**The Battle of the Neighborhoods – Houston, Texas**

Diego González Cárdenas

August/26/2020

1. **Introduction**
   1. **Background**

Every single person has his/her own interests and elements of relevance when trying to find a place to live. As it is a really important decision, agents about to complete such a big change in their life will want to be comfortable and pleased by this change. Moreover, once moving to a new city, state or country, it is important to move into the best neighborhood that one can find and afford. Of course, there are important differences amongst the different areas in the same city of interest.

Houston, Texas is one of the most important cities in the US. Its privileged location, vast amount of culinary, sports and entertainment venues along with its medical and educational services, which are of the best in the world, make it a very attractive city to live. Houston is also a city with a lot of diversity, for this matter very different conditions may be true on the living areas of the city.

**1.2 Problem**

The best way to make the correct decision, of your next place to live is to be fully informed about what the different neighborhoods and areas of the city can offer. A new resident in a city needs to find a home where he/she can get on time to work, have proximity to the places or venues of his/her interest, among other priorities. For this project I will make an analysis which will try to find the best neighborhoods to look for an apartment or house taking in consideration the following characteristics: Proximity to attractive venues such as restaurants, bars, cafes and museums. Proximity to public transportation services and to hospitals/health care services. Also, it is important to live near education services/schools (for this project Universities will be the point of interest). Important aspects such as security, housing prices, among others are also important for such decisions. Nevertheless, such subjects are beyond the scope of this document. An important attempt will be made to construct a methodology which can be replied in any city in the world. I believe similar decision-making procedures are made regardless of the location in study.

**1.3 Interested Audience**

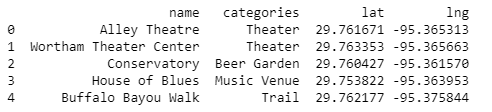
As the project will try to result in a methodology with replicability all around the globe, the target audience will be any person who is considering moving to a new city; any individual who wants to find the best location possible to live in their new city of residence. For this matter, this tool will hopefully be useful for a large amount of people.

1. **Data**

As previously mentioned in the introduction section, there will be several characteristics to take into account to build a proper methodology. For this same matter, there will be several datasets which will play an important role in this project.

* 1. **Attractive Venues**

For this section, the data used will be obtained from Foursquare location data service, thoroughly used through the Applied Data Science Capstone course. Since foursquare gives the relevant data from venues this will solely be our source on this occasion. Foursquare will retrieve a maximum of 100 venues around the indicated location (29.7589382, -95.3676974.) The data will be grouped by categories, no further relevant data manipulations are needed for this dataset. The final result for the data manipulation described was (showing only the first 5 venues):



* 1. **Public Transportation Venues**

For this section, we will take into consideration only the METRORail service in Houston. Although there are several other public transportation services in the city, for simplicity, our scope of study will center exclusively in this light rail system. With a very relevant presence all throughout Houston consisting of 3 lines, with a total of 44 Stations this can be considered as the most important service of transportation in Houston. It is important to mention that there are other 2 lines to be added on the near future to this system.

The location data of the stations from the 3 currently existent lines is gathered from the official METRORail website. The datasets for the 3 separate lines were found, the datasets will be treated separately to eliminate any columns which are not useful and to find the latitude and longitude points with the given Address from the dataset using Geopy. Datasets will be merged after treatment. If Geopy was unable to find such locations, they were ‘dropped’ from the dataset.

Merged and treated final dataset (incomplete for simplicity):



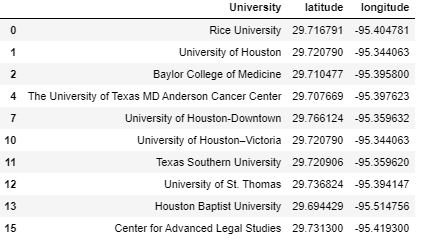
* 1. **Hospitals**

The database for this section was retrieved from the *Texas Health and Human Services* website. The data used is the **2019 DSHS/AHA/THA Annual Survey of Hospitals.** This dataset had the information of allegedly all the hospital or health services centers in Texas. Originally, the dataset contained the name of the Hospital, County, City, Name of administrator, Phone, Number of Beds, and type of ownership. Of course, most of this data is not useful for what we want to achieve. Data was transformed from a PDF to a CSV file, retrieved as a Pandas df, unwanted columns were deleted the whole data frame was treated to show only our sample of interest, the Hospitals in Houston. As done in the previous section, Geopy was used to retrieve the latitude and longitude coordinates of the hospitals according to their address. Locations which were not found were deleted from the dataset. Final dataset:



* 1. **Universities**

Lastly, for the Universities section I web scrapped in the webpage <https://www.unipage.net/en/universities/houston>, which unfortunately contained several non-relevant columns for this project, such as Rank, and the tuition price for a year for both a Bachelor and a Master’s degree. On such dataset there was only one column of relevant information: The name of the university. However, this was the only piece of information needed since this dataset was treated in the same manner as the METRORail stations and the hospitals. Coordinates were found with no problem thanks, once again, to Geopy.



* 1. **Mapping Houston Neighborhoods**

All of the elements mentioned in the Data section will be drawn into a map separately so we can visually see the location such elements on a map. Finally, web scrapping will be done in order to find a GEO JSON file which can allow us to see in a Folium Choropleth map the neighborhoods or boroughs in Houston. This will be done to see graphically which elements correspond to each neighborhood; this will enable us we to make accurate comparisons and conclusions to determine which neighborhood is most attractive according to our preferences.

1. **Methodology**

As mentioned in the Introduction section what we attempt to measure is the proximity or presence of our desired services and amenities in any given neighborhood. Moreover, our decision to label a neighborhood as a good decision to live in will depend on: Proximity and abundance of attractive venues, public transportation services, hospitals, and universities. On the other hand, an important question that will be answered is: Are there any patterns for the distribution of popular venues and public services across Houston? In other words, is there a relevant and observable relationship among the presence of public services and the number of venues nearby? And, does this connection presents a positive or negative correlation? Finally, my own thoughts and conclusions will be presented by the end of this work. With the intention to clarify any questions or inquiries that may still remain on the mind of the reader.

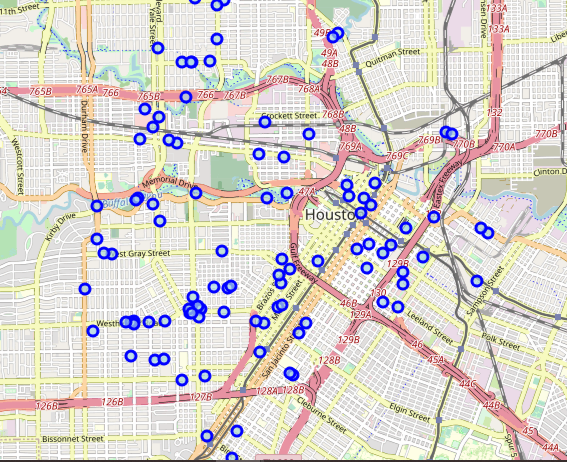
* 1. **Exploratory Data Analysis**

For all the relevant aspects to considered on for our project (all of them stated on the Introduction) a somewhat similar procedure was made. As previously mentioned, Folium was widely used in this project. A visual representation always serves as an important aid to draw relevant conclusions. Maps of all elements were depicted.

Once all the different maps are presented, a map including all the used datasets, with their respective neighborhood areas will be rendered. A map combining all the elements should be the best way to draw accurate and informed conclusions. Finally, relevant conclusions, connections and observations that we gather from our visualization tools will be presented and discussed at the end of this section.

**3.1.1 Venues**

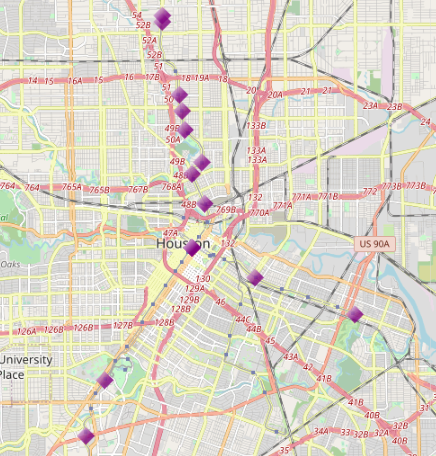
For the first map, when retrieving, treating and mapping the venues dataset, we can observe a somewhat unequal distribution of venues across Houston. There is seemingly no clear data distribution nor clusters, for the most part. There are somehow, some exceptions to this general observation, there are some segments of the map that can be considered as a relevant cluster of venues, namely there can be a formation of an important bundle of venues in Downtown Houston and other cluster Southwest of Downtown Houston. The previous comes as no surprise since there is always an important presence of important venues in the Downtown areas of any city. Nevertheless, the other possible cluster mentioned may be an interesting area to take a closer look in our analysis.



**3.1.2 METRORail Stations**

When mapping the location of METRORail Stations we can draw an important, yet somewhat predictable conclusion: the location distribution of this public transportation service shows that locations are actually following a line rather than showing the formation of clusters. Nevertheless, this is no surprise since such stations are planned to have a somewhat even distribution to reach most places in Houston.

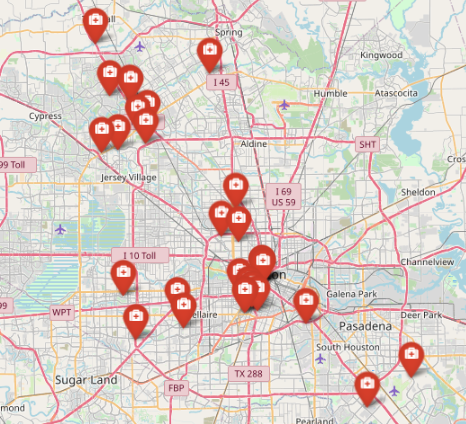
There were several important limitations when working with this data frame, nevertheless this will be thoroughly discussed in the **Limitations** section later on.



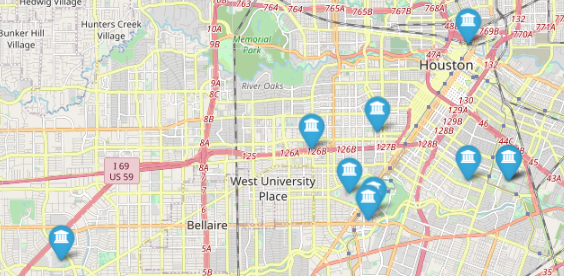
**3.1.3 Hospitals**

Other interesting map to share in this document is the hospitals map. Such map presented characteristics and give us completely different conclusions from the venues map. In this one, we can observe clear clustering for the location of hospitals in the city of Houston. There are several, really marked clusters on the locations of these health care centers. Which lead us to conclude that there are “hospital districts” on the city of Houston. Once again, we can observe an important cluster on Downtown Houston.

Such observation will truly have an impact on the decision to live in one or other neighborhood if what you are looking for is to live near a hospital. Nevertheless, with this visual representation we are unable to determine the possible causes for the formation of such clusters or if there is a relation with this clustering and the presence of other important services.



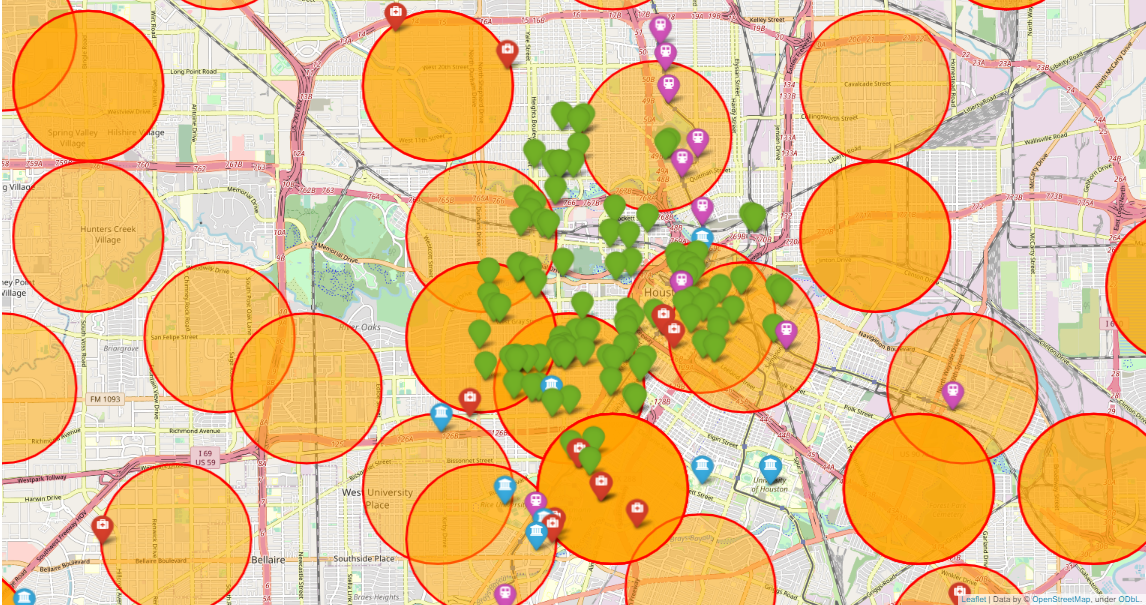
**3.1.4 Universities**

For universities we can observe that there is one important cluster having an important number of our observations, nevertheless the distribution apart for this cluster seems to be unclear and without a visible pattern. There are other duplets of universities which may be considered clusters and two outliers. This data frame by itself does now help us draw any conclusions. What will be interesting is to compare with other datasets to see a relationship of distributions or a possible answer as to why such universities are located where they are. 

**3.1.4 Putting all the pieces together**

Although the previously displayed and discussed maps don’t tell us a lot of information by themselves, it is the combination of all these elements that can lead us to draw important conclusions. Combining all the elements help us to realize both: Why some cities are distributed the way they are; and also, which neighborhood will represent a best place to live, which is the ultimate goal in this project.

First and foremost, I’ll give a brief explanation of the map: The green tags represent the location of the venues found by Foursquare. On the other hand, the purple tags which have a Metro cart on them represent the location of the METRORail Stations. The red tags which have a Medical kit represent the location of the hospitals. Blue tags represent the location of the universities. Lastly, the big orange circles represent the **approximate** location of Houston’s neighborhoods. This will allow us to determine which important venues or services belong to which neighborhood.



As I have previously mentioned several times, there is a really important cluster in Downtown Houston, there is an important presence of venues, and also the presence of public transportation, hospitals and universities. Although this is not a surprising conclusion, I was surprised because I originally believed that the most important clusters of hospitals and universities will be located in this area, and even though there is certainly the presence of this services in Downtown Houston, there are more important clusters in other areas of Houston.

A really interesting aspect which reinforces the initial statement of Houston being a very attractive place to live is the fact that METRORail stations are distributed in a way that no matter where you live, the proximity to Hospitals will never be an issue in the city. If we take a look on the neighborhood areas that have a hospital in them, almost in all cases we will find a public transportation station as well.

On the other hand, there is also an important frequency of neighborhoods which include both Hospitals and Universities. This will make such neighborhoods as good opportunities and probably the right decision to become their new place to live. Even on the cases were both elements seem to be outliers from their respective groups, we see this rule to hold true. Although our final map allows us to see a clear relationship on the location distribution of Hospitals and Universities, the direction of this relation is still unclear. With this information, whether the presence of Universities is a determinant for the construction of a hospital nearby, or the other way around, is still unclear.

The existence of venues, on the other hand, seem to be mostly unrelated to the presence of any other elements on our map. There are several neighborhoods in which we can observe a really high occurrence of venues and absolutely no presence of other services. Although in general the number of venues is much higher than any other feature, the map clearly shows us that there is not a high dependence or relation on whether a venue is established and the proximity to transportation, hospitals or universities. Rule which seems to hold false with all our other features.

1. **Limitations**

It is widely known that it is almost impossible to find the perfect data set for our project, and that in order to be able to appropriately use data, there are several modifications and cleanse techniques that a Data Scientist can use. My project was not an exception to this rule, and on the process of development, I encountered several speedbumps which surely bias or modify our results and conclusions. Such limitations can be opportunities for improvement in further investigations. I believe it is important to mention such limitations, an important part of our formation as Data Scientists is through trial and error. For this reason, I will thoroughly discuss the roadblocks I found on my experiment. Any suggestions or advice from the reader are welcomed and strongly encouraged.

First and foremost, foursquare only shows us 100 venues closer to the indicated coordinate. For this reason, we can see a great occurrence of venue presence in the center of the map, whereas there is no presence of this element as we go further away from our coordinates. This gives us the false illusion that the only venues in Houston are close to this location when the reality is that there are much more venues. Moreover, we get the false sense that there are many neighborhoods which do not have venues in them, when actually they do contain the presence of venues.

Maybe a possible solution would be to first get the location of all the neighborhoods and do a venue search for all the neighborhood center locations. This will ensure we get more than just 100 venues. Also, this will ensure that we get a more accurate distribution of venues. We will also have the presence of venues all throughout the map rather than only on the center.

Other restriction found was that in some data sets Geopy was unable to retrieve latitude and longitude points for a given Address; of course, on the pandas data frame this will show up as a *NaN.* Since there are no further way to treat such data (i.e. it would be completely inaccurate to do state such locations as the mean of all the other points); all the data points in which I encountered this error, were dropped. Unfortunately, this process of data mining takes a lot of important relevant information from our datasets; this of course, may lead to inaccurate results.

There was a dataset in which this problem was particularly abundant. The METRORail stations dataset was originally 3 datasets belonging to different rail lines which were merged together. Our merged, untreated dataset had 44 data points or stations. After treatment, and because Geopy was unable to retrieve the location of 21 of those stations, such number was reduced to only 23 data points. Of course, this has an important impact on our final results. On other datasets this problem was much less present. For instance, in the Hospitals dataset, there were only 2 Houston hospitals which needed to be dropped from the original dataset. The reason why Geopy failed to obtain the locations of such an important amount of METRORail stations is still a mystery to me.

Finally, I tried finding GEO JSON files which will allow us to render a Choropleth map with Houston’s neighborhoods with high accuracy. Although I did find some files on the web, I failed to implement them in my experiment. The alternative which I used was to web scrape for the names of such neighborhoods and their Zip Codes. After such procedure, I used Geopy and the **uszipcode** package in Python to obtain the location coordinates of such neighborhoods. Although the process of rendering the **location** of Houston’s neighborhoods may be accurate, I could only represent them as circles, rather than actually showing their shapes. Naturally this is a big opportunity of improvement for my project.

As mentioned in the introduction to this section: If there are any suggestions and advice to improve my project, please let me know.

1. **Conclusions**

Taking into consideration the limitations and problems I encountered while developing my project, we can still draw some important conclusions: As we can expect, since Houston is one of the most important cities in the U.S., there are many important venues and amenities located throughout the city. Nevertheless, an there is an important concentration of elements near the center of the city, Downtown Houston is a very attractive place to live. Nevertheless, we can expect higher apartment and housing prices in such location.

There are a lot of METRORail stations which are spread throughout Houston; so, if proximity or closeness to one of these stations is the most important point for your decision to move to a specific neighborhood, you really have many neighborhoods to consider. Also, the presence of public transportation stations has a close proximity to Hospitals, which is of course a big point in favor for this city. On the other hand, there are many neighborhoods which have the presence of both Hospitals and Universities. This makes Houston a really attractive city to live.

1. **Final Questions to Answer:**

In the introduction section, I mentioned three important questions which are to be answered with the development of the project: **Are there attractive venues nearby?** There are many attractive venues spreaded through Houston, although we were limited to only 100 venues, we covered a lot of terrain. Also, it is important to consider that due to the important and numerous presences of public transportation stations on Houston, proximity to a desired venue will not become an issue for people living in Houston. **Are there any patterns for the distribution of popular venues and public services across Houston?** Yes, based on our research, there is an important cluster of venues located around public services in the city. We can expect the same to hold true if we go further away from Downtown Houston. **Final recommendations from the information gathered.** My recommendation would be to live near Downtown Houston, nevertheless to avoid living in such neighborhood. I would decide to live either on **Neartown - Montrose or MacGregor neighborhoods.** They both have all the attractive features which we desire and which were described at the beginning of this project. Also, the cost to living in such neighborhoods will be somewhat lower than living on Downtown Houston, plus being less overcrowded.