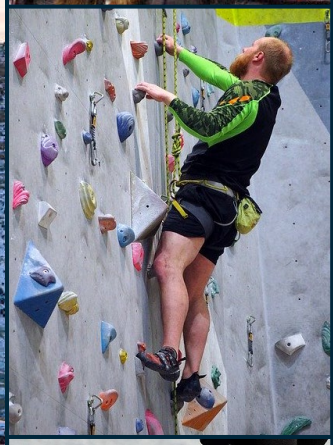
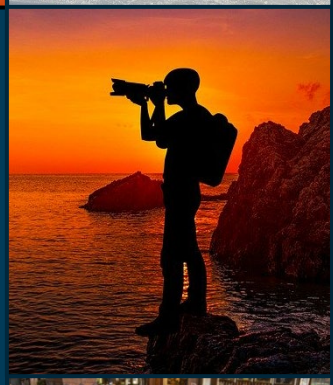
EXPAT ADVICE: IRELAND TO USA





APRIL 24

LC Data & Photography
Authored by Lenka Čížková

Expat Advice:

Relocating from Ireland to USA

IBM Data Science Capstone Project

"A journey of 1000 miles starts
with a single step."
Lao Tzu

Introduction

The Cleirigh family wants to relocate from Ireland to the USA. They wish to find the best place to live for all of them, together. Let's summarize their needs and wishes:

- **Mother Kayleigh** is a professional landscape and wildlife photographer specialized in private, all-inclusive luxury photography workshops. She would like to live somewhere close to a national park. The park should not be too crowded: maximally 10,000 visitors per year per km^2. Kayleigh is also interested in the national parks in general, she would like to know how big are the U.S. natural parks, where are they located and how many visitors they get, to plan future family holidays and/or bigger workshops.
- **Father Brian** is a passionate climber. He would like to live close to some nice rock-climbing area. Except of that, he wants to build a climbing gym specialized in youth training courses. Needless to say, it would be easier in a place where there are not yet many climbing gyms around.
- Their **son Sean** is a very bright student and a multi-talented athlete. Next year will be his final year of secondary school. He has skipped two grades and will graduate at 16. He wants to go to a top US university, preferably one with a higher ratio of international students. His excellent school grades and sport achievements will get him admission into the university.
- **Daughter Orla** attends primary school. She loves animals and figure skating and therefore, she wants to live close to some zoo exhibits and a skating rink. Orla has many friends at her school and consequently, she is less happy than the rest of the family when they discuss the planned relocation. She would like to know whether the US universities really are that good for Sean to want to attend one of them.
- **Grandmother Shauna** is a passionate cook and an art lover. She wants to start her own Irish pub and thus, she would prefer a place without an abundance of Irish pubs or restaurants. In her free time, she loves to visit art galleries.
- The family prefers to live not far away from the university. To get some idea of living in the USA, they would like to know what are the typical venues close by the universities.

Audience

Except of the Cleirigh family themselves, this study can be interesting for any of the following groups:

- students who want to attend one of the top universities in the USA;
- landscape and wildlife photographers, both professional and amateur;
- tour operators as well as individual tourists and rock climbers interested in visiting national parks;
- art lovers;
- anybody who is interested in a sophisticated approach to finding a suitable place to live.

Data

In the Introduction, we have seen that the needs and wishes of the Cleirighs are very diverse. Next, we will present the data that can be used to identify the most suitable living place for the Cleirighs.

Top universities in the USA

To identify the top universities in the USA, we will use one of the well-known university rankings, the *Times Higher Education World University Rankings 2020* [1], available via https://www.timeshighereducation.com/world-university-rankings/2020/world-ranking.
The ranking includes 1397 universities and consists of the following 13 indicators: **Rank** of the university, its **Name**, **Country/Region**, **No. of FTE students**, **No. of students per staff**, [percentage of the] **International students**, **Female:Male Ratio**, as well as the **Overall**, **Teaching**, **Research**, **Citation**, **Industry Income** and **International Outlook** scores.
The geographical coordinates of the universities can be retrieved via the ArcGIS API using the Python's geocoder library.

National parks in the USA

All the necessary information about the national parks can be found in the Wikipedia's *List of national parks of the United States* [2], available via https://en.wikipedia.org/wiki/List_of_national_parks_of_the_United_States.
For each of the 62 national parks, the data consists of the **Name** of the national park, an **Image** (a photograph showing a landscape typical for the park), **Location** including the state and the geographical coordinates of the park, the **Date** [when the park was] **established as** [a national] **park**, the park's **Area (2019)** in acres as well as km^2, the number of **Recreational visitors (2018)** per year and a **Description**.
The geographical coordinates of the parks will be combined with the geographical coordinates of the universities to calculate their mutual distance.

Foursquare

The location data platform Foursquare [3] will be used as a valuable source of information on the remaining aspects.
First of all, we use the venues in the surrounding of each university to look for a possible clustering of the universities (does a 'typical university surrounding' exist?) and to identify possible anomalies.
Further, for each university, we explore its region to find information about rock-climbing areas, climbing halls, zoo exhibits, skating rinks, Irish restaurants and, last but not least, museums and art galleries.
Finally, all the above-mentioned information will be combined to find the most suitable place for the Cleirighs to relocate to.

Data Cleaning, Pre-Processing

National parks

The data on national parks were scraped from Wikipedia. There is no missing data in the dataset. It consists mainly of items combining a raw text with numeric data (see Table 1) and thus, it needs to be pre-processed to extract numerical values of the respective quantitative variables.

As Kayleigh wants 'her' national park not to be too crowded, we need to create a feature describing the 'crowdedness' of the parks. A suitable indicator is *visitor density*, i.e., the number of visitors per year per square kilometer, calculated as a ratio of the park's area and the respective number of visitors per year.

Further, for Kayleigh to be able to spend as much time as possible with the family, and at the same time, to run the long workshops in the national park, the national park should be relatively close to the city. Kayleigh has decided that 150 km is the maximum distance she is ready to travel regularly between the city and the national park. To calculate the distances, we have to extract the geographical coordinates of the parks (and later, of the universities).

Thus, for every park, we need to get its name, geographical coordinates, area and the number of visitors. *Location* was split on '/' and from the last part, we get the latitude and longitude. 'S' in the latitude or 'W' in the longitude means the result needs to be a negative number. From *Area*, we get the part between '(' and 'km2'. Finally, the visitor density (feature named 'Crowded') was calculated as explained above. A part of the resulting data is shown in Table 2 - National parks: Pre-processed data.

	Name	Image	Location	Date established as park[5][10]	Area (2019)[11]	Recreation visitors (2018)[8]	Description
0	Acadia	NaN	Maine44°21′N 68°13′W / 44.35°N 68.21°W	February 26, 1919	49,076.63 acres (198.6 km2)	3537575	Covering most of Mount Desert Island and other...
1	American Samoa	NaN	American Samoa14°15′S 170°41′W / 14.25°S 170...	October 31, 1988	8,256.67 acres (33.4 km2)	28626	The southernmost national park is on three Sam...
2	Arches	NaN	Utah38°41′N 109°34′W / 38.68°N 109.57°W	November 12, 1971	76,678.98 acres (310.3 km2)	1663557	This site features more than 2,000 natural san...

Table 1 - National parks: Raw data

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded
0	Acadia	198.6	3537575	44.35	-68.21	Maine	17812.56
1	American Samoa	33.4	28626	-14.25	-170.68	American Samoa	857.07
2	Arches	310.3	1663557	38.68	-109.57	Utah	5361.12

Table 2 - National parks: Pre-processed data

Universities

The data on the top universities were scraped from the *Times Higher Education World University Rankings 2020* [1]. The content of the website is dynamically generated. The viewer can set, among others, the region they are interested in, and the number of universities to be shown per page; adding `/length/-1` to URL makes it possible to get the data on all the universities in one go. The table consists of two parts: Ranking and Scores, which need to be scraped separately and combined subsequently. There are several universities with missing data in the columns *International Students* and *Female:Male Ratio*; in this case, missing data does not hinder any analyses but we at least have to take care to code it properly.

	Rank	Name	Country/Region	No. of FTE students	No. of students per staff	International Students	Female:Male Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
0	1	University of Oxford	United Kingdom	20,664	11.2	41%	46 : 54	95.4	90.5	99.6	98.4	65.5	96.4
1	2	California Institute of Technology	United States	2,240	6.4	30%	34 : 66	94.5	92.1	97.2	97.9	88.0	82.5
2	3	University of Cambridge	United Kingdom	18,978	10.9	37%	47 : 53	94.4	91.4	98.7	95.8	59.3	95.0
3	4	Stanford University	United States	16,135	7.3	23%	43 : 57	94.3	92.8	96.4	99.9	66.2	79.5
4	5	Massachusetts Institute of Technology	United States	11,247	8.6	34%	39 : 61	93.6	90.5	92.4	99.5	86.9	89.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1392	1001+	Yuan Ze University	Taiwan	8,356	19.5	8%	42 : 58	10.7–22.1	17.3	13.9	15.5	47.0	28.3
1393	1001+	Zagazig University	Egypt	156,419	24.0	1%	53 : 47	10.7–22.1	13.6	7.7	29.6	34.4	38.8
1394	1001+	University of Zagreb	Croatia	68,216	18.9	3%	59 : 41	10.7–22.1	17.8	12.9	25.3	37.4	33.0
1395	1001+	University of Zanjan	Iran	9,980	25.1	0%	54 : 46	10.7–22.1	17.0	12.3	28.5	43.8	18.7
1396	1001+	Zhejiang University of Technology	China	31,228	14.7	8%	n/a	10.7–22.1	16.7	14.2	32.3	51.4	24.4

Table 3 - Universities: Raw data

Similarly to the data on national parks, the resulting data frame contains raw data (see Table 3) that needs to be pre-processed to make analyses possible:

- remove the comma in the 'No. of FTE students'.
- extract the ratio of international students, taking into account that this ratio is missing for some universities, shown just as '%' (contrary to, e.g., '14%'): this missing data will be coded as NaN.
- extract/calculate the ratio of female students, taking into account any missing data; again, this missing data will be coded as NaN.
- retype strings to int (or float) where relevant; this is important for the descriptive statistics that we will use later.

- extracting the values of 'Overall' score. In some cases, the 'Overall' score is not given as a single number (e.g., 50.2) but as a range (e.g., 50.1–53.7). In such a case, we assign the mean value of the range (e.g., (50.1 + 53.7)/2).

Table 4 shows the pre-processed data.

	Rank	Name	Country/Region	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
0	1	University of Oxford	United Kingdom	20664	11.2	0.41	0.46	95.4	90.5	99.6	98.4	65.5	96.4
1	2	California Institute of Technology	United States	2240	6.4	0.30	0.34	94.5	92.1	97.2	97.9	88.0	82.5
2	3	University of Cambridge	United Kingdom	18978	10.9	0.37	0.47	94.4	91.4	98.7	95.8	59.3	95.0
3	4	Stanford University	United States	16135	7.3	0.23	0.43	94.3	92.8	96.4	99.9	66.2	79.5
4	5	Massachusetts Institute of Technology	United States	11247	8.6	0.34	0.39	93.6	90.5	92.4	99.5	86.9	89.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1392	1001+	Yuan Ze University	Taiwan	8356	19.5	0.08	0.42	16.4	17.3	13.9	15.5	47.0	28.3
1393	1001+	Zagazig University	Egypt	156419	24.0	0.01	0.53	16.4	13.6	7.7	29.6	34.4	38.8
1394	1001+	University of Zagreb	Croatia	68216	18.9	0.03	0.59	16.4	17.8	12.9	25.3	37.4	33.0
1395	1001+	University of Zanjan	Iran	9980	25.1	0.00	0.54	16.4	17.0	12.3	28.5	43.8	18.7
1396	1001+	Zhejiang University of Technology	China	31228	14.7	0.08	NaN	16.4	16.7	14.2	32.3	51.4	24.4

Table 4 - Universities: Pre-processed data

Foursquare: Venues in the neighbourhood of the universities

First, we have to find the geographical coordinates of every university. To do so, I have tried several geocoding services and based on their stability, response time and unlimited number of calls, I have decided to use the ArcGIS API for Python.

Using Foursquare API regular calls `explore`, we explore the neighbourhood of each university (only considering the Top500 US universities), within the radius of 1500m from the previously retrieved geographical coordinates of the respective university, and make a list of all venues found nearby. For each venue, we keep record of the name, latitude and longitude of the respective university; name, latitude and longitude of the venue itself and the venue category as listed by Foursquare.

In total, 9411 venues were found in the neighbourhoods of the 121 universities. Inspecting the results reveals that there are no venues close to New Mexico Institute of Mining and Technology. That seems to be quite rare because we would at least expect some restaurants, bars or shops in the neighbourhood of any university. Indeed, for most universities, 100 venues were found (which is the limit imposed by Foursquare), see Table 5.

Next, we have checked whether there are any universities with less than 20 venues: such an anomaly could suggest that the retrieved geographical location of the university is not right. Indeed, there are

9 universities with less than 20 venues (see Table 6) and New Mexico Institute of Mining and Technology with no venue at all.
Manually checking via the Google Maps, we have found several errors in the geographical coordinates of these universities, retrieved via the ArcGIS geocoder. The following list gives the

Name	Venue
American University	100
Arizona State University (Tempe)	100
Boston College	33
Boston University	100
Brandeis University	81
...	...
Washington State University	100
Washington University in St Louis	100
Wayne State University	88
William & Mary	100
Yale University	100

Table 5 - Foursquare: Number of venues within 1500m from the university

	Name	Venue
0	Ohio State University (Main campus)	3
1	Penn State (Main campus)	4
2	University of California, Davis	18
3	University of California, Merced	5
4	University of California, Santa Cruz	4
5	University of Florida	18
6	University of Houston	14
7	University of Missouri-Columbia	2
8	University of Nebraska-Lincoln	14

Table 6 - Foursquare: Universities having less than 20 venues

coordinates of the respective universities as shown by the Google Maps as well as suggested additions or changes to the *Name* of the university in order to get the proper geographical coordinates via new ArcGIS API calls.

- New Mexico Institute of Mining and Technology => add 'New Mexico' or 'Socorro'; 34.0639857, -106.8868248
- Ohio State University (Main campus) => change to 'Ohio State University, Columbus'; 40.006918, -83.030418
- Penn State => change to 'Penn State University'; 40.7983317, -77.8609841
- University of Florida => add 'Gainesville'; 29.6436099, -82.3550447
- University of Houston => add 'Houston, Texas'; 29.719758, -95.342029
- University of Missouri-Columbia => only use 'University of Missouri', or change to 'University of Missouri, Columbia, Missouri'; 38.9397238, -92.3274538
- University of Nebraska-Lincoln => add 'Lincoln'; 40.820151, -96.700544

The other three universities (University of California, Davis; University of California, Merced; University of California, Santa Cruz) seem to be located correctly, having a less busy neighbourhood.

Remark: The coordinates of the University of Iowa and of the University of Washington were also wrong when tried for the first time but they are all right now, probably due to an update in the ArcGIS geocoder.

	Name	Name-edit	Latitude	Longitude
0	New Mexico Institute of Mining and Technology	New Mexico Institute of Mining and Technology,...	34.06014	-106.89187
1	Ohio State University (Main campus)	Ohio State University, Columbus	40.00630	-83.01638
2	Penn State (Main campus)	Penn State University	40.80707	-77.85888
3	University of Florida	University of Florida, Gainesville	29.64973	-82.34113
4	University of Houston	University of Houston, Calhoun Rd, Houston	29.72224	-95.33782
5	University of Missouri-Columbia	University of Missouri, Columbia, Missouri	38.94441	-92.32930
6	University of Nebraska-Lincoln	University of Nebraska-Lincoln, Lincoln	40.81690	-96.70010

Table 7 - Foursquare: Changed names of the universities and the corrected coordinates

Table 7 shows the edited names of the universities as well as the newly retrieved geographical coordinates. Using the corrected geographical coordinates, substantially more venues are found (see Table 8 and Table 9). After replacing the wrong venues by the newly retrieved ones, we have 9849 venues in total, with 438 unique categories.

	Venue
Name	
New Mexico Institute of Mining and Technology, Socorro	23
Ohio State University, Columbus	93
Penn State University	94
University of Florida, Gainesville	52
University of Houston, Calhoun Rd, Houston	58
University of Missouri, Columbia, Missouri	73
University of Nebraska-Lincoln, Lincoln	100

Table 8 - Foursquare: Number of venues when using the corrected coordinates

Foursquare: Family-wishlist venues

Next, we explore the regions surrounding each of the 121 universities that have a national park within 150 km distance (see section Family List) and check whether the wishes of all the family members would be fulfilled. Let's recall:

- Father Brian is a passionate climber. He would like to live close to some nice rock-climbing area. Except of that, he wants to build a climbing gym specialized in youth training courses. Needless to say, it would be easier in a place where there are not yet many climbing gyms around.

	Name	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New Mexico Institute of Mining and Technology,...	34.06014	-106.89187	M Mountain Coffeehouse	34.057322	-106.892443	Coffee Shop
1	New Mexico Institute of Mining and Technology,...	34.06014	-106.89187	Bodega Lounge	34.063017	-106.891849	Sports Bar
2	New Mexico Institute of Mining and Technology,...	34.06014	-106.89187	The Capitol Bar	34.057045	-106.892619	Bar
3	New Mexico Institute of Mining and Technology,...	34.06014	-106.89187	Socorro Springs Brewing Company	34.070112	-106.892552	Brewery
4	New Mexico Institute of Mining and Technology,...	34.06014	-106.89187	Domino's Pizza	34.068281	-106.893383	Pizza Place
...	...	...	...	...	...	...	...
488	University of Nebraska-Lincoln, Lincoln	40.81690	-96.70010	Saltdogs Baseball	40.823220	-96.714153	Baseball Stadium
489	University of Nebraska-Lincoln, Lincoln	40.81690	-96.70010	Hiro 88	40.816543	-96.712091	Sushi Restaurant
490	University of Nebraska-Lincoln, Lincoln	40.81690	-96.70010	Rocket Fizz	40.816880	-96.712283	Candy Store
491	University of Nebraska-Lincoln, Lincoln	40.81690	-96.70010	10 Below	40.815602	-96.709000	Bar
492	University of Nebraska-Lincoln, Lincoln	40.81690	-96.70010	Destinations Coffee House	40.826277	-96.701556	Coffee Shop

Table 9 - Foursquare: Newly retrieved venues

- Daughter Orla attends primary school. She loves animals and skating and therefore, she wants to live close to some zoo exhibits and a skating place.
- Grandmother Shauna is a passionate cook and an art lover. She wants to start her own Irish pub and thus, she would prefer a place without an abundance of Irish pubs or restaurants. In her free time, she loves to visit art galleries.

Using Foursquare API regular calls `search` [4], we explore the region of each university to find venues relevant to all the family members as described above. `radius` of the region depends on the activity and thus, on which category we search for. While a *Skating Rink* and a *Zoo* or *Zoo Exhibit* for Orla as well as an *Art Gallery* for Shauna should be as close as possible to allow for spontaneous relatively short visits in the moment they feel for it (up to 5 km; or up to 10 km for a real *Zoo* – as compared to a smaller *Zoo Exhibit*), a *Rock Climbing Spot* for Brian relates to a planned activity taking several hours anyway and thus, it may be a bit more distant (up to 30 km). Brian wants to build his own *Climbing Gym* and Shauna wants to open her *Irish Pub*, both of them preferably in a place where there are no venues of the same category within the radius of 5 km. When calling Foursquare API, except of `radius` we need to specify `intent`. The default value of `intent` is `checkin`, only returning venues where the user could check at the provided location at the current moment in time, which is not relevant to our business case: maybe the venue is closed just this day of week; we also do not want to have to take into account the local time when running the query. Hence, instead of `checkin`, we need to use `browse` which searches an entire region and also returns the venues that are not open at the moment.

These are the relevant venue category ID's [5]:

- Climbing Gym: 503289d391d4c4b30a586d6a
- Rock Climbing Spot: 50328a4b91d4c4b30a586d6b
- Irish Pub: 52e81612bcbc57f1066b7a06
- Art Gallery: 4bf58dd8d48988d1e2931735
- Zoo: 4bf58dd8d48988d17b941735 (incl. sub-category Zoo Exhibit)
- Skating Rink: 4bf58dd8d48988d168941735 (both ice and non-ice)

In total, there are 2044 retrieved venues. Some of the universities do not have any venues of a particular type within the radius as given above: 11 universities do not have any *Climbing Gym* around; 2 universities, no *Rock Climbing Spot*; 16 universities, no *Irish Pub*; 3 universities, no *Zoo* and 8 universities have no *Zoo Exhibit* within the given radius; finally, there is no *Skating Rink* in the region of 6 universities. Table 10 shows examples of the retrieved venues.

	Name	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Search Category
0	Stanford University	37.42907	-122.16978	Whiting Family Rock Climbing Wall	37.431310	-122.159202	Climbing Gym	Climbing Gym
1	Stanford University	37.42907	-122.16978	Stanford Climbing Wall	37.426398	-122.177369	Climbing Gym	Climbing Gym
2	University of California, Los Angeles	33.92775	-118.37275	Sender One LAX	33.932492	-118.371679	Rock Climbing Spot	Climbing Gym
3	University of California, Los Angeles	33.92775	-118.37275	Hangar 18 Indoor Climbing Gym	33.900628	-118.364514	Climbing Gym	Climbing Gym
4	University of Washington	47.65651	-122.31209	The Mountaineers South Climbing Plaza	47.684769	-122.263698	Climbing Gym	Climbing Gym
...	...	...	...	...	...	...	...	...
50	University of Denver	39.67850	-104.96662	Joy Burns Ice Arena	39.682934	-104.961353	Skating Rink	Skating Rink
51	American University	38.93844	-77.08666	Washington Harbour Ice Rink	38.901434	-77.059980	Skating Rink	Skating Rink
52	Colorado State University, Fort Collins	40.57807	-105.08155	Eagles country	40.592527	-105.032171	Skating Rink	Skating Rink
53	Colorado State University, Fort Collins	40.57807	-105.08155	Old Town Ice Rink	40.587915	-105.075655	Skating Rink	Skating Rink
54	Colorado State University, Fort Collins	40.57807	-105.08155	Edora Pool and Ice Center (EPIC)	40.563206	-105.045343	Gym Pool	Skating Rink

Table 10 - Foursquare: Family-wishlist venues

Some retrieved venues have multiple venue categories and show a (primary) category that seems to be irrelevant, such as *Abc Kids Climbing* in Boulder, Colorado, retrieved as *Climbing Gym* but showing the primary category *Daycare*. Checking online shows that it really is a climbing gym. On the other hand, some venues are misclassified. For example, *Maggie Daley Climbing Walls*, categorized as *Rock Climbing Spot*, is an artificial outdoor climbing wall, no rock climbing at all. Similarly, *Sender One LAX* in Los Angeles is categorized as both *Rock Climbing Spot* and *Climbing Gym* (see Table 10) while it is a climbing gym, no rock climbing. Within *Search Category Zoo*, we have several venues with *Venue Category* equal to, for example, *Science Museum*, *Trail* or *Gift Shop*. These do not qualify as a *Zoo / Zoo Exhibit*.

Inspecting the retrieved results has shown that Foursquare only returns one venue category even for the venues that are included in multiple categories. For example, *Sender One LAX*, mentioned above, shows in the results only the category *Rock Climbing Spot*, independent on which category was used even when search query includes both categories. Therefore, we cannot filter the venues automatically based on multiple category names; comparing results of multiple queries (in both *Climbing Gym* and *Rock Climbing Spot*) also does not help to solve this inexactness. These examples show that some venues need to be verified, especially when there are just a few of them in a given category. The inexactness of the venue categories (e.g., *Daycare*, *Building*, *Recreation Center*) as mentioned above is the reason why, while retrieving each venue, we have added *Search Category* to the retrieved information, to make it possible to sort all venues according to the categories we are interested in.

When cleaning the data, we have to take into account the following:

- There are several art galleries close to every university. Therefore, we do not need to care about art galleries anymore and we can drop them from the list to reduce the amount of data.
- For venues with *Search Category* equal to *Rock Climbing Spot* and the *Venue Category* equal to *Climbing Gym, College Gym, Gym* or *Athletic & Sports*: change the *Search Category* to *Climbing Gym*; drop duplicates.
- Drop all venues with *Search Category* equal to *Rock Climbing Spot* that have another *Venue Category*, except of *Other Great Outdoors*. The rest can be checked manually at a later phase, if needed.
- Drop all venues with *Search Category* equal to *Irish Pub* that have *Venue Category* equal to *American Restaurant*.
- Zoo: We are interested in Zoo's at max. 10 km, or Zoo Exhibits at max. 5 km.
    - Drop all venues with *Search Category* equal to *Zoo* that have *Venue Category* other than *Zoo*.
    - Drop all venues with *Search Category* equal to *Zoo Exhibit* (whereby we have searched, in fact, for a *Zoo* again) that have *Venue Category* other than *Zoo*, *Zoo Exhibit* or *Farm.*

After cleaning the data, there are 590 venues left. Table 11 shows the numbers of relevant family-wishlist venues (per category, per university).

Name	Climbing Gym	Irish Pub	Rock Climbing Spot	Skating Rink	Zoo	Zoo Exhibit
American University	2	4	4	1	4	50
California Institute of Technology	2	0	7	2	0	1
Colorado School of Mines	2	0	8	0	0	0
Colorado State University, Fort Collins	0	1	3	3	0	2
George Mason University	1	0	1	1	0	0
George Washington University	2	11	4	5	3	49
Georgetown University	2	9	4	4	3	49
Howard University	1	9	4	3	4	49
Nova Southeastern University	2	0	9	0	0	1
Stanford University	3	0	7	1	0	1
The University of Tennessee-Knoxville	1	0	3	1	1	2
University of Arizona	0	0	1	1	1	3
University of California, Los Angeles	4	0	10	4	1	2
University of California, Merced	0	0	1	0	1	0
University of California, Riverside	1	0	5	1	1	0
University of California, Santa Barbara	0	0	8	3	1	1
University of California, Santa Cruz	1	0	2	0	0	1
University of Colorado Boulder	3	0	13	8	0	2
University of Denver	1	2	7	1	3	0
University of Maryland, College Park	2	0	5	1	0	0
University of Miami	3	0	6	0	0	0
University of Oregon	2	1	0	1	1	2
University of South Carolina-Columbia	0	0	2	0	2	11
University of Southern California	3	8	12	5	0	1
University of Virginia (Main campus)	1	1	0	1	0	0
University of Washington	4	5	6	3	3	48
Vanderbilt University	1	0	3	5	1	0

Table 11 - Foursquare: Numbers of retrieved family-wishlist venues, per category, per university

Exploratory Data Analysis

Exploratory data analysis helps to understand the data and to identify anomalies.

National parks

Table 12 shows the descriptive statistics (mean, standard variation, minimum, first quartile, median, third quartile and maximum) of *area*, *number of visitors per year* and *visitor density* of the 62 national parks. We can see that there are big differences between the parks. Obviously, there is an extreme outlier, as shown in the descriptive statistics as well as the histogram (Figure 1) and the boxplot of the visitor density (Figure 2).

	Area in km2	Visitors per year	Crowded
count	62.000000	6.200000e+01	6.200000e+01
mean	3420.788710	1.394863e+06	4.433794e+04
std	6646.028274	1.897715e+06	3.197370e+05
min	0.800000	9.591000e+03	3.100000e-01
25%	229.400000	3.121205e+05	3.654875e+02
50%	889.500000	6.506595e+05	9.950700e+02
75%	2989.525000	1.674884e+06	2.554323e+03
max	33682.600000	1.142120e+07	2.520225e+06

Table 12 - National parks: Descriptive statistics

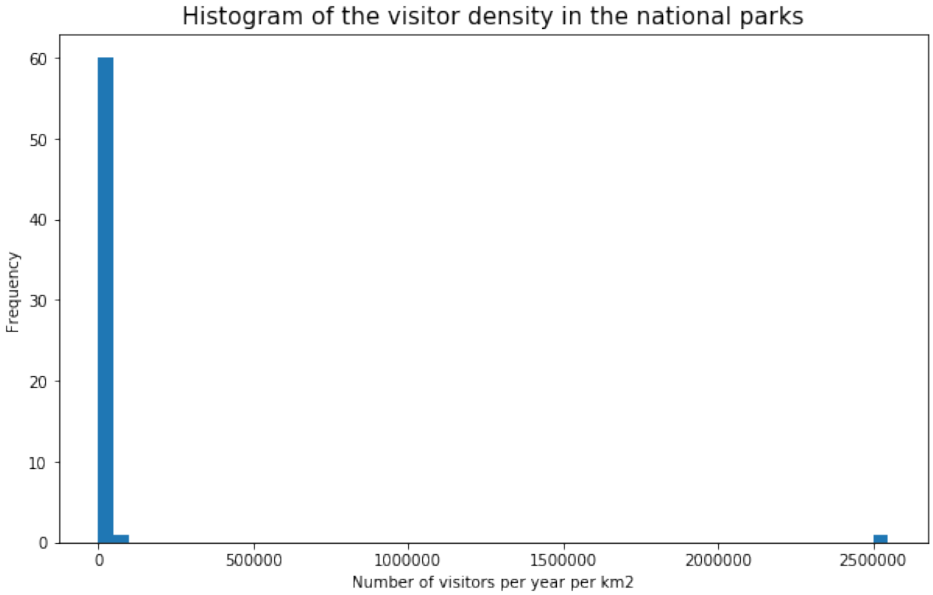


Figure 1 - National parks: Histogram of the visitor density

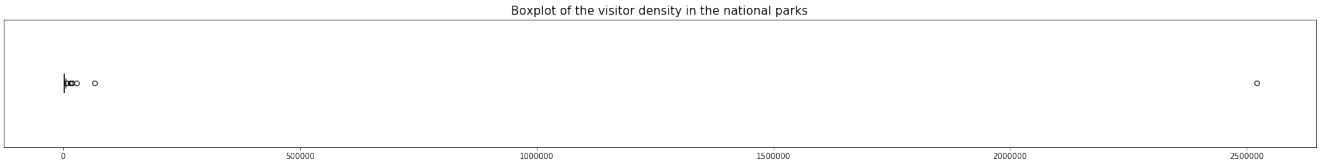


Figure 2 - National parks: Boxplot of the visitor density

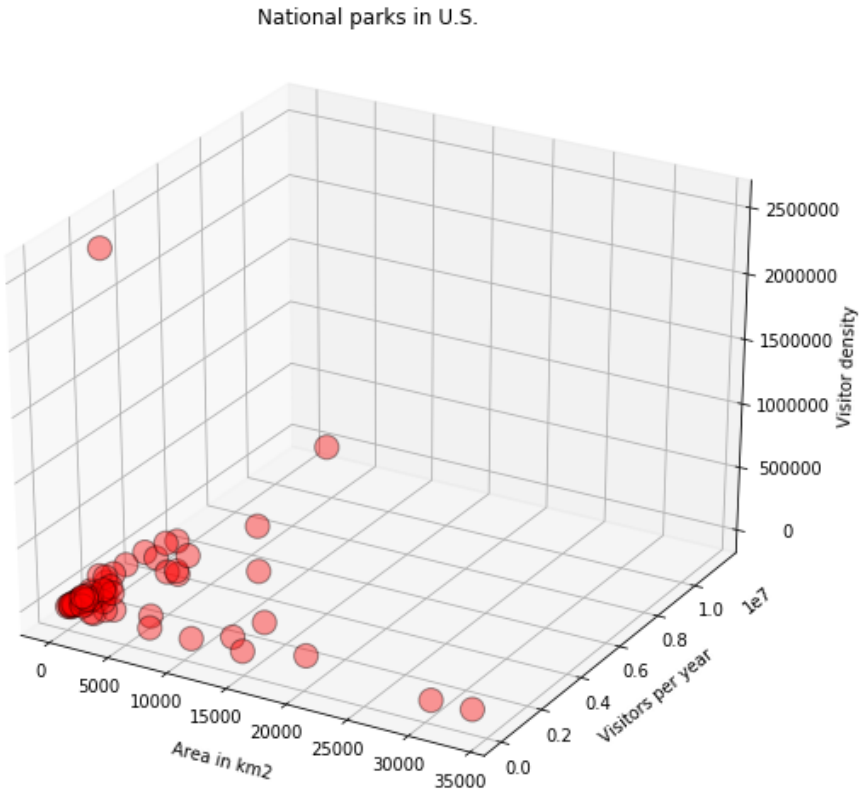


Figure 3 - National parks: Area, number of visitors and visitor density

The park with the highest visitor density, Gateway Arch, has the visitor density about 1000 times as high as the third quartile of the visitor densities of all the national parks. This suggests a possible anomaly. Indeed, the Gateway Arch [6] is not a natural park such as Kayleigh searches for (see Figure 4). Therefore, we drop it from the list of the national parks and re-done the descriptive statistics without it; see Table 13.

The pair plot (Figure 5) shows that there are still quite big differences between the parks. Therefore, we can try to identify and characterize similar groups of the parks, see section Clustering.

Figure 4 - Gateway Arch

	Area in km2	Visitors per year	Crowded
count	61.000000	6.100000e+01	61.000000
mean	3476.854098	1.384678e+06	3749.629508
std	6686.383097	1.911755e+06	9686.138333
min	22.500000	9.591000e+03	0.310000
25%	261.800000	3.089620e+05	362.660000
50%	895.900000	6.449220e+05	971.890000
75%	3082.700000	1.663557e+06	2479.630000
max	33682.600000	1.142120e+07	66972.760000

Table 13 - National parks: Descriptive statistics (Gateway Arch excluded)

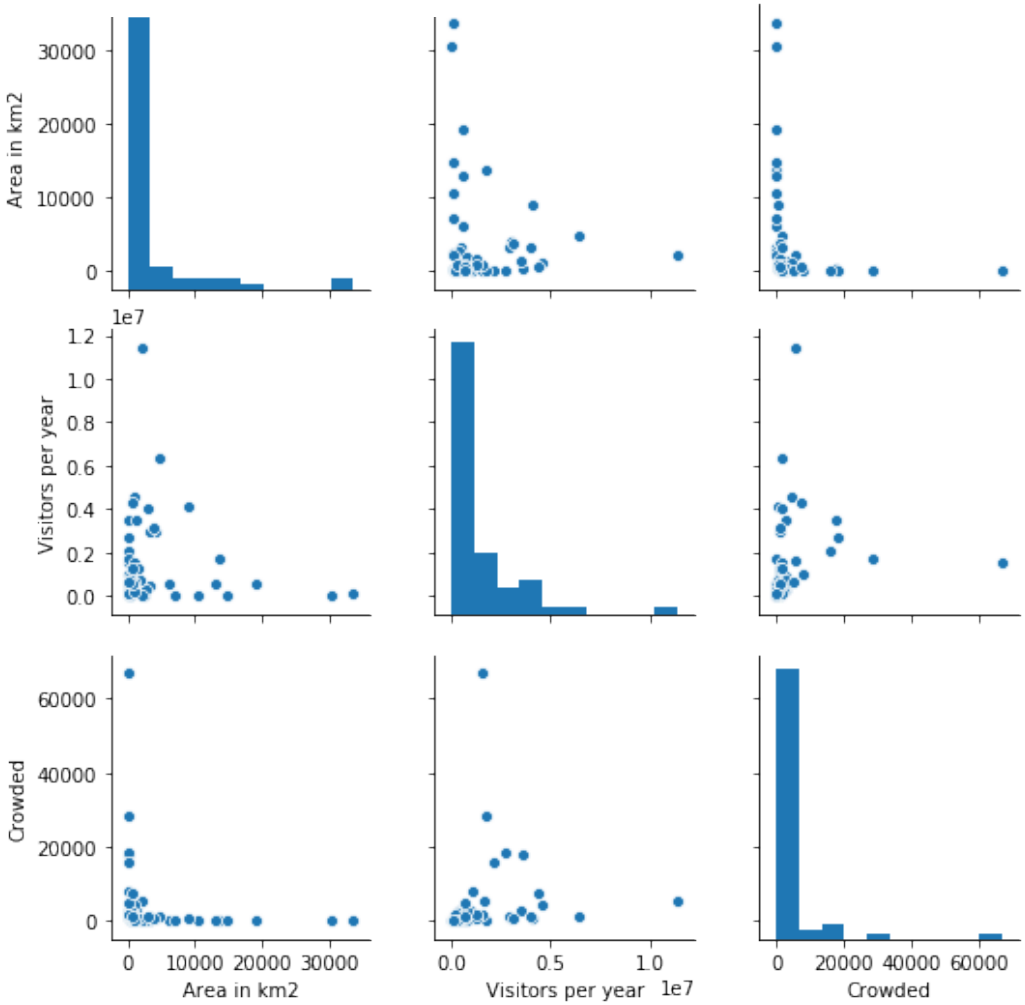


Figure 5 - National parks: Pair plot of the parks' area, number of visitors per year and visitor density

Finally, Figure 6 shows the position of the national parks as well as their area indicated by the size of the marker (radius = $\log_{10}(area) + 4$) and the respective visitor density colour-coded green (less than 10 visitors per year per km^2) to red (more than 10,000 visitors per year per km^2).

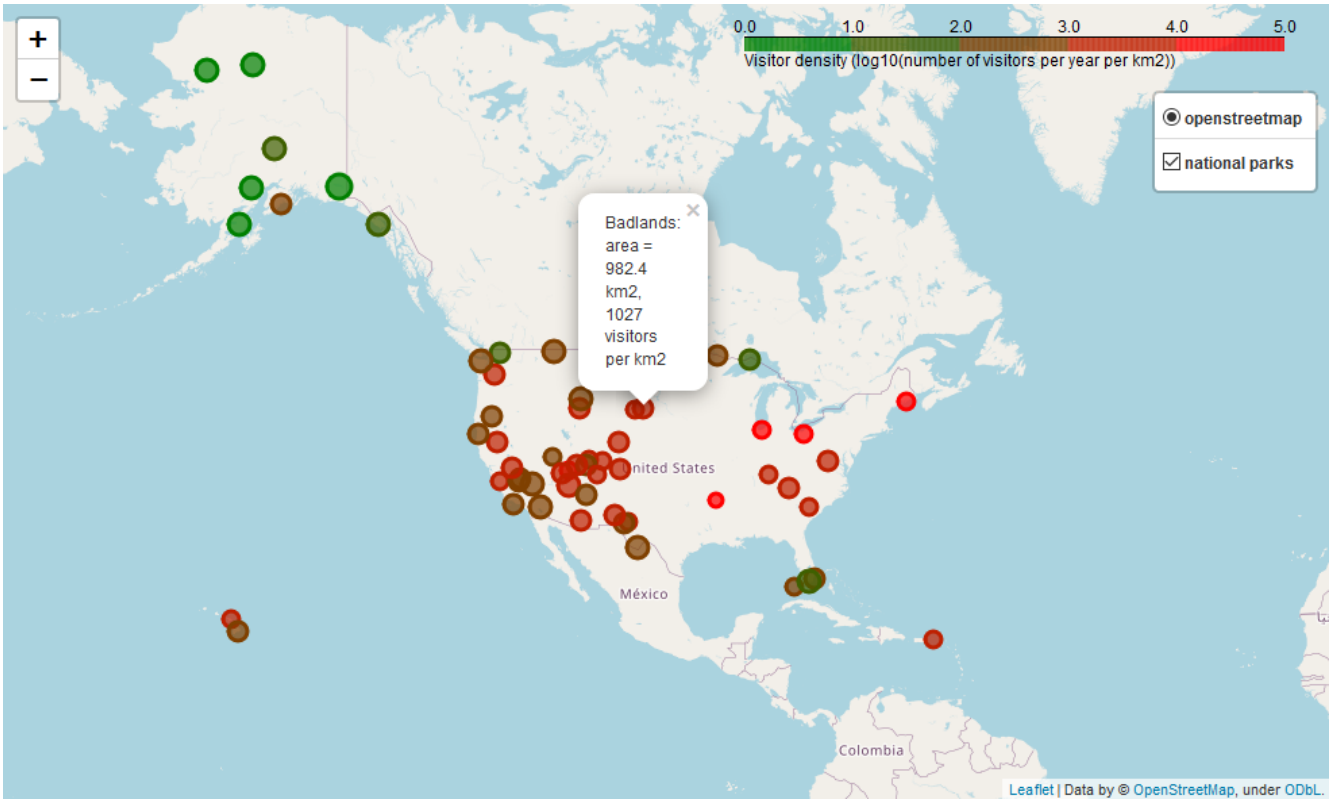


Figure 6 - National parks: Geographical location and visitor density

Universities

Table 14 shows the descriptive statistics (mean, standard variation, minimum, first quartile, median, third quartile and maximum) of the quantitative variables of the 1397 universities. We can see that there are 2 missing values for *International students' ratio*, and 66 missing values for *Female Ratio*.

	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
count	1397.000000	1397.000000	1395.000000	1331.000000	1397.000000	1397.000000	1397.000000	1397.000000	1397.000000	1397.000000
mean	23741.153185	19.012026	0.113577	0.498850	34.588332	28.221832	23.970365	48.091482	46.469435	47.098926
std	32821.296774	16.913747	0.117738	0.123528	17.081937	14.147078	17.535404	27.737472	16.270320	23.287695
min	558.000000	0.900000	0.000000	0.000000	16.400000	11.200000	6.800000	1.700000	34.400000	13.100000
25%	10267.000000	12.400000	0.020000	0.430000	16.400000	18.300000	11.600000	23.300000	35.700000	27.400000
50%	17848.000000	16.400000	0.080000	0.520000	31.750000	23.800000	18.000000	45.500000	39.400000	43.100000
75%	29437.000000	21.900000	0.170000	0.580000	45.650000	33.600000	30.100000	71.900000	49.800000	62.800000
max	830104.000000	493.500000	0.830000	1.000000	95.400000	92.800000	99.600000	100.000000	100.000000	99.700000

Table 14 - Universities: Descriptive statistics

Pairwise relationships in the university data

Figure 7 shows pairwise relationships in the data.

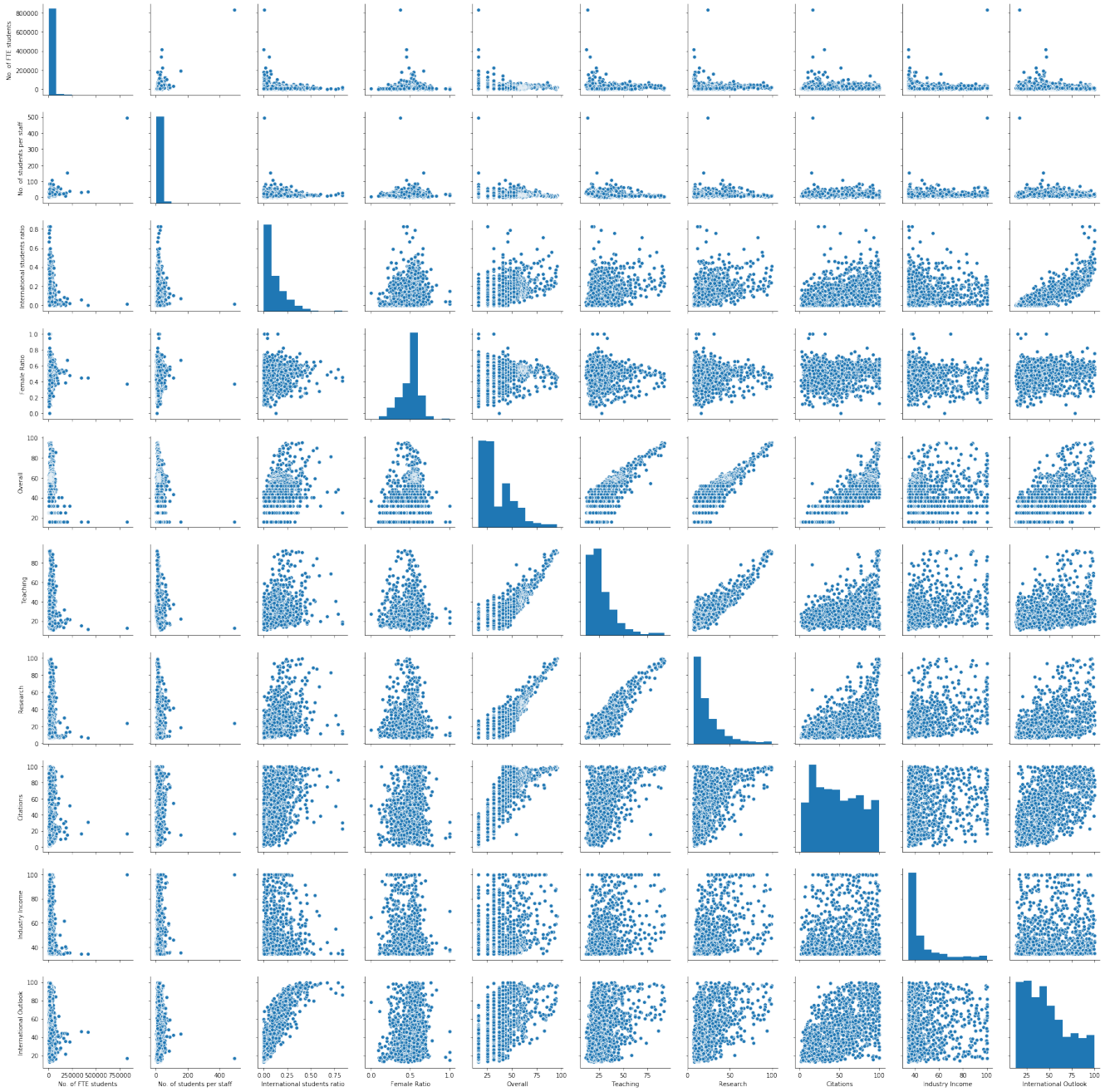


Figure 7 - Universities: Pairwise relationships in the data

The *Overall* score is calculated [7] from the values of the indicators *Teaching* (30%), *Research* (30%), *Citations* (30%), *Industry Income* (2.5%) and *International Outlook* (7.5%), and hence, it is not surprising to see positive correlations between these variables.

Next, we want to find out whether there is a relation between the *Overall* score and any of the variables in the first part of the table: *No. of FTE students*, *No. of students per staff*, *International students' ratio* and *Female Ratio*.

Overall score vs. No. of FTE students

The boxplot of the number of FTE students (Figure 8) reveals the presence of extreme outliers. There are only 3 universities in the upper 70% of the range, shown in Table 15, all of them having more than 300,000 FTE students while the upper quartile of the number of FTE students is equal to 29437 (see Table 14), about 10 times less.

Pearson correlation coefficient between *Overall* and *No. of FTE students* is very low: it is equal to -0.026884 when taking into account all the universities, and 0.001555 with 2 universities with the highest number of FTE students being left out.

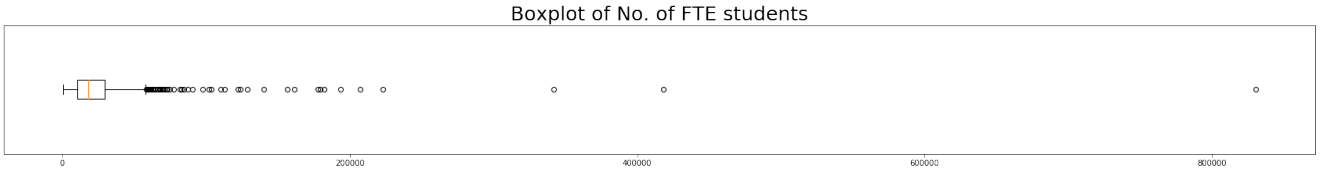


Figure 8 - Universities: Boxplot of No. of FTE students

	Rank	Name	Country/Region	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
1005	1001+	Al-Azhar University	Egypt	342151	32.6	0.06	0.45	16.4	15.2	8.3	16.5	34.5	46.7
1012	1001+	Anadolu University	Turkey	830104	493.5	0.01	0.37	16.4	13.0	24.1	16.5	100.0	17.3
1360	1001+	Tribhuvan University	Nepal	418053	36.0	0.00	0.45	16.4	11.4	6.9	30.8	34.4	45.8

Table 15 - Universities with the highest No. of FTE students

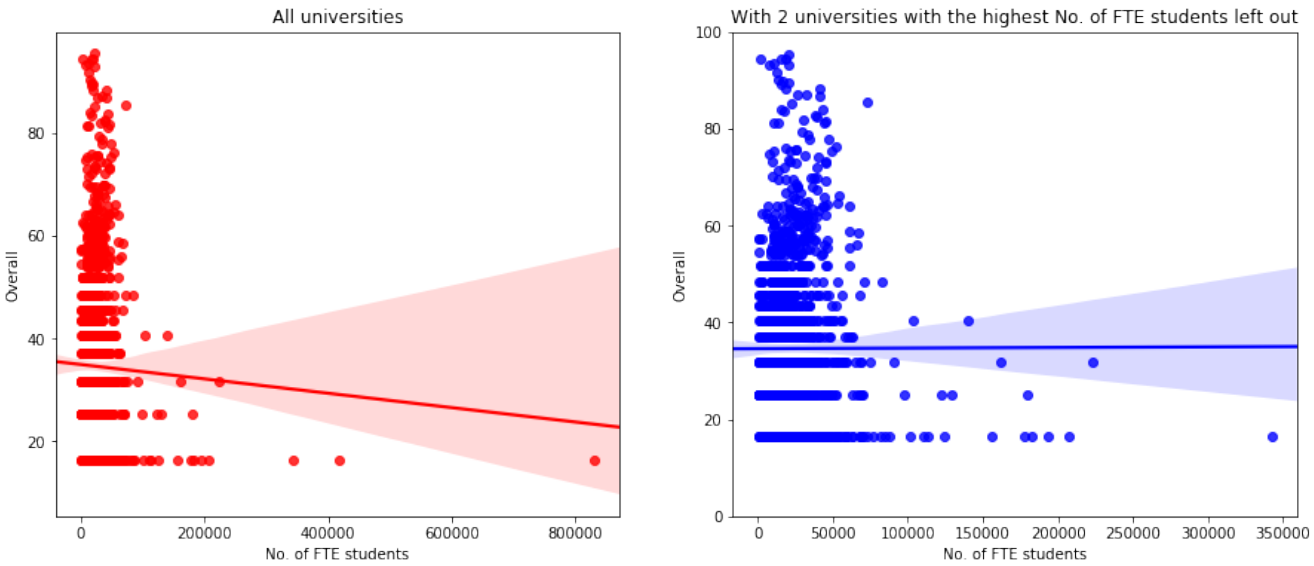


Figure 9 - Universities: Overall vs. No. of FTE students

Overall score vs. No. of students per staff

The boxplot of the number of students per staff member (Figure 10) also shows the presence of extreme outliers. There are only 2 universities in the upper 70% of the range, shown in Table 16, having 493.5 and 155.2 students per staff member, respectively, while the upper quartile of the number of students per staff member of all universities is equal to 21.9.

Pearson correlation coefficient between *Overall* and *No. of students per staff* is also very low: it is equal to -0.02358 when taking into account all the universities, and 0.006415 when the 2 universities with the highest number of students per staff member are left out.

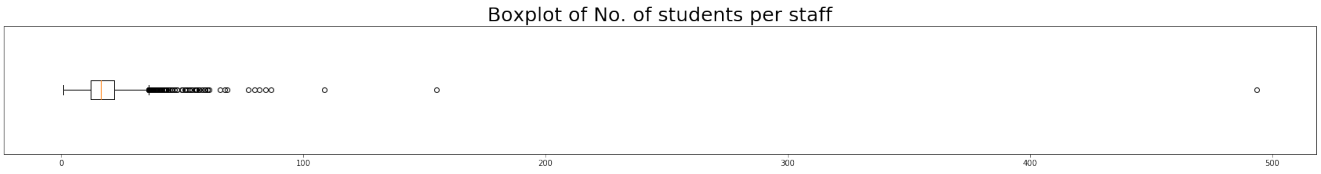


Figure 10 - Universities: Boxplot of No. of students per staff

	Rank	Name	Country/Region	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
1012	1001+	Anadolu University	Turkey	830104	493.5	0.01	0.37	16.4	13.0	24.1	16.5	100.0	17.3
1314	1001+	University of South Africa	South Africa	193874	155.2	0.07	0.67	16.4	22.1	19.1	15.0	35.2	43.7

Table 16 - Universities with the highest No. of students per staff

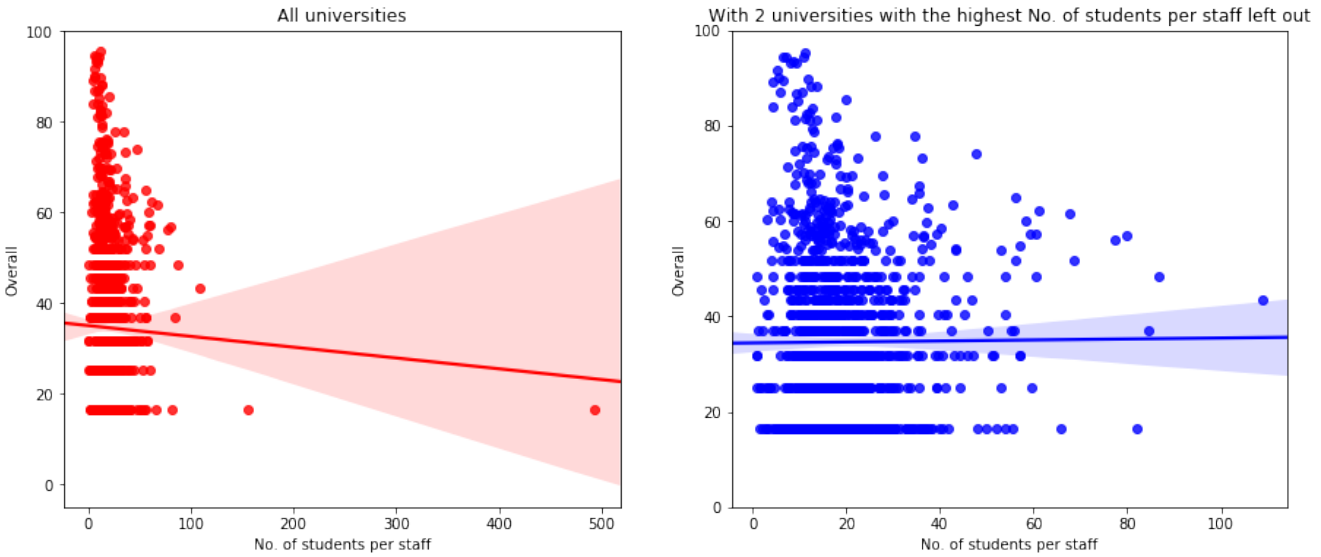


Figure 11 - Universities: Overall vs. No. of students per staff

Overall score vs. International students' ratio

While there are some outliers in *International students' ratio*, as can be seen in Figure 12, they are not that extreme as in the two previous variables.
Overall score and *International students' ratio* are positively correlated, their Pearson correlation coefficient is equal to 0.562808.

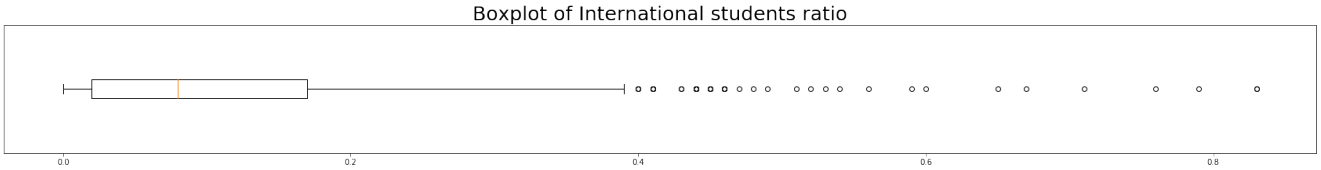


Figure 12 - Universities: Boxplot of International students' ratio

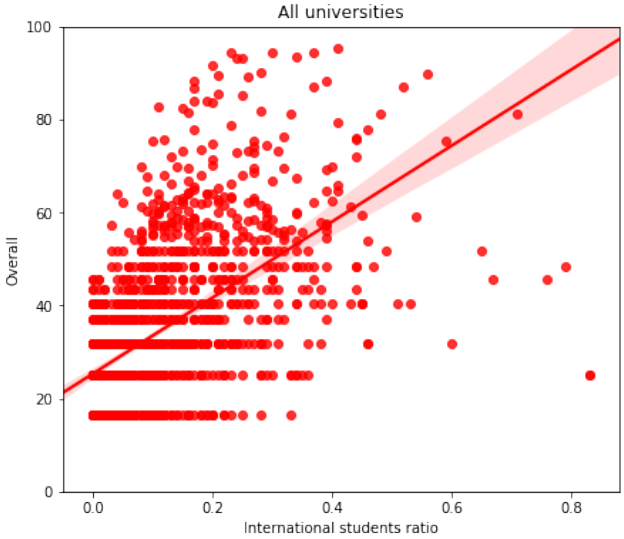


Figure 13 - Universities: Overall vs. International students' ratio

Overall score vs. Female Ratio

Variable *Female Ratio* has the lowest number of outlier and the outliers are least extreme (compared to *No. of FTE students*, *No. of students per staff* and *International students' ratio*), see Figure 14.
Pearson correlation coefficient between *Overall* score and *Female Ratio* is equal to 0.059456.

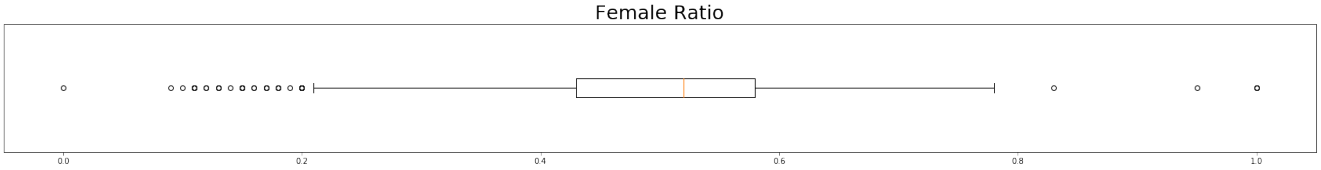


Figure 14 - Universities: Boxplot of Female Ratio

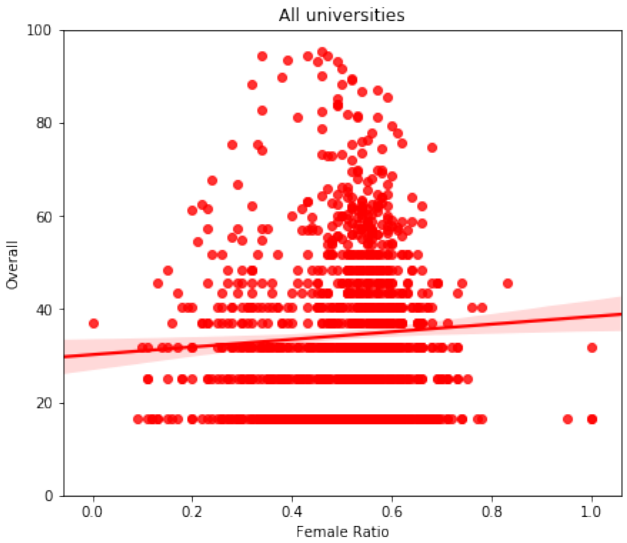


Figure 15 - Universities: Overall vs. Female Ratio

We can see that the variables *No. of FTE students*, *No. of students per staff* and *Female Ratio* are no good predictors for the *Overall* score since the respective regression lines (see Figure 9, Figure 11 and Figure 15) are close to horizontal and the data points are very scattered and far from the fitted line. The respective correlations are low.
On the other hand, both the scatterplot with the fitted regression line and the correlation coefficient show a relation between the *International students' ratio* and the *Overall* score (not surprisingly: except of other influences, *International students' ratio* counts directly towards *International Outlook* and *International Outlook* counts directly towards *Overall score*), but the data points are still quite far from the fitted line.

Are the U.S. universities that good?

Let's compare the descriptive statistics of the (ranked) US universities to the rest of the world.
US universities:

	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
count	172.000000	172.000000	172.000000	166.000000	172.000000	172.000000	172.000000	172.000000	172.000000	172.000000
mean	22829.168605	14.874419	0.134070	0.510783	49.870930	41.159302	36.823837	73.155814	44.927326	48.540698
std	13487.184146	6.052773	0.086766	0.084394	18.829367	19.419580	24.014286	20.590929	13.668793	15.121290
min	1809.000000	1.000000	0.010000	0.180000	16.400000	15.900000	7.700000	23.300000	34.400000	19.700000
25%	12125.000000	10.950000	0.077500	0.480000	37.000000	26.900000	19.475000	58.775000	36.200000	37.550000
50%	21404.000000	15.550000	0.110000	0.520000	45.650000	35.750000	27.950000	74.650000	39.800000	46.100000
75%	30422.500000	18.325000	0.180000	0.560000	61.525000	49.675000	47.000000	93.325000	47.525000	59.675000
max	66872.000000	35.300000	0.480000	0.740000	94.500000	92.800000	98.600000	100.000000	99.900000	89.000000

Table 17 - Descriptive statistics of the US universities

Non-US universities:

	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
count	1225.000000	1225.000000	1223.000000	1165.000000	1225.000000	1225.000000	1225.000000	1225.000000	1225.000000	1225.000000
mean	23869.203265	19.592980	0.110695	0.497150	32.442531	26.405306	22.165633	44.572245	46.685959	46.896490
std	34685.265939	17.844134	0.121212	0.128072	15.679147	12.196524	15.609538	26.787556	16.596379	24.212544
min	558.000000	0.900000	0.000000	0.000000	16.400000	11.200000	6.800000	1.700000	34.400000	13.100000
25%	9986.000000	12.500000	0.020000	0.420000	16.400000	17.800000	11.100000	20.600000	35.700000	25.800000
50%	17432.000000	16.500000	0.070000	0.530000	31.750000	22.700000	16.600000	40.700000	39.300000	42.300000
75%	29311.000000	22.500000	0.160000	0.580000	40.550000	30.800000	28.400000	66.900000	50.400000	63.400000
max	830104.000000	493.500000	0.830000	1.000000	95.400000	91.400000	99.600000	100.000000	100.000000	99.700000

Table 18 - Descriptive statistics of the non-US universities

And more specifically, universities in Ireland:

	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
count	9.000000	9.000000	9.000000	7.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000
mean	13718.444444	22.744444	0.242222	0.540000	43.766667	25.744444	28.011111	67.811111	41.544444	83.155556
std	5603.862443	3.201215	0.167016	0.061101	10.343114	7.798736	8.669695	18.120047	3.827895	8.635553
min	2265.000000	18.000000	0.100000	0.430000	25.200000	15.700000	13.300000	32.900000	35.100000	71.800000
25%	12619.000000	21.100000	0.170000	0.515000	37.000000	20.300000	23.500000	54.600000	39.200000	78.200000
50%	14221.000000	21.900000	0.180000	0.570000	45.650000	24.900000	26.900000	75.400000	42.400000	81.300000
75%	16853.000000	24.000000	0.290000	0.580000	51.900000	30.100000	29.900000	77.500000	43.900000	92.500000
max	22541.000000	28.300000	0.650000	0.590000	56.400000	41.700000	43.900000	90.000000	47.500000	94.400000

Table 19 - Descriptive statistics of the universities in Ireland

In Table 17, Table 18 and Table 19, we can see that taking into account *all* the ranked universities, US universities do not seem to be that much better compared to the universities in Ireland. While in Ireland, there are only 9 ranked universities with *Overall* score ranging 25.2-56.4 with mean equal to 43.8, in USA, there are 172 ranked universities with *Overall* score ranging 16.4-94.5 with mean equal to 49.9. Thus, the scores of the US universities are much more spread.

On the other hand, Sean only considers top universities (while he cannot specify yet what that means to him), and therefore, it is reasonable to only compare the 500 best universities in the ranking.

US universities in Top500:

	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
count	121.000000	121.000000	121.000000	115.000000	121.000000	121.000000	121.000000	121.000000	121.000000	121.000000
mean	24549.958678	13.147934	0.151983	0.511652	58.109504	48.005785	44.854545	83.545455	47.840496	52.735537
std	14433.236336	5.341865	0.083363	0.070883	16.108460	19.086240	24.356755	13.024631	14.977275	14.455967
min	1809.000000	1.000000	0.010000	0.290000	40.550000	17.100000	10.300000	55.500000	34.400000	19.700000
25%	12735.000000	9.400000	0.090000	0.480000	45.650000	33.900000	25.300000	72.700000	38.000000	40.900000
50%	23116.000000	13.000000	0.140000	0.520000	51.900000	42.800000	35.300000	84.400000	42.100000	52.200000
75%	33186.000000	16.600000	0.200000	0.550000	68.100000	57.500000	59.300000	96.400000	50.500000	63.300000
max	66872.000000	27.600000	0.480000	0.740000	94.500000	92.800000	98.600000	100.000000	99.900000	89.000000

Table 20 - Descriptive statistics of the Top500, US universities

Non-US universities in Top500:

	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
count	379.000000	379.000000	379.000000	355.000000	379.000000	379.000000	379.000000	379.000000	379.000000	379.000000
mean	21758.841689	20.975198	0.202427	0.511606	51.848549	37.386807	38.352507	75.102639	54.777045	69.153298
std	15744.823464	13.314893	0.137621	0.116201	10.989742	14.844771	17.601131	15.998098	20.085157	21.622945
min	558.000000	2.600000	0.000000	0.130000	40.550000	11.900000	7.600000	15.600000	34.400000	15.600000
25%	10924.000000	12.700000	0.105000	0.470000	43.400000	27.150000	26.300000	64.650000	38.500000	53.250000
50%	18642.000000	17.200000	0.170000	0.550000	48.450000	35.200000	35.500000	75.800000	45.800000	72.600000
75%	29280.500000	24.850000	0.290000	0.580000	57.500000	44.150000	47.150000	88.000000	66.850000	87.750000
max	140126.000000	108.800000	0.790000	0.830000	95.400000	91.400000	99.600000	100.000000	100.000000	99.700000

Table 21 - Descriptive statistics of the Top500, non-US universities

And more specifically, Top500 universities in Ireland:

	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
count	6.000000	6.000000	6.000000	4.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.00000
mean	13620.333333	23.166667	0.285000	0.575000	49.991667	29.150000	31.683333	78.850000	42.800000	86.65000
std	7047.007270	3.760674	0.192328	0.017321	4.205403	7.234846	7.426013	5.902796	3.439767	7.62804
min	2265.000000	18.000000	0.120000	0.550000	45.650000	20.300000	23.100000	73.600000	37.300000	78.20000
25%	10446.000000	21.300000	0.172500	0.572500	46.350000	25.450000	27.625000	75.625000	41.500000	80.17500
50%	15382.000000	22.500000	0.235000	0.580000	50.175000	28.600000	29.850000	76.900000	43.150000	86.90000
75%	16859.000000	25.725000	0.297500	0.582500	51.900000	30.625000	34.850000	79.600000	44.350000	93.40000
max	22541.000000	28.300000	0.650000	0.590000	56.400000	41.700000	43.900000	90.000000	47.500000	94.40000

Table 22 - Descriptive statistics of the Top500 universities in Ireland

In Table 20 and Table 22, we can see that even the best university in Ireland scores under the mean of the Top500, US universities in *Overall* (56.4 vs. 58.1), *Teaching* (41.7 vs. 48.0) and *Research* (43.9 vs. 44.9) scores. The average *No. of students per staff* over all respective universities is substantially higher in Ireland than in the USA (23.2 vs. 13.1) which could possibly explain a part of the difference in the scores.

Figure 16 and Figure 17 show the boxplots of *Overall* scores per country of all universities and of the Top500 universities, respectively.

Figure 18 shows the location of the Top500 US universities as well as their respective number of FTE students indicated by the size of the marker (radius = square root of the number of FTE students, divided by 20) and ratio of international students colour-coded blue (less than 10%) to yellow (more than 40%). Finally, Figure 19 shows both the national parks and the Top500 US universities.

In the following parts, with universities we mean US universities in the Top500 of the *Times Higher Education World University Rankings 2020*.

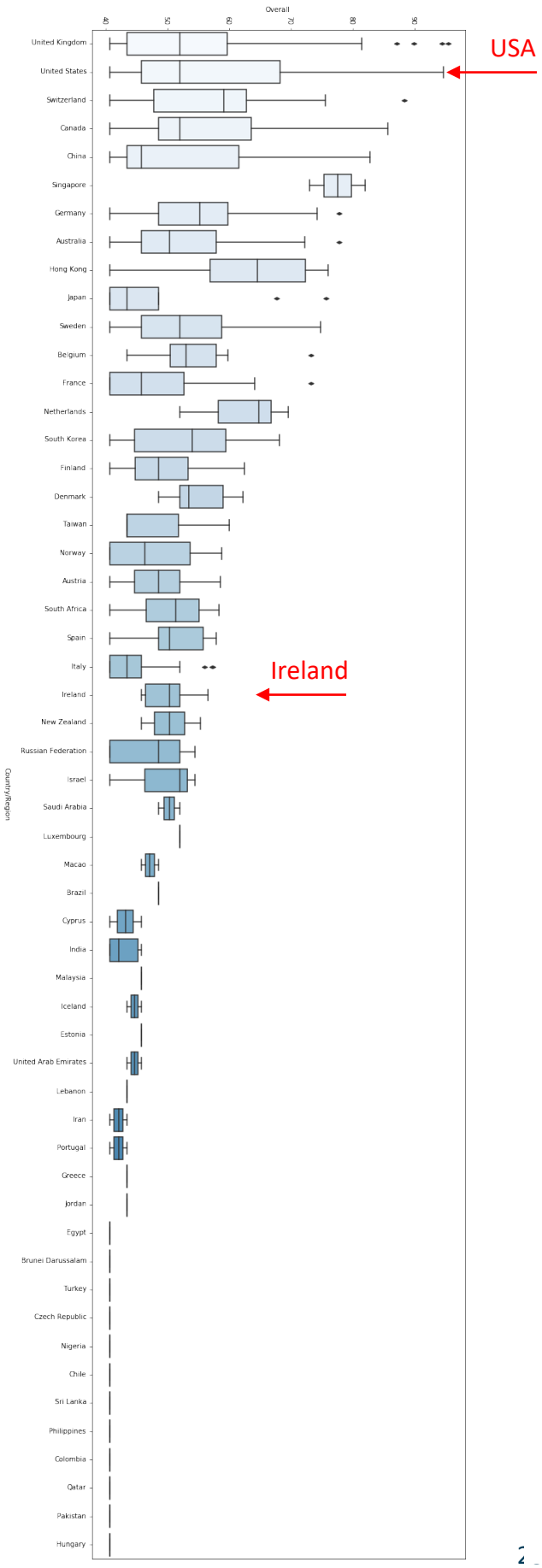


Figure 17 - Boxplots of Overall scores of Top500 universities, per country

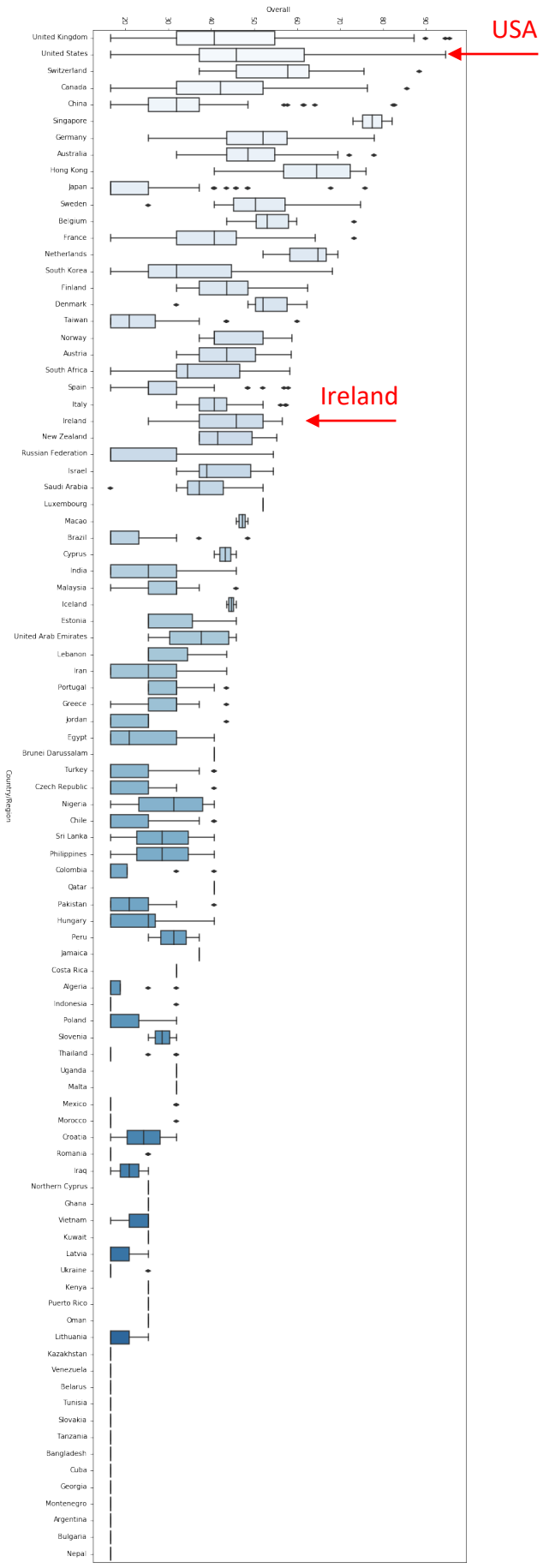


Figure 16 - Boxplots of Overall scores of all universities, per country

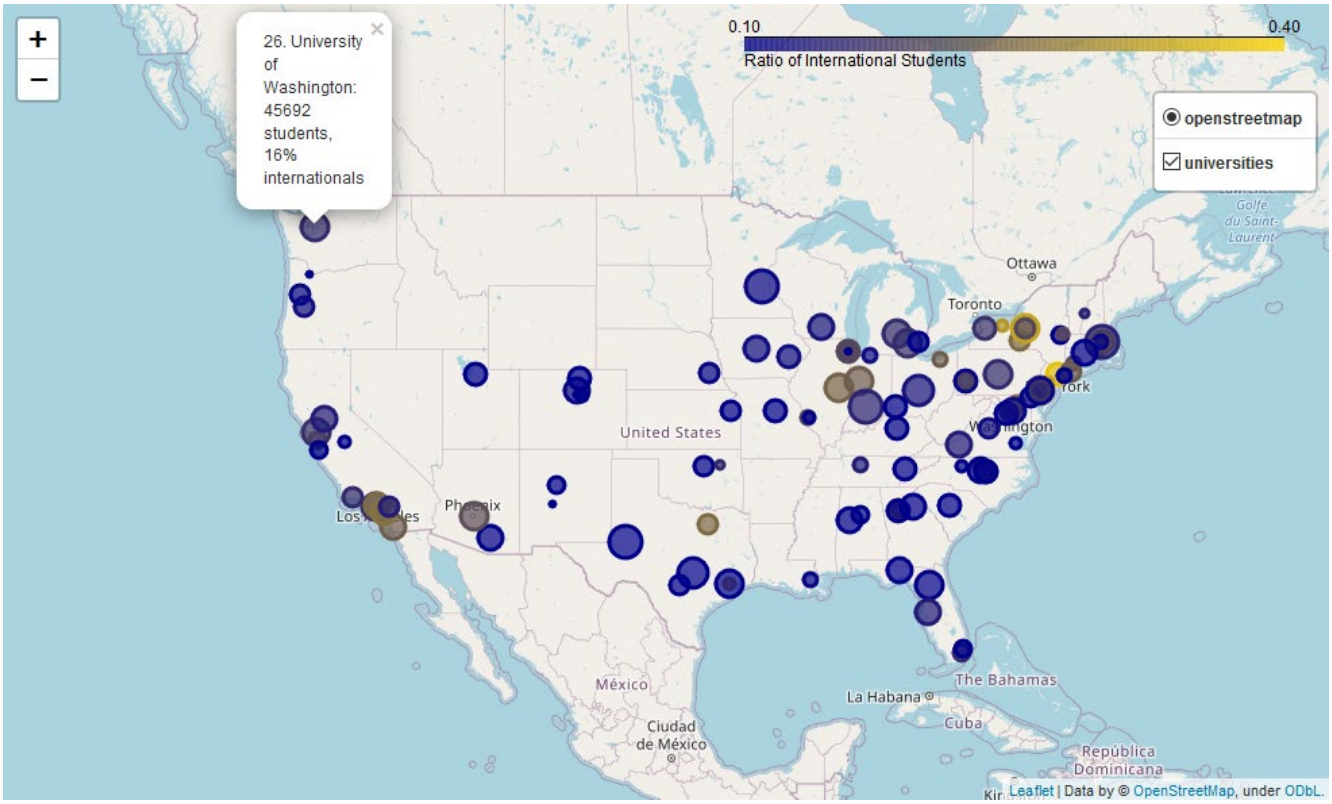


Figure 18 - Top500 US universities: Geographical location and ratio of international students

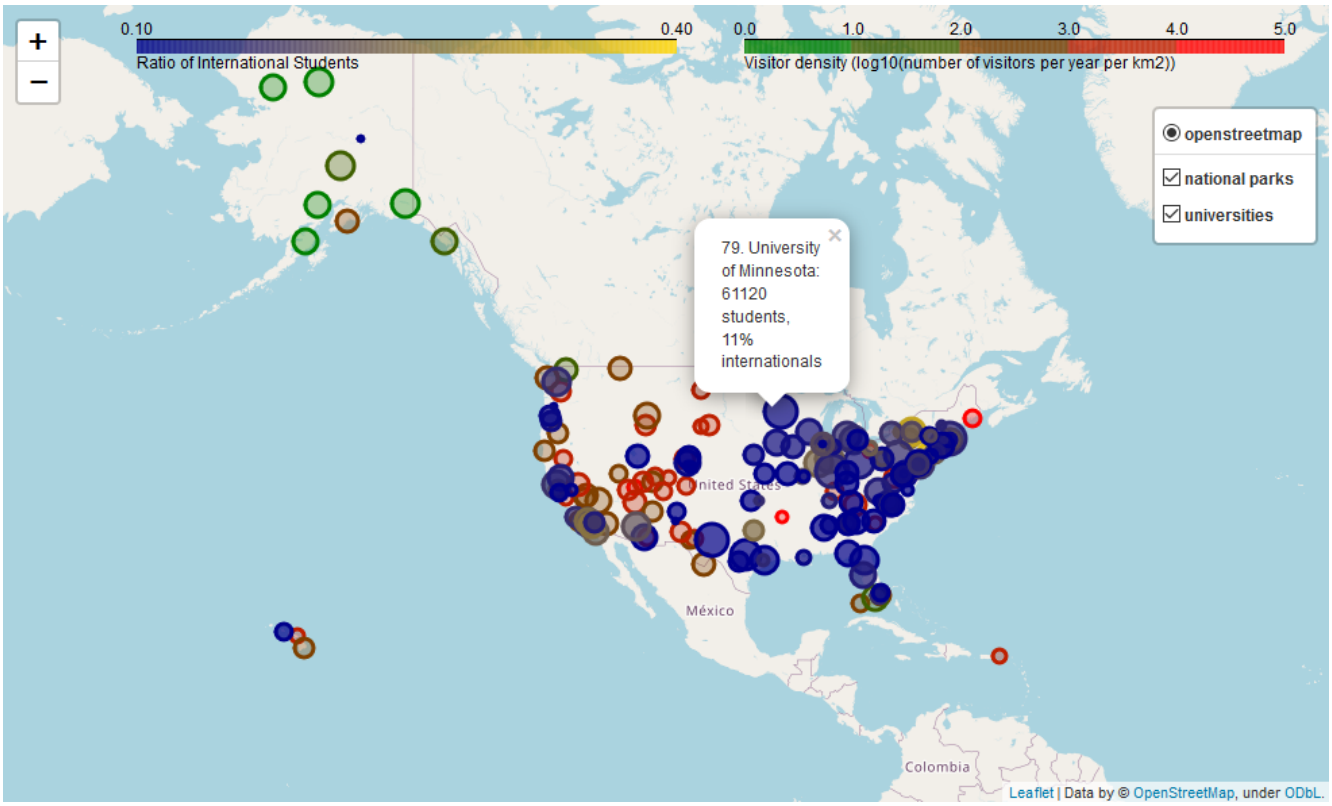


Figure 19 - National parks and Top500 US universities

Foursquare: Venues in the neighbourhood of the universities

Frequencies of the retrieved Foursquare venues within 1500 m from the universities give us an insight in the most typical venues found in the vicinity of the US universities. Figure 20 shows 20 most common venues.

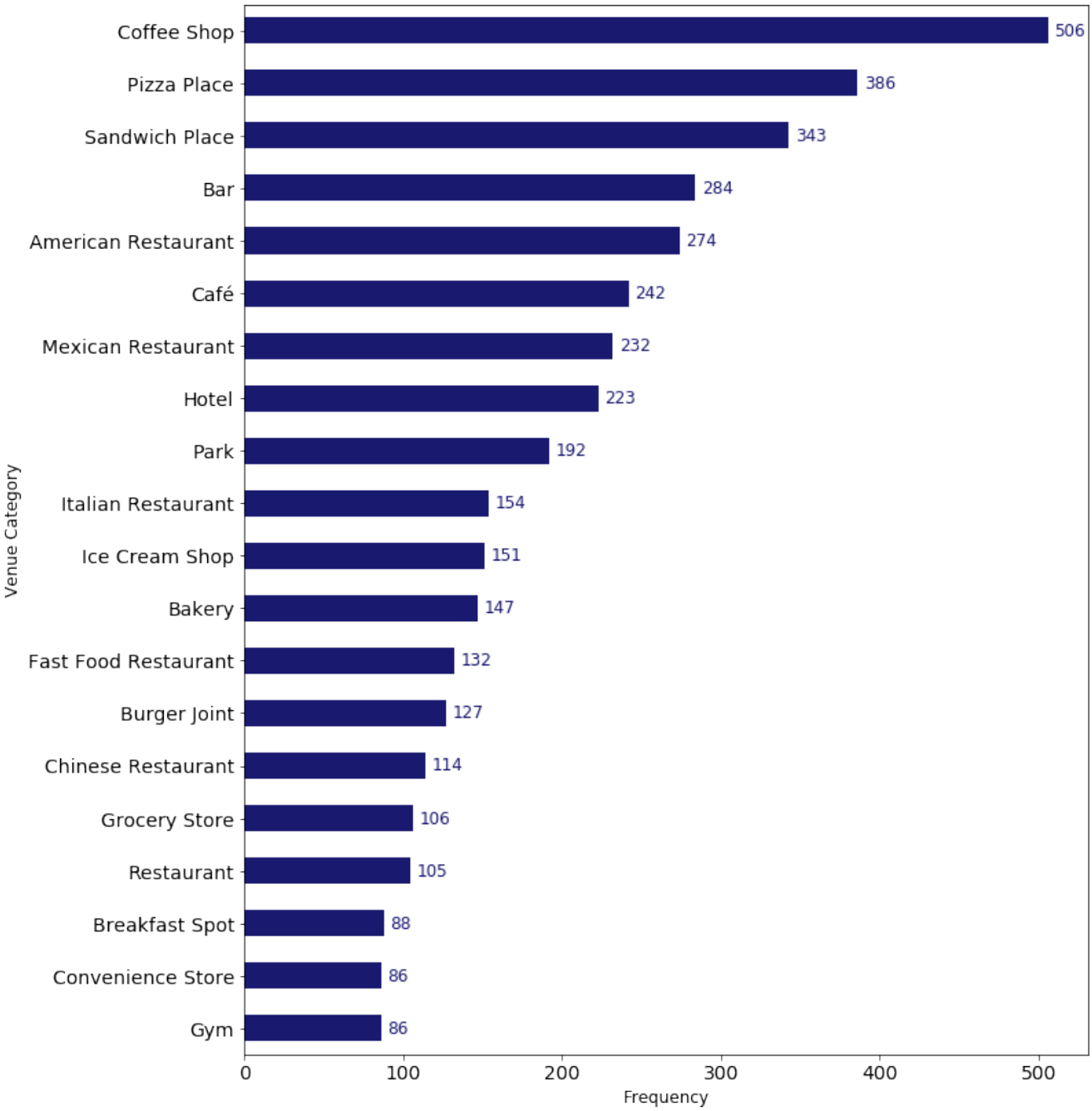


Figure 20 - Foursquare: Frequencies of the most common venues within 1500m from the US universities

Clustering

National parks

Kayleigh is interested in the characteristics of the national parks: are they huge or compact; are they crowded or can you enjoy your solitude and make photographs without being constantly disturbed by tourists walking by? We will use K-Means clustering to group the national parks, based on their respective area, number of visitors per year and visitor density, to get the first insight.

How many clusters?

First, we use the elbow method to find the best value of k for K-Means. However, as shown in Figure 21, there seem to be two elbows: at k=4 and at k=7. So, based on the elbow method, we cannot tell which of these values of k is better.

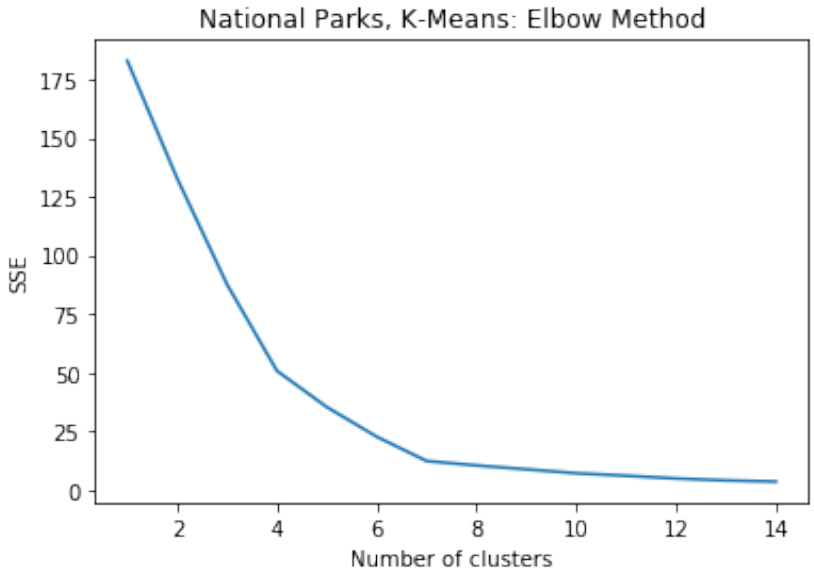


Figure 21 - Clustering the national parks: Elbow method to find the best value of k

Therefore, we also check the silhouette score, which often gives a better idea concerning the optimal value of k than the elbow method. According to [8],

- The silhouette score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.
- The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

Based on the silhouette score (see Figure 22), we can conclude that the optimal value for K-Means is k=7.

Another clustering method, the Affinity Propagation [9], suggests clustering into 8 clusters (see Figure 23) – that is not much different from 7 clusters by K-Means.

Figure 24 shows that the clustering by K-Means is nearly the same as the clustering by Affinity Propagation. The only difference is that one of the K-Means clusters becomes two clusters using the Affinity Propagation.

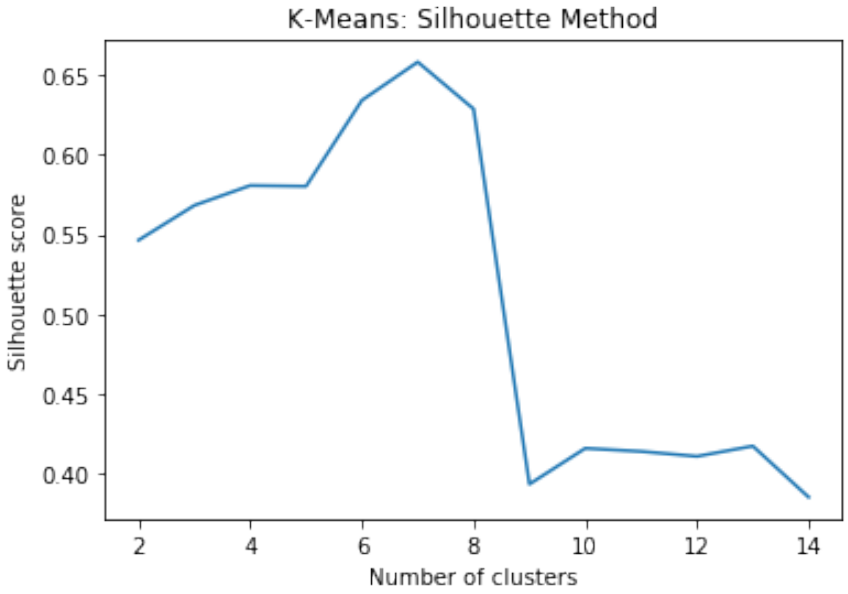


Figure 22 - Clustering the national parks: Silhouette method to find the best value of k

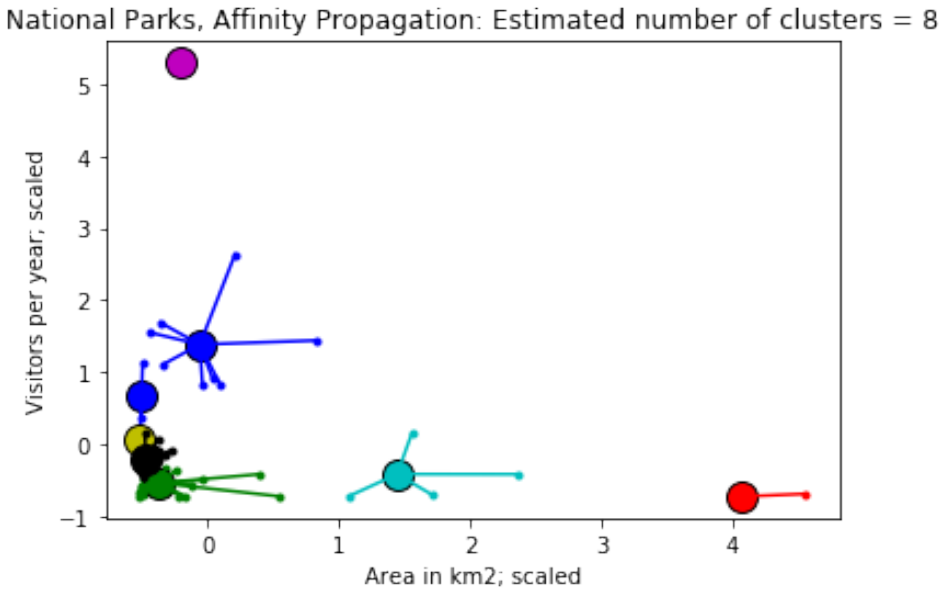


Figure 23 - Clustering the national parks: Affinity propagation

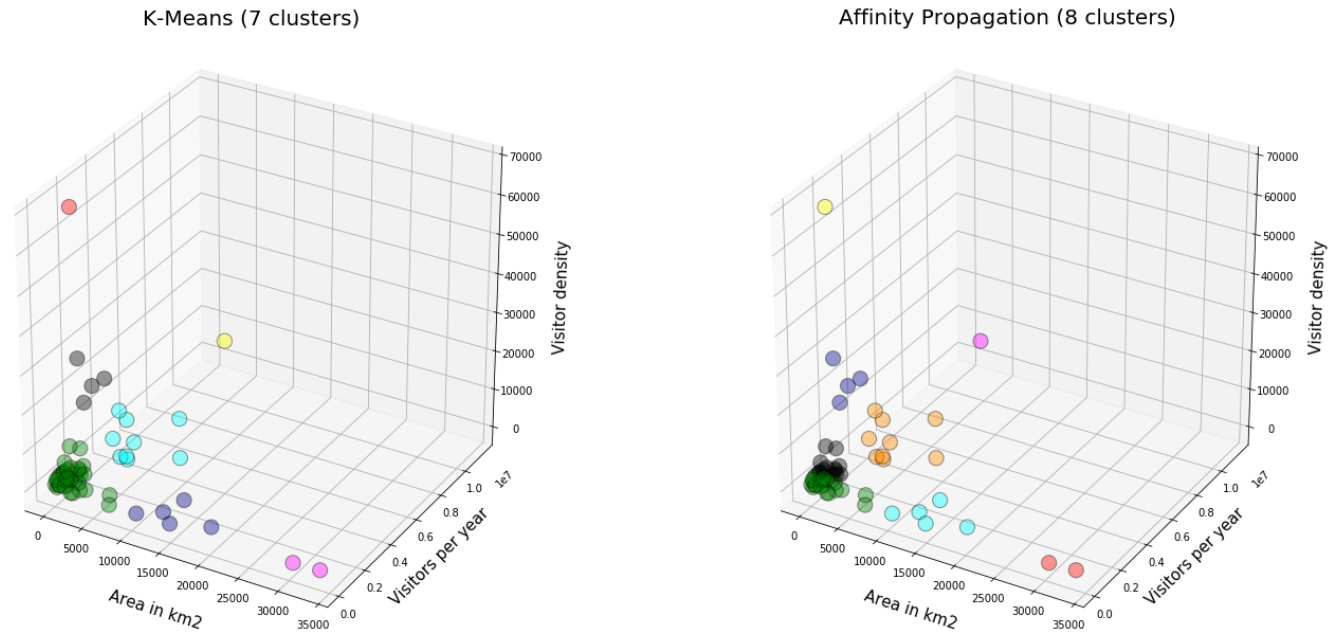


Figure 24 - Clustering the national parks: K-Means vs. affinity propagation

Characterizing the clusters

Figure 25 shows the geographical location and clustering of the national parks.

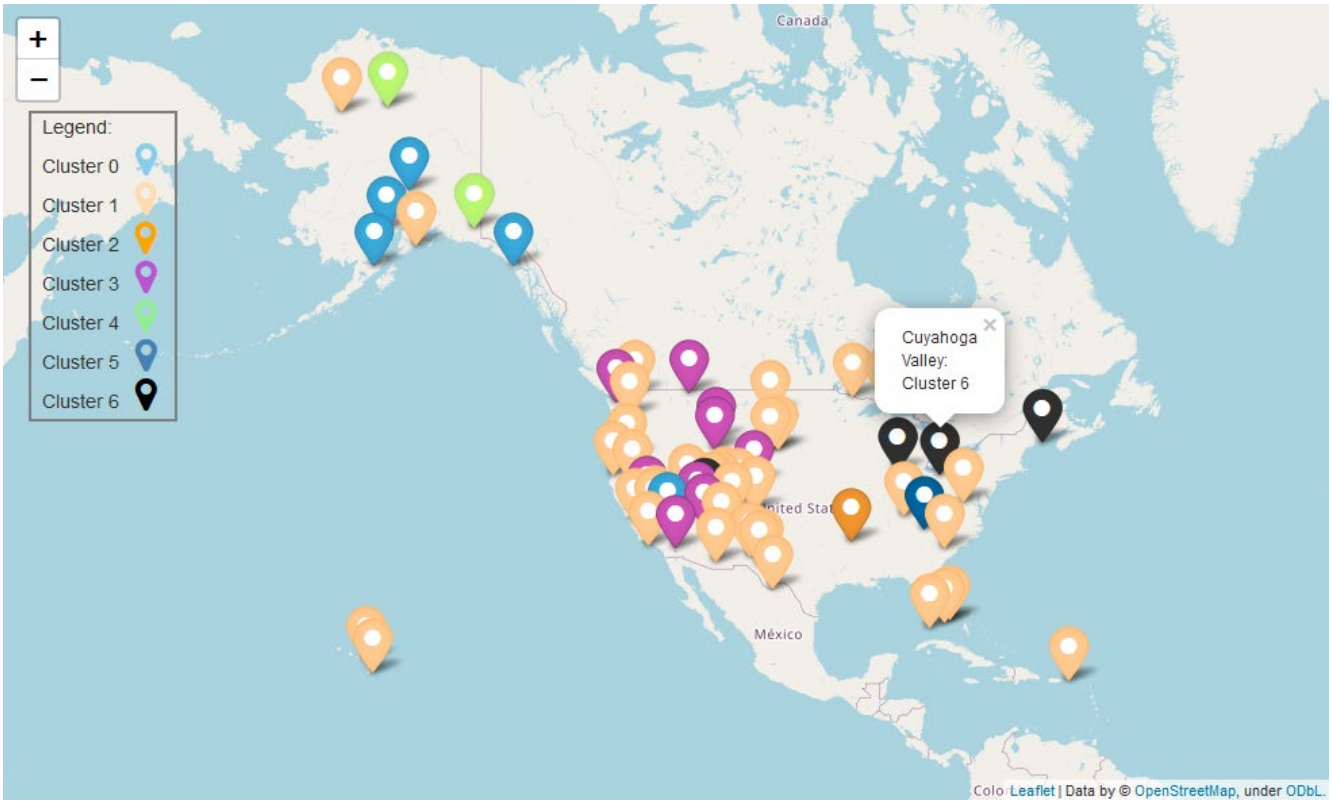


Figure 25 - National parks: Geographical location and clustering

Cluster 0 consists of 5 parks (4 of which in Alaska) with a large area (10000-20000 km^2) and a low visitor density (< 125).

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded	Cluster Labels K-Means
15	Death Valley	13793.3	1678660	36.24	-116.82	California, Nevada	121.70	0
16	Denali	19185.8	594660	63.33	-150.50	Alaska	30.99	0
21	Glacier Bay	13044.6	597915	58.50	-137.00	Alaska	45.84	0
34	Katmai	14870.3	37818	58.50	-155.00	Alaska	2.54	0
38	Lake Clark	10602.0	14479	60.97	-153.42	Alaska	1.37	0

Cluster 1 is the biggest one, consisting of about 2/3 of all the parks, with a relatively small area and a lower number of visitors than cluster 3.

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded	Cluster Labels K-Means
1	American Samoa	33.4	28626	-14.25	-170.68	American Samoa	857.07	1
2	Arches	310.3	1663557	38.68	-109.57	Utah	5361.12	1
3	Badlands	982.4	1008942	43.75	-102.50	South Dakota	1027.02	1
4	Big Bend	3242.2	440091	29.25	-103.25	Texas	135.74	1
5	Biscayne	700.0	469253	25.65	-80.08	Florida	670.36	1
6	Black Canyon of the Gunnison	124.6	308962	38.57	-107.72	Colorado	2479.63	1
8	Canyonlands	1366.2	739449	38.20	-109.93	Utah	541.25	1
9	Capitol Reef	979.0	1227627	38.20	-111.17	Utah	1253.96	1
10	Carlsbad Caverns *	189.3	465912	32.17	-104.44	New Mexico	2461.24	1
11	Channel Islands	1009.9	366250	34.01	-119.42	California	362.66	1
12	Congaree	107.1	145929	33.78	-80.78	South Carolina	1362.55	1
13	Crater Lake	741.5	720659	42.94	-122.10	Oregon	971.89	1
17	Dry Tortugas	261.8	56810	24.63	-82.87	Florida	217.00	1
18	Everglades	6106.5	597124	25.32	-80.93	Florida	97.78	1
24	Great Basin	312.3	153094	38.98	-114.30	Nevada	490.21	1
25	Great Sand Dunes	434.4	442905	37.73	-105.51	Colorado	1019.58	1
27	Guadalupe Mountains	349.5	172347	31.92	-104.87	Texas	493.12	1
28	Haleakalā	134.6	1044084	20.72	-156.17	Hawaii	7756.94	1
29	Hawai'i Volcanoes	1317.7	1116891	19.38	-155.20	Hawaii	847.61	1
32	Isle Royale	2314.0	25798	48.10	-88.55	Michigan	11.15	1
35	Kenai Fjords	2710.0	321596	59.92	-149.65	Alaska	118.67	1
36	Kings Canyon	1869.2	699023	36.80	-118.55	California	373.97	1
37	Kobuk Valley	7084.9	14937	67.55	-159.28	Alaska	2.11	1
39	Lassen Volcanic	431.4	499435	40.49	-121.51	California	1157.71	1

40	Mammoth Cave	218.6	533206	37.18	-86.10	Kentucky	2439.19	1
41	Mesa Verde *	212.4	563420	37.18	-108.49	Colorado	2652.64	1
42	Mount Rainier	956.6	1518491	46.85	-121.75	Washington	1587.38	1
43	North Cascades	2042.8	30085	48.70	-121.20	Washington	14.73	1
45	Petrified Forest	895.9	644922	35.07	-109.78	Arizona	719.86	1
46	Pinnacles	108.0	222152	36.48	-121.16	California	2056.96	1
47	Redwood *	562.5	482536	41.30	-124.00	California	857.84	1
49	Saguaro	371.2	957405	32.25	-110.50	Arizona	2579.22	1
50	Sequoia	1635.2	1229594	36.43	-118.68	California	751.95	1
51	Shenandoah	806.2	1264880	38.53	-78.35	Virginia	1568.94	1
52	Theodore Roosevelt	285.1	749389	46.97	-103.45	North Dakota	2628.51	1
53	Virgin Islands	60.9	112287	18.33	-64.73	U.S. Virgin Islands	1843.79	1
54	Voyageurs	883.1	239656	48.50	-92.88	Minnesota	271.38	1
55	White Sands	592.2	603008	32.78	-106.17	New Mexico	1018.25	1
56	Wind Cave	137.5	656397	43.57	-103.48	South Dakota	4773.80	1

Cluster 2 consists of only 1 park (Hot Springs) with by far the highest visitor density (66973 visitors per year per km^2).

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded	Cluster Labels K-Means
30	Hot Springs	22.5	1506887	34.51	-93.05	Arkansas	66972.76	2

Cluster 3 consists of 9 parks with a large number of visitors per year (3 to 6 million), less crowded than cluster 6 due to a larger area.

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded	Cluster Labels K-Means
20	Glacier	4100.0	2965309	48.80	-114.00	Montana	723.25	3
22	Grand Canyon *	4862.9	6380495	36.06	-112.14	Arizona	1312.08	3
23	Grand Teton	1254.7	3491151	43.73	-110.80	Wyoming	2782.46	3
33	Joshua Tree	3217.9	2942382	33.79	-115.90	California	914.38	3
44	Olympic	3733.8	3104455	47.97	-123.50	Washington	831.45	3
48	Rocky Mountain	1075.7	4590493	40.40	-105.58	Colorado	4267.45	3
58	Yellowstone	8983.2	4115000	44.60	-110.50	Wyoming, Montana, Idaho	458.08	3
59	Yosemite *	3082.7	4009436	37.83	-119.50	California	1300.62	3
60	Zion	595.9	4320033	37.30	-113.05	Utah	7249.59	3

Cluster 4  consists of only 2 parks in Alaska with a much larger area than the other parks. Both of them have a relatively low number of visitors per year and thus, a low visitor density.

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded	Cluster Labels K-Means
19	Gates of the Arctic	30448.1	9591	67.78	-153.3	Alaska	0.31	4
57	Wrangell–St. Elias *	33682.6	79450	61.00	-142.0	Alaska	2.36	4

Cluster 5  consists of only 1 park (Great Smoky Mountains) with by far the highest number of visitors per year.

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded	Cluster Labels K-Means
26	Great Smoky Mountains	2114.2	11421200	35.68	-83.53	North Carolina, Tennessee	5402.14	5

Cluster 6  consists of 4 parks with a high visitor density (> 15000 visitors per year per km^2), having a small area combined with several million visitors per year.

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded	Cluster Labels K-Means
0	Acadia	198.6	3537575	44.3500	-68.2100	Maine	17812.56	6
7	Bryce Canyon	145.0	2679478	37.5700	-112.1800	Utah	18479.16	6
14	Cuyahoga Valley	131.8	2096053	41.2400	-81.5500	Ohio	15903.29	6
31	Indiana Dunes	62.1	1756079	41.6533	-87.0524	Indiana	28278.24	6

Universities

Clustering based on venues close by

First, we will try to group the universities based on the venues in their surroundings. Doing so, we attempt to answer the question 'What is the typical neighbourhood of a university?' or, more precisely, 'Which venues can typically be found in the neighbourhood of a university?'

	Name	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	American University	Coffee Shop	Sandwich Place	Pizza Place	Bus Stop	Indian Restaurant	Mexican Restaurant	Grocery Store	Food Truck	Convenience Store	Park
1	Arizona State University (Tempe)	Pizza Place	Coffee Shop	Sandwich Place	Mexican Restaurant	Bar	American Restaurant	Burger Joint	Hotel	Deli / Bodega	Dessert Shop
2	Boston College	Pizza Place	Sushi Restaurant	Lake	Convenience Store	Chinese Restaurant	Café	Coffee Shop	Trail	Park	Golf Course
3	Boston University	Bakery	American Restaurant	Pizza Place	Gym / Fitness Center	Café	Park	Thai Restaurant	Hotel	Middle Eastern Restaurant	Trail
4	Brandeis University	Trail	Pizza Place	Bar	Donut Shop	Mexican Restaurant	Sandwich Place	Chinese Restaurant	Italian Restaurant	Indian Restaurant	New American Restaurant
...	...	...	...	...	...	...	...	...	...	...	...
116	Washington State University	American Restaurant	Yoga Studio	Café	Bar	Coffee Shop	Beer Store	Sushi Restaurant	Boutique	Mediterranean Restaurant	Mexican Restaurant
117	Washington University in St Louis	Italian Restaurant	Coffee Shop	American Restaurant	Sandwich Place	Thai Restaurant	Ice Cream Shop	Art Museum	Rock Club	Park	Bakery
118	Wayne State University	Coffee Shop	Pizza Place	American Restaurant	Chinese Restaurant	History Museum	Sandwich Place	Art Museum	Dog Run	New American Restaurant	Gift Shop
119	William & Mary	Hotel	History Museum	American Restaurant	Historic Site	Pizza Place	Gift Shop	Ice Cream Shop	Art Museum	Bar	Resort
120	Yale University	Pizza Place	Coffee Shop	Bar	Italian Restaurant	Indian Restaurant	Theater	Café	Bakery	Cocktail Bar	American Restaurant

Table 23 - Universities: The most common venues

Using Foursquare, we were able to retrieve up to 100 venues within 1500m radius from the universities. Table 23 shows the Top10 list of the most common venues per university.
I have tried to use K-Means to group the universities based on the Foursquare data, starting with 2 clusters, increasing the number of clusters and observing how the groups are changing as well as checking whether the resulting clusters can be characterized / interpreted reasonably:

- With 2 clusters, University of California in Merced (with only 4 types of venues and 5 venues in total: Lake (2x), Food, Laundromat, Diner) occupies its own cluster; all other universities are grouped together.
- With 3 clusters, University of California in Merced still occupies its own cluster. Further, there are 2 big clusters without any obvious structure. For example, various pairs of universities in Cluster 0 don't share any common venues in their Top10; various pairs of universities in Cluster 2 seem to only share a Coffee Shop in their Top10, but various universities in Cluster 0 also have a Coffee Shop in their Top10.

- With 4 clusters, University of California in Merced keeps to occupy its own cluster. But now, University of California in Santa Cruz (another one having only 4 venues, namely Trail, Tree, Convenience Store, Farm) also gets its own cluster. It is not easy to characterize the two big clusters: Cluster 0 seems to be dominated by Coffee Shops and Sandwich Places (but both of them can be found in Top3 of many universities in Cluster 1, too); while in Cluster 1, it's mostly Bar and Pizza Place (and both of them can be found in Top3 of many universities in Cluster 0, too).
- With 5 clusters, the two Californian universities named above still keep their own clusters. Now we have another small cluster, composed of Texas A&M University and New Mexico Institute of Mining and Technology, both having Mexican Restaurant, Hotel and Convenience Store in their Top10. In the two big clusters, there are several other universities having Mexican Restaurants as well as Hotel or Convenience Store in their Top10, but not all three of them. There is no obvious difference between the two big clusters.
- With 6 clusters, the same two Californian universities still keep their own clusters. However, there are now four big clusters. All universities in Cluster 0 have Hotel in their Top10, combined with American Restaurant and/or Coffee Shop (however, there are also universities with this combination in other clusters). In Cluster 1, the combination Pizza Place + Bar + Coffee Shop or Sandwich Place seem to dominate. In Cluster 2, most of the universities have at least one Sushi or Japanese or Mediterranean or Italian Restaurant or Ice Cream Place, but some pairs only share Hotel in their Top10. There is no obvious structure in Cluster 5, with many pairs of universities having completely different Top10 most common venues.
- When further increasing the number of clusters, the big clusters keep splitting and it is difficult to find any structure at least in some of them.

Conclusion: Using the naive approach, we were not able to find any reasonable clustering based on the most common venues of the universities.

Finding the best value of k for K-Means clustering

At first, we have chosen the value of k for the K-Means clustering 'manually' based on experimenting with k, trying to find a value giving results that could be interpreted reasonably. Two outliers were identified (universities with just a few venues, lacking the usual venues such as Coffee Shop or Café, Pizza Place, Sandwich Place or American Restaurant) but except of that, we have not found any obvious structure.

Next, we have tried to apply the elbow method and the silhouette score to find the optimal value of k. However, the results did not get better. There is no elbow at all (see Figure 26) and the silhouette scores are very low for all values of k (see Figure 27), both suggesting that there is no clustering in the data. The venues of the universities are simply too diverse and don't show enough similarity to allow any reasonable grouping.

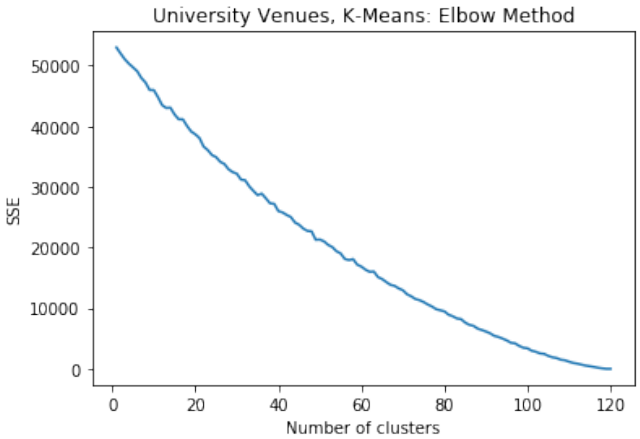


Figure 26 - Clustering universities based on their venues: Elbow method to find the best value of k

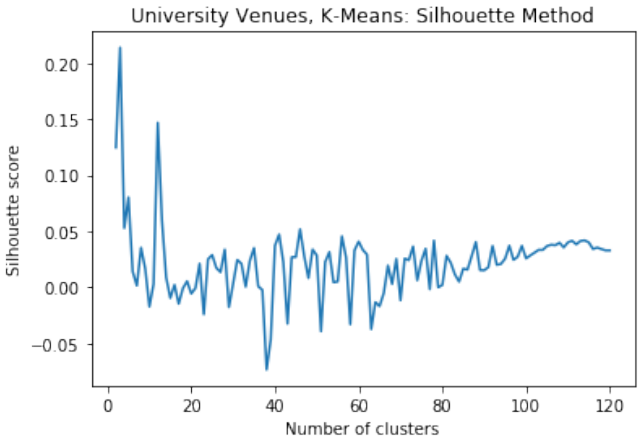


Figure 27 - Clustering universities based on their venues: Silhouette method to find the best value of k

Clustering based on the university scores

Further, I have checked whether the universities could be clustered based on their ranking scores. Again, both the elbow method and the silhouette score suggest there is no clustering in the data (see Figure 28 and Figure 29). In fact, this is not surprising taking into account that the goal of the ranking of the universities is not to show their similarity but to accent their differences.

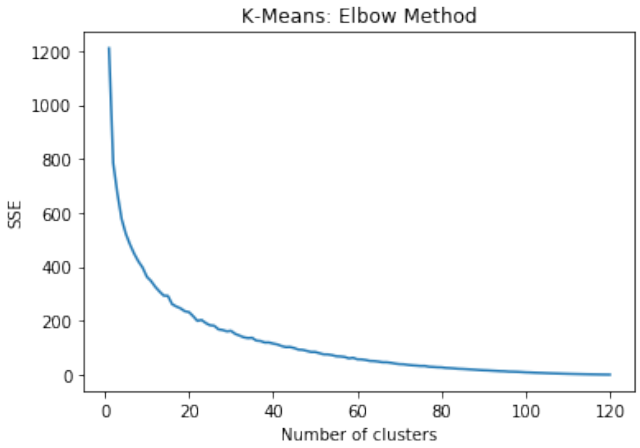


Figure 28 - Clustering universities based on their scores: Elbow method to find the best value of k

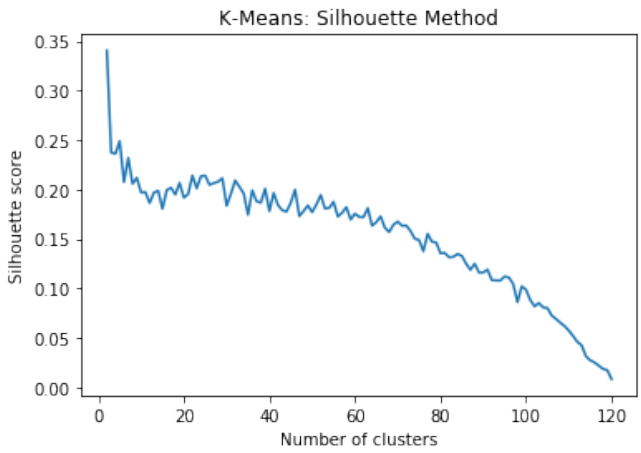


Figure 29 - Clustering universities based on their scores: Silhouette method to find the best value of k

We can conclude that neither the clustering of the universities on the venues close by, nor the clustering based on their scores has revealed any grouping in the data.

Family Wishlist

From a university to a national park: the distances

In this section, we will combine the wishes of all the family members to identify the most suitable location, starting with Kayleigh. Kayleigh is going to regularly travel between their (future) living place close by one of the universities and a national park; she wants to travel maximally 150km. We have to calculate the distances (in km) between the universities and the national parks to find out which pairs satisfy this requirement. For each university, we need to find the minimal distance to a national park and the name of the closest national park.

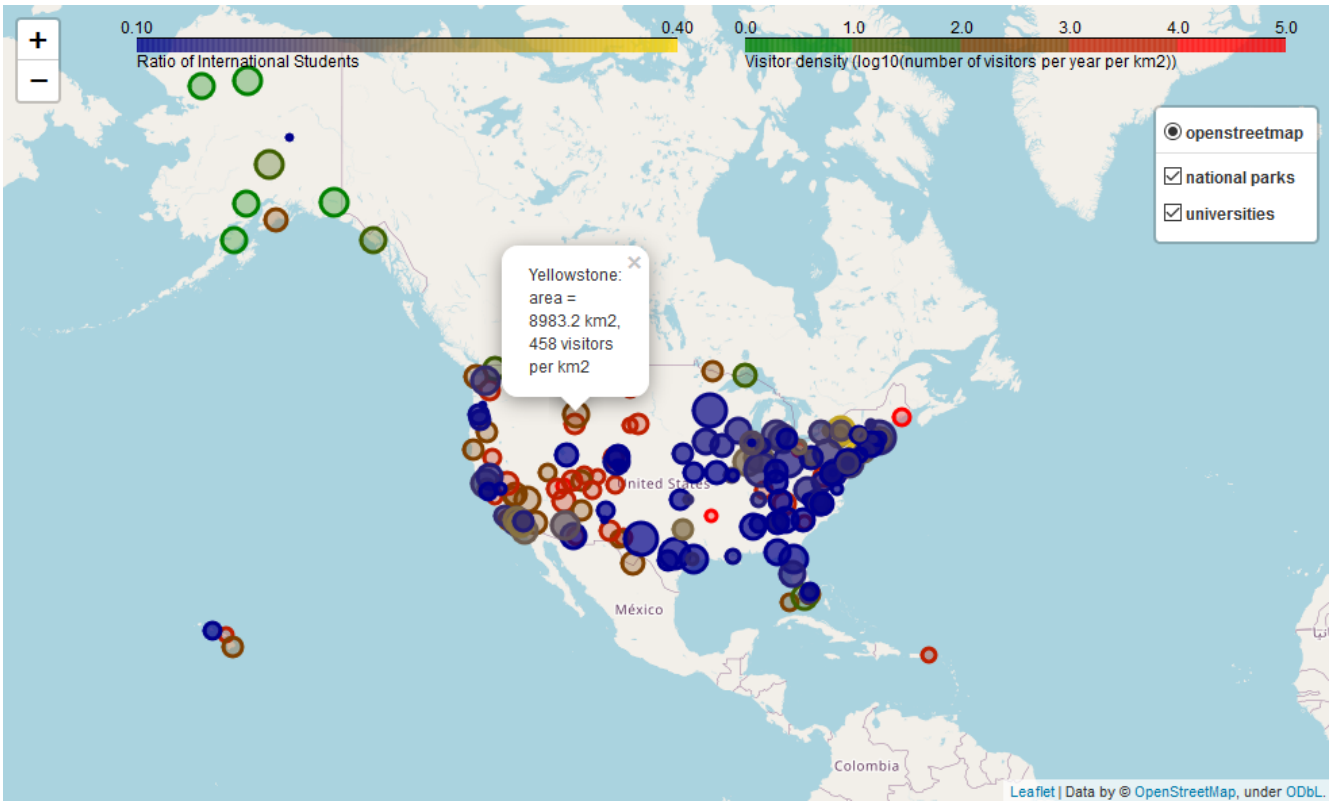


Figure 30 - National parks and Top500 US universities (corrected geographical coordinates)

Earlier, we have identified 61 national parks and 121 universities relevant to our problem. To keep with another wish of Kayleigh, we only consider those national parks that are not over-crowded. To define what we mean by over-crowded, let's recall the descriptive statistics of the national parks: see Table 24. From all these national parks, Hot Springs have the highest visitor density: about 67 thousand visitors per year per km^2; definitely too much, according to Kayleigh. She wants to limit this number to maximally 10 thousand visitors per year per km^2 because together with her clients, she needs to have time to photograph instead of spending it in traffic congestions (see Figure 32; photo courtesy of National Park Service [10]). The third quartile of the visitor density is 2.5 thousand. Hence, visitor density less than 10 thousand (visitors per year per km^2) is not overly limiting. Table 25 shows the parks that will be left out. Table 26 reveals the mutual distances between the universities and the

national parks as calculated using Geopy [11] based on the earlier retrieved geographical coordinates. There are 27 universities having a national park within the distance of 150km. In the next part, we have only considered these 27 universities, to limit the number of necessary Foursquare queries.

	Area in km2	Visitors per year	Crowded
count	61.000000	6.100000e+01	61.000000
mean	3476.854098	1.384678e+06	3749.629508
std	6686.383097	1.911755e+06	9686.138333
min	22.500000	9.591000e+03	0.310000
25%	261.800000	3.089620e+05	362.660000
50%	895.900000	6.449220e+05	971.890000
75%	3082.700000	1.663557e+06	2479.630000
max	33682.600000	1.142120e+07	66972.760000

Table 24 - National parks: Descriptive statistics (Gateway Arch excluded)

	Name	Area in km2	Visitors per year	Latitude	Longitude	State	Crowded
0	Acadia	198.6	3537575	44.3500	-68.2100	Maine	17812.56
7	Bryce Canyon	145.0	2679478	37.5700	-112.1800	Utah	18479.16
14	Cuyahoga Valley	131.8	2096053	41.2400	-81.5500	Ohio	15903.29
30	Hot Springs	22.5	1506887	34.5100	-93.0500	Arkansas	66972.76
31	Indiana Dunes	62.1	1756079	41.6533	-87.0524	Indiana	28278.24

Table 25 - National parks having visitor density higher than 10000 visitors per year per km²

Figure 31 - Acadia NP

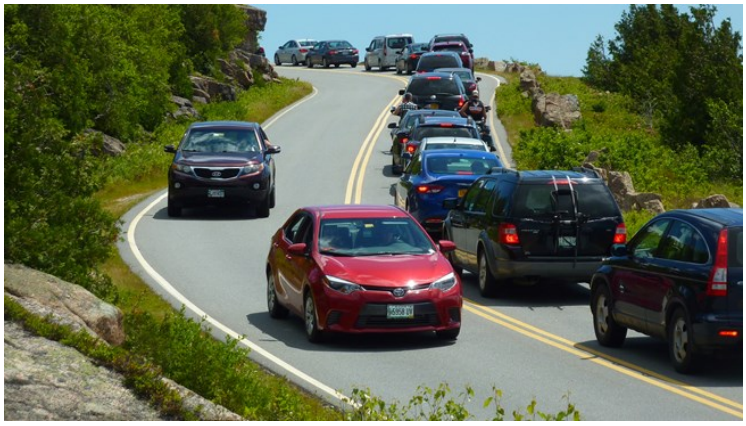


Figure 32 – Acadia NP, the road to the Cadillac Mountain summit

University	American Samoa	Arches	Badlands	Big Bend	Biscayne	Black Canyon of the Gunnison	Canyonlands	Capitol Reef	Carlsbad Caverns *	Channel Islands	...	Wrangell–St. Elias *	Yellowstone	Yosemite *	Zion	Closest NP	Min Distance to NP	Closest NP 2	Distance to NP 2	Closest NP 3	Distance to NP 3
California Institute of Technology	7720	918	1720	1509	3770	1055	864	771	1294	120	...	3442	1333	428	578	Channel Islands	120	Joshua Tree	209	Sequoia	259
Stanford University	7644	1114	1801	1976	4166	1274	1080	972	1721	454	...	2966	1262	240	808	Pinnacles	138	Yosemite *	240	Kings Canyon	329
Massachusetts Institute of Technology	11848	3258	2548	3224	2028	3110	3306	3408	3142	4275	...	5023	3167	4103	3598	Shenandoah	748	Congaree	1275	Great Smoky Mountains	1305
Princeton University	11529	2988	2324	2874	1706	2836	3033	3137	2810	3993	...	5003	2960	3848	3325	Shenandoah	376	Congaree	909	Great Smoky Mountains	935
Harvard University	11845	3255	2545	3222	2028	3108	3303	3405	3140	4272	...	5020	3164	4100	3596	Shenandoah	747	Congaree	1274	Great Smoky Mountains	1304
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Oklahoma State University	9518	1141	965	957	1992	981	1164	1272	809	2047	...	4182	1475	2001	1432	Great Sand Dunes	772	Carlsbad Caverns *	809	Guadalupe Mountains	858
Rensselaer Polytechnic Institute	11640	3043	2332	3035	1983	2896	3091	3193	2943	4061	...	4853	2952	3887	3384	Shenandoah	611	Great Smoky Mountains	1155	Congaree	1171
University of South Carolina-Columbia	10854	2603	2145	2168	930	2442	2633	2741	2190	3525	...	5218	2781	3474	2909	Congaree	33	Great Smoky Mountains	295	Shenandoah	558
University of Texas at San Antonio	9121	1426	1609	451	1879	1302	1415	1501	626	2026	...	4721	1968	2134	1589	Big Bend	451	Carlsbad Caverns *	626	Guadalupe Mountains	652
University of Tulsa	9614	1237	1012	1027	1908	1076	1261	1369	899	2149	...	4236	1548	2100	1531	Great Sand Dunes	869	Mammoth Cave	887	Carlsbad Caverns *	899

Table 26 - Distances between the universities and the national parks

Family-wishlist venues

Table 27 shows the numbers (per venue category, per university) of relevant family-wishlist venues in the region of all the universities selected in the previous step.

Name	Climbing Gym	Irish Pub	Rock Climbing Spot	Skating Rink	Zoo	Zoo Exhibit
American University	2	4	4	1	4	50
California Institute of Technology	2	0	7	2	0	1
Colorado School of Mines	2	0	8	0	0	0
Colorado State University, Fort Collins	0	1	3	3	0	2
George Mason University	1	0	1	1	0	0
George Washington University	2	11	4	5	3	49
Georgetown University	2	9	4	4	3	49
Howard University	1	9	4	3	4	49
Nova Southeastern University	2	0	9	0	0	1
Stanford University	3	0	7	1	0	1
The University of Tennessee-Knoxville	1	0	3	1	1	2
University of Arizona	0	0	1	1	1	3
University of California, Los Angeles	4	0	10	4	1	2
University of California, Merced	0	0	1	0	1	0
University of California, Riverside	1	0	5	1	1	0
University of California, Santa Barbara	0	0	8	3	1	1
University of California, Santa Cruz	1	0	2	0	0	1
University of Colorado Boulder	3	0	13	8	0	2
University of Denver	1	2	7	1	3	0
University of Maryland, College Park	2	0	5	1	0	0
University of Miami	3	0	6	0	0	0
University of Oregon	2	1	0	1	1	2
University of South Carolina-Columbia	0	0	2	0	2	11
University of Southern California	3	8	12	5	0	1
University of Virginia (Main campus)	1	1	0	1	0	0
University of Washington	4	5	6	3	3	48
Vanderbilt University	1	0	3	5	1	0

Table 27 - Foursquare: Numbers of retrieved family-wishlist venues, per category, per university

In the next section, we will describe how all the processed data will be combined to solve this business case.

Results: Final Selection

Now, we have collected all the data we need to choose the best locations according to the wishlist of the Cleirigh family. The universities selected so far belong to the Top 500 according to the *Times Higher Education World University Rankings 2020* [1] and for each of them, there is at least one national park within the distance of 150 km that is not over-crowded.
In the last step, we will filter the remaining universities so that in their region (specified by the radii given earlier), there is/there are:

- at most 2 climbing gyms, and
- at most 1 Irish pub, and
- at least one rock climbing spot, and
- at least one Zoo (at max. 10 km distance) or at least one Zoo Exhibit (at max. 5 km), and
- at least 1 skating rink.

For the Top 100 universities (taking into account the world ranking), we allow for one exception, and for Top 30, maximally two exceptions.
There are 11 universities fulfilling these requirements, as listed in Table 28. The locations of these universities and of the respective national parks are visualized in Figure 33. Out of these 11 universities, 7 of them fulfil all 5 requirements imposed on family-wishlist venues. For 3 of the universities, one of the requirements is not met: at Stanford, there are 3 climbing gyms in its vicinity; at University of California in Los Angeles, there are 4 climbing gyms; and University of Maryland has no Zoo / Zoo Exhibit close by. University of Washington only meets 3 requirements on family-wishlist venues (there are 4 climbing gyms and 5 Irish pubs close by) but it is compensated by its rank 26. On the other hand, the last two universities listed in Table 28 are ranked 300+ and both of them have a very low ratio of international students (less than 10%). Thus, we advise to choose one of the first 9 universities in this table.

To make the final decision, the Cleirigh family will visit the respective cities and their surroundings to get a feel for the region; this is a step that cannot be done on paper or online. Based on these visits, they will choose their next place to live.

	Name	Rank	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	...	International Outlook	Latitude	Longitude	Min Distance to NP	Closest NP	Distance to NP 2	Closest NP 2	Distance to NP 3	Closest NP 3	Family Score
0	California Institute of Technology	2	2240	6.4	0.30	0.34	94.50	92.1	97.2	97.9	...	82.5	34.135900	-118.126530	120	Channel Islands	209	Joshua Tree	259	Sequoia	5
1	Stanford University	4	16135	7.3	0.23	0.43	94.30	92.8	96.4	99.9	...	79.5	37.429070	-122.169780	138	Pinnacles	240	Yosemite *	329	Kings Canyon	4
2	University of California, Los Angeles	17	41066	9.4	0.17	0.54	86.80	83.1	88.6	97.3	...	64.1	33.927750	-118.372750	97	Channel Islands	229	Joshua Tree	279	Sequoia	4
3	University of Washington	26	45692	11.1	0.16	0.53	81.60	72.2	82.2	98.6	...	60.4	47.656510	-122.312090	96	Olympic	99	Mount Rainier	142	North Cascades	3
4	University of California, Santa Barbara	57	24089	27.6	0.16	0.52	69.60	47.9	63.6	96.4	...	68.1	34.416300	-119.847360	60	Channel Islands	247	Sequoia	258	Pinnacles	5
5	University of Maryland, College Park	91	33108	16.6	0.11	0.48	62.70	46.9	59.1	89.6	...	41.2	38.987850	-76.938900	133	Shenandoah	673	Congaree	690	Great Smoky Mountains	4
6	University of Arizona	104	39124	18.4	0.10	0.52	61.80	52.6	53.7	85.3	...	40.2	32.232100	-110.950950	43	Saguaro	333	Petrified Forest	439	Grand Canyon *	5
7	Vanderbilt University	116	12006	3.0	0.14	0.54	60.20	48.7	42.1	95.4	...	43.0	36.148620	-86.804880	131	Mammoth Cave	300	Great Smoky Mountains	610	Congaree	5
8	University of California, Riverside	251	22272	18.0	0.14	0.53	48.45	31.2	30.3	85.9	...	64.7	33.911310	-117.498430	149	Joshua Tree	178	Channel Islands	266	Death Valley	5
9	The University of Tennessee-Knoxville	301	25907	16.8	0.05	NaN	45.65	32.9	23.4	80.3	...	45.7	35.969736	-83.936213	49	Great Smoky Mountains	236	Mammoth Cave	377	Congaree	5
10	Colorado State University, Fort Collins	401	26014	16.3	0.06	0.53	40.55	27.8	25.3	64.1	...	38.4	40.578070	-105.081550	47	Rocky Mountain	318	Great Sand Dunes	318	Black Canyon of the Gunnison	5

Table 28 - Universities: Final selection

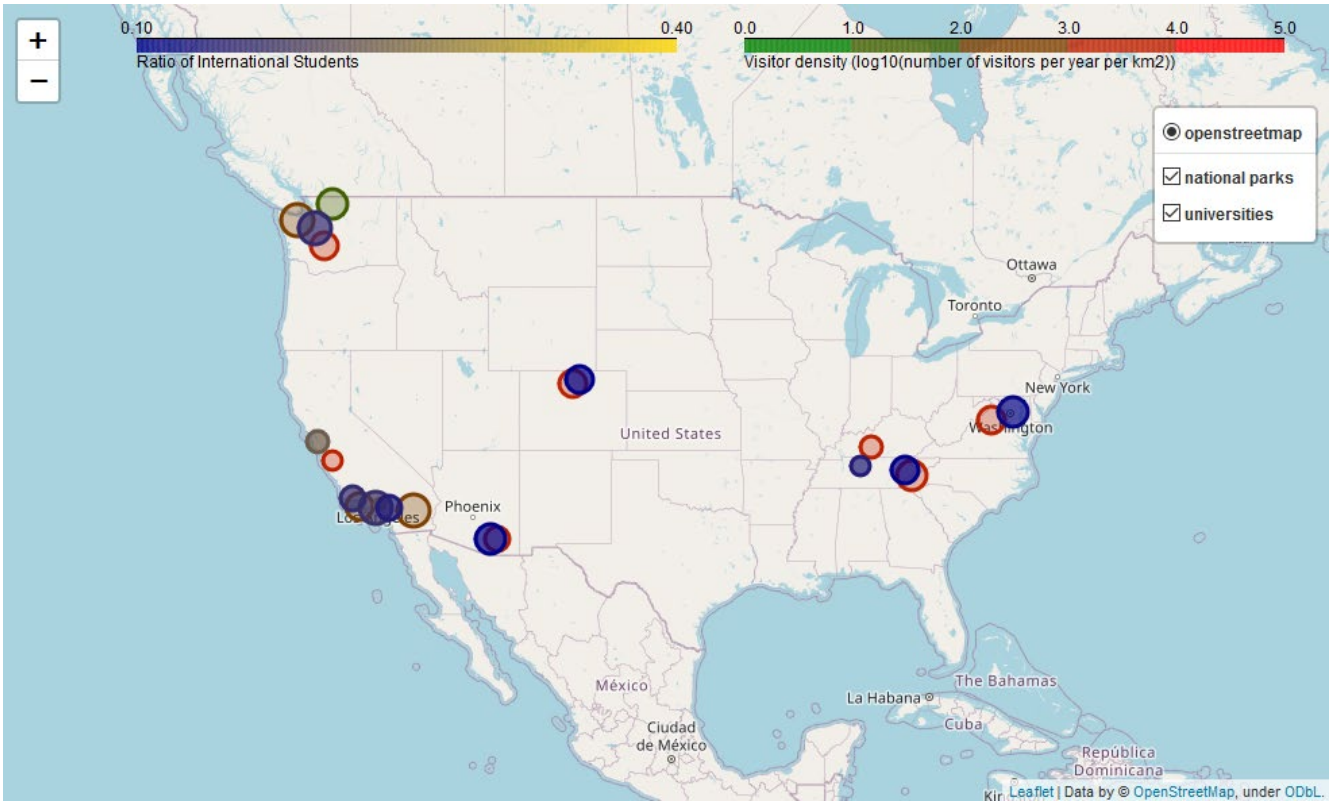


Figure 33 - Universities and national parks: Final selection

Discussion

In this study, we have shown that it is possible to combine various types of data, available via internet, to effectively identify locations fulfilling given requirements. Specifically, to identify the best future living place for the Cleirigh family, we have combined the data on U.S. national parks (their geographical coordinates, area and number of visitors per year) and the THE university ranking with Foursquare venue data to identify which top-scoring universities are within 150 km distance from a national park, have rock climbing and skating places as well as zoo exhibits in their surroundings and, on the other hand, do not have (too many) climbing gyms and Irish pubs close by.

Except of that, we have shown that careful assessment of the data using, e.g., exploratory statistics can – and should! – be used to identify anomalies in the data. Specifically, based on the descriptive statistics, we have identified a national park that is no *natural* park; further, based on the number of venues found via Foursquare, we were able to identify universities with incorrect location information.

On the other hand, it is obvious that some data needs to be interpreted carefully. Credibility of the data source is an important factor. For example, the datasets carefully put together by official institutions, such as the THE university ranking, tend to be correct. On the other hand, the rather random and unsystematic character implies that missing data and errors are quite usual in datasets consisting mainly of voluntary public submissions. For example, some venues retrieved via Foursquare are miscategorized as discussed on page 11 (e.g., climbing gyms categorized as *Rock*

Climbing Spot). In some cases, additional retrieved data can help to identify and correct these errors. Therefore, it is important to always know the structure of available data and based on this knowledge, to retrieve all relevant data. Less credible data should be verified, especially if any important decisions need to be made based on this data.

Conclusion

In this study, I have analysed the data on U.S. national parks and universities and identified the university-park pairs fulfilling the imposed requirements and having a mutual distance at most 150 km. I have calculated visitor density as an indicator of crowdedness of the national parks and based on this, identified an outlier. Further, I have clustered the national parks and based on the result, I was able to characterize similarities and differences between the groups of parks, concerning their area, number of visitors per year and visitor density. These results are useful for Kayleigh but they can also be useful, for example, for tour operators having experience with some of the parks who wish to extend their range of operations to other parks. On the other hand, I was not able to find any interpretable clustering of the universities, neither based on their scores, nor on the venues close by. However, the Foursquare data on venues was useful to identify incorrect geographical coordinates of the universities and to determine the most common venues in the vicinity of the universities. Further, I have used additional Foursquare data to identify places having the desired types of venues close by while rejecting places having too many potential business competitors for Brian and Shauna who are going to start their business from scratch. Finally, I came to a selection of 11 universities that fulfil the requirements of the family in the best possible way.

Future directions

This study and the consequent advice concerning the optimal next living place for Cleirighs could be further improved taking into account more various types of data. For example, Foursquare queries could be used to retrieve *Rating* and/or *Like Count* of the venues; then, we could drop low-rating venues from further analyses. Another example is temperature and humidity data that could be used to reflect possible preference of the family for a particular type of climate. It could be also interesting to conduct a similar study for European and Asian universities and compare the results.

References

[1] "Times Higher Education World University Rankings 2020," [Online]. Available: https://www.timeshighereducation.com/world-university-rankings/2020/world-ranking. [Accessed 21 4 2020].

[2] Wikipedia, "List of national parks of the United States," [Online]. Available: https://en.wikipedia.org/wiki/List_of_national_parks_of_the_United_States. [Accessed 21 4 2020].

[3] "Foursquare," [Online]. Available: https://foursquare.com/city-guide. [Accessed 19 4 2020].

[4] "Venue Search," Foursquare, [Online]. Available: https://developer.foursquare.com/docs/api-reference/venues/search. [Accessed 21 4 2020].

[5] "Venue Categories," Foursquare, [Online]. Available: https://developer.foursquare.com/docs/build-with-foursquare/categories. [Accessed 21 4 2020].

[6] Wikipedia, "Gateway Arch National Park," [Online]. Available: https://en.wikipedia.org/wiki/Gateway_Arch_National_Park. [Accessed 21 4 2020].

[7] "THE World University Rankings 2020: methodology," [Online]. Available: https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2020-methodology. [Accessed 19 4 2020].

[8] "scikit-learn: Machine Learning in Python," [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation.

[9] "Affinity Propagation," [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#affinity-propagation. [Accessed 21 4 2020].

[10] "U.S. National Park Service," [Online]. Available: https://www.nps.gov/index.htm. [Accessed 21 4 2020].

[11] "Python | Calculate Distance between two places using Geopy," GeeksForGeeks, [Online]. Available: https://www.geeksforgeeks.org/python-calculate-distance-between-two-places-using-geopy/. [Accessed 21 4 2020].