**Capstone project - *Data Daddy***

***Note to reader: this is all hypothetical***

**Introduction**

Data Daddy, established by Sir Daniele Capraro, is a company that provides all data analytics outsourcing needs. Established in 2009, Data Daddy was originally focused on the financial institutions and successfully worked with and developed bespoke models to assist these companies during and post financial crisis. Throughout the 2010s, Data Daddy was established as a leader for in Financial industry. With this reputation, towards the end of 2019, Data Daddy attempted to branch out into industries, however, this evolution was halted due to COVID19 (#TotallyRelevant).

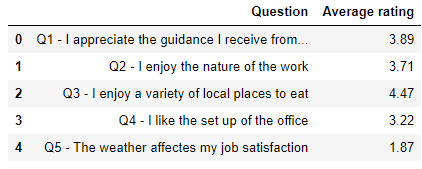
With COVID19 eradicated 6 months ago, we are seeing people re-enter the workforce, unemployment rates dropping, and the general public’s “willingness” to spend is rapidly increasing back to normal. For Data Daddy, this influx translates spike in demand, and the need to open a new San Francisco office.

This report will determine the optimal neighbourhood in San Francisco to open a new branch.

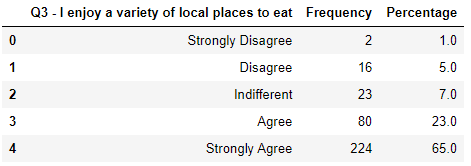
**Data**

In a prior project, in anticipation for opening a new San Francisco office, all 345 Data Daddy employees were asked to fill out a questionnaire, rating on a 5 point Likert scale, which aspect of their job is tied to their job satisfaction. A quick analysis showed the overwhelming majority of employees associated their job satisfaction with local places to eat. See Figure 1 below.

**Figure 1 - Average result of job satisfaction survey**



**Figure 2 - Frequency count of "Q3 - I enjoy a variety of local places to eat" rating**



With 88% of Data Daddy employees agree/strongly agreeing to local places to eat job satisfaction, the free text section of this questionnaire uncovered Data Daddy employees all cited pizza in this section, where:

* Employees enjoy the eating pizza for lunch,
* Employees enjoy ordering pizza for social work events, and
* Employees enjoy their clients taking them out to pizzerias.

With the number of places selling pizza as the primary driver in Data Daddy employees job satisfaction, we will leverage off Foursquare data to determine the optimal neighbourhood to open a new Data Daddy branch.

***Please note: San Francisco neighbourhood are: Central, Southern, Bayview, Mission, Park, Richmond, Ingleside, Taraval, Northern, and, Tenderloin.***

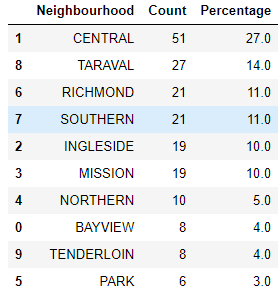
We would assume the Central neighbourhood to have the most pizzerias.

Data we will be extracting from Foursquare is the number of places that sell pizza in San Francisco. We will use this to display the pizza restaurant density by neighbourhood through a choropleth graph.

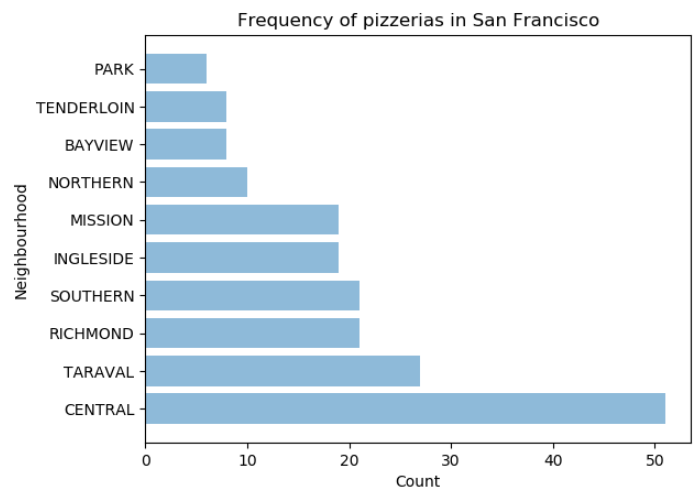
One limitation for Foursquare is that I am not able to extract all the restaurants that sell pizza in one go. To get around this, I will extract the 50 closest restaurants that sell pizza for each neighbourhood, and remove any duplicates for restaurants that have been returned multiple times

Table 3 and Figure 1 (below) displays the frequency of pizzerias by neighbourhood.

#### Table 3 - Frequency of pizzerias by neighbourhood



#### Figure 1 - Bar chart of frequency of pizzerias in San Francisco be neighbourhood



We are able to see that the Central location has appears to have more pizzerias.

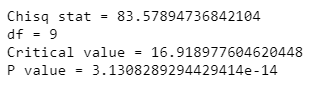
**Methodology – Chi Squared Goodness of Fit test**

We will use a Chi Squared Goodness of Fit test to determine which Neighbourhood has statistically the most pizzerias. This test is used to compare the observed frequency against the expected frequency for a categorical variable. To determine whether or not some neighbourhoods have more pizzerias, we would test against each neighbourhood having the same number of pizzerias.

Null hypothesis: observed = expected

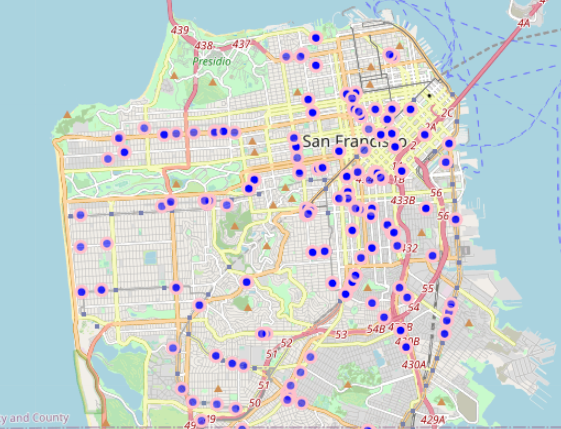
Alternate hypothesis: observed ≠ expected

**Results**



The above p-value is less than 0.05, so we reject the null hypothesis. This suggests that not all Neighbourhoods have the same number of pizzerias.

#### Figure 2 – Map of pizzerias in San Francisco



After inspecting figure 1, and the above map, we are able to conclude that the Central Neighbourhood statistically has the most pizzerias.

**Discussion**

This purpose of this report is to determine the optimal neighbourhood in San Francisco to open a new branch. With data showing the key driving factor for job satisfaction among Data Daddy employees to be proximity to pizzerias, we have analysed data to determine which Neighbourhood in San Francisco has the most pizzerias.

At the beginning of the report we assumed that the Neighbourhood to have the most pizzerias is "Central". Using the Chi Squared Goodness of Fit test, we are able to conclude that this assumption is correct.

**Conclusion**

In this report I analysed geospatial data to determine which neighbourhood contains the most pizzerias. With the number of pizzerias being the prime factor in determining where to open a Data Daddy San Francisco branch, I would recommend opening the San Francisco branch in the Central neighbourhood.

This report looks solely at the frequency of pizzerias and narrows down the neighbourhoods in San Francisco. There are many additional data we could leverage off to inform the executives at Data Daddy on possibly another neighbourhood, or a district within the central neighbourhood for a new Data Daddy branch. Future research could:

* Sample a more comprehensive job satisfaction survey prior to this report
* leverage off real time commercial rental properties in the San Francisco area
* Compare the cost of setting up a branch in a neighbourhood surrounding the central neighbourhood