IBM Data Science – Course 9 Final Project
Cultural outlets as correlated with city characteristics in the United States

Introduction

Different cities across the United States offer different cultural outlets. Some are hubs for sports, some for industrial production, some are known for their foods, some are known for art and entertainment. With the advantage of access to all of these enriching enterprises can often come a cost. City size, population density, and cost of living all rise with the level of activity of a city. For those who value both a rich city life, but who also don't want to drive themselves bankrupt seeking one out, it can be overwhelming and intimidating to determine where to go.

Business Problem

This project is aimed at providing guidance to people who are interested in certain types of artistic outlets - performing arts, music, and comedy. Where can these endeavors be found in large quantities? Are all of the locations where they can be found prohibitively expensive or overwhelmingly busy with city life? Or, are there some cities where large amounts of interesting can be found, without the hustle bustle present in many cities, and at a reasonable cost? This project will provide some helpful information toward answering these questions.

Data

This project essentially involves looking at relations among presence of performance venues in different cities, and some basic socioeconomic characteristics of those cities. Consequently, the data sets required must relate to these factors. Below is a list of the data sets used, with referrals to where they were sourced from. The referrals are available in the references section.

- List of populous United States cities. [1]
- Population and population density of these cities. [1]
- Quantified cost of living in each city. [2]
- Location of each city (latitude / longitude). [1]
- Entertainment venues present in each city, and their locations. The details of the venues are not necessary - however, their locations are, as is how many venues of each relevant type (comedy, music, or performing arts) are in or near each city. [3]

Methodology

Three dataframes, corresponding to the above-described data, were populated. They respectively consisted of

1. US city data: list of populous US cities, population and population density of these cities, location of each city.
2. US city data: cost of living in each city.
3. Performance venues data: total number of comedy, music, and performing arts venues within or near each city that was common to both of the above two data frames.

In populating dataframe 3, two cutoffs were used for the maximum distance from the city center that was allowed for counting a venue as belonging to a city: 15km, and 100km. These distances were loosely interpreted in the remainder of the analysis as corresponding to the city itself, vs. the city's greater metro area.

The data frames were cleaned and merged into a single dataframe which was used for the remainder of the analysis. An image of this dataframe is below.

	2019 rank	city	state	state abbreviation	Cost of Living Index	2019 estimate	2010 census	change	2016 land area (mi^2)	2016 land area (km^2)	2016 population density (mi^-2)	2016 population density (km^-2)	Location	total results - comedy	total results - music	total results - performing
0	1	New York City	New York	NY	131.0	8336817	8175133	+1.98%	301.5 sq mi	780.9 km2	28,317/sq mi	10,933/km2	40°39′49″N 73°56′19″W / 40.6635°N 73.9387°W...	141	289	231
1	2	Los Angeles	California	CA	143.4	3979576	3792621	+4.93%	468.7 sq mi	1,213.9 km2	8,484/sq mi	3,276/km2	34°01′10″N 118°24′39″W / 34.0194°N 118.4108°...	138	186	168
2	3	Chicago	Illinois	IL	102.0	2693976	2695598	−0.06%	227.3 sq mi	588.7 km2	11,900/sq mi	4,600/km2	41°50′15″N 87°40′54″W / 41.8376°N 87.6818°W...	86	201	182
3	4	Houston	Texas	TX	96.9	2320268	2100263	+10.48%	637.5 sq mi	1,651.1 km2	3,613/sq mi	1,395/km2	29°47′12″N 95°23′27″W / 29.7866°N 95.3909°W...	21	150	160
4	5	Phoenix	Arizona	AZ	105.8	1680992	1445632	+16.28%	517.6 sq mi	1,340.6 km2	3,120/sq mi	1,200/km2	33°34′20″N 112°05′24″W / 33.5722°N 112.0901°...	32	137	152

Figure 1: Merged dataframe containing all information used for the analysis.

Several exploratory analyses of the data was then performed. These included:

1. Histograms to understand the distribution of the number of each type of venue in large cities.
2. Scatter plots the total number of each type of venue against city socioeconomic characteristics to get an idea of what sorts of correlations exist, and how tight they are.
3. Linear (ridge) modelling of the relationship between venues and socioeconomic characteristics, to put numbers to how predictive the latter are of the former.
4. Map generation showing city location, number venues, and cost of living, to provide at-a-glance insight into where to look to find lots of culture at reasonable prices.

Results

Below are provided figures from each of the above-mentioned exploratory analyses.

Histograms: Revealed in the histograms is a difference in distribution when only the city proper vs. the larger metro area is taken into account – with a larger area, the distribution for each type of venue evens out, whereas most cities have a small number of venues and only a few have a large number of venues when only the city proper is considered. Additionally, the distributions vary by venue type: Most cities have a small number of comedy venues, with less having a large number. By contrast, there is a more even distribution in the number of music and performing arts venues across cities.

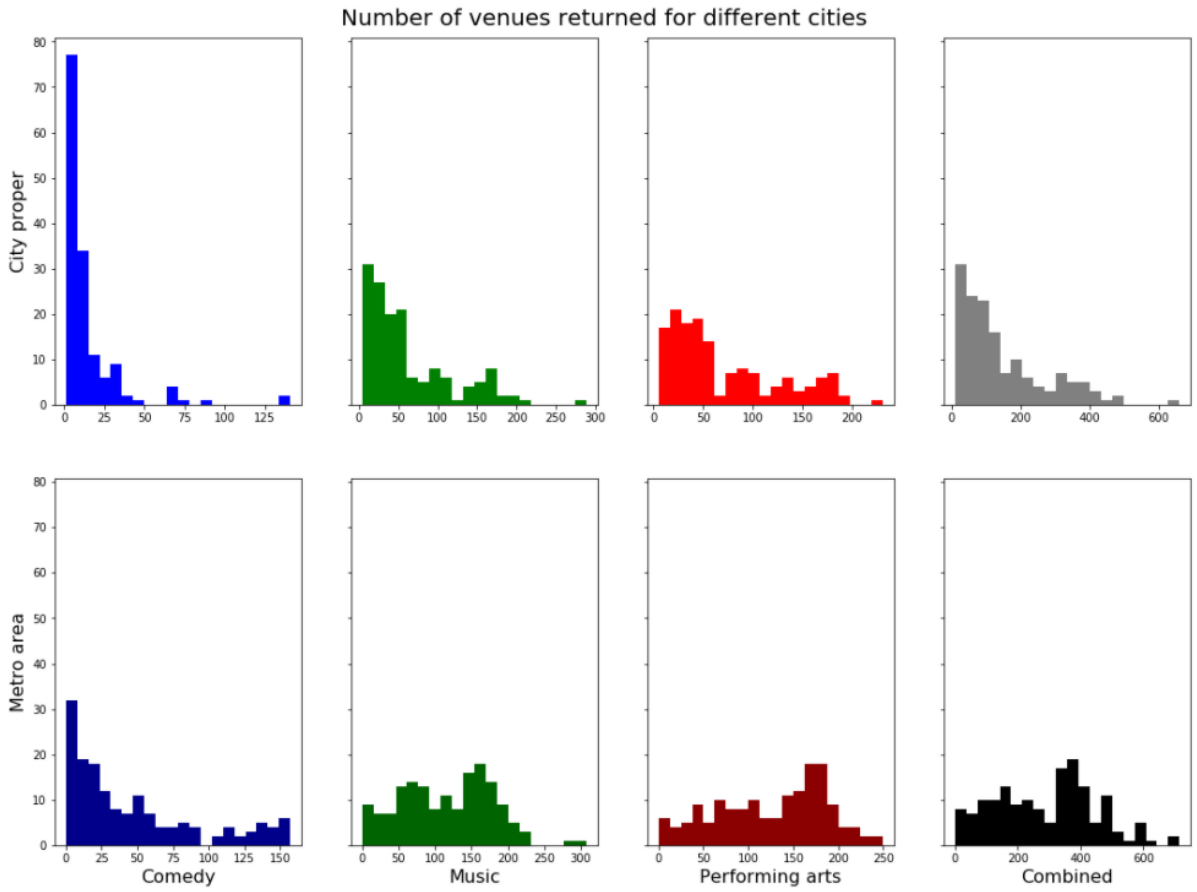

Figure 2: Histograms of total number of venues of different types in different cities. Top-to-bottom: City proper (≤15km from city center), city metro area (≤100km from city center). Left-to-right: Comedy venues, music venues, performing arts venues, all venues combined.

Scatter plots: Below are four figures, each corresponding to a different type of performance venue. (Figures 3, 4, 5, and 6 have y-axes with total number of comedy, music, performing arts, and combined venues, respectively.) While all plots show a positive correlation between availability of each type of venue in a city and that city's socioeconomic characteristics (cost of living, population, and population density), this correlation is extremely loose. Additionally, it seems to be split for cities with large vs. small populations and population densities – this is made visible by columns 3 and 5 in the plot, which are respectively the same as columns 2 and 4, but with only the leftmost points shown.

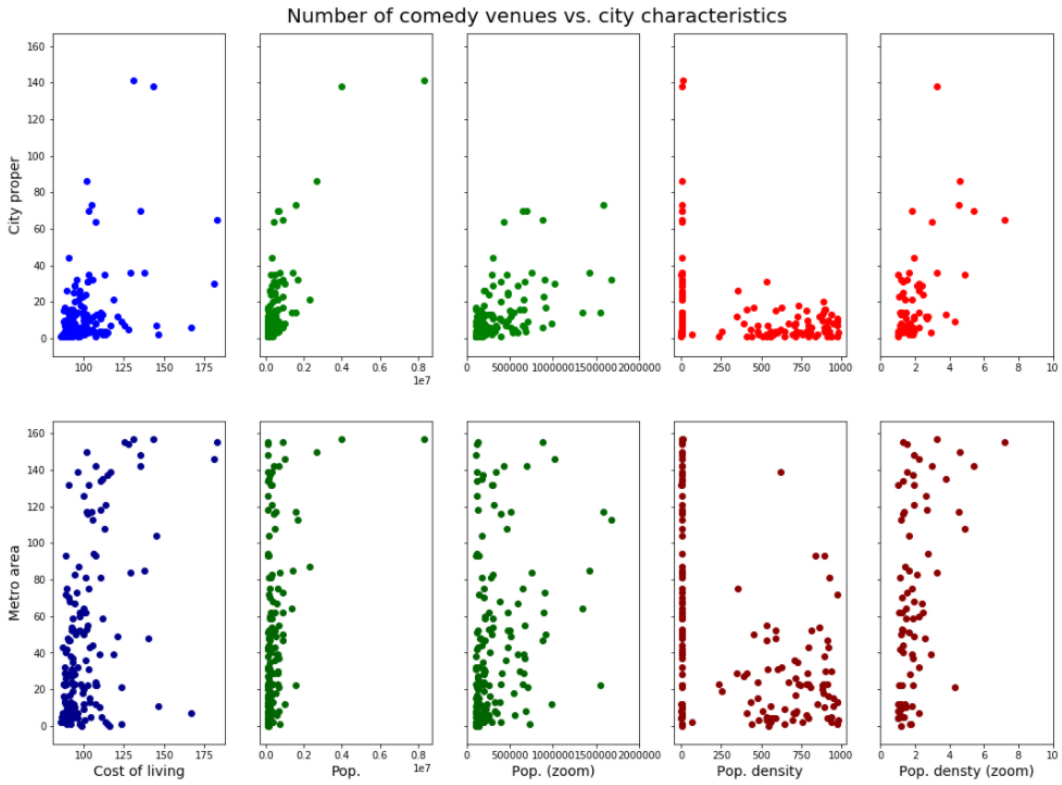

Figure 3: Number of comedy venues vs. different city characteristics.

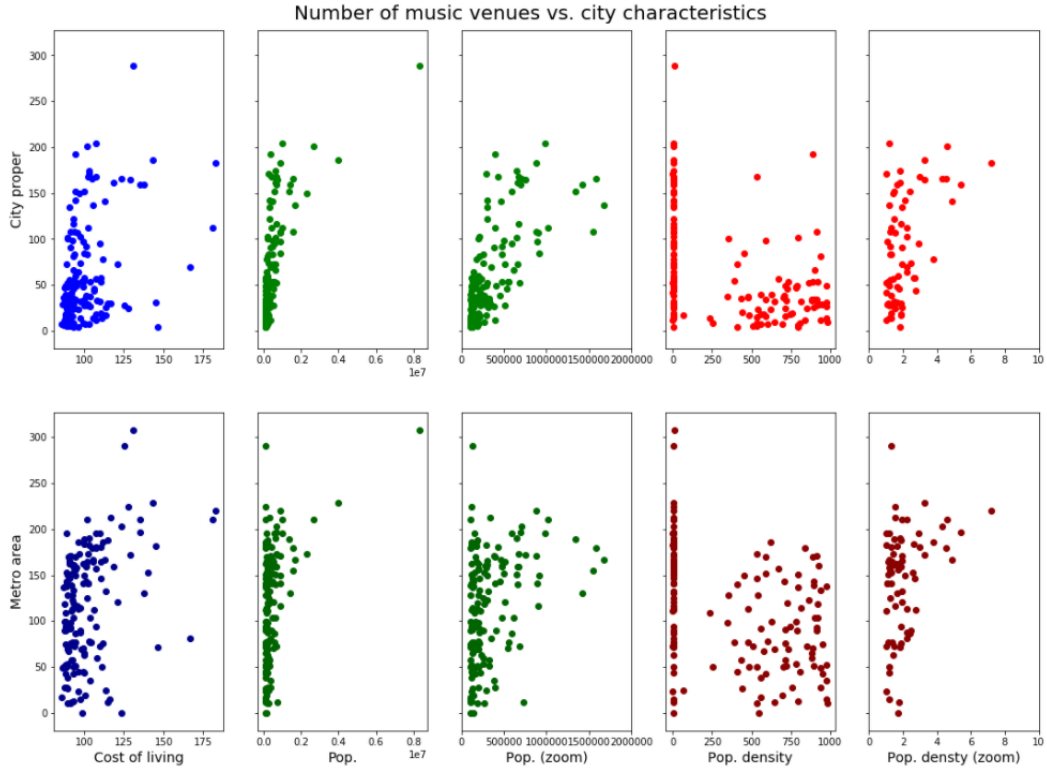

Figure 4: Number of music venues vs. different city characteristics.

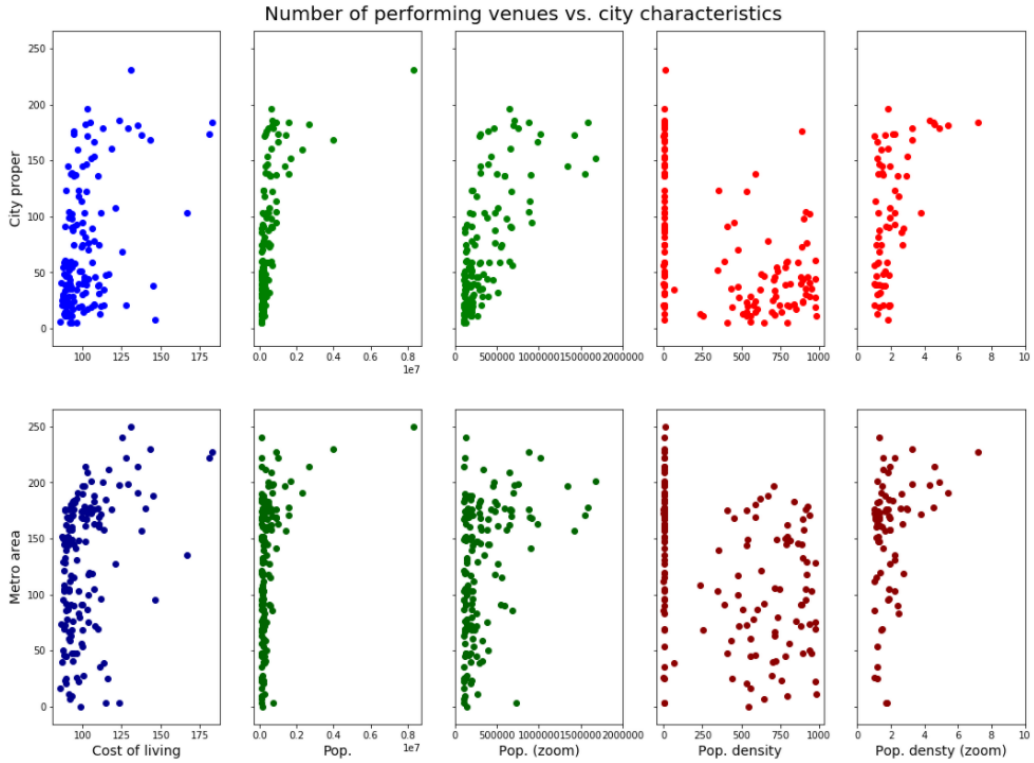

Figure 5: Number of performing arts venues vs. city characteristics.

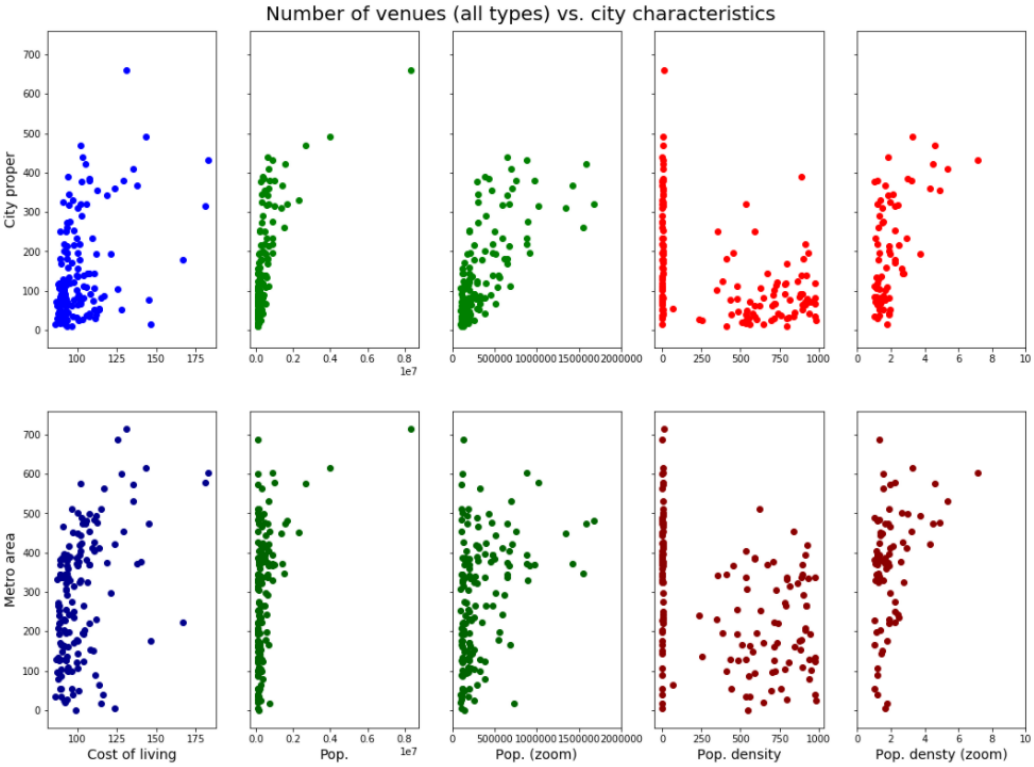

Figure 6: Combined venues vs. city characteristics.

Modelling: Ridge regression with various alpha parameters were used to model the number of each type of venue against city population, population density, cost of living, latitude, and longitude. In general, R^2 values were low, ranging from 0.05 for correlating number of comedy venues in the city proper against the above-mentioned dependent variables, to 0.56 for correlating combined number of variables against the dependent variables.

Maps: Below are four figures. Each has two maps of the United States (left: city proper, right: city metro area), with bubbles located at city locations. The area of each bubble corresponds to the number of venues of a given type in the corresponding city. The color of each venue corresponds to the cost of living for each city. Figures 7, 8, 9, and 10 correspond to comedy, music, performing arts, and combined venues.

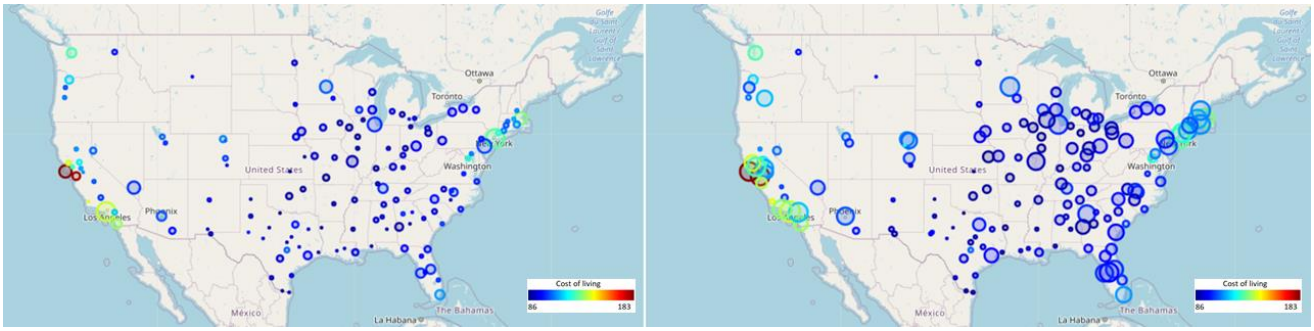

Figure 7: Number of comedy venues in cities across the US. Left-to-right: City proper, greater metro area.

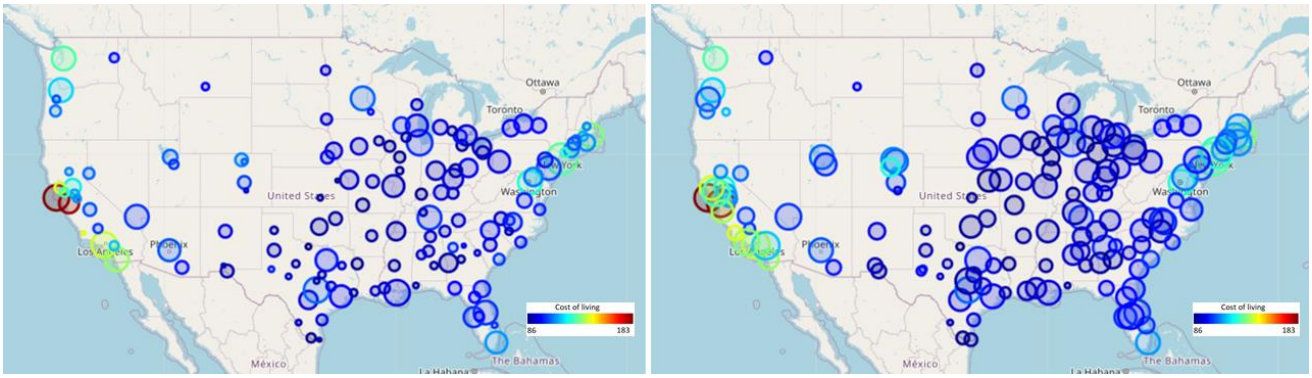

Figure 8: Number of music venues in cities across the US. Left-to-right: City proper, greater metro area.

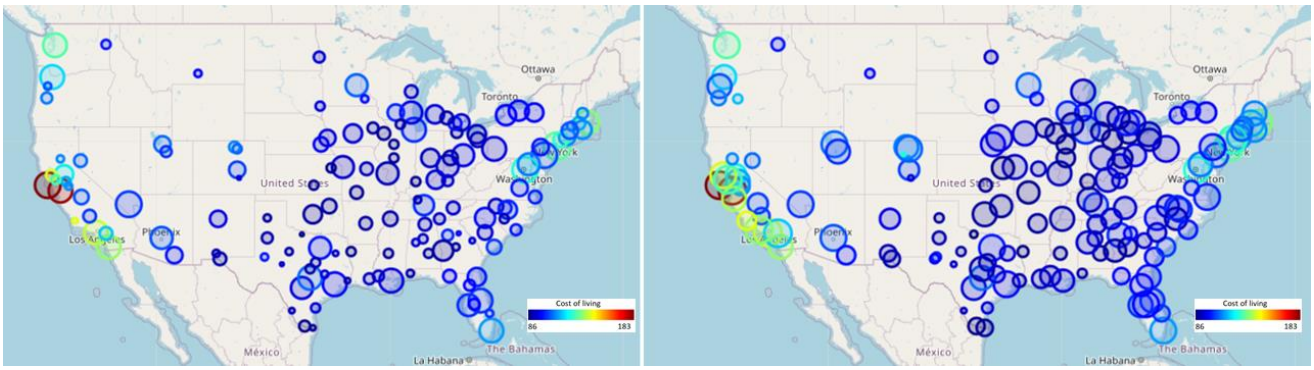

Figure 9: Number of performing arts venues in cities across the US. Left-to-right: City proper, greater metro area.

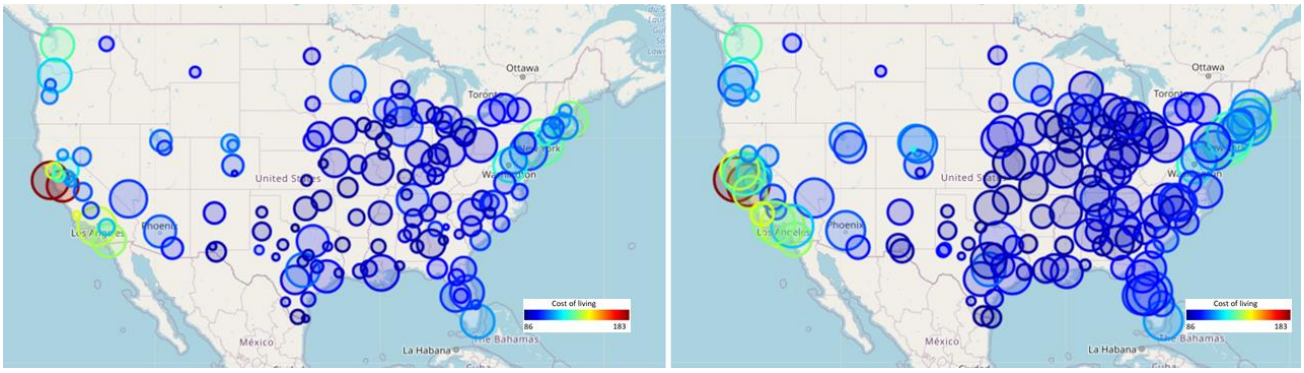

Figure 10: Combined number of venues in cities across the US. Left-to-right: City proper, greater metro area.

Discussion

Much useful information is available from considering the above results.

From the histograms, we know that within city limits, there are only handfuls of cities with large amounts of venues of each type, whereas there are many cities with only a few venues. So, if an artistically-minded person is seeking large numbers of performance-related outlets within city limits, there choices are arguably limited, depending on where they cutoff the number of venues that are acceptable for their interests. However, this becomes less the case if such a person is willing to drive into the greater metro area of a city. Additionally, this issue is most relevant for comedy venues, and less so for music and performing arts.

Regardless to what someone is looking for, the scatter plots (and the poor performance of the regression model) is encouraging. At most cost-of-living ranges and populations / population densities, there are cities that can be found in the US which offer various forms of entertainment and artistic outlet. I.e., one is not forced into a expensive hub or a large or dense area to gain access to fun performances.

As far as where one should look to find venue availability without sacrificing their lifestyle, the maps are very revealing. Large, dark-blue bubbles represent cities with low cost of living, but lots of entertainment. A glance at any of the maps shows that there are plenty of such locations in the United states. In fact, the expensive areas are primarily on the west coast, and to a lesser extent, in the northeast. However, even in these areas, there are some cities which can be found which offer a compromise as far as access to entertainment vs. cost of living. That being said, the region of the US spanning from Texas to the Great Lakes offers a very large number of affordable cities with plenty of entertainment available.

Conclusion

The purpose of this project was to determine to what extent artistic outlets were predicted by socioeconomic characteristics of cities in the United States. More pointedly, it is aimed at providing guidance to people interested in gaining access to performances without sacrificing their quality of life. The results of the analysis are positive for most people. Even if someone does not have much flexibility as far as living location, they can still find entertainment provided that they have the means to leave the city for its greater metropolitan area. More to the point, though, the data showed quite clearly that

socioeconomic city characteristics were only mildly predictive of artistic access. Provided that someone is willing to have fairly reasonable levels of flexibility in choosing where they live, they can find cities with a great deal of fun to be had, without sacrificing their lifestyle (as quantified by cost of living, etc.) to do so.

References

1. Wikipedia, "List of United States cities by population", accessed November 26, 2020. https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population.
2. AdvisorSmith, "AdvisorSmith City Cost of Living Index", accessed November 26, 2020. https://advisorsmith.com/data/coli/.
3. Foursquare Developers API, accessed November 26, 2020. https://developer.foursquare.com/.