry Creek, South Brodie, Slaughter Creek McKinney, Bluff Springs, Onion Creek, and Franklin Park.
4. The other zip codes were not selected because they either fulfilled one of the criteria or were poor on one of them.

C. Venue/Features
1. The top 100 venues in each neighborhood, within a 500 radius were retrieved from the Foursquare API.

D. K-Means Clustering/Elbow Method
1. Venue categories were used as features.
2. The elbow method did not result in a significant k-value, where the sum of squared errors were reduced.
3. Arbitrary k-value of 8 was chosen, just for visualization, but was not considered overall.

E. Venue Category Frequency
1. Venue categories from the 1st most common to the 5th most common were chosen: hotels, food (twice), cocktail bars, and yoga studios.
2. Only the top 5 were chosen due to a desire to limit budget.

V. Discussion

A. Dropped and Renamed Neighborhoods
1. There were some neighborhoods that were not recognized by the OSM database, and a manual search via google maps did not also return a geolocation. These neighborhoods were dropped, specifically 2 neighborhoods.
2. The Hays Wartha neighborhood was dropped since neither google search nor the OSM database recognized the location.
3. The Robinson Ranch neighborhood was dropped since the OSM database did not recognize it and according to google search it was in a different county (not Travis Country but Williamson County i.e. no longer Austin).

4. Some neighborhoods were renamed by a more common name, due to the OSM database not recognizing the string structure/formatting (how it was presented to Nominatim), but the neighborhood was the same.
5. The two dropped can be a source of missed data or missed investment opportunities.

B. Variation of Data depending of Zip Code Changes
   1. In conducting this study zip code assignment to neighborhoods and population data was a key characteristic. However, on further research it was found that zip codes are in fact unreliable sometimes. This is because zip codes apparently are set-up by the USPS, which changes the assignment of zip codes depending on the route that they serve.
   2. In other words the USPS can add neighborhoods into one zip code, or remove a neighborhood into one zip code.
   3. This variation can be a source of error in neighborhood data since the neighborhoods of interest were filtered using zip codes.
   4. This variation can be a source of error in population data since populations were added by zip code, and if neighborhood zip codes changed then so does the sum of the population.

C. Troublesome Elbow Method/Unreliable K-Means Clustering
   1. K-means clustering was used in the hope of finding hidden patterns in venue categories within the chosen neighborhoods. Although finding an appropriate k-value with the elbow method returned a smooth curve (fig.5) , even if the features were normalized via the scikit-learn *StandardScaler( )* function.
   2. A smooth curve could mean clusters had different sizes and feature density. This means some of the neighborhoods had either a higher density of venues than the rest, or some of the neighborhoods had a lower density of venues than the rest, or both.
   3. Further investigation of mapping data in fig.1 revealed that indeed some neighborhoods had little feature count (venue count) such as the suburbs compared to neighborhoods with high feature count such as Downtown Austin.
   4. This is smooth curve is the reason why the result of K-Means clustering was not taken into account.

5. Although k-value of 8 was arbitrarily used just to visualize, but again, the result was not used to shape the final conclusion.
6. It is proposed that zip code 78701 (the Downtown Austin neighborhood area and the like) has the higher feature density, meaning a larger amount of venues. This is important to the REIT company since it could mean that real estate investment opportunities are focused on these locations compared to the rest. Other areas may lack or have little real estate investment opportunities.

D. Venue Types to Build REIT Portfolio Around
   1. The property location choices are between neighborhoods in 78759, 78701, 78748, and 78744. Data from fig.1, fig.2, fig.3 to verify.
   2. Property types to possibly invest in are those concerned in hotel real estate, food (perhaps restaurants?) real estate, cocktail bar real estate, and yoga studio real estate. Data from fig.6 to verify.

VI. Conclusion (Open-Ended)/Suggested Improvements

A. If from this data alone, one was to build an REIT portfolio in Austin, TX, one should choose to invest in neighborhoods in zip code 78759, 78701, 78748, and 78744. This answers the question where in Austin, TX to invest in.

B. If from this data alone, one was to invest in certain property types in these neighborhoods the best types would be hotel real estate, food real estate, cocktail bar real estate and yoga studio real estate. This answers the question on what type of real estate to invest in.

C. **However, even if the elbow method for K-Means clustering came out as a smooth curve, it reveals that the data must be studied again as it suggests that some of these neighborhoods have a either a high density venue count, a low density venue count, or both. Meaning, some income-generating real estate are densely focused in certain neighborhoods and some neighborhoods have minimal or do not have any.**

D. Proposed Improvements:
   1. Instead of using OSM database for geolocation, Google API might be preferable if it is more up-to-date.
   2. Instead of using zip code as a characteristic for neighborhoods and population data, use ZTCA instead. During the end of this study, it was found that the U.S. Census Bureau has a more stringent form of neighborhood assignment called ZTCA or Zip Code

Tabulation Area (note that ZTCA is not the same as the USPS zip code). ZTCA is less variable than the zip code when it comes to time.

3. Use the 2020 population data via the U.S. Census Bureau API, once it has been published in order to find more current population trends.
4. Along with the U.S. Census Bureau API, compare other characteristics that may aid in investment types such as income and household types. For instance a single person household may prefer spending money on cocktail bars but a family household may prefer spending money on hotels and family resorts.
5. Take a look back at the chosen zip codes, it is hypothesized that the neighborhood that has a large feature (venue category) density compared to the rest is Downtown Austin, and that the frequency of venue categories shift depending if this neighborhood is accounted for or not. Perhaps consider only neighborhoods like Downtown Austin or exclude it all in all. The choice is up to the REIT company's intention whether to build in city centers or in suburbia.