Performing Arts Theater
Hotspots around the United States

by John Jackson

Introduction/Business Problem

In this report, we will look at Performing Arts Theaters around the United States, to identify communities that have a thriving performing arts theater community, and locations that have a relatively deficient presence. This report may be of interest to those considering relocating to areas with thriving communities, such as actors, managers, designers, patrons or investors, or for those looking to establish theater in areas that are currently underrepresented.

For example, cities with either a high, or a higher than predicted, number of theaters will be considered hotspots, and those with less will be considered "growth opportunities." Additional insights may be gathered while analyzing the data.

Data Sources

The following data sources were used:

I gathered population-related data for all large cities in the United States, getting data from:

1) https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

which lists cities with populations over 100,000, growth rate (percent growth from 2010 to 2018), and population density (from 2016) available from the same source. The data looks like this:

Rank	City	State[c]	2018 estimate	2010 Census	Change	2016 land area		2016 population density		Location
1	*New York*[d]	New York	8,398,748	8,175,133	+2.74%	301.5 sq mi	780.9 km²	28,317/sq mi	10,933/km²	40.6635°N 73.9387°W
2	*Los Angeles*	California	3,990,456	3,792,621	+5.22%	468.7 sq mi	1,213.9 km²	8,484/sq mi	3,276/km²	34.0194°N 118.4108°W

Rank	City	State[c]	2018 estimate	2010 Census	Change	2016 land area		2016 population density		Location
3	*Chicago*	Illinois	2,705,994	2,695,598	+0.39%	227.3 sq mi	588.7 km²	11,900/sq mi	4,600/km²	41.8376° N 87.6818° W

The population-related data required a bit of wrangling, to remove commas, footnote labels, percent signs and other text. There are 314 cities with a population of 100,000 or more in the database.

2) Foursquare (https://foursquare.com/) was used to identify the number of performing arts theaters within 10km of each city center. From foursquare, I used, under "categories", the performing arts id, {'id': '4bf58dd8d48988d1e5931735'}, and then filter to include {'name': 'Theater'} but not other names, such as {'name': 'Music Venue'} or {'name': 'Concert Hall'} which show up under the same id.

The raw (json) data returned from foursquare looks like this:

{'meta':
{'code': 200, 'requestId': '5e8e6135618f43001bebae82'},
'response': {'venues': [
{'id': '4a0b05a7f964a520bb741fe3',
'name': 'The Walter Kerr Theatre', '
location': {'address': '219 W 48th St',
'crossStreet': 'btwn Broadway & 8th Ave',
'lat': 40.76061320906541, 'lng': -73.98572244899813,
'labeledLatLngs': [{'label': 'display',
'lat': 40.76061320906541,
'lng': -73.98572244899813},
{'label': '?', 'lat': 40.760562, 'lng': -73.985672}],
'distance': 327,
'postalCode': '10036',
'cc': 'US',
'city': 'New York',
'state': 'NY',
'country': 'United States',
'formattedAddress': ['219 W 48th St (btwn Broadway & 8th Ave)',
'New York, NY 10036',
'United States']},
'categories': [{'id': '4bf58dd8d48988d137941735',
'name': 'Theater',
'pluralName': 'Theaters',
'shortName': 'Theater',

```
                'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/
                        arts_entertainment/performingarts_theater_',
                'suffix': '.png'},
                'primary': True}],
    'referralId': 'v-1586389276',
    'hasPerk': False}
```

3) geocoder.arcgis (import geopy.geocoders) was used to get latitude and longitude for each city. For example:

```
g = geocoder.arcgis('New York, New York, USA')
lat_lng_coords = g.latlng
```

4) Because foursquare returns at most 50 venues per call, and the number of performing arts venues in many cities is greater than this, I had to figure out a way to get performing arts venues in distributed areas throughout the city. To accomplish this, I used www.zipcodeapi.com, which returns all of the zip codes in a city. The code looks like this:

```
APIkey = 'cUFEF53I6cj2mYOySZXHKVFvwPFNRFCekd5rcd4ijL1P
          dGEMlkPg7wAeOdoM3Z2y'
city = Austin
state = 'TX'
url = f"https://www.zipcodeapi.com/rest/{APIkey}/city-zips.json/
      {city}/{state}"
response = requests.get(url)
```

At most 10 calls can be made in an hour, and the number of cities requiring zip codes was over 100, so many hours were required to get all of the necessary zip codes. Zipcode data was saved to a file after each usage.

When using foursquare to get the venues in a city, if the number returned was 50 or more, then I repeated the call to get data for each zip code. There are over 100 zipcodes in New York City and in Los Angeles, so it took many calls to get all of the venues in many of the larger cities. Venues names then had to be filtered to ensure that they were unique, so no venue was counted multiple times.

To use the tool, I also had to write up a conversion for state names to their two-letter abbreviations.

4) http://www.univsearch.com/family-search.php?family=50
lists the colleges and universities in the United States with Visual and Performing Arts Schools and Department, with their city and state. An example is here:

COLLEGE/ UNIVERSITY NAME	CITY	STATE

A1 BUSINESS AND TECHNICAL COLLEGE	CAGUAS	PR
ABILENE CHRISTIAN UNIVERSITY	ABILENE	TX
ACADEMIA SERRANT INC	PONCE	PR
ACADEMY EDUCATION CENTER INC	MINNEAPOLIS	MN
ACADEMY OF ART COLLEGE	SAN FRANCISCO	CA

There are 2065 colleges and universities in that database.

After combining all of the data, I got the dataframe as shown below:

```
city_df.head(10)
```

	City	State	Population	Growth	Density	Latitude	Longitude	Theaters	Colleges
0	New York	New York	8398748	2.74	28317	40.71455	-74.00714	78	30
1	Los Angeles	California	3990456	5.22	8484	34.05349	-118.24532	163	20
2	Chicago	Illinois	2705994	0.39	11900	41.88425	-87.63245	169	17
3	Houston	Texas	2325502	10.72	3613	29.76078	-95.36952	106	12
4	Phoenix	Arizona	1660272	14.85	3120	33.44825	-112.07580	65	6
5	Philadelphia	Pennsylvania	1584138	3.81	11683	39.95222	-75.16218	79	16
6	San Antonio	Texas	1532233	15.43	3238	29.42458	-98.49461	79	7
7	San Diego	California	1425976	9.07	4325	32.71568	-117.16171	84	11
8	Dallas	Texas	1345047	12.29	3866	32.77822	-96.79512	57	7
9	San Jose	California	1030119	8.90	5777	37.33865	-121.88542	40	5

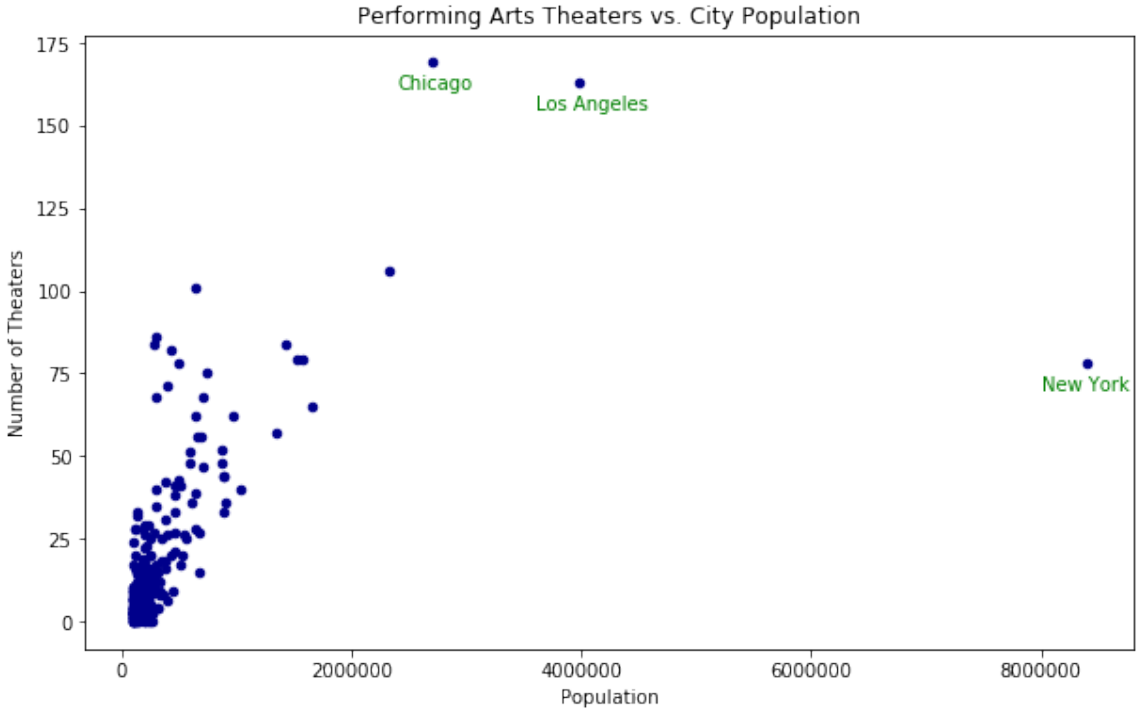
```
city_df.tail(10)
```

	City	State	Population	Growth	Density	Latitude	Longitude	Theaters	Colleges
304	Davenport	Iowa	102085	2.41	1631	41.53826	-90.57602	9	2
305	South Bend	Indiana	101860	0.68	2457	41.67907	-86.25405	10	2
306	Vista	California	101224	7.88	5436	33.20239	-117.23505	10	0
307	Tuscaloosa	Alabama	101113	11.77	1399	33.21042	-87.56625	9	3
308	Clinton	Michigan	100800	4.14	3573	42.07273	-83.97033	0	0
309	Edison	New Jersey	100693	0.73	3389	40.53055	-74.38114	3	1
310	Woodbridge	New Jersey	100450	0.87	4351	40.55452	-74.27851	1	0
311	San Angelo	Texas	100215	7.53	1681	31.46150	-100.44242	4	1
312	Kenosha	Wisconsin	100164	0.95	3577	42.58808	-87.82326	7	3
313	Vacaville	California	100154	8.36	3449	38.35218	-121.99263	3	0

A multivariate linear regression model was used to predict the number of performing arts theaters based on the city population, growth, density, and the number of local colleges with arts departments.

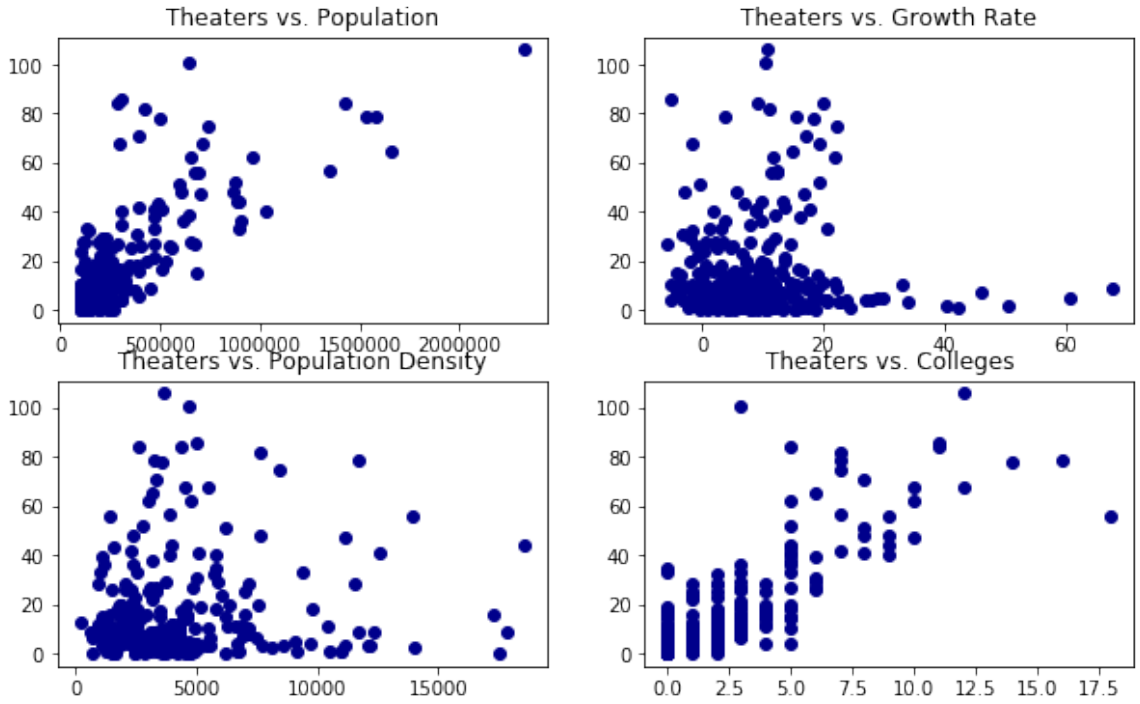
Methodology

The number of theaters as a function of population is shown in this scatterplot:



The three largest cities, New York, Los Angeles, and Chicago, are much larger and/or have many more theaters than other locations, so we remove them as outliers, although we also remember that these are the most prominent theater areas in the country!

We now consider a multivariate regression to estimate the influence of population, growth rate, population density, and local colleges on the number of performing arts theaters.



The plots seem to infer that there is correlation between the number of theaters and city population, and with the number of colleges, but not with growth rate or population density.

Indeed, running multivariate regressions using the following combinations of variables generate the associated R^2 values.

Variables	R^2
Population, Growth, Density, Colleges	0.764
Population	0.591
Growth	0.000
Density	0.015
Colleges	0.666
Population, Colleges	0.762

Based on these results, we discard the cities growth rates and density as being uncorrelated with the number of theaters, whereas both the population and number of colleges are correlated with the number of theaters, and a regression with both of them provides the highest predictive value, capturing 73.4% of the variance.

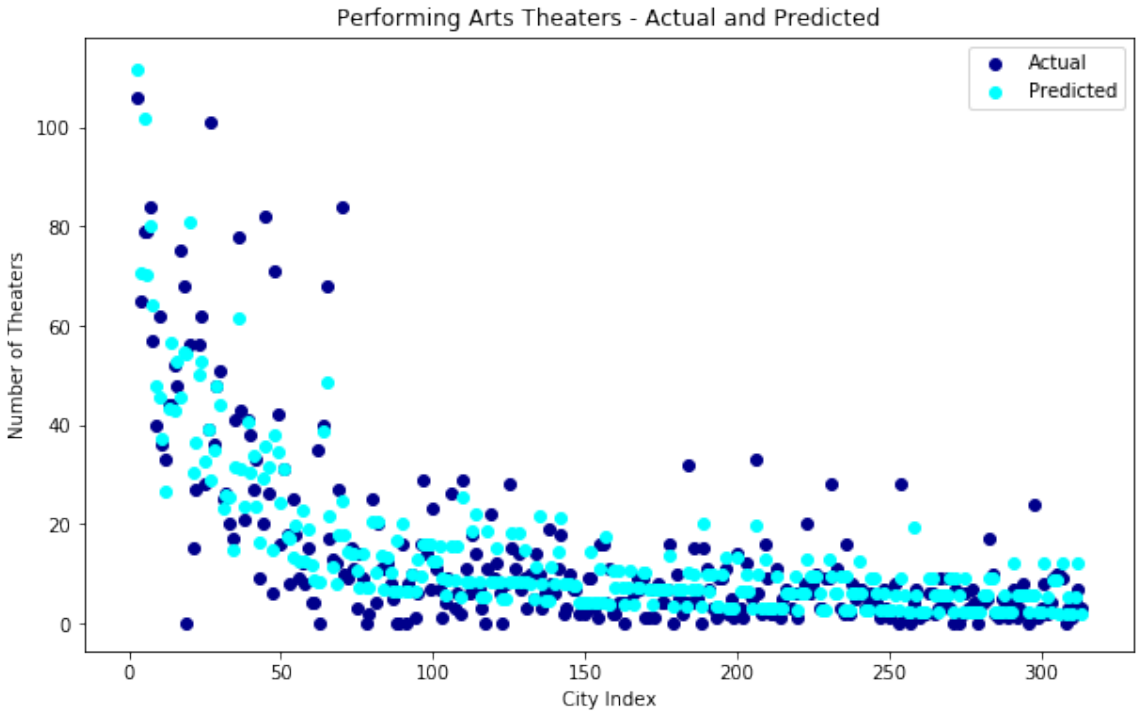
The regression equation with population and colleges is:

predicted_theaters = -1.095 + 3.1086e-5 * Population + 3.3608 * Colleges

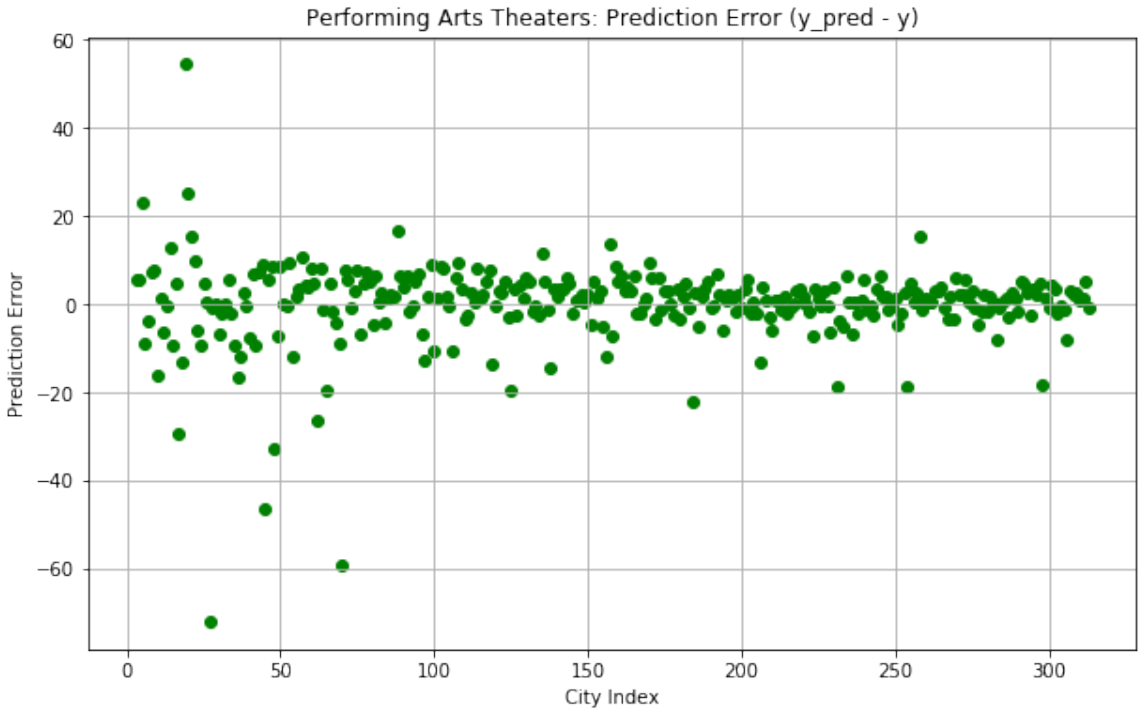
Interpreting, there are about 3.1 theaters for every 100,000 in population, plus 3.4 theaters for every local college.

Inferences:

So, based on the regression equation, what cities have significantly more theaters? and which have significantly fewer?



This is fairly cluttered. Look instead at the prediction error (theaters_predicted – theaters_actual):



I added the predicted number of theaters, and the prediction error, to the dataframe and sorted by error, to highlight the cities with more than and fewer than the expected number of theaters.

Results

The cities with theater numbers that are higher than typical accounting for population and the number of local colleges are:

	City	State	Colleges	Theaters	Pred_Theaters	Pred_error
27	Las Vegas	Nevada	3	101	29.037787	-71.962213
70	Orlando	Florida	5	84	26.054163	-57.945837
45	Minneapolis	Minnesota	7	82	37.697352	-44.302648
63	St. Louis	Missouri	11	86	49.192861	-36.807139
48	Tampa	Florida	8	71	40.516322	-30.483678
17	Seattle	Washington	7	75	47.071260	-27.928740
62	Saint Paul	Minnesota	0	35	7.835374	-27.164626
2	Chicago	Illinois	17	169	142.324660	-26.675340
184	Syracuse	New York	2	32	10.542209	-21.457791
125	Columbus	Georgia	1	28	8.277602	-19.722398

and cities with fewer theaters than might be expected are:

	City	State	Colleges	Theaters	Pred_Theaters	Pred_error
135	Worcester	Massachusetts	5	10	23.125521	13.125521
78	St. Petersburg	Florida	2	0	14.131259	14.131259
14	San Francisco	California	9	44	58.675140	14.675140
157	Jackson	Mississippi	4	4	18.723424	14.723424
21	El Paso	Texas	3	15	30.153233	15.153233
258	Manchester	New Hampshire	5	4	20.973774	16.973774
5	Philadelphia	Pennsylvania	16	79	105.642816	26.642816
1	Los Angeles	California	20	163	191.321923	28.321923
20	Boston	Massachusetts	18	56	87.093586	31.093586
0	New York	New York	30	78	358.364343	280.364343

Discussion

Perhaps not surprisingly, the cities with thriving numbers of performing arts theaters include Las Vegas, Nevada (top of the list) and Orlando, Florida, both very popular tourist destinations. Other thriving theater programs exist in Minneapolis, Minnesota, St. Louis, Missouri, Tampa, Florida, and Seattle Washington.

Areas that could likely support more theaters include New York City, Boston, and Los Angeles (although real estate and other factors are also bigger concerns in these locations). Other cities that could likely support more theaters include Philadelphia, Pennsylvania, Manchester, New Hampshire, and El Paso, Texas.

Conclusion

Finding and pulling data from various databases can be difficult and time consuming. Different formats for data must be carefully studies and reconciled. Nonetheless, interesting insights into areas to consider for starting a business or where one could find a job or career working in an apparently thriving community can be found based on empirical data. I'll have to check out a theater or two the next time I am in Minneapolis or St. Louis!