Franchise Opportunity in Arlington, VA

## Introduction

My brother and I are looking for a franchise opportunity in Arlington, Virginia, and want to decide which to buy and where to place it. The analysis in the paper of competitive forces in Arlington franchise markets would also be valuable to commercial real estate brokers interested in finding tenants for retail space in offices and shopping centers.

## Background

Arlington County is across the Potomac River southwest of Washington, DC. When George Washington and Pierre Le Enfant laid-out the Federal District to house the national government, they planned a city 10 miles on each side of a square, which included the port cities of Alexandria in Virginia, and Georgetown in Maryland. During the Civil War, the previously-Virginia part was returned to Virginia, and that area, approximately 26 mi2, became the County (and city) of Arlington. The Pentagon, home of the Department of Defense, and Arlington National Cemetery, as well as Reagan National Airport (DCA), are located in Arlington.

Arlington is the smallest self-governing county in the United States.[[1]](#endnote-1) The city has a young, well educated workforce. Almost 3 in 4 people more than 25-years-old have a college degree[[2]](#endnote-2). About 93% of High School graduates in the county attend college[[3]](#endnote-3). The median income in Arlington is $112,138, and the per capita income is $67,061.

The Orange Line of the Washington Area Metropolitan Commuter Rail, known as “the Metro” has been a major source of development Arlington in the previous 30 years. Arlington Metro Stations along the Orange line include Ballston-MU, Virginia Square-GMU, Clarendon, Court House and Rosslyn. The Blue Line stations in Arlington include Reagan National Airport, Crystal City, Pentagon City, Pentagon, Arlington Cemetery and Rosslyn (transfer to Orange Line).

## The problem

The expectation is that a logistic regression algorithm should be able to identify the metro stations because of their unique patterns of venues, relative to the more residential neighborhoods. Moreover, there may be neighborhoods that look very much like Metro stations, (i.e., false positives) that may be interesting.

An examination of the venues at Metro stations (and look-alike neighborhoods) should guide decisions about franchise placement.

## Data

#### Data Gathering

Data was gathered in three steps: 1) Identify Arlington neighborhoods; 2) get geo-coordinates of the neighborhoods; and 3) find venues in each neighborhood. After that, we will look at the patterns of neighborhood, again, expecting Metro stations to be different from other neighborhoods.

Arlington neighborhoods were identified from Wikipedia, ["List of Neighborhoods in Arlington County, Virginia"](https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Arlington_County,_Virginia)[[4]](#endnote-4). This data was compared with zip codes, but zip codes were not used, essentially because the neighborhoods were more finely grained data. Census tract data was also reviewed by not used, except in summary from the Census.gov/QuickFacts/. In future projects, I’ll look forward to exploring the Census API.

The neighborhoods’ geo-coordinates were identified using “geocoders” in the Python package “Geopy”.

#### Data Cleansing

Several issues arose with the neighborhood list from Wikipedia. First, three areas identified in the list are not in Arlington: Forest Hills, Crystal Gateway, and Williamsburg; Crystal Gateway sounds like Crystal City, which *is* in Arlington, and Williamsburg Village, not Williamsburg, is in Arlington. In addition, Randolph Square is a condominium in Shirlington, so it was deemed redundant to Shirlington. Shirlington was used rather than Shirlington Crest, from the list. Ballston was used, and Gates of Ballston was omitted.

Some of the neighborhoods listed did not yield coordinates or yielded obviously incorrect coordinates. At a high level, Arlington is at 38.x latitude and -77.x longitude. Anything too far from that is an obvious error. The following neighborhoods did not yield valid coordinates: Arlington Ridge, Ashton Heights, Boulevard Manor, Country Club Hills, Forest Glen, Lacey Forest, Station Square, Willet Heights. My assumption is that these are the names apartments or housing developments, not recognizable by the geocoder.

Some of the listed neighborhoods had multiple names, and only one of them gave valid coordinates. For example, for “Waycroft-Woodlawn”, only Woodlawn gave an address in Arlington. For the neighborhood named “Nauck (Green Valley A.K.A. The Valley)”, Green Valley and The Valley gave identical coordinates (so the Valley was omitted), but Nauck and Green Valley gave non-overlapping venues. High View Park and Halls Hill gave identical coordinates: High View Park was omitted. Greenbrier and Garden City gave different neighborhood coordinates, but they were so close together, they gave identical neighborhood venues.

The final issue of duplicates arises from the simple fact that two distinct neighborhoods may be adjacent to the same venues on their borders. I sorted venues by latitude and longitude, and then by venue name, and finally by neighborhood name, and identified duplicates among those values. I kept “the first one”, essentially the one earliest in the alphabet. For example, I kept Ballston and let Carlin Springs be deleted; however, Carlin Springs venues that were not also in Ballston were retained. Again, with Ballston as an example, this procedure resulted in Metro Stations being retained, and non-Metro stations being deleted.

The neighborhood of Old Glebe is an established residential neighborhood, in the North of Arlington. The geocoder gave it coordinates identical to Prospect House, which is nowhere near it. So, I looked-up Old Glebe on Google, and overrode the coordinates in the data.

#### The Final Data

When the data cleansing was complete, there were 789 unique venues, in 54 neighborhoods. The venues fell into 186 categories, but which I manually summarized into

|  |  |
| --- | --- |
| Bank | 15 |
| Construction & Landscaping | 4 |
| Food Truck | 16 |
| Health & Recreation | 167 |
| Hotel | 21 |
| Market | 53 |
| nothing | 17 |
| On-premise Food/Drink | 366 |
| Rental Car Location | 6 |
| Store | 118 |
| Theater | 4 |

On-premise Food/Drink is a fancy way of denoting a restaurant. I plan to break these down further in the results section.

The Poop-Deck is listed as a “Harbor / Marina”; it is somebody’s apartment that he (it must be a he) has created into a social media phenomenon. It is no more a “Harbor / Marina” than it is a Movie theatre. I kept it, but I listed it as if it were a bar: Whatever, Pal.

Only two of the x neighborhoods that are in fact Metro stations were so-designated; I deleted the two and designated all Metro stations manually. That is, the data set will be manually labeled. A c classification algorithm, logistical regression, will be used to try to identify the features that designate whether or not a neighborhood is a classification algorithm, and the extent to which the features are indicative.

We’ve rejected all the usual candidates, because, as we’ve been surprised to learn, the city is extremely overdeveloped with fast-food and other franchises. Thus, in Arlington, we’ve rejected: McDonalds (3), Taco Bell (3), Wendy’s (3) Burger King (2), Chick-fil-A (2). Chipotle is not burgers, but it has 6 units in Arlington. California Tortilla, which I’ve never heard of, has two units in Arlington.

In addition, we’ve rejected the “upscale burger” market, because these too are over-represented: Shake Shack (3); Elevation Burger (2), which are fantastic and very tempting despite being over-represented, Big Buns (2), which I like, and I can not lie; it’s is such a great name, I might buy it just for that. Five Guys has only one unit in Arlington, but is all over northern Virginia; BRG-The Burger Joint has only one unit in Arlington,

For bakeries, there are … ; I have no idea what CAVA is, but it’s a competitor.

awhere you discuss the business problem and who would be interested in this project.

## Data

* + where you describe the data that will be used to solve the problem and the source of the data.

## Methodology

* + section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

## Results

* + section where you discuss the results.

## Discussion

* + section where you discuss any observations you noted and any recommendations you can make based on the results.

## Conclusion

* + section where you conclude the report.

1. <https://projects.arlingtonva.us/data-research/fast-facts/>, Downloaded 9/22/2019 [↑](#endnote-ref-1)
2. <https://www.census.gov/quickfacts/arlingtoncountyvirginia>, Downloaded 9/22/2019 [↑](#endnote-ref-2)
3. <https://projects.arlingtonva.us/data-research/fast-facts/>, Downloaded 9/22/2019 [↑](#endnote-ref-3)
4. <https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Arlington_County,_Virginia>, Date referenced 9/22/2019

   List of Neighborhoods in Arlington County, Virginia [↑](#endnote-ref-4)