Neighborhood Analysis in the Washington, D.C. Area for the Future Amazon HQ2

Capstone Project - Battle of the Neighborhoods

May 4, 2020

Contents

1 Introduction and Business Problem

On November 13, 2008 Amazon announced their plans to establish a second headquarter office (named HQ2) in the United States, located in New York city and Washington D.C. ([1, 2]). This represents a great opportunity for the greater D.C. area to experience growth and to attract a significant number of both skilled workers and other businesses, which are expected to populate the DMV (D.C., Maryland and Virginia) area.

The National Landing in Arlington County was chosen as the future site of HQ2. This county houses multiple national landmarks (Arlington National Cemetery, the Pentagon), a thriving business economy (with strong presence of both private companies and federal agencies), and along with the neighboring Alexandria routinely top national rankings of best places to live and retire. All these characteristics make it a very desirable area to live and work on.

However, the D.C. metropolitan area already suffers from severe issues that are likely to worsen with the expected heavy affluence of workers. Specifically, this report will look into the housing prices (rent and sales), and the commuting time. This last issue is a major concern in the area, as the D.C. area commute regularly ranks within the national top 5 work commute times (regardless of whether the commute is analyzed in terms of actual time spent commuting, or just the time spent sitting in traffic) ([3, 4]).

For all the reasons above, this document aims to analyze the different regions in the DMV area in order to help understand what is the current situation prior to the arrival of Amazon's HQ2, in order to be able to better plan for the changes that the arrival will trigger, and in this way, allow all the interested stakeholders, from city planners to realtors and small business owners, to identify potential issues and areas to improve.

1.1 Goal

The goal of this analysis will be **to improve the understanding of the DMV area in terms of amenities, commute time, and housing costs**. These are significant aspects that will drive the decision of many of the new workers that will arrive to the area when deciding where to settle down, which will, in turn, affect back to those same issues (amenities, commuting, and housing costs).

It shall be noted tha the goal of this analysis is not to perform a prediction on trends or metrics, as the arrival of HQ2 is expected to be significantly disruptive, and therefore using current trends to forecast the future seems not appropriate. Therefore, we will focus this report on the analysis and understanding of the current state of the DMV areas.

1.2 Objectives

In order to achieve the main goal described previously, we will be performing descriptive analyses on the DMV area. Especifically, the follwing objectives will be targeted:

- *Amenities Similarity*: In order to identify neighborhoods and areas that offer similar amenities to National Landing, we will look for areas that offer similar types and quantities of amenities in the whole DMV area.
- *Commute Time Distribution*: As National Landing sits inside the I-495 Beltway, and is bordered by the Potomac river, the time to commute for new potential workers may be significant. To analyze this, we will study not only the average expected commute time from each neighborhood or area to National Landing, but we will also focus the analysis on finding patterns regarding the distribution of this commute time.
- *Housing Costs*: The study of the costs of renting and purchasing a home in the different areas will also be analyzed, and, as with the commute time analysis, patterns and distributions will be displayed and analyzed.
- *Desirability Distribution*: After individually analysing each of the features previously described, a further study will be performed attempting to combine all these individual aspects in a single metric that tells us how similar in terms of desirability the different areas are.

2 Data

In this chapter we will review the data that will be used for achieving the goals described in Chapter 1. We will look at what type of data is needed (including the formats, ranges, and types), the sources we can use, the understanding of how the data fits into the project, and any preparation that may be needed.

2.1 Data Requirements

In order to achieve the goals of this project, we need several pieces of data:

- **Geographical Description of the Areas to Analyze**: The first piece we will need is the geographical definition of the areas that we will be using for the analysis. These areas should be separate areas (i.e. with no overlapping) that provide full coverage of the areas of District of Columbia, Maryland, and Virginia. Furthermore, we will need to be able to replace each area with a point of reference that can be used to query FourSquare for amenities. Given the large amount of National Parks and water bodies in the area, we will need to be careful to ensure that some areas do not get misrepresented because of the presence of one of these areas (which, for example, would have no amenities nearby if querying from the center of a National Park).
- **Amenities in Each Area**: In order to simplify the analysis, we will be focusing on the top-10 most popular types of amenities in an area, as provided by FourSquare.
- **Commute Times**: The analysis of the commute times will require the availability of the required time to commute from one point to the expected location of HQ2, as it is usually reported by apps like Google Maps, or Waze. Given that this expected commute time may vary significantly from one day of the week to another (e.g. depending on school hours or telecommuting in the area), we will need to consider the average commute time across a number of days and hours, to get a better representation of this metric. In order to simplify the analysis, we will look only at car commute times only, leaving walking, biking, or public transportation for a future extended analysis.
- **Housing Costs**: Finally, we will need to acquire a dataset that provides us with information regarding the cost of renting or buying a home in each of the areas analyzed. This information should be normalized to account for differences in terms of size, number of rooms, etc.

2.2 Data Acquisition and Cleaning

2.2.1 Areas Definition (ZCTAs)

The main issue in collecting the data for this project may lay in the definition of the areas for the analysis. While D.C. defines neighborhoods, which seem to be a good partition for this type of analysis, it tursn out that the definition of those neighborhoods is not clear, and there are overlaps ([5]). For this reason, and in order to have a division that can apply not only to D.C. but also to Maryland and Virginia, we decided to proceed with divisions based on ZIP codes tabulation areas (ZCTA), as defined by the U.S. Census Bureau ([6]). These ZIP code tabulation areas are full-coverage, non-overlapping areas that cover all of the U.S., and we can use the datasets available from the same U.S. Census Bureau to get the coordinates from a representative point of each area. The datasets are provided as Shapefiles, which can be easily converted to CSV files using any GIS application. For this documentation, we will use QGIS ([7]), a free open source GIS package available for Windows, Linux, BSD, and macOS.

When we load the downloaded shapefile into QGIS, we see the graphical representation of the geographic objects in the file, as seen in Figure 2.1. We can see that there are over 33 thousand entries (i.e. ZIP code areas) in the file. For each one of them, besides the geography coordinates we have several data fields. From these, we are interested in

ZCTA5CE10 (the ZIP code for each area), and *INTPTLAT10* and *INTPTLON10* (the latitute and longitude of an internal point at or close to the center of the area). The rest of the fields we are not interested in.

We can see there is no indication of which state each entry belongs to. We could try to make do with this information, load a shapefile with the states' coordinates into Python and use the coordinates for the internal point to query the state for each entry. However, QGIS allows us to do this same process and with less processing and memory resources: For each entry, the states file contains the FIPS State Code [8] (STATEFP, a two digit code that uniquely identifies each state and territory), the US Postal Service two-letter code for the states (STUSPS), and the fully spelled name (NAME). We can use any of these three fields in combination with the geography of the state to identify the ZIP codes that belong to each one of them. To do this, we will create a new layer (i.e. a new shapefile in memory) that will be the result of merging the fields of the state shapefile into the ZCTA shapefile, with the logic for the merge being that the ZCTA area is WITHIN the state area. At this point, we can use QGIS filtering features to select and erase all the entries that are not part of the District of Columbia, Maryland, or Virginia, by looking at the value of the *STUSPS* field. The resulting map (shown in Figure 2.2) contains a much smaller set of entries.

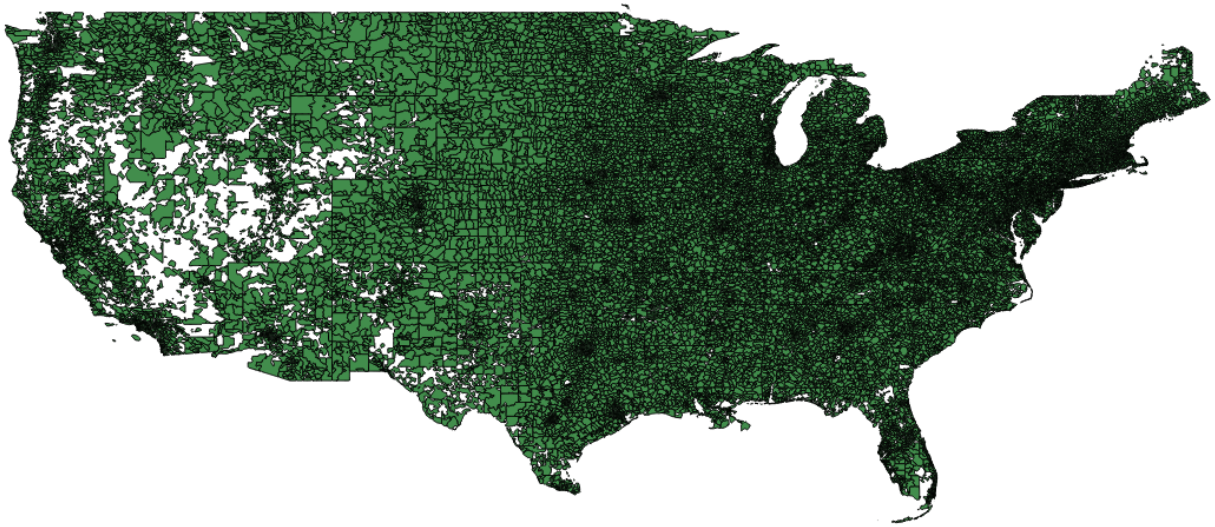


Figure 2.1: ZCTA shapefile loaded in QGIS

Figure 2.2: ZCTA shapefile loaded in QGIS

The resulting CSV file contains 1395 entries with the ZCTAs that belong to the District of Columbia, Maryland, or Virginia. This is a large number of areas to analyze, so we will use the acquisition of the next sets of data to reduce

this number.

2.2.2 Commute Times

In order to obtain the commute times for each ZCTA we will use the Bing Maps API [9, 10]. This API provides free accounts with generous rate limits, so we can test the code and collect the data we need wihtout reaching the maximum number of queries. Using this API we only need to build an HTTP requests (GET or POST) providing our API key, the start and end coordinates of the commute, and the timestamp for the start or end of the commute. The query can contain multiple start points, and/or end points, and the server response will contain a matrix or individual responses that will account for those combinations. We will not use that syntax in our code, as the response times may increase significantly (although the number or billable transactions will decrease).

The commute information we are looking forward to obtain is the commute time from the internal point of each ZCTA to National Landing. From OpenStreetMaps ([11]) we can see that the coordinates used for the map marker of National Landing are '(38.8548783, -77.0517428)' ([12]). This will be our commute end. The start will be the internal point that the ZCTA shapefile provided. Regarding the start time for the commute, we will evaluate the commute time at 30 minutes intervals, starting at 6:00am until 9:30am (included). We will also get the commute time over a work week (Monday through Friday), and then collapse the daily results into the average value for each time slot. That way we can capture the variation of commute times while still keeping the amount of features reasonable.

The server response is formatted as a JSON document, from which we want to extract the commute time. We will do this by using the 'requests' package to build and send the GET request, and the JSON capabilities of the 'pandas' package to parse the response and access the information we want.

Once with the commute times collected, we can try to trim our work data a bit more. If we look at the map of ZCTAs (with the coordinates of the National Landing marked as a red circle in the center) (Figure 2.3), working with all the ZCTAs in Maryland and Virginia means that we are covering a huge area.

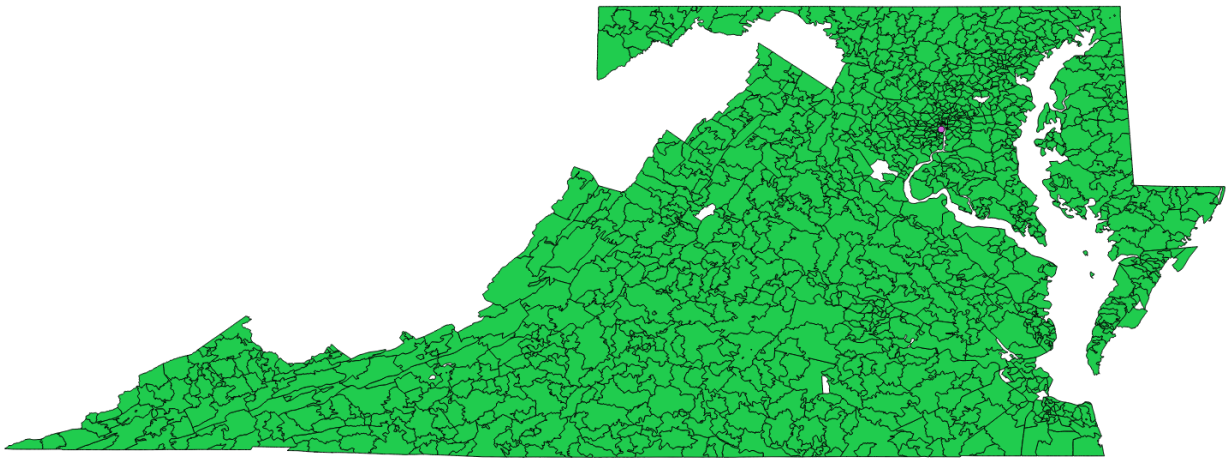


Figure 2.3: Map of all the ZCTAs in the District of Columbia, Maryland, and Virginia

Most of those locations are not really feasible locations to commute to the National Landing on a daily basis, as we can see in Figure 2.4, with the ZCTAs colorized based on the maximum commute time of all the time slots. As we can see, there are huge regions with maximum commute times over 2 hours, not counting traffic delays, accidents, etc. So let's trim the selection of ZCTAs and let's work only with those that have a maximum commute time of 60 minutes. This reduces the number of areas to process from almost 1400 to only 181.

The map we can look at now (Figure 2.5) is more compact, allows us a finer granularity regarding the commute time clustering, and seems to line up better with natural features (we can tell the impact of crossing the river, or leaving the district), highways (the ZCTAs immediately next to the two main higways in the area, I-95 and I-270) show lower commute times than the neighbors), which is a better foundation for our analysis.

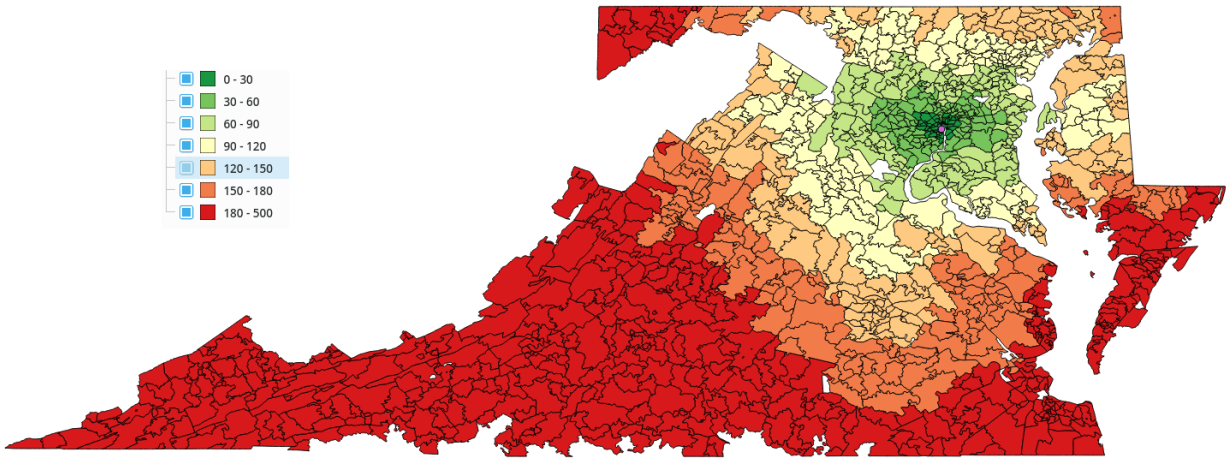


Figure 2.4: Map of all the ZCTAs in the District of Columbia, Maryland, and Virginia colorized by the maximim commute time

2.2.3 Housing Costs

The data that we will use for the housing costs will be obtained from Zillow ([13]). Zillow is a real estate database company that provides some of its datasets publicly available for free. Some of these datasets are the 'Zillow Home Value Index (ZHVI)' and 'Zillow Rent Index (ZRI)' time series. The indexes account are defined as 'smoothed, seasonally adjusted measure of the typical home value' and 'smoothed measure of the typical estimated market rate rent', and are provided as CSV files showcasing the evolution of those indexes for geographical areas over time.

As Zillow provides several indexes depending on the type of housing, number of bedrooms, etc, we have selected this initial set of features to run the analysis on:

- ZHVI Time Series for Single Family Homes.
- ZHVI Time Series for Condos/Co-op.
- ZHVI Time Series for 1 Bedroom Houses.
- ZHVI Time Series for 2 Bedroom Houses.
- ZHVI Time Series for 3 Bedroom Houses.
- ZHVI Time Series for 4 Bedroom Houses.
- ZHVI Time Series for 5 or More Bedroom Houses.
- ZRI Time Series for Multifamily, Single Family Residences, and Condos/Co-op.

As not all of the ZCTAs have housing information available, some more of these areas as discarded for the analysis. We can see a visual representation of the values of the indexes in Figures 2.6 and 2.7

2.2.4 Area Amenities

The information about the available amenities in each of the ZCTAs will be obtained using the FourSquare API ([14]). As we have the coordinates for an internal point inside each of the ZCTAs, we will use those coordinates to query FourSquare for venues in a 500 m radius, with a maximum of 100 venues per area. We will then condense the results in a one-hot encoding like dataframe that will be subsequently aggregated per area, resulting in a dataframe listing all the types of venues for each area, along the percentage of total venues returned for the area that belong to that class.

Some of the areas queried returned no results, and therefore those entries have been removed from the list of areas for the analysis. The final map of the ZCTAs ready for the analysis, colorized using the total count of venues per area can be seen in Figure 2.8.

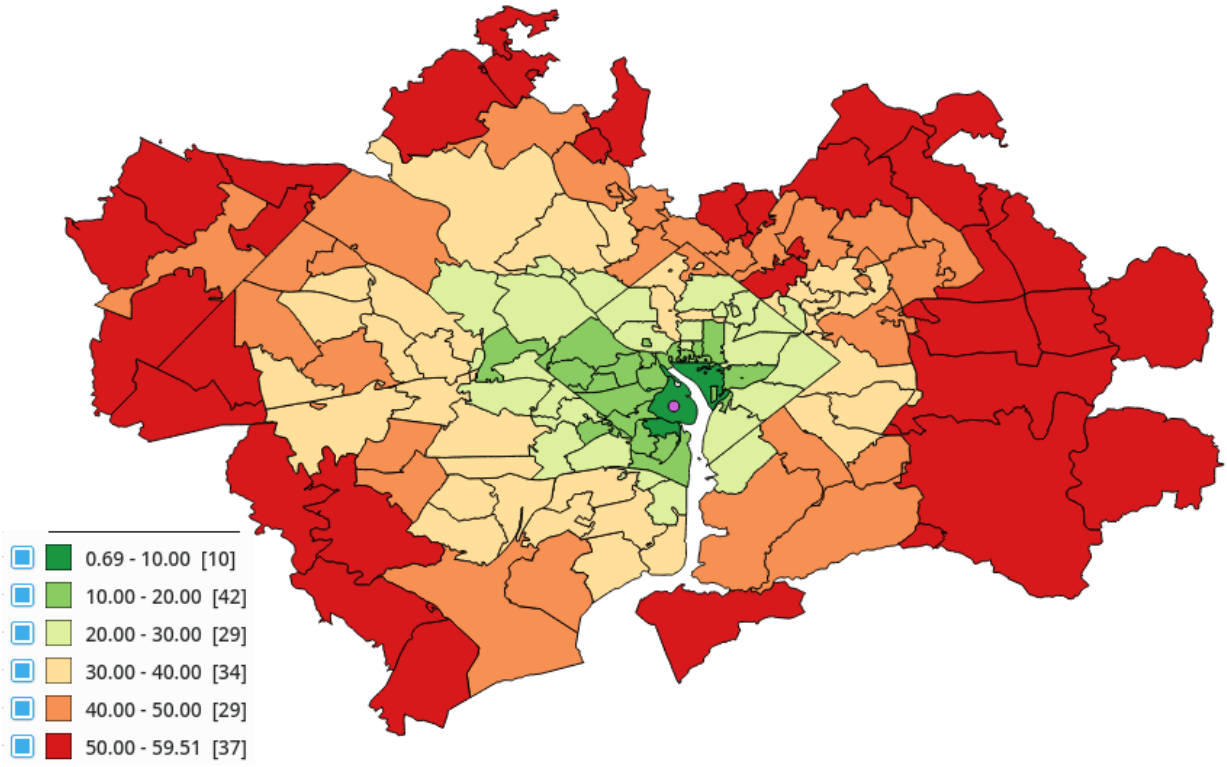


Figure 2.5: Map of the reduced selection of ZCTAs in the District of Columbia, Maryland, and Virginia colorized by the maximim commute time

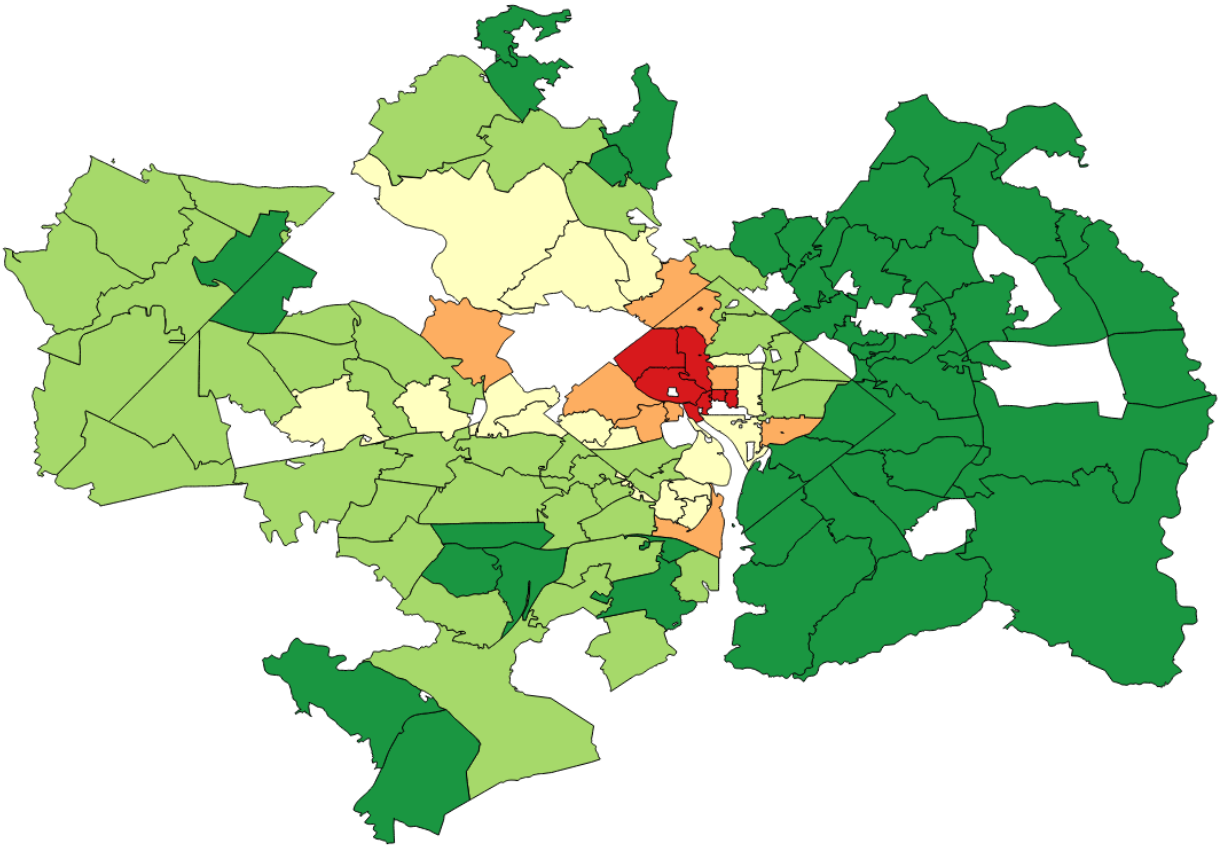


Figure 2.6: Map of the ZCTAs colored by the highest house buying index

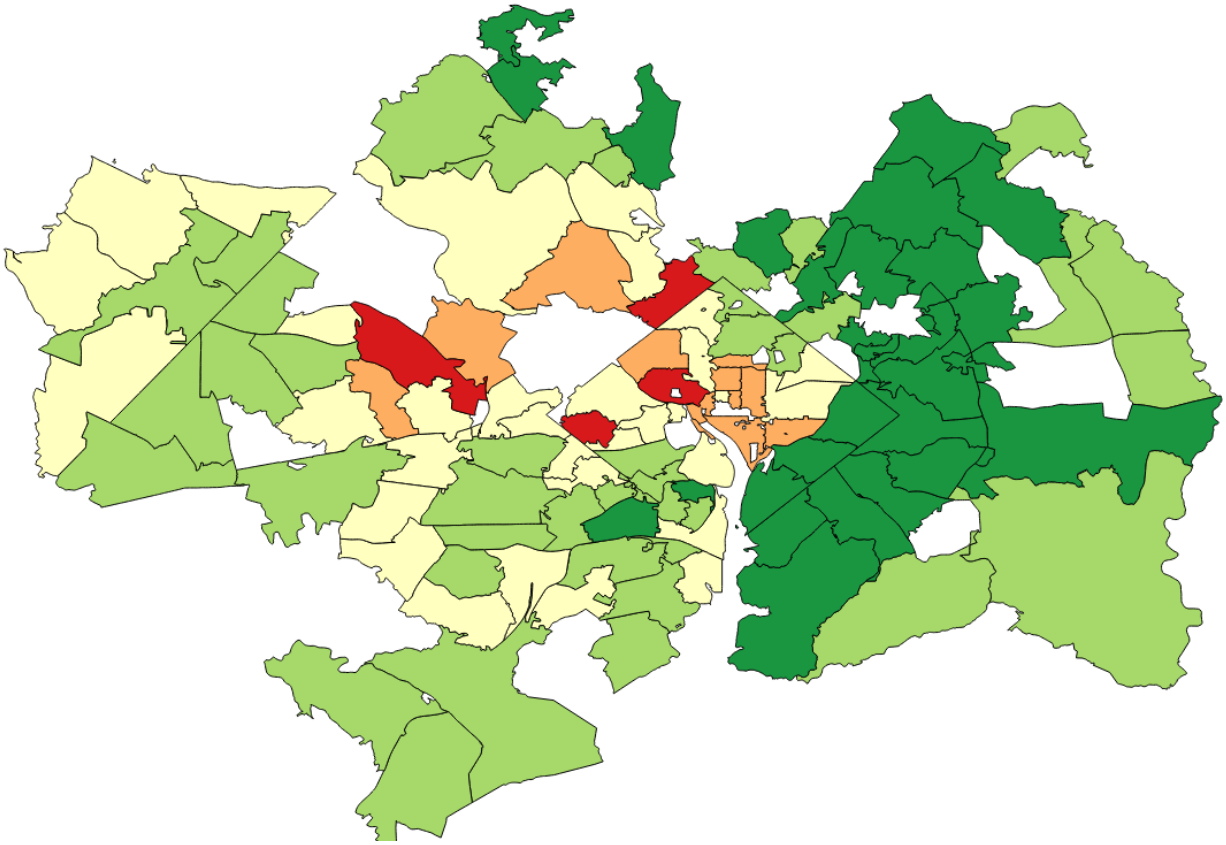


Figure 2.7: Map of the ZCTAs colored by the house rent index

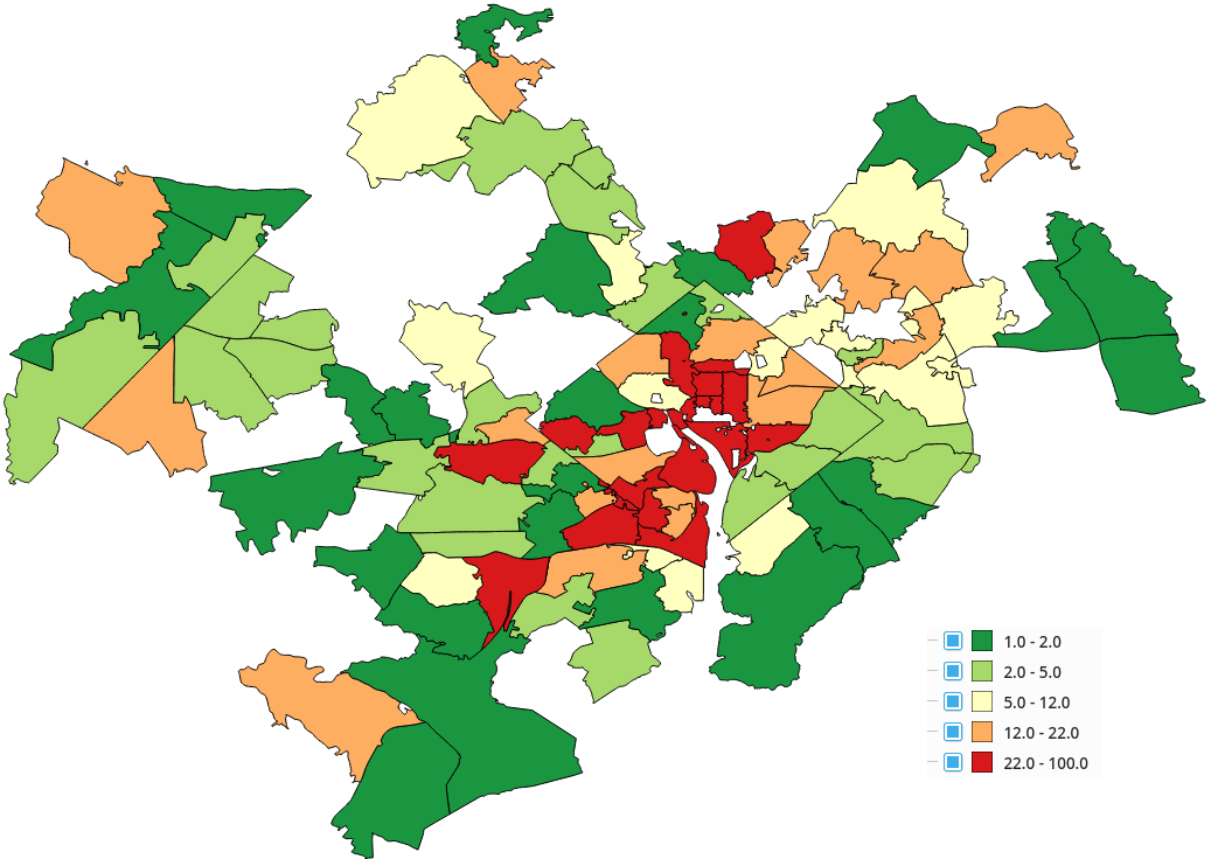


Figure 2.8: Map of the ZCTAs colored by the venue count

3 Methodology

3.1 Exploratory Analysis

3.1.1 Correlation Between Commute Times

One of the first questions we need to answer is whether some of the features we have obtained for the analysis are reduntant, or if all of them are relevant. The commute time features are a prime subject of analysis in this regard, as common sense would dictate that the commute time would increase or decrease similarly for all the start times (that is, that there is a direct relation between the difference in commute time at 6:00am and any other time of the day). However, the fact that traffic demand is different at different times of the day (e.g., early in the morning there is a lot of cargo traffic from warehouses to stores, while later in the morning, when the schools are about to open, traffic increases between residential areas and schools). If we can prove that this relation exists, we can replace the 8 features of the analysis with a single one, making the rest of the analysis significantly easier.

In order to validate this hypothesis, we have computed the Pearson Correlation Coefficient and the P-value between the commute times starting at different times. We can see in Table 3.1 how all the Pearson correlation coefficients are very close to 1, indicating a positive correlation. Additionally, all the P-values are extremely small, indicating a strong confidence on the results.

	Correlation Coefficient	P-Value
06:30	0.9834	$5.40e^{-76}$
07:00	0.9731	$1.35e^{-65}$
07:30	0.9651	$4.98e^{-60}$
08:00	0.9755	$1.41e^{-67}$
08:30	0.9745	$1.05e^{-66}$
09:00	0.9693	$1.01e^{-62}$
09:30	0.9616	$5.87e^{-58}$

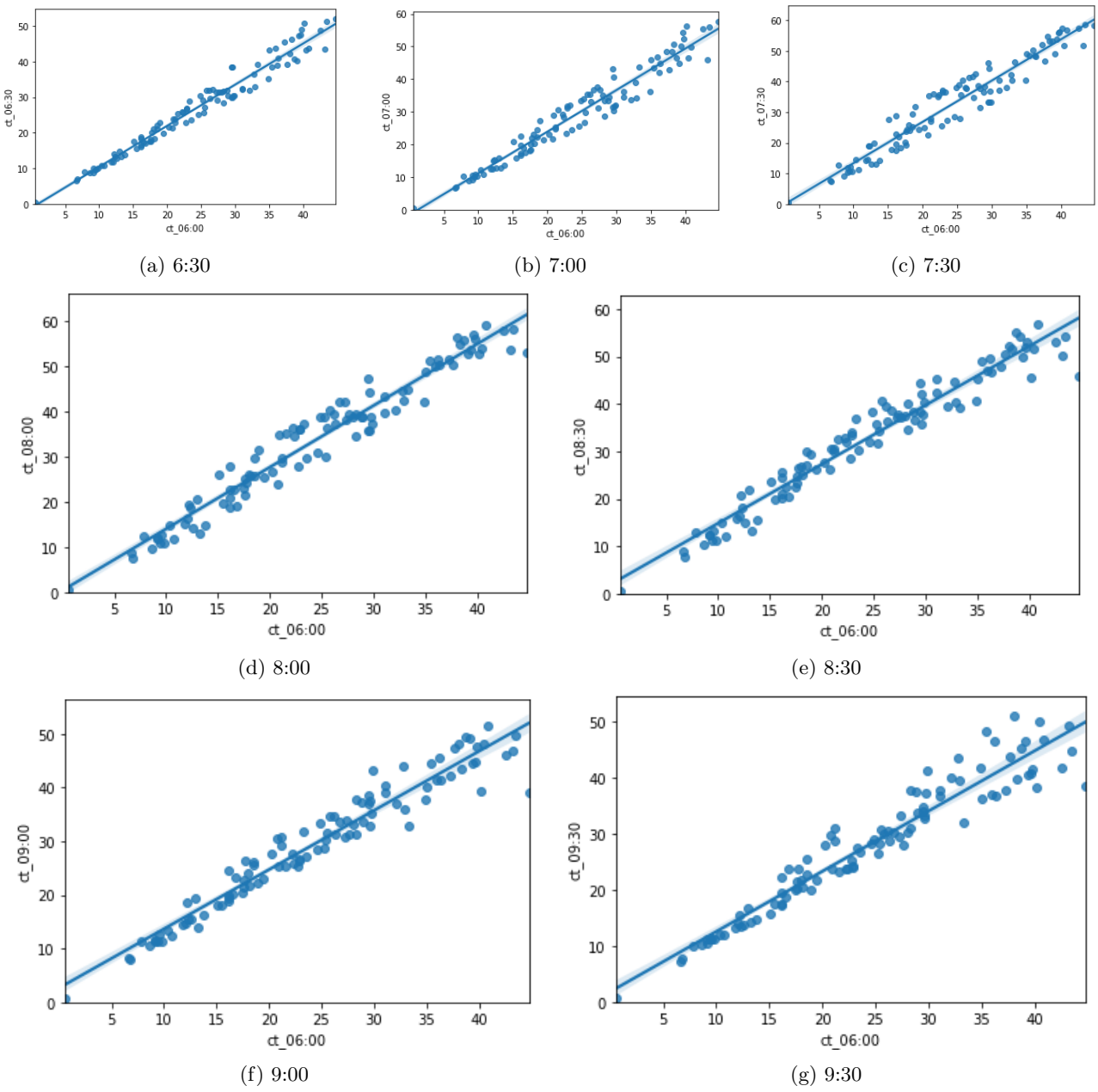
Table 3.1: Pearson Correlation Coefficients and P-Values of the commute time at 6:00 with the rest of the hours

One interesting thing of note that we can miss if we only look at the correlation factor, but it is clearly shown in the correlation plots in Figure 3.1 is that the correlation is stronger with the lower values. This makes sense, as shorter commute times indicate shorter commute distances, and therefore, lower changes for traffic jams, roads of different speeds, etc. This also means that if we hadn't set a limit to the maximum commute time, we may have not found this correlation to be so clear. Additionally, we can see how the correlation is stronger with closer commute times, which also makes sense, as the road conditions are more likely to be similar within 30 minutes, than when 4 hours have passed.

With all this information, for further analyses, we will replace the all the commute times columns with the commute time at 8:00.

3.1.2 Correlation Between Housing Prices

Similarly to what happened with the commute times, it would also seem reasonable to think that the home indexes of the different types of homes have some relation between each other. And if this is the case, we can also reduce the number of features used for our analysis. In Table 3.2 we can see that in this case the correlation is not that clear. While some features have a high correlation index, others do not. In this case it seems clear that the single family index is largely driven by large houses with 3 or more bedrooms, while the rent index seems to behave closer to the smaller homes.



(a) 6:30 (b) 7:00 (c) 7:30

(d) 8:00 (e) 8:30

(f) 9:00 (g) 9:30

Figure 3.1: Correlation plots between the commute time at 6:00 and the rest of the start times

If we look at the histograms of the different indexes, shown in Figure 3.2, we can see how the Multi-Family Homes and the 1- and 2-Bedroom homes have a significantly different distribution than the rest. If we compute the correlation between the Multi-Family Homes indexes and the other two, we obtain the values in Table 3.3, which show a much better correlation for the 1-Bedroom homes, similar for 2-Bedroom, and worse correlation for Renting.

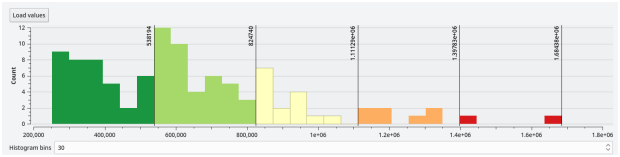
With all this information, for further analyses, we will replace the all the housing index columns with only the single-family and multi-family indexes.

3.1.3 Correlation Between Commute Time and Housing Prices

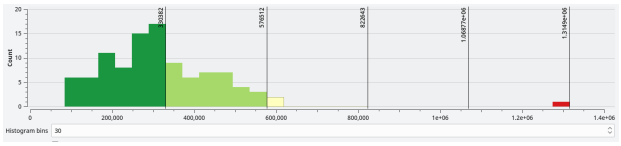
The next item we need to explore is whether there is a correlation between and commute times. On the one hand it would seem that higher commute times would lead to lower housing indexes, as it is likely that we are getting further

	Correlation Coefficient	P-Value
Multi-Family Homes	0.6123	$7.98e^{-12}$
1-Bedroom	0.4814	$3.03e^{-07}$
2-Bedroom	0.7802	$4.24e^{-22}$
3-Bedroom	0.8769	$1.41e^{-33}$
4-Bedroom	0.9001	$5.77e^{-38}$
5-Bedroom	0.9290	$5.94e^{-45}$
Renting Index	0.7994	$7.50e^{-24}$

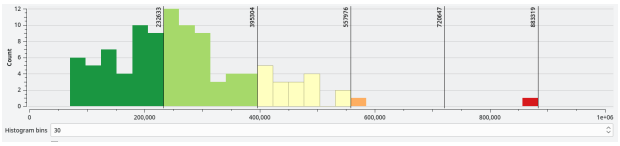
Table 3.2: Pearson Correlation Coefficients and P-Values of the ZHVI for single family homes, with the rest of types of homes



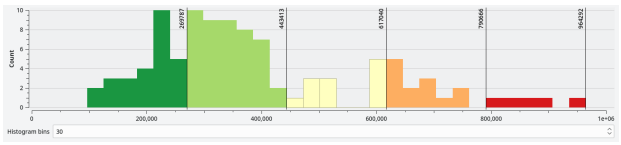
(a) ZHVI Single Family Homes



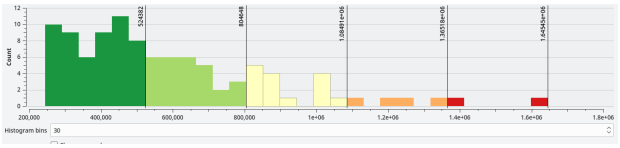
(b) ZHVI Multi-Family Homes



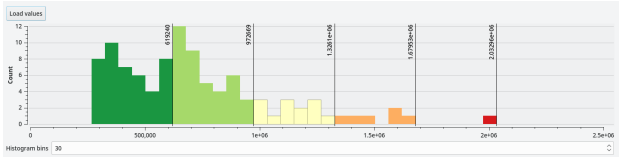
(c) ZHVI 1-Bedroom Homes



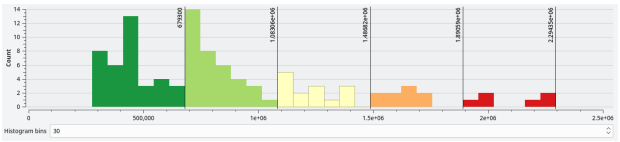
(d) ZHVI 2-Bedroom Homes



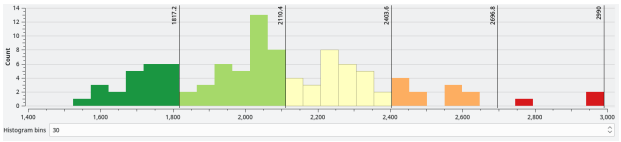
(e) ZHVI 3-Bedroom Homes



(f) ZHVI 4-Bedroom Homes



(g) ZHVI 5-Bedroom Homes



(h) ZRI All Homes

Figure 3.2: Histograms of the housing indexes

	Correlation Coefficient	P-Value
1-Bedroom	0.8644	$1.28e^{-31}$
2-Bedroom	0.7762	$9.46e^{-22}$
Renting Index	0.7691	$3.71e^{-21}$

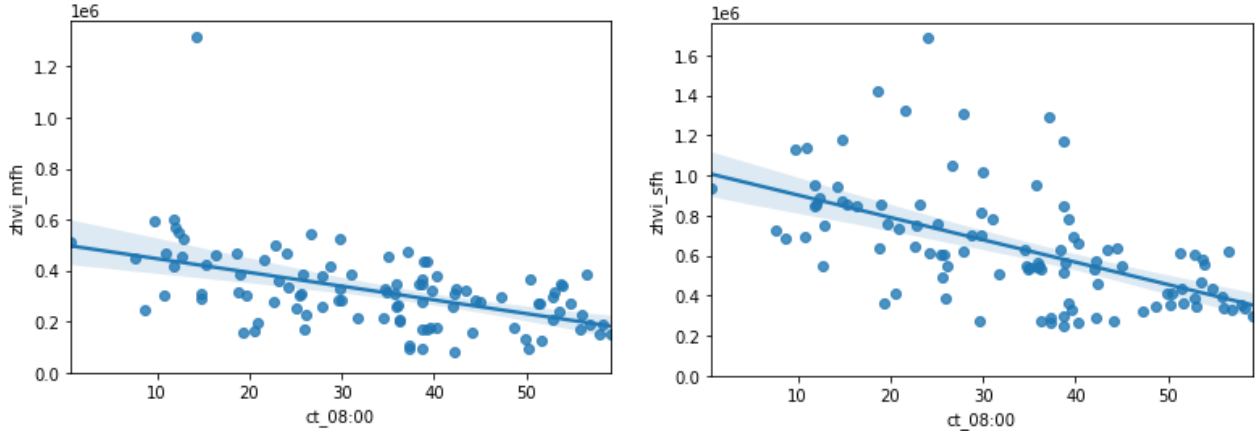
Table 3.3: Pearson Correlation Coefficients and P-Values of the ZHVI for single family homes, with the rest of types of homes

away from the urban area and therefore the prices should be lower. However, looking at the coefficient indexes in Table 3.4 and the correlation plots in Figure 3.3 we can see that the Pearson Correlation Coefficient is negative (which makes sense: we expect that as the commute time increases the index decreases), but the correlation is not great. We can find an explanation for this by looking at the maps colorized with the commute times and housing indexes (Figure 3.4) we can see that while the commute time follows a clear distribution increasing with the distance to the DC area, the housing indexes do not behave that way. For single family homes we can see how the indexes are high not only in the

DC orban area, but also in the North West corridor that constitutes the population clusters in Montgomery County, while the multi-family homes shift more towards DC's North East region, getting close to Baltimore. In both these cases the variation with distance is very inconsistent, with areas like the peninsula in the East having very low values, but the South region of the map in Virginia, which is further away, still has moderate indexes.

	Correlation Coefficient	P-Value
ZHVI Single-Family	-0.5609	$8.74e^{-10}$
ZHVI Multi-Family	-0.4930	$1.07e^{-07}$

Table 3.4: Pearson Correlation Coefficients and P-Values of the commute times with the types of homes



(a) Single Family Homes

(b) Multi-Family Homes

Figure 3.3: Correlation plots between the commute time and the types of homes

With these results in hand, we cannot conclude that there is any significant correlation between the commute times and the housing indexes.

3.1.4 Correlation Between Housing Indexes and Venue Count

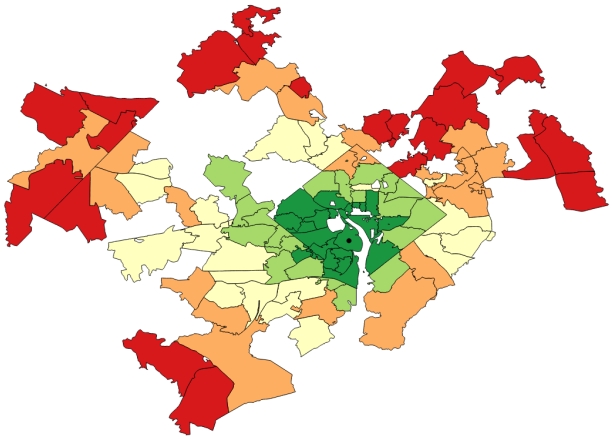
Another aspect of the data that needs to be analyzed is whether the housing indexes are somehow correlated with the venue count for each ZCTA. In Table 3.5 we can see the Pearson Correlation Coefficiens and P-Values for the number of venues in each ZCTA and the types of homes. As we can see, these are the lowest values obtained so far, both in term of the coefficients and the P-Values. Looking at the map in Figure 3.6 we can see that the distribution of the venue count is the most idiosyncratic we have seen so far, alternating close areas with very high number of venues with areas with very low.

	Correlation Coefficient	P-Value
ZHVI Single-Family	0.3447	$3.89e^{-4}$
ZHVI Multi-Family	0.4340	$5.21e^{-6}$

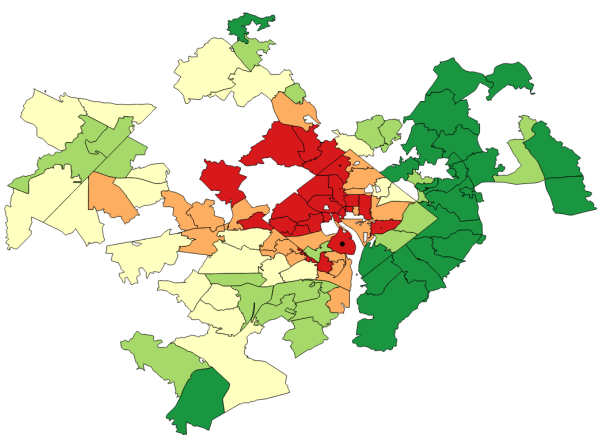
Table 3.5: Pearson Correlation Coefficients and P-Values of the venue counts with the types of homes

3.2 Clustering Analysis

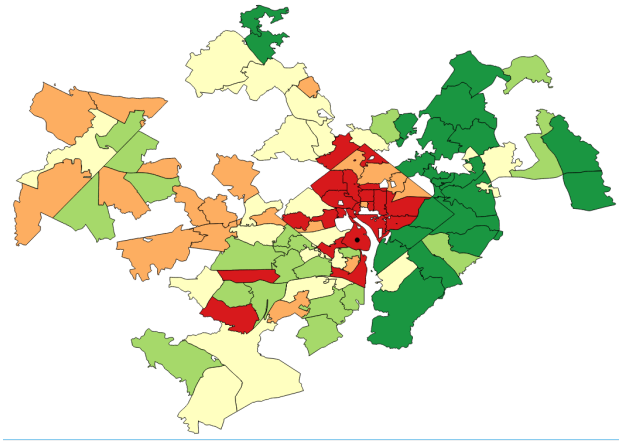
The main goal of this study is to identify areas around the Washington DC greater metropolitan area that show similar characteristics to the National Landing site. Therefore, the tools indicated for this type of work are clustering tools, that will group the different areas of study according to how similar they are. We do not want to use classification



(a) Commute Time

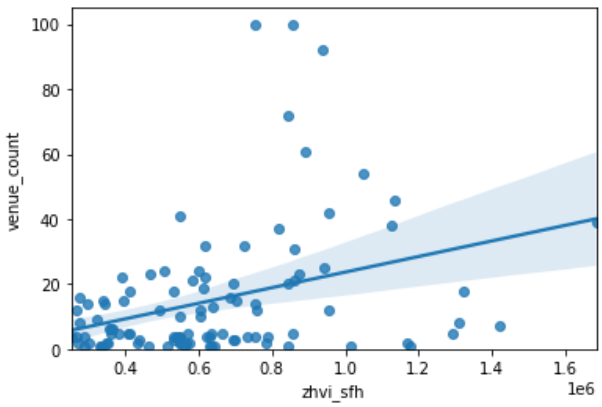


(b) Single Family Homes

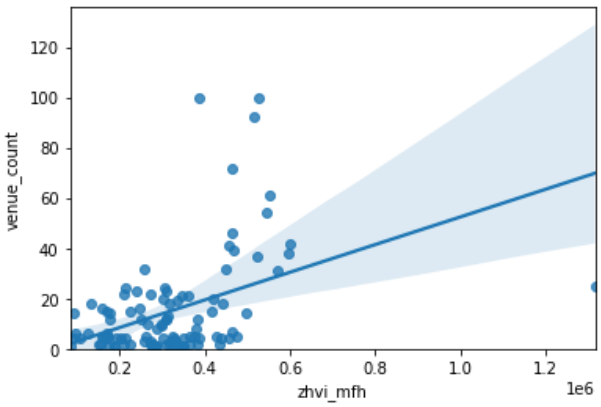


(c) Multi-Family Homes

Figure 3.4: Geographical maps colorized according to the commute time and housing indexes



(a) Single Family Homes

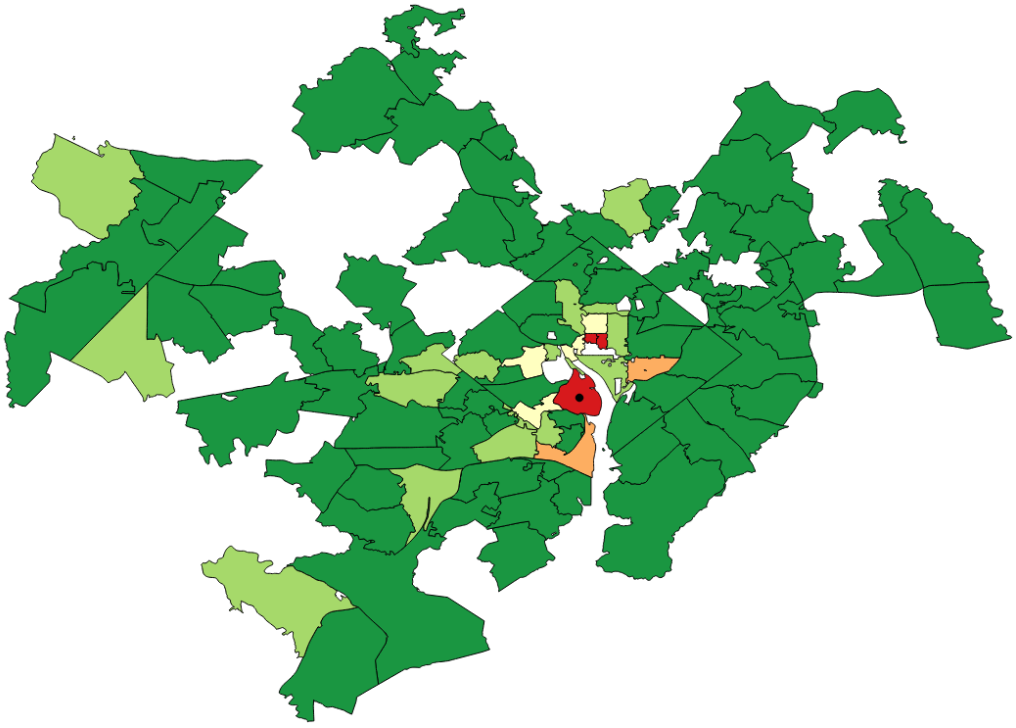


(b) Multi-Family Homes

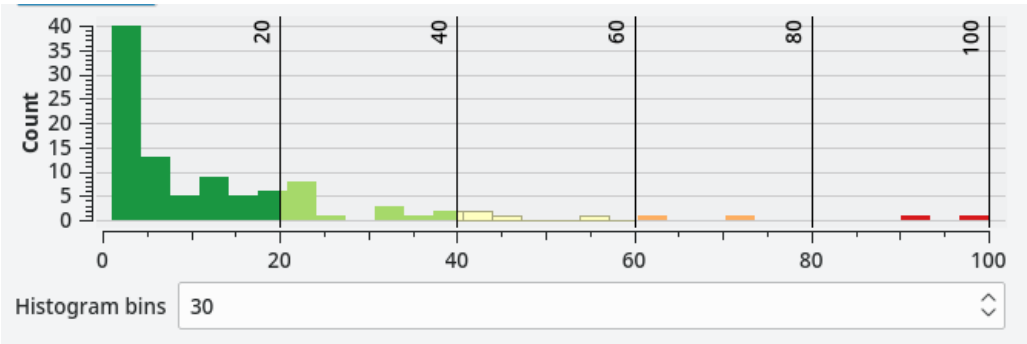
Figure 3.5: Correlation plots between the number of venues and the types of homes

techniques, as we do not have labels or categories beforehand that we can define in order to identify the different types of areas. Therefore, clustering techniques seem to be the most appropriate tool to tackle this problem.

Regarding the specific tools used, we consider using both K-Means and DBSCAN. K-means is attractive for us



(a) Geographical map of the distribution of the number of venues



(b) Histogram of the distribution of the number of venues

Figure 3.6: Map and histogram of the number of venues in each ZCTA

because being unsupervised, we can let the algoroithm work without having to worry about us doing a good join training it. On the other hand DBSCAN has the advantage of not forcing all the nodes being classified into one of the clusters, and that when not having prior knowledge about the shape of the clusters (as in this case), it adapts better to arbitrary shapes. Seeing how we do not know which one will yield better results, we will then proceed to analyze the data using both methods and we will compare the results afterwards.

Additionally, we have to account for the fact that the different features use scales and ranges very different from each other. Besides that, not all of them show linear behavior (like the 'number of venues' feature, that as we saw in the histogram in Figure 3.6b it is distributed almost logarithmically). To overcome this problem we will normalize the different metrics before applying the algorithm.

In order to determine the number of clusters we will run the clustering algorithm for multiple values of cluster number, and plot the inertia (the sum of the squared distances from each node to the cluster center) to determine when adding more clusters only yields minimal gains.

4 Clustering Results and Discussion

4.1 Clustering Results

In this section we will review the results obtained from the clustering algorithms and compare them.

First of all, for the K-Means clustering, the first notable issue is the election of the number of clusters to use. For this task we carried out the classification process with different number of clusters, computing the inertia (the sum of the squared distances from each node to the center of the cluster) at each step. Then we plotted it hoping to find an 'elbow' that would indicate when adding more clusters only yielded diminishing returns. However, as we can see in Figure 4.1 the curve does not have a clear elbow. While the rate of improvement clearly slows down after 6 clusters, the absolute improvement is still significant. For this reason, we decided to carry out the analysis using 2 values for the number of clusters: 6 (the closest to an elbow that could be identified in the figure) and 30 (a significant absolute improvement compared with 6).

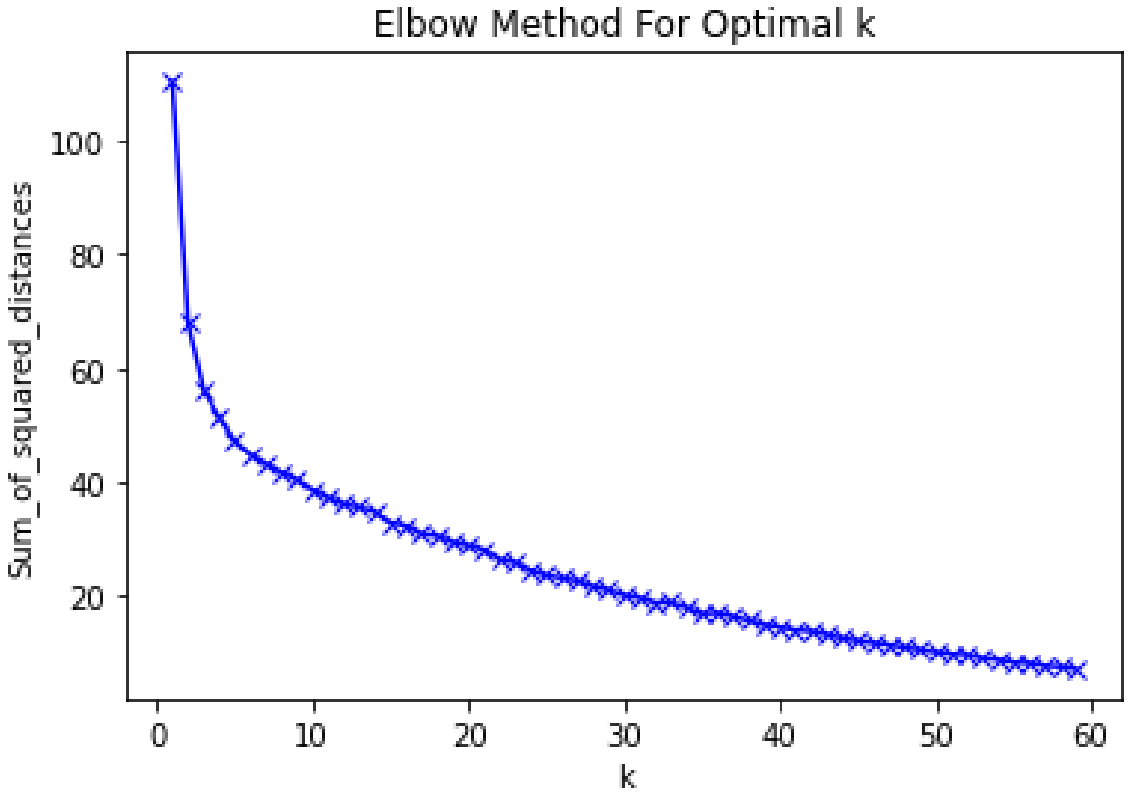
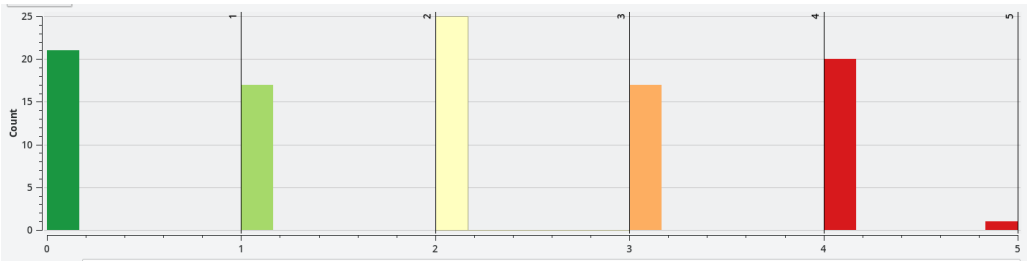


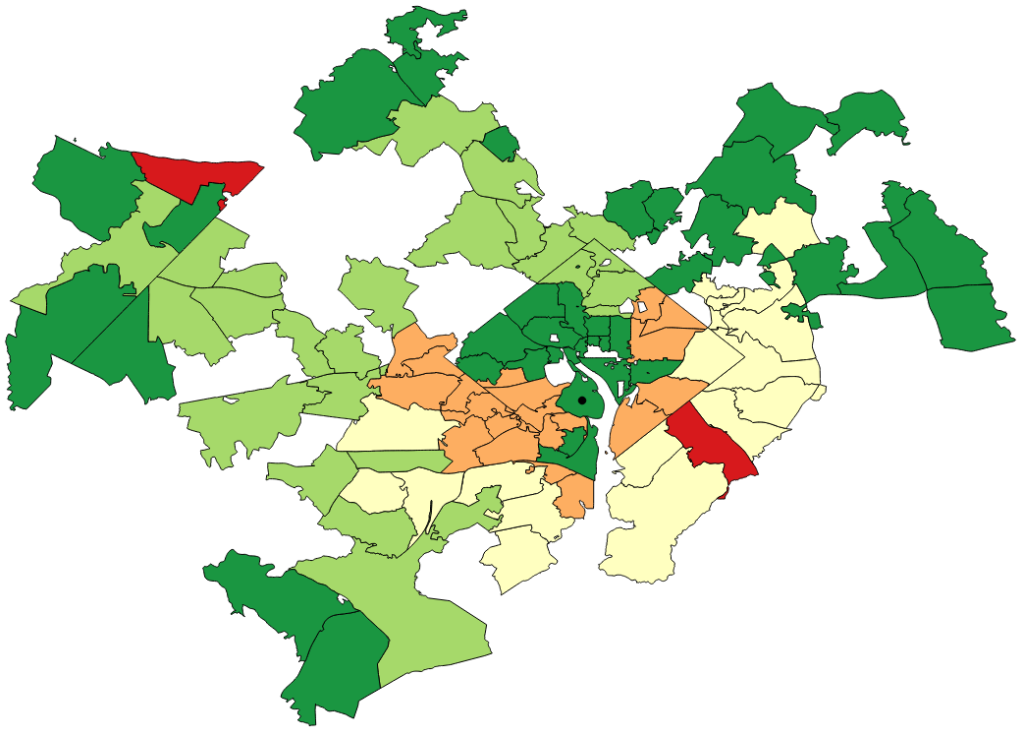
Figure 4.1: Inertia plot for K-Means

Looking at the results of the KMeans clustering with 6 clusters, we can look both at the histogram (Figure 4.2a) and the geographical map (Figure 4.2b). We can see how the clusters are evenly populated except for the last one, which can be identified in the map as an outlier, with both areas in that cluster being on the limits of the area being considered. Regarding the quality of the clustering, it is really not that great, because if we look at the map we can see how the DC area is all included in the same cluster (which is mostly expected), but also included are areas far away from DC in the rural and mountain areas of Maryland and Virginia. It is clear that this clustering method didn't have

enough clusters to differentiate all the specifics.



(a) Histogram of the KMeans Clustering with 6 clusters

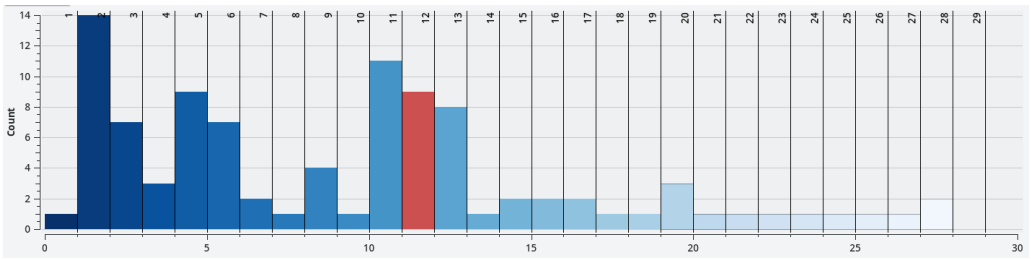


(b) Geographical map of the KMeans Clustering with 6 clusters

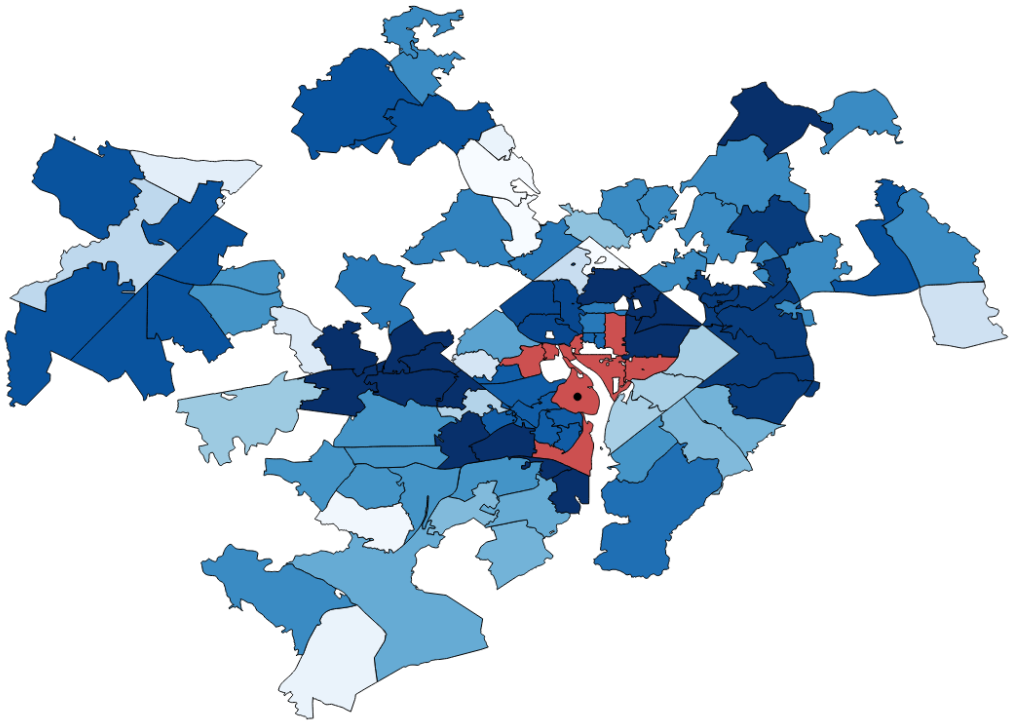
Figure 4.2: Results for the KMeans Clustering with 6 clusters

That conclusion is more obvious if we look at the results of the Clustering with KMeans and 30 clusters (Figure 4.3). In these figures we plotted everything in blue except the cluster with the National Landing, which is plotted in red. This allowed us to clearly present the results without causing confusion with multiple similar shades of the same color. As we can see in the histogram, the allocation of areas per cluster in this case is nowhere as uniform as before. However, the map shows that the cluster that contains the National Landing area is composed of areas in the DC and Northern Virginia. This is a much better result than before, not only because of the National Landing cluster, but also the rest of the areas colored similarly appear to have relevant connections in terms of type of area (urban, suburban, rural), distance to the DC area, and housing indexes.

Finally, the DBSCAN clusteing results can be observed in Figure 4.4. As we can see, most of the areas were considered outliers. We tried to refine and improve these results by modifying the values of epsilon and the minimim number of nodes in a cluster. We can see in Figure 4.5 how the number of clusters and outliers changed when tweaking the values of epsilon and the minimum number of samples in a cluster. We can see how there are no good combinations, and the one with the results shown was actually the most promising of all. However, none of the combinations tried managed to allocated the National Landing in a cluster. Even looking at the areas that have been classified, we can see how suburban areas near DC have been allocated in the same cluster as mountain areas in the Appalacchian Mountains.



(a) Histogram of the KMeans Clustering with 6 clusters



(b) Geographical map of the KMeans Clustering with 30 clusters

Figure 4.3: Results for the KMeans Clustering with 30 clusters

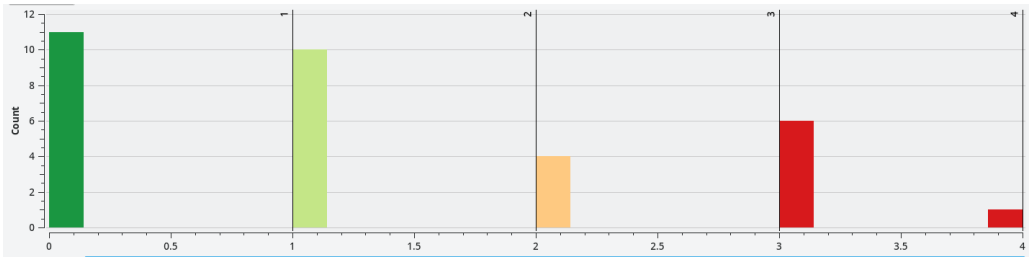
Therefore, DBSCAN did not provide us with satisfactory results.

4.2 Discussion

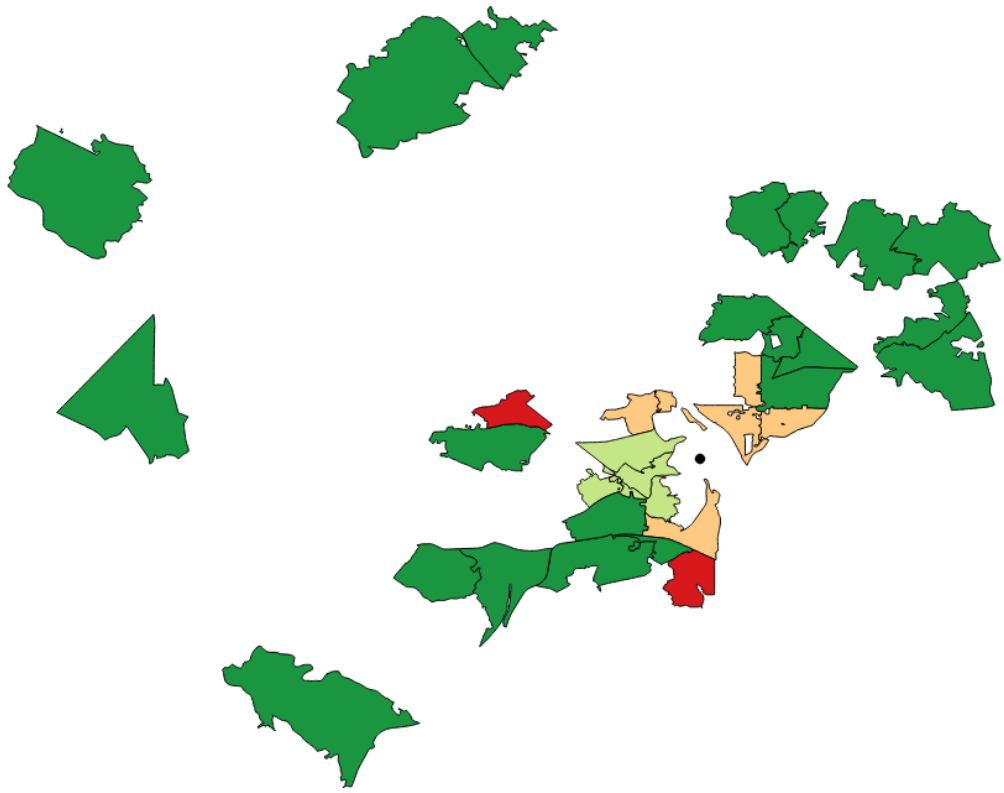
The main observation that we can make based on the previous results is that the area under analysis is very diverse in terms of the metrics selected. This has led to a situation where most of the areas could be classified as their own clusters. While KMeans managed to get around this by using more clusters, DBSCAN failed to properly identify which areas are actual outliers, resulting in very confusing clustering results.

In general, for these results KMeans has yielded more consistent and robust results, and the clustering resutls, when plotted as a map, can be explained, or, at the very least, do not shock as something completely unexpected. This was supported by the ability of running the algorithm with multiple amounts of clusters, and obtaining meaningful different results each time. As a contrast, DBSCAN was also run with different parameters, but the results in most cases were clustering all the areas as outliers, or all of them in a single cluster.

One aspect that came to our mind when analyzing the results was the possibility that, by organizing the features in the way we did, we may have given more weight to the venue availanbility, as that aspect ended up comprising over 279 dimensions of the final 294. It would be interesting to find a way to balance the weight of the different metrics



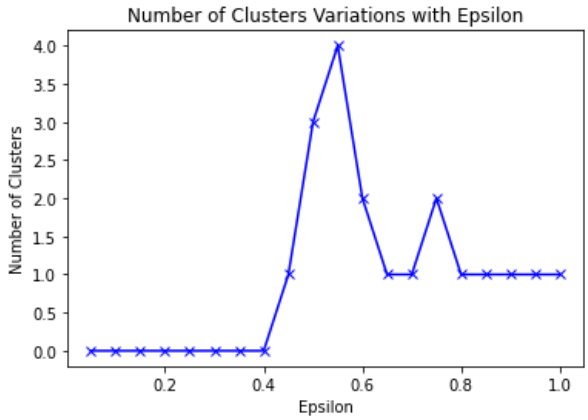
(a) Histogram of the DBSCAN Clustering



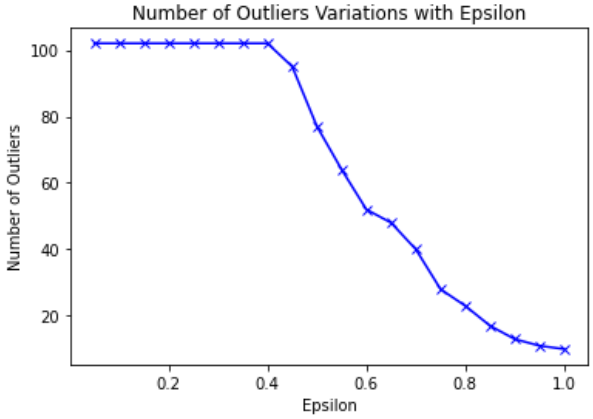
(b) Geographical map of the DBSCAN Clustering

Figure 4.4: Results for the DBSCAN clustering

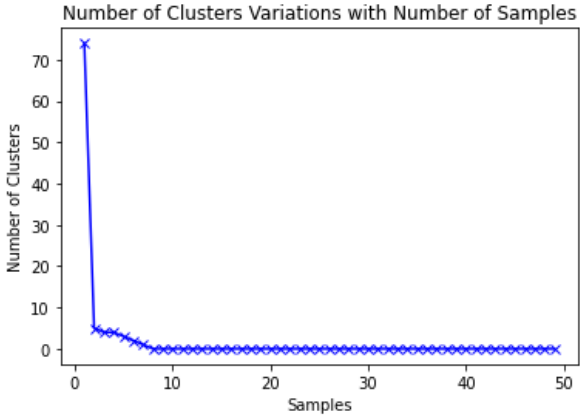
and re-run the classifications, in order to compare the results.



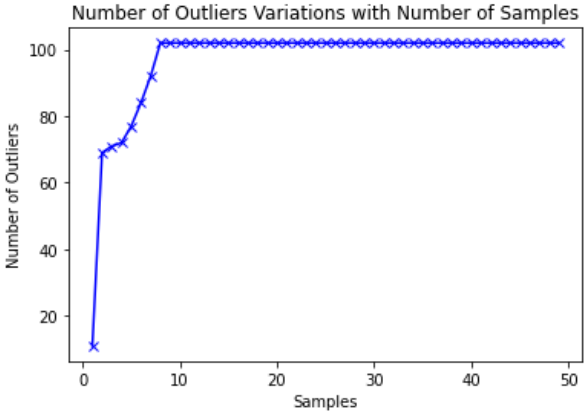
(a) Evolution of the number of clusters with values of epsilon



(b) Evolution of the number of clusters with values of epsilon



(c) Evolution of the number of clusters with the minimum number of samples in a cluster



(d) Evolution of the number of clusters with the minimum number of samples in a cluster

Figure 4.5: DBSCAN tweaking tests

5 Conclusions

In this report we have studied the areas in DC, Maryland, and Virginia, and characterized them in terms of dommute time to the National Landing site, housing costs indexes, and venue availability. We managed to acquire the data, pre-process to clean and trim entries that would only slow down the study, and process the resulting datasets with several clustering algorithms. Some of the results we got seem satisfactory (KMeans with 30 clusters), while others did not work out at all (DBSCAN). As a first analysis of the problem, the work presented here is a solid foundation to continue refining and improving the study, incorporating new data and metrics as they become available or relevant.

As the immediate future work, we would recommend perform the clustering processes individually per feature, and then analyzing how much they differ from the results we obtained. It is very possible that one (or a set of) metrics appear as dominant factors in the analysis, and in this case, it shall be studied how to proceed with them.

Another step would be incorporating services available in the area such as police, firefighters, and hospitals per capita. Alternatively, the distance to hospitals, fire and police stations, etc. could be used for the analysis, thus providing another relevant and important piece of information into the analysis.

Bibliography

[1] "Amazon's Grand Search For 2nd Headquarters Ends With Split: NYC And D.C. Suburb." https://www.npr.org/2018/11/13/665646050/amazons-grand-search-for-2nd-headquarters-ends-with-split-nyc-and-d-c-suburb Accessed: 2020-03-01.

[2] "Amazon Picks New York City, Northern Virginia for Its HQ2 Locations." https://www.wsj.com/articles/amazon-chooses-new-york-city-and-northern-virginia-for-additional-headquarters-1542075336. Accessed: 2020-03-01.

[3] "Washington is No. 3 in traffic congestion, study says." https://www.washingtonpost.com/local/trafficandcommuting/its-a-waste-of-time-washington-is-no-3-in-traffic-congestion-study-says/2019/08/22/e6602e0e-c4d6-11e9-b72f-b31dfaa77212_story.html. Acccessed: 2020-03-01.

[4] T. L. David Schrank, Bill Eisele, "2019 urban mobility report," tech. rep., Texas A and M Transportation Institute, 2019.

[5] "Neighborhoods in Washington, D.C.." https://en.wikipedia.org/wiki/Neighborhoods_in_Washington,_D.C. Acccessed: 2020-03-01.

[6] "2019 TIGER/Line Shapefiles (machinereadable data files) / prepared by the U.S. Census Bureau, 2019." https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html. Accessed: 2020-03-01.

[7] "QGIS Geographical Software." https://qgis.org. Accessed: 2020-03-01.

[8] W. Foundation, "Federal information processing standard state code," tech. rep., 2020.

[9] "Distance Matrix API." https://www.microsoft.com/en-us/maps/distance-matrix. Accessed: 2020-03-01.

[10] "Microsoft Documentation - Calculate a Distance Matrix." https://docs.microsoft.com/en-us/bingmaps/rest-services/routes/calculate-a-distance-matrix. Accessed: 2020-03-01.

[11] "Open Street Maps." https://www.openstreetmap.org. Accessed: 2020-03-01.

[12] "National Landing at Open Street Maps." https://www.openstreetmap.org/node/6061028362. Accessed: 2020-03-01.

[13] "Zillow Home Value Index." https://www.zillow.com/research/data/. Acccessed: 2020-03-01.

[14] "FourSquare API." https://developer.foursquare.com/. Accessed: 2020-03-01.