Capstone Project Final Report - AG

Introduction

In recent years, the city of Tulsa in the mid-east of Oklahoma has been rapidly changing. In the past, it has been famous for its abundance of wealthy oil companies which made the city prosper. However, throughout the 90's the city's oil industry began to decline, and more and more companies started to move south to Texas. This left the city of Tulsa in a very low state. The population was aging, the economy was declining and as it had very little to offer, it felt like a dying city. To combat this trend, local wealthy philanthropists started investing in the city, and with an excellent multi-year plan of the city's leadership it began reviving.

As a result, Tulsa is now a great place to live. Housing prices are relatively very low (almost half of the national average), and new restaurants, retailers and other venues for past time activities are constantly popping up. The city is full of green lawns, beautiful parks and lush trees throughout. Tulsa is now experiencing a growth period, and construction is evident in many places. As such, opportunities to capitalize on this momentum are plentiful, and the real estate market is booming.

This project will focus on assisting a local construction and development company in Tulsa, Oklahoma. The data will help the company determine recent places of growth, despite the ongoing Covid-19 pandemic, where new homeowners would best utilize mixed use real estate, urban density, walkability to retail and food/beverage concepts, along with other amenities. Their aim would ultimately be double: To identify prime locations for investment in housing, and to find opportunities for developing the surrounding up-and-coming areas.

Data

To complete the project, a few sources of data were needed. However, Tulsa doesn't have readily available data like New York or Toronto. While there are sources to find Tulsa's list of neighborhoods, none (that I could find) come in table form or include their coordinates. As such, I had to resort to using zip code data, instead. This meant that the area represented by each row in the dataframe will be bigger, but at least it could be shown on a map. It also allowed me to find much more information on it.

As for the sources I used:

1. Realtor.com – This website provided most of the usable columns for my main dataset. This includes such columns as "median listing price", "median listing price per square feet", "new listing count" and "total listing count". Additionally, Realtor.com had historic data for each month in the last 4 years. This historic information was used to show trends in each zip code.

2. Zipatlas.com – This website provided pinpoint coordinates of each zip code in Tulsa. It was also where I found the population and population density for each zip code area. However, this data is dated to 2010 as it was gathered by The United States Census Bureau, and the results of the 2020 census were not yet publish (as of writing these lines).
3. FourSquare API – By using FS I obtained the list of most popular venues in each zip code area, in a 1,000 meter radius (approximately 0.62 US miles).

Methodology and Results

For the completion of this project many steps in data wrangling had to be taken. The raw data that was obtained had significant problems that needed to be addressed and corrected. The dataset containing the coordinates for each zip code was especially important since as far as I know, Geocoder can't be used to find the coordinates of zip codes, only addresses. Column names had to be changed, data types and formats were adjusted, many columns and rows were dropped due to missing values and/or redundancy and, lastly, the two sources of data were merged.

The total number of rows in the first dataset was 28. This coincided with the number of total zip codes, as each row represented a different one. However, as mentioned in the Data Section, this data is dated to 2010. One of the limitations of this project is a lack of updated data. With that being said, since there is a census every 10 years, a repeat of this project within the coming year (when the census data is published) could be supported by more accurate data. As such, the 3 columns (containing information on: Population, people per square mile and average income per household) are to be regarded as mere reference points, for a "generalized idea" of the situation today: **This project assumes that no significant changes occurred *between* the zip codes. That even though things changed, the change was proportionate to the other zip codes (i.e. if 3 of the zip codes had values 30, 10, 20 in 2010 – the values today could be 60, 20, 40 - but not 40, 60, 20)**.

The second dataset had more recent data, up to a month prior to when it was received – June 2020. However, it included rows for each zip code in the US, and each month in the four years prior to June 2020. When it was merged with the first dataset only the rows containing information regarding the 28 zip codes of Tulsa County remained. The merged dataset contained 1024 rows.

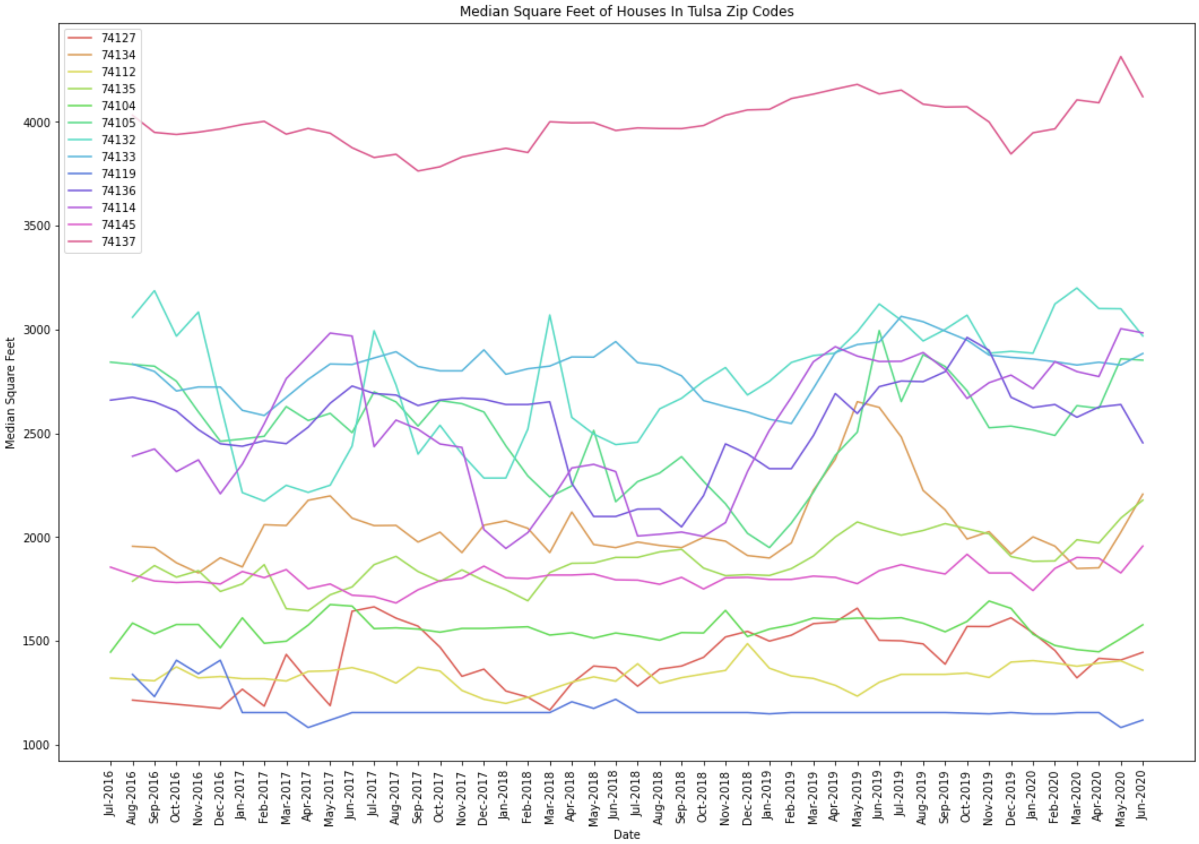
While the original dataset had 28 zip codes, some had missing data. This made them unusable for the purpose of this project and they had to be dropped. The final dataset only contained information on 13 zip codes. Even though this is less than half, it was first discussed with the local development and construction company and was approved by them.

Following is a list of the main variables in the final dataset. These variables will later be used for comparisons between and within the zip codes.

Column	Description
Median Listing Price	The median listing price within the specified geography during the specified month.
New Listing Count	The count of new listings added to the market within the specified geography. The new listing count represents a typical week's worth of new listings in a given month. The new listing count can be multiplied by the number of weeks in a month to produce a monthly new listing count.
Price Increase Count	The count of listings which have had their price increased within the specified geography. The price increase count represents a typical week's worth of listings which have had their price increased in a given month. The price increase count can be multiplied by the number of weeks in a month to produce a monthly price increase count.
Price Decrease Count	The count of listings which have had their price reduced within the specified geography. The price decrease count represents a typical week's worth of listings which have had their price reduced in a given month. The price decrease count can be multiplied by the number of weeks in a month to produce a monthly price decrease count.
Median List Price Per Sqft	The median listing price per square foot within the specified geography during the specified month.
Median Listing Sqft	The median listing square feet within the specified geography during the specified month.
Avg Listing Price	The average listing price within the specified geography during the specified month.
Total Listing Count	The total of both active listings and pending listings within the specified geography during the specified month. This is a snapshot measure of how many total listings can be expected on any given day of the specified month.

Once the main dataset was established, it was segmented by zip codes. This enabled further analysis and exploration by visualizations. Two main types of graphs were created: Comparisons between zip codes and comparisons between the variables within each zip code. Following are examples of the graphs:

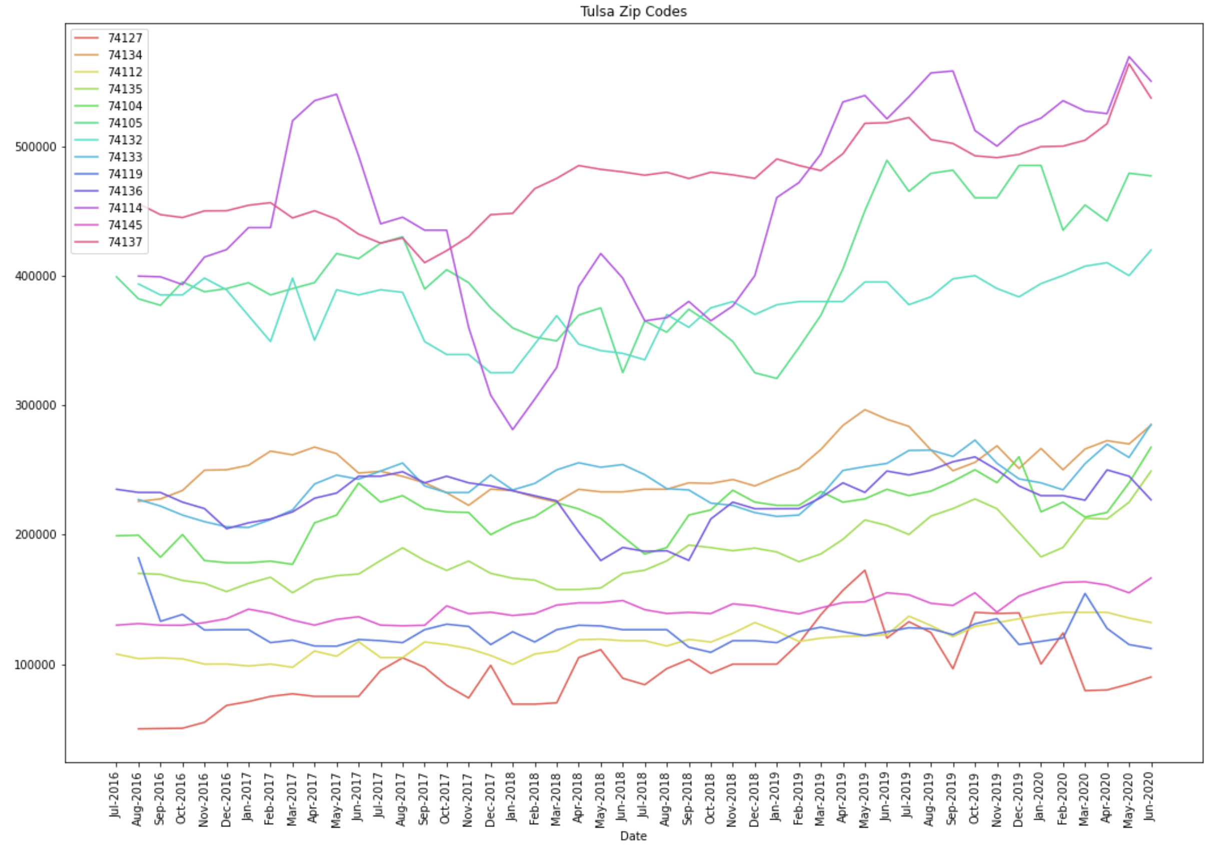
This first graph illustrates the differences in the median square feet of each zip code, in each of the months in our dataset (June 2016 to June 2020).



Notice how the homes in 74137 are significantly larger than those in the rest of the sampled zip codes. On the other hand, those in 74119, 74112 and 74127 are the smallest.

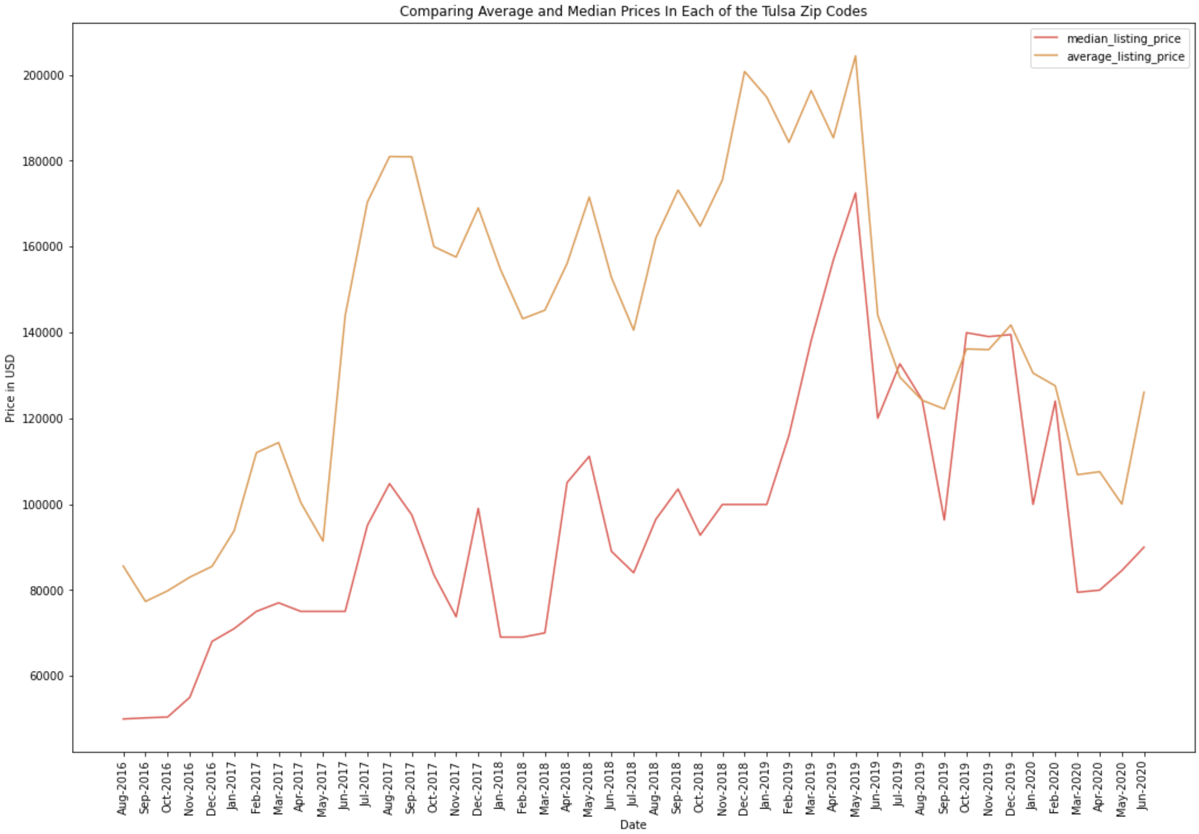
The following 3 graphs are all interactive. Since it is not possible to have a Word or PDF document create the graphs interactively, a sample figure is presented as an example.

In the first of the three there is a comparison between the zip codes on each of the variables in the dataset. The interactive function allows the user to choose a column from the drop down list and the graph will compare the zip codes on that basis. The example presented here is of the median listing price.



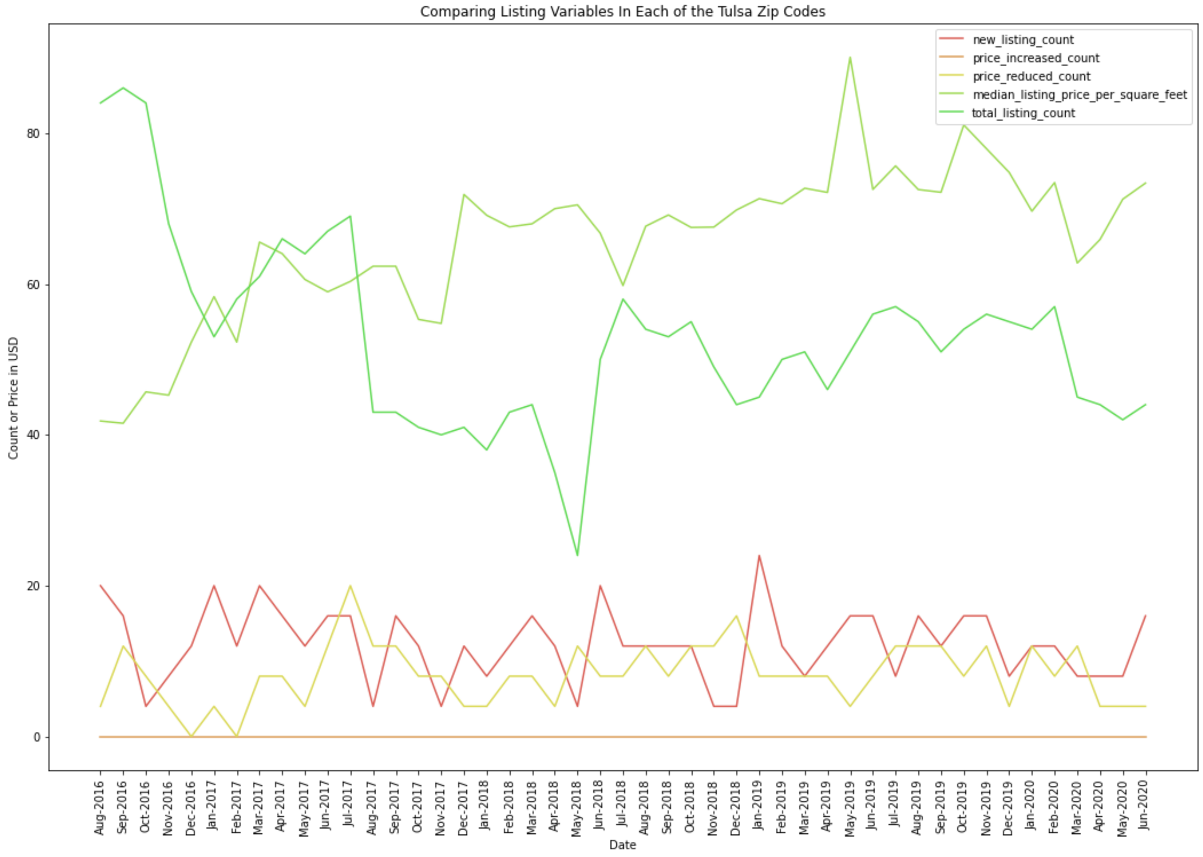
Once again, it seems that although no longer as clearly "a cut above the rest", the houses in 74137 are higher than those in the rest of the sampled zip codes – meaning they are listed in a higher price. However, they are relatively close in price to the houses in 74114, 74132 and 74105. On the other hand, once again those houses in 74119, 74112 and 74127 are the lowest on the graph, meaning they are listed for the lowest prices.

The second interactive graph shows a comparison between the average and median listing price. The interactive function allows the user to choose a zip code from the drop down list and the graph will compare the two variables on that basis. The example presented here is of zip code 74127.



Notice the spike in the average listing price in the period between June 2017 and May 2019. Interesting how the median listing price remained much lower during that period, aside from May 2019 and the few months preceding it. This suggests that for most of those two years only a small part of the houses had very high listing prices, which raised the average significantly. Closer to May 2019 the median listing price began increasing, which suggest more and more houses were listed at higher prices. Regardless of this spike, overall the houses in zip code 74127 show a trend of increase in price in the past 4 years.

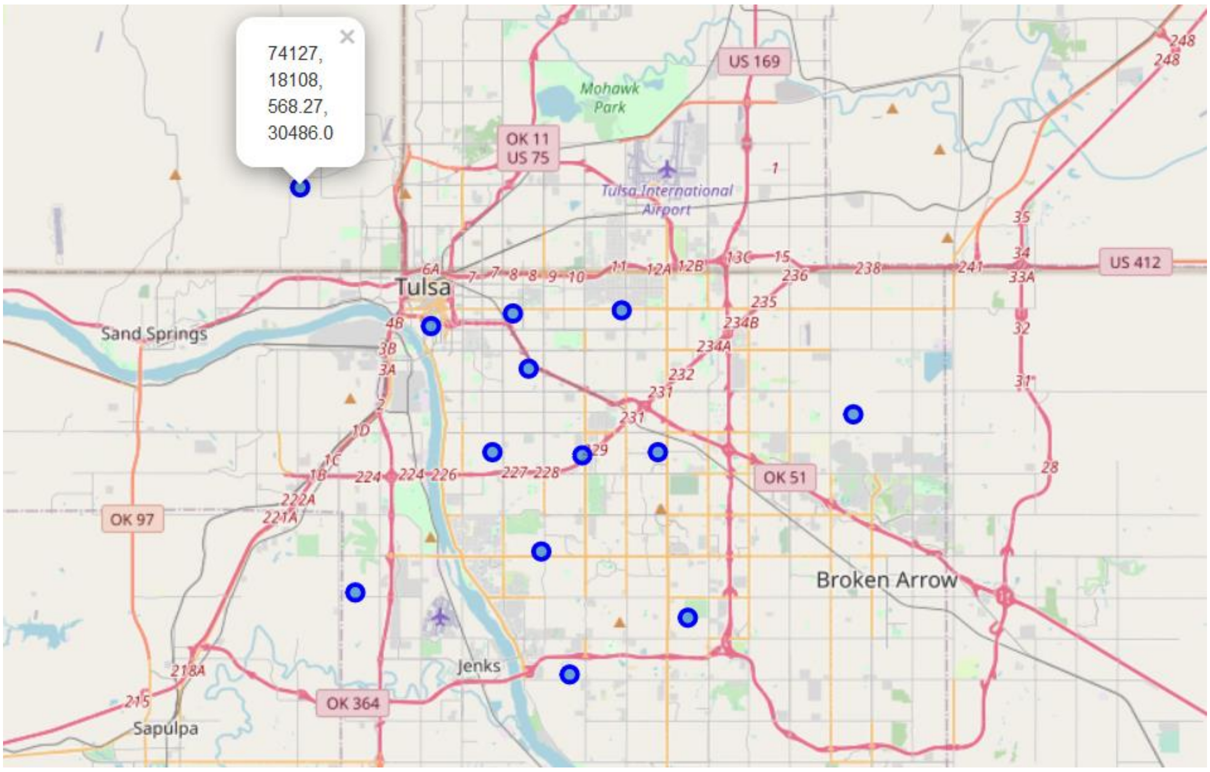
In the last of the three interactive graphs there is a comparison between the remaining variables in the dataset – those with smaller values. The interactive function allows the user to choose a zip code from the drop down list and the graph will compare the variables on that basis. The example presented here is of zip code 74127 once again.



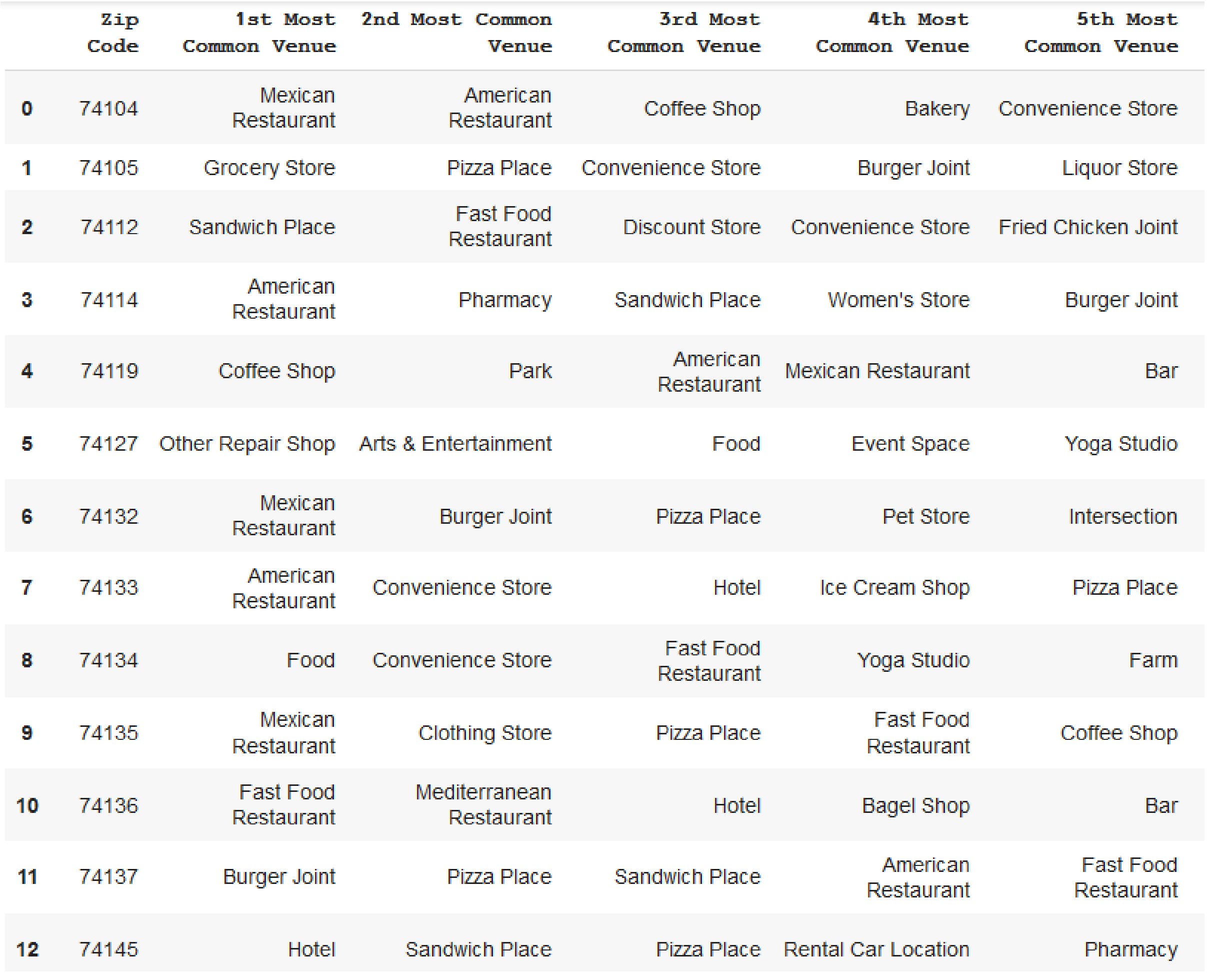
The first thing that can clearly be inferred from this figure is the "price increased count" line, which is a constant 0. This means that throughout these 4 years none of the listings were ever changed to be listed at a higher price than they were initially. While there is no data on the actual selling price, which is likely to have been different than the listing price, it is indeed worth noting how the price listed was not increased even once. Comparatively, the "price reduced count" line has values between 0-20.

Other notable variables are the "total listing count" and "median listing price per square feet" variables. Notice how the total amount of houses for sale is showing a decreasing trend, while the price per square feet is on the rise. This means that while there are fewer and fewer houses for sale in zip code 74127, the larger ones will likely be sold for higher now than they would have 4 years ago.

The next step in this project was to create maps using Folium. This enables the reader to better understand the information provided by the figures above. A visual representation of the location of each zip code, along with the population, the population per square mile and the average household income (from 2010) was displayed. The notebook allows the viewer to click and see the mentioned information, while Word and PDF do not support this feature. A screen caption is presented below, with the pop up information on zip code 74127.



At this final point in the project, Foursquare API was utilized. This allowed for each zip code to be explored in greater depth. From the coordinates of each zip code center, Foursquare API was used to identify the most commonly visited venues. Since zip codes cover a relatively large area, the radius of each explored zip code was set to 2000 meters (which is 2 kilometers or about 1.25 miles). In other words, Foursquare disclosed the most visited venues in a 2000 meter radius from each zip code center. This information was later refined to a table containing only the top 10. For the sake of viewer comfort, following is a table of only the top 5 most visited venues in each zip code (the complete table is at the end of the notebook).



	Zip Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	74104	Mexican Restaurant	American Restaurant	Coffee Shop	Bakery	Convenience Store
1	74105	Grocery Store	Pizza Place	Convenience Store	Burger Joint	Liquor Store
2	74112	Sandwich Place	Fast Food Restaurant	Discount Store	Convenience Store	Fried Chicken Joint
3	74114	American Restaurant	Pharmacy	Sandwich Place	Women's Store	Burger Joint
4	74119	Coffee Shop	Park	American Restaurant	Mexican Restaurant	Bar
5	74127	Other Repair Shop	Arts & Entertainment	Food	Event Space	Yoga Studio
6	74132	Mexican Restaurant	Burger Joint	Pizza Place	Pet Store	Intersection
7	74133	American Restaurant	Convenience Store	Hotel	Ice Cream Shop	Pizza Place
8	74134	Food	Convenience Store	Fast Food Restaurant	Yoga Studio	Farm
9	74135	Mexican Restaurant	Clothing Store	Pizza Place	Fast Food Restaurant	Coffee Shop
10	74136	Fast Food Restaurant	Mediterranean Restaurant	Hotel	Bagel Shop	Bar
11	74137	Burger Joint	Pizza Place	Sandwich Place	American Restaurant	Fast Food Restaurant
12	74145	Hotel	Sandwich Place	Pizza Place	Rental Car Location	Pharmacy

The information in this table can be used to make subjective assumptions on the population in each zip code. While I might have my own thoughts on the matter, this project is meant to assist the development construction company in making a decision, not to influence them in any way. As such, any subjective assessments and assumptions should be made exclusively by them.

Note:

No machine learning algorithms were used in the scope of this project. This was due to the nature of the business problem, which did not require any predictions to be made.

Discussion

Upon review of the methodology and results section, most of the observations that were to be made were already discussed. However, in regards to zip code 74127, by noting that the listings are on a trend of an increase in price, with an increase in the price per square feet, an apparent recommendation comes to mind. It might be worthwhile to build a house prioritizing a large square footage, while minimizing other costs as deemed appropriate. With the trends displayed, future house sells could make this a worthy investment.

At this point I'd like to address the table noting the most frequented venues in each zip code. As mentioned, I recommend that the development and construction company make their own assumptions on the nature of the population living in each zip code. However, with their assumptions in mind, the company can better fit a new potential building to their desired area. For instance, in zip code 74137 the top 5 venues are food related establishments. Perhaps building a place of business that holds within itself different food vendors should be considered.

Conclusion

As the interactive figures in the first part of the results section only displayed one option from their respective drop down menus, clearly information is missing from this report. Since the main solution to the business problem is solved by allowing the use of the notebook, this report's main purpose was to provide greater background and explanations for the items in the notebook. By allowing the reader to better understand the reasoning and methods used, it is my hope that this report will become an integral reading to accompany the notebook.

Finally, it is important to discuss the limitations of this project. First off, as mentioned in the data section, the age of the data: Tulsa is known for having gone through a booming growth and improvement in recent years. Lacking the data to reflect that is a big issue, since many changes are very likely to have occurred. A future repeat of this project could use more recent data once the 2020 census is published. Secondly, most of the zip codes had missing rows, which made them unusable. Access to more reliable databases, or a means of collecting missing values would improve a future project greatly. Lastly, I had obtained a json file containing coordinates of the borders of all the zip codes in Oklahoma. I planned to use it for a clearer visualization of the zip codes on the Tulsa map. However, after many hours of attempts, I could not find the right way to load it into Folium. Either due to not using the right keys or not using the right code, all I got were errors. An expert would have managed to present better visualizations for the maps.

With all the being said, this project was very challenging, and I learned a lot while working on it. It is my hope that it has achieved its goals, and that the local Tulsa company will benefit greatly from the insights this project can provide for them.