Grant Hile, Kat Molostvova
Dr. Golshan
May 8, 2019

CIT460 - Final Project Report

Introduction

Finding a new place to live is always a daunting task, and even more-so if you have the entire United States to select from! Given numerous cities and towns, that vary in near every possible way, selecting a new home from scratch would be difficult. Thankfully, with the wealth of public data that has become accessible in recent years, and use of R and data science techniques, we can narrow down the options.

Cities

We acquired our initial cities data from simplemaps.com. The table available there included city name, state, county id, lat & long, population, and a set of zip codes, all of which proved useful in later analysis. To select our cities of interest, we sorted by population, and selected the 3 most populous cities from each state.

Economic Data

We downloaded a table containing county level economic data from the US census. We decided to look at mean commute time, median family income, unemployment rate, and the percent of the population with health insurance (public or private).

Taxes

Wallethub.com has information on average state + local taxes on a median income family for each state, which we scraped and imported to use in our ranking. This information is statewide, but should still be relatively accurate for each city.

Housing Data

We obtained county level housing data from the US Census. We decided to look at the median cost of mortgage and rent for each city. To account for the differences in income in different cities, we calculated the ratio between median family income and median mortgage/rent for use in our metric. Some major cities are experiencing housing crisis' so this seems a good

Libraries

We collected 2016 library information from the Institute of Museum and Library Services, using a table they have available for download. We counted the number of libraries per city, and then calculated the number of people each library served. We had some issues working with this data - neither going by city, or the more complex to work with set of zip codes seemed fully accurate.

Crime

We collected crime statistics from the FBI, using a table they have available for download. This information has a number of fields Not Available, that we had to work around in our analysis. The data consists of raw counts of crimes reported, so to ensure that our rating of criminality would be

applicable across different city sizes, we calculated the rate of reported crime per person in each city. It should be noted that the FBI has warnings posted regarding any conclusions drawn through use of this data, due to differing reporting standards, and other issues.

Our Combined 'Happiness' Metric

Thinking about the things we valued in a city, and the importance of our different variables, we developed a system of assigning either 3, 6, or 9 points to each variable. The percentile rank of each city was calculated for each variable, then multiplied by that variable's number of points to calculate the city's score. Total score was then summed, giving us a total point value for each city out of 51 possible points.

'Happiness' Points

Metric	**Points**
Population	9
Median family income	9
% with health insurance	6
Unemployment rate	6
Mean commute time	3
Median mortgage compared to income	3
Median rent compared to income	3
State & local tax rate	3
Violent crime per capita	3
Property crime per capita	3
Libraries per capita	3

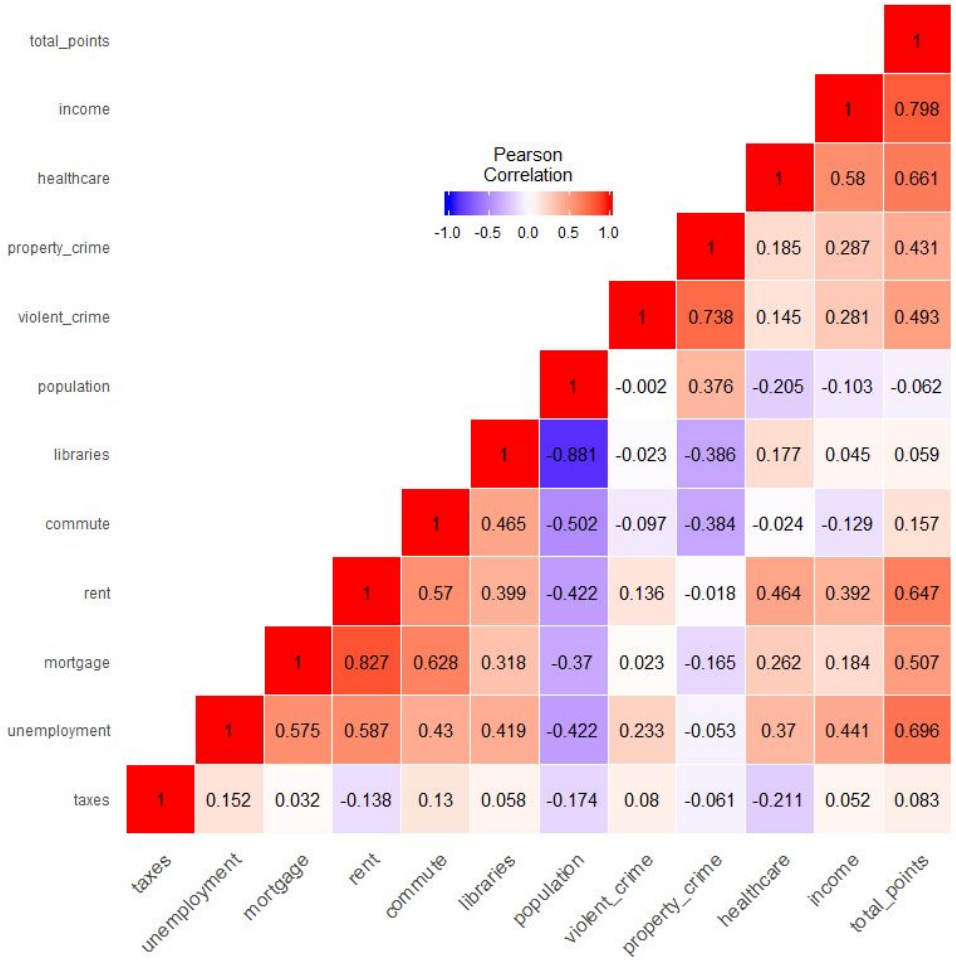
Social Media Sentiment Analysis

Using Twitter, we looked at social media posts made in the vicinity of our top 10 cities. After authenticating in Twitter, we created a loop that searched for 500 tweets near each of the top 10 cities. After cleaning up the data, we then performed sentiment analysis on the text gathered, and calculated the percentage of tweets that had 'positive' sentiment, by taking a ratio of positive sentiments to total.

Correlation Analysis

To find out the top three most important factors in the happiness metric, we performed correlation analysis of its components. First, we converted the data to a numeric matrix and created the correlation matrix. After plotting the matrix and using Pearson correlation, we were able to determine that happiness points correlate the closest to income (.798), employment (.696) and healthcare (.661).

As income, employment or healthcare increase, “happiness” also increases. It is interesting that while population is given a value of 9 points, its correlation is near zero.



Packages Used

We made extensive use of the data.table package, ggplot and related packages, tidyverse packages, and twitteR and wordcloud. Data.table acts as an extension to the default data.frame class, the fact that it can be treated as a data.frame, while having additional functionality was great. Specifically, being able to store our zips data as a numeric vector, in a column in our table, was helpful. Ggplot was used for our correlation chart an our map, with the addition of ggrepel for clean labeling of our map points.

Results

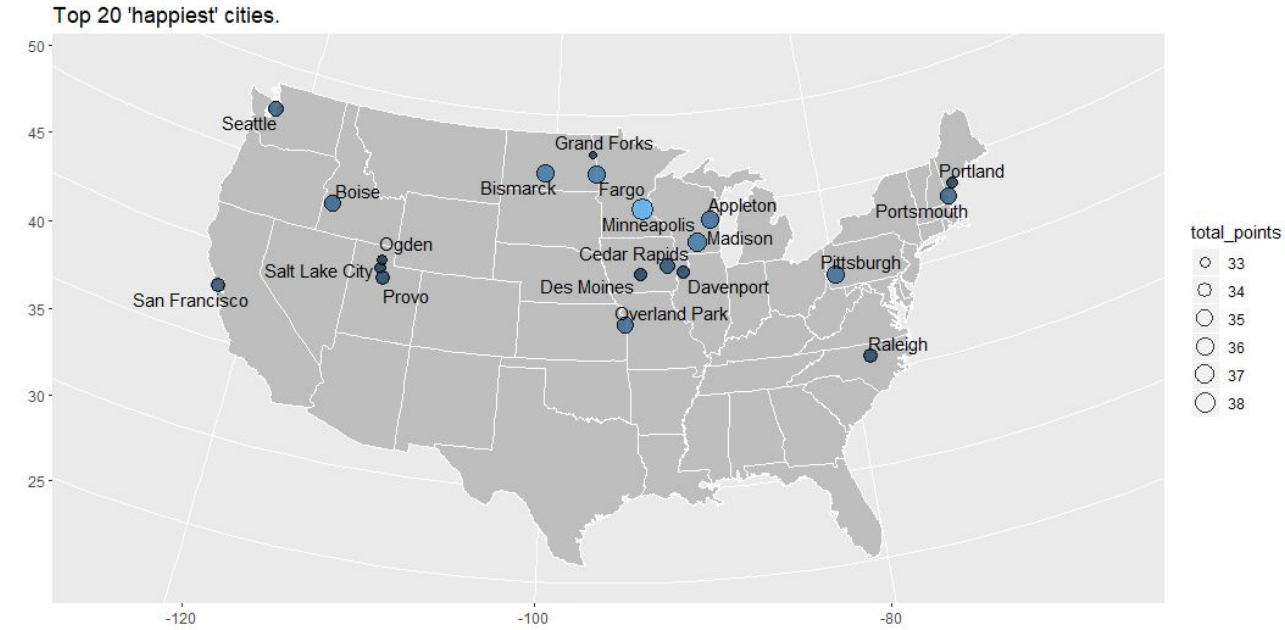
Summary results follow, full tables and data are available in the R project.

Rankings, Top 20

	City	State	Total Points
1	Minneapolis	MN	38.6
2	Madison	WI	36.3

3	Bismarck	ND	36.2
4	Fargo	ND	36.1
5	Appleton	WI	35.8
6	Pittsburgh	PA	35.7
7	Overland Park	KS	35.5
8	Boise	ID	35.3
9	Portsmouth	NH	35
10	Seattle	WA	34.6
11	Cedar Rapids	IA	34.4
12	San Francisco	CA	33.9
13	Raleigh	NC	33.8
14	Provo	UT	33.8
15	Davenport	IA	33.6
16	Des Moines	IA	33.5
17	Salt Lake City	UT	33.3
18	Portland	ME	33.1
19	Ogden	UT	32.9
20	Grand Forks	ND	32.7

Map, Top 20



Sources:

https://simplemaps.com/data/us-cities

https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey/explore-pls-data/pls-data

https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/tables/10tbl08.xls/view

https://wallethub.com/edu/best-worst-states-to-be-a-taxpayer/2416/

https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/DP03/0100000US.05000.004

https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/DP04/0100000US.05000.004