Investigation of best placement for The Tech Museum

Lars S. Madsen

Applied Data Science Capstone

May 2020

Contents

Abstract

This report contains an investigation of the ten biggest cities in USA in order to find the optimum location for a technology museum. The report contains comparisons between 90 localities in the target cities on factors such as population size, existing museum locations, higher learning institutions and other venues such as restaurants, as well as public transportation possibilities.

Introduction

"The Web as I envisaged it, we have not seen it yet. The future is still so much bigger than the past."
Tim Berners-Lee, Inventor of the World

We want to identify the best location for our Technology Museum. This is the task I will try to solve in this report. What is given is that it has to be in one of the ten largest cities in USA, that there should not already be a science- or technology museum present within a 7-mile radius and that the general composition of the city, and the neighborhood should be similar to where science- or technology museums are already present.

Obviously there are a lot of factors to consider when determining if a given neighborhood is similar to another. Here I will focus on population size, population density, higher learning institutions, restaurant proximity as well as public transportation possibilities.

On the map below the ten largest cities of USA are marked, the marker size indicates population size.

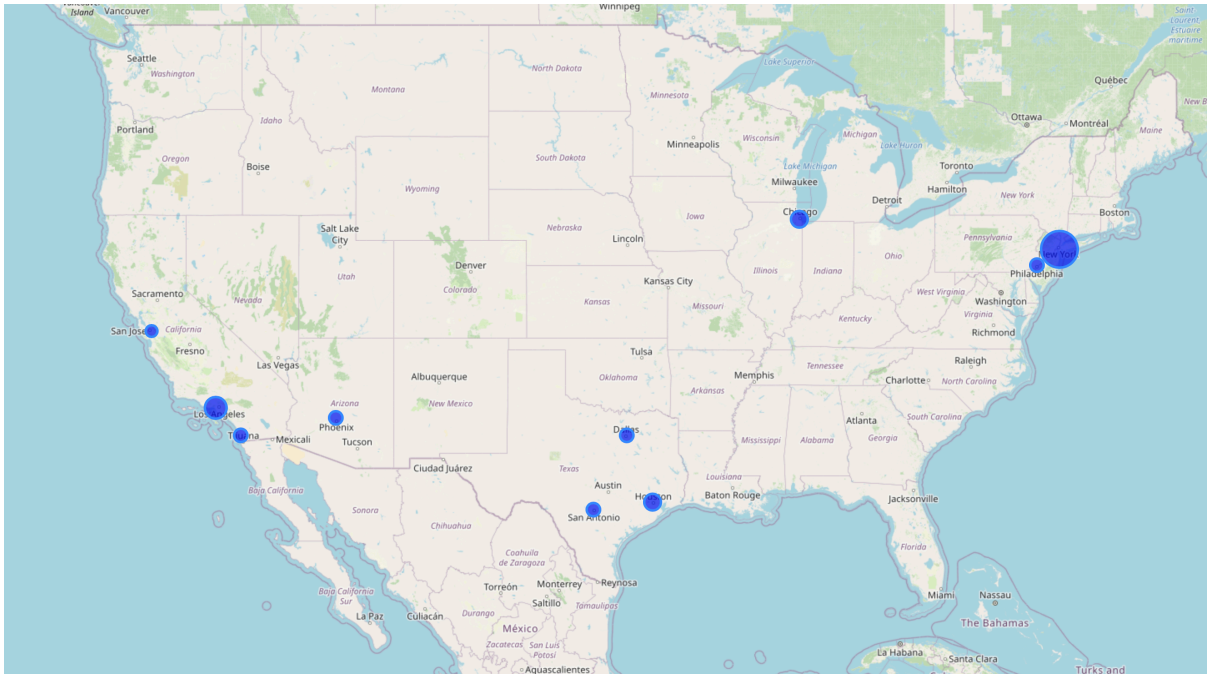

Figure 1

The data used

The data that has been used in this report has been sourced form wikipedia.org and combined with detailed venue information from foursquare.com. For ease I have defined a city as the location and a 21 mile radius (approximately the radius of New York City).

The data will be processed using python in a Jupyter Notebook which can be found here at https://github.com/TheRealLSM/Coursera_Capstone/blob/master/Applied%20Data%20Science%20Capstone%20-%20%20final.ipynb . The notebook will also be used to generate informative charts and maps using the libraries Matplotlib and Folium as well as ML models using Tensorflow / Keras and scikit-learn. Most of the data wrangling will be done using Pandas.

By using Beautifulsoup I have acquired the city list that the report is based upon from https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population. The cities in scope and the corresponding information are:

city	url	population	density	area	latitude	longitude
New York City	https://en.wikipedia.org//wiki/New_York_City	8336817	28,317/sq mi	301.5 sq mi	40.6635	-73.9387
Los Angeles	https://en.wikipedia.org//wiki/Los_Angeles	3979576	8,484/sq mi	468.7 sq mi	34.0194	-118.4108
Chicago	https://en.wikipedia.org//wiki/Chicago	2693976	11,900/sq mi	227.3 sq mi	41.8376	-87.6818
Houston	https://en.wikipedia.org//wiki/Houston	2320268	3,613/sq mi	637.5 sq mi	29.7866	-95.3909
Phoenix, Arizona	https://en.wikipedia.org//wiki/Phoenix,_Arizona	1680992	3,120/sq mi	517.6 sq mi	33.5722	-112.0901
Philadelphia	https://en.wikipedia.org//wiki/Philadelphia	1584064	11,683/sq mi	134.2 sq mi	40.0094	-75.1333
San Antonio	https://en.wikipedia.org//wiki/San_Antonio	1547253	3,238/sq mi	461.0 sq mi	29.4724	-98.5251
San Diego	https://en.wikipedia.org//wiki/San_Diego	1423851	4,325/sq mi	325.2 sq mi	32.8153	-117.1350
Dallas	https://en.wikipedia.org//wiki/Dallas	1343573	3,866/sq mi	340.9 sq mi	32.7933	-96.7665
San Jose, California	https://en.wikipedia.org//wiki/San_Jose,_Calif...	1021795	5,777/sq mi	177.5 sq mi	37.2967	-121.8189

Figure 2

For each of these ten cities I have selected 9 locations for closer examination. The nine locations consist of the center location for the city as well as 8 locations forming a circle around the center location, in a 21 mile radius. Below is a map of the New York City locations analyzed.

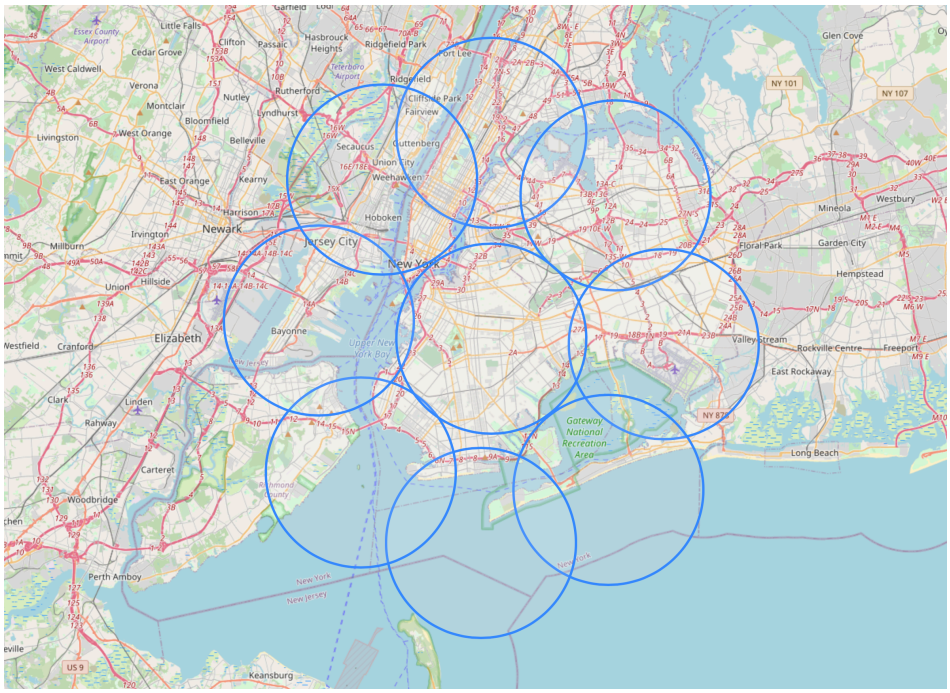

Figure 3

Using the Foursquare APIs I have built a list shown to the right containing number of tech museums present in the specific localities. This list contains the field **“valid”** which specifies if there are *any* tech museums present which will thus disqualify that location. The museum count will be used later in the report for graphics and for similarity comparison.

Following that I have augmented the Pandas dataframe with counts of public transport venues, restaurants and universities within 1/3 mile from the centre of these localities (using the latitude and longitude fields.

city	latitude	longitude	valid	museum_count
New York City	40.798500	-73.938700	0	22
New York City	40.757555	-73.831097	0	4
New York City	40.659558	-73.788764	0	5
New York City	40.563952	-73.837381	0	1
New York City	40.528730	-73.947456	1	0
New York City	40.575258	-74.052220	1	0
New York City	40.675312	-74.088125	0	12
New York City	40.768201	-74.033390	0	30
New York City	40.663500	-73.938700	0	15
Los Angeles	34.154400	-118.410800	0	3
Los Angeles	34.113455	-118.303197	0	20
Los Angeles	34.015458	-118.260864	0	19
Los Angeles	33.919852	-118.309481	0	16
Los Angeles	33.884630	-118.419556	0	2
Los Angeles	33.931158	-118.524320	0	2
Los Angeles	34.031212	-118.560225	0	1
Los Angeles	34.124101	-118.505490	0	1
Los Angeles	34.019400	-118.410800	0	7
Chicago	41.972600	-87.681800	0	10
Chicago	41.931655	-87.574197	0	18

Figure 4

In summary the data I have used is:

Source	Content	Structure
https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population	Top City List	HTML Table
https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population	Population count	HTML Table
https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population	Population density	HTML Table
https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population	Area information	HTML Table
https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population	City location	HTML Table
Foursquare API	Museum count	JSON structure
Foursquare API	Learning institution count	JSON structure
Foursquare API	Public transportation count	JSON structure
Foursquare API	Restaurant count	JSON structure

Figure 5

Methodology

In order to find the right candidates I have used two different machine learning models, a shallow, densely connected, neural network created in Keras and a linear regression model created with scikit-learn. The training material for both models was the same; each sample had city population, number of restaurants within 1/3 mile, number of higher learning institutions within 1/3 mile and number of public transportation stations within 1/3 mile. Since population size and restaurants vary greatly a decided to normalize the input for both models. The target was also the same; number of science/technology museums within the location. The neural network was trained with a stochastic gradient descent optimizer and trained for 15 epochs on the 66 samples (locations containing science museums).

The output from both models is therefore the number is museums the model predicts should be in the input location.

Since the training material has been limited (66 samples) and the input parameters have been rather simple I opted for a very small, shallow, densely connected network instead of a deeper one. Also, due to the nature of the input, the r2 scores very not overly impressive as seen on figure 8 to the right (since several rows had the same input values, but with different result values). Even if the deciding values were perhaps too simple, the models more or less agree on the best locations given the criteria (rows **dense** for the neural network and **linear** for the linear regression model) even if they don't agree on the actual number of museums the locations should have.

```
Dense
-0.007872558401637031
Linear
0.1327352668970364
```

Figure 6

	city	latitude	longitude	valid	museum_count	restaurant_count	transportation_count	institution_count	dense	linear
4	New York City	40.528730	-73.947456	1	0	0	0	0	5.942764	13.218598
5	New York City	40.575258	-74.052220	1	0	0	0	0	5.942764	13.218598
22	Chicago	41.702830	-87.690556	1	0	1	1	1	4.965051	6.145692
27	Houston	29.921600	-95.390900	1	0	0	0	0	3.471725	5.726592
32	Houston	29.698358	-95.504420	1	0	0	0	0	3.471725	5.726592
33	Houston	29.798412	-95.540325	1	0	22	0	1	4.829288	0.015684
34	Houston	29.891301	-95.485590	1	0	1	0	0	3.538155	5.418541
36	Phoenix, Arizona	33.707200	-112.090100	1	0	0	0	0	3.213642	4.930544
43	Phoenix, Arizona	33.676901	-112.184790	1	0	0	0	1	3.429383	5.996766
44	Phoenix, Arizona	33.572200	-112.090100	1	0	3	0	0	3.385561	4.006390
52	Philadelphia	40.114101	-75.227990	1	0	0	0	0	3.172983	4.809847
54	San Antonio	29.607400	-98.525100	1	0	9	0	1	3.836331	3.057767
55	San Antonio	29.566455	-98.417497	1	0	6	0	1	3.682623	3.981921
60	San Antonio	29.484212	-98.674525	1	0	2	0	0	3.275054	4.147906
61	San Antonio	29.577101	-98.619790	1	0	4	0	10	5.263879	14.194014
64	San Diego	32.909355	-117.027397	1	0	0	0	0	3.108586	4.610344
65	San Diego	32.811358	-116.985064	1	0	2	0	0	3.221378	3.994241
69	San Diego	32.827112	-117.284425	1	0	0	0	0	3.108586	4.610344
75	Dallas	32.693752	-96.665181	1	0	0	0	0	3.078145	4.510379
76	Dallas	32.658530	-96.775256	1	0	0	0	0	3.078145	4.510379
77	Dallas	32.705058	-96.880020	1	0	0	0	0	3.078145	4.510379
79	Dallas	32.898001	-96.861190	1	0	1	0	0	3.130698	4.202328
83	San Jose, California	37.292758	-121.668964	1	0	0	0	0	2.956127	4.109691
84	San Jose, California	37.197152	-121.717581	1	0	0	0	0	2.956127	4.109691
85	San Jose, California	37.161930	-121.827656	1	0	0	0	0	2.956127	4.109691

Figure 7

Data analysis

The 90 candidate locations was quickly brought down to 25 qualified locations when removing locations where science or technology museums were already present. The distribution of valid locations at this point amongst the candidate cities is shown on figure 7 below.

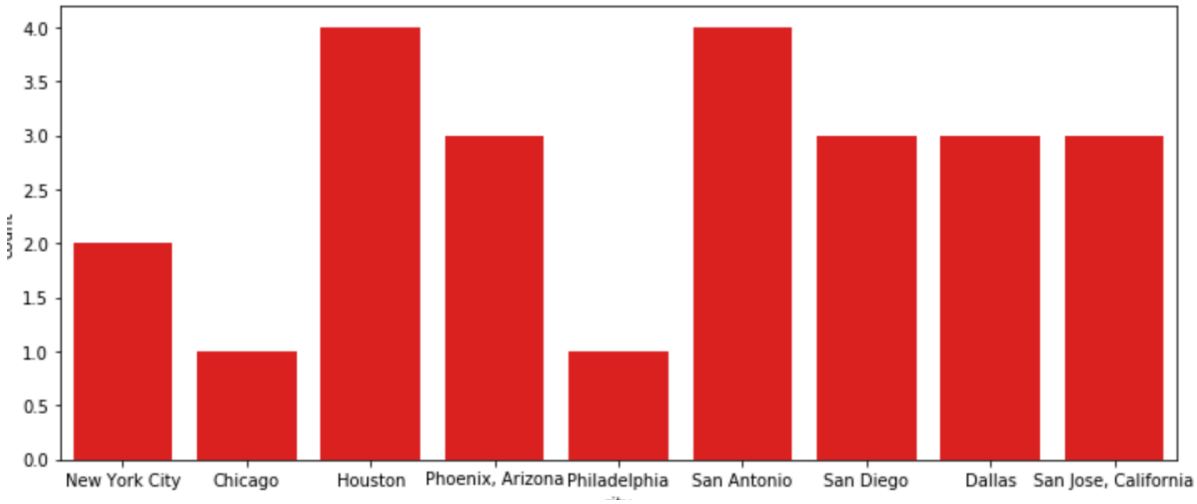

Figure 8

On closer inspection it turns out that most of the area in the two locations in New York City is in the water however and there are thus no restaurants, institutions or public transportation on these sites. I therefore decided to ignore suggestions with no restaurants the from the models. This especially had an influence on the results from the linear regression model which seemed to favor city population and institution count over other venue counts. The densely connected network however seemed to be a bit more balanced between all input values. For the densely network a custom loss function could also have been used to penalize locations with no venues. Also, more diverse input data for the models, would have resulted in a more nuanced picture as well.

Results

The aim of this report has been to highlight how data sources from like Wikipedia and Foursquare can be combined into training data for different machine learning models that could then be used to find the optimum location for our imaginary The Tech Museum. The thesis was that population size, as well as nearby venues would have an impact in the quality of a location. Since it must be assumed that locations where there are already several technology museums present are the best locations, the models were trained in a way so as to suggest locations that were similar to those with many technology museums. It goes without saying that the 4 success criteria used in the models are not telling the full truth when it comes to deciding the quality of a location. Therefore it is suggested to use more information than what was used here, in order to get a more clear picture. It seems however, that both models agree on the best locations, given the criteria as shown in figure 9.

`Dense model top 3 candidate locations:`

	city	latitude	longitude	valid	museum_count	restaurant_count	transportation_count	institution_count	dense	linear
61	San Antonio	29.577101	-98.619790	1	0	4	0	10	5.263879	14.194014
22	Chicago	41.702830	-87.690556	1	0	1	1	1	4.965051	6.145692
33	Houston	29.798412	-95.540325	1	0	22	0	1	4.829288	0.015684

`Linear regression model top 3 candidate locations:`

	city	latitude	longitude	valid	museum_count	restaurant_count	transportation_count	institution_count	dense	linear
61	San Antonio	29.577101	-98.619790	1	0	4	0	10	5.263879	14.194014
22	Chicago	41.702830	-87.690556	1	0	1	1	1	4.965051	6.145692
34	Houston	29.891301	-95.485590	1	0	1	0	0	3.538155	5.418541

Figure 9

Conclusion

So there you have it! The optimum location for The Tech Museum was decided to be in San Antonio, just south of The University of Texas of San Antonio based on city population size, restaurants, higher learning institutions and public transportation stations nearby. The location is shown on figure 9 below.

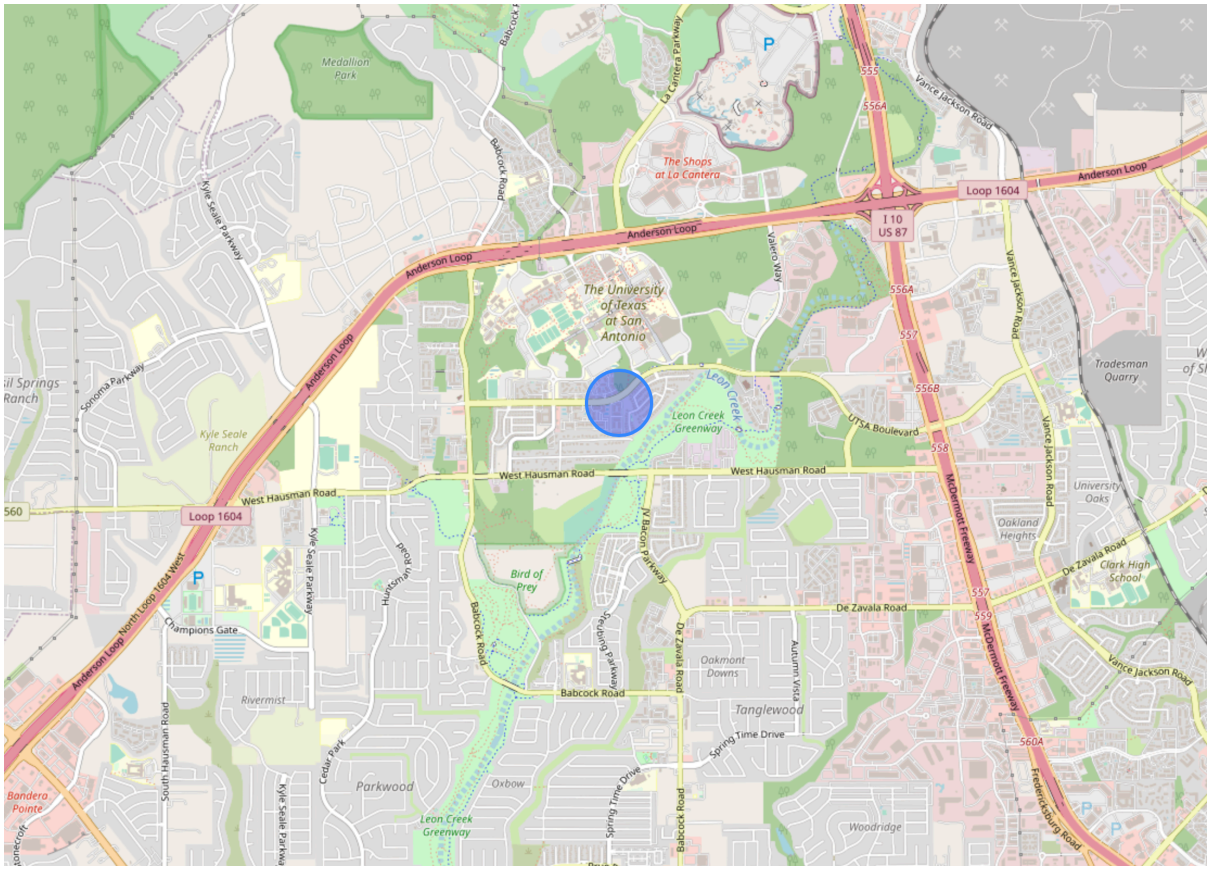

Figure 10

If that is not possible then both models suggest Chicago as the second best solution.

Of course if we do decide to create a floating technology museum then both models agree that New York City is the right place for it.