Clustering the world’s largest cities

Millennials – a mobile generation

- Millennials (people born between 1980 and 2000) comprise over one third of the global professional workforce.
- >50% of millennials expect to have 2-5 employers in their lifetime.[1]
- For more than 60% of millennials, seeing or experiencing the world is their topmost priority.[2]
- International moves are primarily driven by career progression and salary, but several other factors such as expenses, quality of life and cultural immersion are also important.

[1]: PWC, Millennials at Work – Reshaping the workplace

[2]: Deloitte Global Millennial Survey 2019

This exercise aims at clustering the world’s largest cities based on the following parameters

Demographic

Population

Area/Size

GDP per capita

Population density

GDP

Cost of living

Non - Demographic

Venue distribution around city center

Sources of data

Wikipedia

- List of the world's 80 largest cities, along with their size and population[1]
- City – wise GDP data[2]

- Population density = Population/Area
- GDP per capita = GDP/Area

[1]: The largest population and area measures within a row were considered
[2]: The PPP adjusted GDP was used for calculations

Numbeo

- Cost of living indices

- Cost of living, Rent, Grocery and Restaurant price indices were considered.[3]

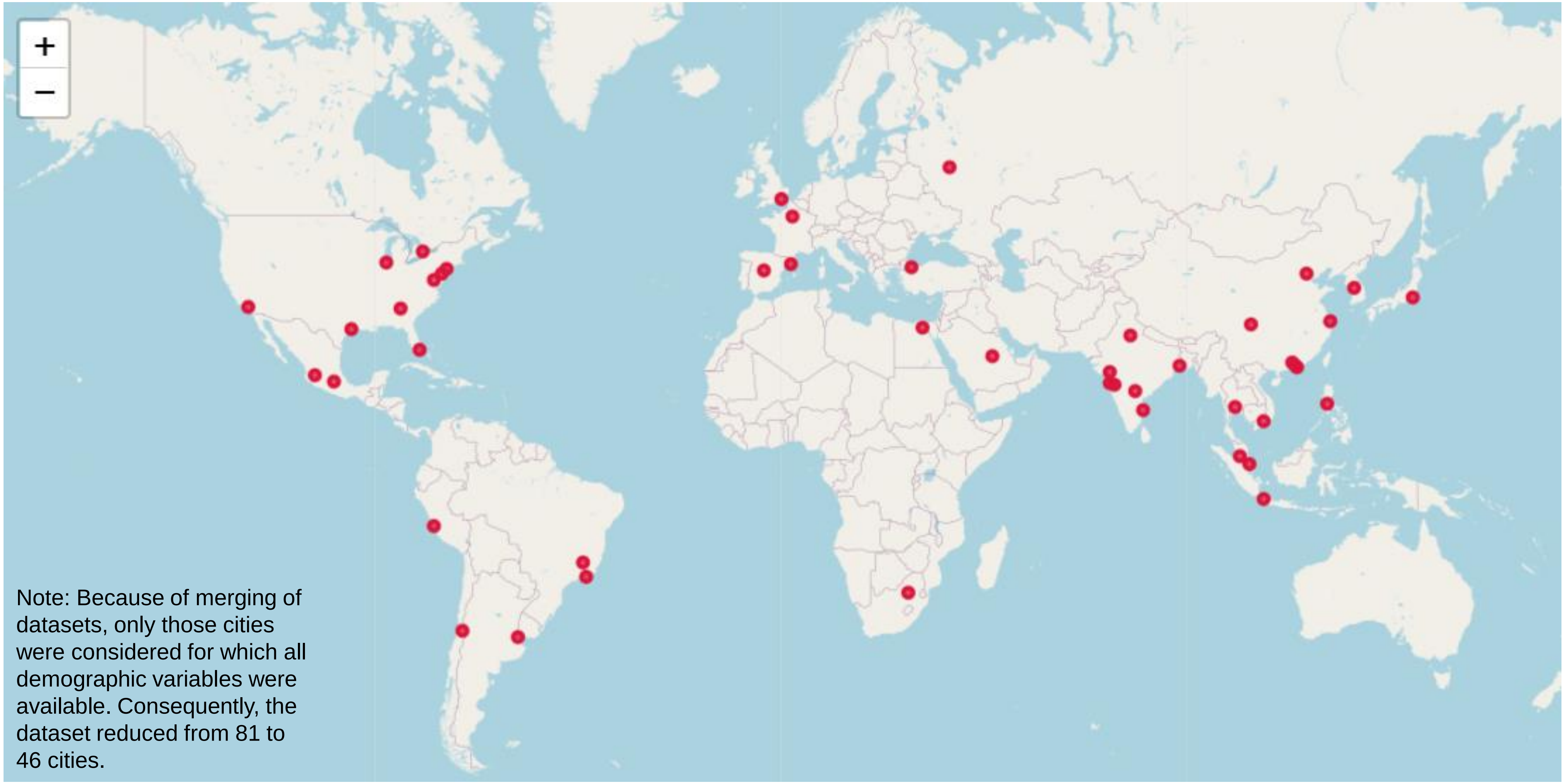
[3]: Cost indices are defined based on NYC prices – a cost of living index of 75 means the city is 25% cheaper than NYC.

FourSquare

- The FourSquare API was used to query a list of 500 locations within a 10-kilometer radius from the city centre[4]. All venue types were considered.

[4]: The Nominatim method from the geopy library was used to return the latitude and longitude values of each of these cities

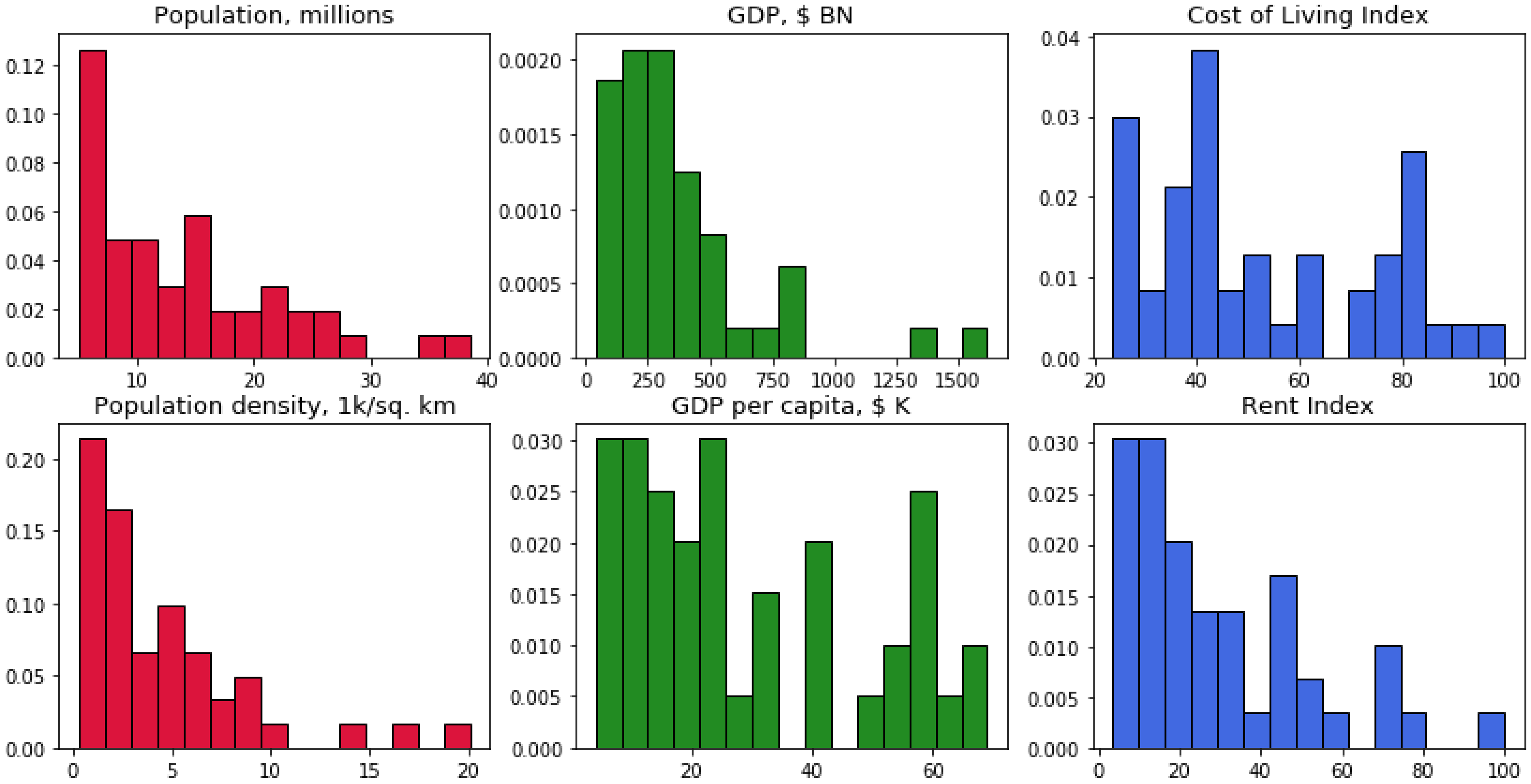
Where are our cities located?



Note: Because of merging of datasets, only those cities were considered for which all demographic variables were available. Consequently, the dataset reduced from 81 to 46 cities.

Analysis of Demographic Data

Distribution of population, population density and rent index is mostly skewed towards the left

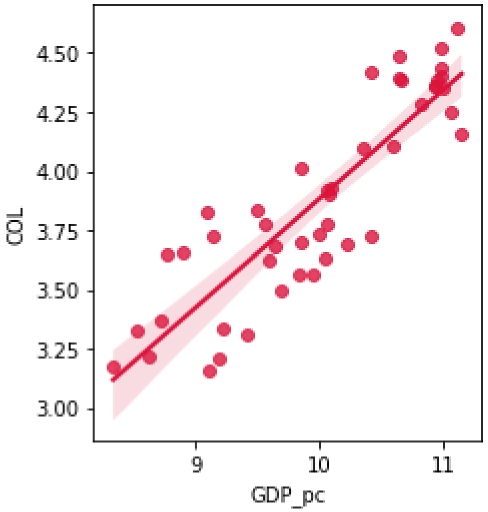


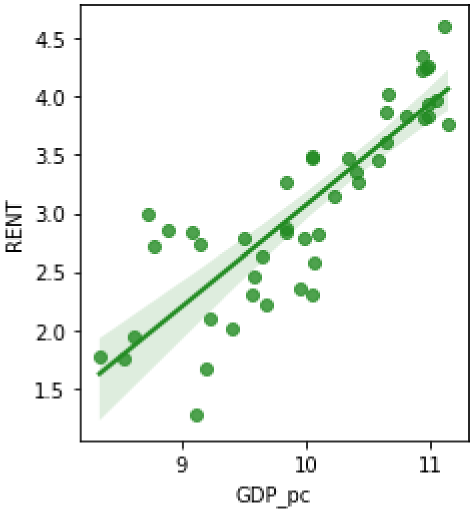
Cities with

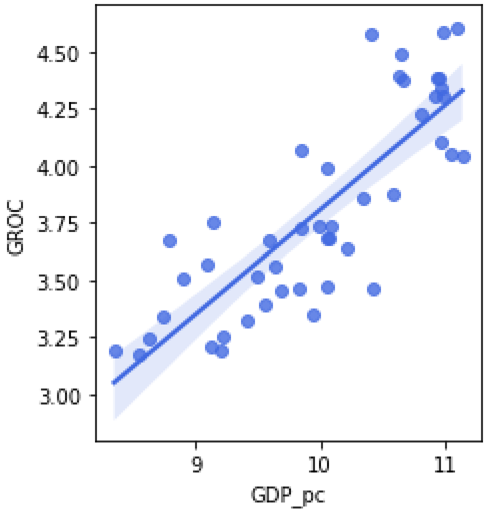
- Population more than 30 million: Tokyo and Jakarta
- Population density >10k/km^2: Manila, Chennai, Surat and Philadelphia
- GDP > 1 Trillion: Tokyo and New York

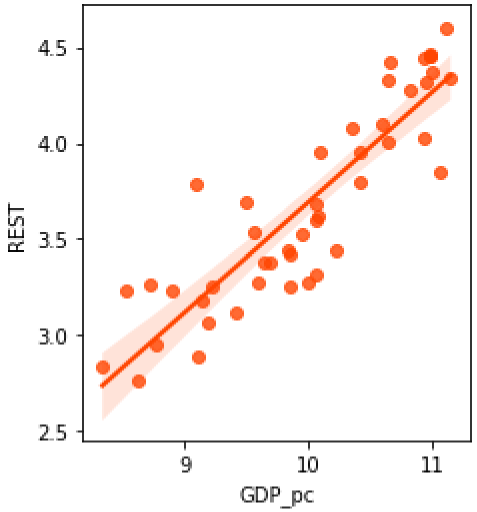
All plots are density plots, ie the area under the histograms is 1.0.

The wealthier cities are also more expensive









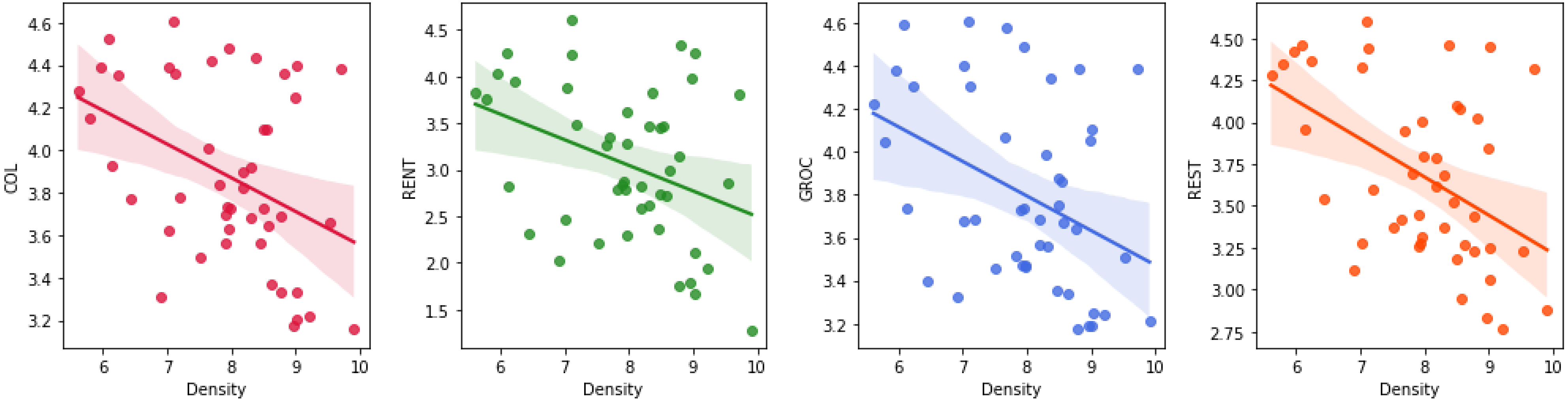
- Strong (R^2>0.7) positive correlation observed between GDP per capita and all cost indices

- Rent seems to be the most affected by GDP per capita – for a 1% increase in GDP pc, the rent index increases by 0.87%

Index	**Coefficient**	**R^2**
Rent	0.871431	0.723006
Restaurants	0.577262	0.789755
Cost of Living	0.461144	0.769275
Groceries	0.456155	0.789755

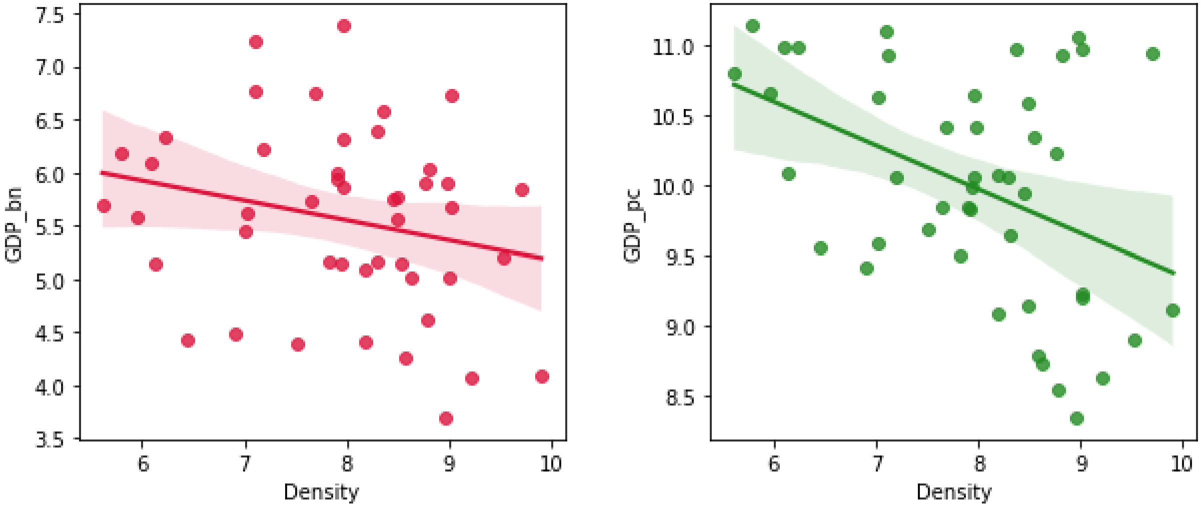
All plots and correlations were built on a log – log scale, as the orders of magnitude of both the x and the y variables were very different

No correlation[1] between density and cost indices



Index	Coeff.	R^2
Cost of Living	-0.16	0.17
Groceries	-0.16	0.23
Restaurants	-0.23	0.23
Rent	-0.28	0.13

Or between density and GDP



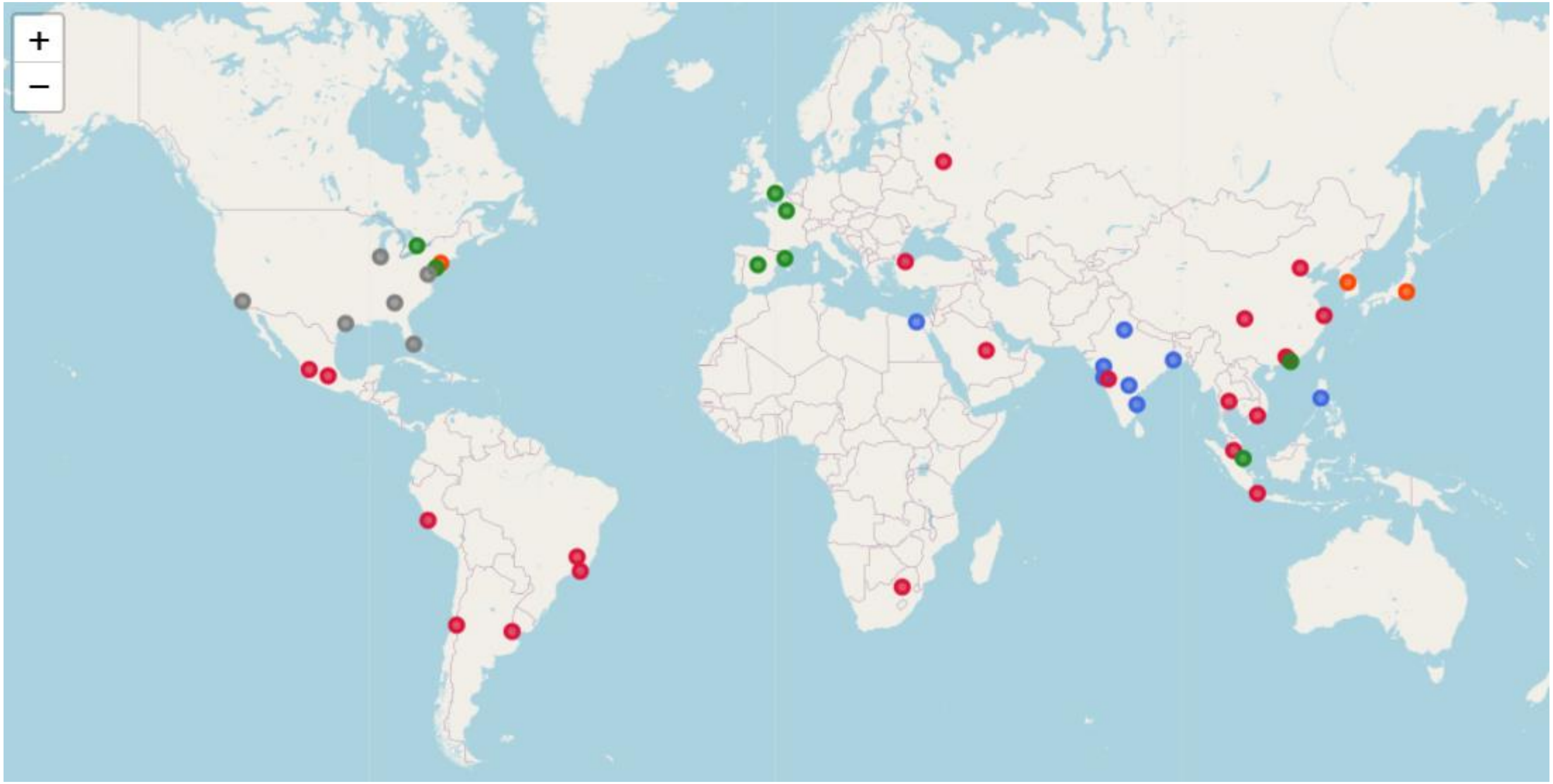
GDP measure	Coefficient	R^2
Total	-0.19	0.06
Per Capita	-0.31	0.18

All plots and correlations were built on a log – log scale, as the orders of magnitude of both the x and the y variables were very different

[1]: While the coefficients are negative, the R2 values are too small to conclude a reasonable correlation between the variables

Demographic clustering

Demographic clustering - results

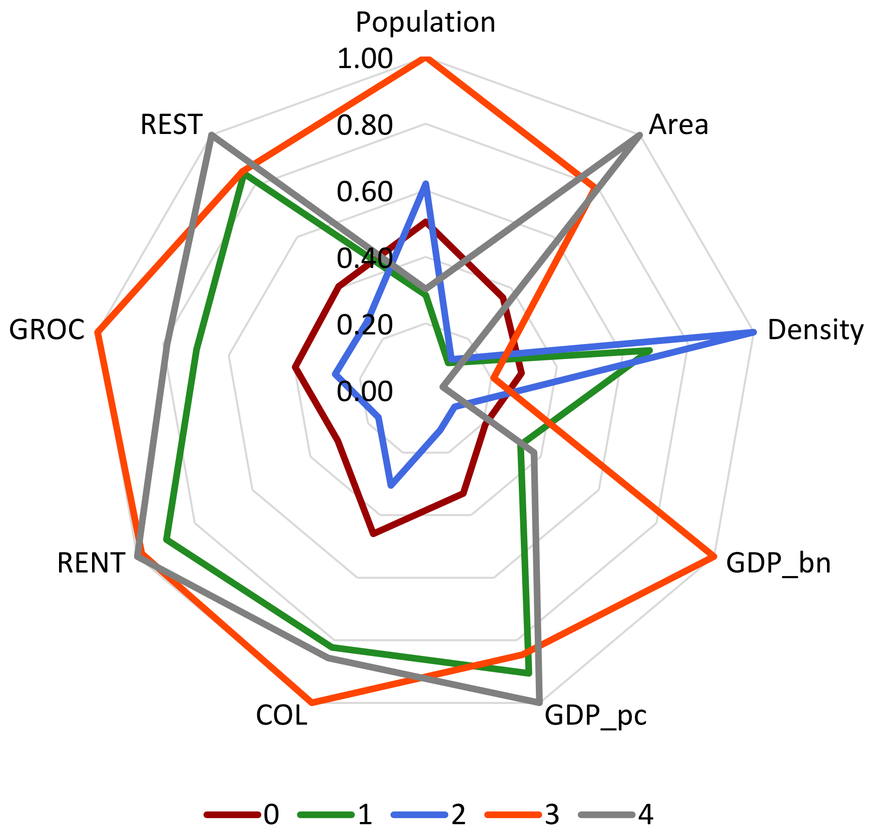


Demographic clustering – cluster characteristics

Cluster	Population	Area	Density	GDP(bn)	GDP(pc)	COL	RENT	GROC	REST	Cities
0	14322316.1	6464.5	2936.7	266.4	18480.8	41.6	17.1	37.9	32.7	Shanghai, Mexico City, Beijing, Buenos Aires, Istanbul, Rio de Janeiro, Guangzhou, Moscow, Shenzhen, Jakarta, Lima, Bangkok, Chengdu, Ho Chi Minh City, Kuala Lumpur, Riyadh, Santiago, Pune, Belo Horizonte, Johannesburg, Guadalajara,
1	8088271.5	1890.4	6862.6	423.6	50600.3	74.4	50.4	66.6	68.3	Paris, London, Hong Kong, Madrid, Toronto, Singapore, Philadelphia, Barcelona,
2	17557500.0	2149.1	10051.3	129.8	7175.8	27.6	9.2	26.3	21.8	Delhi, Cairo, Mumbai, Kolkata, Manila, Chennai, Hyderabad, Surat,
3	28354666.7	14157.0	2085.9	1288.6	47271.9	90.4	55.2	95.4	68.9	Tokyo, New York, Seoul,
4	8593312.5	17863.8	528.3	484.4	55842.4	77.4	56.1	75.2	80.5	Los Angeles, Chicago, Houston, Miami, Atlanta, Washington, D.C.,

- Cluster 0 has the most number of cities (almost 50% of the dataset)
- The location of the clusters also seems influenced by their geography. For example, cluster 2 contains only Indian cities.

Demographic clustering – cluster descriptions



- **Cluster 0**: Bulk of the cities fall into this category. They do not have any marked characteristic, as all the parameters are average, and no distinction is seen. From a geographical perspective, these cities are all located in developing countries.
- **Cluster 1**: These cities are marked by small size, moderately high population density, high GDP per capita and high cost indices. Most of them are European.
- **Cluster 2**: These cities are characterized by their small size, low GDP, low cost indices and extremely high population density. One can expect them to be high poverty, congested cities. Most of them are located in India.
- **Cluster 3**: These cities have large populations, extremely high GDPs and cost indices, but low population density. Only three cities - NYC, Tokyo and Seoul fall in this category.
- **Cluster 4**: These cities are similar to cluster 4 but are even bigger in size. They can best be described as rich, large cities. They are all located in the USA.

Since all demographic variables have different orders of magnitude, they were divided by the maximum value in their column to reduce their range to [0,1]. This makes visualization on a radar chart easier.

Further improvements

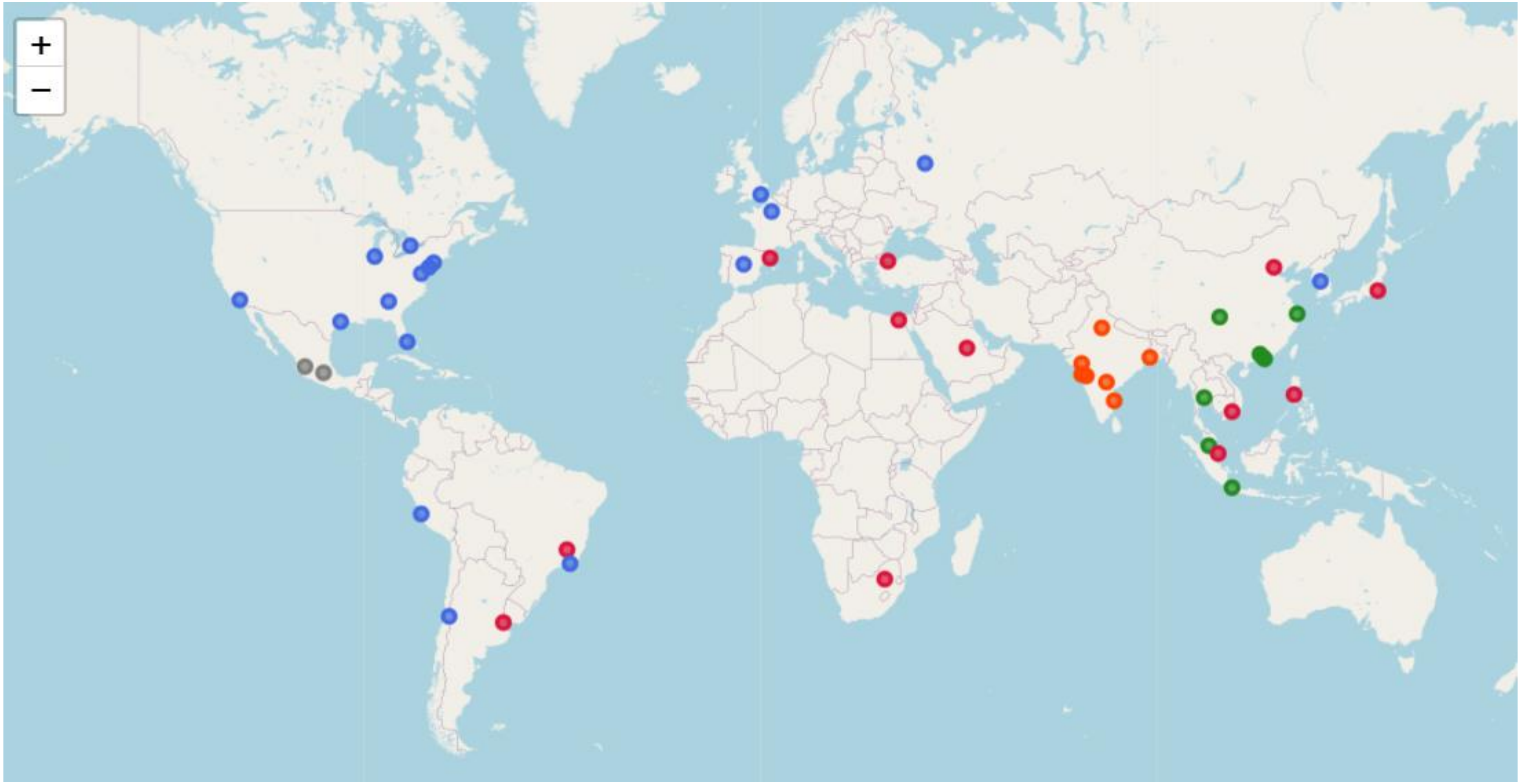
- Use actual income levels in the city (the GDP per capita was used as a proxy for wealth, but is not entirely accurate as it includes government spending as well)

- Also consider the strength of the local currency.

- As far as the foursquare data is concerned, the current exercise did not distinguish between types of venues. However, when deciding to move overseas, some venues are more important than others. For example, banks, schools and medical facilities might be considered more important than restaurants. For a better analysis, 4 or 5 such macro categories can be defined and the number of venues that fall under each of them can then be used for comparison

Venue - based clustering

Venue based clustering - results



Venue based clustering – cluster characteristics

Cluster	Cities
0	Barcelona, Beijing, Belo Horizonte, Buenos Aires, Cairo, Ho Chi Minh City, Istanbul, Johannesburg, Manila, Riyadh, Singapore, Tokyo,
1	Bangkok, Chengdu, Guangzhou, Hong Kong, Jakarta, Kuala Lumpur, Shanghai, Shenzhen,
2	Atlanta, Chicago, Houston, Lima, London, Los Angeles, Madrid, Miami, Moscow, New York, Paris, Philadelphia, Rio de Janeiro, Santiago, Seoul, Toronto, Washington, D.C.,
3	Chennai, Delhi, Hyderabad, Kolkata, Mumbai, Pune, Surat,
4	Guadalajara, Mexico City,

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Hotel	Hotel	Park	Indian Restaurant	Ice Cream Shop
Coffee Shop	Shopping Mall	Coffee Shop	Hotel	Mexican Restaurant
Café	Coffee Shop	Hotel	Café	Taco Place
Park	Park	Bakery	Multiplex	Seafood Restaurant
Bakery	Chinese Restaurant	Trail	Restaurant	Bakery

As compared to demographic clustering, venue-based clustering results are also grouped by geography. This may be because of the kind of venue data provided by FourSquare.

- Cluster 0 is located across all continents with no special characteristics.
- Cluster 1 is mostly present in SEA and China, marked by Chinese restaurants.
- Cluster 2 has cities in the US and Europe, marked by a high density of parks.
- Cluster 3 contains cities that are all located in India, with Indian restaurants and multiplexes among the top 5 venue categories.
- Cluster 4 contains two Mexican cities, marked by taco places and Mexican restaurants.