**Capstone Project – Battle of the Neighborhoods** **Week 2**

Olufunmi Olugbode

April 2021

# **Executive Summary**

There are many factors to consider when purchasing a house, budget being one of the controlling factor. However, for this exercise, the many other additional considerations to make when looking for a home such as safety, economics, school district and lifestyle would be considered.

I am moving to Texas and I have family in Houston and Duncanville and using python, visualization tools and foursquare, these cities are evaluated to narrow down my preferred location.

Data is sourced from different websites and python is used to scrape the relevant tables regarding crime, economics, school district and lifestyle (restaurants and venues). Following, the data is cleaned to remove non numerical data, rename headers then merge with location geospatial data. Foursquare will be used to leverage location data to find venues nearby.

Results indicates that the Duncanville area have low crime rates and the average income is relatively comparable to Houston. However, the school districts are not highly ranked but there are private schools that are within the Dallas area that are accessible. Duncanville appears to be a small town with very few activities around, however, the neighboring cities do provide the preferred lifestyle.

# **Introduction**

Decision making, especially when trying to make huge purchases requires a lot of research and a huge amount of data based on factors specific to the individual. This task will take several hours and tons of printed paper, however, data science has streamlined these tasks and allows for more efficient and a quick turnaround.

## **Business Problem**

Using this opportunity to carry out a personal project pertaining to house shopping, this project will utilize sourced data, data science - visualization tools, and location data, to considered specific factors to narrow down a location. Two cities in Texas will be evaluated to select a location (Duncanville and Houston, TX). The top features considered are as follows:

* **Safety:**  Crime rates in Texas will be evaluated as a neighborhood that feels safe is very / extremely important.
* **School district:**  for future planning and family planning. In additional, poor school district can impact future resale value of the home
* **Neighborhood:** A Community where the inhabitants are professional and have similar lifestyle is preferred.
* **Proximity to shopping, services and/or leisure activities:** As an social individual is important that some amenities or venues that I frequently are within walking distance.

# **Data Acquisition and Cleaning**

When using data, analysis are only as good as the data that is being utilized. Therefore, this section will detail the data sources for the structured data and discuss pre-processing measures where data will be transformed into a format that is understandable for the business problem.

## **Data sources**

Most of the information is sourced from statistics within the past 5 to 6 years (2014 – 2020) and is detailed as:

* **Texas Crime Statistic:** Information is taken from the FBI website for the year 2016

Source: <https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-6/table-6-state-cuts/texas.xls>

* **Texas School District:** Information is taken from the Texas Education Agency assessments for the 2018-2019 school year.

Source: <https://www.schooldigger.com/go/TX/cityrank.aspx?t=name>

* **Neighborhood:**Median household income is sourced from U.S. Census Bureau, American Community Survey (ACS) and Puerto Rico Community Survey (PRCS) for a 5 year estimate. The median income considers inflation and estimates the household income for 2018.

Source: https://www.indexmundi.com/facts/united-states/quick-facts/texas/median-household-income/cities#table

* **Proximity to shopping, services and/or leisure activities:** Foursquare API will be used to leverage location data to find venues related to the neighborhood.
* **Geospatial Coordinates:** sourced from maps of the world.

Source: https://www.mapsofworld.com/usa/states/texas/lat-long.html

## **Data cleaning**

Data cleaning involves fixing incorrect, corrupted, incorrectly formatted, duplicate or incomplete data within a dataset. In addition, removing data that does not belong in the dataset. Data cleaning for this project involves mostly removing data that is not relevant to the business problem and an example is shown in Figure 3.1 and Figure 3.2. The following functions were applied to clean the data:

* Deleting irrelevant columns and rows: using the *drop()* function
* Changing header titles: using the *rename()* function
* Dropping null value columns to avoid errors: using the *dropna()* function
* Selecting specific rows: using the *iloc()* function
* New data frame with split value columns: using *string* *split()* function

Text, table, email

Description automatically generated

**Figure 3.1: Original Dataset**

**Table

Description automatically generated**

**Figure 3.2: Clean Dataset**

# **Methodology**

This section details the procedure used to perform the battle of the neighborhood analysis.

## **Web Scraping**

The internet is a rich resource for any field of research or personal interest. To perform analysis for the battle of the neighborhood, all the dataset is taken from the internet and to effectively harvest data, web scraping was carried out. Web scraping is a process of gathering information such as tables from the internet and python utilizes BeautifulSoup to assist with this task.

## **Data Visualization**

## **Generating Graphs**

Bar plot is a graphically representation that describes the comparisons between the different categories. The matplotlib in python provide the bar() function.

## **Foursquare API**

To find the popular places near the cities I have selected to purchase a house, one of the best way to visualize / explore the venues within a particular radius is to utilize Foursquare API in python.

## **Folium**

To visualize data on a map, the folium concept is utilized.

## **K means**

To group data point K means was carried out, using to the Silhouette analysis to find the optimum number of clusters. K-Means dataset we used Silhouette analysis to tell us there was a lot of similarity between neighborhoods and the most common venues.

# **Results and Discussion**

In order to understand the data, the results are place in a visual context so that patterns, trends and correlation can be visible. Based on my requirements to narrow down a location to purchase a home, I initially compared Houston and Duncanville as some of my family members reside in this area. However, given the population (see Figure 5.1) in Duncanville it is more likely that the social lifestyle I require will not be found in the area. Therefore, the surrounding areas close to Duncanville (Dallas and Irving) will be assessed going forward.

Chart, bar chart

Description automatically generated

**Figure 5.1: Texas Cities vs Crime Rate**

Concerning the safety of the neighborhood, Figure 5.2 shows the crime rate for the selected Texas cities vs crime rate and crime rate in Houston is a concern.

Chart

Description automatically generated

**Figure 5.2: Texas Cities vs Crime Rate**

Figure 5.3 shows that Irving has a better school district and the Dallas and Houston area educational performance are average in comparison. Duncanville comes in last, however, there are several private schools in the area shown XXXX.

Chart, bar chart

Description automatically generated

**Figure 5.3: Texas Cities vs School District Ranking**

Using Foursquare API, Table 5.1 is generated to show the popular venues in the area. The list of neighborhoods with Restaurant venues show that most neighborhoods were similar and the greatest concentration of restaurants was in the Dallas area. This might seem obvious as the population in Duncanville is considerably lower and will impact traffic for these services, so there appears to be correlation.

Households in Duncanville, TX have a median annual income of 55,224 US dollars in 2018 which is comparable to the income in Houston but less than the median annual income of 60,293 US dollars across the entire United States (XXXXX).This was surprising considering that population in Duncanville is considerable less than Dallas and fewer establishments are located here.

Since the crime rate was a concern in the Houston area, comparable educational performance and comparable median income, a decision was made to remove the Houston dataset before considering the venues in the neighborhood.

**Table 5.1: Texas Cities and Common Venue**

Table

Description automatically generated

Figure 5.4 plots the clusters and includes the optimum location within Dallas, Duncanville, and Irving city limits that fit my social lifestyle.

A picture containing chart

Description automatically generated

**Figure 5.4: Optimum Location**

# **Conclusion**

The data collected and the analysis gives some confidence in deciding what Texas city to purchase a home which is within the Dallas city limits. This allows for proximity to the social life I desire. Note that I am better informed about these Texas cities prior to carrying out this study, of course, with more dataset more can be discovered.