New Restaurant Location Recommendation in Orlando, FL

Liwen Huang & Ziqian Wang | September 20, 2021

⚠ **Interactive Report Link: https://bit.ly/3tliMzu**

Motivation

According to the National Restaurant Association, the U.S. restaurant industry was worth $659 billion in 2020, which was $240 billion below pre-pandemic sales estimates. COVID has caused more than 110,000 restaurant businesses to shut down or shift operational models completely, raising deep disturbance about dining out in general. According to new research, the majority of closed restaurants had been operating for more than 16 years, with 16% operating for more than 30 years (Kelso, 2021). The pandemic brought a deep sense of crisis, not only to new restaurants but also restaurants that had been operating long term. Although the Association predicts the growth of restaurant industry sales will be double-digits in 2021, that "[won't be] nearly enough to make up for the substantial losses experienced in 2020" (National Restaurant Association, 2021).

The pandemic and deuterogenic work-and-study-at-home policies have fueled large-scale digital transformation use within the restaurant industry. Panasonic conducted a series of comprehensive industry surveys and research about "How COVID-19 is transforming Food Service & Food Retail," with 100% of operators believing that the pandemic has increased the urgency to adopt transformational technologies; 71% of operators indicated that digital transformation is now significant. It is particularly essential—and urgent—that restaurants provide curbside takeout and delivery services to consumers. The application and generalization of digital transformation has derived neo-catering models, such as ghost kitchens, that offer takeout and delivery catering services by cooperating with third-party delivery companies but does not include dine-in seating or typical operations. The pandemic and digital transformation simultaneously brought challenges and opportunities to new investors and businesspeople in the restaurant industry.

Orlando, a tourism city in the State of Florida, is one of the most-visited cities in the world, famous for its theme parks, cultural sites, and the Walt Disney World Resort. Orlando attracts millions of visitors every year; in 2018 alone, more than 75 million tourists visited. The pandemic inflicted heavy losses on Orlando's tourism and restaurant industries, despite state and city governments providing a series of financial relief policies in 2020. The industry is recovering slowly with the advent of COVID vaccines.

This paper makes three contributions: First, it provides an overall exploratory data analysis on Orlando's demographics features. Second, it analyzes the correlation among restaurant types, demographics features, and restaurant performance. Third, it classifies regional categories of restaurants and provides comprehensive location recommendations for new restaurants and aspiring investors and operators.

Data Sources

Yelp Business Datasets (Primary Data Source)

The Yelp Business Datasets contain two datasets: the business, and the check-in. The original business datasets include data, attributes, and categories. Each business has eight variables, including business ID, name, ZIP code, coordinates, stars, review counts, and open status. Each business also has 41 attributes, including but not limited to takeout, drive through, access to Wi-Fi, parking, and so on. This information can be downloaded from Yelp Open Dataset.

Orlando, FL Population and Demographics Datasets

The second dataset contains population density, income density, crime data, and home value data in Orlando, FL from 2015 to 2021. The population data are from “The 2010 US Census Population by Zip Code” project by Jon Bittner. This dataset contains ZIP code, population, and ZIP code area of City of Orlando, FL. The household income data are from the University of Michigan’s Population Studies Center.

Orlando, FL Neighborhood Datasets

The third dataset contains Orlando neighborhoods from Wikipedia and properties from the FourSquare API interface. This dataset includes official-only neighborhoods as defined by the Orlando government. The reference of the neighborhood project is based on the IBM Data Science capstone by Liwen Huang.

Orlando, FL Crime Datasets

The fourth dataset contains Orlando crimes data from OPD Crimes, Orlando Police Department records management system.

Dataset	Yelp Business/Checkin	Population & Demographics	Neighborhood	Crime
Source	Yelp.com	US Census/University of Michigan Population Studies Center/Zillow	Wikipedia/Foursquare	OPD Crimes
Format	JSON	CSV/Shapefile	CSV	CSV
Access Method	Download	Download	API	Download
Variables	Business id, business name, rating, review counts, checkin counts, price tags, categories	Population, Income, Zillow Home Index, ZIP Code, Geo	Neighborhood, properties, coordinates	Offense category, offense type, status, location
Size	118MB/379MB	3.23MB/2KB/52.3MB	148KB	37.7MB
Date	2020	2010/2020/2021	2021	2021

Data Manipulation Methods

Yelp Business Datasets

First, we removed non-Orlando and non-restaurant businesses by filtering cities and categories.

Second, we explored **missing values** in the datasets by missingno library. Figures 1 and 2 show the missing values in each column before and after filtering out closed businesses. The Yelp datasets contain 3,748 rows (businesses) including closed businesses, 2,565 rows (businesses) excluding closed businesses, and 346 businesses that do not have price range value (attributes.RestaurantsPriceRange2). The missing value rates are 9% and 13.5%, respectively. Author/researcher Joseph L. Schafer considered a missing rate of 5% or less as inconsequential. As a result, we conducted a deep investigation into the 346 missing data of price range value. The result of the investigation revealed that in the missing data, 21 restaurants were either closed or cannot be found on Yelp.com, 7 are not restaurants, 29 have multiple locations with price ranges, 2 changed business names, and 24 complemented prices range from Yelp.com. After missing data handling, the Yelp dataset price range column reduced the missing value from 346 to 185. The total number of open businesses is 2,534. The **final missing value rates are 4.9% and 7.3%**, respectively.

Figure 1 Missing values before cleaning

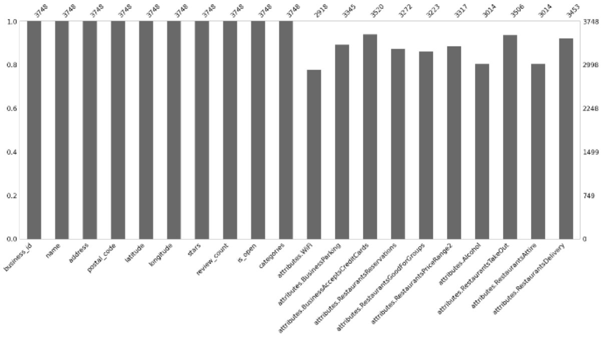
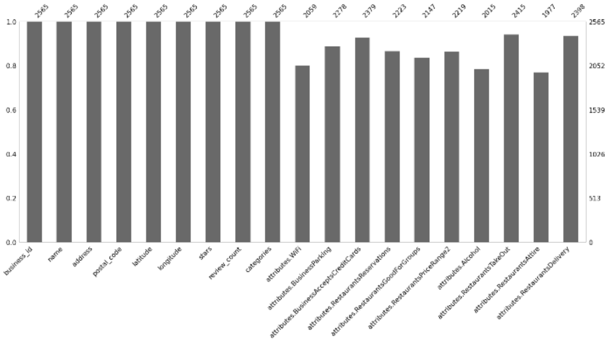


Figure 2 Missing values after cleaning



Lastly, we dropped unnecessary columns, including but not limited to address, city, state, HasTV, and so on. Additionally, we renamed columns to informal names, for example, changing “attributes.WiFi” to “wifi.”

Orlando, FL Population and Demographics Datasets

The population dataset contains four columns: Zip/ZCTA, 2010 Population, Land Sq Mi, and Density/sq mi. The income dataset contains four columns: ZIP code, Median Income, Mean Income, and population.

Population datasets: First, we renamed columns to informal names, such as “Zip/ZCTA” to “zipcode”. We dropped the unnecessary column “Land Sq Mi”. Second, we left Orlando, FL, population data only by filtering ZIP code. We reset the index and sorted by ZIP code.

Income datasets: First, we dropped unnecessary columns, such as mean and population, and renamed columns to informal names, such as “ZIP” to “zipcode”. Second, we left Orlando, FL, income data only by filtering ZIP code. We reset the index and sort by ZIP code.

Last but not least, we merged population datasets and income datasets by ZIP code.

Orlando, FL Neighborhood Datasets

Neighborhood is the area around you or around a particular place or people who live there (“Neighborhood,” 2020). To explore and target recommended locations across different venues (according to the presence of amenities and essential facilities), we will access data through the FourSquare API interface and arrange them as a dataframe for visualization.

First, we discarded the official neighborhood list from Wikipedia and instead got coordinates’ information. We dropped unnecessary columns and mapped coordinates to each neighborhood.

Second, we get nearby venues through the FourSquare API interface. The venue dataframe contains seven columns, street, street latitude, street longitude, venue, venue latitude, venue longitude, and venue categories. We have 259 unique categories and 1,210 venues in Orlando.

Third, we counted quantities of venues in each neighborhood. Additionally, we defined a function to return the most common venues to nearby neighborhoods. We created a new dataframe containing all neighborhoods with the top ten common venues for each of them.

Orlando, FL Crime Datasets

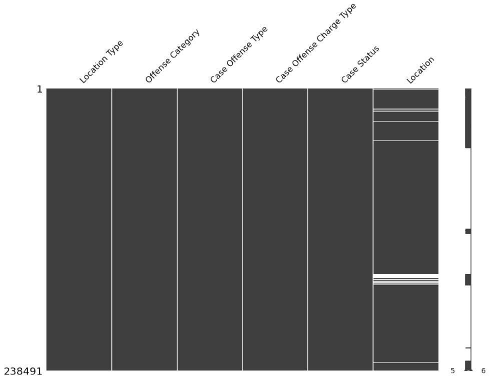
The crime dataset contains 10 columns, case number, case datetime, location, location type, offense category, offense type, case disposition, and status.

First, we dropped unnecessary columns, such as datetime and case number. We renamed columns to casual names, such as Case Offense Location Type to Location Type.

Second, we checked the missing value rate; Figure 3 shows it as significantly low. We dropped null values. We split the location column into latitude and longitude columns. We created a new dataframe including latitude, longitude, location type, offense category, and case status.

Third, we explored offense counts of each offense category and offense counts of each location.

Figure 3 Mssing values in crime datasets



Analysis and Visualization

There are three questions we would like to answer:

- What factors are correlated to restaurant performance in Orlando?
- What types of restaurants are outperforming? Moreover, in which area(s) are they outperforming?
- Given a specific region, what insight can we provide to the stakeholder as suggestions and references?

There are two significant issues to resolve before the analysis:

- How to define and construct the restaurant performance metric.
- How to divide all restaurants into meaningful and manageable types.

Features Engineering

To begin with, we would need to assign all categories to a nationality type of each restaurant to ensure that all businesses are labeled with **at least one of the types** and with as little overlap as possible. From 218 restaurant tags extracted from the original category column, we filtered out 130 tags related to our dataset. Therefore, to generate meaningful insights from the dataset, we need more comprehensive categories from this list. The criteria to generate the desired categories are as follows:

- Each of the generated categories must contain sufficient (5% or above) data.
- All categories with more than 5% of the data should be preserved.
- Any generated category is allowed to have less than 5% of the data if the category is generated from multiple smaller categories.

In research from Banerjee and Poddar (2021) and enhanced according to our exploratory data analysis, we generated seven major restaurant types. They are:[1]

- Latin America
- North American
- Ottoman Cuisine
- Quick and Greasy

[1] See details on Appendix 1.

- European
- Asian
- Cafes & Desserts
- Other(residual)

In addition, we generate the **performance index score**. It is an ordinal score that could present the relative performance ranking and non-linear magnitude of performance differences.

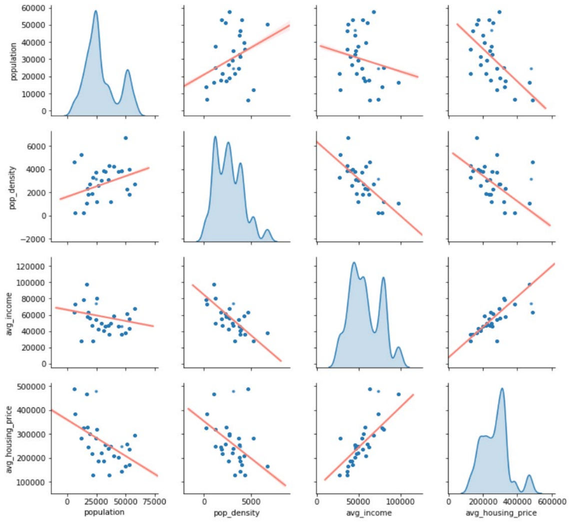
$$\begin{aligned}&Restaurant\ Performance\ Index\\&\qquad = \left((Volume\ + Total\ Review\ Counts) \div Total\ Operation\ Days\right)^{(1+0.08\times Stars)}\end{aligned}$$

The rationale behind the score is:

- From our previous exploratory data analysis, we find that both check-in volume and total reviews are **linearly correlated** with the customer volume for the restaurant. However, there is **no clear linear relationship** between check-in volume and total reviews.
- There is an inconsistency of the timespan among the restaurants' data; we would want to standardize and constrain this inconsistency.
- "A one-star increase leads roughly to a 9% increase in revenue" (Luca, 2016). We adjusted the article's estimation of 0.089 to 0.08. There is no description of the price distribution in the article, while we found a heavy right skewness of our data's price range that would weaken the boosting effect of Yelp stars. To reflect this dataset's characteristics, we decided to round down to 0.08 rather than round up to 0.09.
- All other approaches to approximate revenue have been proven ineffective or misleading.[2]

Community features correlations

Figure 4 Scatter matrix among community features



The scatter matrix figure, Figure 4, provides some key insights about community features correlations:

- All community feature data are likely to be of **non-Gaussian** distribution
- We **expected strong positive correlation** between two pairs:
  - Average income and housing price
  - Population and population density
- **Strong negative correlation** between:
  - Average income population density
  - Average housing price and population density
  - Average housing price and population
- **Moderate negative correlation** between average income and population

None of the features in the checklist has a Gaussian distribution via the Shapiro test and normal test from scipy.stats.

By using the Pearson correlation coefficient, we summarized the strength of the linear relationship between two data samples and Spearman's correlation coefficient to define the nonlinear relationship. The Asian and North American types show a **relatively stronger overall positive correlation** to the performance index score (0.141 and 0.218, respectively),

[2] See details on Appendix 2.

while Quick and Greasy shows an overall **solid negative correlation** (-0.31). The causality is straightforward: People with higher incomes are less likely to reside in high-density residential areas, instead preferring upmarket areas with less population and density. Nevertheless, no causal inference can be concluded at this point.

At this stage, based on the general overview, several conclusions were drawn: The Quick and Greasy feature has a **relatively prominent negative correlation** with the performance index score. North American and Asian features have a **relatively noticeable positive correlation** with the performance index score. We assumed that business types should be negatively correlated with each other: There is a **trivial positive correlation** among European, Cafes & Desserts, and Quick and Greasy, suggesting certain common traits might exist. We will further examine the data by applying conditions and filters.

Correlations between community features and restaurant types by conditions

Figure 5 Correlation Coefficient between Community Features & Performance Score, by restaurant type

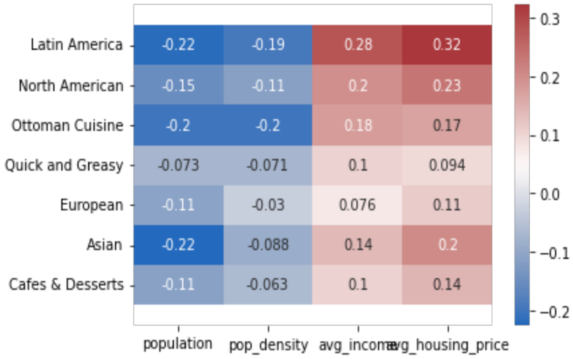


Figure 6 Correlation Coefficient between Community Features & Restaurant Type, compare to coefficient average

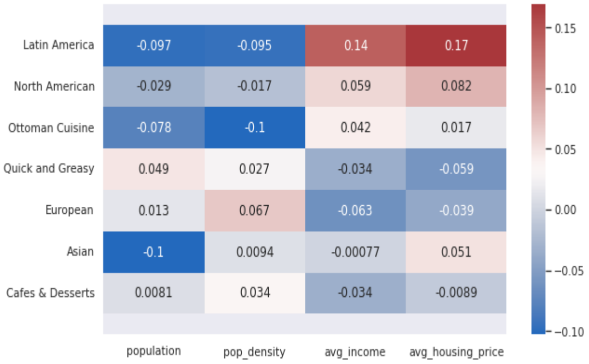


Figure 5 shows that none of the community features have a significant or notable correlation to allocating certain types of restaurants. The finding from Figure 5 and Figure 6, showing community feature effects on performance index score by restaurant type, is applicable through different areas with different community attributes. Through Figure 5 and Figure 6, concerning data restriction, come to the following conclusions:

- Population and population density have a general **negative correlation** with restaurant performance.
    - Among all restaurant types, Quick and Greasy, European, and Cafes & Desserts types of restaurant show above par performance under the **negative correlation**.
    - Among all restaurant types, Latin American and Ottoman Cuisine types of the restaurant show significant below-par performance under the **negative correlation**.

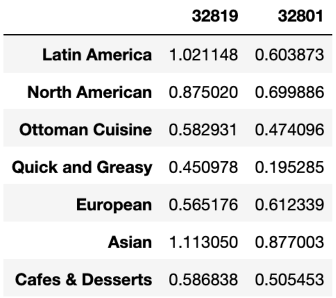
- Average income and average house price have a generally **positive correlation** with restaurant performance.
    - Among all restaurant types, Latin American, North American, and Ottoman Cuisine types show **above-par** performance under the **positive correlation**.
    - Among all restaurant types, Quick and Greasy and European restaurant types show **below-par** performance under the **positive correlation**.

- There is one uncommon finding that the Asian restaurant shows **inconsistent** performance under **different correlations**.

Scenario Analysis – ZIP code 32819 & 32801

Two postal code areas have very different community features: 32819 has a high average house price and income but low population density; meanwhile, 32801 has precisely the opposite features. We will look into the scenario and provide the last layer of our analysis to answer the stakeholder's question of "given a region, what types of restaurants would be recommended?"

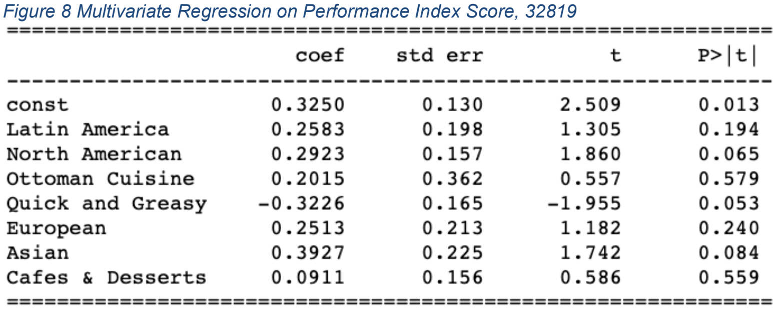
Figure 7 Average performance score comparison



	32819	32801
Latin America	1.021148	0.603873
North American	0.875020	0.699886
Ottoman Cuisine	0.582931	0.474096
Quick and Greasy	0.450978	0.195285
European	0.565176	0.612339
Asian	1.113050	0.877003
Cafes & Desserts	0.586838	0.505453

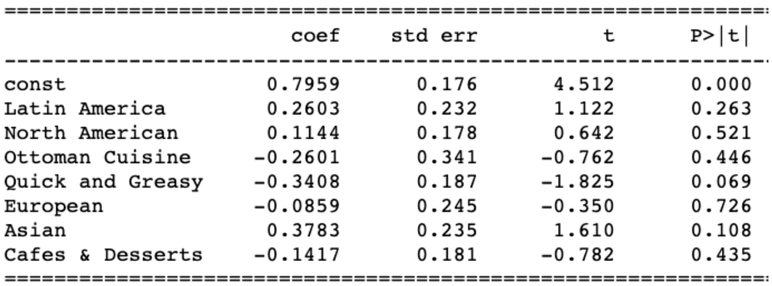
We constructed a comparison between two regions' average performance index scores by restaurant types (Figure 7). The results align with our previous findings and early conclusions. Meanwhile, it is surprising that the **Quick and Greasy** type of restaurant also performed **significantly better** in 32819. This might be a result of the skewness of the performance index scores. After examining the median, we conclude that the actual differences of **Quick and Greasy** type between the two regions are **trivial or insignificant**.

Figure 8 Multivariate Regression on Performance Index Score, 32819



```
==============================================================================
                        coef    std err          t      P>|t|
------------------------------------------------------------------------------
const                 0.3250      0.130      2.509      0.013
Latin America         0.2583      0.198      1.305      0.194
North American        0.2923      0.157      1.860      0.065
Ottoman Cuisine       0.2015      0.362      0.557      0.579
Quick and Greasy     -0.3226      0.165     -1.955      0.053
European              0.2513      0.213      1.182      0.240
Asian                 0.3927      0.225      1.742      0.084
Cafes & Desserts      0.0911      0.156      0.586      0.559
==============================================================================
```

Figure 9 Multivariate Regression on Performance Index Score, 32801



```
==============================================================================
                        coef    std err          t      P>|t|
------------------------------------------------------------------------------
const                 0.7959      0.176      4.512      0.000
Latin America         0.2603      0.232      1.122      0.263
North American        0.1144      0.178      0.642      0.521
Ottoman Cuisine      -0.2601      0.341     -0.762      0.446
Quick and Greasy     -0.3408      0.187     -1.825      0.069
European             -0.0859      0.245     -0.350      0.726
Asian                 0.3783      0.235      1.610      0.108
Cafes & Desserts     -0.1417      0.181     -0.782      0.435
==============================================================================
```

The next step is to see if the correlation is significant given the current assumptions and datasets. We conducted separate regressions on both regions to see if the restaurant type strongly affects performance score (Figure 8 and Figure 9).

- At alpha = 0.05 significant level, **none** of the restaurant type coefficients is statistically significant. Therefore, we would accept the null hypothesis that they have a litter to next to litter effect on the performance score.

- At alpha = 0.10 significant level, we find that Quick and Greasy have a **significant negative relation** with the performance score; we also have Asian have a **moderate positive correlation** with performance score. Both new findings align with our early conclusions.

Now we look at **other** postal code areas.

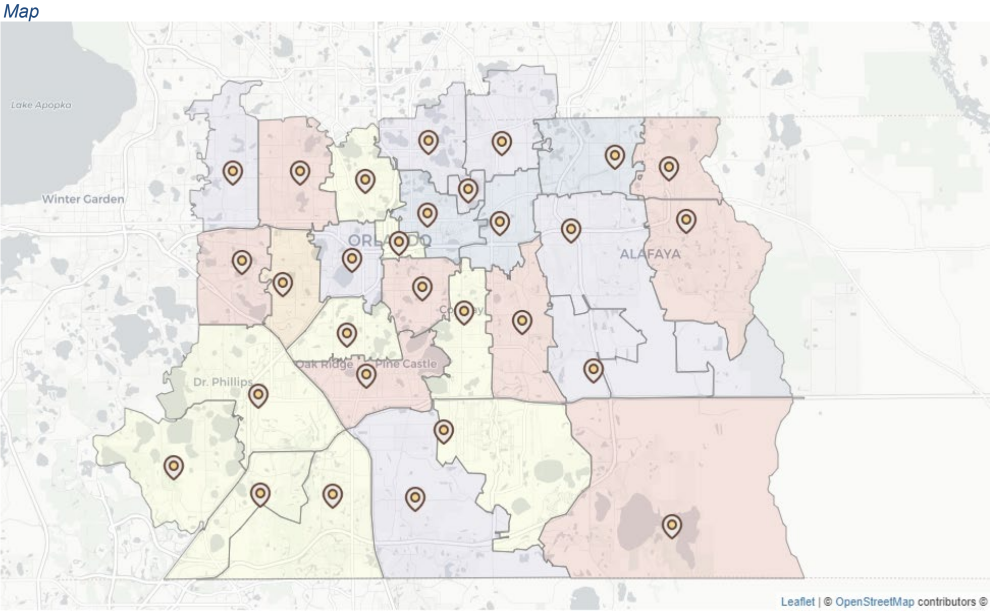
- At alpha = 0.05 significant level, **none** of the restaurant type coefficients is statistically significant. Therefore, we would accept the null hypothesis that they have a litter to next to litter effect on the performance score.

- At alpha = 0.10 significant level, we find that Quick and Greasy have a **significant negative relation** with the performance score; Asian and North American have a **moderate positive correlation** with performance score.

The new findings align with early conclusions. We generate and confirm the following conclusion based on the findings above, with respect to the dataset and methods limitation:

- Restaurant types, in general, are **less significant** than community features to correlate with performance scores.
- There is a **compound effect between population density and population**. There is a positive correlation between population density and performance for almost all restaurant types, which cohere to our common sense and daily observations. However, population feature is negatively correlated with performance throughout the entire analysis.
    - After examining Orlando's geo map for postal code regions with high population and density, the average income and the average house price are relatively low.
- Restaurant types have a noticeable yet not statistically significant correlation on performance, and the correlation level varies with the region's geo feature.
    - Compared to other types, Quick and Greasy show **resistance negatively correlated** with population, especially within the high population and population density areas. This is the most common type of restaurant in Orlando.
    - There are **no significant differences** in the correlation between a Cafes & Desserts restaurant and performance. We could conclude that the only influential factor is population density: **the higher, the better**. This is the second typical restaurant in Orlando.
    - North American restaurants show a **positive correlation** with performance among almost all postal code regions. This is the third most common type of restaurant in Orlando.
    - Asian restaurants have a **relatively high positive correlation** for almost all regions. However, the negative correlation with population is significantly above average. We might need an additional dataset to examine this pattern.
    - **Ottoman Cuisine restaurant is discovered to be the most fragile type.** The positive correlation between average income and the average house price is relatively weak. The negative correlation with population is average, plus it is the only type that has a negative correlation with the population density, unexplained at the moment. This is the least typical restaurant in Orlando.

Figure 10 Orlando Map



Limitations and Future Directions

Dataset Completeness

- **Inconsistency of the timespan among data for each restaurant:** When we examined the structure and distribution of the checkin.json, we noticed that the dataset is of subpar quality, and its incompleteness is beyond our expectations. The starting dates and the ending dates of the check-in data are different for almost every restaurant. The check-in amounts are dramatically different among restaurants. We tried to use the total day sum to normalize the check-in volume, but these methods could also be essentially inaccurate if there is a systematic or biased pattern in the check-in data collection.
- If a **considerable dataset** is available, we could conduct various, more complicated analyses on each level of our topics. If we were conditioning on specific features and clustering on the current dataset, the data size of each group would be at the edge of adequate for any statistical inferences. With the limitation of the Yelp dataset, our analysis could only provide a **less robust and effective description and conclusion** of the zip code areas that have insufficient restaurant data. However, we could still generate insights and recommendations for these areas based on general restaurant types analysis and the result of other zip code areas with high similarity.
- **Compound relationships** between population, population density, and restaurant performance: "Correlation does not indicate Causation." After excluding the population factor, we observed strong positive correlations between population density and restaurant performance score, consistent with our common sense. When combined with population density, we believe that the population feature contains certain hidden variables that are associated with income or wealth level. More data is required to draw further conclusions.

Measurement Accuracy

- **Mean:** We confirmed the non-Gaussian distribution of the data and examined the results using the median rather than the mean. The results were not drastically different, so our non-quantitative conclusion still holds. However, we could conduct further analysis to compare the results from the median or weighted-average metric. We decided not to go further down this topic regarding adherence to the primary goal.
- **Performance index score:** The rudimental thought was to approximate the restaurants' revenue. The straightforward formula is:

  $$Total\ Sales\ per\ Day = Sales\ per\ Bill \times\ Volume\ per\ Day$$

  We did not accomplish the approximation due to insufficient data to approximate the revenue per bill. We attempted to use Tips.json to approximate the average sales, using an assumption of average tips percentage of 15%. However, this could yield significant bias and inaccuracy, for many restaurants have no or next-to-no tips, such as fast food, some cafes, and so on. Various attempts proved inefficient to fix the issues above; therefore, we decided to formulate a qualitative performance index score to present the performance ranking and the intuitive magnitude of performance differences.

  We will explore and collect further and extended data that would allow performing better approximation, especially the approximation profit rather than revenue. For instance, among all restaurant types, the Quick & Greasy type is the most common regarding worst performance correlation. Fast food and similar restaurants have a higher profit margin (7%~9%) than do full-service restaurants (3%~5%). This information was not reflected in our analysis because an adequate dataset is unavailable for public use.

Ethical Considerations

All datasets were imported and collected from various public platforms or resources. After examination, we determined no data aggregation effect; therefore, we could consider no notable privacy concerns. However, the datasets are biased by nature for their underlying skewness toward restaurants and users with high internet technology adoption levels. Unfortunately, we have no resources or toolkits to examine further the dataset's robustness concerning its collection and processing. Thus, we must assume the data is sufficient to provide insights with acceptable accuracy and variances.

Statement of Work

Liwen Huang	Ziqian Wang
● Basic data cleaning and manipulation. ● Orlando, FL population analysis. ● Orlando, FL Zillow house index analysis. ● Orlando, FL neighborhood analysis and visualization. ● Restaurants feature engineering. ● Drafted final report. ● Generated the interactive report.	● Restaurants data manipulation. ● Restaurants exploratory data analysis and visualization. ● Restaurants feature engineering. ● Restaurants correlation analysis ● Restaurants correlation table visualization ● Comprehensive analysis and visualization for stakeholders ● Drafted final report

References

National Restaurant Association. (2021, January 26). *Association Releases 2021 State of the Restaurant Industry Report.* https://restaurant.org/news/pressroom/press-releases/2021-state-of-the-restaurant-industry-report

Kelso, A. K. (2021, January 26). *U.S. Restaurant Industry Finished The Year $240 Billion Below Pre-Pandemic Sales Estimates*. Forbes. https://www.forbes.com/sites/aliciakelso/2021/01/26/the-us-restaurant-industry-finished-the-year-240-billion-below-pre-pandemic-sales-estimates/?sh=4f15c624ebfa

Panasonic. (2020, November). *How COVID-19 is transforming Food Services & Food Retail?* https://ftp.panasonic.com/industries/food-retail/PNA_Food_Tech_Report.pdf

Huang, Liwen (2020). *Capstone Project - The Battle of Neighborhoods.* https://github.com/alisonhuang1988/Coursera-capstone/blob/master/Capstone%20Project%20-%20The%20Battle%20of%20Neighborhoods.ipynb

Cheema, Jehanzeb R. (2014) "Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research," *Journal of Modern Applied Statistical Methods:* Vol. 13 : Iss. 2 , Article 3.

Schafer JL. Multiple imputation: a primer. *Stat Methods in Med.* 1999;8(1):3–15. doi: 10.1191/096228099671525676.

Neighborhood. (2020). In *Merriam-Webster's Dictionary and Thesaurus* (p. 760). Merriam-Webster Inc.

Luca, Michael. "Reviews, Reputation, and Revenue: The Case of Yelp.com." Harvard Business School Working Paper, No. 12-016, September 2011. (Revised March 2016. Revise and resubmit at *the American Economic Journal - Applied Economics.*)

Banerjee, S., & Poddar, A. (2021). Run-of-the-Mill or Avant Garde? Identifying restaurant category positioning and tastemakers from digital geo-location history. *Journal of Business Research, 130,* 436–443. https://doi.org/10.1016/j.jbusres.2020.01.060

APPENDIX

Appendix 1

The goal of our Restaurant Type feature engineering is to:

- Reduce the number of labels we need to deal with
- Reduce occurrences of label overlapping
- Construct meaningful restaurant type for further multivariate analysis

Our early approach was keeping the top 5 restaurants with most restaurant counts and combining the resting into 3-5 categories by specific rules. This approach is abandoned for two reasons:

- Nearly half of the top 2 (American New & American Traditional) types is the same restaurant of overlapping tags.
- The tags other than top 5 is too many (110+) to be summarized into 3-5 category if we do not want it to be too vague and lose meaning.
- The top 5 tags are too inclusive. A restaurant could have fast food, American New, American Tradition, Mexican and more tags simultaneously; it is difficult to perform any analysis with this level of compound effects.

Therefore, we desire a more distinct restaurant type, if not strictly mutually exclusive.

We first extract all tags from our object data frame and manually label them into two clusters, nationality type (American, Chinese, Indian, etc.) and business type (Steakhouses, fast food, Pizza, etc.).

Then, we referred to the category frame from Author/Researcher Syagnik Banerjee and Amit and Poddar's work and made some adjustments. For example, we added more tags into the Quick & Greasy and Ottoman Cuisine category to adjust to our dataset. In addition, we replaced the Mexican type with Latin America for the high overlapping between Mexican tags and other Latin American tags.

We could separate some primary Asian types as we have some very distinct and familiar varieties such as Indian food, Chinese food, Korean food, etc. The primary reason we pool all Asian (except the tags in Ottoman Cuisine) tags into one category is that there are indecisive differences in restaurant numbers for us to decide which tags should stand alone and which tags share join the 'other Asian' group.For example, we have Chinese tag and Indian tag at similar quantity level, Korean food, and Japanese food at same quantity level; and restaurant numbers of all four tags combined is less than the American New types significantly.

Finally, there are several tags we have put into the other type, for they cannot fit into any of our current categories and have too litter instances for further analysis.

Appendix 2

The primary approach of our approximate is originated from the conventional revenue calculation:

$$Restaurant\ Revenue = Cutomer\ Volumes \times Spending\ per\ Customer$$

We naturally split the equation and the approximation into two parts: the customer volume and the per customer bill/spending.

In the beginning, we would like to use the check-in data from Check-in.json as a measure of customer volume and the tips from tips.json as an estimate of the per check-in spending level. However, after throughout examination on the object datasets, this approach was discovered to be critically flawed due to the following issues:

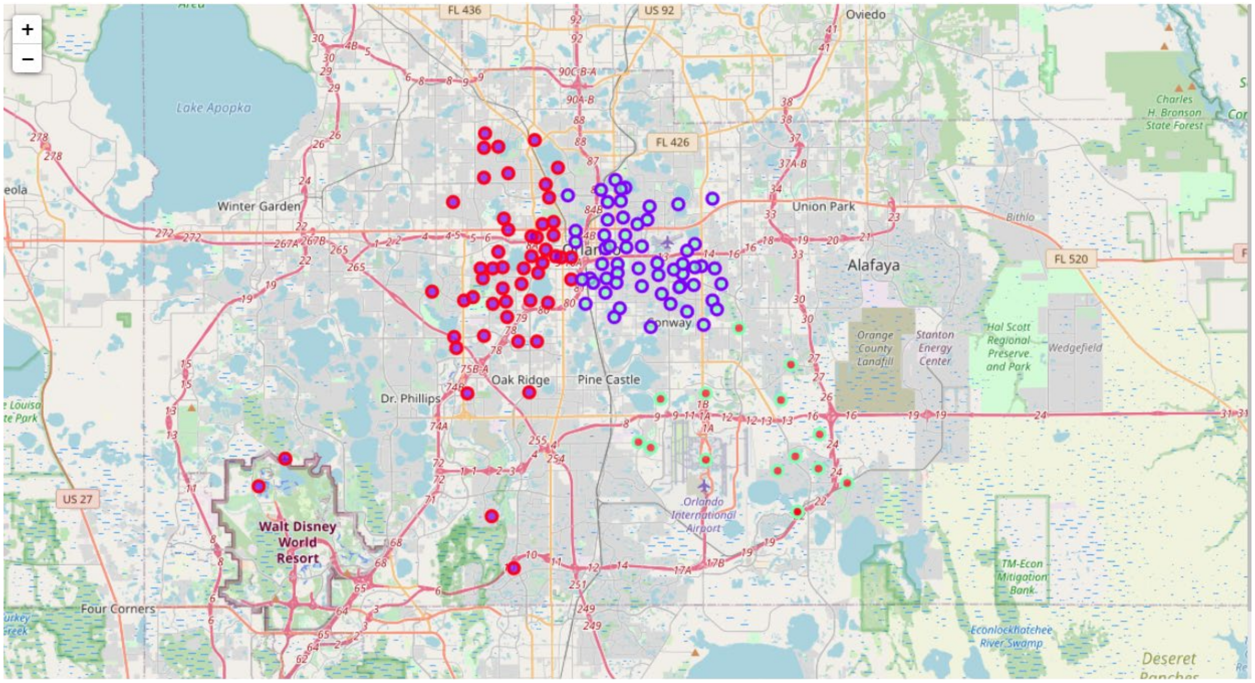
- Data completeness
    - We expected the checkin.json dataset could match and cover the records in the tips.json dataset. However, after mapping and join on the multiple different record id, we found neither dataset could provide a recording space that would contain the instances from the other dataset. If we only take the intersection of datasets, the valid intersection subset contains only an inadequate amount (less than 60% of the original dataset) of instances for further cleaning and manipulation.

- Data consistency
    - Besides the incompletes, data consistency appears to be another problem that is beyond our reach to fix.
        - Multiple restaurants that still open as of today while only have check-in records from 2017 to 2018;
        - A few restaurants have opened since 2016, yet all check-ins are centered around mid-2019, and nothing before or after;
        - A Non-fast food type restaurant has 170+ check-in records and 0 records of tips. And vice versa, there is a restaurant with less than 10 check-in records yet has more than 50 tips records.

- Unrepresentative metric
    - On top of the previous two issues we have revealed, we determined that it is inaccurate to approximate per customer spending leveraging the tips.json dataset. We could not establish a convincing connection between the tips given and the actual spending level. Even though we may have a reasonable average level (15%, for example), the customers' tipping behavior varies significantly among different restaurants. For example, people tend to tip substantially more in a Bar & Grill restaurant and significantly less (or even none) in a fast-food restaurant. In addition, we found the completeness of the tips.json is in question.

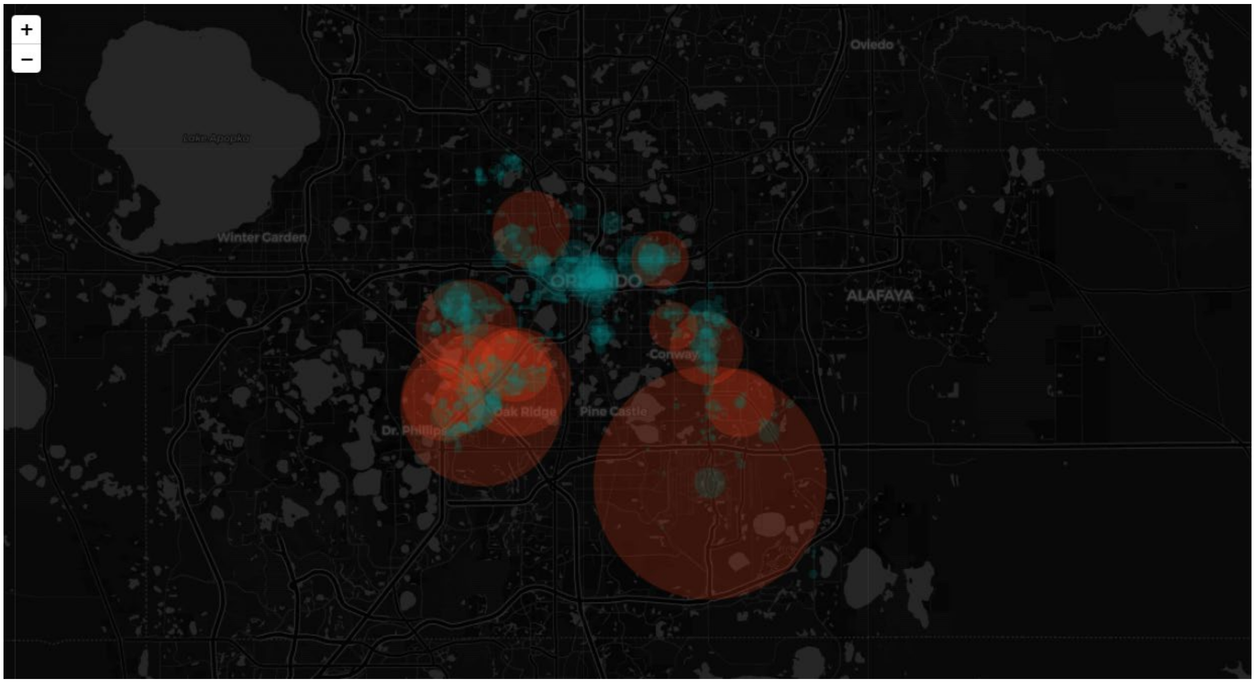
Therefore, we concluded that it is inapplicable to construct a quantitative measure of performance in terms of revenue estimates; we can generate an ordinal, qualitative metric or score to present the performance ranking.

Orlando Neighborhood Map

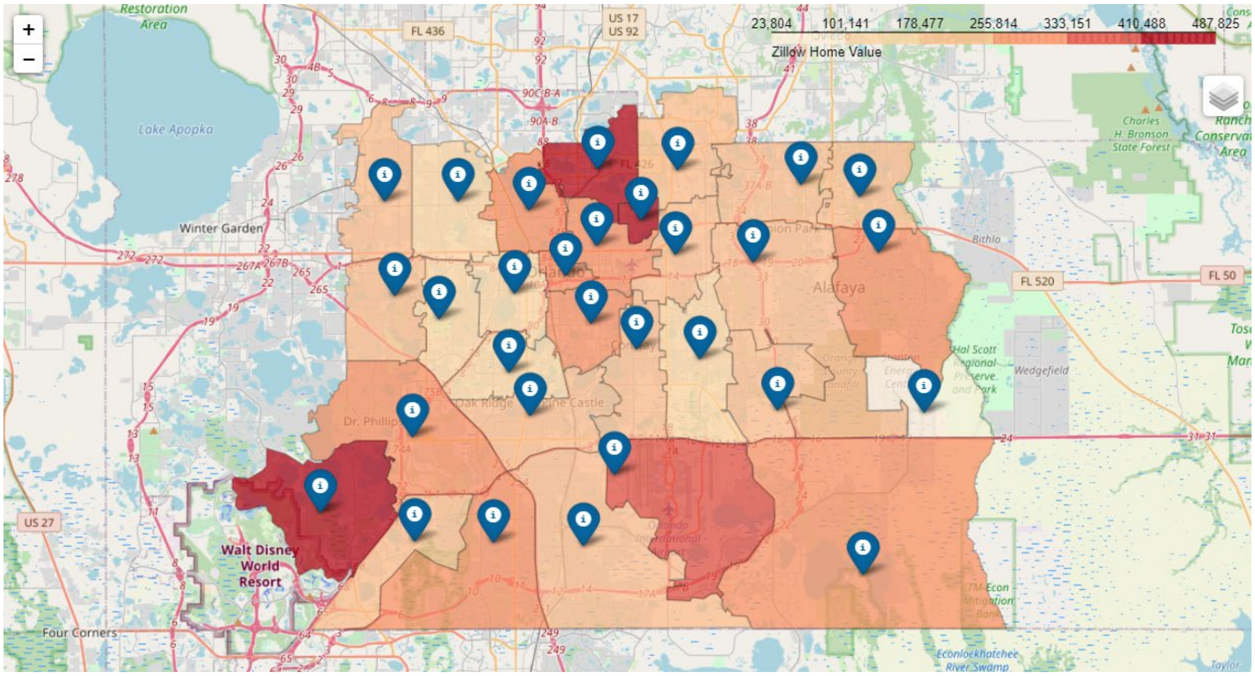


We may analyze our results according to the three clusters we have produced. Even though all clusters could praise an optimal range of facilities and amenities, we have found two main patterns. The first pattern we are referring to, i.e., Cluster 0, have more Park and Gym, the second pattern we are referring to, i.e. Cluster 1 and 2, have more restaurants and Lounge.

Orlando Crime Map

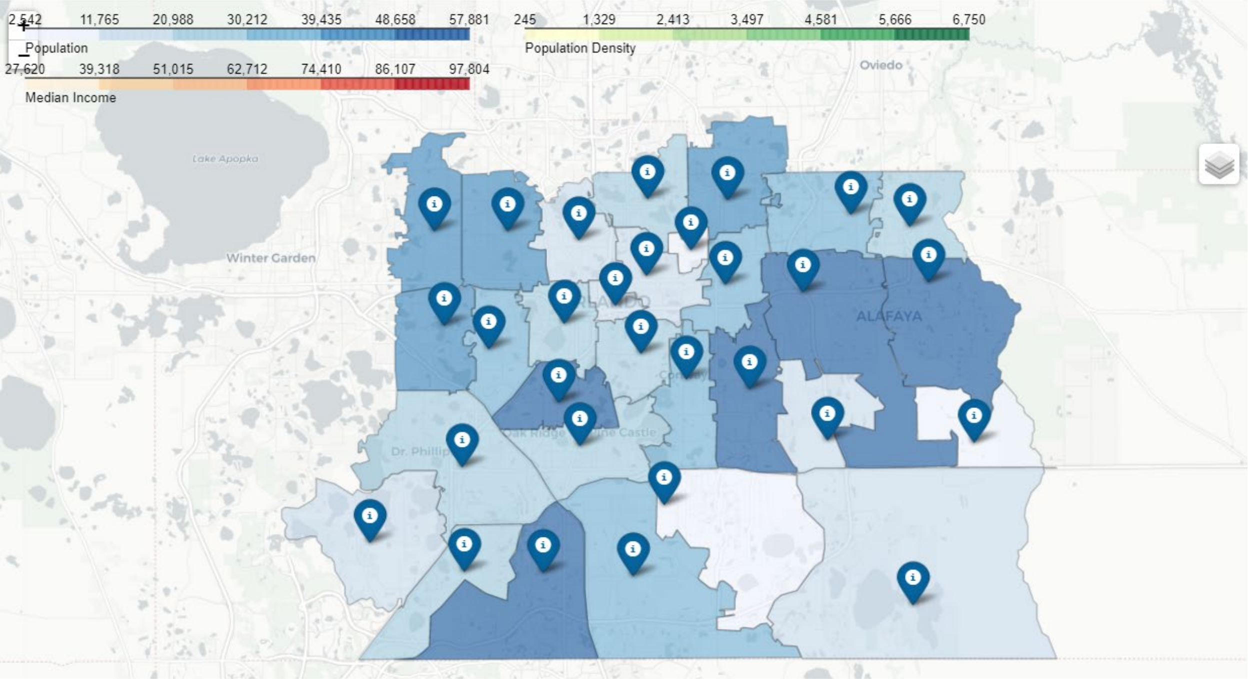


Orlando Home Value Map

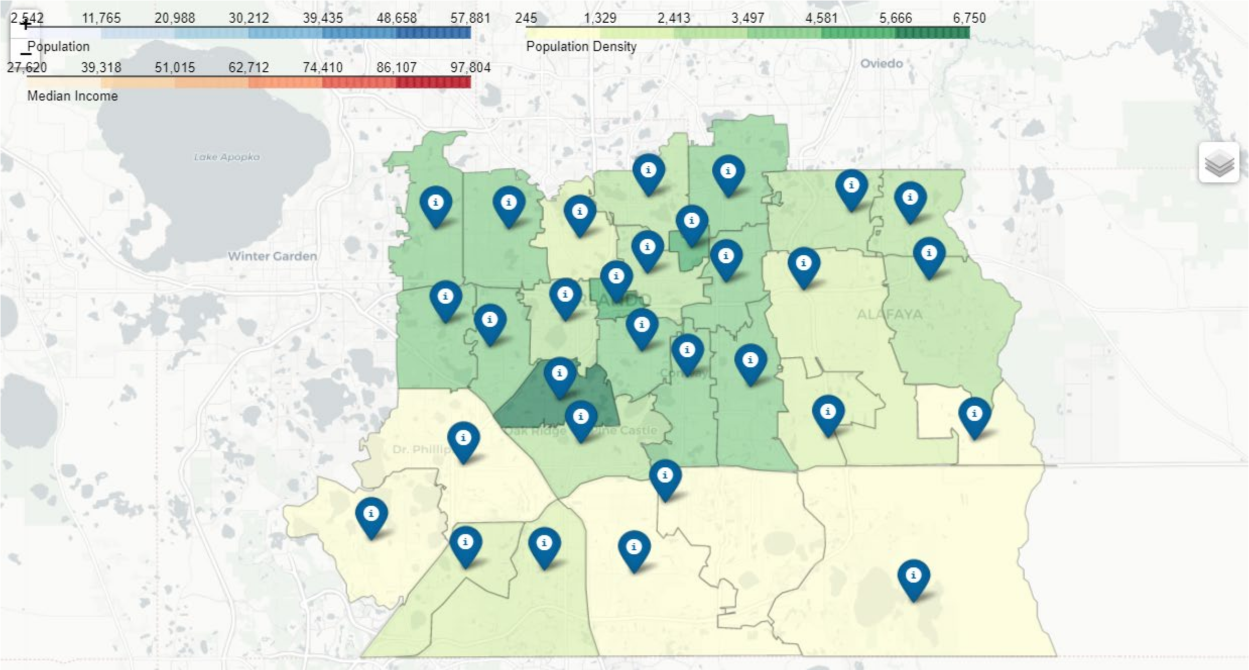


Orlando Geographics Maps

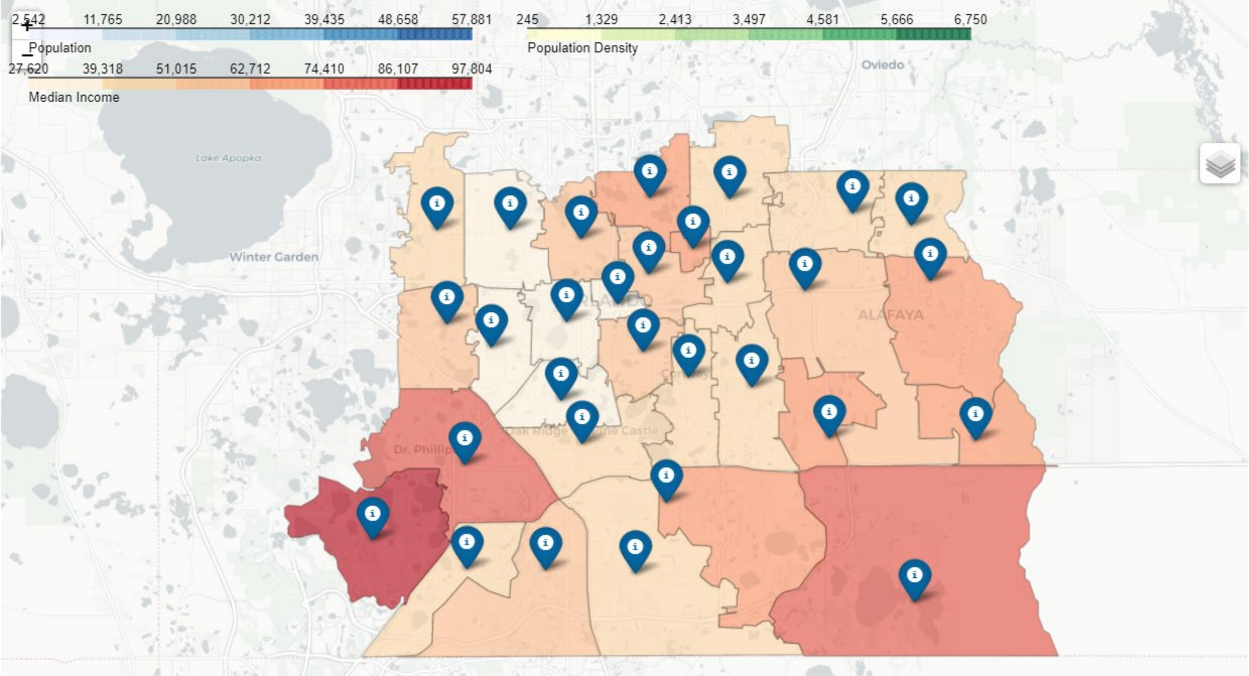
Population Map



Population Density Map



Median Income Map


2,542
11,765
20,988
30,212
39,435
48,658
57,881
Population
27,620
39,318
51,015
62,712
74,410
86,107
97,804
Median Income
245
1,329
2,413
3,497
4,581
5,666
6,750
Population Density
Oviedo
Lake Apopka
Winter Garden
ALAFAYA
Dr. Phillips
Oak Ridge
Pine Castle