Capstone Project –
The Battle of Neighborhoods (Week 1)

1. Introduction/Business Problem

In this project, I assume some friend asked me: is it fine to open a Chinese restaurant in Ford Bend, TX; and if it is, which region (zip code) for this restaurant. The selections of the business type (Chinese restaurant) and the county (Ford Bend, TX) for consideration are only due to my personal curiosity. Any other business types and locations can be studied (given the required data is available). To recommend the business locations, a classification model is developed based on the data from a nearby county Harris, TX, and many relevant factors, such as the neighborhood demographics, average house value, average income per household are included in the model development.

The location of a business unit, especially for a restaurant, often bears significant influence on the prosperity and profitability of that business. Therefore, I believe this topic is worth investigation and the developed workflow, maybe seem inmmature currently but upon continuous improvement, will have real business application in the future.

2. Data Sets

In order to train a model for restaurant location selection, the main factors influencing the profitability of a restaurant should be identified. There are many discussions on what factors should be considered, such as visibility, parking, space size, crime rates, surrounding businesses/competitor analysis, accessibility, affordability, safety[1]. Yet parameterizing these factors and finding the corresponding data sets are not easy. I spent a lot of time in selecting and

appraising the appropriate data set. Finally, I concentrated my attention on the data contained in zip-codes.com. The specific webpage is https://www.zip-codes.com/county/tx-harris.asp. Fig.1 shows an excerption of the table contained in this webpage. The first column lists all the zip codes within Harris County, TX. By clicking each zip code, the browser is then directed to another webpage containing the detailed demographical data and other statistical data corresponding to the zip code, such as Fig.2 to Fig.4 show the Zip Code data, 2010 Census Demographics and other demographics for Zip Code 77002.

HARRIS County, TX Covers 230 ZIP Codes					
ZIP Code	Classification	City	Population	Timezone	Area Code(s)
ZIP Code 77001	P.O. Box	Houston	0	Central	832/713/281/346
ZIP Code 77002	General	Houston	16,793	Central	832/713/281/346
ZIP Code 77003	General	Houston	10,508	Central	832/713/281/346
ZIP Code 77004	General	Houston	32,692	Central	832/713/281/346
ZIP Code 77005	General	Houston	25,528	Central	713/832/346
ZIP Code 77006	General	Houston	19,664	Central	713/832/346
ZIP Code 77007	General	Houston	30,853	Central	713
ZIP Code 77008	General	Houston	30,482	Central	713
ZIP Code 77009	General	Houston	38,094	Central	713
ZIP Code 77010	General	Houston	366	Central	832/713/281/346
ZIP Code 77011	General	Houston	19,547	Central	713
ZIP Code 77012	General	Houston	20,719	Central	713
ZIP Code 77013	General	Houston	17,602	Central	713
ZIP Code 77014	General	Houston	28,684	Central	281/832/346
ZIP Code 77015	General	Houston	53,621	Central	713/281
ZIP Code 77016	General	Houston	26,989	Central	281/713
ZIP Code 77017	General	Houston	32,561	Central	713
ZIP Code 77018	General	Houston	25,563	Central	713
ZIP Code 77019	General	Houston	18,944	Central	713/832/346
ZIP Code 77020	General	Houston	25,464	Central	713

Fig.1 Excerption of Harris County Table

ZIP Code 77002 Data	
Zip Code:	77002
City:	Houston
State:	TX [Texas]
Counties:	HARRIS, TX
Multi County:	No
City Alias(es):	Houston Clutch City
Area Code:	832 / 713 / 281 / 346
City Type:	P [Post Office]
Classification:	[Non-Unique]
Time Zone:	Central (GMT -06:00)
Observes Day Light Savings:	Yes
Latitude:	29.750209
Longitude:	-95.367693
Elevation:	38 ft
State FIPS:	48
County FIPS:	201
Region:	South
Division:	West South Central
Intro Date:	<2004-10

Fig.2 Zip Code 77002 Data

ZIP Code 77002 2010 Census Demographics	
Current Population:	9,099
2010 Population:	16,793
Households per ZIP Code:	3,080
Average House Value:	$233,000
Avg. Income Per Household:	$72,306
Persons Per Household:	1.31
White Population:	8,842
Black Population:	6,687
Hispanic Population:	3,248
Asian Population:	366
American Indian Population:	82
Hawaiian Population:	16
Other Population:	1,001
Male Population:	13,839
Female Population:	2,954
Median Age:	33.40 years
Male Median Age:	33.50 years
Female Median Age:	32.90 years

Fig.3 Zip Code 77002 Census Demographical Data

ZIP Code 77002 Other Demographics	
# Residential Mailboxes:	6,946
# Business Mailboxes:	3,410
Total Delivery Receptacles:	10,209
Number of Businesses:	2708
1st Quarter Payroll:	$4,202,146,000
Annual Payroll:	$12,990,674,000
# of Employees:	97,041
Water Area:	0.049 sq mi
Land Area:	2.019 sq mi
113th Congressional District:	02|18
113th Congressional Land Area:	308.75|235.2 sq mi
Single Family Delivery Units:	54
Multi Family Delivery Units:	6,388
# Residential Mailboxes:	6,946
# Business Mailboxes:	3,410
Total Delivery Receptacles:	10,209

Fig.4 Zip Code 77002 Other Demographical Data

Since there is no ready-to-download file summerizing these data provided by website. A python code applying modules of BeautifulSoup and requests is developed to scrape the information from the specific webpage and all the pages linking the zip codes. The codes for scraping and processing the data are summarized in read_data.py and process_data.py. Process_data.py processes the data frame created by read_data.py, such as dropping the unnecessary columns, removing the "$" and "," in currency columns, and others.

The final data set acquired is shown in Fig.5 as an example. There are 131 zip codes and 25 attributes, such as longitude, current population, average income per household, racial/age distribution, and others.

	Zip Code	Latitude	Longitude	Current Population	2010 Population	Households per ZIP Code	Average House Value	Avg. Income Per Household	Persons Per Household	White Population	...
0	77002	29.750209	-95.367693	9099	16793	3080.0	233000.0	72306.0	1.31	8842	...
1	77003	29.748829	-95.343842	13997	10508	3894.0	272800.0	59575.0	2.41	5494	...
2	77004	29.727170	-95.361846	34014	32692	12802.0	247300.0	48592.0	2.02	9752	...
3	77005	29.717416	-95.418732	25502	25528	9548.0	940800.0	180758.0	2.43	21639	...
4	77006	29.739568	-95.388252	24930	19664	11809.0	430600.0	82878.0	1.62	16389	...

...	Other Population	Male Population	Female Population	Median Age	Male Median Age	Female Median Age	# Residential Mailboxes	# Business Mailboxes	Total Delivery Receptacles	Number of Businesses
...	1001	13839	2954	33.4	33.5	32.9	6946.0	3410.0	10209.0	2708
...	2111	5709	4799	31.8	33.0	30.6	5808.0	760.0	7003.0	398
...	1814	16368	16324	29.4	29.8	29.1	16839.0	1264.0	20051.0	725
...	394	12528	13000	38.7	38.0	39.3	10495.0	1018.0	13659.0	1014
...	1285	11111	8553	35.5	37.0	33.6	15389.0	1145.0	16673.0	1028

Fig.5 Example of data set acquired from zip-codes.com.

Since we have the latitudes and longitudes of zip codes in the data set, we can plot them on a map through the method we learned, which is shown in Fig.6.

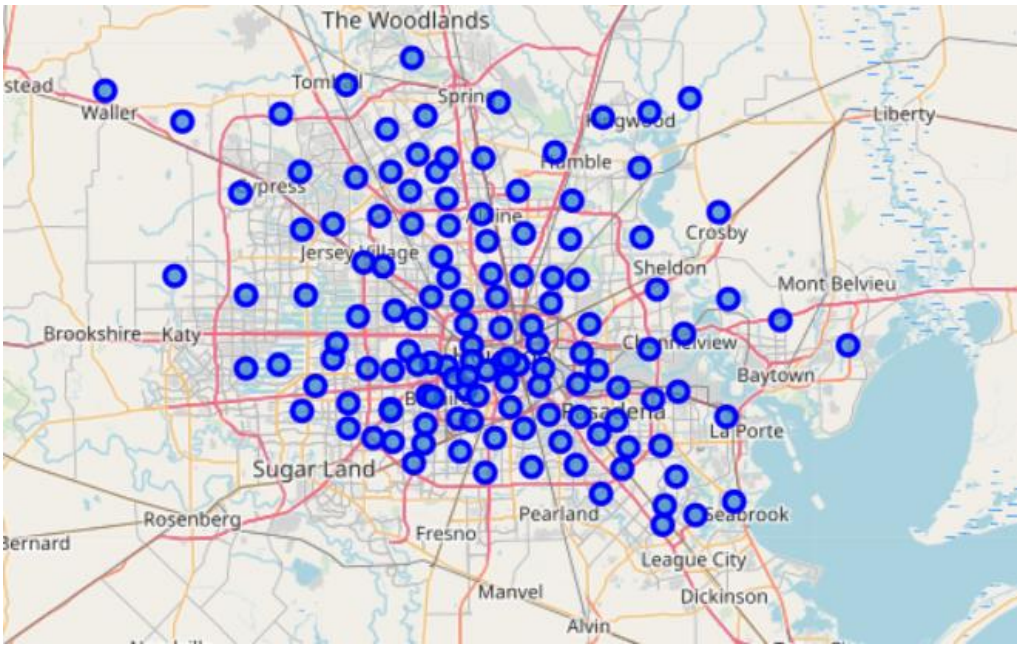

Fig.6 Regions (within Harris, TX) in data set acquired from zip-codes.com

We will develop our classification model to determine which location is good for a Chinese restaurant based on the data set shown in Fig.5, also the data set acquired through Foursquare for each zip code. Detailed discussions are covered in later parts of this report.

3. References

[1] Tom Larkin (2017, September). 8 Factors for Choosing a New Restaurant Location. Retrieved from https://www.foodnewsfeed.com/fsr/vendor-bylines/8-factors-choosing-new-restaurant-location.