Clustering Residential Areas Near Universities in Cebu City

Joseph L. Herrera

July 22, 2021

1 Introduction

It is common for students to pick a top university in earning their Bachelor's degree, often of which are located far from their hometown[1]. Because of this, students often find themselves spending a significant amount of time searching for a place to stay. When considering possible areas to stay for college, one would often consider commercial residential areas, such as dormitories, apartments, condominiums, and housing developments. The main consideration when choosing a residential area is proximity to the university, lessening travel time to and from the university. However, some students may also consider other factors like the common neighboring venues, such as the number of restaurants, parks, and the like nearby. In this case study, we explored nearby residential areas to some of the universities in Cebu City. we then clustered these residential areas based on their distance to the nearest university and the frequency occurrence of various nearby venue categories. The results of this case study could prove to be beneficial to students in making the right decision when choosing a suitable place to stay for college. Variations of this case study could be used to cater specifically to one's needs, but is beyond the scope of this paper.

2 Data Acquisition and Cleaning

2.1 Data Sources

In this case study, we retrieved the necessary location data from Foursquare's Places API[2], which offers real-time access to Foursquare's global database of venue data and user content. Some use cases of Foursquare's Places API include searching for venues based on a query or category with a specified radius of a given latitude and longitude, or getting venue recommendations in an area, or even see the trending venues given the time of day. Furthermore, one can even retrieve statistics for a specific venue, and even retrieve information about similar venues and users who liked a venue. To our knowledge, there is currently no publicly-available structured collection of information about the various universities in Cebu City. This being the case, we used the Places API to search for the various universities in Cebu City, wherein we can determine the name, coordinates, and other information about each university. From there, we can use the Places API to determine nearby residential areas to each university, and also information about nearby venues/locations to each residential area.

2.2 Data Collection and Data Wrangling

Data wrangling is first performed after the first call using the Places API to determine the information about different universities in Cebu City. In this step, domain knowledge about the local information of Cebu City is heavily relied on to remove any trivial duplicates, like "Cebu

Doctors' University" and "CDU" which refer to the same university, adjust the coordinates, and remove non-universities that were incorrectly categorized as a university. The resulting dataset contains some of the universities located in Cebu City and its corresponding latitude and longitude coordinates. Afterwards, the Places API is used to determine nearby residential areas to each university. In particular, we searched for the locations with the "Housing Development" category or "Residential Building (Apartment / Condo)" category from the Places API Venue Categories[3]. Duplicates were then removed, relying again on local information about Cebu City. Lastly, the Places API was used to determine nearby locations and venues to each residential area. Seeing that these locations could have different names but fall under the same category, we decided to retrieve information about their categories and coordinates only.

3 Methodology

3.1 Feature Selection and Feature Engineering

In this section, we will extract some features about each residential that we would like to further explore and work on. We will also perform exploratory data analysis when necessary. We first begin by taking into account the fact that we have information about the coordinates of both the univeristies and nearby residential areas in Cebu City. Knowing this, we could calculate the distance of each residential area and determine which university is closest to it and the distane to that university. We do this by using the haversine formula[4]. The assumption here is that we treat Earth as a perfect sphere. There formula is expressed as

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\left(\varphi_1\right)\cos\left(\varphi_2\right)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{1}$$

where φ_1 and φ_2 are the latitudes of points 1 and 2 measured in radians, λ_1 and λ_2 are the longitudes of points 1 and 2 measured in radians, and r is the radius of the sphere. In our case, we treat r as the volumetric mean radius of Earth measured in meters, i.e. $r = 6371000$ m. We would like to consider the surrounding venues to each residential area. This could give us an idea as to what locations are close to each residential area, and therefore an idea of what kind of neighborhood it belongs to. To do this, we utilize the Places API by Foursquare and use the explore endpoint[5] of the API to see venue recommendations near a given latitude and longitude.

3.2 Exploratory Data Analysis

After the necessary data have been retrieved, exploratory data analysis (EDA) is performed better understand the dataset being dealt with. In particular, the following are looked into:

- Distribution of distances to the nearest university through the use of a box plot.
- Top 20 venue categories near all residential areas based on frequency of occurrence through the use of a bar plot.

As seen in Figure 1, it is clear that many of the residential areas are within 500 meters to the closest university. Tearest residential area is only about 50 meters away, whereas the farthest is over 1.3 kilometers away. To be more precise about the statistics of the distances to the nearest university, we generate a table listing the minimum and maximum values, quartiles, mean, and other descriptive statistics of the distance attribute, and can be seen in Table 1.

We then proceed to visualize the top 20 venue categories near all residential areas based on the frequency of occurrence, which refers to the regularity with which something happens[6].

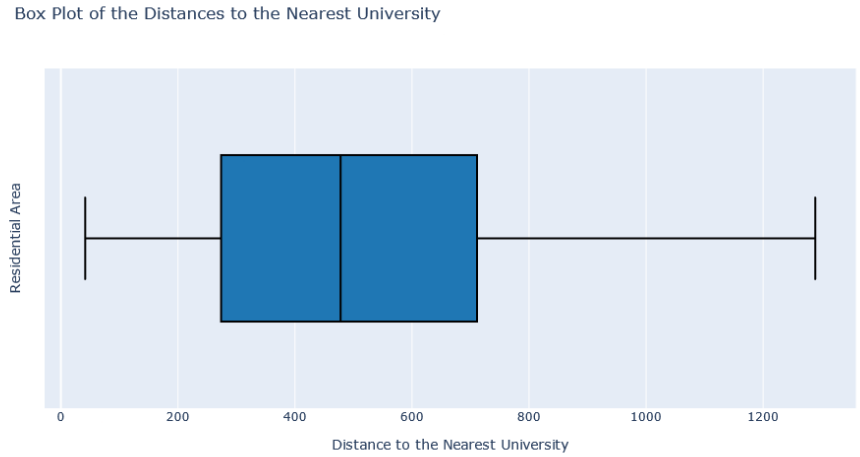


Figure 1: Box plot of the distances to the nearest university

Table 1: Descriptive statistics of the distances to the nearest university. All values are measured in meters.

	Count	Mean	Std	Min	25%	50%	75%	Max
Distance	257	504.3	277.7	41.6	274.0	478.0	710.5	1289.2

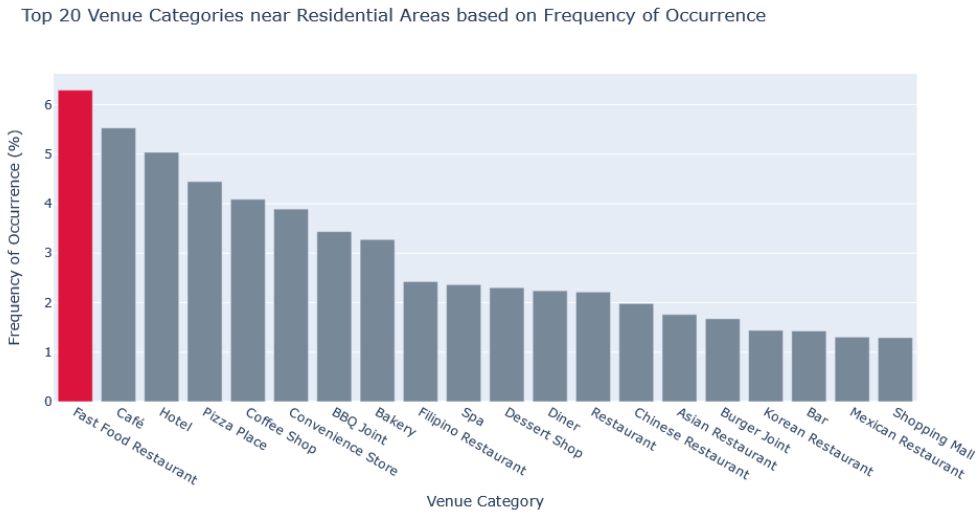


Figure 2: Top 20 venue categories near residential areas based on frequency of occurrence

In Figure 2, it is evident that many of the venues near residential areas fall under categories that are food-related, suggesting a correlation between residential areas and eateries. Of all the listed venue categories, we see that the fast feed restaurant is the highest, with a frequency of occurrence rate of 6%.

3.3 Model Training

The feature set for training included the distance to the nearest university and the frequency of occurrence of nearby venue categories. The K-Means Clustering algorithm was used as the model, wherein the elbow method[7] was use in selecting the best value of K. Feature scaling was utilized on the distance attribute using min-max normalization[8], given by Equation (2). In Figure 3, we see that the optimal value for our K-means clustering algorithm would be $K = 5$.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

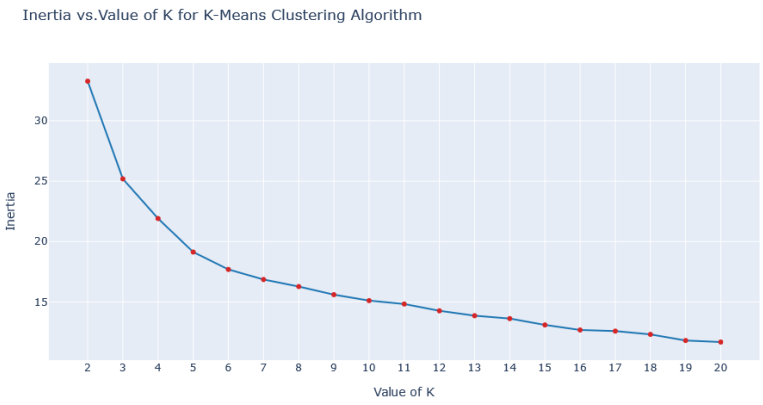


Figure 3: Inertia vs value of K for K-Means clustering algorithm.

4 Results and Discussion

After labels were generated from the 5 clusters, visualizations are generated to better understand the similarities of residential areas within each cluster, and the dissimilarities between different clusters. A Folium map was generated to visualize the different clusters on a map, box plots were generated to compare the distribution of distances to the nearest university for each cluster, and bar charts were generated to visualize the top 20 venue categories based on frequency of occurrence for each cluster.

In Figure 4, we can easily visualize the distribution of the residential areas in Cebu City based on the cluster they belong to. It can be seen that the distance to the nearest university is an important factor when clustering the residential areas.

Key takeaways from Figure 5 are:

- Cluster 1 and cluster 5 are very similar based on the distance to the nearest university. This suggests that residential areas belonging to either cluster should be given high priority when selecting a place to stay that is near a university.
- Clusters 4, 2, and 3 follow in decreasing priority, respectively.
- The clustering algorithm may have separated the clusters 1 and 5 based on the nearby venues to both clusters.

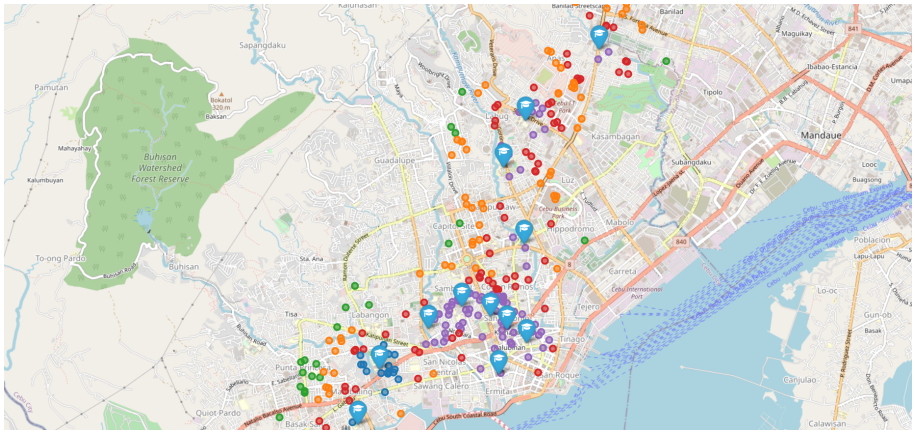


Figure 4: Residential areas in Cebu City per cluster.

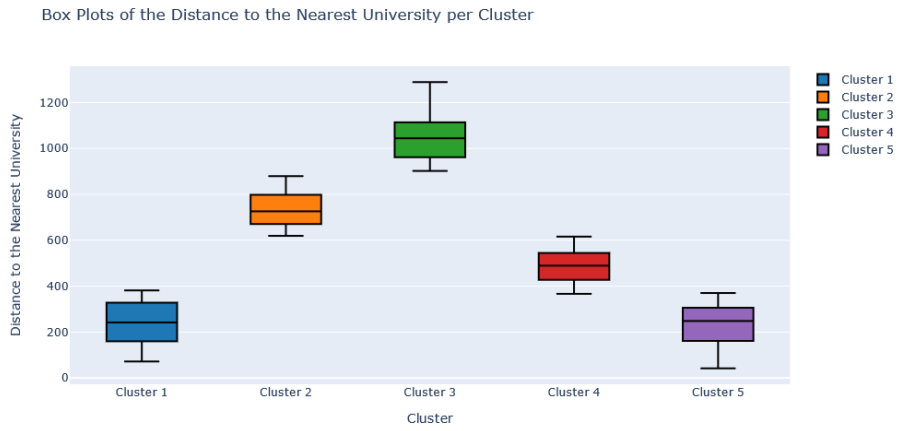
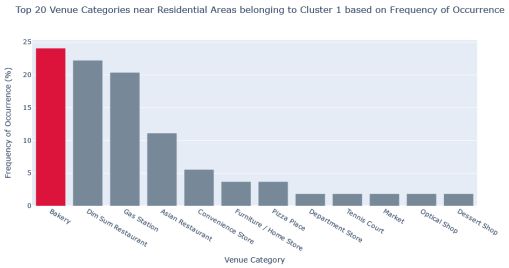


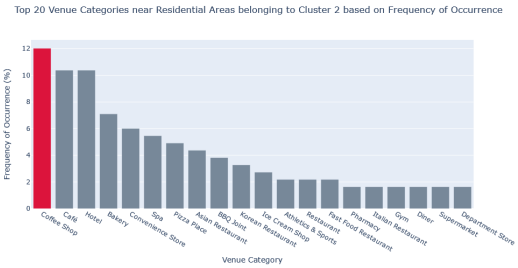
Figure 5: Box plots of the distances to the nearest university per cluster.

Key takeaways from Figure 6 are:

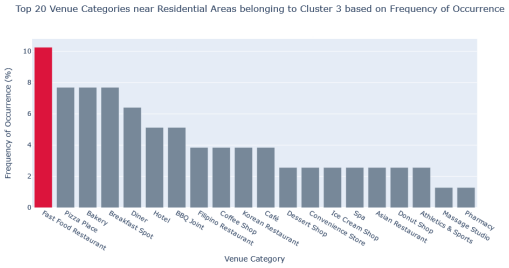
- There are fewer unique categories of venues near residential areas belonging to cluster 1 than those belonging to all other clusters.
- Eateries were the most frequently occurring venue categories, suggesting that there may be a pattern between residential areas and eateries.
- The most common eateries include fast food restaurants, bakeries, pizzarias, and cafés.



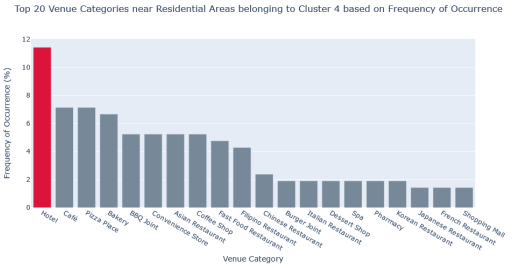
(a) Top venue categories near cluster 1.



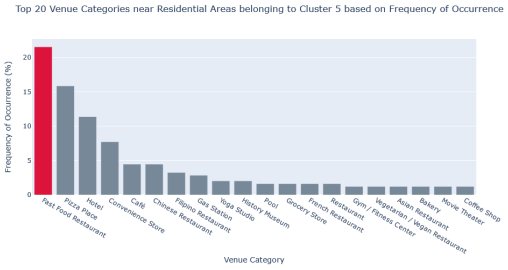
(b) Top venue categories near cluster 2.



(c) Top venue categories near cluster 3.



(d) Top venue categories near cluster 4.



(e) Top venue categories near cluster 5.

Figure 6: Top venue categories for each cluster based on frequency of occurrence.

5 Conclusion and Recommendations

Clustering of residential areas near universities based on their distance to the nearest university and frequency of occurrence of venue categories was studied in this project. Because there was no publicly-available information about the universities in Cebu City together with its coordinates and the number of students enrolled, Foursquare's Places API was used to retrieve location data about the universities in Cebu City. Data wrangling was then performed to clean that retrieved data, relying heavily on local knowledge about the universities in the city. Furthermore, location data was also collected about the residential areas near the universities, dropping any duplicates. Once done, feature engineering was performed to determine the closest university to each residential area and the corresponding distance, measured in meters, to that university. Location data was also retrieved about nearby venues to each residential area, to which the frequency of occurrence of each venue category was calculated. Therefore, the feature set for training included the distance to the nearest university and the frequency of occurrence of nearby venue categories. Model training was then performed using the K-Means Clustering algorithm, wherein the elbow method was used to determine the best value of K. Visualization of

the residential areas based on the cluster they belong to was then performed using Folium. Box plots were also generated to compare the distances to the nearest university per cluster. It was evident that clusters 1 and 5 were closest to a university, showing no statistical difference in their distances. Clusters 4, 2, and 3 followed with increasing distance to a university, respectively. Bar plots were also generated to determine the top 20 frequently occurring categories of venues near residential areas per cluster. It was evident that eateries were the most frequently occurring venue categories, such as fast food restaurants and bakeries.

In conclusion, students choosing to study in a university in Cebu City should give high priority on residential areas belonging to clusters 1 and 5 when choosing a place to stay. Comparing the average of the median distances for both clusters, we see that residential areas belonging to these clusters are only approximately 245 meters away from the nearest university. If the average walking speed of a young adult is 4.83 km/h, then that would translate to approximately 3 minutes in walking distance.

It is highly recommended to perform further studies focusing on the correlation of eateries and residential areas. It is also recommended to further clean the datasets used, add more universities, and explore more venues near residential areas. One could also use more general categories for the venues, e.g. fast food restaurants and bakeries can both be categorized as eateries. Other algorithms may also be used to compare different means of clustering residential areas.

References

[1] *Percentage of Out-of-State Students at Public Universities*. URL: https://www.collegexpress.com/lists/list/percentage-of-out-of-state-students-at-public-universities/360/ (visited on 07/20/2020).

[2] *Foursquare Places API*. URL: https://developer.foursquare.com/docs/places-api/ (visited on 07/20/2020).

[3] *Foursquare Places API: Venue Categories*. URL: https://developer.foursquare.com/docs/build-with-foursquare/categories/ (visited on 07/21/2020).

[4] *Haversine Formula*. URL: https://en.wikipedia.org/wiki/Haversine_formula (visited on 07/22/2020).

[5] *Foursquare Places API: Endpoints*. URL: https://developer.foursquare.com/docs/places-api/endpoints/ (visited on 07/22/2020).

[6] *Frequency of Occurrence*. URL: https://www.encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/frequency-occurrence (visited on 07/22/2020).

[7] *Elbow Method*. URL: https://en.wikipedia.org/wiki/Elbow_method_(clustering) (visited on 07/22/2020).

[8] *Min-Max Normalization*. URL: https://en.wikipedia.org/wiki/Feature_scaling#Rescaling_(min-max_normalization) (visited on 07/22/2020).