FINAL ASSIGNMENT – CAPSTONE PROJECT

What makes US cities successful?
A study of Livability.com's Top-100 places to live in 2020

1. Introduction

City officials have always been interested in improving the living conditions of their fellow citizens, while making their cities more appealing for national and international tourism. While the reasons behind this could lie within the selfless interest in making others' lives better, there are usually more "selfish" reasons for this including the interest in getting re-elected for the position, the generation of more tax revenue, and the recognition attained when doing an outstanding job by media and other officials nationwide.

The availability of more and more data to quantify the impact of several factors in the living standards of the population, assess the efficiency of implemented measures and ordinances, and to compare the overall performance to other cities across the nation have proven a very valuable asset that continues to be mined by the most modern administrations.

This report will study data from the 100 cities identified in www.livability.com as 2020's Top-100 Best Places to Live in the US, and try to extract information that could provide guidance for city officials to replicate the success formulas of these "top places to live". In particular, the lessons learned through the analysis of readily available data will be applied to provide recommendations to three major cities in the State of Oklahoma (Oklahoma City, Tulsa and Stillwater) in order to become the first Oklahoman city to make it into the list of Top-100 places to live in the US.

2. Data

The initial data frame was collected manually from a report put together by the website livability.com (https://bit.ly/2K4df3x). The report is a result of the statistical analysis of polling data from 1,000 US cities, a study performed in collaboration with Ipsos. This report provides a series of numerical indicators of different aspects of city live (civics, demographics, economy, education, health, housing and infrastructure), calculated using raw data from a wide array of sources, such as the US Census Bureau, US Department of Housing and Urban Affairs, USDA, and many more.

Using these 7 individual indicators, an overall "Livability Score" (LivScore) is calculated and used to rank the top 100 places to live. There is not a proper description explanation of either how the individual indicators are calculated, or how these are used to come up with the final score.

A sample of one of the 100 pages is shown next:

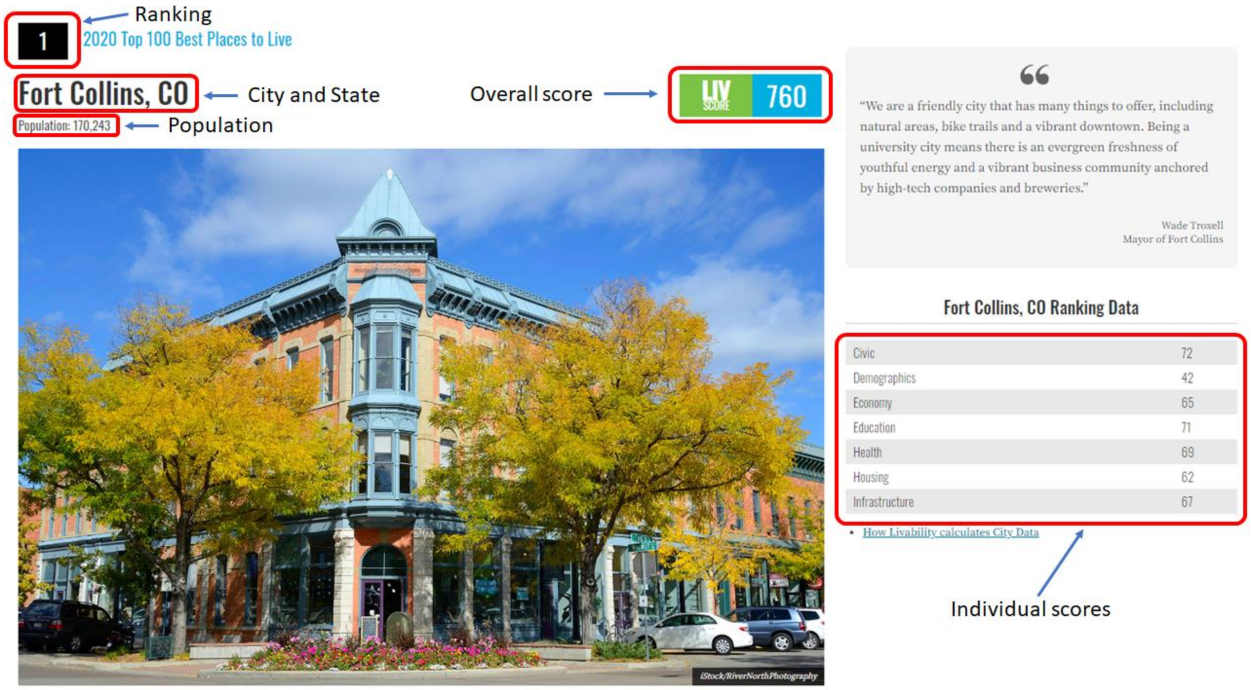

Figure 1. Page of the 1st-ranked city according to Livability.com's TOP 100 Best Places to Live: Fort Collins, Colorado

Having the information spread through 100 individual pages, it was considered faster to browse through them manually collecting the few bits of information required from each page than to develop a code to automatize the data collection. The information was collected in a comma-separated-values (csv) file titled "Capstone project livability ratings.csv", which will be available in the Jupyter Notebook. The first five rows of data along with the column names can be seen in the following image:

Capstone project livability ratings.csv - Notepad

File Edit Format View Help

```
RK,City,State,Population,LIV,CI,DE,EC,ED,HE,HO,IN
1,Fort Collins,Colorado,170243,760,72,42,65,71,69,62,67
2,Ann Arbor,Michigan,119980,759,84,67,52,73,69,66,65
3,Madison,Wisconsin,259680,758,75,58,61,62,78,47,55
4,Portland,Maine,66215,744,66,54,64,67,71,56,71
5,Rochester,Minnesota,118935,727,66,42,49,60,79,55,72
```

Figure 2. 5 first rows of the csv file containing the raw data

While a few insights can be extracted from this raw data, the core of this project will deal with analyzing these cities using data from the different categories of venues in their city centers using the FourSquare API.

FINAL ASSIGNMENT – CAPSTONE PROJECT

The Livability report has given us the names of 100 cities throughout the United States that are worth investigating further. We can feed these names to a Python program that will first extract relevant data about the types of venues present in these cities, and then will help us interpret and extract conclusions from this data.

How do we intend to do this? The general categories of the venues present in each of the cities will be extracted using the popular FourSquare API, and their relative frequency used to identify various "types of cities" (i.e. a nightlife hotspot, an outdoors paradise, an artistic center, etc.). The different general categories, as described in the FourSquare website, are the following:

a. Food: Restaurants, food trucks, delis, etc.
b. Drinks: Places to meet and have a drink, such as breweries, clubs, pubs, etc.
c. Coffee: Places to sit and have a warm beverage, such as a café.
d. Shops: Places to buy different products, such as a shopping mall, an antique shop, etc.
e. Arts: Music venues, museums, galleries, etc.
f. Outdoors: Parks, trails, gardens, etc.
g. Sights: Landmarks such as statues, monuments, etc.

A program will be built that extracts and stores the total number of venues of a particular type for each of the cities using the FourSquare API. This data set will then be evaluated to determine the different shares of each type of venue on each of the cities, and how they compare to each other (using a clustering technique such as K-Means).

Then, the case study entities of Oklahoma City, Tulsa and Stillwater will be evaluated to determine which type of city (i.e. which cluster) they belong to using another machine learning technique (K-Nearest Neighbors). Finally, we will determine which of the Top 100 cities are more similar to the three Oklahoman cities using Euclidean distance.

With the insights accumulated throughout this process, a series of recommendations will be drafted that could help the three Oklahoman cities to climb up in the ranking, maybe reaching one year the Top 100 cities ranking.

The following image shows a sample of the type of data that will be available after collection and cleanup:

	City	Food venues	Drinks venues	Coffee venues	Arts venues	Outdoors venues	Sights venues	Population	LIV	CI	DE	EC	ED	HE	HO	IN	Latitude	Longitude
0	Fort Collins, Colorado	53	26	10	18	18	74	170243	760	72	42	65	71	69	62	67	40.550853	-105.066808
1	Ann Arbor, Michigan	92	38	34	28	37	100	119980	759	84	67	52	73	69	66	65	42.268157	-83.731229
2	Madison, Wisconsin	65	32	24	12	38	85	259680	758	75	58	61	62	78	47	55	43.074761	-89.383761
3	Portland, Maine	100	64	35	36	47	100	66215	744	66	54	64	67	71	56	71	43.661028	-70.254860
4	Rochester, Minnesota	35	17	13	4	14	32	118935	727	66	42	49	60	79	55	72	44.023439	-92.463018

Figure 3. Top 5 rows of the cleaned Python data frame used in this study.