Data

2.1 Data Scraping and Cleaning

In this section we will first retrieve the geographical coordinates of the three cities (San Francisco, Los Angeles, and San Diego). Then, we will leverage the FourSquare API to obtain URLs that lead to the raw data in JSON form. We will speerately scrape the raw data in these URLs in order to retrieve the following columns: "name", "categories", "latitude", "longitude". and "id" for each city. We can also provide another column ("city") to indicate which city the restaurants are from.

It is important to note that the extracts are not of every restaurant in those cities but rather all of the restaurants within a 1000KM range of the geographical coordinates that geolocator was able to provide. However, the extraction from the FourSquare API actually obtains venue data so it will include venues other than restaurants such as concert halls, stores, libraries etc. As such, this means that the data will need to be further cleaned somewhat manually by removing all of the non-restaurant rows. Once this is complete, we have a shortened by cleaned list to pull "likes" data. The reason the cleaning takes precedence is mainly that pulling the "likes" data is the computing process which takes the longest time in this project so we want to make sure we are not pulling information that will end up being dropped anyways.

The "id" is an important column as it will allow us to further pull the "likes" from the API. We can retreive the "likes" based on the restaurant "id" and then append it to the data frame. Once this is complete, we finally name the dataframe 'raw_dataset' as it is the most complete compiled form before needing any processing for analysis via machine learning.

2.2 Data Preparation

The data still needs some more processing before it is suitable for model training and testing. Mainly, the "categories" column contains too many different types of cuisines to allow a model to yield any meaningful results. However, the different types of natural cuisines have natural groupings based on conventionally accepted cultural groupings of cuisine. Broadly speaking, all of the different types of cuisine could be reclassified as European, Latin American, Asian, North American, drinking establishments (bars), or casual establishments such as coffee shops or ice cream parlours. We can implement manual classification as there really aren't that many different types of cuisines.

As this project will compare both linear and logistic regression, it makes sense to have "likes" as both a continuous and categorical (but ordinal) variable. In the case of turning into a categorical variable, we can bin the data based on percentiles and classify them into these ordinal percentile categories. I tried different ways of binning but in the end, splitting the sample into three different bins proved to yield the best classification results from a prediction standpoint.

As the last stage of data preparation, it is important to note that the regressors are categorical variables (3 different cities and 6 different categories of cusines). Hence, they require dummy variable encoding for meaningful analysis. We can accomplish this via one-hot encoding.

	name	categories	lat	lng	id	city	likes
7	Birba	Wine Bar	37.777750	-122.424159	551b7760498e1612b67f33f9	San Francisco	132
8	Blue Bottle Coffee	Coffee Shop	37.776286	-122.416867	5560dbdb498e91a2bcde84f6	San Francisco	457
9	Blue Bottle Coffee	Coffee Shop	37.776430	-122.423224	43d3901ef964a5201f2e1fe3	San Francisco	941
12	Philz Coffee	Coffee Shop	37.781266	-122.416901	5151a10ce4b06ae7735335db	San Francisco	480
14	Ritual Coffee Roasters	Coffee Shop	37.776476	-122.424281	4dd94350e4cd37c893d8146f	San Francisco	823
16	a Mano	Italian Restaurant	37.776917	-122.423856	5907aa1c178a2a76066e0f59	San Francisco	347
19	Saigon Sandwich	Sandwich Place	37.783084	-122.417650	43eb7d31f964a520392f1fe3	San Francisco	439
24	Biergarten	Beer Garden	37.776013	-122.424247	4dd7e48fd22d38ef42f35bd8	San Francisco	801
26	George and Lennie	Coffee Shop	37.781701	-122.415213	55da2db5498eb79ab95580cb	San Francisco	79
28	Hinata	Sushi Restaurant	37.783090	-122.420732	584a0d7b349355671db7eab2	San Francisco	69
29	Gioia Pizzeria	Pizza Place	37.776328	-122.425859	5cfaca87a2c00b002bd3242e	San Francisco	47
32	City Beer Store	Beer Bar	37.778493	-122.412244	5b83612295a722002ccd6e80	San Francisco	113
34	Suppenküche	German Restaurant	37.776324	-122.426382	42c5d900f964a520d5251fe3	San Francisco	817

➔ Data above is before the data pre processing

	name	american	asian	bar	casual	euro	latino	Los Angeles	San Diego	San Francisco	ranking	likes
7	Birba	0	0	1	0	0	0	0	0	1	2	132
8	Blue Bottle Coffee	0	0	0	1	0	0	0	0	1	3	457
9	Blue Bottle Coffee	0	0	0	1	0	0	0	0	1	3	941
12	Philz Coffee	0	0	0	1	0	0	0	0	1	3	480
14	Ritual Coffee Roasters	0	0	0	1	0	0	0	0	1	3	823
16	a Mano	0	0	0	0	1	0	0	0	1	3	347
19	Saigon Sandwich	0	0	0	1	0	0	0	0	1	3	439
24	Biergarten	0	0	1	0	0	0	0	0	1	3	801
26	George and Lennie	0	0	0	1	0	0	0	0	1	2	79
28	Hinata	0	1	0	0	0	0	0	0	1	2	69
29	Gioia Pizzeria	0	0	0	0	1	0	0	0	1	1	47
32	City Beer Store	0	0	1	0	0	0	0	0	1	2	113
34	Suppenküche	0	0	0	1	0	0	0	0	1	3	817
35	Petit Crenn	0	0	0	0	1	0	0	0	1	3	260

➔ Data after data pre processing