GV-CAMBRIDGE CAPSTONE

This report analyzes the city of Cambridge Massachusetts neighborhoods around the Harvard,

Coursera Data Science Professional Certificate
Guido Viarisio June 1st, 2020

Introduction

1. Cambridge City

Cambridge is located across the Charles River from Boson. It has a diverse population of over 100,000 residents. Cambridge is known as one of the leading intellectual hubs of the U.S., attracting college students from all over the world to study at Harvard, the Massachusetts Institute of Technology and Lesley University.
Cambridge attracted a great number of technology-based enterprises like software and biotechnology. Kendall Square is a Cambridge area, a neighborhood in Cambridge that has been dubbed "the most innovative square mile on the planet."

2. The targeted people for this study

Because of the famous schools, Cambridge attracts many students from around the world. Many of the thousands of students of MIT, Harvard and Lesley University have not been in Cambridge before and therefore they are interested in learning about the city:

- Interesting places
- How safe if the city
- Differences in preferences between different neighborhoods
- Potential places to work because Cambridge is "the most innovative square mile on the planet."
    - Range of salaries etc

Data

Because I live in Boston and I know the Cambridge are very well I selected the following areas to be used as the Neighborhoods to be analyzed which are the most useful to the students:

"Lesley University" : "29 Everett St, Cambridge, MA 02138"
"Biogen" : "225 Binney St, Cambridge, MA 02142"
"MIT" : "77 Massachusetts Ave, Cambridge, MA 02139"
"Cambridge Innovation Center" : "1 Broadway, Cambridge, MA 02142"
"Harvard Square" : "1400 Massachusetts Ave, Cambridge, MA 02138"
"Central Square" : "581 Massachusetts Ave, Cambridge, MA 02139"
"Kendal Square" : "325 Main St, Cambridge, MA 02142"

Venues data for each area is obtained using FourSquare APIs

The following additional data is obtained as follows:

Zip codes latitude and longitude:
data from https://public.opendatasoft.com
From the Cambridge Wikipedia scraping the data:
https://en.wikipedia.org/wiki/Cambridge,_Massachusetts
Income by zip code
Main Employers in Cambridge
From Cambridge Open Data:
Statistics on crimes by Neighborhood
https://data.cambridgema.gov/widgets/xuad-73uj

From www.payscale.com

Salaries information by type of job

from https://www.payscale.com/research/US/Location=Cambridge-MA/Salary

From www.cambridgema.gov

Cambridge popular employers range of salaries

Neighborhood polygons geo data to be able to draw the city of Cambridge in a map

https://data.cambridgema.gov/Planning/Cambridge-Neighborhood-Polygons/4ys2-ebga

Methodology

Everything was developed in a Jupiter Notebook and it is divided in the following sections:

0- Setup imports

All the imports that might be necessary are imported here

1- Definition of constants

- FourSquare
  i. URLs that are used with FourSquare APIs
  ii. Client ID, Client Secret and Version
- Cambridge data
  i. Cambridge address to get latitude and longitude
  ii. Python Data dictionary defining the Neighborhoods explained in the Data section

addressesDic = {"Lesley University":{"address":"29 Everett St, Cambridge, MA 02138"},
"Biogen":{"address":"225 Binney St, Cambridge, MA 02142"},
"MIT":{"address": "77 Massachusetts Ave, Cambridge, MA 02139"},
"Cambridge Innovation Center":{"address":"1 Broadway, Cambridge, MA 02142"},
"Harvard Square":{"address":"1400 Massachusetts Ave, Cambridge, MA 02138"},
"Central Square":{"address":"581 Massachusetts Ave, Cambridge, MA 02139"},
"Kendal Square":{"address":"325 Main St, Cambridge, MA 02142"} }

2- Define Generic Functions

- This Python functions do not include visualization functions which are defined later in the Visualization section
  i. populateVenues()
     1. Invokes FourSquare APIs for each Neighborhood to get the venues

2. Stores the information in the Python dictionary addressesDic adding the data frame for the neighborhood created from the data returned by the API

ii. getLatitudeLongitude(address)
1. finds the latitude and longitude for an address

iii. getCategoryType(row)
1. finds the category type for one venue

iv. filterDF(df)
1. filter data returned from API keep only a few columns:
a. name, categories, location and id

v. setupAreasDF()
1. Create the dfAreas data frame with the basic information of the Neighborhoods to be able to show them in a map

vi. showVenuesInfo ()
1. Displays the data stored in the addressesDic for each neighborhood (area)

vii. getNearbyVenues()
1. Uses the FourSquare APIs to explore each neighborhood to be able to get the 10 most common venues in neighborhood.

- Functions that are used to *apply()* to data frames to modify/create some data
  i. findCentroid()
  1. Finds the centroid of an area defined by a polygon

3- Get Data for Cambridge
- Get the latitude and longitude of Cambridge, MA
  i. Using functions defined in section 2
- Populate all the areas defined with FourSquare APIs for venues in each area
  i. Using functions defined in section 2
- Get nearby Venues
  i. Using functions defined in section 2
- Retrieve zip codes longitude and latitude
  i. From the Cambridge open data described before
- Scrape from the Cambridge wikipedia the following
  i. Income by zip code
  ii. Main employers in Cambridge
- Load crime reports from Cambridge Open data
  i. Create two data frames
  1. Total crimes by neighborhood
  2. Total crimes ty type of crime
- Load the Cambridge definition of the polygones
- Load average salary information by type of job
- Load the Cambridge geo data in geojson format
  i. Find the centroid for each Neighborhood

ii. Add crime data to each neighborhood to be able to show in a map the neighborhood and to total number of crimes

4- Data Visualization

- Define Utility functions for visualization
  i. displayMap()
     a. Using folium.features.CircleMarker to show on the map
  ii. plotBarDF()
     1. Plots a bar chart for data in a data frame passed as argument
  iii. showBoundaries()
     1. Adds the Cambridge boundaries in a map using the geo data in geojson format
     2. Displays using the folium choropleth
  iv. showAreasOfInterest()
     1. Displays the areas of interest which are the Neighborhoods and highlighting them as circles on top of the Cambridge boundaries map
     2. Each neighborhood has a marker that is displayed when the circle is clicked
  v. showNeighborhoodCrimesBound()
     1. Displays the Cambridge boundaries and on top the that a circle which is proportional to the number of crimes in the neighborhood.
     2. When the circle is clicked shows the neighborhood name and the total number of crimes in that neighborhood
- 4-2 Show Visualizations
  i. Show the Cambridge city boundaries
     1. Using the function defined in previous section
  ii. Show the venues for each one of the areas (neighborhoods)
     1. The center of the area is in read and the venues in blue
  iii. Display a bar plot of the number of crimes by Neighborhood
  iv. Display a bar plot with the number of crimes by type of crime
  v. Display the a bar chart of employers and number of employees
  vi. Show the average salary by employee in a horizontal bar chart

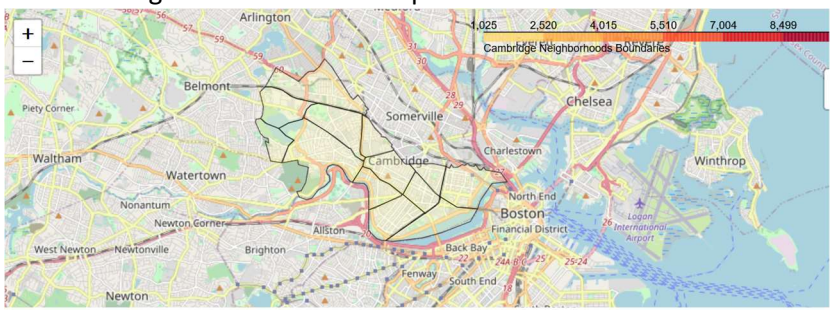
- # Results

The following are the results from the city of Cambridge, MA

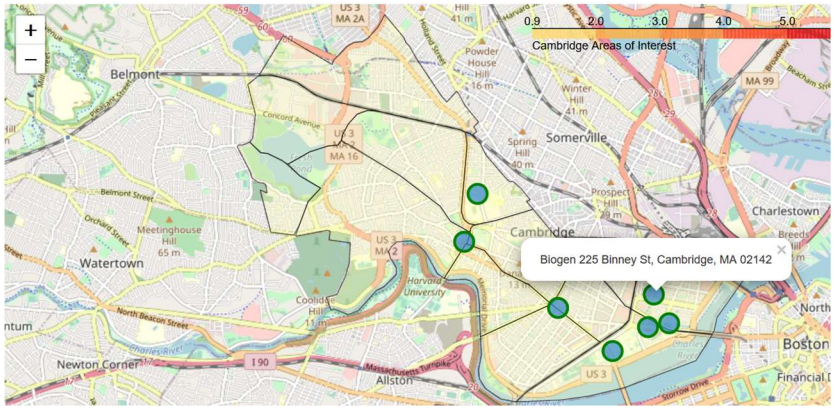
1- Report of most common venues by Neighborhood

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Biogen	Greek Restaurant	Garden	Skating Rink	French Restaurant	Event Space	Breakfast Spot	Café	Jewish Restaurant	Gym	Movie Theater
1	Cambridge Innovation Center	American Restaurant	Vegetarian / Vegan Restaurant	Bakery	Skating Rink	Garden	French Restaurant	Event Space	Café	Gym	Mediterranean Restaurant
2	Central Square	Cocktail Bar	Yoga Studio	Martial Arts Dojo	Bookstore	Comedy Club	Dance Studio	Vegetarian / Vegan Restaurant	American Restaurant	Supermarket	Record Shop
3	Harvard Square	Café	Arepa Restaurant	Bakery	Salad Place	Rock Club	Plaza	Pizza Place	Gastropub	Indie Movie Theater	Mexican Restaurant
4	Kendal Square	Coffee Shop	American Restaurant	Hotel Bar	Garden	Event Space	Gym	Vegetarian / Vegan Restaurant	Bookstore	Bar	Burrito Place
5	Lesley University	Ice Cream Shop	History Museum	Record Shop	College Technology Building	Italian Restaurant	Pub	American Restaurant	Science Museum	Sporting Goods Shop	Rock Club
6	MIT	College Gym	Pizza Place	Park	Burrito Place	Concert Hall	Café	Gym / Fitness Center	Bakery	Breakfast Spot	Arepa Restaurant

2- Display the Cambridge boundaries in a map

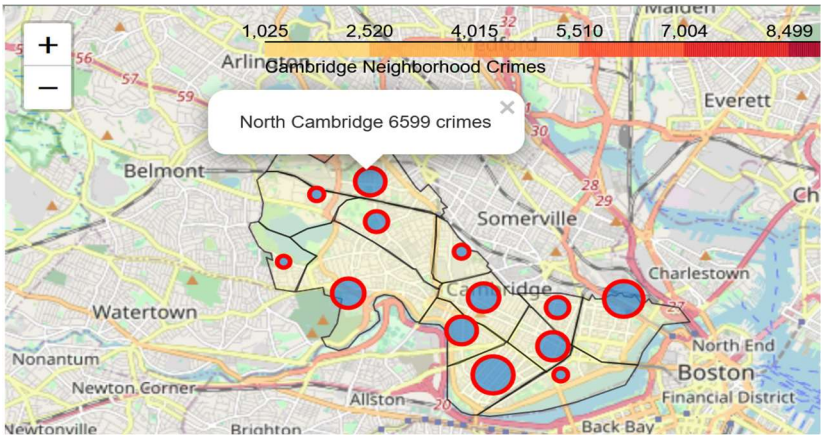
a. 

3- Display the Selected areas (Neighborhoods). A tip show the name and address

a. 

4- Display the number of crimes in each city Neighborhood

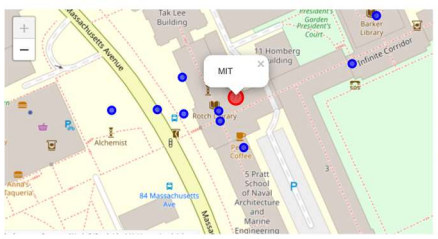
a. The size of the circle is proportional to the number of crimes in that neighborhood

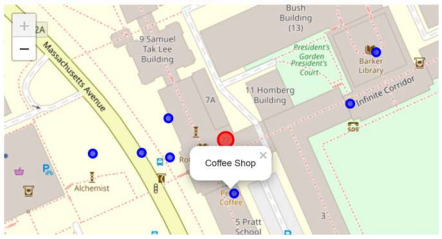
b. 
1,025
2,520
4,015
5,510
7,004
8,499
Cambridge Neighborhood Crimes
North Cambridge 6599 crimes
Belmont
Somerville
Everett
Charlestown
Watertown
Nonantum
Newton Corner
Allston
North End
Boston
Financial District
Back Bay

5- Display the venues for each Neighborhood (area)

a. MIT

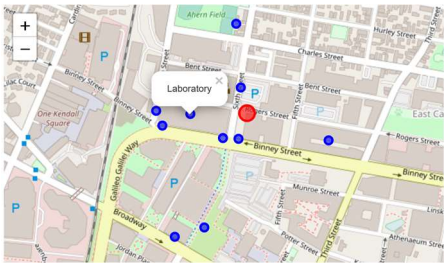
b.





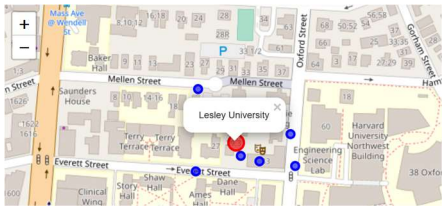
c. Biogen

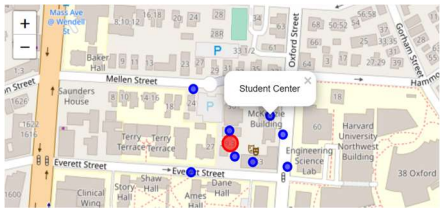
d.



e. Lesley University

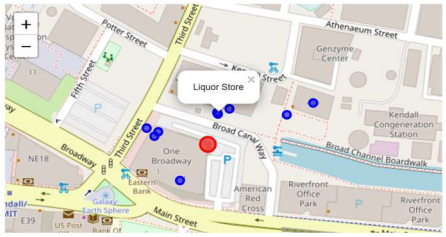
f.





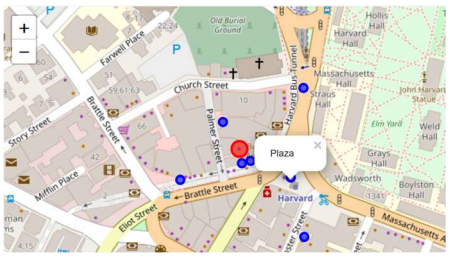
g. Cambridge Innovation Center

h.



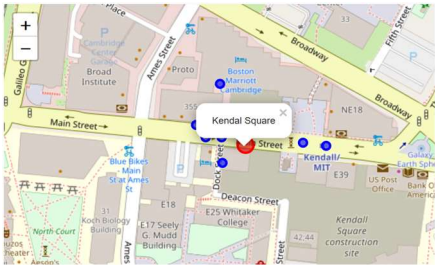
i. Harvard Square

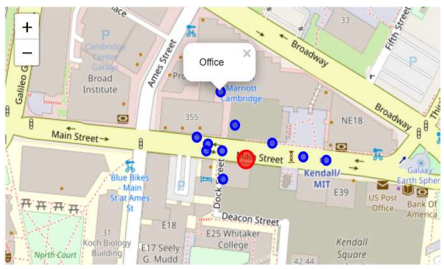
j.



k. Kendal Square

l.



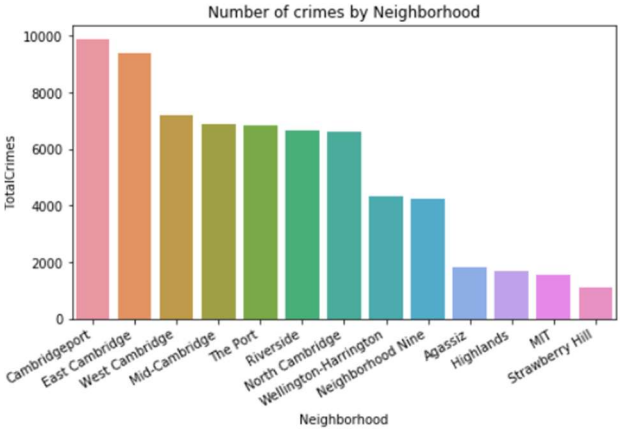


m. Central Square

n.

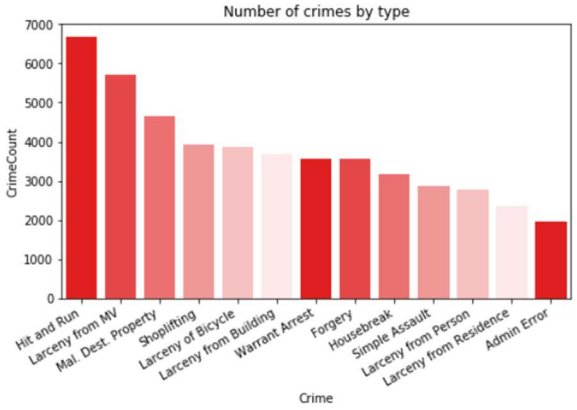
6- Display plot

a. Number of crimes by neighborhood



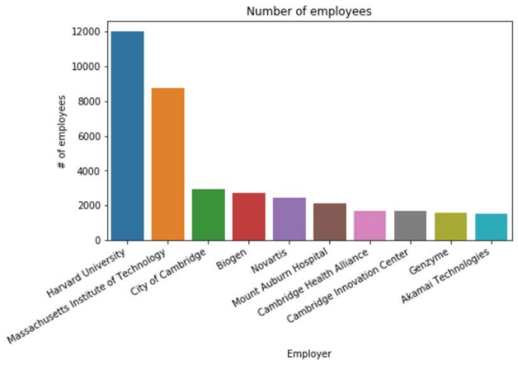
b.

c. Number of crimes by type



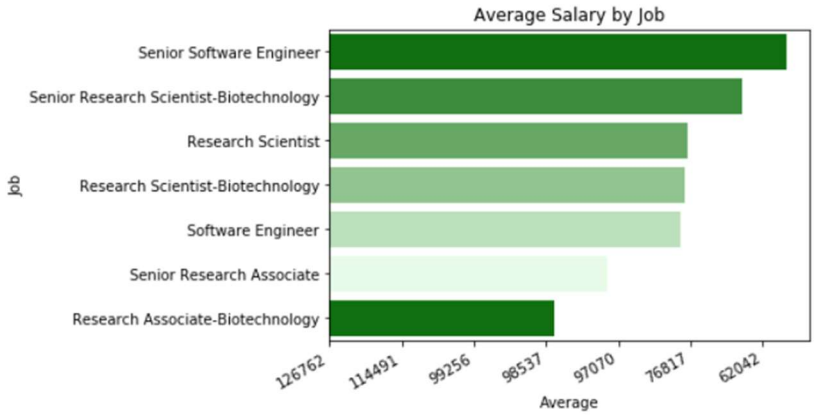
d.

e. Number of employees by employer



f.

g. Average Salary by job type



h.

- ## Discussions and observations

From the results we can observe the following

1- From the most common venues
   a. From the first most common
      i. 5 out of 7 neighborhoods chose a place to eat
      ii. Lesley University students (80% women) selected an ice-cream place
      iii. MIT students selected a Gym
   b. From the second most common
      i. 4 out of 7 selected a place to eat
      ii. Biogen selected a Garden
      iii. Lesley University selected History museum
      iv. Central Square a Yoga studio

   Looking at Lesley University (80% women) it seems to indicate that women have different preferences in most cases

   One could conclude that if one wanted to create a new business the area with the most possibility of success would be in a place to eat (restaurant, cafeteria, etc)

2- From the report on crimes
   a. The two areas less safe are:
      i. Cambridgeport with 9906 crimes
      ii. East Cambridge with 9415
   b. The most frequent crime
      i. Hit and Run with 6690 cases
         1. It could be because many people are from other countries and they might try to avoid any encounter with the law

The hit a run cases seems to be serious and I would recommend that the police set up more street cameras to be able to identify the perpetrators and capture them.

3- From the reports on Employers and Salaries
    a. Harvard with 11,997 employees and MIT with 8763 employees have a significant part of the employment in Cambridge
    b. Cambridge is one the centers of technology in the world and therefore has attracted many companies in software and biotechnology.
        i. The top paying jobs are
            1. Software Engineer average $126,762
            2. Senior Research Asistant-Biotechnology $ 114,491

Cambridge employment is highly oriented to technology and therefore if a student is not technologically oriented it probably Cambridge would not be the first place to look for a position

Conclusion

- This study concentrated on some of the Cambridge neighborhoods near the universities of Harvard, MIT and Lesley Universities and the results and discussions are shown in the previous sections
- A future study could be done considering all neighborhoods in Cambridge which would be of interest for the whole population and not just students
- Also a future study could study the impact of the transient population (many of the residents are students) during the summer