IBM DATA SCIENCE CAPSTONE PROJECT

SCOPING THE MOST DESIRABLE AREAS TO OPEN A RESTAURANT IN CALGARY, AB

IAN AUSTIN

SEPTEMBER 2021

LOCATION, LOCATION, LOCATION…..

- Opening a restaurant, or any business, one of the bigger decisions a person can make
- CNBC Article: 60% of restaurants fail in year 1, 80% by year 5
- Number one reason cited for failure: location
  - Getting it right really is ‘do or die’

LOCATION – CALGARY, AB

- COVID-19 pandemic has changed numerous things
  - ~30% of pervious vehicle/person traffic in City Center
  - Trending towards more work from home/hybrid arrangements
- City design philosophy pushing towards more walkable communities
  - Constituents asking for it
  - Intersection with post-COVID reality of work

RESTAURANT – WHERE TO LOCATE ONE

- Analysis: looking for neighborhoods/areas with
  - Higher Median Income: more disposable income leads to more being spent on dinners etc
  - Walkability: people pushing for walkable communities with closer amenities
  - Restaurant Density: try to fins a sweet spot
    - Too few – maybe not a desirable area (more analysis required)
    - Too many – too much potential competition for disposable $
- Analysis will look for areas with a good combination of each variable

DATA SOURCES

- Mostly web scraped
- Open-source data
  - Not always the best source (quality or layout)
- Collected data for analysis
  - Postal Codes / Neighborhoods in Calgary
  - Population by Postal Code
  - Median Income by Neighborhood / Postal Code
  - Walk Scores by Neighborhood / Postal Code
  - Restaurant Data (Foursquare API – from center of postal codes)
- Some data required grouping/averaging (income, walk score data) to specify by postal code

DATA ANALYSIS

- Data web scraped, cleaned/preprocessed, and combined in sequence to a single analysis data frame
- Exploratory Analysis
  - Look for relationships between the 3 parameters
    - Income
    - Walkability
    - Restaurant Density
  - Plotted individually and correlation matrix
    - Moderate negative correlation (Income with both Walkability and Restaurant Density)
    - Moderate positive correlation (Walkability and Restaurant Density)

[47]:

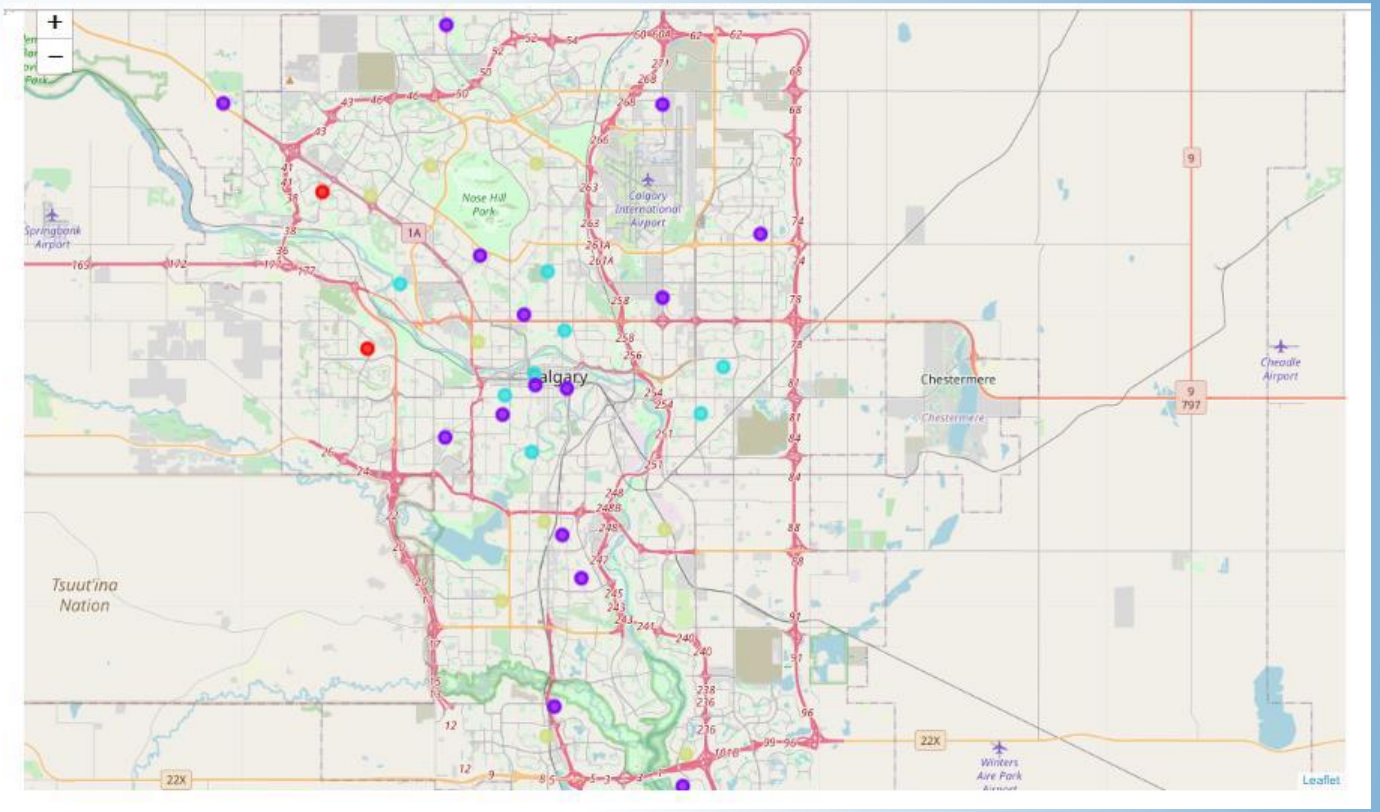
	Median Household Income After Taxes	Walk Score	Restaurants/1000 residents
Median Household Income After Taxes	1.000000	-0.454394	-0.366446
Walk Score	-0.454394	1.000000	0.556149
Restaurants/1000 residents	-0.366446	0.556149	1.000000

DATA ANALYSIS

- Clustering analysis (K-Means) selected
  - No targets
  - Scoping out best area characteristics to open a new restaurant
- 3 steps
  - First Clustering/Grouping of clusters
  - Second Clustering/Grouping of clusters
  - Final Analysis

FIRST ANALYSIS

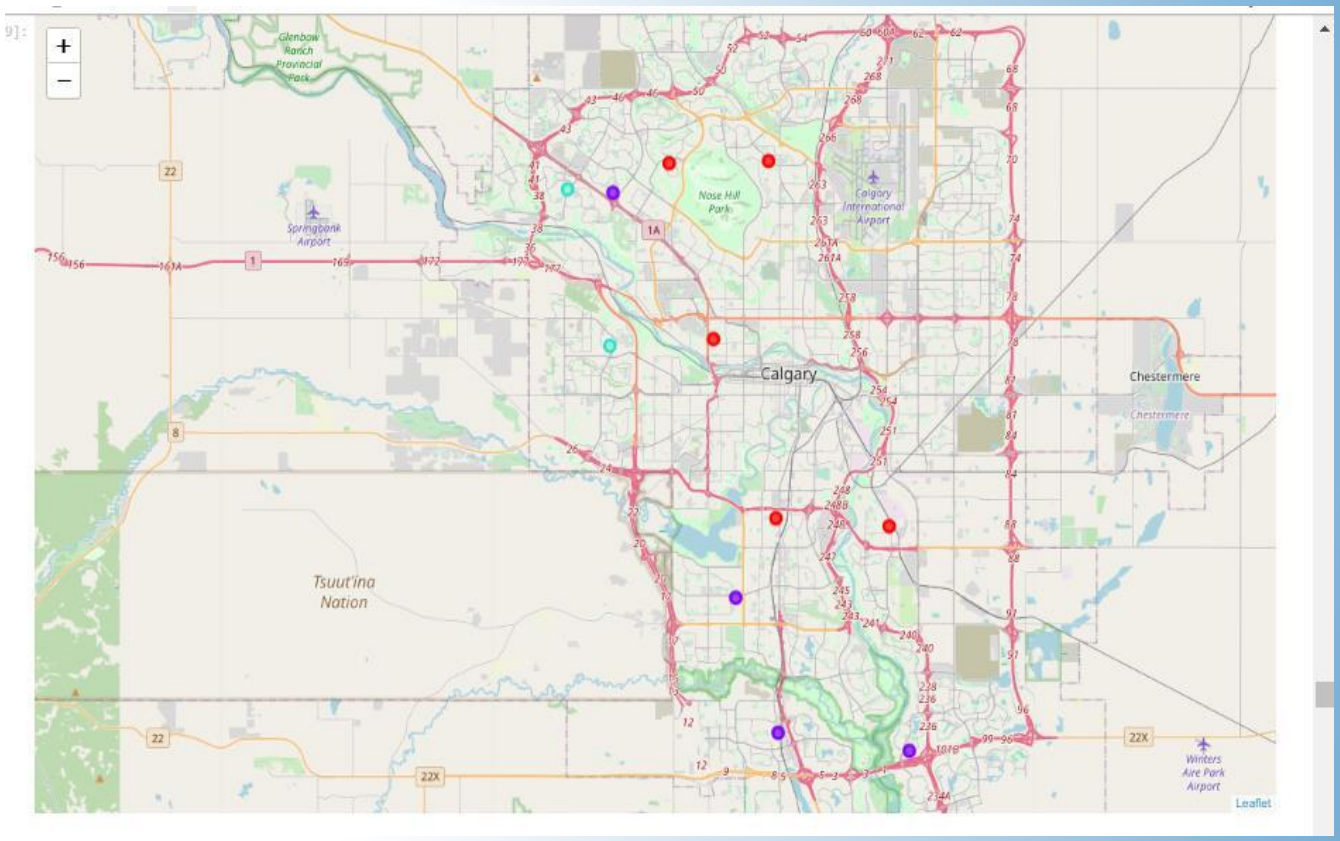
- After iteration, 4 clusters
  - Cluster 1:
    - West side, High Income, low walkability, low restaurant density
  - Cluster 2:
    - medium income, full range walk score, medium – high restaurant density
  - Cluster 3:
    - low – medium income, low – medium walkability, medium restaurant density
  - Cluster 4:
    - Higher income, low – medium walkability, full spectrum restaurant density



- Considering all variables, postal codes from clusters 1 and 4 selected for further analysis
  - Trying to find a sweet spot for income, medium walkability, lower density

SECOND ANALYSIS

- Second Clustering - postal codes from First analysis Clusters 1 and 4
  - Cluster 1: higher income, low – medium walkability, full spectrum restaurant density
    - Postal code T2V chosen for further analysis
  - Cluster 2: higher income, low walkability, low restaurant density
    - All chosen – difficult to separate
  - Cluster 3:
    - Same as Cluster 1 f/ First Analysis
    - Postal code T3H chosen for further analysis



FINAL ANALYSIS

- Most promising postal codes f/ Second Analysis
- Weighted comparison of the 3 measures
- Applying weights, Total Score computed
  - Weighted Income + Weighted Walk Score – Weighted Restaurant Density
    - - on the Restaurant Density to reflect a lower number being more desirable
- Decision made to name a top 3
  - Scoping analysis, not definitive

Measure	Weighting	Weighting Sensitivity	Explanation
Income	0.4	0.33 – 0.5	Income given higher weighting, as disposable income is the highest predictor as to whether a person will spend money in a restaurant
Walkability	0.4	0.33 – 0.4	Walkability given high weighting, as with changes in traffic flows people will look to eat (when possible) closer to home when going out. Also considered traffic from take out services (Doordash etc)
Restaurant Density	0.2	0.1 – 0.33	Given slightly lower weighting. Areas with high density have been filtered out in previous steps. The assumption is the areas left can absorb another restaurant under the right conditions

CONCLUSIONS / FURTHER ANALYSIS

- Following 3 communities noted as most promising by analysis

[87]:

	Postal Code	Neighbourhood	Income - Weighted	Walk Score - Weighted	Restaurants/1000 residents - Weighted	Total Score	Latitude	Longitude	Rank
0	T3H	Discovery Ridge, Signal Hill, West Springs, Ch...	0.798361	-0.279459	-0.164943	0.683844	51.0566	-114.1815	1
1	T2Z	Douglas Glen, McKenzie Lake, Copperfield, East...	0.096639	-0.035346	-0.164512	0.225805	50.9023	-113.9873	2
2	T2V	Oak Ridge, Haysboro, Kingsland, Kelvin Grove, ...	-0.510321	0.863706	0.327472	0.025914	50.9909	-114.0740	3

- Potential Further Analysis
  - Better (pay for service) data
    - Neighborhood level analysis
    - Possible demographic analysis
    - Eliminate grouping/averaging
  - More up to date data
    - Census data analysis lags by years