Roadmap of files on GitHub Repository for Foodies – Project 1

Group Members: Carly Russell, Loba Quasem, Phillip Choi, Natalie Stanislov

Please note: all files are on the master, unless specified that they are on another branch

1. **Project Planning**

Cities.csv in the API Calls/Resources folder, to find the cities in the bay area, and help us to generate YELP API queries

Cities\_by\_Zip, file of the bay area zips, cities and counties

Counties\_by\_zip was a different version of the file, only has the zips and the counties

1. **Yelp API Calls**

calls\_by\_city.ipynb –first Yelp API calls were performed by city (Located in the API Calls Folder – Please reference the README information and output files)

calls\_by\_city\_ratings.ipynb Yelp queries (Located in the API Calls Folder – Please reference the README Folder for information and output files)

calls\_by\_zip.ipynb – Yelp search by zip code (Located in the API Calls Folder – Please reference the README Folder for information and output files)

coordinates\_grid.ipynb is a jupyter notebook creating lat & lng coordinates through out the bay area. The first jupyter notebook had the coordinates spaces every 5 miles, the second every 3 miles. (Located in the API Calls Folder – Please reference the README Folder for information and output files)

calls\_by\_coord\_ratings\_ipynb – calls by coordinates 5 miles apart with 5 mile search radius, query return is a sorted results from YELP. (Located in the API Calls Folder – Please reference the README Folder for information and output files)

Hypotenuse theory calls\_by\_coordinate\_radius\_4023\_testing.ipynb – A test file with Yelp API to see if reduced graphical coordinates and shorter search radius would be better. If one thinks about the geographical coordinates spaced out every 5 miles, then the search radius selected was 2.5 miles and covering 78% of bay area while the search went on for the restaurants. (Located in the Test API Calls Folder)

Hypotenuse theory calls\_by\_coordinate\_radius\_5690\_testing.ipynb – A test file with Yelp API to see if reduced graphical coordinates and a shorter search radius would be better. If one thinks about the geographical coordinates spaced out every 3 miles, then we want to grab all restaurants in the coordinate grid rectangular spacing with minimum restaurant duplicates returned, in order to reduce the number of API calls over 1K. (Located in the Test API Calls Folder)

Half way between 3 miles = 1.5

1.5 = y

Y^2 + Y^2 = r^2

Whereby, r will give the search radius of returning all restaurants in a geocoordinate block while minimizing duplicates.

calls\_by\_coord.ipynb - calls by coordinates 3 miles apart with a smaller search radius

Located in the API Calls Folder – (Please reference the README file in the API Calls Folder for more information on this file and the contents of the API Calls Folder)

calls\_for\_2\_coords\_rating.ipynb – running code to get back restaurants sorted by ratings to 2 coordinates that were over 1K (Located in the API Calls Folder – Please reference the README Folder for information and output files)

calls\_by\_coord\_ver2.ipynb – calls by coordinates 3 miles apart with a reduced radius with updated code for category parameters (Located in the API Calls Folder – Please reference the README Folder for information and output files)

1. **Data Cleaning**

Merged\_zip.csv file, was the output from the first merge in coordinates 2 data set

Cleaning\_Yelp\_Data\_On\_Coodinates\_2.ipynb, First Pass at Cleaning with Coordinates 2 Data set, which was supposed to be prep for cleaning Coordinate 3 Data Set, the final Yelp API.

Loba working on zips.ipynb – loba’s file of her work for cleaning zips codes

Zip\_of\_9\_counties.csv – starting file of bay area counties, and zip codes to use in the Cleaning\_Yelp\_Data\_on\_Coordinates3\_Calls.ipynb in the resources folder in GitHub master

Cleaning\_Yelp\_Data\_On\_Coordinates3\_Calls.ipynb cleaning notebook

analysis&merging. ipynb - Cleaning data with the “Resources/restaurant\_data\_coords3.csv”, dropping duplicates, matching zips and concatenating

Cleaned\_zip.csv – Final Clean Bay Area Zip Codes to merge with the data set to get the bay area counties on the data table by matching with their zip codes, an output from Cleaning\_Yelp\_Data\_On\_Coordinates3\_Calls.ipyn

restaurant\_data\_coords\_rating3.csv – file for looking at category and cleaning, in the Resources folder

Cleaning\_Yelp\_Data\_On\_Coordinates3\_Calls.ipynb. Final version of cleaning for Coordinates 3 Data Set (final data set)

Cleaning\_2.csv, located in the GitHub project-1 master resources file was an output file used in Cleaning\_Yelp\_Data\_On\_Coordinates3\_Calls.ipynb to merge with counties\_by\_zip.csv file

Unclean\_2.csv – located in the resources folder, file used to clean up the zips and put all the unwanted zips on this file to ensure there were no lines we didn’t want to keep.

Restaurants\_cleaning\_1.csv – file in the Resources folder, an intermediate file used in cleaning generated from the Cleaning\_Yelp\_Data\_On\_Coordinates3\_Calls.ipynb file

Restaurants\_cleaning\_2.csv – file to find categories of cuisines that our group approved of

Original Categories.list.xlsx – first pass of categorizing cuisine types

Category Lumped.csv -file helped categorize cuisine types

Category Lumped2.xlsx – a pass at categorizing cuisine types

restaurant\_data\_coords\_rating3.csv – file for looking at cleaning categories, in the Resources folder

restaurants\_with\_county.csv – file generated from the Cleaning\_Yelp\_Data\_On\_Coordinates3\_Calls.ipynb file, an intermediate file

Preliminary\_Category\_Sorting\_Lumping\_by\_Counties.ipynb – pass at cleaning the categories that are unwanted and then creating a list of categories by county by restaurant ratings on final clean data set.

Categories.ipynb – file of initial/draft category cleaning, in the Resources folder on master GitHub

Final\_Clean\_Restaurants.csv – in the Resources folder, the first final cleaned data set

unique\_categories\_by\_county\_csv - file was going to be used to find out the unique number of cuisine categories per category, this was generated from the Preliminary\_Category\_Sorting\_Lumping\_by\_Counties.ipynb

unique\_categories\_by\_county\_rating3.5&up.csv - file was going to be used to find out the unique number of cuisine categories per category, this was generated from the Preliminary\_Category\_Sorting\_Lumping\_by\_Counties.ipynb

Reanalysis performed in excel to lump categories a different way in excel, file is “cleaned\_category\_19848.csv” located on the Nat branch

Appr\_cat\_list.csv was the file of the final categories after lumping, located on the Nat Branch.

1. **Data Analysis & Graphing**

Plot\_coord.ipynb Creates scatter plot using lat/lng and total restaurants at each coordinate pair

Pricing\_&\_Rating.ipynb file for analysis on the Final Clean Data Set

Restaurantsbyprice.png analysis graph overview of the bay area on the Final Clean Data Set

**Bayarea.mostpop10.png – analysis was performed on Final Clean Restaurants.csv in the Resources GitHub file**

**Bayarea\_piechart.mostpop10.png – Analysis performed on Final Clean Restaurants.csv**

Percentages of the 3.5 to 5 Star Ratings by county, on Final Clean Data.xlsx – An analysis to look at the distribution of the ratings by county.

County\_graphs.pptx on Nat Branch – county graphs generated of a reanalysis performed in excel, “cleaned\_category\_19848.csv”. This file is located in Nat branch

Much of the graphing was performed in Nats brach: a bar graph, a bell curve of ratings, box whisker, county analysis, etc. either off the Final Cleaned Data or off the cleaned\_category 19484.csv file (located in Nat branch)

1. **Project Presentation & Conclusions**
   1. All\_food\_in\_the\_hood.pptx – final presentation