# **Trending FourSquare Venues in COVID-19 Hot Spot Zip Codes in Virginia, USA**

### H. Diana McSpadden

### December 30, 2020

## **1. Introduction**

**1.1 Background**

According to the Virginia Department of Health COVID-19 is spread through respiratory droplets, and is more likely spread when people are in close contact. "Community spread" is the public health term used to describe an infection when the causative contact with a sick person is unknown. Community spread may be occuring Virginia communities at public venues such as restaurants, bars, coffee shops, and grocery stores.

**1.2 The Question**

Determining what categories of venues are correlated with above average per capita COVID-19 infection rates could be of great utility for communities to determine risk and possible locations of community spread. My project uses FourSquare venues by Virginia zip codes to determine if such a correlation exists. Machine Learning, specifically the Support Vector Machine (SVM) and Logistic Regression methods, are used to determine if above average COVID-19 per capita infections can be predicted based on types of FourSquare venues by zip code.

## **2. Data Acquisition and Preparation**

**2.1 Data Sources**

Four data sources are utilized:

**2.1.1 World Population Review**: Population by Virginia Zip Code

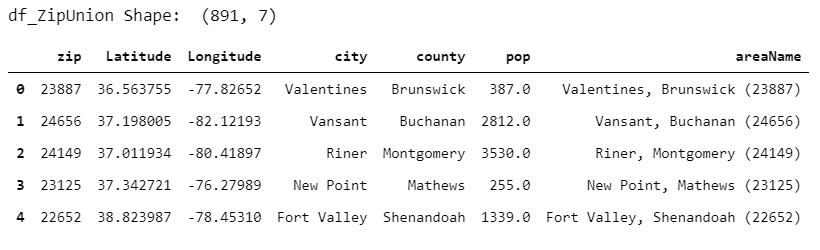
World Population Review provides a csv download of Virginia Zip Codes and estimates of the 2020 population for each zip code. These data are used to seed the dataset of Virginia zip codes, and to determine per capita COVID-19 infections rate by zip code. The csv was loaded into a pandas DataFrame that included zip code, city and county names, and population. The zip code was cast as a string data type to facilitate later merging with additional data sources. **892** records were collected.

**2.1.2 Open Data Standard API**: Zip code Latitude and Longitude

The FourSquare API does accept zip code as a location identifier, but I wanted practice at retrieving latitude and longitude values. Open Data Standard provides an API to acquire a latitude and longitude for each Virginia zip code. The data was retrieved via API as a JSON response and loaded into a pandas DataFrame.

1275 records were retrieved and the results were merged with the zip-population DataFrame resulting in an 891 record DataFrame due to some zip codes not being represented in both data sets. An “areaName” column was calculated for readability: CITY, COUNTY (ZIP CODE).

**Figure 1: Location DataFrame**



**2.1.3 Virginia's Open Data Portal:** COVID-19 Counts by Zip Code

The Virginia Department of Health provides cumulative COVID-19 counts by Virginia zip code. These data are available as a csv download and are used to calculate the per capita COVID-19 infections per zip code. The download included case counts for the date range 5/15/2020 - 12-27-2020. The data also included counts of number of testing encounters and number of PCR testing encounters.

I first cast the date column as a datetime. Then, I dropped records before December 1, 2020 and the testing encounter counts. There were records that had been suppressed to protect patient privacy in zip codes with low population counts. I removed rows with suppressed infection counts.

To calculate the number of cases per zip code from December 1, 2002 to December 27, 2020 I grouped the DataFrame by zip and used the minimum count and maximum count for each zip to calculate the zip code infection count for the date range.

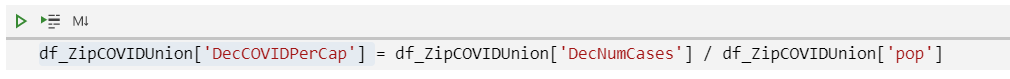
**Figure 2: Calculating Number of December 2020 COVID Infections by Zipcode**

|  |  |
| --- | --- |
|  |  |

I then merged the December 2020 case count DataFrame with the location DataFrame. Since some Virginia zip codes had no infections in the December timeframe and were not present in the DataFrame, a left join from the location DataFrame was used, and NaN’s were assigned 0 infections.

The per capita December infection rate was calculated by dividing DecemberNumCases by population.

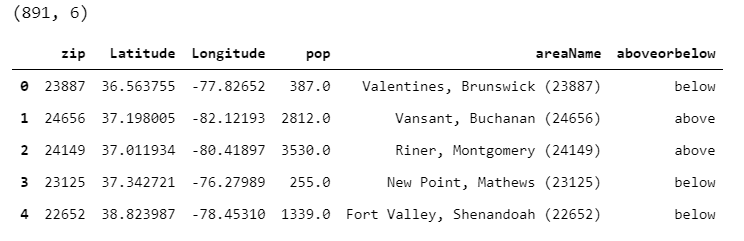
**Figure 3: Calculating Per Capita December Cases**



The mean per capita December infection rate was calculated for Virginia zip codes and was **0.010908115737950074.**

A column was calculated for the merged DataFrame that assigned “above” to zip codes with above average December infections, and “below” for zip codes with below average December infections. The ‘city’, ‘county’, ‘DecNumCases’, and ‘DecCOVIDPerCap’ columns were dropped from the DataFrame.

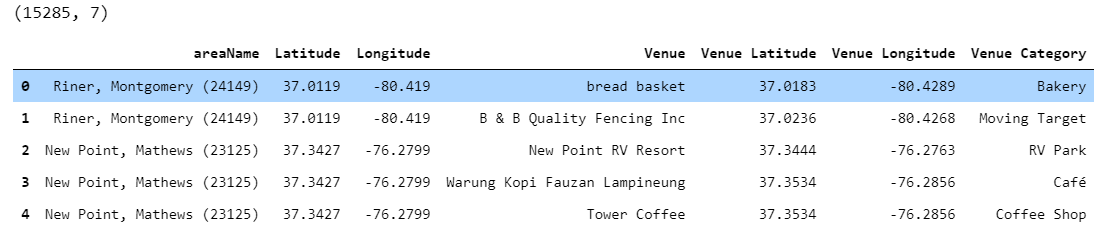
**Figure 4: Virginia Zip Code COVID Infection DataFrame**



**2.1.4 FourSquare Venues Endpoint:** Exploring Venues Near Zip Code

FourSquare provides venues within 2000m of a zip code. The FourSquare API returns the categories for each venue. 15,285 venues were retrieved near Virginia zip codes.

**Figure 5: FourSqaure Venues DataFrame**



## **3. Exploratory Data Analysis**

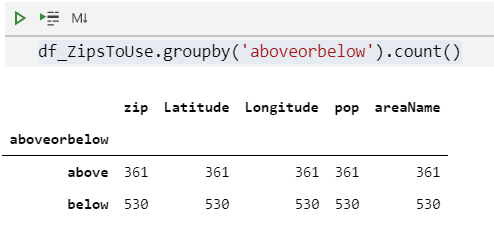
Before training and testing the machine learning algorithms I investigated features of the data.

**3.1 Exploring Virginia Zip Codes With Above Or Below Average December per capita COVID infections**

**3.1.1 QUESTION: How many Virginia zip codes were above or below the per capita average COVID infection rate?**

As seen in Figure 6, **361** zip codes have above average December per capita infection rates, and **530** Virginia zip codes have below average per capita infection rates.

**Figure 6: Above and Below per capita Average**



**3.1.2 QUESTION: How are these zip codes spread throughout the state?**

The folium library was utilized to visualize the above and below zip codes on a map of Virginia. The red dots are zip codes with above average December per capita infections and the blue dots are below average December per capita infections.

**Figure 7: Map of Above And Below Average December 2020 per capita COVID-19 Infections**

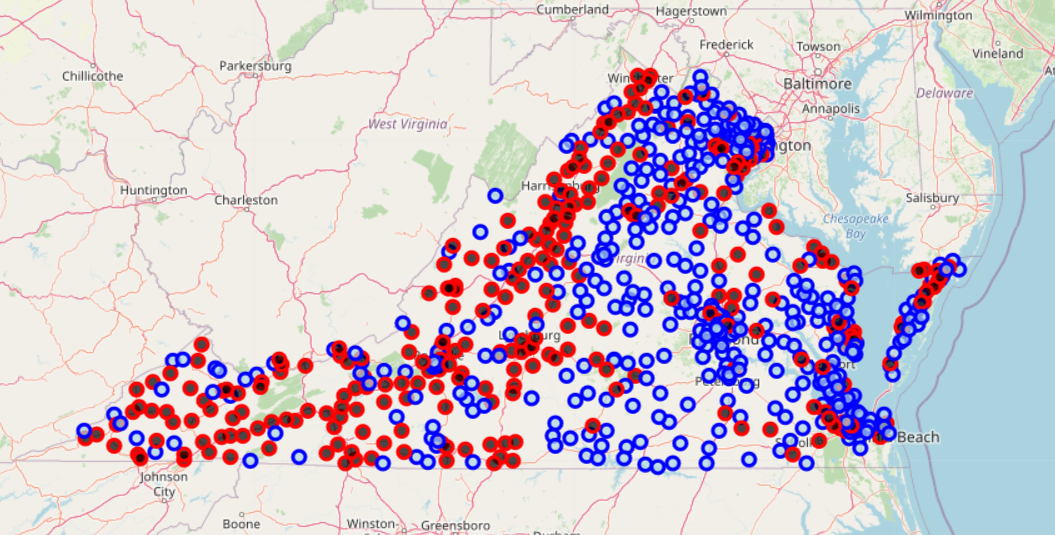


Figure 7 clearly shows that zip codes with above and below average infections are intermingled; however, the rural Southwest of the state has a majority of above average infections.

**3.1.3 QUESTION: Did the infection rate in Virginia change throughout December 2020?**

**Figure 8: Virginia case counts throughout December 2020**

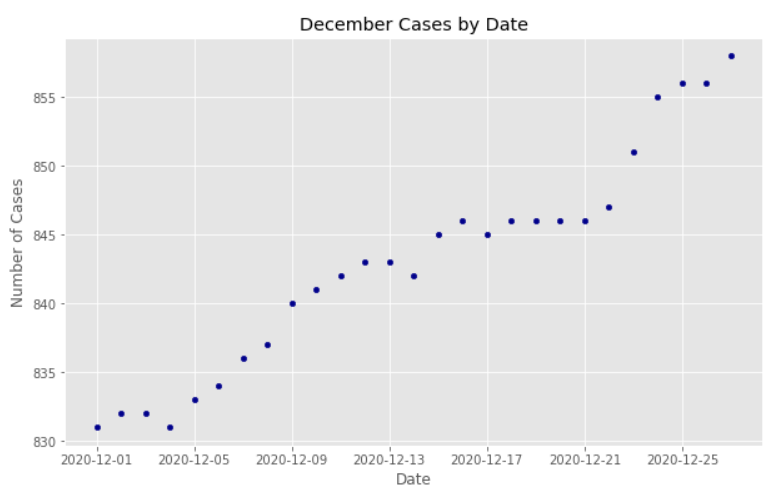


Figure 8 clearly shows that cases increased at a relatively rapid rate between December 5th and 12th, slowed between the 13th and 22nd, then again increased at a relatively rapid rate between December 22nd and 27th.

**3.1.4 QUESTION: What were the zip code population distributions between the above and below average per capita infection rate counties?**

**Figure 9: Zip Code Population By Above And Below Per Capita Infection Rate**

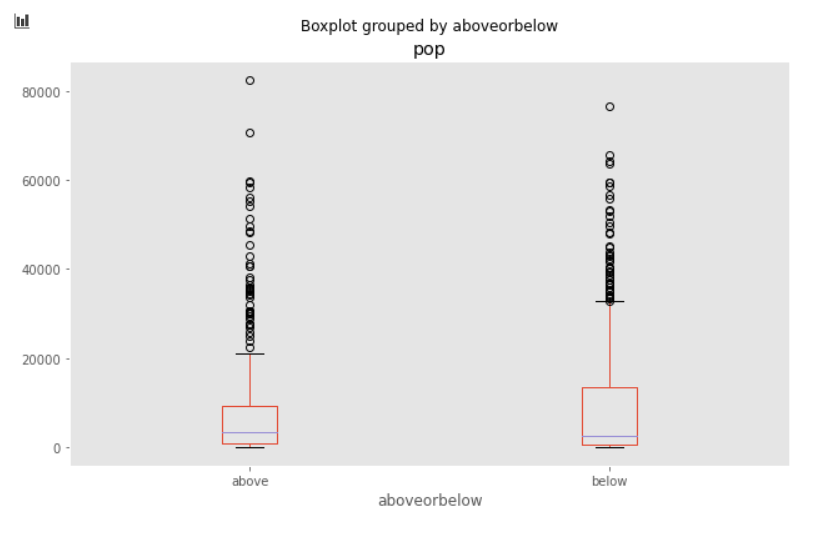
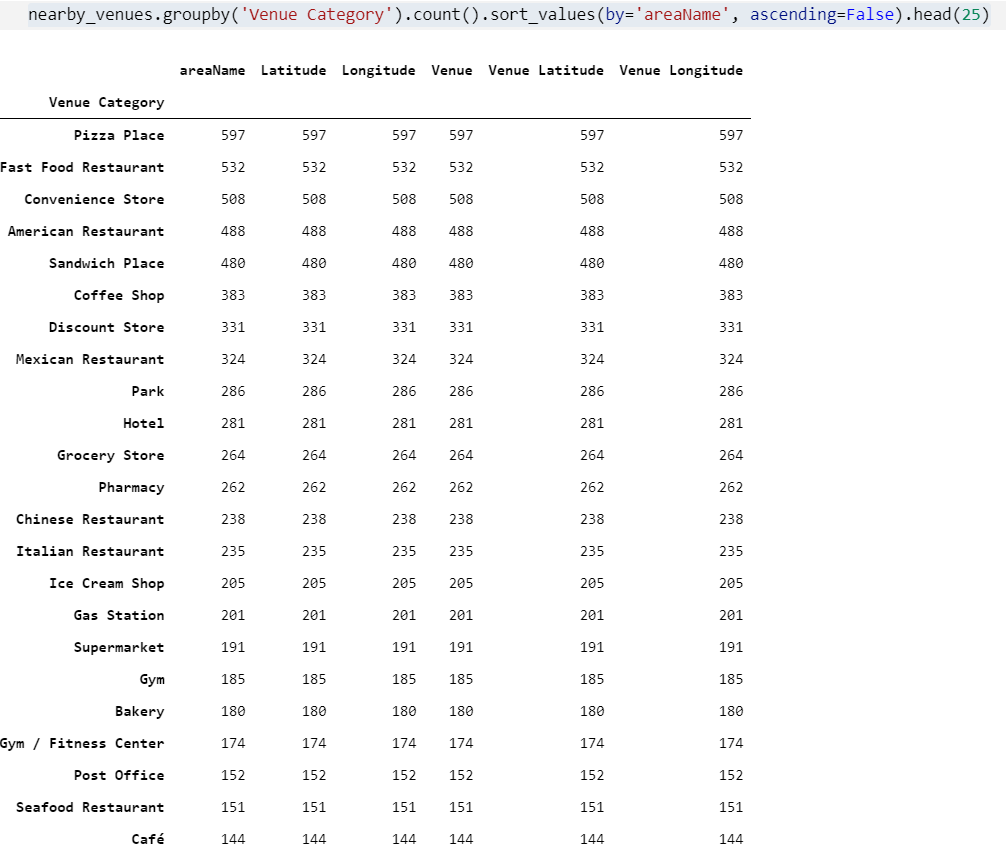


Figure 9 shows that the population distributions are similar in both the above and below per capita infection rate segments.

**3.2 Exploring Venue Categories**

Venues in **462 venue categories** were represented in the data retrieved from FourSquare. Figure 10 displays the top 25 venue categories retrieved. Pizza Places (597 Pizza Places within 2km of our zip codes), Fast Food Restaurants (532), and Convenience Stores (508) are the top three venue categories in the data set.

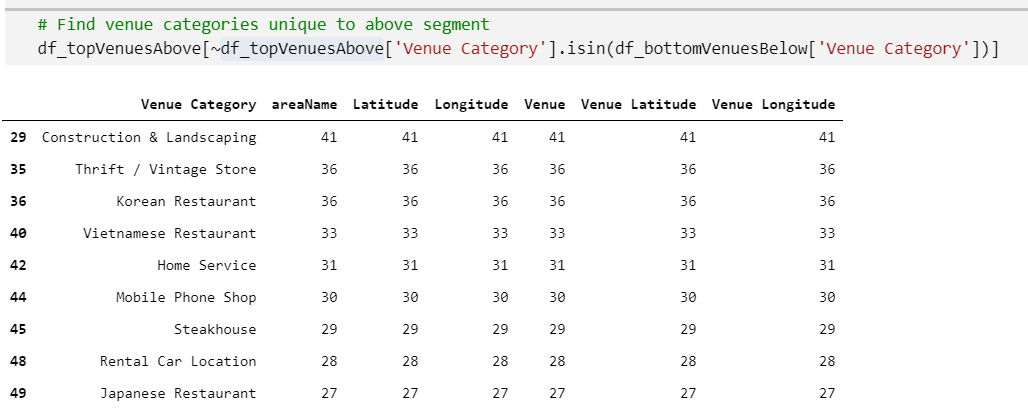
**Figure 10: Top 25 Virginia Venue Categories**



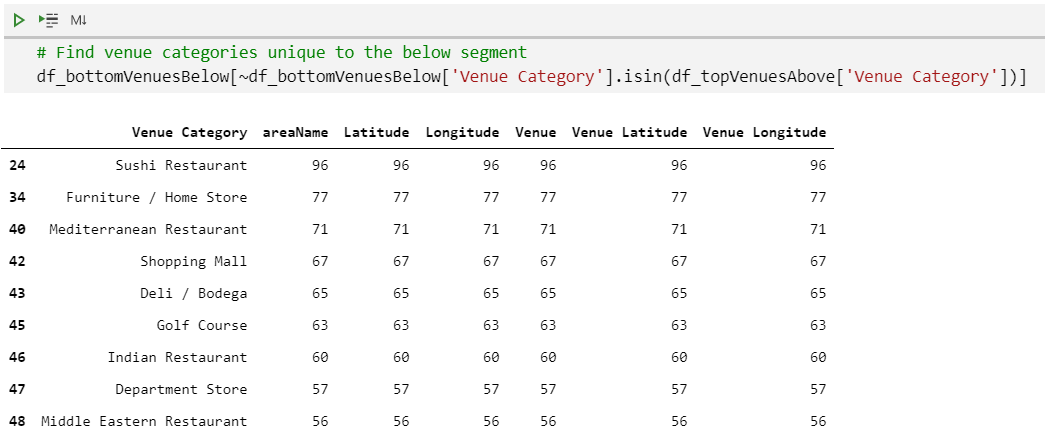
The top 50 Venue Categories represented in the above average infection rate zip codes, and below average infection rate zip codes were identified. There is significant overlap in the top venue categories represented in both the “above” and “below” segments. There are only 59 unique venue categories in the union of the top 50 of the above and below segments.

There are venue categories unique to the top 50 in the above average infection segment as seen in Figure 11. Figure 12 displays venue categories unique to the top 50 in the below average infection segment.

**Figure 11: Venue Categories unique in the Top 50 Above Average Infection Zip Codes**



**Figure 12: Venue Categories unique in the Top 50 Below Average Infection Zip Codes**

****

**Interesting Finding From Unique Infection Segment Venue Categories**

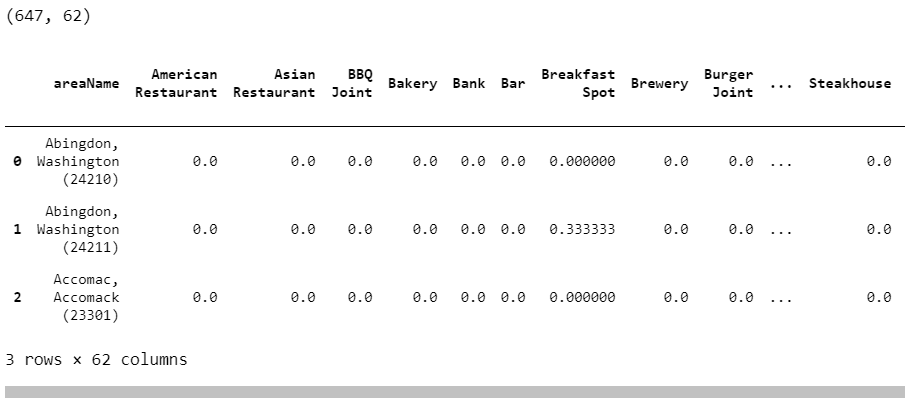
Figures 11 and 12 display the information that may be key to whether a machine learning algorithm can accurately predict whether a zip code is above or below the average December infection rate. Of interest in these two segments are that the retail locations in the below segment are higher dollar establishments, i.e. furniture stores, department stores, and shopping malls, while the above average segment’s retail locations include thrift stores, mobile phone retailers, and rental car locations. Also of interest is that golf courses, another indication of higher SES, are represented in the below average infection segment. Also interesting, ‘Sushi’ restaurants are represented in the below average segment while ‘Japanese’ restaurants are unique to the above average segment.

The question remains, are there enough differences between the above and below segments to allow a machine learning algorithm to accurately predict a zip code category.

**4. Predictive Modeling**

The data was then prepared for machine learning algorithms. The top 50 venue categories in both the above and below segments were used for training and test data. Dummy variables were created for the venue categories, and then were grouped and the mean() was used to standardize the venue categories by zip code. The resulting data included **647 zip codes, and 59 venue categories**. The DataFrame included the areaName, venue category dummy variables, population, and the above or below categorization.

**Figure 13: DataFrame Ready For Machine Learning**



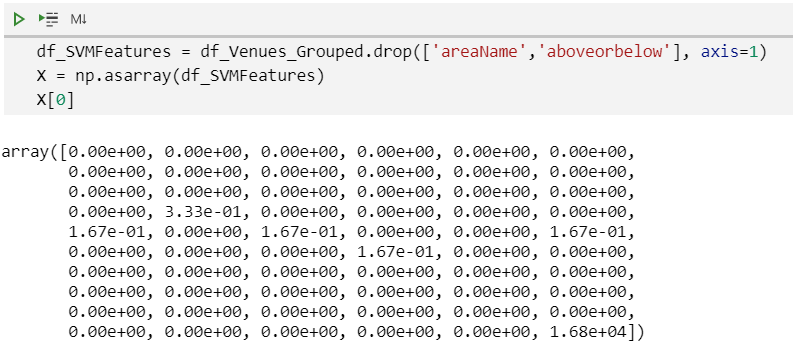
**4.1 A Support Vector Machine**

From IBM's Skills Development Network, "SVM's work by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data is transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong."

With 59 different venue categories and similar zip code population distributions there is not a linearly separable line in a hyperplane to separate the above and below infection rate zip codes. I attempted to train an SVM to distinguish between above and below zip codes by venue categories and population.

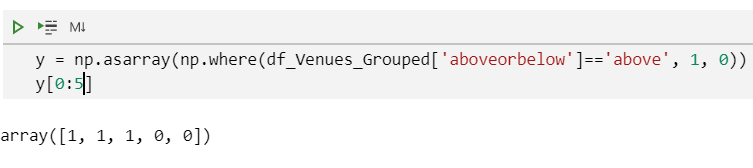
The features DataFrame was created.

**Figure 14: SVM: Features DataFrame**



The array of true values was created, where a value of 1 identified an above average per capita infection rate, and 0 identified a below average per capita infection rate.

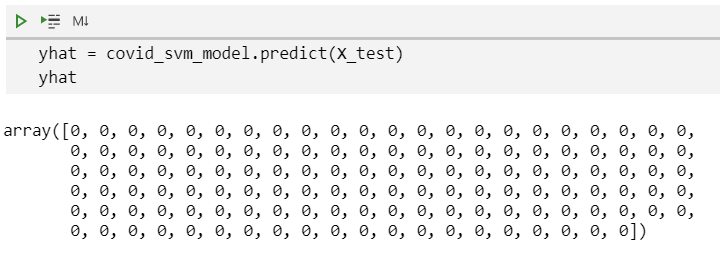
**Figure 15: SVM: True Values**



Training and test data were separated with an 80/20 split, and the SVM was trained on the train data using the default kernel function: RBF (Radial Basis Function).

The model was then tested with the test data.

**Figure 16: SVM: Testing The Model**



As shown in Figure 16, the SVM was totally unable to predict an above average infection rate, i.e. there are no 1’s in the yhat array.

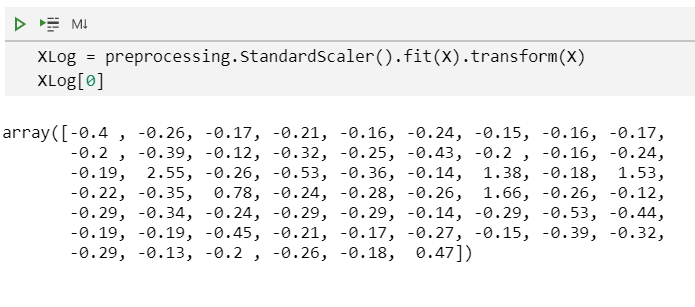
The Support Vector Machine was unable to predict above or below average per capita infection rates based on FourSquare venue categories and population.

**4.2 Logistic Regression**

Because the SVM was unsuccessful I attempted a logistic regression model using the same data. From IBM Skills Development Lab, “Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, y, is categorical.” Logistic regression uses predictor variables, in my case, venue categories and population, to produce a formula predicting the category of a response variable, in my case, above or below infection rate.

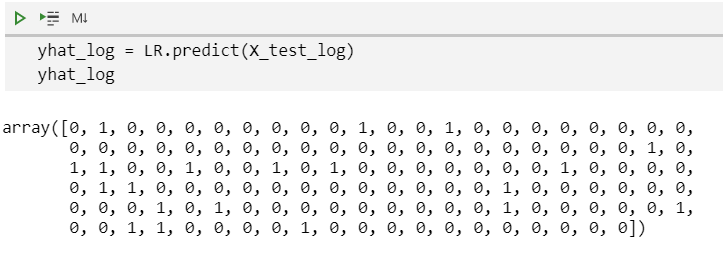
Again, I split the data into 80/20 train/test DataFrames. As seen in Figure 17, for logistic regression I used the Standard Scaler preprocessing to prepare the feature data.

**Figure 17: Logistic Regression: Scaling The Data For Logistic Regression**



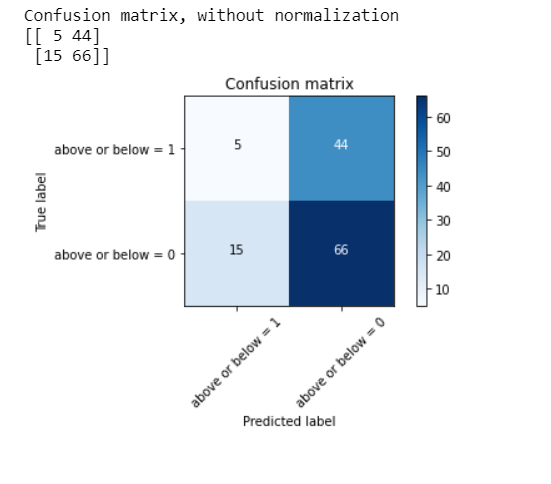
I then used the liblinear function to create and train the logistic regression model using the training data. I used the test data to evaluate. The logistic regression model was able to predict both above and below categories as seen in Figure 18.

**Figure 18: Logistic Regression: Predictions**



However, as seen in the resulting confusion matrix, the model did not perform well. The Jaccard Score was a disappointing **0.078125**.

**Figure 19: Logistic Regression: Confusion Matrix**



The logistic regression model does not perform well at predicting true above average per capita COVID infections. It performs better at predicting below average per capita COVID infections. In other words, it has a **high false negative rate** which is unacceptable.

**5. Conclusions**

While there are interesting differences between the top venues in both the above and below average per capita infection rate zip codes in Virginia, the data is not useful to predict, using either a SVM or logistic regression model.

However, it may be useful to use trending venues, i.e. venues that people are actually visiting for this analysis. Despite multiple attempts the FourSquare trending data was unavailable to me. It may be that due to social pressure to quarantine that people are not checking into the FourSquare app. It should also be noted that many venues will be beyond the 2000m radius of the FourSquare results.

This project is for example purposes only.

## **References**

*Endpoints | Places API*. (n.d.). FourSquare. Retrieved December 27, 2020, from <https://developer.foursquare.com/docs/places-api/endpoints>

Open Data Standard. (2020, December 27–30). *Open Data Standard: Latitude and Longitude of Virginia Zip Codes* [Retrieve all Virginia zip codes and their corresponding latitude and longitude]. <https://public.opendatasoft.com/api/records/1.0/search/?dataset=us-zip-code-latitude-and-longitude&q=VA&rows=1275&facet=state&facet=timezone&facet=dst&refine.state=VA>

*VDH-COVID-19-PublicUseDataset-ZIPCode | Virginia Open Data Portal*. (2020, December 29). Virginia Open Data Portal. <https://data.virginia.gov/Government/VDH-COVID-19-PublicUseDataset-ZIPCode/8bkr-zfqv>

Virginia Department of Health. (2020, December 28). *Virginia Department of Health Coronavirus FAQ*. <https://www.vdh.virginia.gov/coronavirus/frequently-asked-questions/general-questions/>

*Zip Codes in Virginia 2020*. (n.d.). World Population Review. Retrieved December 27, 2020, from <https://worldpopulationreview.com/zips/virginia>