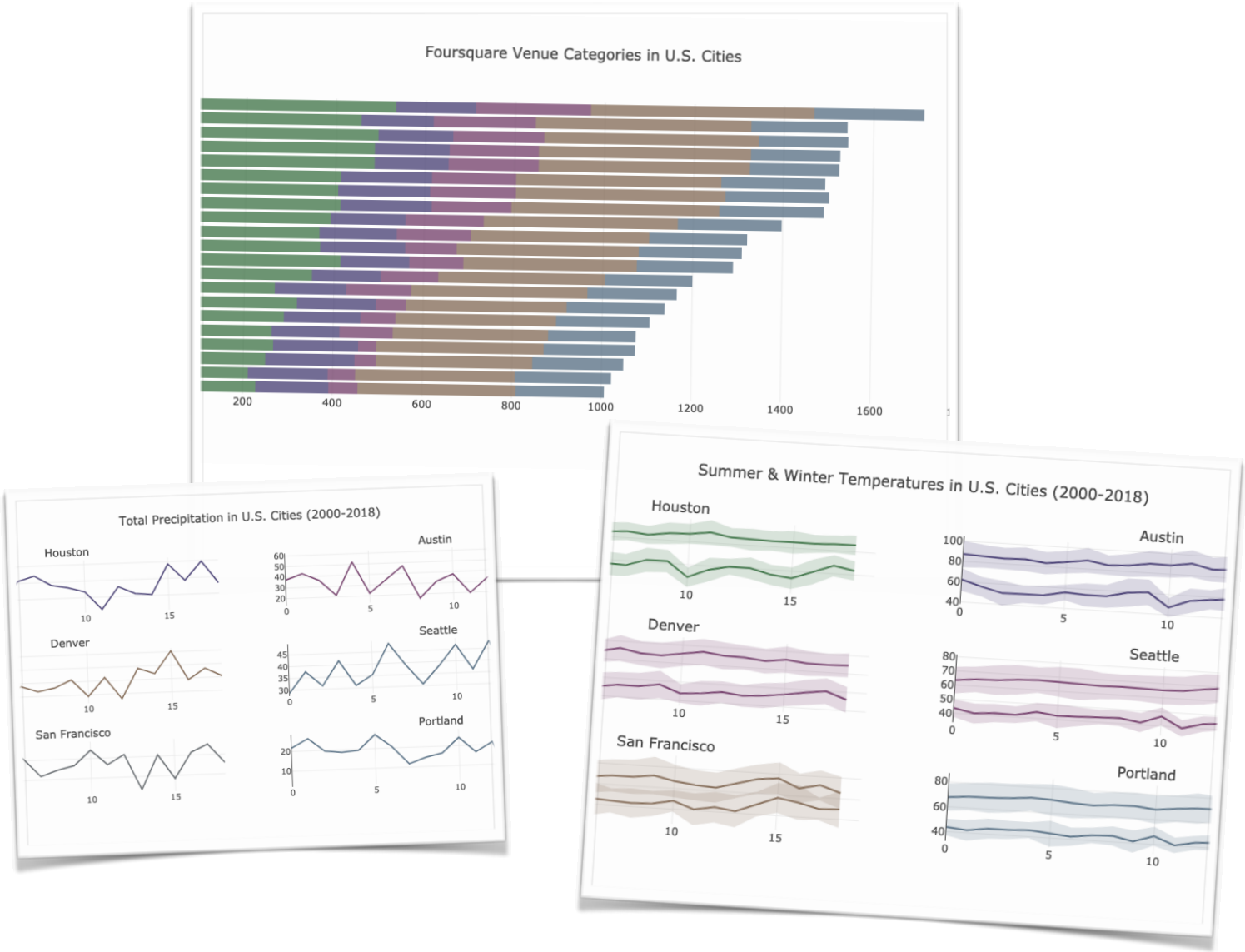
Valentina Fontiveros
April 20, 2019

Using Data Science

To Choose A City To Live in the U.S.A



Valentina Fontiveros
April 20, 2019

Using Data Science

To Choose A City To Live in the U.S.A

Introduction

Pursuing a Data Science career is an exciting career option for scientists and business people wanting to break into the technology industry.

One common consideration when transitioning into a tech career is whether or not it would be beneficial to move into a tech hub such as San Francisco.

Although San Francisco is an attractive option, there are other considerations that should be weighed in order to make the best decision.

The **objective** of this exercise is to evaluate several U.S. cities based on cultural, climatological, & financial data to determine which cities would be a good fit to personal preferences.

This project may be of interest to any person trying to figure out where to move. In order to make an objective, responsible decision, one must research and weigh pros and cons.

The personal preferences used in this instance include:

- Weather: mild.
- Scenery: near mountains.
- Urbanization and beautification: A city with a large number of parks.
- Outdoors: availability of hiking trails and outdoor venues.
- Career: tech hub.
- Economics: a reasonable affordable place.

Finally, a classification model will be designed in order to predict whether an unknown city is a good fit to personal preferences.

Data

Data were downloaded and scraped from various sources including Foursquare and U.S. government sites.

The Foursquare data will be used to determine how many venues fit in my personal interests per city.

The climatological data will be used to record high and low temperature extremes for each city as well as yearly precipitation volume.

The cost of living data will be used to compare key cities after the initial analysis and filtering of cities.

Salary differential data for each U.S. geographical area will be explored and considered in the final analysis.

Links:

Venue Data: https://foursquare.com/

Temperature & Precipitation Data: https://www.ncdc.noaa.gov/cag/city/time-series/

Effective tax rates https://smartasset.com/taxes/oregon-tax-calculator#D3qPPxj0kR

Cost of Living https://www.nerdwallet.com/cost-of-living-calculator

Salaries for data scientist in different U.S. regions: https://datasciencedegree.wisconsin.edu/data-science/data-scientist-salary/ https://learning.oreilly.com/library/view/2016-data-science/9781492049029/ch03.html#idm140536273113840

Methodology

I. Import libraries.

II. Create a list with cities of interest (at least 40)*.

III. Extract and Explore Foursquare data.

    I. Browse Foursquare.com and find venue categories of interest.

    II. Record Category IDs and order them in various lists.

    III. Create a function to connect to Foursquare, extract the relevant data, and store it in a dataframe.

    IV. QC and filter the Foursquare data.

        I. Drop duplicates.

        II. Check venue types and ensure relevant categories are being used.

        III. Find the total venue count per city. Cities with the highest venue counts will be used as a proxy of good "fit" to personal interests.

        IV. Find venue types with least common numbers < 20, and filter them out.

        V. Prepare data for plotting

        VI. Create horizontal bar plots to demonstrate cities with the highest and the lowest number of venues.

IV. Scrape Temperature & Precipitation Data for Each City

    I. Find government site with predictable data structure to extract summer and winter temperatures from 2000-2018.

    II. Find city codes for each city and store them in a list (this is a manual step).

    III. Create a function to generate a URL list for the temperature and precipitation data.

    IV. Store data for average temperatures, high temperatures, and low temperatures in a list of dataframes.

    V. Extract data for each city and store in a dataframe.

VI. Create interactive plots for data exploration for key cities. Since plots are onerous to produce, we will export only the most salient cities.

V. Cost of Living Data

I. Create a cost of living dataset for key cities selected after initial analysis.

VI. Find the best possible classification model for whether or not I would like a city based on tested features.

I. Create training and test datasets based on Like data.

Results

Ranking Cities Based on Foursquare Data

After extracting venue data from each city, they were organized and plotted based on the following categories:

- Outdoors: Outdoors and Recreation Venues: Trails, Bike Trail, Botanical Gardens, Forest, Mountain, Nature Preserve, National Park.
- Professional: Tech Startup, Convention Center, Observatory.
- Startups: Tech Startup, Convention Center, Observatory.
- Cultural: Spiritual Center: Buddhist Temple, Hindu Temple, Synagoge, Winery.
- Food: Farmers Market, Health Food Store, Organic Grocery, Fruit and Vegetable Store, Juice Bar.
- Beauty: Park, Flower Shop, Art Studio.

The city with the most venues was New York City, NY (Fig. 1). It is not surprising, since New York City was urbanized to be a pleasant place to live in. I have been in New York, and I thoroughly enjoyed the stroll around Central Park and the beautiful gardens and public spaces. It is also one of the most multicultural cities of the world. I have visited New York, however, the great numbers of people and high cost of living would be definitely a minus.

New Haven is a close second, and this was a surprising result. I have not visited New Haven.

The next few results are **Palo Alto, San Francisco, San Jose and other cities in California.** This result makes sense, since those are some of my favorites cities.

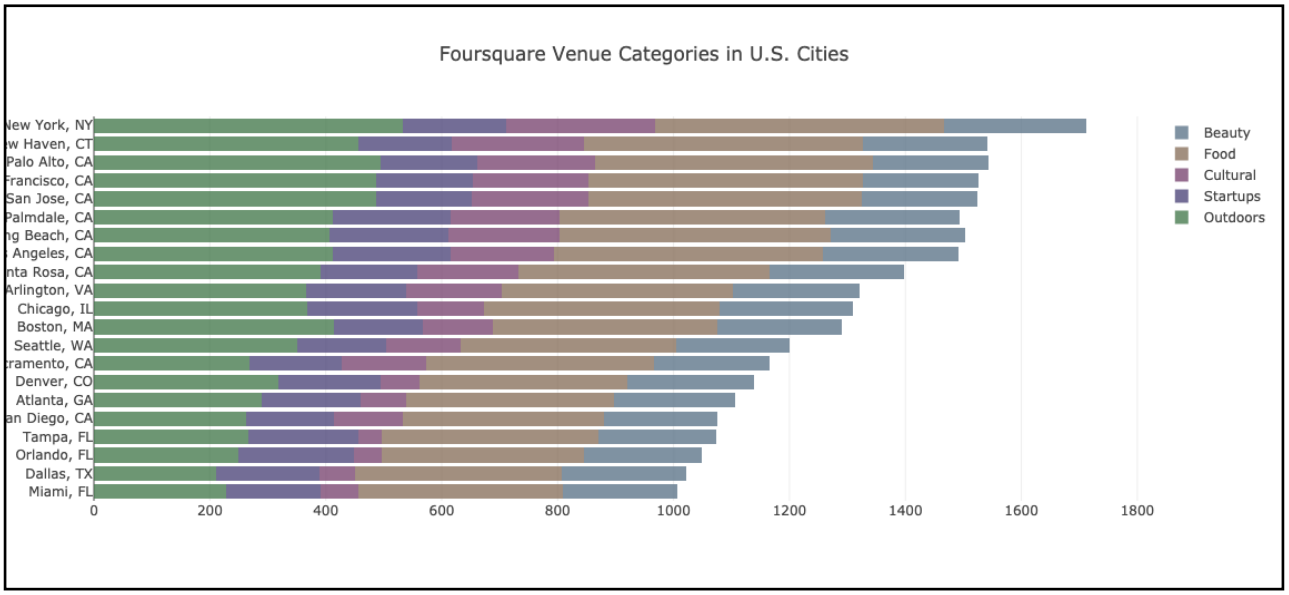


Fig 1a. Number of venues matching personal interests in 41 cities in the U.S.

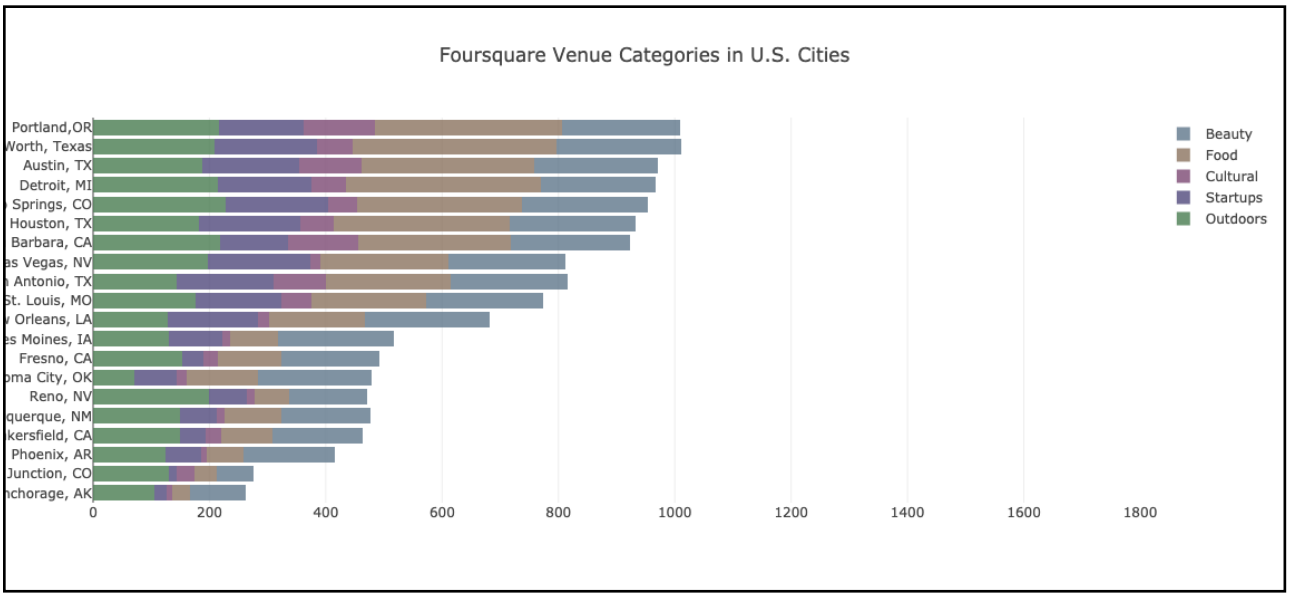


Fig 1b. Number of venues matching personal interests (continued).

The city where I currently reside in is **Houston, Texas, ranks as #27** in this list. I've been interested in exploring **Austin, Texas, and it ranks as #24.** I have also been interested in exploring **Denver, Colorado (#15)** and **Seattle, Washington (#13).**

While I have enjoyed living in California in the past, there are other *considerations* to take into account.

Temperature and Precipitation Data

In order to further discriminate between the various cities, climatological data were extracted analyzed.

Since extracting and plotting this type of data was an onerous process, a subset of the 41 cities was selected for QC and visualization (Fig. 2, 3).

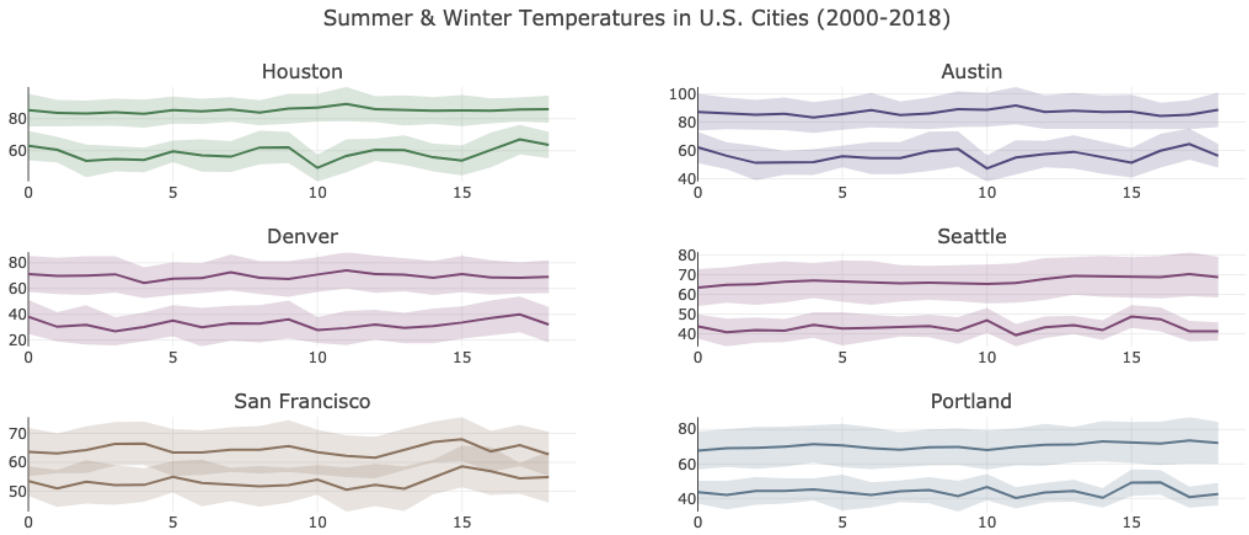


Fig 2. Temperature plots for various cities of interest for the 2000-2018 period (Farenheit). Upper lines show average temperature for the summer season (month of August) and lower lines show the average for the winter season. Shaded areas show the range of temperatures seen during that month (maximum and minimum).

A surprising result is that **Austin, Texas is warmer than Houston, Texas during the summer, and that Denver, CO is colder than Seattle, WA during the Winter.** San Francisco, true to its fame, shows comfortable temperatures year round.

The precipitation data show that it **rains more (in total inches) in Houston than in Seattle**, even though Seattle's reputation would indicate otherwise. It rains less in Denver than it does in either San Francisco or Denver. Another plot of interest would indicate the rain intensity and frequency per city.

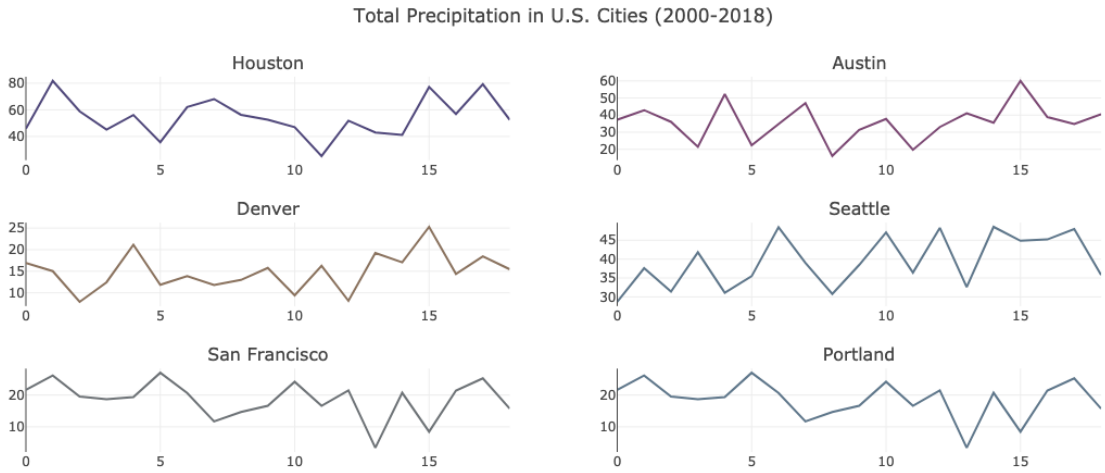


Fig 3. Total annual precipitation plot for various cities of interest between 2000-2018 (Inches).

Cost of Living

Everyone would like to live in Utopia (San Francisco, anyone?). Yet, to live comfortably in this real world of ours, one must ask… how much does it cost?

Effective tax rates for the subset of cities were less variable than originally thought (near 35% for a salary of $120k), and thus, should not be a driving factor in determining whether to live in one location or another.

Cost of living, however, was highly variable and a significant factor. For instance, to enjoy the same standard of living in Houston and San Francisco, one person must earn twice as much in San Francisco.

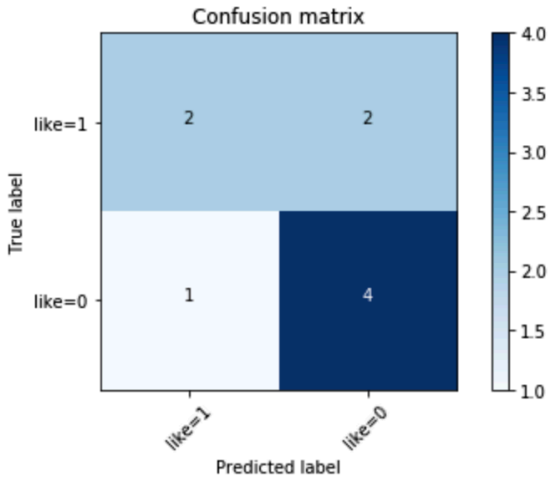
Classification Model

Three features from the Foursquare dataset (total number of outdoors, cultural, and health food venues) were used to train and test a classification model. Forty-one cities were included as samples, with the outcome of Like (Yes/No, 0,1) (Fig. 4). The training set was chosen randomly from 32 samples, while the test set was comprised of 9 samples.

	City	Outdoors	Cultural	Healthy Foods	Like?
0	Anchorage, AK	108	9	32	0
1	Grand Junction, CO	130	30	39	0
2	Phoenix, AR	125	10	64	0
3	Bakersfield, CA	149	28	87	0
4	Albuquerque, NM	149	13	98	0
5	Reno, NV	199	14	59	0
6	Oklahoma City, OK	69	17	122	0
7	Fresno, CA	154	26	109	0
8	Des Moines, IA	131	14	82	0
9	New Orleans, LA	128	19	166	0
10	St. Louis, MO	175	52	177	1

Fig. 4. The first few cities input in the classification model.

After testing the model, 6 out of the 9 samples were classed correctly (See the confusion matrix in Fig.5). The log loss calculate turned out to be ~0.67.




```
print(classification_report(y_test, yhat))
```

```
             precision    recall  f1-score   support

          0       0.67      0.80      0.73         5
          1       0.67      0.50      0.57         4

avg / total       0.67      0.67      0.66         9
```

```
from sklearn.metrics import log_loss
log_loss(y_test, yhat_prob)
```

```
0.66072069826403601
```

Fig. 5. Confusion matrix for the built model.

Discussion

This log loss value is pretty high. Our model is not doing a great job of classifying which cities we would prefer. However, we improved on the chance probability (50%) by predicting correctly 6 out of the 9 cities.

In order to improve this model, one should increase the number of samples significantly and find other more prescriptive parameters.

Data is king. No matter how complicated an algorithm is, if the sample space is limited, classification results will often fail to be satisfactory. It's very time consuming to amass the data required and time pressure will always be a factor. In the near future, we will improve on this model and republish the results in my personal blog.

In addition, predicting preferences is to begin with a very ambitious project to undertake as a budding data scientist with a very limited time frame.

Notwithstanding, one can make good use of the insights extracted from the Foursquare data. The number of total venues seemed to be highly correlated to preference even though this was not strongly supported by the classification model.

Other Sources and Statistics to Consider:

I will have to think more deeply about where to find reliable data regarding these statistics and how to integrate them into my analysis: Healthiest US Cities, Best standard of living, cost of living, demographics.

This website contains open government data. https://cities.data.gov/

Future Work: Other Datasets to Consider

Pollen and Mold Data¶

http://pollen.aaaai.org/nab/index.cfm?p=AllergenCalendar&stationid=188&qsFullDate=10/1/2018

Conclusion

We had some success with a limited dataset to predict preferences. Predicting preferences is extremely difficult since there are so many variables and many are subjective. I thought this exercise would come up with a more clear-cut result!

Extending this example to other applications, marketers love to segment populations based on preferences, but this is a very complicated exercise. No wonder the tech companies want all the data they can gather about us.

It takes 10*(# of independent variables)/(smallest portion of yes/no preference) samples to be able to start having some success with classification algorithms.

To improve this exercise, many more samples need to be taken. Therefore, I need to get a job so I am able to start traveling to all these cities and have some extra time to devote to this analysis!