DASC 5300/CSE 5300

Foundations of Computing

Instructor: Sharma Chakravarthy

Project 2: Python Programming and Graph Data Analysis

By

Aditya Vardhan Marri (1002028892)

Ashik Maraliga Venkatesh (1002023308)

Table of Contents:

Python Programming and Data Analysis for Yelp Dataset

Abstraction:

In this study, we intend to share our effort, in which we processed, cleaned, and analyzed the Yelp dataset related to the city “Exton”. Data analysis is the process of gathering, converting, cleaning, and modelling data to discover the underlying information needed. Data visualization is often used to represent data to find relevant patterns in the data.

First extract the necessary information into lists from the given json files containing data related to businesses, reviews and users individually by reading them line by line and create their respective data frames. During this process, get only the businesses which are in the city of Exton, next using the business ids from business json extract the review information of the reviews related to these businesses and finally use these review ids and extract the user information containing any of the review ids. While doing the above extraction process for businesses and users create their unique number and codes during the initial preprocessing for further use. Next merge the columns needed and create a new data frame completeDataFromExton.

Next remove the repeated reviews by users to same business in the completeDataFromExton and take their average value of stars as new data frame finalGraphDataFrame. Next, dummy user information and create a record in the userDataFrame. Next add the dummy user and business relation information to the finalGraphDataFrame data frame. Extract the business_number and user_number columns from the finalGraphDataFrame to the input.txt file to get the characteristics of the graph by passing the file as parameter to the provided the python files.

Next find the users and businesses with most number of reviews and draw their respective histograms. Now plot the bi-partite graph based on the mapping information gathered from initial processing of the data. Also calculate the maxflow for business and user nodes.

File Descriptions:

1. yelp_academic_dataset_business. json is the input file for businesses data.

2. yelp_academic_dataset_review. json is the input file for reviews data.

3. yelp_academic_dataset_user. json is the input file for users data

4. input.txt is the file created consisting of all the business and the user mapping using their unique numbering done during the preprocessing.

5. businessDataFrame, reviewDataFrame and userDataFrame are the data frames created from the data extracted during the initial processing of the related yelp json files.

6.finalGraphDataFrame is the data frame which we use to map the graph using networkx.

Division of Labor:

With the knowledge and experience gained from the project 1 we have started working on the project 2. The entire project was performed in a group setting. We had gone through some important libraries like Pandas, folium, networkx, seaborn to work on this project. After studying logic in each phase of the project we brainstormed our ideas and came up with the effective algorithms to use and we utilized Google, YouTube, and Kaggle to figure out the syntax and analysis the data.

Problems Encountered:

1.Combining the data from different Json files and create a usable data frame with necessary information for analysis. We resolved it using the merge functions by proper initial processing of the json files in the start itself.

2.Taking only one edge if we have multiple reviews given by the same user to same business, so we implemented group by function on business_code and user_code to get the average of the stars.

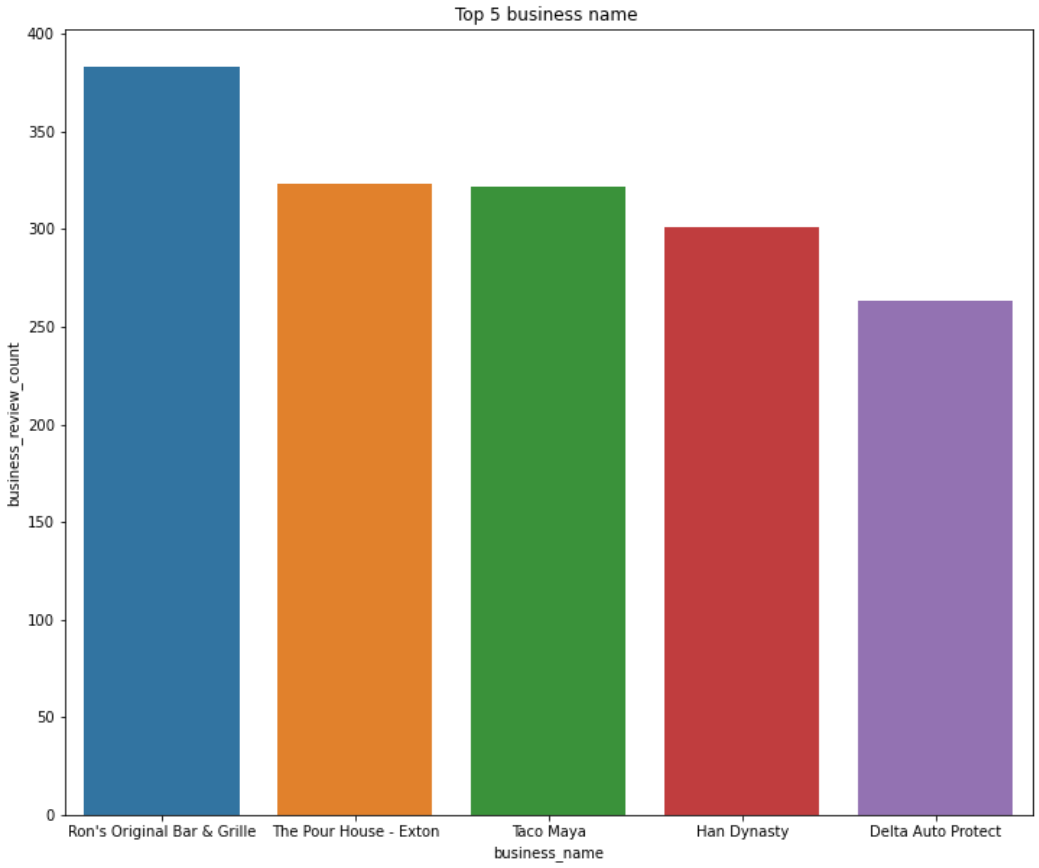
3.Creating the bipartite graph and to make visual distinction of the business and user nodes. Another hurdle in the graph is to highlight the nodes based on their edge count/connectivity. We used node_size function which increases the size based on number of edges.

4. In the last project, using the latitude and longitude we placed the markers on map but in this project we need to show the names and number of reviews along with the markers on the map for the users plot as well as the business plot. We used the tooltip and popup functions belonging to folium library to display the information on the marker when clicked.

Analysis:

Business related Analysis:

We have a total of 419 businesses in the city of Exton. Among the businesses in the city, businesses b20(Ron's Original Bar & Grille),b86(The Pour House - Exton),b216(Taco Maya),b87(Han Dynasty),b340(Delta Auto Protect) have received the most number of reviews in their respective order. Combined all the reviews received by the businesses we have a total of 13315, on an average each business in the city has 31.77 reviews.

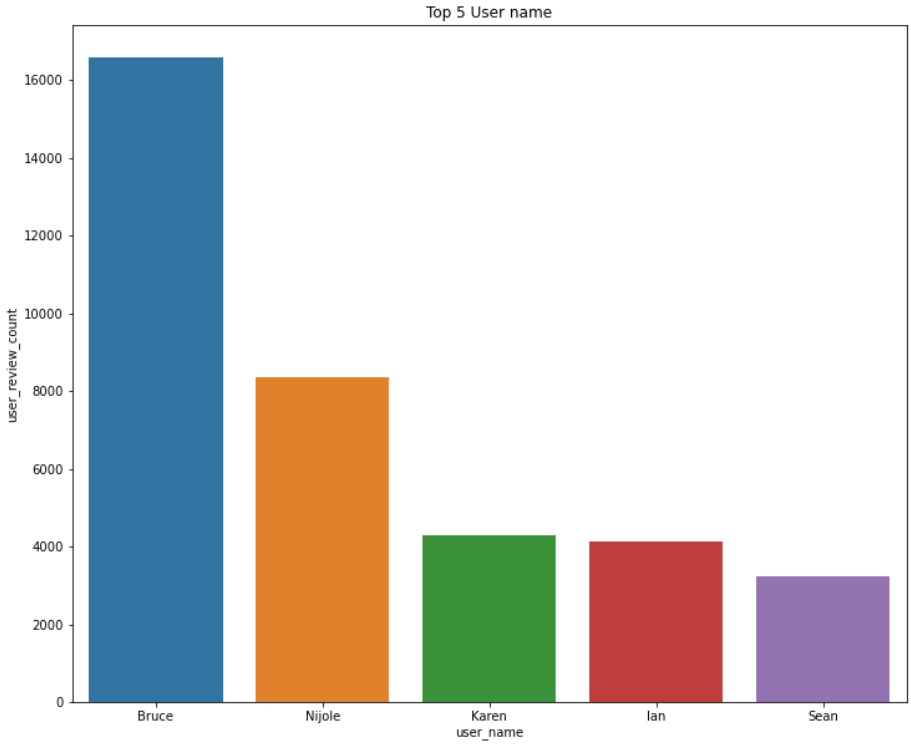


It does not always indicate having more reviews they have the best rating, the businesses having the highest rating the city are b67(Blue Buddha Healing Arts) - 4.791, b333(Maridadi Couture) – 4.687, b95(Damas Mediterranean Grill) – 4.649,b373(Th- Back Space) – 4.615,b204(Kisses Sweet Spa) – 4.517 in the respective order.

Users related Analysis:

In the city of Exton, we have a total of 8155 total users who had given their review to different businesses. The above users have given a total of 13315 reviews in the city, on an average that is 1.64 reviews per person. The people who have given the greatest number of reviews in the city are u812(Cassie), u1111(J.), u81(Matthew), u1574(John), u2753(Zack).

Also, among the people who have given reviews in the city, these are the user who had given reviews in other cities and when we consider that the people with most reviews given are u56(Bruce), u49(Nijole), u13(Karen), u33(Ian) and u29(Sean) in the respective order.

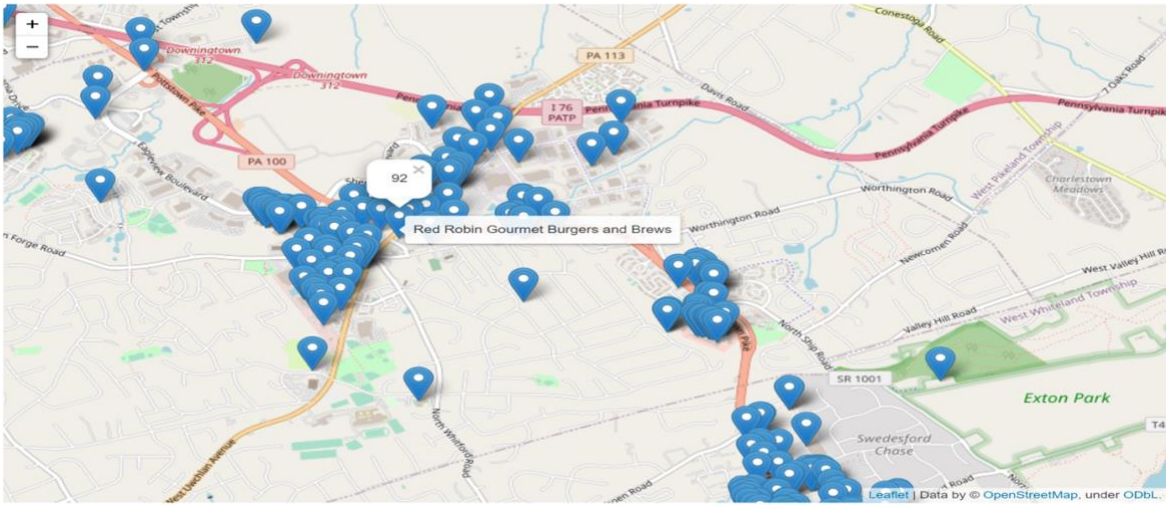


Mapping of Businesses and Users:

We used a python library called folium, which is best for mapping markers and so many.

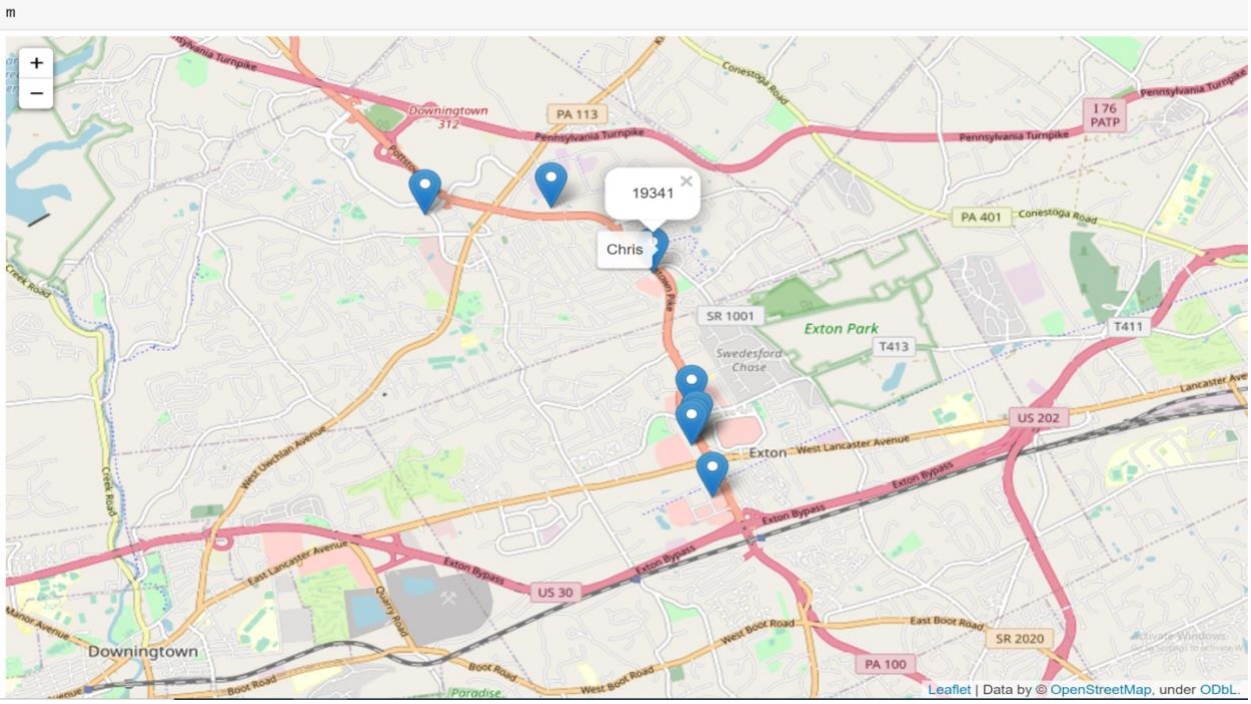
We used popup and tooltip functions of folium to include the information of the name and their unique numbers for business and users markers on their respective map plots.

Business Map:



Users Map:

We added only top 8 users where they live, as most of the users have same pin code ,longitude and latitude, it will be clumsy.



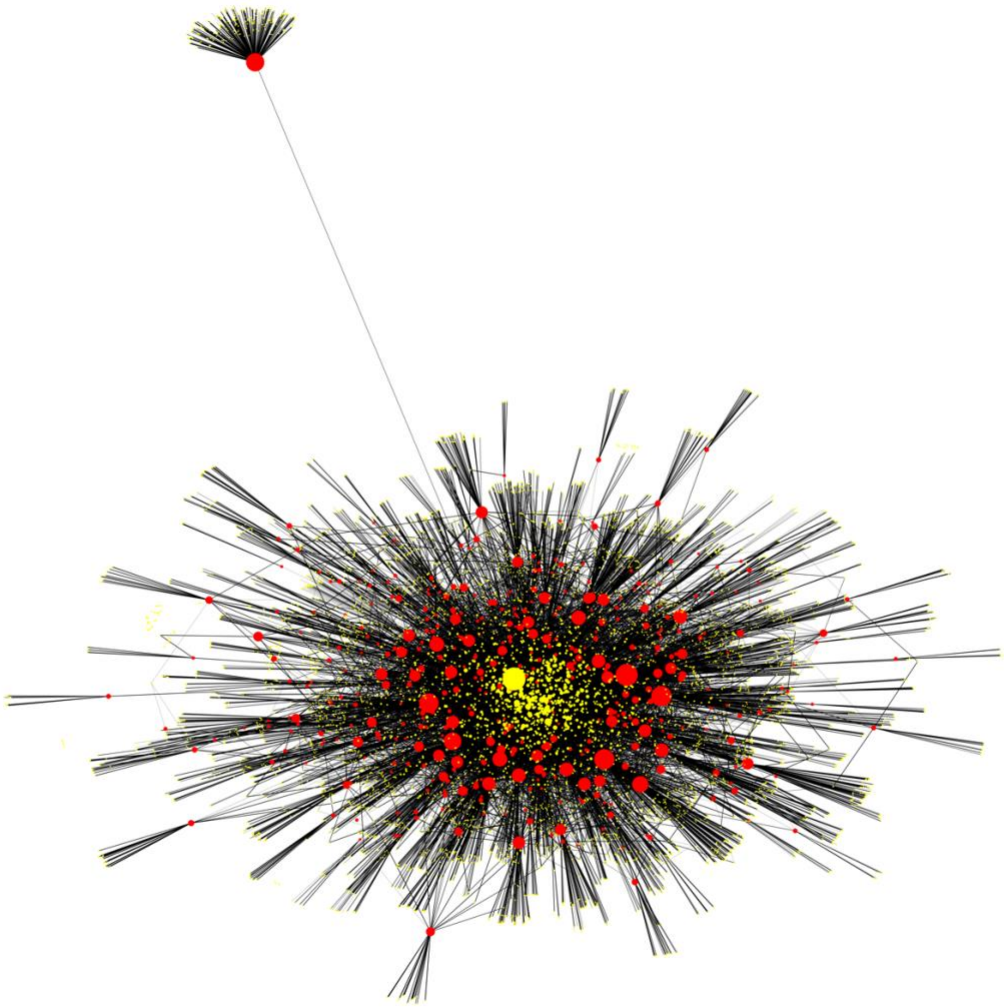
Graph Analysis:

Since the analysis is on single city the graph is disconnected. So, we added a dummy user with id ‘ux’ and mapped to every business in the city to create a connected graph. If we have multiple reviews from a particular user to a single business, we have taken the average of the all the given stars and stars are considered as edge eight.

Using networkx we found the following business nodes b20, b86, b216, b37 and b340 have the highest centrality among businesses. Similarly, among the user nodes, users u812, u81, u1111 and u2753 have the highest centrality excluding the user ‘ux’ as it is a dummy user we introduced. Using the package called Edmonds_karp we calculate max for the 50 user nodes and business nodes.

In the below graph red color nodes are business nodes and yellow ones are the user nodes.

The more the edges the bigger and brighter the node is.



We used the input.txt file created with the mapping data of businesses and users as a parameter to the given python files (MLN.py, MLN_IO.py, network_summary.py) to get the graph characteristics and the below result is generated.

Number of Nodes:8575

Number of Edges :13184

Density:0.000358

Number of connected components:1

Connected Components :8575

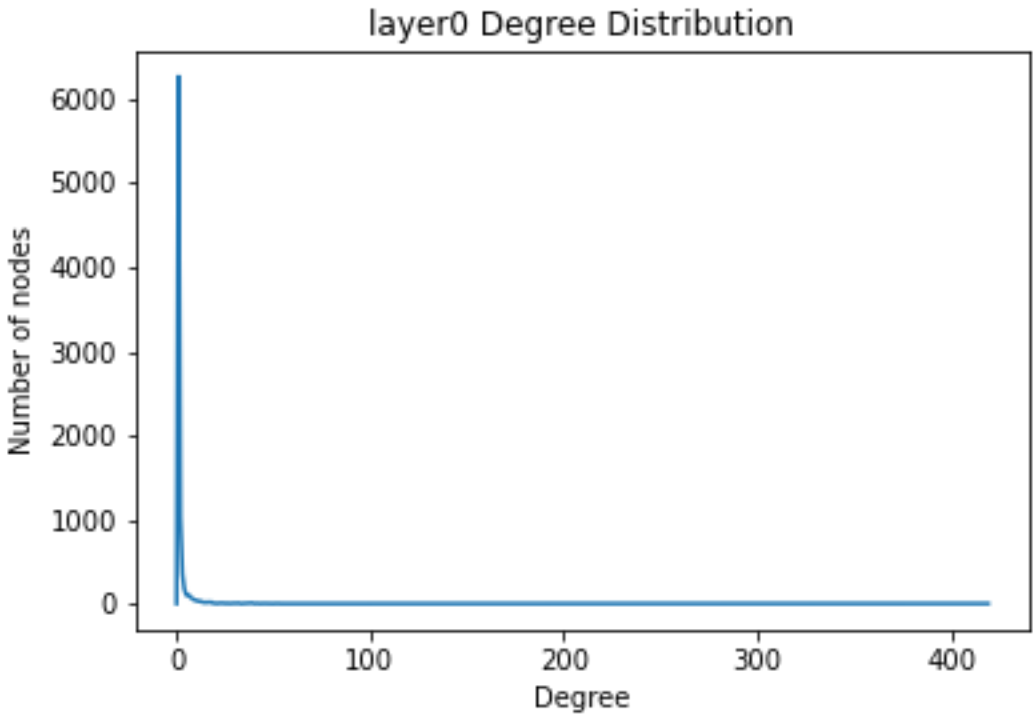
Diameter :4

Min Degree:1

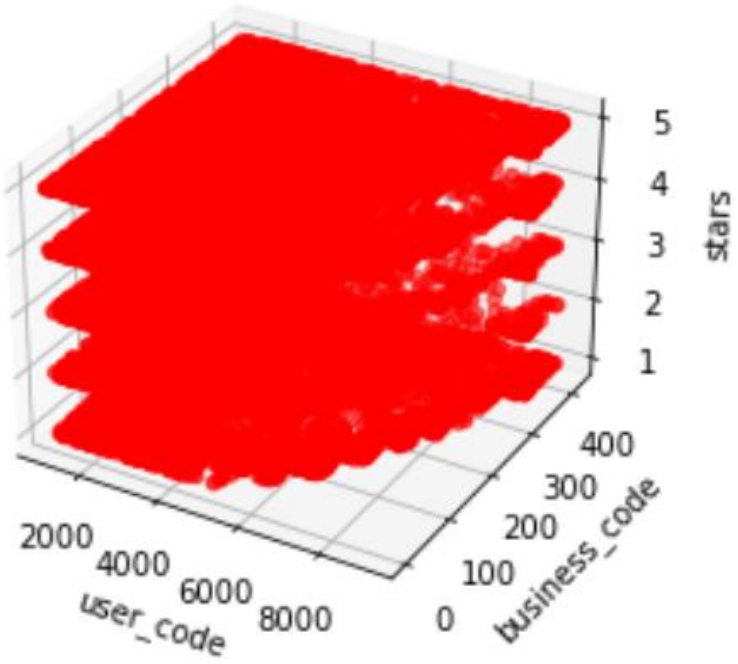
Max Degree:419

Avg Degree:3.07

Std Deviation :13.35



We have plotted the 3D graph using the business_code, user_code and stars and the result is as below.



Conclusion:

After the analysis of the given yelp data set we can observe that most of the businesses are performing well in the city and the average rating for the restaurants is also above 2.5 with a rating of 3. 216.The number of reviews given per person is too low at 1.6 and it would help the businesses to develop more with more reviews to help them understand customers to give better customer service.

References

We looked at several websites and watched a lot of YouTube videos to better comprehend the data and analysis, and we have included the sources below.

1. Reference — NetworkX 2.7.1 documentation
2. seaborn: statistical data visualization — seaborn 0.11.2 documentation (pydata.org)
3. https://www.kaggle.com/code
4. User Guide — pandas 1.4.1 documentation (pydata.org)