Data Science Capstone Project - The best city for a new business (Movie Theater)

1. Introduction to the business problem

To establish a new business it requires a comprehensive preparation. A sound preparation is key to maximize the chances for a successful and prospering development in a competitive environment. A vital ingredient is the right location for the business to be found and its well-examined assessment towards its selection.

The key addressee of this project is a founder of a new business and, thus, will be highly interested in an assessment in order to identify a well-suited location for an upcoming business. The area of the survey comprises the largest cities within the United States. The recommendation for a certain site should rely on the venues in the vicinity of a possible location. The key addressee is aiming at establishing a single (or a few) movie theater(s), but with the intention to expand to multiple theaters in a city. Therefore, the goal of the project is to identify the city and neighborhood in the US that is well suited to host the business, i.e. nearby venues that are supportive for the business together with a privileged site concerning its competitors.

In consequence, the approach is to identify an area that exhibits an enhanced density of entertainment establishments but at the same time shows a lower than average density of movie theaters, thus, displaying an undersupply of movie theaters. In a first step, the most promising cities are identified, which are investigated in a second step. In the latter, the best-suited neighborhoods/ZIP code areas of the chosen cities are investigated to identify the best-suited neighborhood/ZIP code area withing a certain city.

The key addressee of the study is, therefore, the founder of a new business who receives key information about possible locations. However, apart from identifying a well-suited location to establish the business, the projects approach and its outcome can further be used by the founder as supplementary documentation for the communication with different stakeholders such as investors and banks. Furthermore, the latter group, banks and investors themselves, are addresses for this study, as they can use as the project findings for preparatory work in the context of investments.

Hence, the project will address a topic that is of interest for various groups of persons with different perspectives on the result.

2. Data and Approach

2.1. Approach

2.1.1. Step1

As a first step, US cities are investigated regarding their population, their growth and their location to find a well-suited city for a new movie theater. Larger cities above a defined limit (in this case population $>$ 500,000) are identified. As a second selection criterion an increasing population is used. The five cities with the highest growth rates are chosen for further investigation.

2.1.2. Step1

As a second step, the ZIP codes of the selected cities are acquired and the venues and population for each of the different ZIP codes areas are requested. Using Foursquare.com with each zip code, the venues are downloaded and used to compute an index that is considered for the identification of the best-suited neighborhood. Thus, I request, of course, the number of movie theaters in the area and supplement this information with the numbers for coffee shops and bars, that will be considered as representative for the social life in the area. Subsequently, a comparison is conducted for the selected cities and their neighborhoods that depict a higher than average density of such venue categories (coffee shops and bars) showing a vivid public life and a good customer base. Finally this is compared to their movie theater density. The final evaluation results in a recommendation for the best-suited city and corresponding neighborhood to host the upcoming business.

2.2. Data

2.2.1. Overview of the required data

1. List of cities and their population:
`https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population`

2. Zip codes for cities in the USA:
`zip_code_database.csvfromhttps://www.unitedstateszipcodes.org/zip-code-database`

3. List of US state abbreviation (opt.):
`https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations`

4. Detailed information of a zip code area:
`http://www.city-data.com/zips`

5. Venue data from Foursquare.com:
`https://foursquare.com`

3. Methodology

3.1. Index of the code

Part 1 - Select US city	1.1	Download list of US cities
	1.2	1st selection of US cities
	1.3	Selection of the top5 US cities
Part 2 - Select neighborhood	2.1	Zip codes for the top5 US cities
	2.2	Venues for zip code via Foursquare.com
	2.3	Analysis of the Neighborhoods

3.2. Code section - Part1

3.2.1. Code section 1.1

For the selection of the US cities to investigate, first the list of cities is downloaded from `https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population`. This

list encloses the population data from the census in 2010 and an estimate for the population in 2018. From these a population growth during that period is computed and already present in the table. In the course of this report, when I talk about population growth, I refer to this column of the table, which is labeled as Change [%]. An extract of the table is shown in Fig.1, displaying the five largest cities in the US regarding their total population.

	Num	City	2018 estimate	2010 census	Change [%]	Area [km^2]	Pop. Density [/km^2]
0	1	New York	8,398,748	8,175,133	+2.74	780.9	10,933
1	2	Los Angeles	3,990,456	3,792,621	+5.22	1,213.9	3,276
2	3	Chicago	2,705,994	2,695,598	+0.39	588.7	4,600
3	4	Houston	2,325,502	2,100,263	+10.72	1,651.1	1,395
4	5	Phoenix	1,660,272	1,445,632	+14.85	1,340.6	1,200

Fig. 1.— Extract of the data of US cities received from wikipedia.

3.2.2. Code section 1.2

Concerning the numbers from this table that are required for the analysis, i.e. the entries for the population from the census in 2010 and the estimate for 2018 as well as the change of the population has been converted to integer type.

Subsequently, cities exceeding a population of 500,000 are selected, ordered by their population growth and coordinates for the remaining 35 cities are appended using the geocoder library. The top of the list including the five cities with the largest growth are shown in Fig.2. The complete list of the 35 largest US cities can be found in Fig.8 in Sec.7

3.2.3. Code section 1.3

To finish Part 1 of the analysis, the five cities that exhibit the highest growth rate and exceed the limit of 500,000 residents are chosen for further investigation. These are displayed in the table in Fig.2 as well as on the map in Fig.3.

	Num	City	2018 estimate	2010 census	Change [%]	Area [km^2]	Pop. Density [/km^2]	Latitude	Longitude
0	18	Seattle	744955	608660	22.39	217.0	3,245	47.60357	-122.32945
1	11	Austin	964254	790390	22.00	809.9	1,170	30.26759	-97.74299
2	13	Fort Worth	895008	741206	20.75	888.1	962	32.75095	-97.33086
3	19	Denver	716492	600158	19.38	397.0	1,746	39.74001	-104.99202
4	16	Charlotte	872498	731424	19.29	791.0	1,064	35.22286	-80.83796

Fig. 2.— Details of the five best-suited cities for further investigation.

3.3. Code section - Part2

3.3.1. Code section 2.1

To prepare the analysis of the selected cities, the zip codes of the US are taken from `https://www.unitedstateszipcodes.org/zip-code-database/`. The table includes the relevant information of the zip codes: their corresponding city, state and location coordinates; the first entries in the table are displayed in Fig.4. I have extracted the zip codes for each of the five cities that are selected as described in Sec.3.2.3, but excluded the zip code types that are not 'Standard' (Standard, PO Box and Unique are present in the list). The list had further to be cleaned from cities that have the same name, but are located in different federal states of the US. The mode-function has been used to identify the most frequent zip codes of the considered cities. Furthermore, I have worked on the latitude and longitude taken from the database and replaced them by the more precise values that I obtained using the geocoder library. This information is stored in csv-files for each city for later use ($CityName\_PostalCodeInformation.csv$).

3.3.2. Code section 2.2

Using the acquired information described above, I requested the venue information via Foursquare.com to obtain information on the number of the venues 'Movie Theaters', 'Coffee Shops' and 'Bars' for each zip code in an area around the its coordinates with radius of 1 km. This represents roughly a 15 min walk, which is considered as an acceptable walking distance for and by the customer from a bar or coffee shop to the movie theater and vice-versa. Due to the limitation of daily calls of the Fousquare-account, I have stored also this information in a csv-files for each city and for later use ($CityName\_VenueInformation.csv$). This csv-files store the zip code, type of zip code, city, state, coordinates and the number

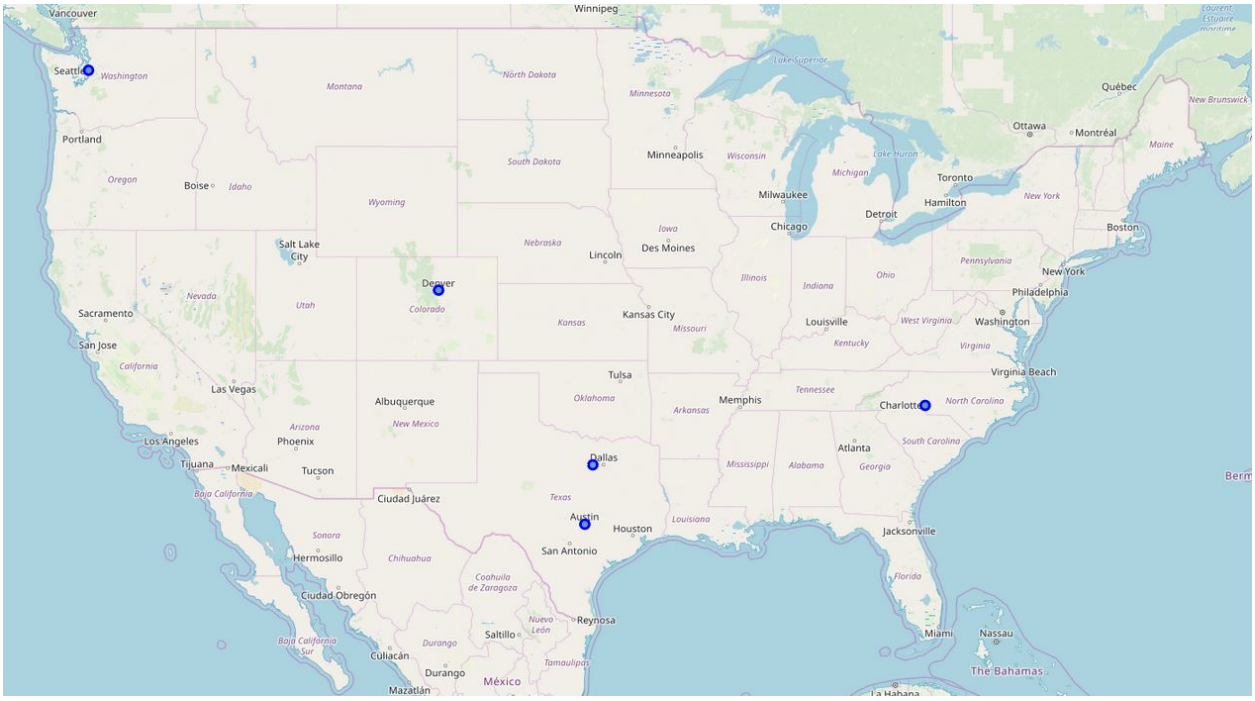

Fig. 3.— Map of the USA with the five selected cities marked as blue dots.

of Movie Theaters, Coffee Shops and Bars. I have supplemented this information with the population of each zip code area taken from `http://www.city-data.com/zips/` using the BeautifulSoup library. that is the sum of the number of coffee shops and bars divided by the number of movie theaters.

3.3.3. Code section 2.3

For the analysis of the attractiveness of a neighborhood or city, I have established an 'evaluation index',

$$Evaluation\ index = \frac{(coffee\ shops + bars)}{movie\ theaters} \tag{1}$$

The sum of the number of coffee shops and bars is considered as representative for the social life and a vivid environment. The division by the number of movie theaters further aims at the attractiveness to establish a movie theater, i.e. finding an area with a relatively low number of present theaters.

Hence, if the sum of coffee shops and bars rises, so does the evaluation index. On the

	zip	type	primary_city	state	latitude	longitude
0	501	UNIQUE	Holtsville	NY	40.81	-73.04
1	544	UNIQUE	Holtsville	NY	40.81	-73.04
2	601	STANDARD	Adjuntas	PR	18.16	-66.72
3	602	STANDARD	Aguada	PR	18.38	-67.18
4	603	STANDARD	Aguadilla	PR	18.43	-67.15

Fig. 4.— Extract of the data for all US zip codes received from wikipedia.

other side, if the number of movie theaters is reduced, the evaluation index is increasing. In consequence, a high evaluation index represents a high attractiveness to establish the business within the considered area. The five zip codes with the highest evaluation index of each city have been stored in a single dataframe (Fig.9 in Sec.7). A plot depicting the evaluation index against the population and number of movie theaters of these zip code areas has been created (Fig.5). It also includes the SVM model that separated well-suited

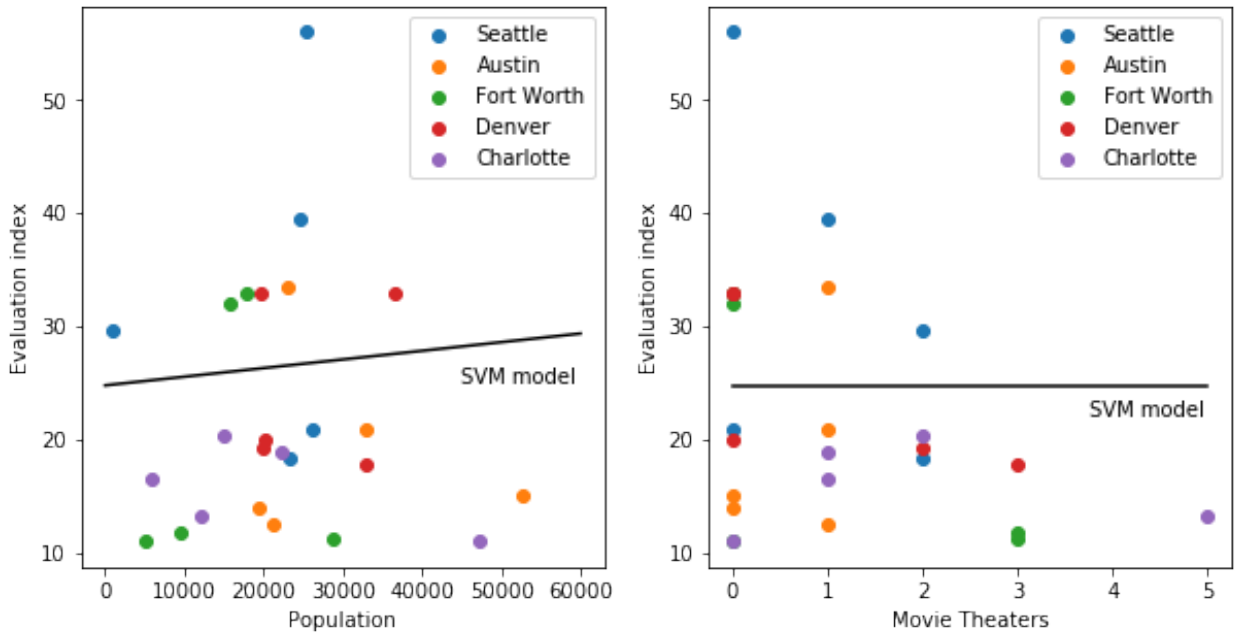

Fig. 5.— Scatter plot showing the five highest evaluation indices for each of the five investigated cities as a function of the population in the area and as a function of the number of movie theaters present in the region. The SVM model is shown as the solidblack line.

neighborhood from these with lower atractiveness. A box plot is shown in Fig.6 depicting the distribution of the evaluation index among the different cities and, hence giving a good overview for the goodness of the citites. A map of the zip code areas of the best-suited city with a circle radius representing the evaluation index is finally shown Fig.7.

4. Results

The selection procedure to obtain the most adequate cities in the US has been described in Sec.2. The limit of 500,000 residents reduces the number of cities to 35. The second condition of the growth rate shows a growth of almost 20% between 2010 and 2018 with values between 19.3% and 22.4% for the five mostly growing cities in the US during that period. The sixth most adequate city has growth rate that is already 2.5% less than the fifth highest in the sample. Therefore, I selected this five cities, Seattle, Austin, Fort Worth, Denver and Charlotte for further investigation (see also Fig.2 and 3).

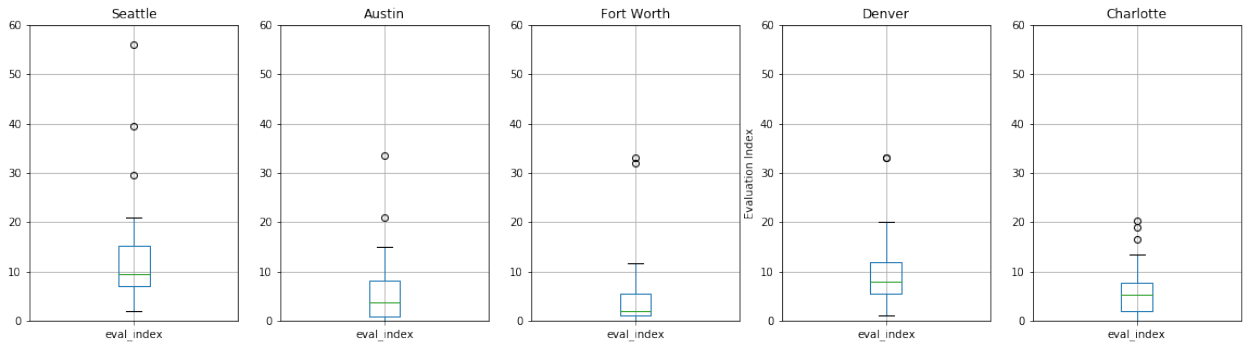

Fig. 6.— Boxplot for the evaluation indices of the zip areas in the five considered cities. The plots display the advantageous environment that is present in Seattle.

The venue information obtained from Foursquare.com and the subsequent calculation and analysis of the evaluation index has revealed that Charlotte has only below average evaluation indices among the top 25 zip code areas (top 5 in evaluation index for each of the 5 cities). Zip code areas with an indices above the average evaluation index have been found in the other cities, in Seattle (3/5), Fort Worth and Denver (2/5) and Austin (1/5). Furthermore, the values for the parameter on which the boxplot in Fig.6 rely are listed in Table4

City	Lo. Quartile	Median	Up. Quartile	3rd Highest	2nd Highest	Highest
Seattle	7.04	9.56	15.13	29.67	39.50	56.00
Austin	0.88	3.75	8.10	15.00	21.00	33.50
Fort Worth	1.00	2.00	5.50	11.75	32.00	33.00
Denver	5.50	8.00	11.97	20.00	33.00	33.00
Charlotte	2.00	5.17	7.71	16.50	19.00	20.33

Table 1: Table showing the lower and upper quartile as well as the median and the three highest values of the evaluation index for the five investigated cities.

The Support Vector Machine (SVM) analysis further supports this results and the SVM results and approach can later be used to repeat the analysis for other venue categories or use the present one to get a quick answer on the attractiveness of other zip code areas in relation to the considered ones.

5. Discussion

From the Sec.4 we see, that Seattle shows the most promising parameters for a new Movie Theater. It shows the highest growth rate together with the zip code areas having the highest evaluation index. Furthermore, the numbers for the venues in the Seattle area show the highest frequency for reaching the limit for Foursquare request (50). In such cases the evaluation index delivers only a lower limit of its real number. Thus, the area is even better suited than visible in this analysis. The analysis further displays more 'above average' indices and the map in Fig.7 shows that the considered five best zip code areas in Seattle are well distributed over the city area which support a later expansion. In addition, all relevant parameters, displayed in the boxplot, such as lower and upper quartile, median and the highest evaluation indices are all supportive for Seattle to be the city recommended for hosting the movie theater(s).

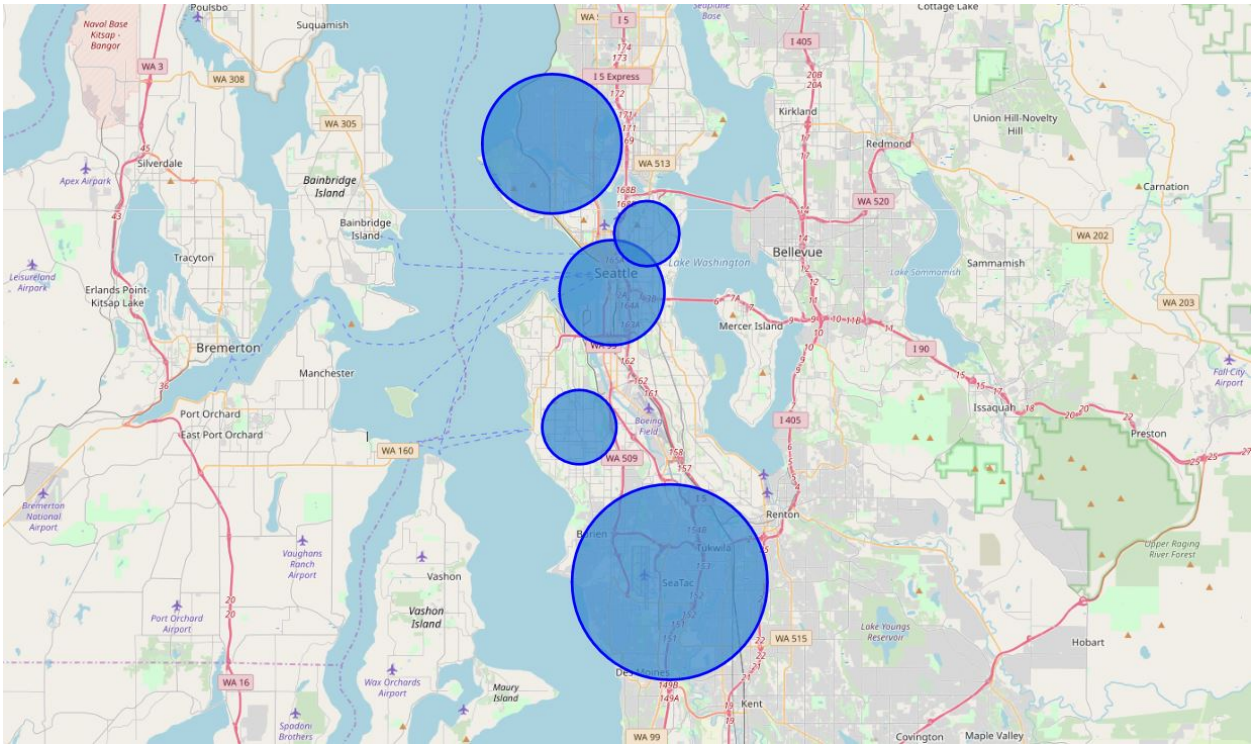

Fig. 7.— Map of the selected city Seattle and the five assessed neighborhoods. They are well distributed among the city supporting a later expansion.

6. Conclusion

The analysis of the US cities has revealed Seattle as the best-suited city for establishing a new movie mheater together with the aim of an expansion towards the city. The parameter are all in favor of this city and even the expansion plans can be well following in the largest city of Washington. The first theater should open in the area with the zip code 98188 and, if further theaters are to be opened, they should be located next in the area with 98107 as zip code. The work further offers a model to support expansion plans and quickly evaluate other zip code areas. Therefore, this work offers a clear statement towards a location for the investment and is, thus, supportive for proposals for investments. Not even a good location has been spotted, the results further emphasizes to establish that new business rather than only recommending a city.

7. Appendix

	Num	City	2018 estimate	2010 census	Change [%]	Area [km^2]	Pop. Density [/km^2]	Latitude	Longitude
17	18	Seattle	744955	608660	22.39	217.0	3,245	47.60357	-122.32945
10	11	Austin	964254	790390	22.00	809.9	1,170	30.26759	-97.74299
12	13	Fort Worth	895008	741206	20.75	888.1	962	32.75095	-97.33086
18	19	Denver	716492	600158	19.38	397.0	1,746	39.74001	-104.99202
15	16	Charlotte	872498	731424	19.29	791.0	1,064	35.22286	-80.83796
19	20	Washington	702455	601723	16.74	158.2	4,304	38.89037	-77.03196
34	35	Mesa	508958	439041	15.92	357.2	1,357	33.41704	-111.83146
6	7	San Antonio	1532233	1327407	15.43	1,194.0	1,250	29.42458	-98.49461
4	5	Phoenix	1660272	1445632	14.85	1,340.6	1,200	33.44825	-112.07580
13	14	Columbus	892533	787033	13.40	565.9	1,520	39.96199	-83.00275
20	21	Boston	694583	617594	12.47	125.1	5,381	42.35866	-71.05674
8	9	Dallas	1345047	1197816	12.29	882.9	1,493	32.77815	-96.79540
26	27	Oklahoma City	649021	579999	11.90	1,570.3	407	35.47203	-97.52107
24	25	Portland	653115	583776	11.88	345.8	1,851	45.51179	-122.67563
23	24	Nashville	669053	601222	11.28	1,232.6	536	36.16784	-86.77816
3	4	Houston	2325502	2100263	10.72	1,651.1	1,395	29.76058	-95.36968
27	28	Las Vegas	644644	583756	10.43	348.1	1,818	36.17193	-115.14001
11	12	Jacksonville	903889	821784	9.99	1,935.8	455	30.33147	-81.65622
14	15	San Francisco	883305	805235	9.70	121.5	7,170	37.77712	-122.41964
7	8	San Diego	1425976	1307402	9.07	842.3	1,670	32.71568	-117.16171
35	36	Sacramento	508529	466488	9.01	253.6	1,953	38.57944	-121.49085
9	10	San Jose	1030119	945942	8.90	459.7	2,231	37.33865	-121.88542
33	34	Fresno	530093	494665	7.16	296.3	1,762	36.74084	-119.78552
16	17	Indianapolis	867125	820445	5.69	936.3	914	39.76691	-86.14996
1	2	Los Angeles	3990456	3792621	5.22	1,213.9	3,276	34.05349	-118.24532
21	22	El Paso	682669	649121	5.17	665.1	1,030	31.75916	-106.48749
32	33	Tucson	545975	520116	4.97	597.8	888	32.22155	-110.96976
5	6	Philadelphia	1584138	1526006	3.81	347.6	4,511	39.95222	-75.16218
28	29	Louisville	620118	597337	3.81	682.5	903	38.25489	-85.76666
0	1	New York	8398748	8175133	2.74	780.9	10,933	40.71455	-74.00714
31	32	Albuquerque	560218	545852	2.63	487.4	1,147	35.08423	-106.64905
25	26	Memphis	650618	646889	0.58	822.1	794	35.14976	-90.04925
2	3	Chicago	2705994	2695598	0.39	588.7	4,600	41.88425	-87.63245
30	31	Milwaukee	592025	594833	-0.47	249.2	2,388	43.04200	-87.90687
29	30	Baltimore	602495	620961	-2.97	209.5	2,934	39.29058	-76.60926
22	23	Detroit	672662	713777	-5.76	359.5	1,871	42.33168	-83.04800

Fig. 8.— List of the 35 largest US cities and their considered data.

	zip	type	primary_city	state	latitude	longitude	Movie Theaters	Coffee Shops	Bars	Population	eval_index
0	98188	STANDARD	Seattle	WA	47.44519	-122.28984	0.0	38.0	18.0	25269	56.000000
1	98107	STANDARD	Seattle	WA	47.66951	-122.37919	1.0	29.0	50.0	24525	39.500000
2	98134	STANDARD	Seattle	WA	47.59346	-122.33368	2.0	39.0	50.0	859	29.666667
3	98106	STANDARD	Seattle	WA	47.52462	-122.35845	0.0	8.0	13.0	26216	21.000000
4	98112	STANDARD	Seattle	WA	47.62364	-122.30724	2.0	36.0	19.0	23413	18.333333
5	78702	STANDARD	Austin	TX	30.26352	-97.72638	1.0	17.0	50.0	22941	33.500000
6	78705	STANDARD	Austin	TX	30.29006	-97.74504	1.0	26.0	16.0	32878	21.000000
7	78741	STANDARD	Austin	TX	30.23666	-97.72794	0.0	6.0	9.0	52726	15.000000
8	78734	STANDARD	Austin	TX	30.34850	-97.96376	0.0	5.0	9.0	19269	14.000000
9	78703	STANDARD	Austin	TX	30.27802	-97.75816	1.0	11.0	14.0	21113	12.500000
10	76104	STANDARD	Fort Worth	TX	32.73414	-97.33845	0.0	10.0	23.0	17855	33.000000
11	76164	STANDARD	Fort Worth	TX	32.78621	-97.35597	0.0	2.0	30.0	15835	32.000000
12	76102	STANDARD	Fort Worth	TX	32.75041	-97.33563	3.0	11.0	36.0	9523	11.750000
13	76107	STANDARD	Fort Worth	TX	32.75204	-97.35929	3.0	9.0	36.0	28703	11.250000
14	76155	STANDARD	Fort Worth	TX	32.82317	-97.05273	0.0	4.0	7.0	5036	11.000000
15	80210	STANDARD	Denver	CO	39.67675	-104.95945	0.0	13.0	20.0	36503	33.000000
16	80212	STANDARD	Denver	CO	39.76567	-105.04615	0.0	12.0	21.0	19700	33.000000
17	80237	STANDARD	Denver	CO	39.63250	-104.89660	0.0	8.0	12.0	20158	20.000000
18	80223	STANDARD	Denver	CO	39.71644	-104.99405	2.0	20.0	38.0	19956	19.333333
19	80205	STANDARD	Denver	CO	39.75517	-104.98266	3.0	21.0	50.0	33012	17.750000
20	28203	STANDARD	Charlotte	NC	35.21398	-80.85774	2.0	13.0	48.0	15032	20.333333
21	28209	STANDARD	Charlotte	NC	35.16916	-80.84899	1.0	5.0	33.0	22155	19.000000
22	28204	STANDARD	Charlotte	NC	35.21243	-80.83416	1.0	12.0	21.0	5930	16.500000
23	28202	STANDARD	Charlotte	NC	35.22660	-80.84135	5.0	30.0	50.0	12165	13.333333
24	28205	STANDARD	Charlotte	NC	35.21324	-80.79635	0.0	2.0	9.0	47174	11.000000

Fig. 9.— Data of the five best suited zip code areas for each of the top five cities in the US including their population and evaluation index.