ETL Project Report:

Summary:

The project applied the ETL work flow to store data collected from the Department of Public Health and Yelp API request to display the customer-based rating and the public health inspection rating. One of the policy questions could be looking into the correlation between the restaurants' inspection scores and its customers rating. Good customer-based rating restaurant may not provide the healthier food environment to their customers. We are about to discover answer in this ETL project.

Note:

1. Even though we are only using SF data in this project, the scope of the analysis can be expanded to the national level, or even international level.
2. Yelp has partnered with the local government agents developed the Local Inspector Value Entry Specification (LIVES) system, which allow people to look for the inspection score and history on their mobile device through a separated app. However, the system is partnered with other local web developers, which has no link to Yelp database and each app can only display inspection history in the local region.

Resources:

The data are coming from two sources,

1. Restaurant Inspection Scores, San Francisco Department of Public Health
   (After cleaning the data, n = 54,314)

2. Customer-Based Rating Scores, Yelp API Request
   (After cleaning the data, n = 4049)

ETL Work Flow:

Step 1:

Downloading the SF local inspection record from SF Department of Public Health web site.

Inspection Record Data Frame:

	business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude
0	93268	Harbor Court Hotel	165 Steuart St.	San Francisco	CA	94105	NaN	NaN
1	70142	Rosa Mexicano	30 Mission St	San Francisco	CA	94105	NaN	NaN
2	2538	KATES KITCHEN	471 HAIGHT St	San Francisco	CA	94102	37.772163	-122.429927
3	98974	Brickhouse	426 BRANNAN ST	San Francisco	CA	94107	NaN	NaN
4	99342	LAI HONG RESTAURANT	1416 POWELL ST	San Francisco	CA	94133	NaN	NaN

5 rows × 23 columns

Step 2:

Using Pandas cleaning the data before importing to a database.

- Change business name and address to lower case for joining in the database
- Change the header names
- Change the zip codes to 5-digit string
- Change the business phone to 11-digit string

	business_id	name	address	business_city	business_state	inspection_id	inspection_date	inspection_score	inspection_type
0	93268	harbor court hotel	165 steuart st.	San Francisco	CA	93268_20180411	4/11/2018 0:00	NaN	New Ownership
1	70142	rosa mexicano	30 mission st	San Francisco	CA	70142_20190408	4/8/2019 0:00	NaN	Reinspection/Followup
2	2538	kates kitchen	471 haight st	San Francisco	CA	2538_20170608	6/8/2017 0:00	73.0	Routine - Unscheduled
3	98974	brickhouse	426 brannan st	San Francisco	CA	98974_20190321	3/21/2019 0:00	NaN	New Ownership
4	99342	lai hong restaurant	1416 powell st	San Francisco	CA	99342_20190222	2/22/2019 0:00	NaN	New Ownership

Step 3:

Sending request to Yelp API for the customer rating data and store into a data frame.

	name	rating	reviews	address	city	state	zip_code	phone
0	Mr Szechuan	4.5	81	890 Taraval St	San Francisco	CA	94116.0	1.415754e+10
1	El Chango Salteño	4.5	2	Avenida Santa Fe 380	Alberdi	X	5000.0	5.435143e+11
2	Che Fico Alimentari	4.5	3	834 Divisadero St	San Francisco	CA	94117.0	1.415417e+10
3	Tuna Kahuna	5.0	64	1117 Burlingame Ave	Burlingame	CA	94010.0	1.650637e+10
4	New England Lobster Market & Eatery	4.5	3084	824 Cowan Rd	Burlingame	CA	94010.0	1.650443e+10

Step 4:

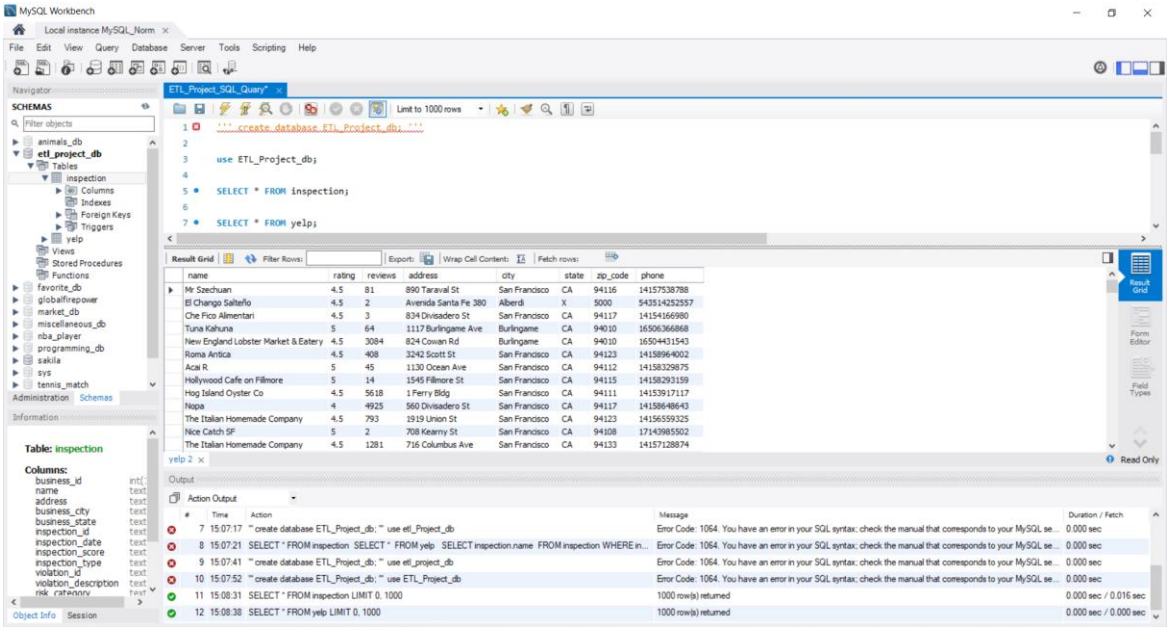
Using Pandas cleaning the data before importing to a database.

- Change business name and address to lower case for joining in the database
- Change the header names
- Change the zip codes to 5-digit string
- Change the business phone to 11-digit string

	name	rating	reviews	address	city	state	phone	zip
0	mr szechuan	4.5	81	890 taraval st	San Francisco	CA	14157538788	94116
1	el chango salteño	4.5	2	avenida santa fe 380	Alberdi	X	543514252557	5000
2	che fico alimentari	4.5	3	834 divisadero st	San Francisco	CA	14154166980	94117
3	tuna kahuna	5.0	64	1117 burlingame ave	Burlingame	CA	16506366868	94010
4	new england lobster market & eatery	4.5	3084	824 cowan rd	Burlingame	CA	16504431543	94010

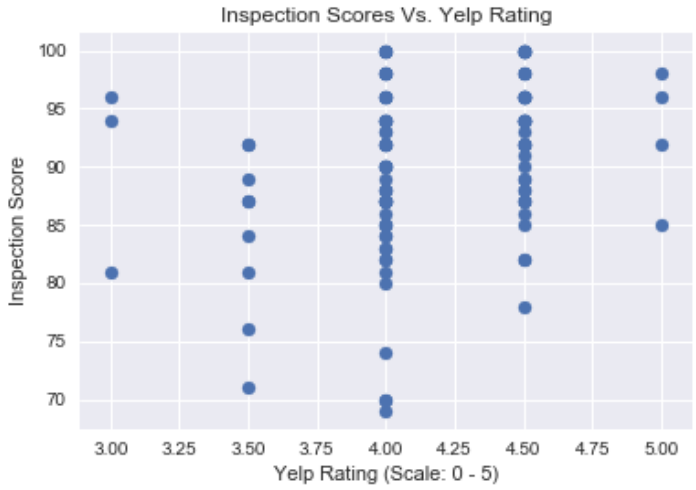
Step 5:

Importing the two data frames into SQLite database. (Optional: MySQL database)



Step 6:

Using Pandas joining the two data frames and creating graphs for the analysis.



Conclusion:

In this project, we do not find any evident suggest that there is a strong correlating between customer-based rating and public health rating. However, the take away from this project is that we can gather data from both the public and private sectors, then store it into a database, so that we can display the search results in place by connect the database to a web page.