The Best Location for a New Daycare in Fairfax County, Virginia

By Gustave Muhoza

July 2020

Executive Summary

In this report, I show how a businessperson can take advantage of data science techniques to solve the most pressing problem for anyone trying to start a new in-person business: where should the new business be located? Guided by the CRISP-DM methodology, I used daycare as a case study to present a roadmap for solving this problem. My objective was to conduct an exploratory analysis that leaves a manageable number of neighborhoods with which the businessperson can start. After defining this objective, I collected demographic and geospatial data in different formats from the Census Bureau and Fairfax County websites. I combined this data with daycare location data returned by a Foursquare API request, and prepared it for modeling. A model was generated using agglomerative clustering, and I deployed it to obtain four clusters of similar neighborhoods from which two clusters and one neighborhood from each of the two clusters were selected.

I. Introduction

I.1. Project Motivation

In the last two decades, the child care services industry has been steadily growing. In 2019, for example, it was estimated that this industry was bringing in 47 billion dollars in revenues.[1] Although child care services is currently ranked as the second best franchise in the United States[2], the vast majority of child care services is still provided by family members or within a home and often without any monetary benefits for the provider.[3] Clearly, there is room for new daycares.

Given this size and steady growth, providing childcare services can be fulfilling not just for the type of services provided--serving children is a worthy goal in its own right--but also because of financial sustainability. As is the case for many other small businesses, however, long term financial sustainability is not guaranteed: starting a child care services business requires extensive market research.

I.2. The Business Problem

In this project, I explore a way data science can help answer perhaps the main question for any in-person business market research, namely, the question of location. Specifically, I present the case of a person in Fairfax County, Virginia who is in the process of making a decision about the neighborhood in which to establish a new daycare. Using Census Bureau, Fairfax County, and Foursquare data, my goal is to answer the following question: Which Fairfax neighborhoods present the best opportunity for a new daycare?

[1][1]"Child Care Provides Nearly $100 Billion Economic Impact ...." https://blogs.edweek.org/edweek/early_years/2019/02/child_care_packs_nearly_100_billion_economic_impact_report_finds.html. Accessed 26 Jul. 2020.

[2] "The Top Franchising Categories of 2020." https://franserve.com/2020/02/the-top-franchising-categories-of-2020/. Accessed 20 Jul. 2020.

[3] "Nearly 30 Percent of infants and Toddlers Attend Home-based Child Care as Their Primary Arrangement." https://www.childtrends.org/nearly-30-percent-of-infants-and-toddlers-attend-home-based-child-care-as-their-primary-arrangement. Accessed 22 Jul. 2020.

I.3. Objectives and Possible Application

To answer the above question, I use CRISP-DM methodology to guide the process. My objective is to present at most two neighborhoods with population characteristics that indicate need for additional daycares. Using python and many of its libraries, I acquire data from different sources and formats, prepare that data and use hierarchical clustering to obtain 4 clusters of neighborhoods. From these four, two clusters are chosen as most favorable, and then one neighborhood from each of the two clusters is selected as best for a new daycare.

Results from this analysis can be useful not just for daycare businesses, but also for similar small businesses in the service industry as well as industry advisors.

II. Data Part I

II.1. Data Description

Choosing a daycare location may depend on many factors (see US Small Business Administration guidance[4]. I explore a few of these factors including the geographic concentration of children ages 0 to 4, family work structure, the number of daycares available in the area, and income level in the neighborhood. I base my analysis and modeling on the following data sources:

- **Census Designated Places (CDPs) data[5]**: The data is available through Fairfax County Open Geospatial Data and is provided in many formats. I chose the csv version. In addition to the CDPs names that will be used as neighborhoods for the purpose of my analysis, this data provides latitudes and longitudes of these neighborhoods for mapping purposes, and for obtaining nearby daycares. Knowledge of the area confirms that Fairfax County CDPs can indeed be considered neighborhoods.

[4] "How to Start a Quality Childcare Business." https://www.sba.gov/sites/default/files/files/pub_mp29.pdf. Accessed 20 Jul 2020.
[5] "Fairfax County Geospatial Data." https://www.fairfaxcounty.gov/maps/open-geospatial-data. Accessed 10 Jul 2020.

- **Census Designated Places (CDPs) data (2)**[6]: This file, which is available in several formats through data.gov replaced the file above in order to obtain neighborhood shape coordinates for mapping needs. I initially downloaded it in its shape (.shp) format and transformed it into a dataframe, and then into csv thinking it would make it easier to access. Unfortunately, the key geometry data format did not come out alright. As a result, I decided to do more study on IBM Watson and was able to figure out how to install geopandas and import to read geojson file. This allowed me to also read it directly from Fairfax GIS site in its geojson format that includes a well-formatted geometry field with Multipolygon coordinates representing all Fairfax CDPs.
- **Census 2010 ZCTA to Place Relationship File**[7]: This csv file is provided by the US Census Bureau and contains population data by zip code tabulation area (ZCTA) and how the ZCTAs match county's CDPs. For simplification, I use "Zip Code" (postal code) and "ZCTA" interchangeably even though I am aware of issues that may result.[8] Lack of geographic precision, however, should not undermine this analysis because a daycare established in one geographic area can provide services to children residing in other nearby geographic areas. This file makes it possible to represent businesses by zip code (approximately) and neighborhood at the same time. I read it directly from the Census Bureau and merge it with the CDPs data from Fairfax County using their shared GEOID (a US Census Code given to each census geographic area). The result is a zip code with the corresponding neighborhoods.
- **American Community Survey Age and Sex (2018 5-Year Estimates Data)**: I obtained this csv data by exploring the US Census Bureau's many tables.[9] This file has an estimated count of children under five years old per CDP. Due to time constraint, I was not able to learn the Census API to obtain the file programmatically. Instead, I downloaded the file and saved it for reading.
- **American Community Survey Comparative Economic Characteristics (2018 5-Year Estimates)**: This csv data available for download at the US Census Bureau Explore

6 "Census 2010 Designated Places." https://catalog.data.gov/dataset/census-2010-designated-place-c3875. Accessed 8 Jul. 2020.
7 "Census 2010 ZCTA to Place Relationship File." http://www2.census.gov/geo/docs/maps-data/data/rel/zcta_place_rel_10.txt?#. Accessed 8 Jul. 2020.
8 "On the use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data."
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762013/. Accessed 15 Jul. 2020.
9 "Explore Census Data." https://data.census.gov/cedsci/. Accessed 9 Jul 2020.

Census Data.[10] The key data points from this table are the number of families with children under 6 and families in which all parents are working.

- **Foursquare Data**: Using Foursquare Places API's explore endpoint and the daycare category (4f4532974b9074f6e4fb0104), I rely on the census data above to pull a list of daycares by each neighborhood and zip code.

In the next sections, I discuss my wrangling with these data. The majority of my work was conducted using IBM Studio, but after several technical difficulties including difficulties with needed library package installation, I moved back to using jupyter notebook on the Skills Network Labs with back up on the notebook on my own pc.

II.2. Neighborhood Geospatial Data Acquisition and Preparation

In this section, I explore two main sources of geographic data I used, namely the Fairfax County GIS data and the Census Bureau Data obtained through data.gov. When I initially read the file from the Fairfax County GIS page, it was because I had not realized that there was no column for the polygon geometry. I later obtained a new file from data.gov and I attempted to read this file and upload it into Watson without success. I decided to download it on my pc and read it using shapefile to convert it into a pandas dataframe[11]. I then created a csv version of the dataframe which I uploaded, first in Watson Studio when I was using the platform.

FIGURE 1. Fairfax County Neighborhoods' Geographical Coordinates Part I

```
s/64191f811f284007b27c6d29a9266411_3.csv' #Variable path to the online file
```

NAMELSAD	LSAD	CLASSFP	PCICBSA	PCINECTA	MTFCC	FUNCSTAT	ALAND	AWATER	INTPTLAT	INTPTLON	SHAPE_Length	SHAPE_Area
Herndon town	43	C1	N	N	G4110	A	11076010	17595	38.970412	-77.387073	0.146095	0.001153
Clifton town	43	C1	N	N	G4110	A	635487	10116	38.780128	-77.385990	0.033160	0.000067
Vienna town	43	C1	N	N	G4110	A	11415119	17193	38.898576	-77.258323	0.202853	0.001187
Annandale CDP	57	U1	N	N	G4210	S	20348939	30218	38.832801	-77.196229	0.236998	0.002114
Crossroads CDP	57	U2	N	N	G4210	S	5311343	717	38.848262	-77.131959	0.124045	0.000551

[10] Same as above

[11] Here are the instructions I followed these troubleshooting steps
https://towardsdatascience.com/mapping-geograph-data-in-python-610a963d2d7f

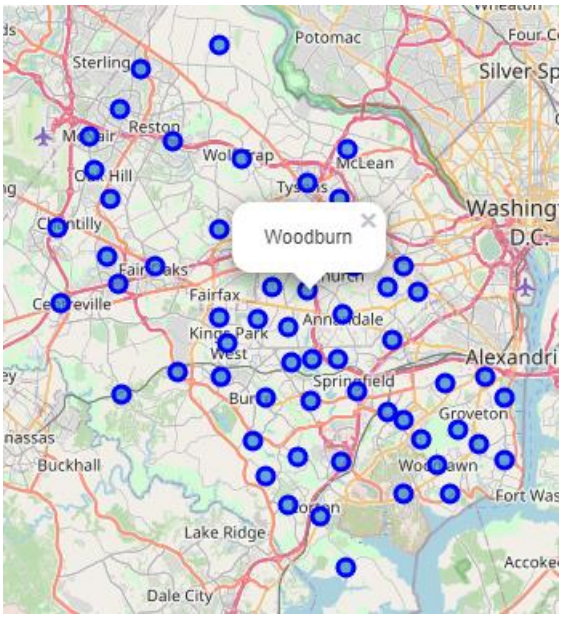
Figure 1 above shows data without multipolygon geometry coordinates. I later obtain a new file with these coordinates as shown below in Figure 2.

FIGURE 2. Fairfax County Neighborhood's Geographical Coordinates Part II

NAMELSAD	LSAD	CLASSFP	...	PCINECTA	MTFCC	FUNCSTAT	ALAND	AWATER	INTPTLAT	INTPTLON	SHAPE_Leng	SHAPE_Area	coords
Herndon town	43	C1	...	N	G4110	A	11076010.0	17595.0	38.970412	-77.387073	0.146095	0.001153	[(-77.41359241699996, 38.97163478600004), (-77...
Clifton town	43	C1	...	N	G4110	A	635487.0	10116.0	38.780128	-77.385990	0.033160	0.000067	[(-77.39284639799996, 38.78154174900004), (-77...
Vienna town	43	C1	...	N	G4110	A	11415119.0	17193.0	38.898576	-77.258323	0.202853	0.001187	[(-77.28476537199998, 38.900528777000034), (-7...
Annandale CDP	57	U1	...	N	G4210	S	20348939.0	30218.0	38.832801	-77.196229	0.236998	0.002114	[(-77.22214534799997, 38.820930763000035),

Although missing these coordinates, the data in Figure 1. was sufficient for an initial folium-generated visualization of the location of the 61 target neighborhoods.

FIGURE 3. Fairfax County 61 Neighborhoods on the Map



II.3. Population Characteristics Data

The first data acquisition for this part of the work consisted of a relationship file from zip code and CDPs. In acquiring this file that did not have the demographics information of interest, I had

assumed that more population data would be more available by zip code than it would be by CDPs. My assumption was based on the fact that zip codes are more popular than CDPs. In the end, I realized that data was more available by CDPs than by zip codes/ZCTA.

FIGURE 4. Zip Code to CDPs Relationship Data

```
#Obtain the file with relationship of zip to county places from the US Census Bureau
zip_c_relat = 'https://www2.census.gov/geo/docs/maps-data/data/rel/zcta_place_rel_10.txt?#'

zip_c_relat_df = pd.read_csv(zip_c_relat)
zip_c_relat_df.head()
```

	ZCTA5	STATE	PLACE	CLASSFP	GEOID	POPPT	HUPT	AREAPT	AREALANDPT	ZPOP	ZHU	ZAREA
0	601	72	358	U1	7200358	4406	1968	1942319	1942319	18570	7744	167459085
1	602	72	616	U1	7200616	3212	1648	955285	955285	41520	18073	83734431
2	602	72	47873	U1	7247873	2886	1205	4238649	4238649	41520	18073	83734431

Please note that the above ZCTAs include codes from other counties in Virginia. To obtain zip codes in Fairfax County, I merged the file with the Fairfax County file on GEOID. [12]

The second data I collected was the number of children under 5 by zip code. This population is a main target population for any daycare. All things being equal, a greater number of children under 5 means higher need for daycares. After obtaining this data, I merged it with While obtaining this data, a few neighborhoods had missing data either because it was not available for the period or because of what appears to be a GEOID mismatch.

FIGURE 5. Number of Children Under 5 in the Neighborhood

[12] There are 154 zip codes corresponding to the 61 neighborhoods.

```
#Read data on population by age to extract number of children under 5

df_data_children = pd.read_csv('VirginiaPlacesAgeandSex2018.csv')
df_data_children.head()
```

)01E	S0101_C04_001M	S0101_C05_001E	S0101_C05_001M	S0101_C06_001E	S0101_C06_001M	S0101_C01_002E	S0101_C01_002M	S0101_C02_002E	S0
'cent Total ation	Margin of Error!!Percent Male MOE!!Total popul...	Estimate!!Female!!Total population	Margin of Error!!Female MOE!!Total population	Estimate!!Percent Female!!Total population	Margin of Error!!Percent Female MOE!!Total pop...	Estimate!!Total!!Total population!!AGE!!Under ...	Margin of Error!!Total MOE!!Total population!!...	Estimate!!Percent!!Total population!!AGE!!Unde...	
(X)	(X)	4138	216	(X)	(X)	281	88	3.5	
(X)	(X)	197	50	(X)	(X)	17	19	3.8	
(X)	(X)	860	159	(X)	(X)	85	61	5.1	

On this file, the key field is the estimated total number of children under 5 living in the neighborhood (column S0101_C01_002E). After cleaning this data, I merge it with the geospatial data for later mapping. Similarly, I obtain the number of families with children under 6, the percentage of these families where all parents are working, and the median income for each neighborhood from the population characteristics file. Finally, I obtain the total population number for each neighborhood and merge it with the geospatial data. Figure 6 on the next page shows the data transformations that I performed on each file's data before merging the data with the data in the dataframe above.

FIGURE 6. Data Transformation

There were cdps without data as well as inconsistency in GEOIDs on the Census Bureau files which led to the NaN. I will have to drop all the NaN with hope I can fill those NaN in the future.

```
#Get Total Population Numbers
df_data_popt = pd.read_csv('TotalPop.csv')
df_data_popt.head()
```

	GEO_ID	NAME	B99123_001E	B99123_002E	B99123_003E
0	id	Geographic Area Name	Estimate!!Total	Estimate!!Total!!Allocated	Estimate!!Total!!Not allocated
1	1600000US5157531	North Shore CDP, Virginia	2841	115	2726
2	1600000US5131200	Glen Allen CDP, Virginia	13042	1156	11886
3	1600000US5143464	Lakeside CDP, Virginia	10174	531	9643
4	1600000US5103000	Arlington CDP, Virginia	196340	5633	190707

Here I just need B99123_001E. (see Census Bureau for allocation definition).

```
#Drop the extra heading
df_data_pot = df_data_popt.drop(0, axis=0)

#Edit the neighborhood GEOID to match the GEOIDs in the first dataframe
df_data_popt['GEO_ID'] = df_data_pot['GEO_ID'].str.split('S').str[1] #Take only characters after US

#Rename column GEO_ID to allow merging
df_data_popt.rename(columns={'GEO_ID':'GEOID'}, inplace= True)

#Add the column with estimated population count
ffx_neighborhoods_with_cwt = pd.merge(ffx_neighborhoods_with_cw,df_data_popt[['GEOID','B99123_001E']], how='left')

#ALso rename the column to make it more meaningful
ffx_neighborhoods_with_cwt.rename(columns={'B99123_001E':'pop_total'}, inplace= True)
ffx_neighborhoods_with_cwt.head()
```

After the merge, I performed further transformations to drop unneeded columns and obtain columns that will be the basis of modeling features selection.

FIGURE 7. Final Population Characteristics Table

```
# Removing the rows with NaN for now
ffx_neighborhoods_pop_char=ffx_neighborhoods_pop_char.dropna().reset_index(drop=True)
ffx_neighborhoods_pop_char.head()
```

	GEOID	NAME	INTPTLAT	INTPTLON	coords	num_ch_u_five	working_f_ch_u_six	pct_all_parents_work	median_income	pop_total
0	5136648	Herndon	38.970412	-77.387073	[(-77.41359241699996, 38.97163478600004), (-77...	1966	2116	66.3	112835	19333
1	5181072	Vienna	38.898576	-77.258323	[(-77.28476537199998, 38.900528777000034), (-7...	957	1319	68.2	155490	12732
2	5101912	Annandale	38.832801	-77.196229	[(-77.22214534799997, 38.820930763000035), (-7...	2947	3311	73.6	90545	35586
3	5104088	Bailey's Crossroads	38.848262	-77.131959	[(-77.14751432699995, 38.85212377200003), (-77...	2086	2121	47.8	64736	19401
4	5105928	Belle Haven	38.777339	-77.057780	[(-77.08112830099998, 38.77294475900004), (-77...	322	405	79	106491	5622

```
#Change columns to numerical values
cols = ['num_ch_u_five','working_f_ch_u_six','pct_all_parents_work','median_income', 'pop_total']
ffx_neighborhoods_pop_char[cols]=ffx_neighborhoods_pop_char[cols].apply(pd.to_numeric, errors='coerce')
```

Please note that I dropped the neighborhoods (rows) that had missing data. There were six neighborhoods that had missing key data such as the number of children under 5. Because I could not identify the best way to estimate a number to assign to the neighborhoods and also because some files had mismatch on GEOIDs on some of the neighborhoods, I decided to remove missing values from the entire table. Figure 8 shows summary statistics for the data emerging from acquisition and preparation stage.

Figure 8. Data Summary Statistics

```
#Obtain summary stats
ffx_neighborhoods_pop_char.describe()
```

	INTPTLAT	INTPTLON	num_ch_u_five	working_f_ch_u_six	pct_all_parents_work	median_income	pop_total
count	55.000000	55.000000	55.00	55.000000	55.000000	55.000000	55.000000
mean	38.832388	-77.242380	1218.60	1434.654545	68.209091	131019.545455	15166.927273
std	0.078143	0.101741	915.12	1065.660536	10.051023	39999.372786	11263.015360
min	38.698722	-77.440879	206.00	259.000000	43.200000	56706.000000	3886.000000
25%	38.776267	-77.302098	559.00	648.500000	61.800000	106402.000000	6915.000000
50%	38.831565	-77.239668	976.00	1236.000000	67.900000	123074.000000	12537.000000
75%	38.889326	-77.171332	1583.50	1897.500000	75.750000	147927.000000	18185.500000
max	39.012386	-77.057664	5014.00	5932.000000	86.900000	228836.000000	59921.000000

Next, I perform some exploratory analysis on these data points. But before this, recall that the geospatial data I used above had problems with within geometry coordinates. This made it impossible to render choropleth maps. To solve this problem, I installed geopandas (easier on Skills Network Labs) and was successful in reading the file in its geojson format directly from the Fairfax County Website.

Figure 9. Obtaining Geospatial Data

```
import geopandas as gpd
#Create variable url with path to the file
geourl='http://data-fairfaxcountygis.opendata.arcgis.com/datasets/64191f811f284007b27c6d29a9266411_3.geojson'

#Examine a few rows
ffx_geo=gpd.read_file(geourl)
```

```
ffx_geo.head()
```

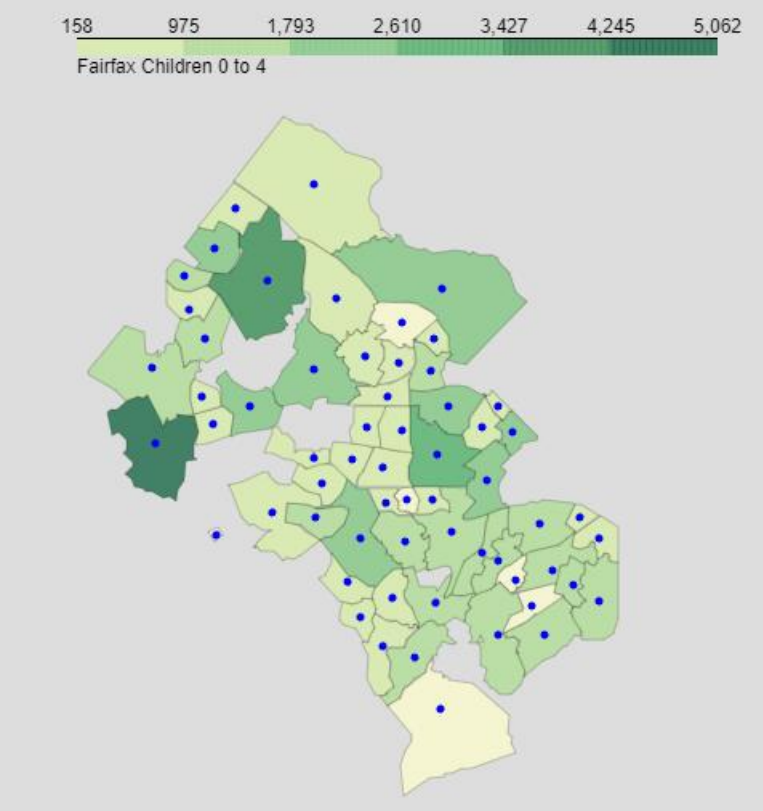
	FID	STATEFP	PLACEFP	PLACENS	GEOID	NAME	NAMELSAD	LSAD	CLASSFP	PCICBSA	PCINECTA	MTFCC	FUNCSTAT	ALAND	AWATER	INTPTLAT	INTPTLON	SHAPE_Length	SHAPE_Area	geometry
0	1	51	36648	02390244	5136648	Herndon	Herndon town	43	C1	N	N	G4110	A	11076010.0	17595.0	+38.9704119	-077.3870725	0.146095	0.001153	MULTIPOLYGON (((-77.41359 38.97163, -77.41292 ...
1	2	51	17376	02390803	5117376	Clifton	Clifton town	43	C1	N	N	G4110	A	635487.0	10116.0	+38.7801282	-077.3859899	0.033160	0.000067	MULTIPOLYGON (((-77.39285 38.78154, -77.39138 ...
2	3	51	81072	02391461	5181072	Vienna	Vienna town	43	C1	N	N	G4110	A	11415119.0	17193.0	+38.8985761	-077.2583226	0.202853	0.001187	MULTIPOLYGON (((-77.28477 38.90053, -77.28440 ...

III. Exploratory Analysis of Population Characteristics Data

After acquiring the data, I began by exploring a few key data of interest. I chose the number of children under 5 and neighborhood median income as these appear to be the most unrelated by comparison (e.g., the number of children under 5 would seem to be related to the number of families with children in the area).

Figure 5 is a geographical representation of where children under 5 live. The greater the number of children, the darker the green.

FIGURE 10. Children Under Five in Fairfax County Neighborhoods

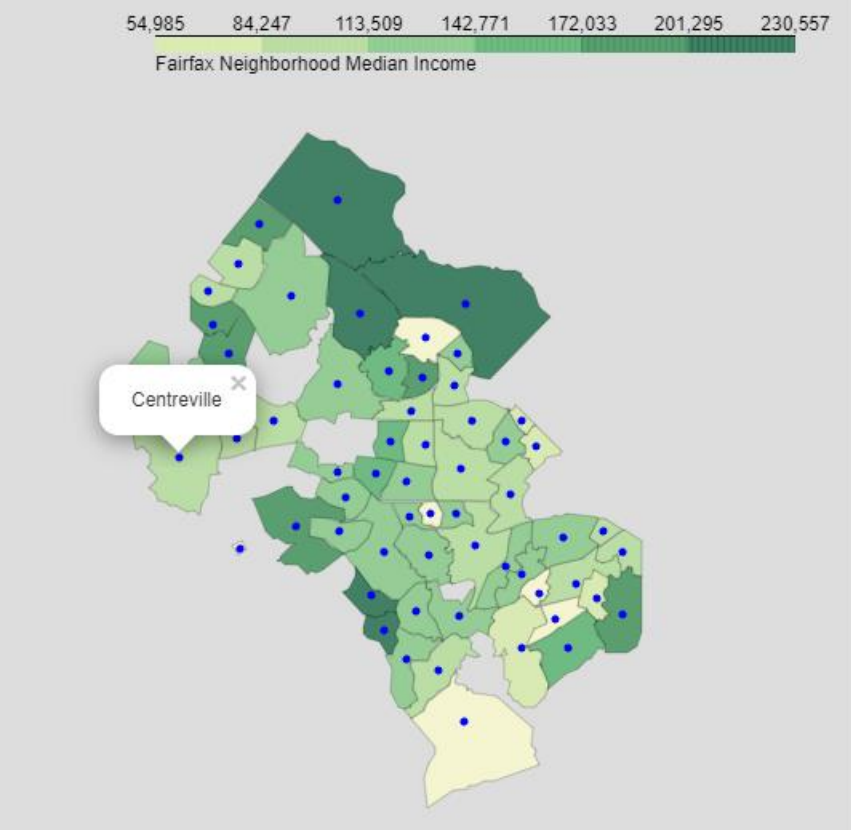


If the businessperson is choosing a location based on just where children live, Reston and Centreville would be the primary candidates followed by Annandale, of course, assuming the ratio of daycares to population under 5 is small. Obviously, other factors will come into play, not least the fact that families in these areas may not have both parents working because one of the parents is taking care of their under five children.

Following the analysis that includes other factors (i.e., other things not equal), it will be interesting to know if neighborhoods with greater number of children under 5 are also the most favorable for a new daycare.

Another question to explore is whether children live where the money is because money is an important factor for any businesses.

FIGURE 11. Incomes in Fairfax County Neighborhoods



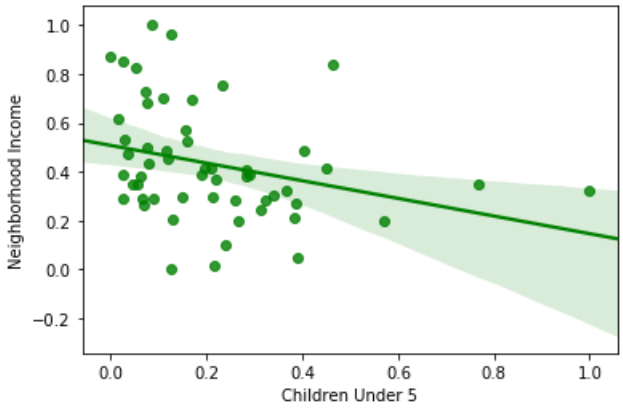
It is not clear whether children live where the money is. In fact, some of the neighborhoods that stood out above like Centreville and Reston when it comes to number of children under 5 do not appear to be the greenest when it comes to income. Neighborhoods such as Great Falls without children under 5 emerge. However, there are also neighborhoods such as Mason Mack in the South that are both yellow on the first map and the second.

Using MinMax to normalize the columns of interest, I further explore this relationship between income and number of children under 5.

FIGURE 12. Do Children Under Five Tend to Live Where the Money Is?

	Nmzed Neighborhood Income	Nmzed Children Under 5
0	0.326085	0.366057
1	0.573892	0.156198
2	0.196590	0.570092
3	0.046651	0.391015
4	0.289229	0.024126

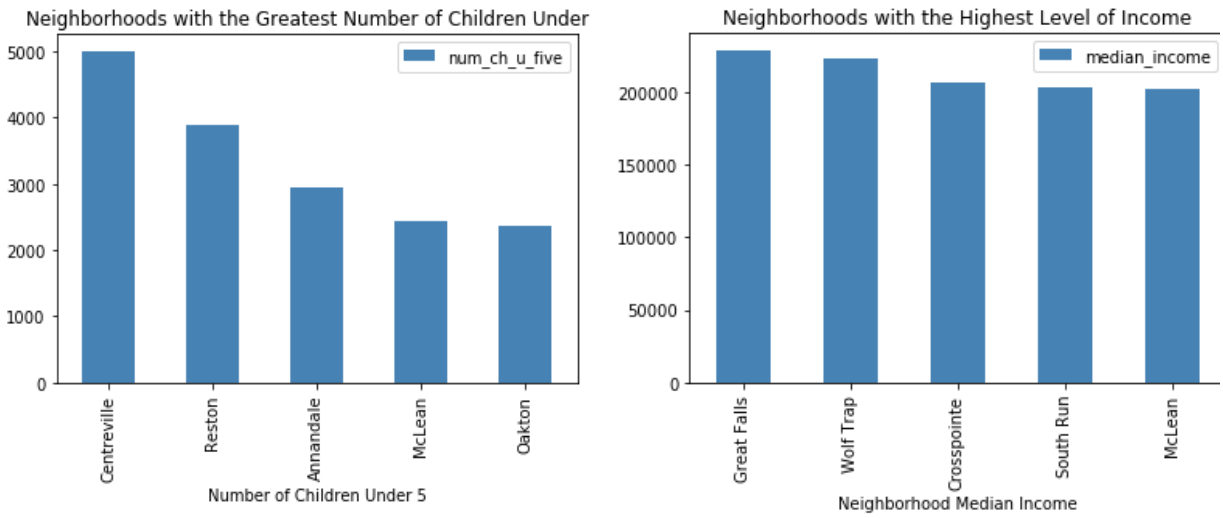
```
import seaborn as sns
ax = sns.regplot(x='Children Under 5', y='Neighborhood Incc
```



Even though the relationship is not straightforward, it sure seems like more children live in areas with middle income by area standard. My guess (also backed by superficial knowledge of the area) is that in terms of areas with smaller numbers of children, areas with higher incomes have more empty nesters, whereas the areas with lower income have a younger population. This hunch can be validated by additional demographic data, but that is beyond the scope of this analysis.

Below are bar graphs to keep in mind when clusters are complete. Will the best neighborhoods be selected among high income neighborhoods or among neighborhoods with the greatest number of children?

FIGURE 13. Top Neighborhoods by Income and by Number of Children



IV. Data Part II: Foursquare Business Data

Foursquare is the only location data I use to obtain the listing of daycares in the neighborhoods. Using the Foursquare's Explore Endpoint and the daycare category ID, I obtain daycares within a 5-kilometer radius of each neighborhood latitude and longitude. I made this choice of radius with the assumption that 5 kilometers is a good distance for a family with children under 5 who would like to choose a daycare.

FIGURE 14. Daycares in Fairfax County

Neighborhood	Daycare	d_id	Venue Latitude	Venue Longitude	Venue Type
Herndon	Gold's Gym	4b2acb5bf964a520d0af24e3	38.952288	-77.349164	Gym / Fitness Center
Herndon	Herndon KinderCare	4e60de9bd164ddd5e5125c7c	38.968380	-77.388485	Daycare
Herndon	Mommy Sanna Home Daycare	5ad47d37b9b37b6832eedf04	38.982489	-77.378671	Daycare
Herndon	Kumon Math and Reading Center of Herndon - Eld...	4dbb193d4df044e524c1cf9e	38.967200	-77.397565	Daycare
Herndon	Herndon Parkway KinderCare	4da3225cd686b60c1874c628	38.964812	-77.400936	Daycare

Neighborhood	Daycare	d_id	Venue Latitude	Venue Longitude	Venue Type
Herndon	Herndon KinderCare	4e60de9bd164ddd5e5125c7c	38.968380	-77.388485	Daycare
Herndon	Mommy Sanna Home Daycare	5ad47d37b9b37b6832eedf04	38.982489	-77.378671	Daycare
Herndon	Kumon Math and Reading Center of Herndon - Eld...	4dbb193d4df044e524c1cf9e	38.967200	-77.397565	Daycare
Herndon	Herndon Parkway KinderCare	4da3225cd686b60c1874c628	38.964812	-77.400936	Daycare
Herndon	Common Ground Child Care Center	4bcef43ab6c49c74d5de9791	38.965489	-77.352892	Daycare

The API call returned **615** daycares. These included gym-based daycares which I excluded. The result was a list of **566** located in all of the 61 Fairfax neighborhoods.

Data preparation included merging the data with the population characteristics data, and expressing the number of daycares in a neighborhood as a ratio of key population characteristic.

FIGURE 15. Obtaining the Number of Children per Daycare

```
#Add column with number of children under 5 per daycare
ffx_neighborhoods_pop_char['num_c_per_dayc']=(ffx_neighborhoods_pop_char['num_ch_u_five']/ffx_neighborhoods_pop_char['num_dcs']).astype('int')
#Add column with number of families with children under 6 in which all parents are working
ffx_neighborhoods_pop_char['num_f_c_u_w']=(ffx_neighborhoods_pop_char['pct_all_parents_work']*ffx_neighborhoods_pop_char['working_f_ch_u_six']/100).astype('int')
ffx_neighborhoods_pop_char.head()
```

	GEOID	NAME	INTPTLAT	INTPTLON	coords	num_ch_u_five	working_f_ch_u_six	pct_all_parents_work	median_income	pop_total	num_dcs	num_c_per_dayc	num_f_c_u_w
0	5136648	Herndon	38.970412	-77.387073	[(-77.41359241699996, 38.97163478600004), (-77...	1966	2116	66.3	112835	19333	22	89	1402
1	5181072	Vienna	38.898576	-77.258323	[(-77.28476537199998, 38.900528777000034), (-7...	957	1319	68.2	155490	12732	15	63	899
2	5101912	Annandale	38.832801	-77.196229	[(-77.22214534799997, 38.820930763000035), (-7...	2947	3311	73.6	90545	35586	10	294	2436
3	5104088	Bailey's Crossroads	38.848262	-77.131959	[(-77.14751432699995, 38.85212377200003), (-77...	2086	2121	47.8	64736	19401	15	139	1013
4	5105928	Belle Haven	38.777339	-77.057780	[(-77.08112830099998, 38.77294475900004), (-77...	322	405	79.0	106491	5622	11	29	319

A key data point emerged from these calculations: a neighborhood's number of children per daycare. Instead of looking at the number of daycares in the area, I will use this new number when choosing the best neighborhood for a daycare. In the future, I hope to use business license data and other data that would indicate business success in a particular neighborhood.

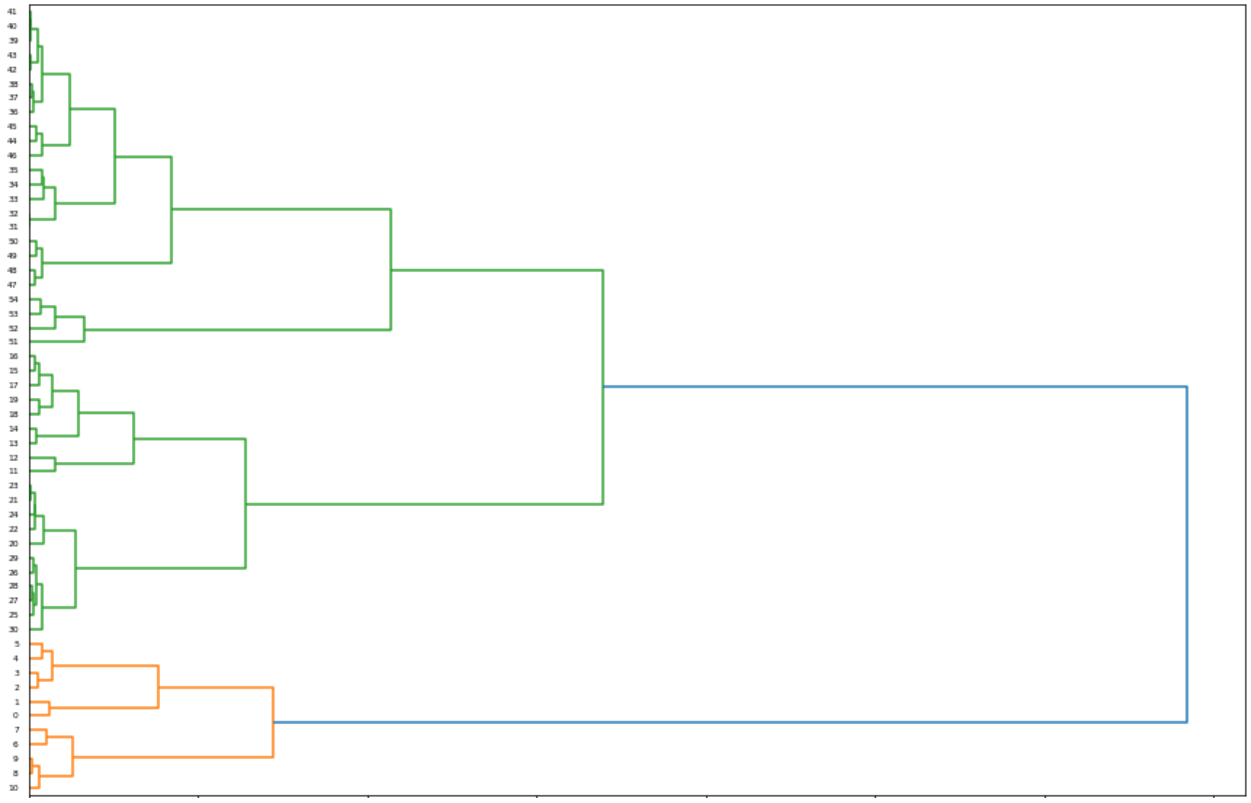
V. Modeling

For modeling, I use AgglomerativeClustering from python's sklearn library to put the neighborhoods into 4 clusters. To choose the number of clusters, I use the dendrogram method to settle on the 4-clusters level (see Figure 16 below).[13] In general, clustering was my choice because the question to answer is that of neighborhood segmentation. Specifically, I chose this type of hierarchical clustering because of the size of the dataset.

The agglomerative algorithm returned a model which I then used to cluster neighborhoods into groups of similar neighborhoods.

[13] Note that 3 clusters could have been chosen as well but given the data and the purpose of this analysis, I decided on 4 clusters.

FIGURE 16. Using the Dendrogram to Obtain the Optimal Number of Clusters



V.1. Results

Interesting labels or groups emerge. Looking at the population data, the two clusters of neighborhoods that emerged stand out by average characteristics.

Figure 16. Summary Characteristics of the neighborhoods

clusters_	num_ch_u_five	median_income	num_f_c_u_w	num_c_per_dayc
0	535	148496	453	50
1	3072	115129	2469	282
2	1386	107936	1040	148
3	1457	205779	1210	350

- Cluster 0: Neighborhoods have a small number of children, incomes above the area average, a small number of families with children under 5 where both parents are working, and the number of childcares appears sufficient. All things being equal, neighborhoods in this cluster would not be on the top of the list for consideration.
- Cluster 1: Neighborhoods have the highest number of children under 5, incomes below area average and median, the highest number of families with both parents working and a high density of children under 5 per daycare. This area would be very attractive especially for someone interested in finding opportunities to serve low income working families.
- Cluster 2: Neighborhoods have a relatively smaller number of children, the lowest level of income, a relatively smaller number of working families with children, and a smaller number of children per daycare compared to cluster 2. I would choose cluster 1 over cluster 2 because they are almost similar in terms of income level but cluster 2 has a higher density of children (almost double) in daycare by comparison.
- Cluster 3: Neighborhoods have a number of children under 5 toward the average, by far the highest level of income, the second highest number of working families with children under 6 where all parents are working, and the highest density of children under five per daycare. At first glance, this area presents a good opportunity for someone whose highest priority is financial stability, although, of course, given the high area income, start-up costs may be high, and facilities may be expensive.

V.2. Choosing the Best Neighborhood

In the section above, neighborhoods were placed into 4 clusters. I further identified two clusters of neighborhoods that appear to present the best opportunity (cluster 1 and 3) for a businessperson depending on the businessperson priority. In this section, I explore the two priority clusters to recommend one neighborhood within each. This selection would simply be based on the number of children by number of daycares nearby the neighborhood.

FIGURE 17. The Best Neighborhoods from the Best Clusters

```
#Examine neighborhoods in cluster 1
low_income_priority= ffx_neighborhoods_pop_char[ffx_neighborhoods_pop_char.clusters_==1].reset_index(drop=True)
low_income_priority
```

	GEOID	NAME	INTPTLAT	INTPTLON	coords	num_ch_u_five	working_f_ch_u_six	pct_all_parents_work	median_income	pop_total	num_dcs	num_c_per_dayc	num_f_c_u_w	clusters_
0	5111464	Burke	38.777769	-77.262617	[(-77.30126837099994, 38.79752775500003), (-77...	2149	2601	77.0	140229	34540	9	238	2002	1
1	5158472	Oakton	38.889444	-77.302226	[(-77.33867038799997, 38.877674770000056), (-7...	2371	2926	63.2	128504	29911	8	296	1849	1
2	5166672	Reston	38.948449	-77.342215	[(-77.39326140899993, 38.94523878200005), (-77...	3896	4540	64.7	116375	49986	20	194	2937	1
3	5114440	Centreville	38.840650	-77.438462	[(-77.47956242999999, 38.86152976200003), (-77...	5014	5932	66.8	112174	59921	12	417	3962	1
4	5184368	West Falls Church	38.865559	-77.187699	[(-77.22043935099998, 38.87583577400005), (-77...	2056	2369	68.8	102952	23824	8	257	1629	1
5	5101912	Annandale	38.832801	-77.196229	[(-77.22214534799997, 38.82093076300035), (-7...	2947	3311	73.6	90545	35586	10	294	2436	1

Clearly **Centerville** presents the best opportunity with a small number of daycares, highest number of children per daycare, and a very high number of families with both parents working. This neighborhood corresponds to the neighborhood with the greatest number of children that were discussed at the beginning.

```
#Examine neighborhoods in cluster 2
high_income_priority= ffx_neighborhoods_pop_char[ffx_neighborhoods_pop_char.clusters_==2].reset_index(drop=True)
high_income_priority
```

	GEOID	NAME	INTPTLAT	INTPTLON	coords	num_ch_u_five	working_f_ch_u_six	pct_all_parents_work	median_income	pop_total	num_dcs	num_c_per_dayc	num_f_c_u_w	clusters_
0	5132496	Great Falls	39.012386	-77.301969	[(-77.37097840599995, 39.01461679600004), (-77...	610	829	69.1	228836	12537	2	305	572	2
1	5148376	McLean	38.943545	-77.192913	[(-77.28777337699995, 38.966140790000054), (-7...	2430	2964	58.2	201570	37220	8	303	1725	2
2	5129136	Fort Hunt	38.735462	-77.057664	[(-77.08219029899999, 38.72865875000008), (-77...	1332	1656	80.6	186931	12885	3	444	1334	2

In this cluster, **Fort Hunt** comes on top with the highest number of children per daycare and the highest percentage of families with children under 6 with all parents working. I will have to investigate this cluster further because it appears to be an outlier.

VI. Conclusion

In this report, I have shown how a businessperson can take advantage of data science advance location search tasks for a planned new business. Guided by the CRISP-DM methodology, my objective was to conduct an analysis that leaves a manageable number of neighborhoods with which the businessperson can start. After defining this objective, I collected demographic and geospatial data in different formats from the Census Bureau and Fairfax County websites. I combined this data with daycare location data I collected through a Foursquare API call, and prepared it for modeling. A model was generated using agglomerative clustering, and I

deployed it to create five clusters of similar neighborhoods from which two clusters and one neighborhood from each of the two clusters was chosen.

Key Takeaways:

The analysis presented here was exploratory in nature and the model obtained should be further evaluated. However, even in this preliminary stage, I achieved my objective and answered the business question was answered. The success of this analysis is three-fold:

- Two priority neighborhoods, namely, **Centreville** and **Fort Hunt**, emerged. This reduces the amount of work necessary for location exploration. With further model evaluation of and other algorithms, an even more tight model may emerge with even more actionable clusters.

- The analysis clearly points to areas for further market research opportunities. New data such as data on recent marriage licenses and additional age groups, can help identify even more areas of opportunity. Business license data would show how successful daycares in given neighborhoods have been, which can further enhance similarity measurement within neighborhood clusters.

- This work shows that data science is not just for big businesses and big data: even small businesses and individuals can benefit from data science and recent advances in machine learning.

With decennial census data coming in the next few months, building on this inquiry will help any new business, especially small businesses without resources to develop expensive analytical capabilities.