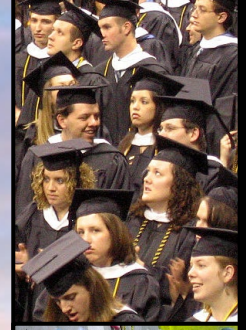
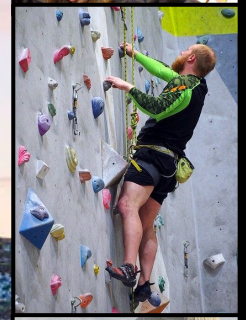
EXPAT ADVICE:
IRELAND TO USA


IBM Data Science Capstone Project

Lenka Čížková

Cleirigh family: Ireland to USA

- **Mother Kayleigh:** luxury photography workshops, wants to live close to a NP (max. distance 150 km, not crowded)
- **Father Brian:** climber, wants to live close to a rock climbing area, start a climbing gym (from scratch)
- **Son Sean:** bright student (last year secondary) & talented athlete, wants to go to a top US university
- **Daughter Orla:** primary school, loves animals and figure skating, wants to have a zoo and a skating rink close by
- **Grandmother Shauna**: cook and art lover, wants to have an art gallery close by, start an Irish pub (from scratch)

Data acquisition and cleaning: top universities

- data scraped from *THE World University Ranking 2000*: rank, name, country, no. of FTE students, no. of students per staff, % international students, female : male ratio, scores: overall, teaching, research, citation, industry income, international outlook
- 1397 universities worldwide
- filter: Top 500, US $\Longrightarrow$ 121 universities
- geographical coordinates via ArcGIS API

Data acquisition and cleaning: U.S. national parks

- data scraped from Wikipedia's *List of national parks of the United States*: name, location (extract latitude, longitude), date established as park, area, number of visitors per year
- 62 national parks
- calculate visitor density (per year per km^2)
- anomalous visitor density $\Longrightarrow$ Gateway Arch (no natural park) $\Longrightarrow$ drop

- calculate distances: 121 universities $\longleftrightarrow$ 61 national parks

Data acquisition and cleaning: Foursquare – venues close by univs

- venues within 1500m from the universities
- name, latitude and longitude of the respective university; name, latitude and longitude of the venue; venue category
- 121 universities $\Longrightarrow$ 9411 venues close by
- anomaly: 10 universities have less than 20 venues close by
- 7 of these universities: incorrect geographical coordinates
- after correction coordinates $\Longrightarrow$ 9849 venues close by

Data acquisition and cleaning: Foursquare – family wishlist

- search radius depending on type of the venue / planned activity:
  - Rock Climbing Spot, 30km
  - Climbing Gym, 5km
  - Skating Rink, 10km
  - Zoo, 10km; or Zoo Exhibit, 5km
  - Art Gallery, 5km
  - Irish Pub, 5km
- 2044 venues, partially misclassified (e.g., Climbing Gym as Rock Climbing Spot)
- cleaning; drop art galleries (enough everywhere) $\Longrightarrow$ 590 venues

Top-500 Universities: Exploratory stats

Why US university:

best score Ireland < mean score USA

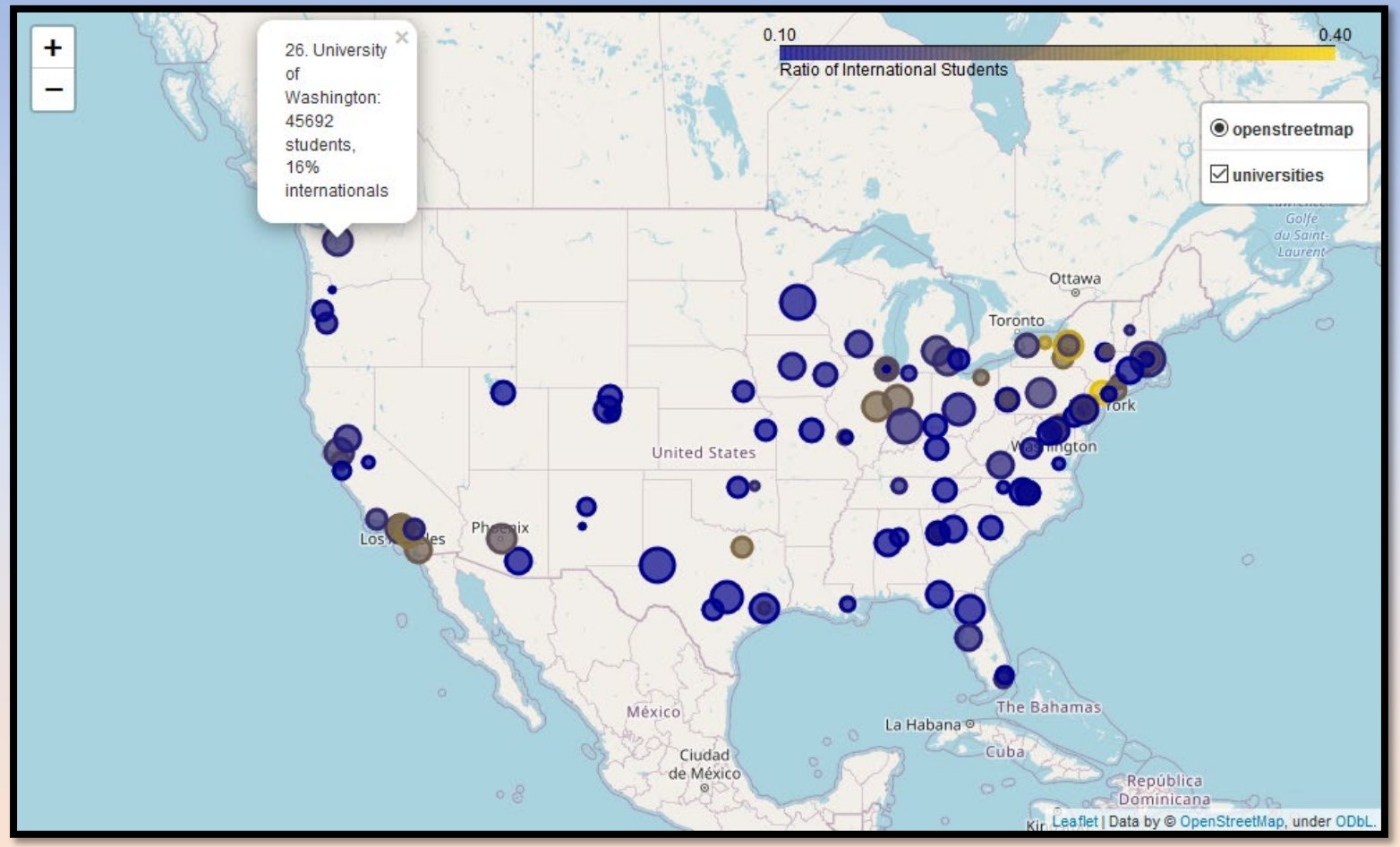
USA →

	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

Ireland →

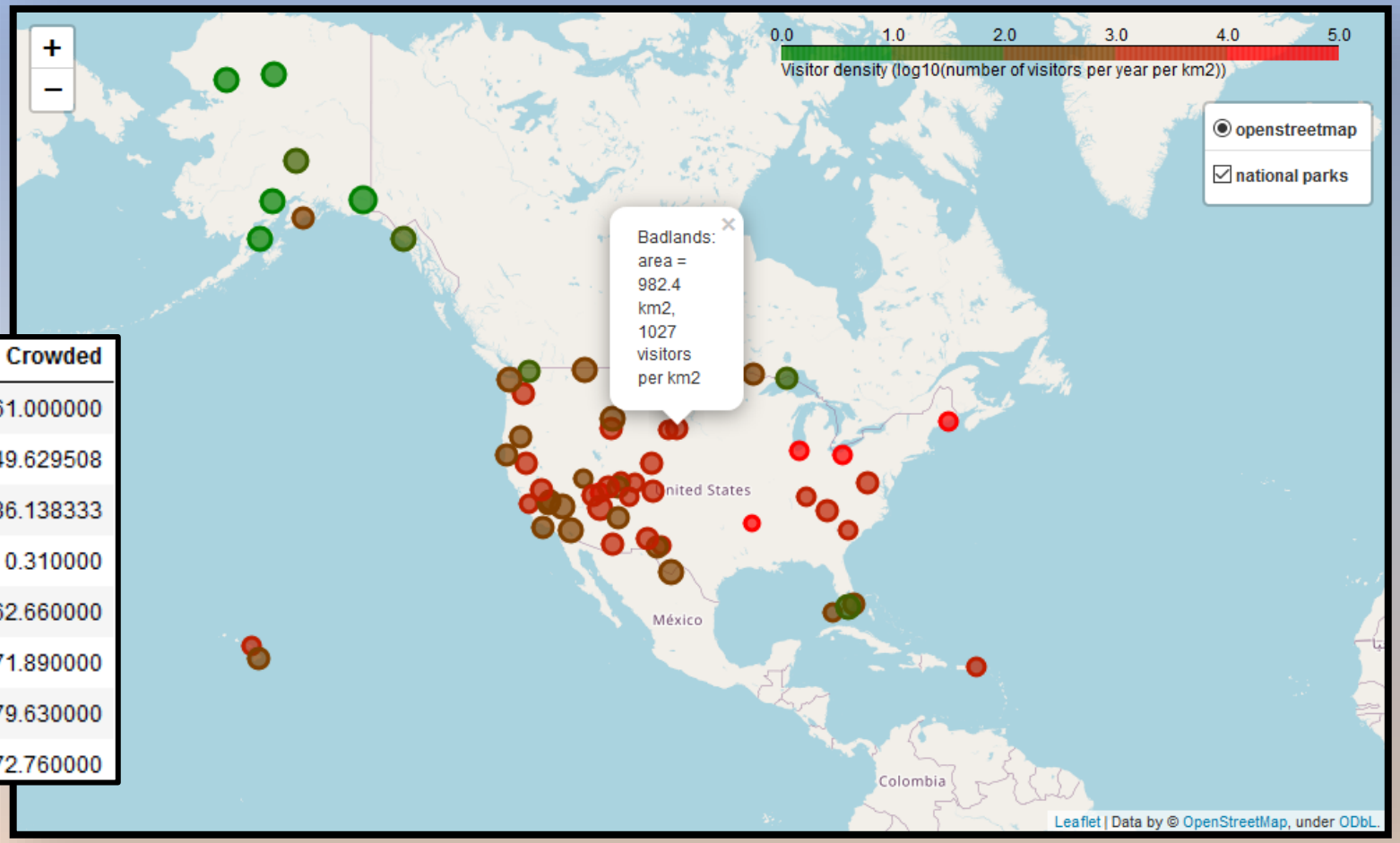
	No. of FTE students	No. of students per staff	International students ratio	Female Ratio	Overall	Teaching	Research	Citations	Industry Income	International Outlook
count	6.000000	6.000000	6.000000	4.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.00000
mean	13620.333333	23.166667	0.285000	0.575000	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

Top-500 Universities:
Exploratory stats

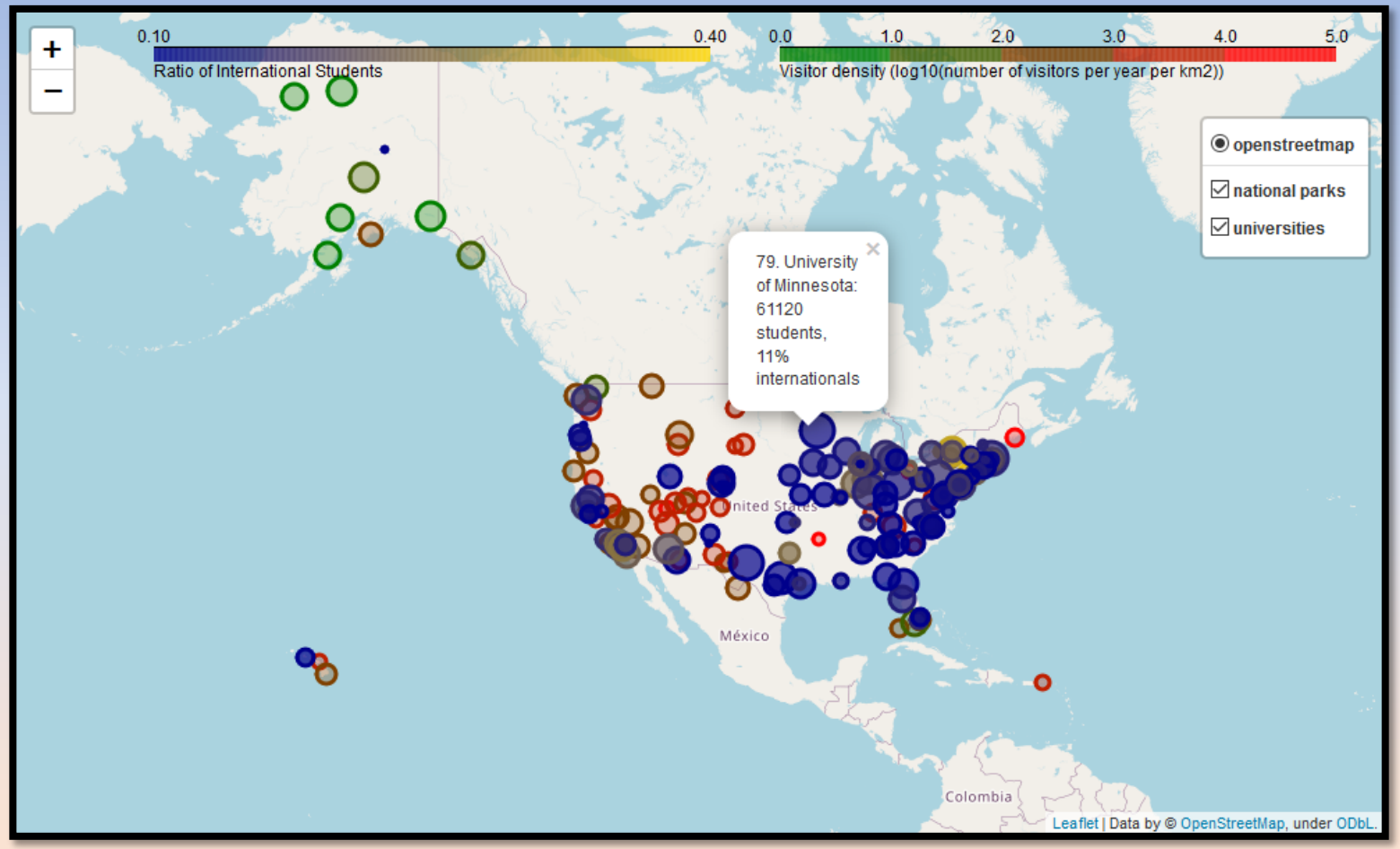


National parks: Exploratory stats

	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

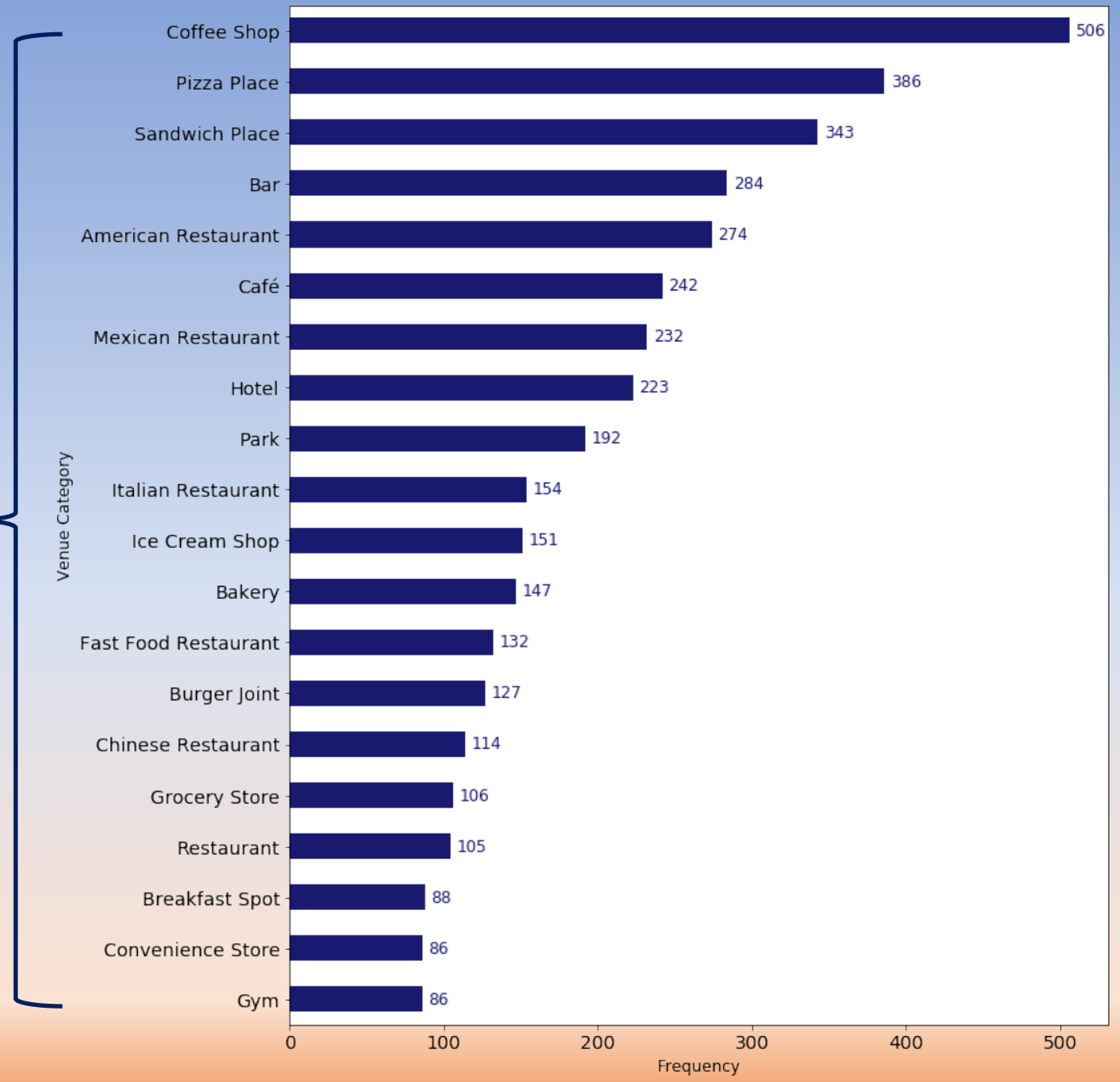


National parks and Universities

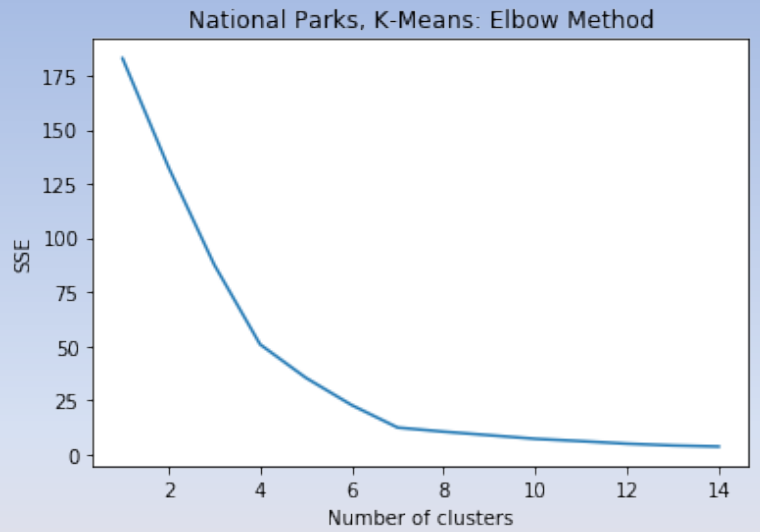


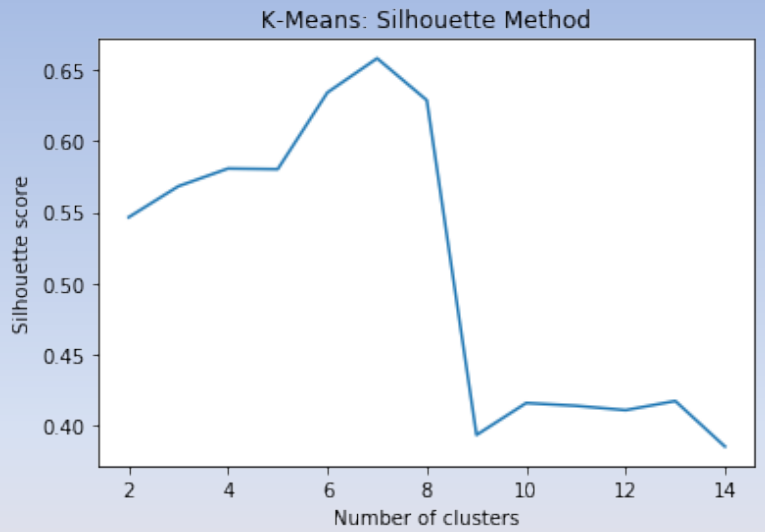
Foursquare venues: Exploratory stats

20 most common venues within 1500m from universities:

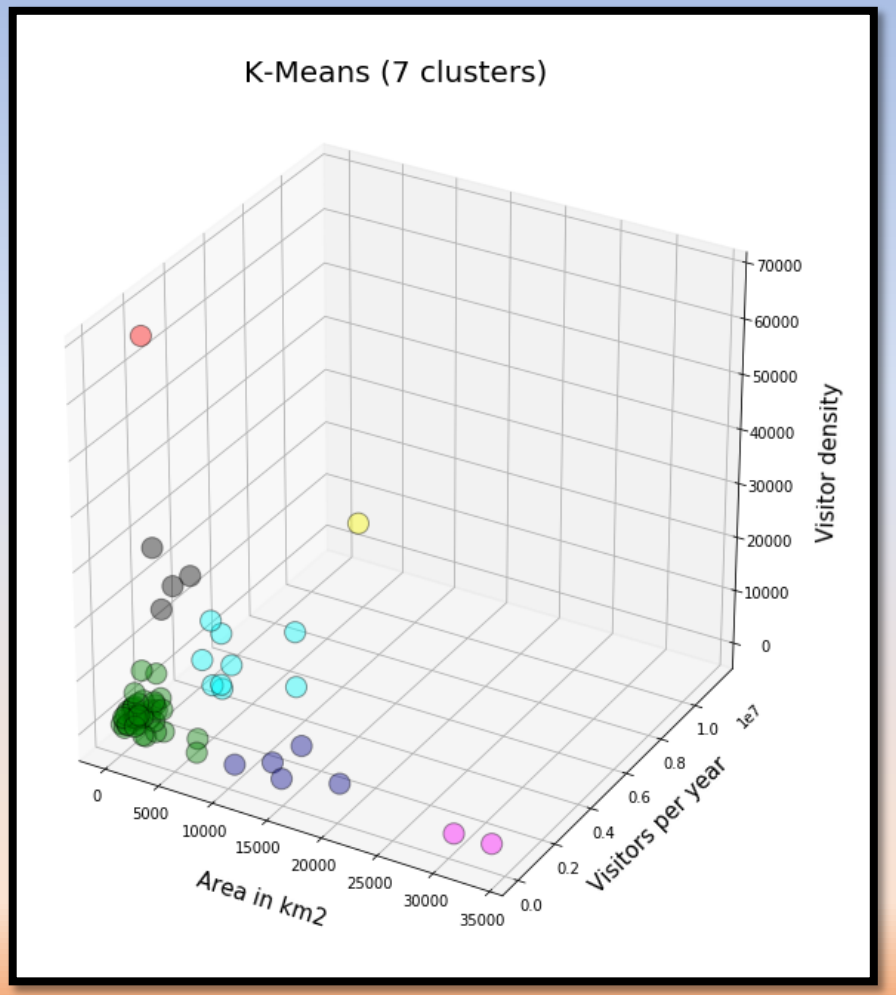


Clustering: National parks

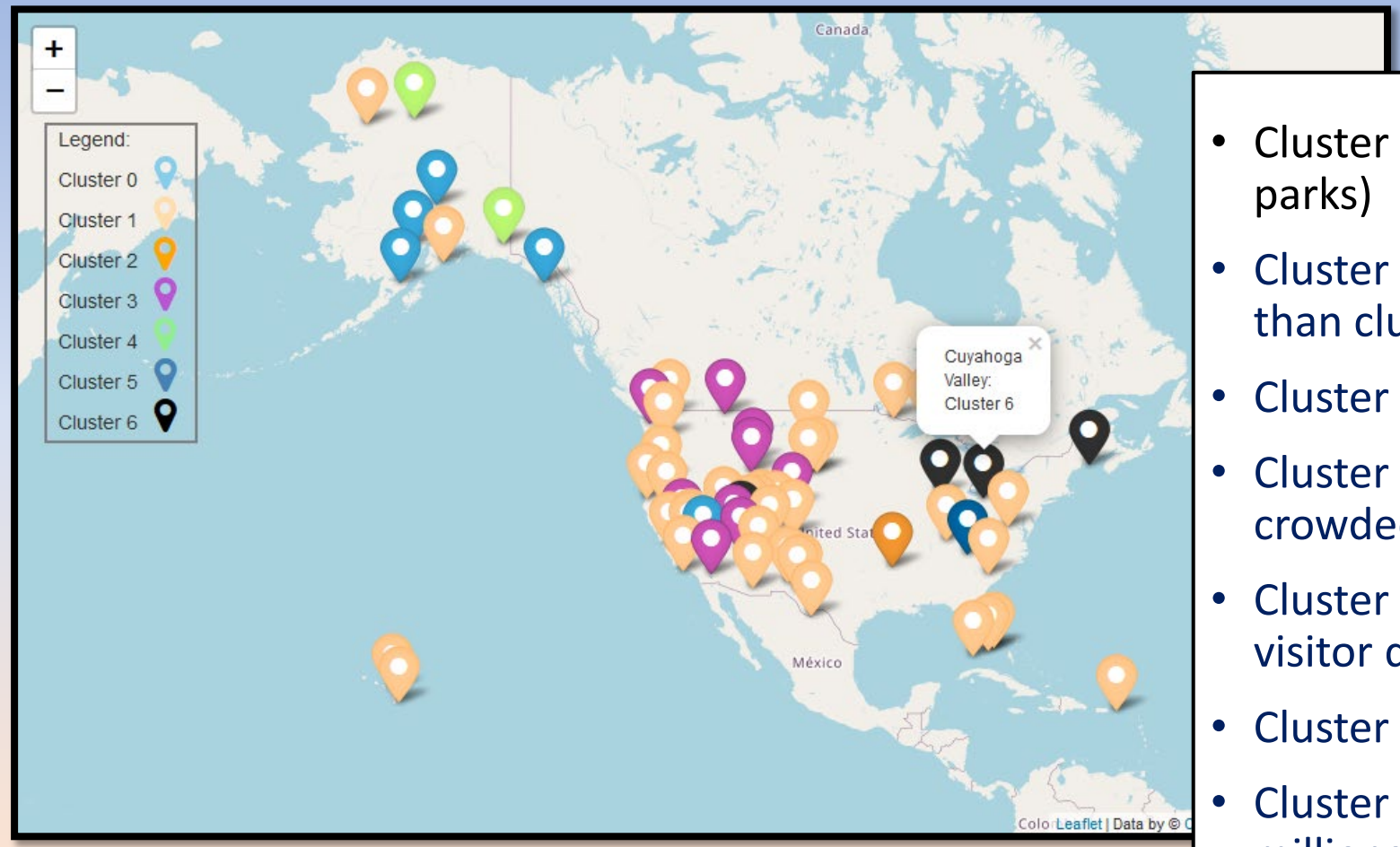




$\Rightarrow$ 7 clusters:

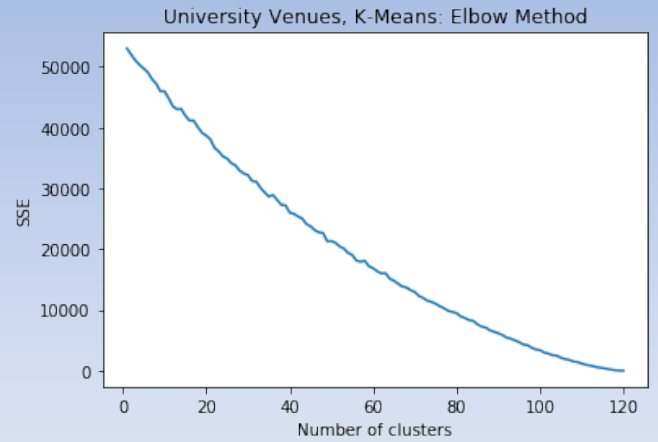


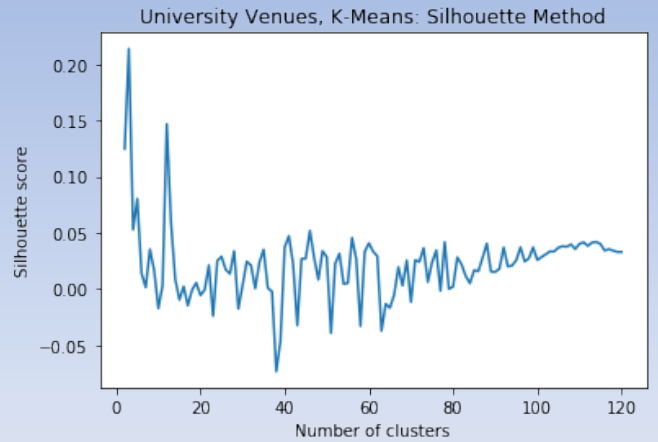
Clustering: National parks



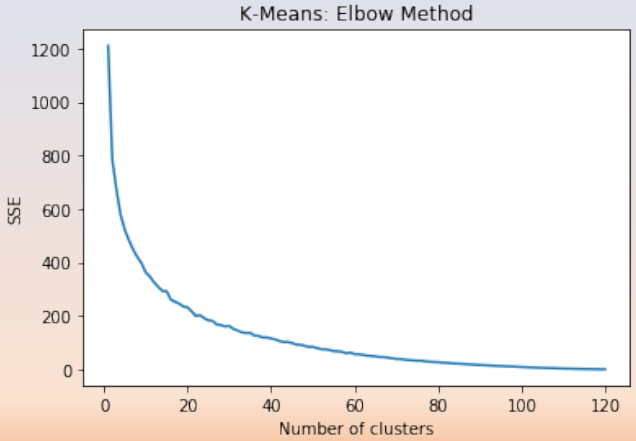
- Cluster 0: large area, low visitor density (5 parks)
- Cluster 1: relatively small area, less visitors than cluster 3
- Cluster 2: highest visitor density (1 park)
- Cluster 3: high number of visitors, less crowded than cluster 6
- Cluster 4: much larger than the rest, low visitor density (2 parks)
- Cluster 5: highest number of visitors (1 park)
- Cluster 6: high visitor density, small area & millions visitors

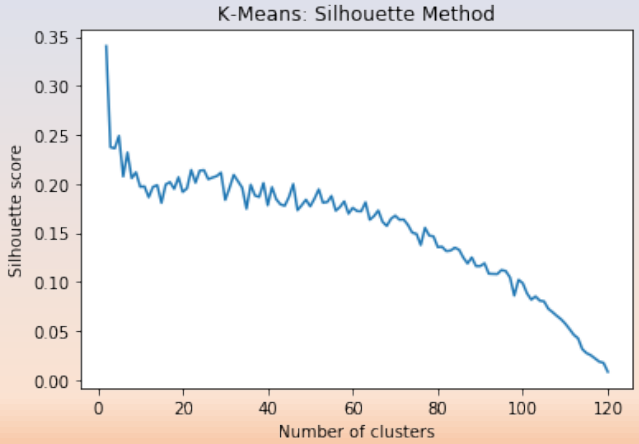
Clustering: Universities





based on venues





based on scores

⇒ no clustering!

Family wishlist, combined

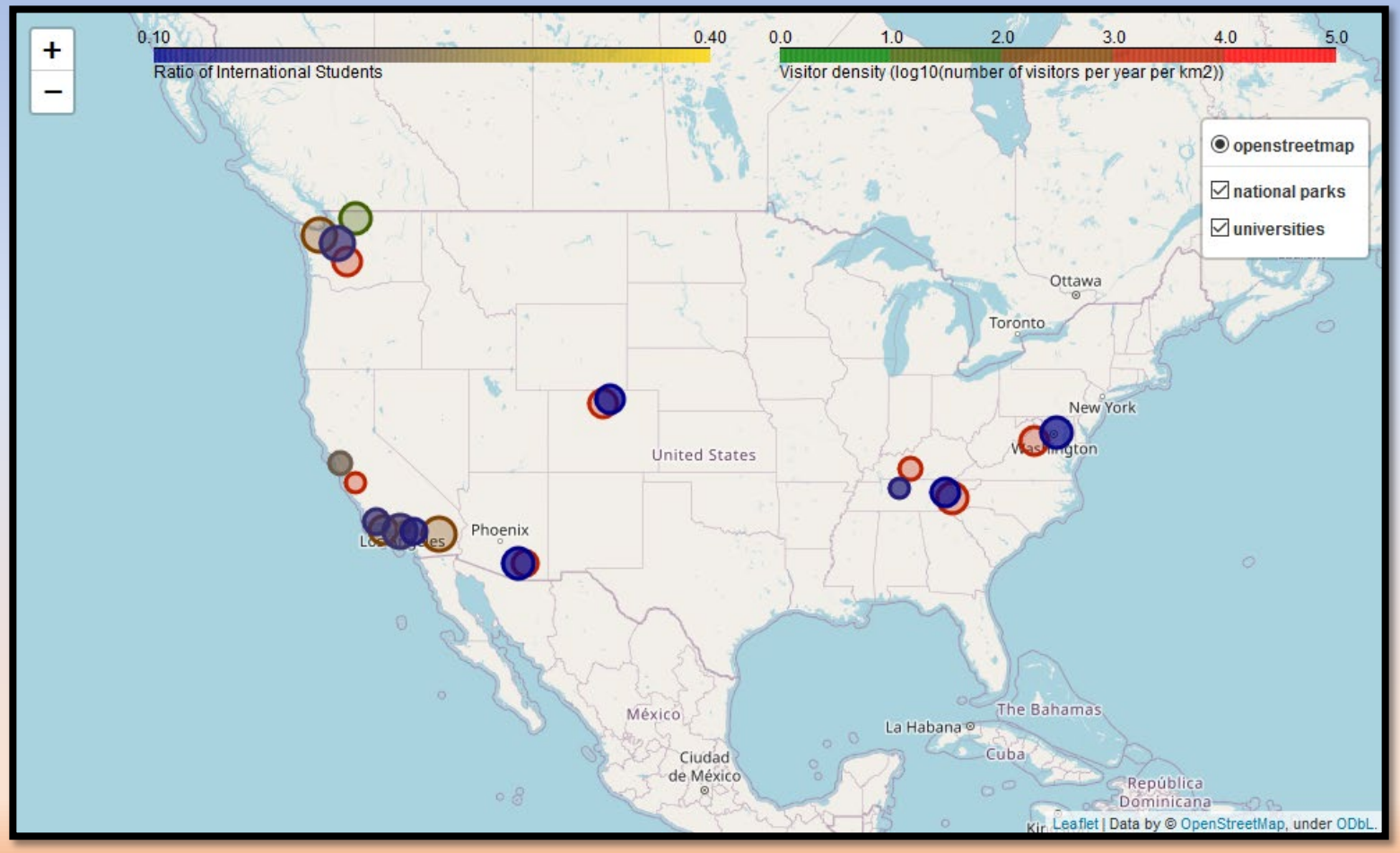
- max. distance university ⟷ park: 150km
- parks not over-crowded: max. 10,000 visitors per year per km^2
- venues – requirements:
  - at most 2 climbing gyms
  - at most 1 Irish pub
  - at least 1 rock climbing spot
  - at least 1 skating rink
  - at least 1 zoo (max. 10 km) or 1 zoo exhibit (max. 5 km)
  - (Top 100: max. 1 exception / Top 30: max. 2 exceptions)

Results – Final Selection

11 universities

	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

Results – Final Selection



Conclusion, Future direction

- analysed US national parks, universities and venues close by
- clustered national parks, characterized clusters (area, number of visitors, visitor density)
- identified most common venues close by the universities
- combined requirements of all family members
- selected 11 universities that best fulfil the requirements
- ideas:
  - venue rating: drop low-rated
  - climate data: filter on personal preferences of the family
  - similar study for Europe/Asia, compare the results