Do Similar Cities have Similar Traffic

Urban planners make decisions every day that change how a city grows. They choose to allow patterns of development, including various types of commercial and industrial development, as well as educational and transport facilities. Often, these are considered in isolation: an educational facility increases the education of the residents, and a transit facility improves transit times. However, what if there are hidden side-effects? Could development that affects the “feel” of a city change the way that development occurs within it, putting the cities on course for unforeseen problems related to poor planning?

For this project, I can’t directly consider “feel”, because it is impossible to define or measure. Because I can’t classify cities based on an objective measure, I will use an unsupervised machine learning, specifically the K-Means clustering algorithm. The cities will be grouped together based on a proxy for feel: the types of venues they host. I think that these could provide a glimpse at the makeup of a city, as the venues found in a city are the result of the same historical and geographical factors that give it its feel. Similarly, I don’t directly define the “goodness” of urban planning. Instead, I consider the level of congestion in a city.

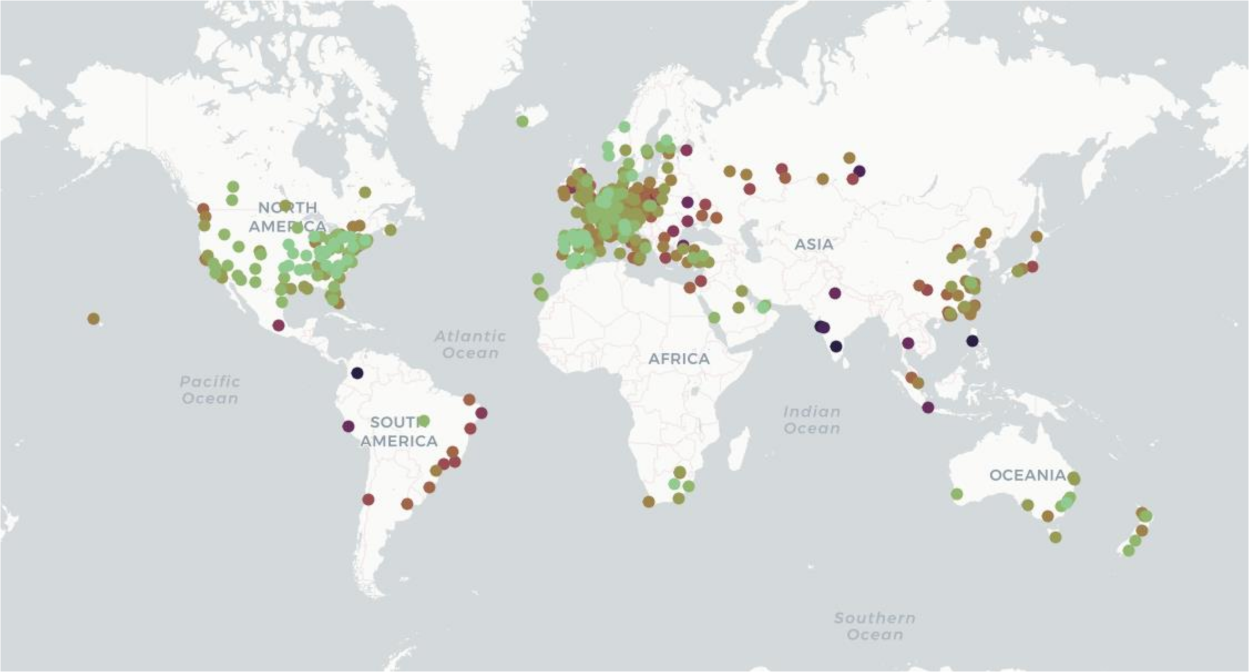
By clustering the cities, and examining the relationship between each cluster and the level of congestion present in each city in a cluster, we will be able to see if similar cities have similar levels of congestion. If this is the case, it would suggest that actions taken by urban planners may have unforeseen side-effects that can cause problems in the future, such as bad traffic congestion.

The Data

To carry out the analysis, we first have to find some data. For this project, I used data from Tomtom, a GPS and mapping company, and Foursquare, a geolocation data provider.

Since looking at traffic data is the focus of this project, it was important to find some measure of congestion in various cities. Luckily, Tomtom have done the heavy lifting here, analyzing the data they get from the vast array of their GPS systems travelling around the world. Here, they have calculated a congestion score for each of 416 cities around the world. They have a simple explanation of this congestion score: “A 53% congestion level in Bangkok, for example, means that a 30-minute trip will take 53% more time than it would during Bangkok’s baseline uncongested conditions.” I used the Beautifulsoup Python package to scrape the information from Tomtom’s website.

The congestion data for each city is shown on the world map below. The least congested city was Greensboro High Point, US with a congestion level of 9%. The most congested city was Bengaluru, India with 71%.



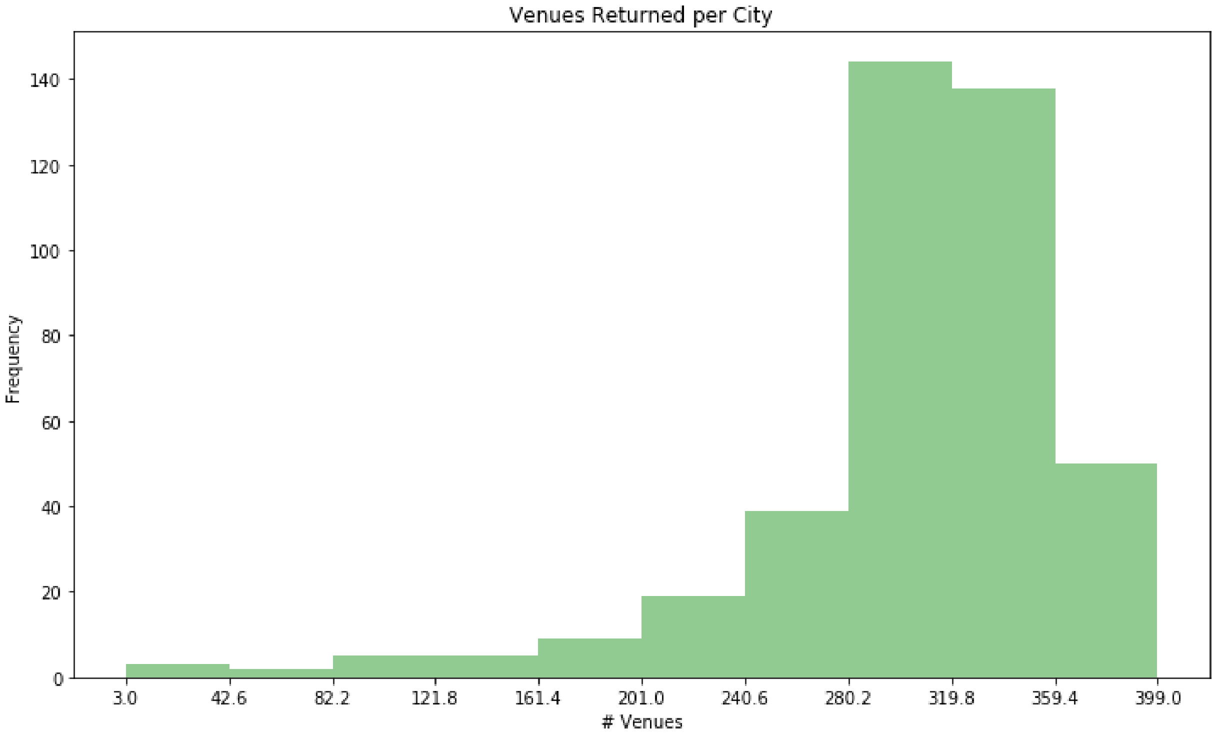
City Congestion (Green = Least Congested, Purple = Most)

To gauge how similar different cities might be, I needed to find some type of data that gave an insight into what types of venues were in each city. For this I used the Requests Python package to access Foursquare's API and download lists of venues for each city. The API limited responses to a maximum of 50 venues under each of 9 categories:

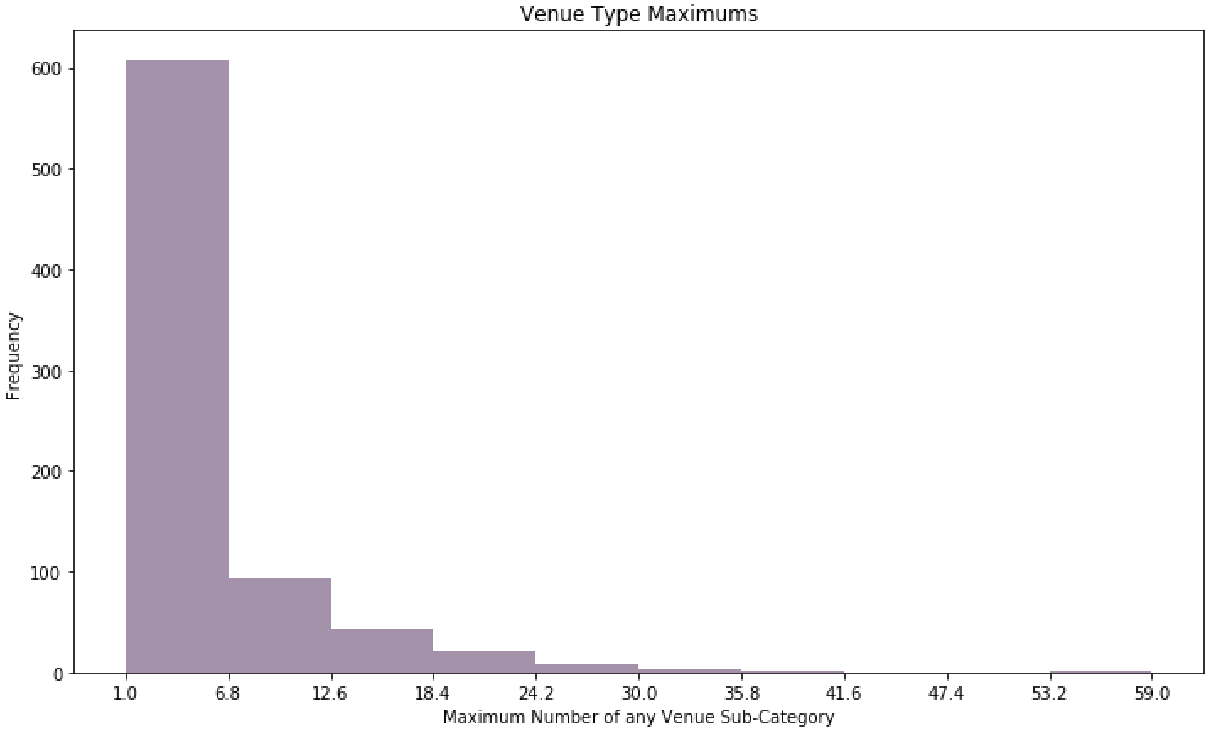
- Arts & Entertainment
- College & University
- Event
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional & Other Places
- Residence
- Shop & Service
- Travel & Transport

The data returned from the API includes a list of venues. Each entry has the venue name, it's sub-category within one of the 9 categories above (e.g. "Coffee Shop" would fall under the "Food" category), and the Latitude and Longitude of the venue.

The data returned on the venues found in each city is visualized below. Most cities returned at least 280 venues each:



Also, there were a great many types of venues returned, and most types were returned less than 12 times, so there was a good variety of venues returned.



Data Processing

The data processing started with cleaning the data set, including the review and correction of data types, removing observations with missing values and reviewing duplicates. Some inferential statistics of the data set were reviewed, for example:

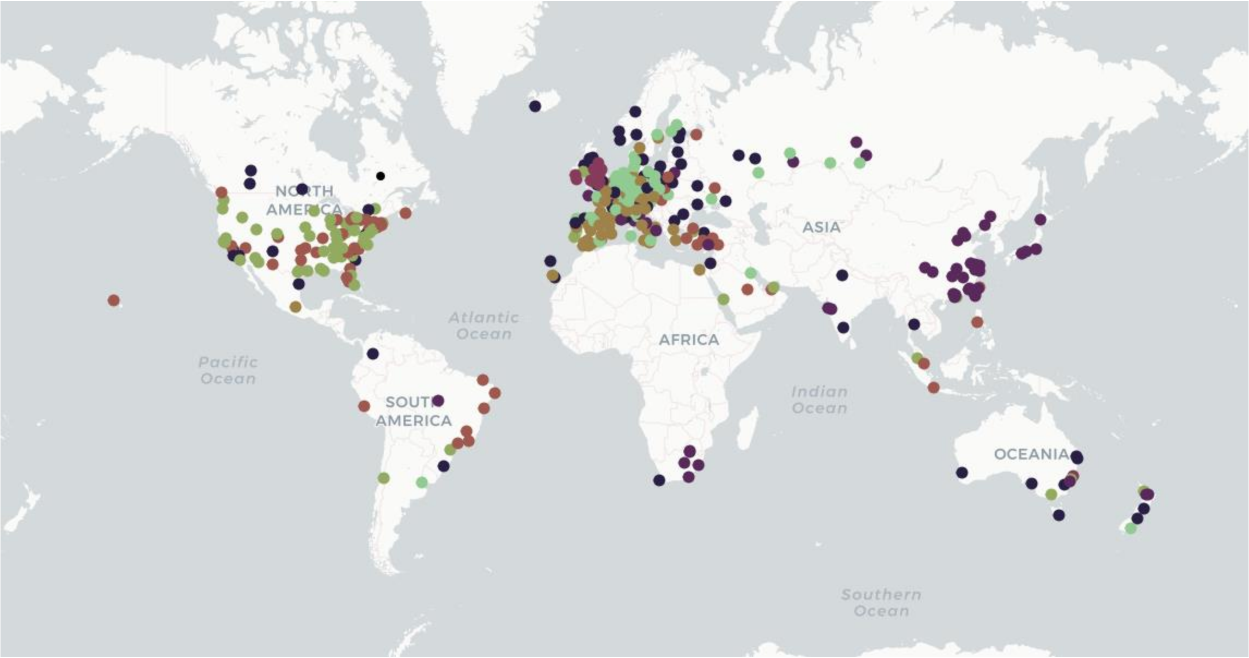
- Average congestion level was 26%.
- Standard deviation of congestion level was 10%.
- Maximum congestion level was 71% (Bengaluru, India)
- Minimum congestion level was 9% (Greensboro High Point, US)

Next, the data had to be normalized in order to use the K-Means algorithm. In this case, as can be seen in the chart above, some types of venue sub-category occurred many times more than others. To illustrate what this means, consider that a city with many airports may have 3, whereas a city with many coffee shops may have 1,000 (though in this case Foursquare limits us to viewing 60). We want both of these metrics to be comparable, so that the measurement of a city with many airports is similar to that for a city with many coffee shops. To achieve this, the number or each sub-category (think coffee shop) was divided by the maximum number of that sub-category in any city. The resulting metric is a number between 0 and 1.

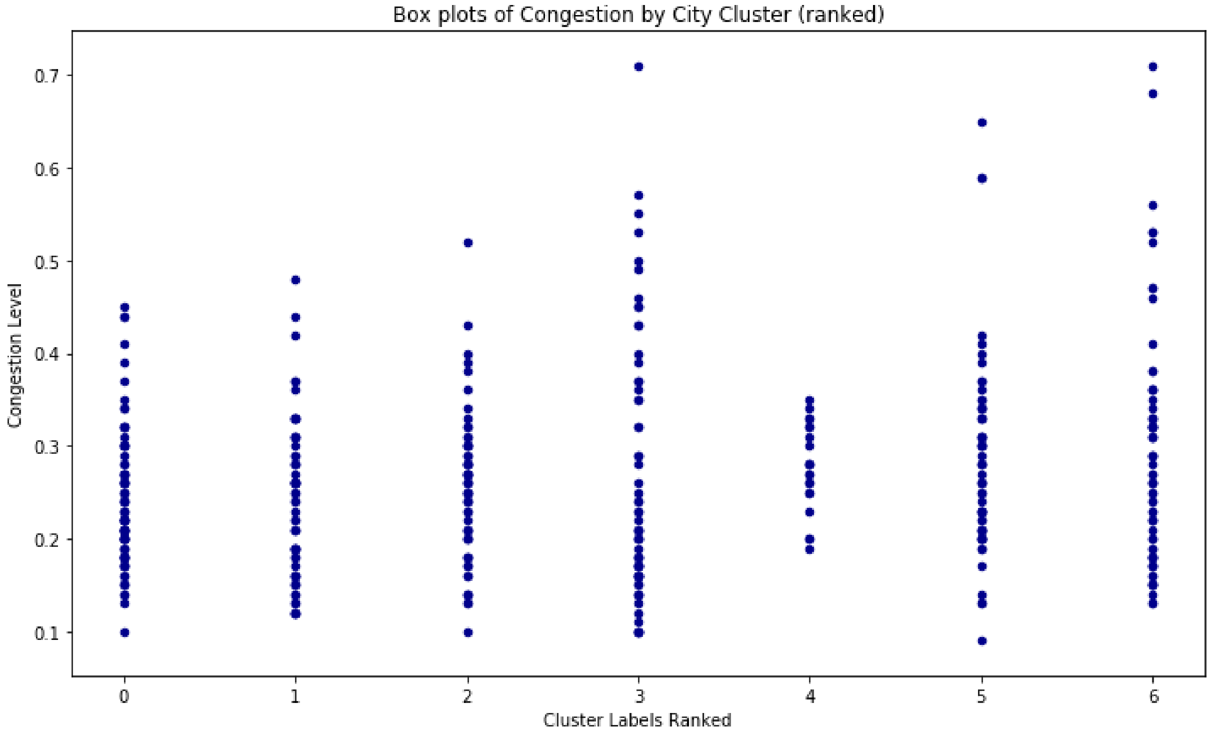
Clustering of the cities was then carried out based on the number of each venue in each sub-category. The resulting clusters were then ranked from lowest average congestion of the cluster to highest, to make visualization a little clearer.

Results

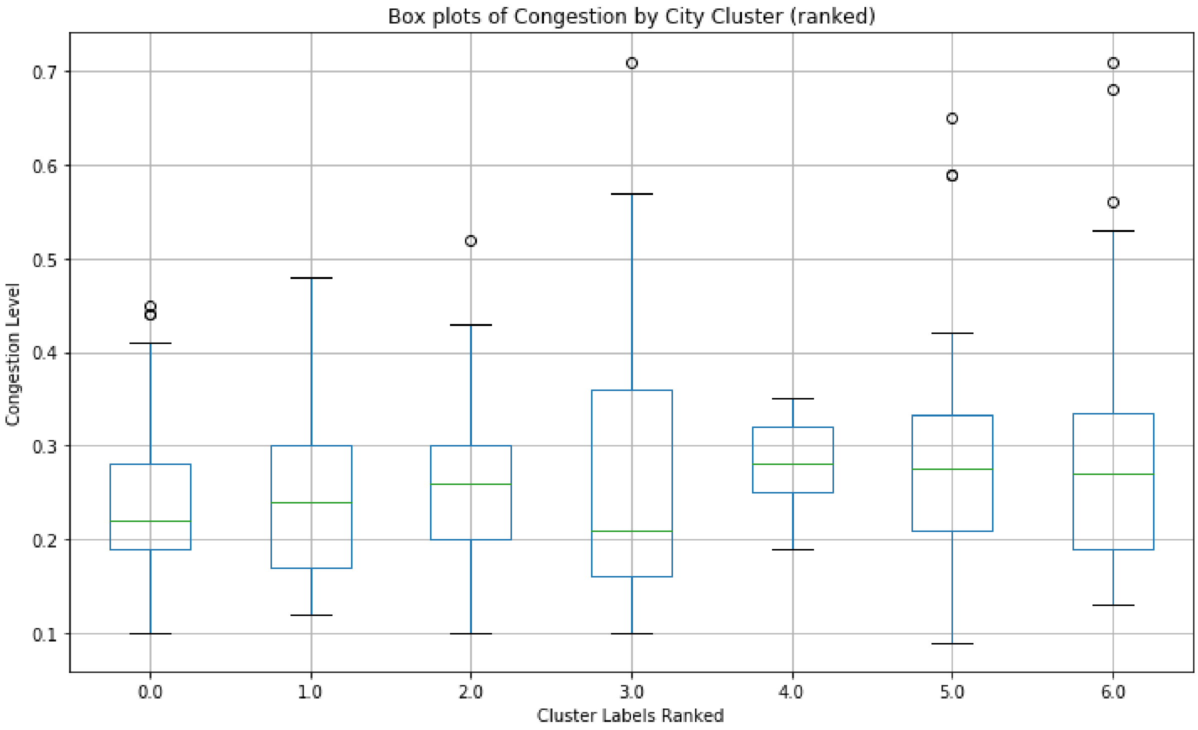
The results are visualized in the plots below. The cities are coloured according to the cluster to which they belong. The clusters in turn are ranked from least congestion (in green) to most congestion (in purple). The congestion level in each city is also plotted below.



City Clusters, Ranked by Congestion Level (Green = Least Congested Cluster, Purple = Most Congested Cluster)



Congestion Level of Each City within each Cluster



Congestion Level of Each City within each Cluster

The ranked clusters appear to show a difference, which appears to indicate that different types of cities (per the output of K-Means clustering) may lead to different levels of congestion. In fact, a basic statistical assessment of the lowest cluster, 0, and the highest, 6, suggests that they are statistically different with a p value of 0.997. However, there are some issues recognizable in the visualizations that call this result into question, as discussed below.

Discussion

Recall the initial question posed at the beginning of this piece: “Do certain types of city just have bad traffic?”. Types of cities were considered to have a different “feel” as measured by a proxy value: the types of venues found within each.

At first glance, the results seem to suggest that this is so, but on closer inspection we find that there are likely other stronger factors at play.

First, the City Clusters show a strong geographical bias. This suggests that the venues found in each city more strongly reflect the geographical location of the city than they do the “feel” of the city. While these factors may be related, we have not sufficiently demonstrated that it is so.

Second, by ranking the clusters output by the K-Means algorithm, we have by definition sorted the cluster with the lowest congestion from that with the highest, and in so doing, may well have created a data-mining bias in our result. One could imagine that the results of a random number generator, similarly sorted into 7 groups and ranked, could show a similar trend. A thorough statistical analysis would be required to demonstrate that the results are free of this bias.

Finally, the number of venues available for each city was limited to 60 for each category of venue by Foursquare. While the results returned were the “most popular”, we do not know how popularity is determined, nor do we know how much confidence there is in that popularity from city to city. Even if we were comfortable with the calculation methodology, there is also the issue of sample size in that calculation. For example, while we might have some confidence in numbers where Foursquare has many users, we cannot have the same confidence in a place that might have only a handful. What if the popular venues in a city were determined only by 5 travellers from out of town? We do not have enough information to evaluate this question.

Results

In summary, we cannot be confident in the outputs of the model.

Nonetheless, it seems obvious that the actions of Urban Planners often have unintended effects, that we must live with for years, maybe even decades. We can always urge our local governments to place a high value on Urban Planning, and contribute to the process when we can.