Child Welfare Referrals and Neighborhood Features

Introduction

Predicting the abuse of a child before it occurs is the Holy Grail of social work. Being a budding data scientist with a background in social work made this promise an irresistible quest, but child welfare data and the predictability of a referral to the child welfare system can be, like the Holy Grail, very elusive. It was with this understanding that I attempted to make, at times, the illogical into logical. My social worker side told me this was a fool’s errand, but the data scientist in me was determined to prove otherwise. It was with this foundation that I formed the hypothesis that the predominant features of a neighborhood could be used to determine the likelihood of the occurrence of child abuse referrals. This hypothesis was shaped by the project requirement that Foursquare data was to be used in some fashion.

Methodology

I began with a year’s worth of child welfare data for a single county in California. The basic data set was the number of referrals, or reports, of child abuse by ZIP code. The idea being that I could take this occurrence and compare the density of referrals, based on overall ZIP population, with the top five features of the ZIP code as extracted from Foursquare. The idea was that certain features would correlate with more dense referral counts. These features could then be used to predict the relative probability of referrals to the child welfare system in other counties of California, or other states. If a consistent set of features was found to significantly correlate, a model could be built and used for staffing, funding and service planning functions.

<Sample of Welfare Data>

|  | **ZIP** | **Referrals** | **City** | **Population** | **Latitude** | **Longitude** | **Density** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 95608 | 131.0 | Carmichael | 59418.0 | 38.621360 | -121.332191 | 0.002205 |
| **1** | 95610 | 212.0 | Citrus Heights | 44147.0 | 38.689802 | -121.267360 | 0.004802 |
| **2** | 95621 | 163.0 | Citrus Heights | 39819.0 | 38.693052 | -121.309461 | 0.004094 |
| **3** | 95624 | 128.0 | Elk Grove | 61989.0 | 38.441480 | -121.307142 | 0.002065 |
| **4** | 95626 | 3.0 | Elverta | 5975.0 | 38.713790 | -121.462730 | 0.000502 |

Density was determined by dividing the number of referrals by the population of the ZIP. The larger this number was, the more referrals per capita the ZIP code contained.

The latitude and longitude for each ZIP was obtained via [www.california-demographics.com](http://www.california-demographics.com). This data set was merged with the child welfare data to produce the basis of the study.

Once the foundational data set was built, the latitude and longitudes were passed through to Foursquare using a 500m radius from the location of the center of the ZIP code, as represented by the latitude and longitude. A maximum of 100 venues was returned and counted to develop the top 5 list.

The data was retuned in a list similar to this sample from the city of Carmichael. Some ZIPs had hundreds of venues listed, such as the Sacramento, and some had one or two.

|  | **City** | **City Latitude** | **City Longitude** | **Venue** | **Venue Latitude** | **Venue Longitude** | **Venue Category** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | Carmichael | 38.62136 | -121.332191 | Roma's Pizza | 38.619713 | -121.328264 | Pizza Place |
| **1** | Carmichael | 38.62136 | -121.332191 | Papa Murphy's | 38.619083 | -121.328062 | Pizza Place |
| **2** | Carmichael | 38.62136 | -121.332191 | 99 Cents Only Stores | 38.618010 | -121.329878 | Discount Store |
| **3** | Carmichael | 38.62136 | -121.332191 | Les Schwab Tire Center | 38.620269 | -121.328096 | Automotive Shop |
| **4** | Carmichael | 38.62136 | -121.332191 | Yianni's | 38.621531 | -121.328074 | Greek Restaurant |

The number of venues per category was then summed and the top five venues per ZIP were put into a data set.

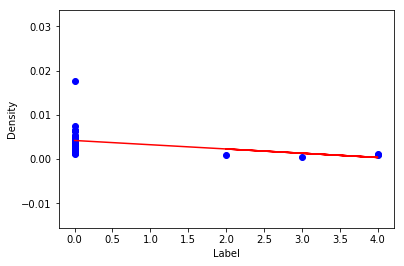
|  | **City** | **1st\_Most\_Common\_Venue** | **2nd\_Most\_Common\_Venue** | **3rd\_Most\_Common\_Venue** | **4th\_Most\_Common\_Venue** | **5th\_Most\_Common\_Venue** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** | Antelope | Pizza Place | Mexican Restaurant | Wings Joint | Supermarket | Ice Cream Shop |
| **1** | Carmichael | Pizza Place | Antique Shop | Martial Arts Dojo | Greek Restaurant | Automotive Shop |
| **2** | Citrus Heights | Cosmetics Shop | Bank | Convenience Store | Park | Pizza Place |
| **3** | Elk Grove | Fast Food Restaurant | Pizza Place | Cosmetics Shop | Rental Car Location | Italian Restaurant |
| **4** | Elverta | Pet Store | Home Service | Women's Store | Donut Shop | Fishing Store |
| **5** | Fair Oaks | Liquor Store | Home Service | Convenience Store | Construction & Landscaping | Women's Store |

The data was then clustered by venue to see if there was a commonality between the type of cluster and the density of referrals.

|  | **Cluster** | **City** | **1st\_Most\_Common\_Venue** | **2nd\_Most\_Common\_Venue** | **3rd\_Most\_Common\_Venue** | **4th\_Most\_Common\_Venue** | **5th\_Most\_Common\_Venue** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | Antelope | Pizza Place | Mexican Restaurant | Wings Joint | Supermarket | Ice Cream Shop |
| **1** | 0 | Carmichael | Pizza Place | Antique Shop | Martial Arts Dojo | Greek Restaurant | Automotive Shop |
| **2** | 0 | Citrus Heights | Cosmetics Shop | Bank | Convenience Store | Park | Pizza Place |
| **3** | 0 | Elk Grove | Fast Food Restaurant | Pizza Place | Cosmetics Shop | Rental Car Location | Italian Restaurant |
| **4** | 3 | Elverta | Pet Store | Home Service | Women's Store | Donut Shop | Fishing Store |
| **5** | 0 | Fair Oaks | Liquor Store | Home Service | Convenience Store | Construction & Landscaping | Women's Store |
| **6** | 4 | Folsom | Gym | Trail | Women's Store | Donut Shop | Flea Market |
| **7** | 0 | Galt | Fast Food Restaurant | Motel | Intersection | Steakhouse | Print Shop |
| **8** | 2 | Herald | Construction & Landscaping | Women's Store | Donut Shop | Flea Market | Fishing Store |
| **9** | 0 | Isleton | Mexican Restaurant | Convenience Store | Fishing Store | Steakhouse | Electronics Store |
| **10** | 0 | Mcclellan | Convenience Store | Grocery Store | Fast Food Restaurant | Café | Sandwich Place |
| **11** | 0 | North Highlands | Mexican Restaurant | Cuban Restaurant | Asian Restaurant | Sandwich Place | Resort |
| **12** | 0 | Orangevale | Spa | Plaza | Pool | Electronics Store | Dive Bar |
| **13** | 0 | Rancho Cordova | Mexican Restaurant | Automotive Shop | Fish & Chips Shop | Dance Studio | Historic Site |
| **14** | 0 | Rio Linda | Pizza Place | Grocery Store | Convenience Store | Coffee Shop | Fast Food Restaurant |

Results

The clusters were then plotted by density to see if there was a linear relationship to indicate a significant correlation.



As can be seen, there is a flat relationship between the cluster, or predominant venues, and the occurrence of child welfare referrals.

Since the line is pretty flat for the relationship between Venue and Density, I decided to look at the weighted values for the venues, giving a score of 5 to 1st most common descending down to a score of 1 to 5th most common venue. I then summed the weighted ranks and grouped by venue to give each venue a weight based on it’s common order (1-5) and it's density for that location or ZIP code. For example, each time “Coffee Shop” was the 1st most common venue, the density for that ZIP was multiplied by 5. Likewise, if “Coffee Shop” was the 2nd most common value, the density was multiplied by 4. These products were summed up by venue to provide a weighted score for the sample county by venue.

| **Venue** | **Density** | **wt** |
| --- | --- | --- |
| **Coffee Shop** | 0.109395 | 0.534942 |
| **Park** | 0.114279 | 0.439326 |
| **Hotel** | 0.105384 | 0.316151 |
| **Mexican Restaurant** | 0.121408 | 0.289027 |
| **Convenience Store** | 0.033477 | 0.137124 |
| **American Restaurant** | 0.105384 | 0.105384 |
| **Fast Food Restaurant** | 0.029269 | 0.095118 |

The idea here is that a particular ZIP code could be analyzed, using Foursquare data, to determine the top 5 venues for that region, then these weighted scores for the venue could be applied and give a relative score. If 20 ZIP codes were to be analyzed, sorted by this weighted score, the corresponding referral density scores should sort in a similar manner. If this is the case, then further analysis would be done to statistically prove the relationship.

The top 20 counties in California by population were selected and child welfare data extracted to produce the following data set for analysis.

|  | **ZIP** | **Population** | **Latitude** | **Longitude** | **City** | **State** | **County** | **Referrals** | **Density** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 90011 | 108051.0 | 34.01 | -118.26 | Los Angeles | CA | Los Angeles County | 1522 | 0.014086 |
| **1** | 90650 | 106404.0 | 33.91 | -118.08 | Norwalk | CA | Los Angeles County | 749 | 0.007039 |
| **2** | 91331 | 105696.0 | 34.26 | -118.42 | Pacoima | CA | Los Angeles County | 1238 | 0.011713 |
| **3** | 90201 | 102878.0 | 33.98 | -118.17 | Bell | CA | Los Angeles County | 947 | 0.009205 |
| **4** | 92335 | 99791.0 | 34.09 | -117.47 | Fontana | CA | San Bernardino County | 1254 | 0.012566 |
| **5** | 90250 | 97371.0 | 33.91 | -118.35 | Hawthorne | CA | Los Angeles County | 779 | 0.008000 |
| **6** | 91342 | 96177.0 | 34.33 | -118.38 | Sylmar | CA | Los Angeles County | 841 | 0.008744 |
| **7** | 90805 | 96069.0 | 33.86 | -118.18 | Long Beach | CA | Los Angeles County | 1094 | 0.011388 |
| **8** | 90280 | 95420.0 | 33.94 | -118.19 | South Gate | CA | Los Angeles County | 808 | 0.008468 |
| **9** | 90044 | 94680.0 | 33.94 | -118.29 | Los Angeles | CA | Los Angeles County | 1742 | 0.018399 |
| **10** | 92503 | 94523.0 | 33.88 | -117.44 | Riverside | CA | Riverside County | 1150 | 0.012166 |
| **11** | 92336 | 94327.0 | 34.15 | -117.46 | Fontana | CA | San Bernardino County | 823 | 0.008725 |
| **12** | 94565 | 93549.0 | 38.00 | -121.92 | Pittsburg | CA | Contra Costa County | 1021 | 0.010914 |
| **13** | 92683 | 91758.0 | 33.75 | -117.99 | Westminster | CA | Orange County | 651 | 0.007095 |
| **14** | 92704 | 90525.0 | 33.72 | -117.91 | Santa Ana | CA | Orange County | 1251 | 0.013819 |
| **15** | 91710 | 88862.0 | 34.00 | -117.68 | Chino | CA | San Bernardino County | 700 | 0.007877 |
| **16** | 92804 | 88065.0 | 33.82 | -117.97 | Anaheim | CA | Orange County | 1267 | 0.014387 |
| **17** | 95076 | 87781.0 | 36.98 | -121.75 | Watsonville | CA | Santa Cruz County | 1190 | 0.013556 |
| **18** | 92154 | 87218.0 | 32.56 | -117.01 | San Diego | CA | San Diego County | 1255 | 0.014389 |
| **19** | 91744 | 86982.0 | 34.03 | -117.94 | La Puente | CA | Los Angeles County | 670 | 0.007703 |

Further Validation

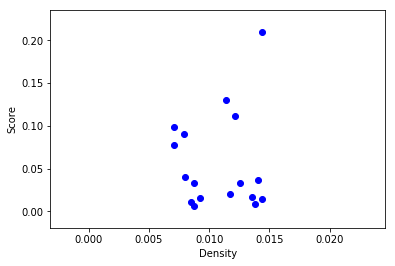
The previous analytical process was applied to the top 20 most populous ZIP codes in California, presuming to remove irregularities that could come from smaller data sets in less populous counties. Taking the weighted venue scores and the top 5 venues per ZIP, the weighted score was applied and summed per ZIP to give a single score. It was hypothesized that this score should represent the relative likelihood of a referral in that ZIP and therefore the known density of the ZIP should align with this score when the tables were sorted.

To calculate the “Score”, each ZIP had it’s top 5 venues determined via Foursquare, just like the sample set. These venues were then translated into their weighted scores, using the table created in the first part of the research. Since the reference table was already weighted based on order (1st to 5th most common), the values were summed to provide the scores shown below.

|  | **ZIP** | **Population** | **Latitude** | **Longitude** | **City** | **State** | **County** | **Referrals** | **Density** | **Score** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **14** | 92804 | 88065.0 | 33.82 | -117.97 | Anaheim | CA | Orange County | 1267 | 0.014387 | 0.209652 |
| **8** | 90805 | 96069.0 | 33.86 | -118.18 | Long Beach | CA | Los Angeles County | 1094 | 0.011388 | 0.129729 |
| **10** | 92503 | 94523.0 | 33.88 | -117.44 | Riverside | CA | Riverside County | 1150 | 0.012166 | 0.111119 |
| **11** | 92683 | 91758.0 | 33.75 | -117.99 | Westminster | CA | Orange County | 651 | 0.007095 | 0.098595 |
| **13** | 91710 | 88862.0 | 34.00 | -117.68 | Chino | CA | San Bernardino County | 700 | 0.007877 | 0.089915 |
| **1** | 90650 | 106404.0 | 33.91 | -118.08 | Norwalk | CA | Los Angeles County | 749 | 0.007039 | 0.078060 |
| **6** | 90250 | 97371.0 | 33.91 | -118.35 | Hawthorne | CA | Los Angeles County | 779 | 0.008000 | 0.039928 |
| **0** | 90011 | 108051.0 | 34.01 | -118.26 | Los Angeles | CA | Los Angeles County | 1522 | 0.014086 | 0.036635 |
| **4** | 92335 | 99791.0 | 34.09 | -117.47 | Fontana | CA | San Bernardino County | 1254 | 0.012566 | 0.033298 |
| **5** | 92336 | 94327.0 | 34.15 | -117.46 | Fontana | CA | San Bernardino County | 823 | 0.008725 | 0.033298 |
| **2** | 91331 | 105696.0 | 34.26 | -118.42 | Pacoima | CA | Los Angeles County | 1238 | 0.011713 | 0.020186 |
| **15** | 95076 | 87781.0 | 36.98 | -121.75 | Watsonville | CA | Santa Cruz County | 1190 | 0.013556 | 0.016486 |
| **3** | 90201 | 102878.0 | 33.98 | -118.17 | Bell | CA | Los Angeles County | 947 | 0.009205 | 0.015206 |
| **16** | 92154 | 87218.0 | 32.56 | -117.01 | San Diego | CA | San Diego County | 1255 | 0.014389 | 0.014441 |
| **9** | 90280 | 95420.0 | 33.94 | -118.19 | South Gate | CA | Los Angeles County | 808 | 0.008468 | 0.011288 |
| **12** | 92704 | 90525.0 | 33.72 | -117.91 | Santa Ana | CA | Orange County | 1251 | 0.013819 | 0.009031 |
| **7** | 91342 | 96177.0 | 34.33 | -118.38 | Sylmar | CA | Los Angeles County | 841 | 0.008744 | 0.006103 |

As can be readily seen in the sorted table, the “Score” column and the “Density” column do not align, with some higher density ZIPs occurring nearer the bottom of the table sorted by Score.

As can also be seen in a scatter plot of the same data, there is no clear linear correlation between the Density and the Score by ZIP code.



To further reinforce the lack of a relationship, the Pearson score was calculated to be 0.064, indicating a statistical lack of correlation.

Conclusions

The obvious conclusion is that there is no relationship between the physical venues in a given area and the occurrence of child abuse referrals. In other words, just because there are a lot of a certain type of restaurant or store or park in a given area, it doesn’t correlate to a higher or lower instance of child welfare involvement. While this doesn’t not help with the issues of staffing and resource allocation, it is helpful to address some common misconceptions about welfare and abuse. By showing that there are no features or venues, often stereotypical, that denote a “bad” neighborhood, preconceived notions about child welfare can be broken down, possibly leading to a more individual based approach to services and future care. It is evident from the hypothesis that there are some ideas that a “bad” neighborhood probably contains some common feature, or venues, and that these features contribute in some way to the likelihood of abuse. This study serves to disprove these notions. Like any human behavior, child abuse does not fall into a neatly defined set of characteristics and should be approached on a case by case basis.

This project did have many possible issues, possibly stemming from the method and the requirements as stated. One possible issue is the unit of area used, ZIP Code. A ZIP code is an arbitrary geographical unit and does not represent a homogenous population. Something that might be more accurate, but much harder to define would be a neighborhood. A neighborhood would be defined by its population and common attributes, rather than its inclusion in an given geographical boundary. By clustering the venues and analyzing the clusters, it was hoped that this deficiency might be mitigated, but that did not prove to be the case.

Another issue was the sparseness of some of the venues and the resulting weights. Since the availability of venues was wholly dependent on user entered data in Foursquare, some areas, or venue types, could be vastly under represented. This could have led to an arbitrary preference for one venue type over another, or a coincidently high density for a venue because no other venues were entered in for that ZIP.

The third observation is that one county probably does not account for a reliable model. A better approach may have been to look at a larger data set, such as statewide, and then distill for a single county for testing the model.

This effort was an interesting exercise in gathering data from disparate sources and compiling it into a single view of a given area and subject. While the hypothesis was ultimately proven to be incorrect, the opportunity to address ideas about environment and possibly culturally biased views of neighborhoods was invaluable. In breaking down the notion that neighborhoods that contain certain features, are more likely to have child welfare involvement, we start to see the clients as individuals, instead of part of a demographic group, and we can better address the needs of each one on his or her unique needs, rather than addressing a population as we thought we understood it.

For all reference materials, please see - [Final Jupyter Notebook](https://github.com/schnagga/Coursera_Capstone/blob/master/Social%20Services%20and%20Neighborhood%20Characteristics%20(FINAL).ipynb)