University California in analysis - The Battle of Neighborhoods

Author: Hua Zhu

Apr, 2019

Capstone Project: Review criteria

- This capstone project will be graded by your peers. This capstone project is worth **70%** of your total grade. The project will be completed over the course of **2 weeks**. Week 1 submissions will be worth **30%** whereas week 2 submissions will be worth **40% of your total grade**.

- For this week, you will required to submit the following:

- A description of the problem and a discussion of the background. (**15 marks**)

- A description of the data and how it will be used to solve the problem. (**15 marks)**

- For the second week, the final deliverables of the project will be:

- A link to your Notebook on your Github repository, showing your code. (**15 marks**)

- A full report consisting of all of the following components (**15 marks**):
    - Introduction where you discuss the business problem and who would be interested in this project.
    - Data where you describe the data that will be used to solve the problem and the source of the data.
    - Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.
    - Results section where you discuss the results.
    - Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
    - Conclusion section where you conclude the report.

- 3. Your choice of a presentation or blogpost. (**10 marks**)

Capstone Project - The Battle of Neighborhoods Requirements

- In this week, you will continue working on your capstone project. Please remember by the end of this week, you will need to submit the following:

- 1. A full report consisting of all of the following components (**15 marks**):
  - Introduction where you discuss the business problem and who would be interested in this project.
  - Data where you describe the data that will be used to solve the problem and the source of the data.
  - Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.
  - Results section where you discuss the results.
  - Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
  - Conclusion section where you conclude the report.

- 2. A link to your Notebook on your Github repository pushed showing your code. (**15 marks**)

- 3. Your choice of a presentation or blogpost. (**10 marks**)

- **Here are examples of previous outstanding submissions that should give you an idea of what your report would look like, what your notebook would look like in terms of clean, clear, and well-commented code, and what your presentation would look like or your blogpost would look like:**

- **Report: https://cocl.us/coursera_capstone_report**

- **Notebook: https://cocl.us/coursera_capstone_notebook**

- **Presentation: https://cocl.us/coursera_capstone_presentation**

- **Blogpost: https://cocl.us/coursera_capstone_blogpost**

Business Problem identification: University California deep analysis for international students

- Each year, many international students [in 2018, it is 28,556] apply their application into University California. There are 9 undergraduate campuses throughout California with 150 disciplines. More academic departments ranked in the top 10 nationally than any other public or private university

- For better decide which campus is most safe and convenient for their future study. Each students and their parents, they need to know:
  - Activities around the campus [Venues of choice]
  - Neighborhood crime-incident data

- This report try to analyze the 9 campus in UC system for undergraduate education, provide intended audience for above information
  - Applicants and their parents/Relatives
  - Every one who Interesting in UC system neighborhood

Data Source and usage

- https://www.universityofcalifornia.edu/infocenter ; https://datausa.io/search/?kind=university and https://nces.ed.gov/collegenavigator/ related campus website for basic information

- Folium map rendering for OpenStreetMap data

- FourSquare API for venue related data

- https://data.chhs.ca.gov/dataset/ for California crime-incident related data

- Python reverse address to geolocation packages for, geolocation determination based on address. http://www.city-data.com/

Data analysis methodology

- Python Anaconda Suite with SciKit-Learn, SciPy and Pandas for

  - Data Cleaning and Selection [read provided 9 campus, use geolocator to add latitude and longitude]
  - Applying Machine Learning algorithms and evaluation metrics [use Foursquare API to get nearby venues, to get top 5/10 venues, K-means to cluster 9 campuses into 3 cluster, and find similarity between campuses]
  - Statistical Correlations and visualization
  - For crime data, get data from https://data.chhs.ca.gov/dataset/, clean it and calculate the safety index for 9 campus, and visualized it.
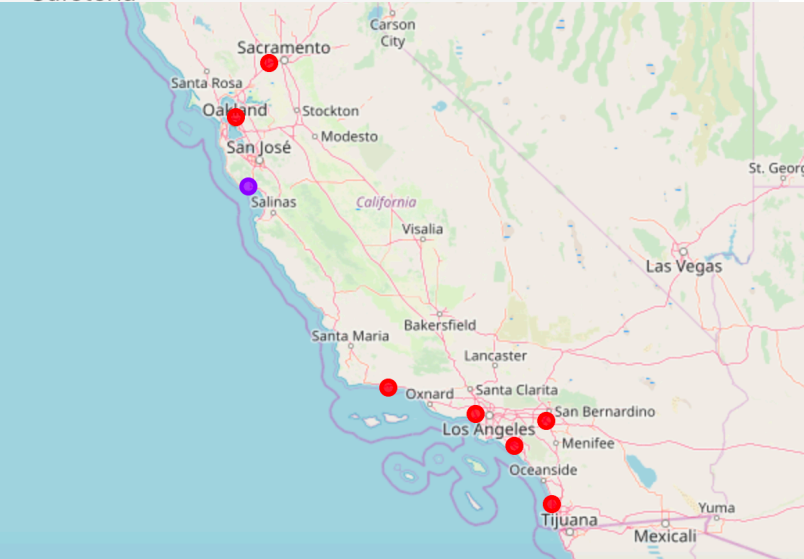
Result for venues around the 9 campuses

- The most popular venue around campus are restaurant, aka coffee or pizza shop.   UCSD is unique in convenience store and theater while UCSC is Gym.

```
----UC Berkeley----
             venue  freq
0             Café  0.15
1   Sandwich Place  0.07
2      Coffee Shop  0.07
3     College Quad  0.05
4      Pizza Place  0.05


----UC Davis----
               venue  freq
0        Pizza Place  0.11
1   Indian Restaurant  0.08
2  American Restaurant  0.05
3     Ice Cream Shop  0.05
4             Bakery  0.05


----UC Irvine----
                 venue  freq
0          Coffee Shop  0.18
1       Sandwich Place  0.12
2        Burrito Place  0.06
3           Juice Bar  0.06
4  Fast Food Restaurant  0.06


----UC LA----
                 venue  freq
0          Coffee Shop  0.12
1  Fast Food Restaurant  0.09
2  American Restaurant  0.06
3          Pizza Place  0.06
4       Medical Center  0.03
```

```
----UC Merced----
                     venue  freq
0       Fast Food Restaurant  0.33
1                    Market  0.33
2                     Plaza  0.33
3  Mediterranean Restaurant  0.00
4                      Park  0.00


----UC Riverside----
              venue  freq
0       Coffee Shop  0.14
1   College Library  0.07
2              Café  0.07
3       Pizza Place  0.07
4              Park  0.07


----UC San Diego----
              venue  freq
0  Convenience Store  0.12
1           Theater  0.12
2              Park  0.06
3        Food Truck  0.06
4       Coffee Shop  0.06


----UC Santa Barbara----
                venue  freq
0      Sandwich Place  0.11
1  American Restaurant  0.06
2  Chinese Restaurant  0.06
3           Juice Bar  0.06
4               Lake  0.06


----UC Santa Cruz----
                venue  freq
0        College Gym  0.2
1   Athletics & Sports  0.2
2               Café  0.2
3   Convenience Store  0.2
4  Fast Food Restaurant  0.2
```

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	UC Berkeley	Café	Sandwich Place	Coffee Shop	Pizza Place	College Quad	Indian Restaurant	Burrito Place	Hot Dog Joint	College Library	Clothing Store
1	UC Davis	Pizza Place	Indian Restaurant	American Restaurant	Bakery	Bar	Mexican Restaurant	Sandwich Place	Ice Cream Shop	Comic Shop	Irish Pub
2	UC Irvine	Coffee Shop	Sandwich Place	American Restaurant	Fast Food Restaurant	Juice Bar	College Quad	Café	Pub	Bus Station	Burrito Place
3	UC LA	Coffee Shop	Fast Food Restaurant	Pizza Place	American Restaurant	Noodle House	Salad Place	Medical Center	College Administrative Building	Museum	Chinese Restaurant
4	UC Merced	Plaza	Market	Fast Food Restaurant	College Auditorium	College Bookstore	College Cafeteria	College Gym	College Library	College Quad	College Theater
5	UC Riverside	Coffee Shop	Student Center	Park	Chinese Restaurant	Café	College Library	Sandwich Place	Burger Joint	Convenience Store	Pizza Place
6	UC San Diego	Theater	Convenience Store	Food Truck	Coffee Shop	New American Restaurant	Park	College Cafeteria	Public Art	Restaurant	Scenic Lookout
7	UC Santa Barbara	Sandwich Place	American Restaurant	Bus Station	Food Court	Hotel	Juice Bar	Lake	College Cafeteria	Mediterranean Restaurant	Mexican Restaurant
8	UC Santa Cruz	Convenience Store	Athletics & Sports	College Gym	Café	Fast Food Restaurant	Concert Hall	College Bookstore	College Cafeteria	College Library	College Quad

Result for safety index around the 9 campuses

- K-means 3 cluster layout

	Alias	Cluster Labels	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	UCB	0	Café	Coffee Shop	Sandwich Place	Pizza Place	College Library	Hot Dog Joint	College Quad	Clothing Store	New American Restaurant	Burger Joint
1	UCD	0	Pizza Place	Indian Restaurant	American Restaurant	Bakery	Bar	Mexican Restaurant	Sandwich Place	Ice Cream Shop	Grocery Store	Italian Restaurant
2	UCI	0	Coffee Shop	Sandwich Place	Fast Food Restaurant	Juice Bar	College Quad	College Auditorium	Chinese Restaurant	Pub	Burrito Place	American Restaurant
3	UCLA	0	Coffee Shop	Fast Food Restaurant	American Restaurant	Café	Pizza Place	Plaza	Garden	College Theater	Hotel	College Bookstore
5	UCR	0	Park	Burger Joint	Coffee Shop	Pizza Place	Food & Drink Shop	Chinese Restaurant	Student Center	Deli / Bodega	Café	College Library
6	UCSD	0	Theater	Convenience Store	Food Truck	Steakhouse	Nature Preserve	Park	College Cafeteria	Public Art	Restaurant	Scenic Lookout
7	UCSB	0	Burger Joint	Sandwich Place	American Restaurant	Restaurant	Food Court	Hotel				

	Alias	Cluster Labels	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 Common
8	UCSC	1	Botanical Garden	College Gym	Athletics & Sports	Mexican Restaurant	Taco Place	Café	F... Re...

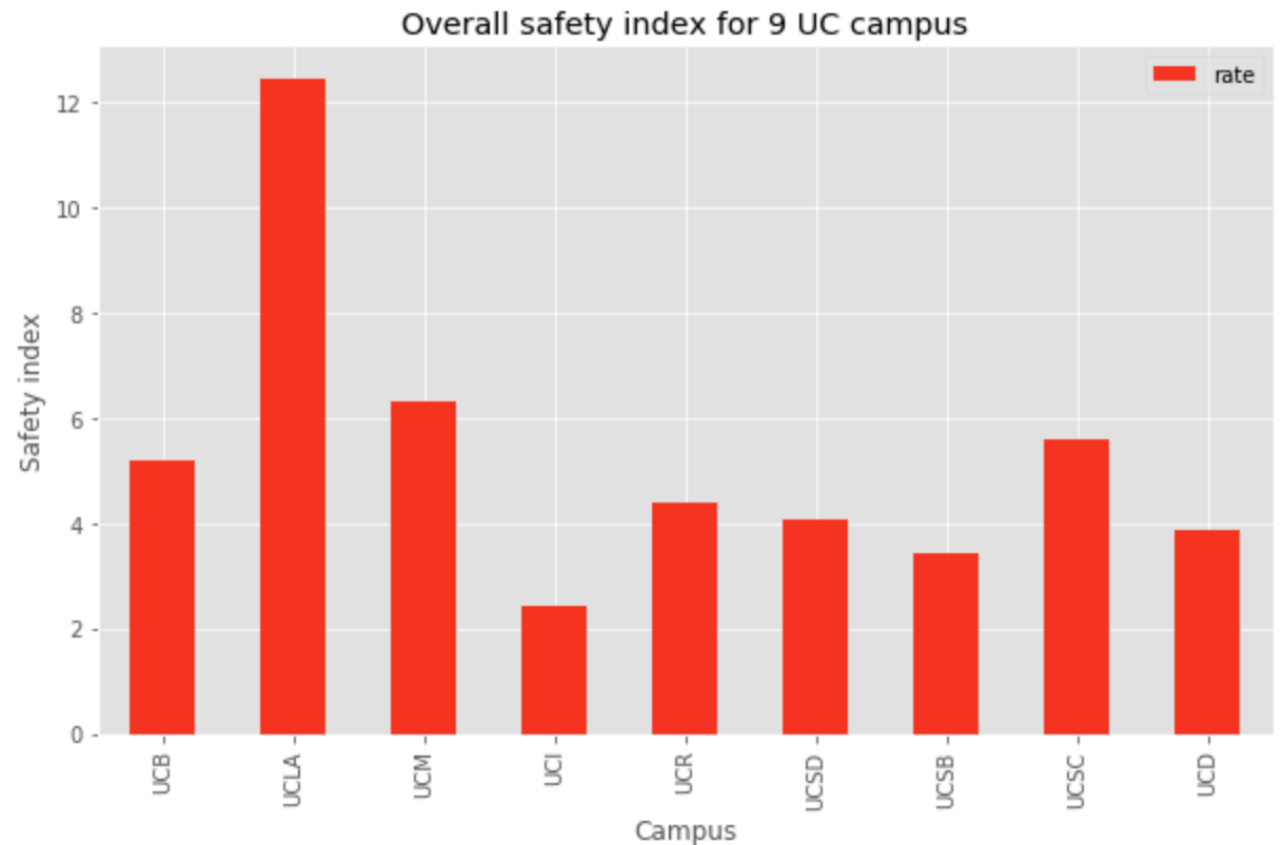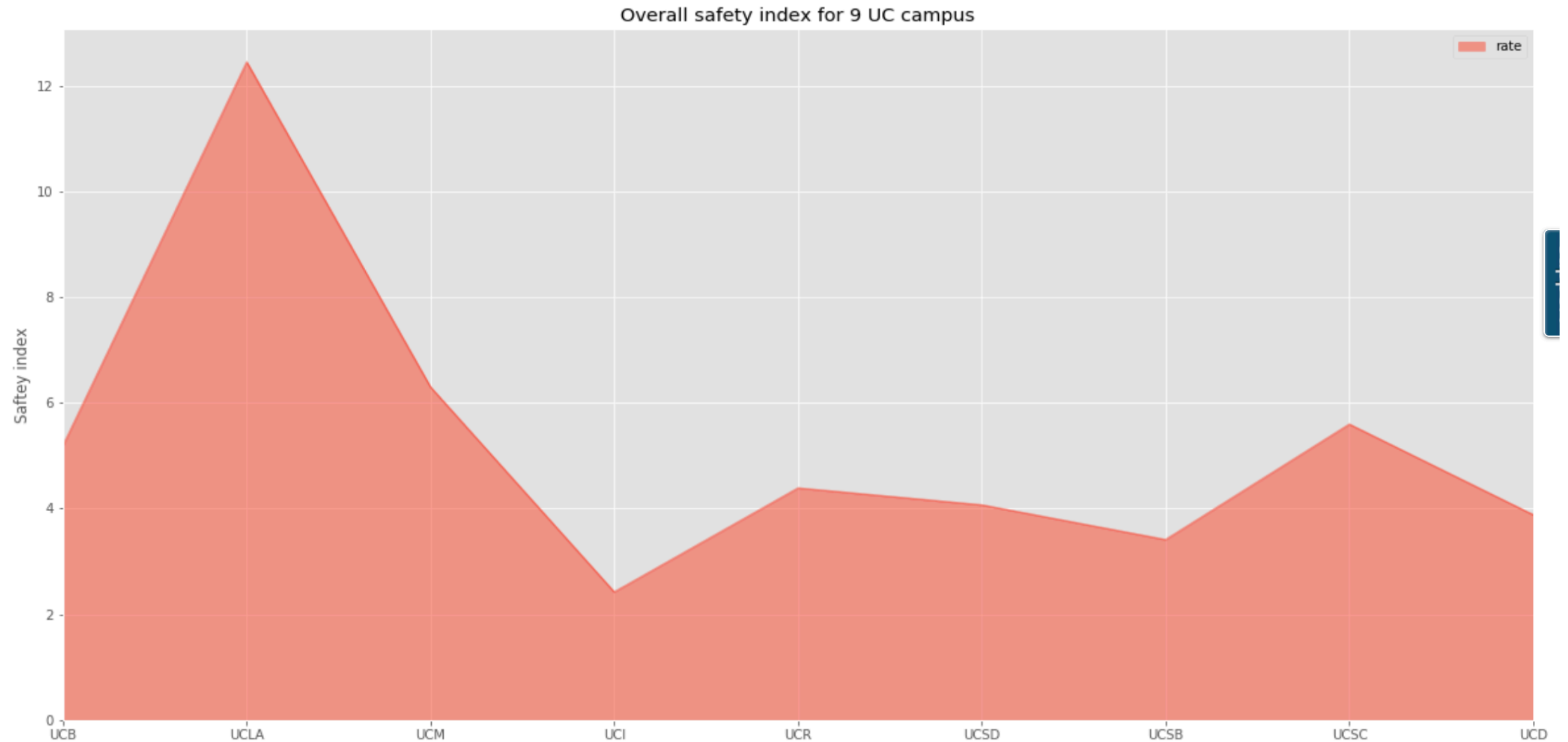|	Alias	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
4	UCM	2	Market	Plaza	Fast Food Restaurant	Wine Bar	College Bookstore	College Cafeteria

Result for safety index around the 9 campuses

The campuses safety ranking are:

- UC Irving/Orange; [2]
- UCSB [3]
- UCR & UCSD, UCD [4]
- UCB [5]
- UCSC, UCM [6]
- UCLA [12]

Overall safety index for 9 UC campus

Result for safety index around the 9 campuses

Overall safety index for 9 UC campus

Discussion

- Activities around campus
  - There are many people working/Living in campus, in return, different kind of food facility is top venues there.
  - UCSD is easy to access for convenient store and theater in top 2
  - UCSC Gym is the top venue
  - For Chinese international student, UCR and UCI is very friendly whilst UCSB is very closing
- Safety Index
  - The most safety campus is UCI, UC Irving/Orange; The second is UCSB [3] and UCR & UCSD, UCD [4] are very good as well
  - UCB[Berkeley] is not very good in frame in the past, but from 2000~2013 average, UCB is actually okay in the middle
  - There are some challenge in UCLA compare to others in safety condition

Conclusion

- Some information and conclusion in this report. But due to the nature of unsupervised learning, metrics can only guide, a best cluster determination algorithm is always arbitrary and should suit the project goal and intended audience,

- There are other factors no consider this time, such as campus itself like application admit rate, yield rate and cost per year. This could be put into future work steams,

- Crime type could be further group by subtype, DBSCAN could be very good start point. It worth a trial in future as well.