# Capstone Project Final Report

## Introduction

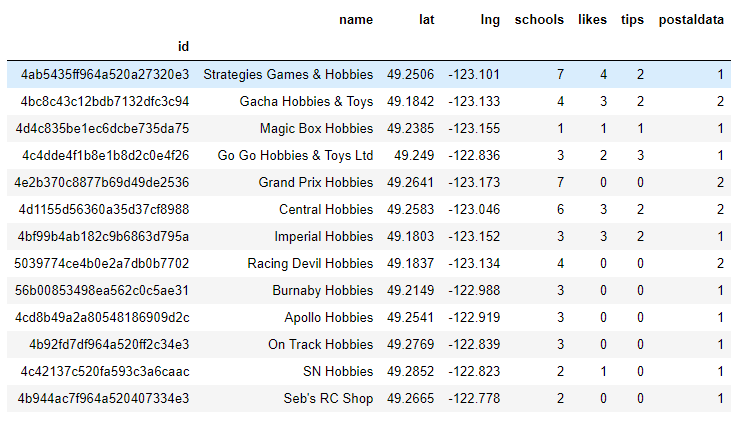
The question I've chosen to answer is: "Which is more important: the number of nearby schools or the lack of competitors to table-top and trading card game store?"

As a board game, trading card game enthusiast myself; I’ve seen stores close within a year of opening to some that stand the test of time. The people who frequent these stores are very diverse and possess diverse interests.  
  
My audience is a store owner. My analysis will target younger audiences by focusing on venue to school data. I will use Foursquare data for Vancouver BC in Canada in my analysis. My analysis will use a combination of regression and Kmeans clustering. I predict that the best location to open wont be an area with few or no similar stores, but one that is relatively near to a relevant venue (school) and not to far from the largest cluster of players.

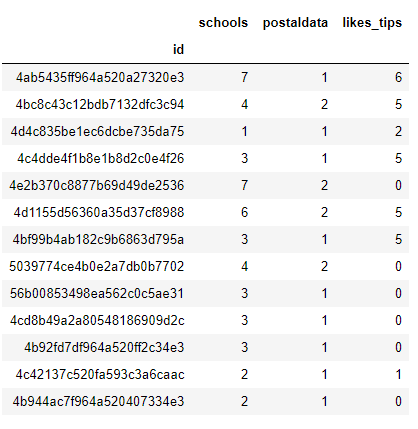
## Data

* What are the board game stores in the area?
  + Foursquare doesn't have a category for this type of store, so I'll need to search from similar stores and manually clean for the correct stores. The endpoint "listed" may be useful to find more venues.
  + where are they located?
  + how many schools are within walking distance?
  + how many similar venues are nearby?
  + how many likes?
  + how many tips?
* Foursquare Endpoints used
  + Search
  + Likes
  + Listed
  + Tips
  + Checkins – was scrapped due to authorization issues
* Changes to data
  + Added missing postal data
  + Removed duplicates with no data
  + Combined duplicates with data
  + Removed venues that were outside Vancouver. I chose the Fraser River (southern arm) as cut off point.
  + Combined like and tip data due to strong relationship

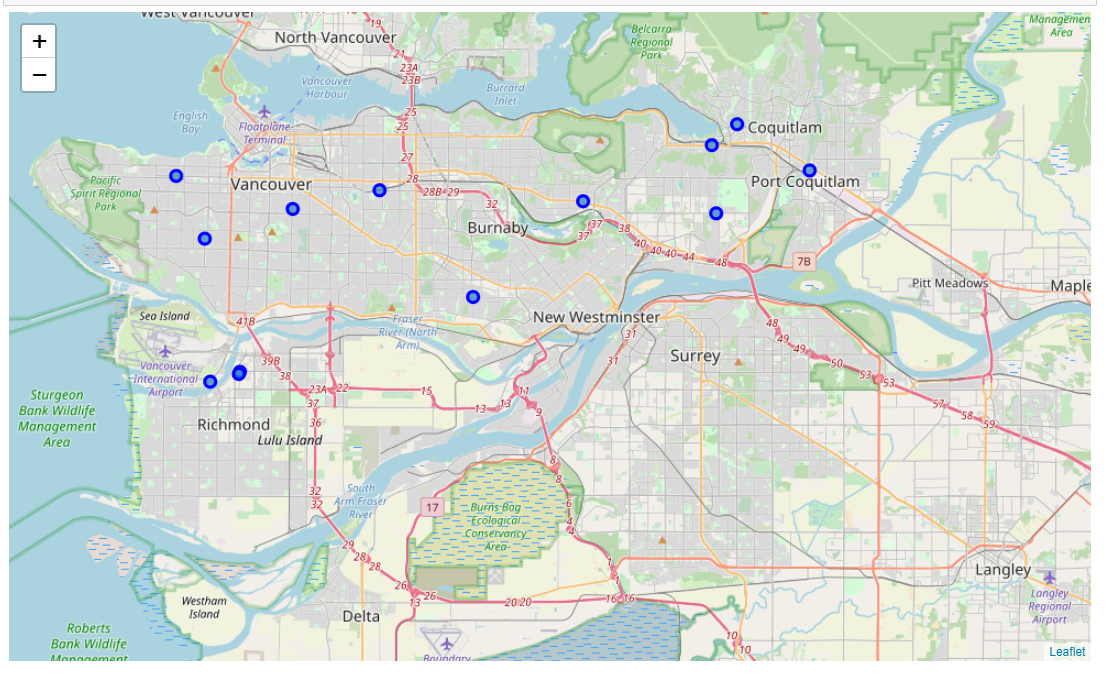
### Cleaned Data



### Test Dataframe



### Map with Venues Removed



## Methodology

### Multiple Linear Regression

Regression is an analysis method for determining the relationship between variables. Linear Regression attempts to create a straight vector through the data. The distance between the predicted vector data and each true data point is minimized. Think of the linear graphs from back in math class, the "y = b + mx". That's what linear regression is trying to do. In the case of multiple linear regression, it would be "y = b + m1x1 + m2x2" or

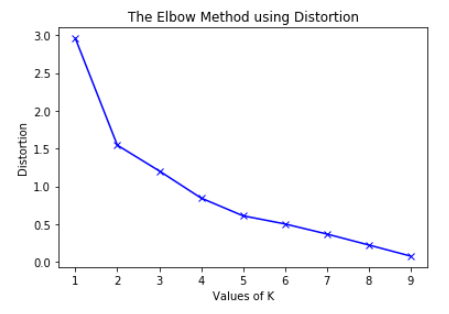
"likes/tips = intercept + m1(number of nearby schools) + m2(number of similar venues in the same postal area)."

### K-Means

K-Means Clustering is an unsupervised machine learning algorithm. The algorithm places data into a given number of clusters ("k") around central points. The average distance between the clustered data and the centroids, distortion, is minimized.

Choosing the "k" value is important. A higher "k" will always reduce distortion, but the effect is weaker at values. A popular method to choose "k" is called the "elbow method." Numerous iterations of the algorithm are ran and the amount of distortion is recorded and graphed. The "k" value that reduces the greatest amount of distortion before diminishing returns.

### Elbow Method

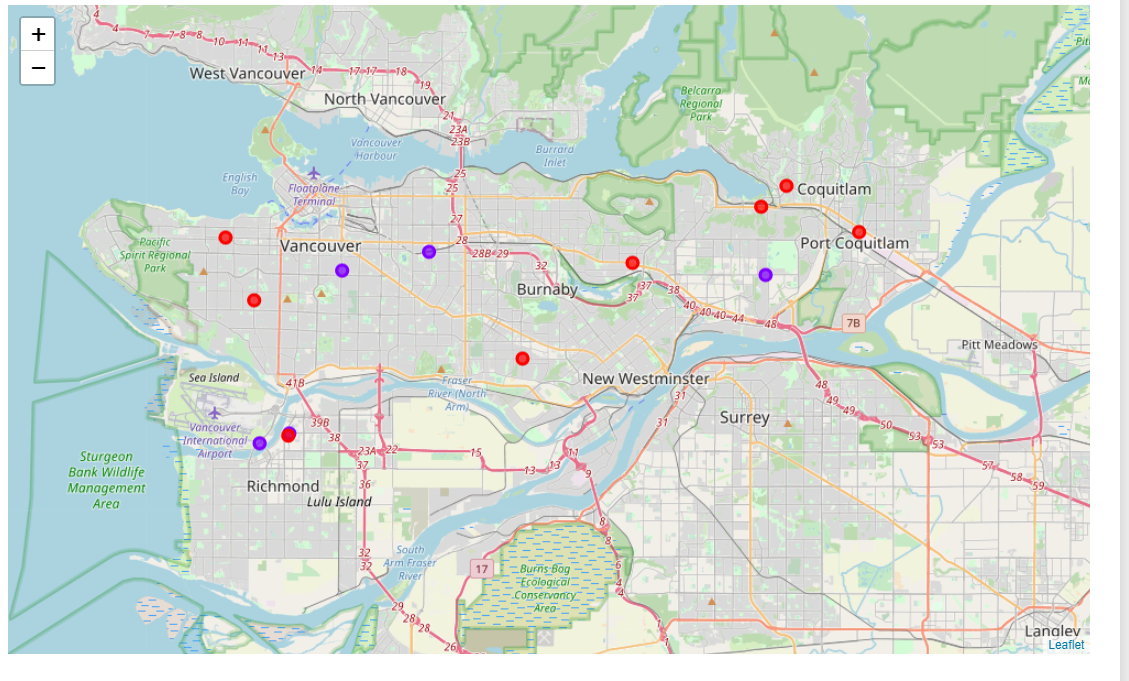


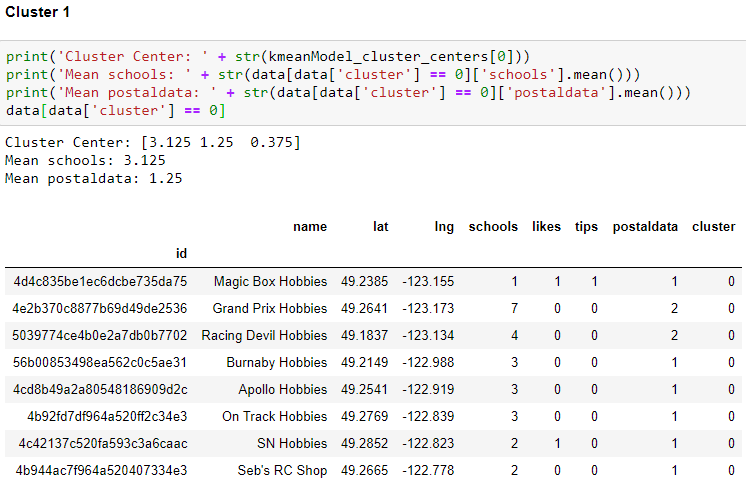
## Results

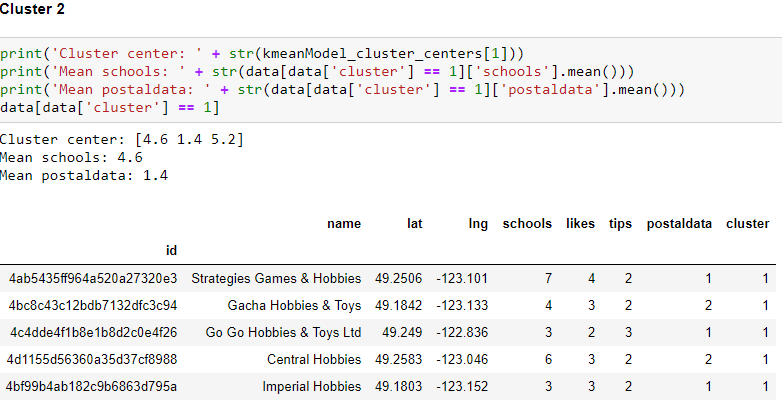
### Multiple Linear Regression



### K-Means







## Discussion

The data is pretty limited. A difference of one like or tip would swing the concludes for a location in a completely different direction. Other important data points, such as what products are sold, are not represented in the data.

With the given data:

* Regression analysis shows that the number of nearby schools has a weaker impact on likes and tips than the number of other hobby stores within the same postal code.
* The number of nearby schools has a positive impact. For about every two schools, one can expect a tip or like.
* The number of venues within the same postal code has a negative impact. For each similar venue, one can expect one less tip or like (when otherwise expected).
* The Kmeans clusters analysis has segmented the venues into high and low like/tips clusters. Given the regression analysis, this is expected as likes to tips has a strong impact. Still, the high like/tips cluster averages one more school. Again, as expected. The confounding piece of information is the nearby similar venues data. The high like/tips cluster averages slightly more nearby venues despite having a negative impact. This is likely due to the fact that there are so few venues compared to schools that the negative effect is overwhelmed.
* There is a pretty glaring outlier in Cluster 1. Grand Prix Hobbies, despite being near to 7 schools, has no likes or tips. 7 nearby schools is the highest value in the dataset. Strategies Games & Hobbies also has 7 nearby schools, but has the most likes and tips.

## Conclusion

Fundamentally, there is not enough data to confidently support any conclusion on the impact of nearby schools or the number of nearby similar venues. However, the data does follow what one would expect: that having nearby schools is a positive effect while nearby competitors has a negative effect.