Contents

[Goal of this analysis 1](#_Toc33728138)

[Big discrepancy in data 1](#_Toc33728139)

[Pre-processing 1](#_Toc33728140)

[Deduplication 1](#_Toc33728141)

[Assignment of requests to POIs 1](#_Toc33728142)

[Calculate POI wise standard deviation and mean on distances 2](#_Toc33728143)

[Plot requests with POI as circle and find radius and density 2](#_Toc33728144)

[***Note:*** 2](#_Toc33728145)

[Outliers 2](#_Toc33728146)

[Scaling 4](#_Toc33728147)

[Conclusion 4](#_Toc33728148)

[Why only distance and not include other parameters? 5](#_Toc33728149)

[Why is scaling required? 5](#_Toc33728150)

[Alternative approach 5](#_Toc33728151)

# Goal of this analysis

I guess the requests are analyzed based on the POIs to check how popular the POIs are, the traffic, the places which contribute to producing most requests, unusual requests as well as (probably) whether more POIs are needed to facilitate efficiency.

# Big discrepancy in data

Data shows the below latitude longitude co-ordinates from Montreal, Canada. While geocoding on Tableau, I found out that the below co-ordinates are from East Asia. There are 7 of them corresponding to POI1 which are wrongly marked as from Canada. There are 15 more corresponding to POI4 which are scattered around France, Iran and so on, but labelled as Canada.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CA | QC | Montreal | 6.92742 | 122.02588 | POI1 | 11531.84 |

# Pre-processing

### Deduplication

* Format column names
* Deleted duplicates in request file based on timestamp, latitude and longitude
* Deleted duplicates in POI file based on latitude and longitude = makes it 3 POIs

### Assignment of requests to POIs

* Used haversine distance to calculate distance between 2 pairs of longitudes and latitudes in kilometers
* Selected the POI with the shortest distance for one request and assigned to it

### Calculate POI wise standard deviation and mean on distances

SD Mean

POI1 412.430602 301.907275

POI3 223.350867 451.528179

POI4 1472.939834 497.279414

Since distance cannot be negative, POI4 having a huge standard deviation means it has a considerable number of data points which are very spread out from the mean (outliers)

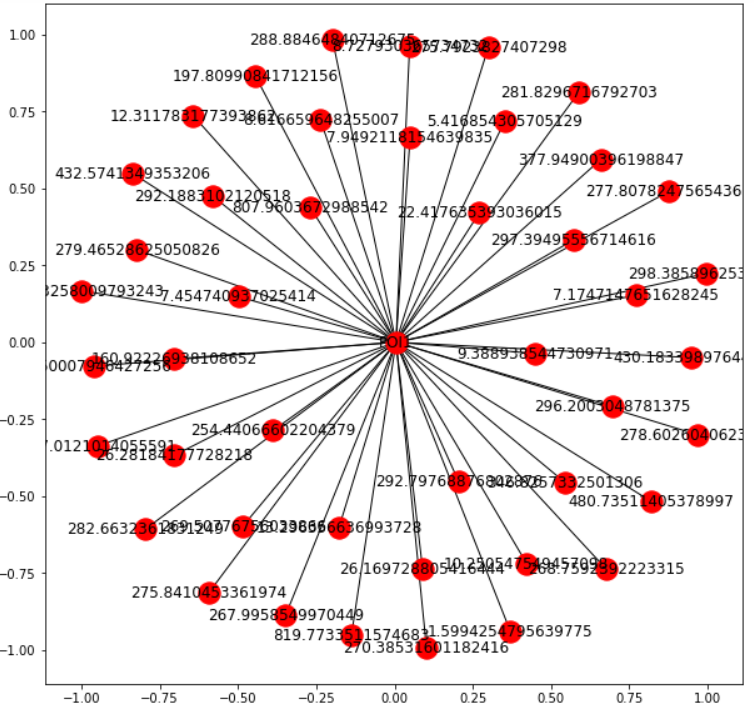
### Plot requests with POI as circle and find radius and density

For poi **POI1**, number of requests assigned: 9727, radius: 11531.836760265789 and density: 2.328263184907089e-05

For poi **POI3**, number of requests assigned: 9795, radius: 1474.5829988065425 and density: 0.0014

For poi **POI4**, number of requests assigned: 477, radius: 9349.585684667305 and density: 1.7369358307478257e-06

To plot a circle with the POI as the center and the distances as edges from the center, it is impossible to visualize all the data points clearly due to large number of data points. So, I minimized the number of edges and it gets me the below figure:

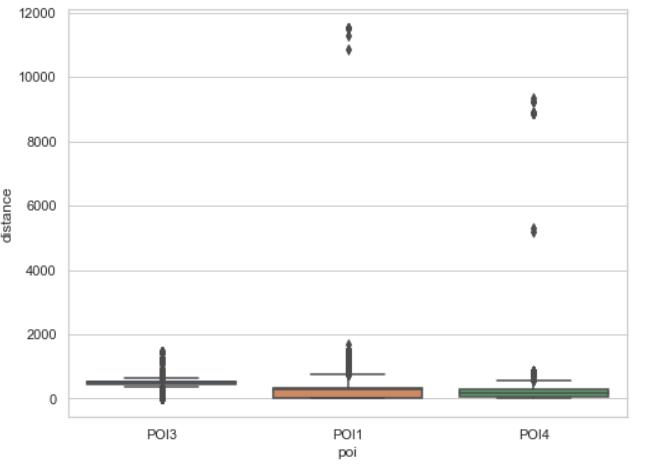


#### **Note:**

*What is the purpose of doing this? While plotting the graph automatically scales the distances, yet we are unable to find any information related to what these distances denote or the direction of bigger distances (outliers, if any) or the different clusters of points based on different distance ranges*

# Outliers

Each of the POIs have requests within 100 meters to tens of thousands of kilometers. While the close ones cannot be considered as outliers, I put focus on the distant ones. Box-plotting them according to the POIs:



**POI1** - The visible outliers lie outside the range of 10000kms and they are only 7 in number. The spread of the rest of the data is within 2000kms. Even after eliminating the outliers, I notice that 75% of the data points lie within the 400 kms range and within this range standard deviation becomes very low. Yet, the 400-2000kms requests should not be considered as outliers, I believe.

**POI4 -** Outliers lie in the range of 1000-10000kms and they are 7 in number. Also, unlike POI, the rest of the data are clustered within 900kms. Since the total count of requests are quite low (477) for POI4, contrary to POI1(9727), the effects of outliers are big here.

**POI3** - I find no outliers outside the 2000kms range. While 75% of the data lies within 600kms yet filtering them based on it does not decrease the standard deviation by 2-3 times.

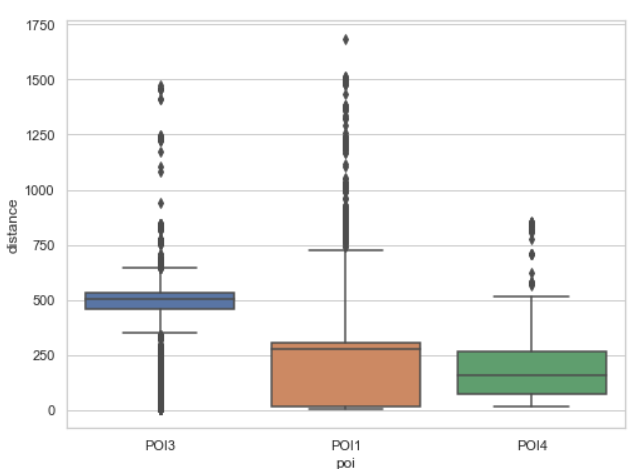
poi count mean std min 25% 50% 75% max

POI1 9727 301.9072 412.4306 0.3484 15.2459 277.5748 314.191 11531.8367

POI3 9795 451.5281 223.3508 0.8093 457.9798 504.9911 532.0725 1474.5829

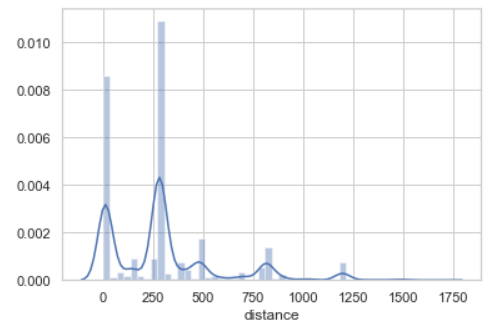
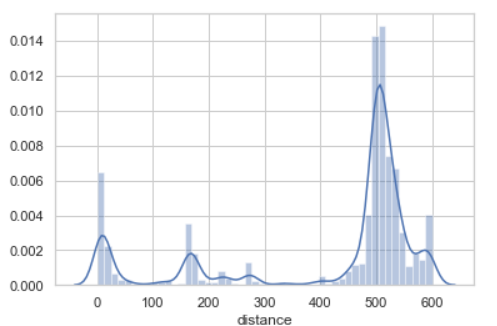
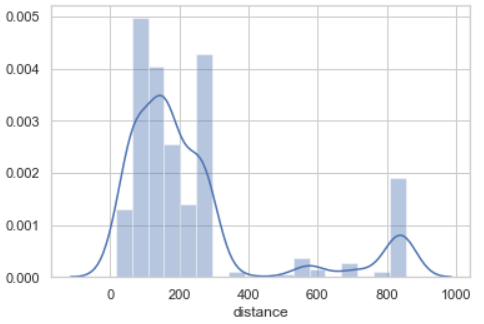
POI4 477 497.2794 1472.9398 16.2290 75.2764 157.0311 275.3848 9349.5856

The boxplot without outliers is the best for POI3 because of the reasons given above but still skewed for POI1 and POI4 due to unequal distribution of still larger distances, which cannot be considered as outliers.



# Scaling

I checked the distribution of the distance column after eliminating the outliers. All of them exhibit multi-modal behavior.

POI1 POI3 POI4

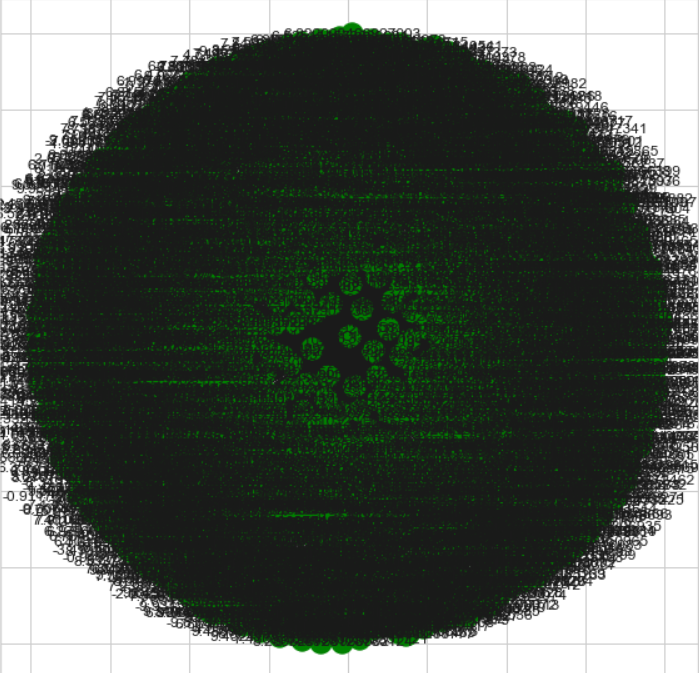
Since z-score normalization is not advisable to apply on non-normal data, I did the linear scaling with min-max normalization. The minimum and maximum are bounded from -10 to 10 by multiplying with a scaling factor: (range2 – range1)/(max – min), where range2 = 10, range1 = -10, max = max(distance) and min = min(distance). The formula looks like below:

factor = (range2 – range1)/(max – min)

scaled\_val = ((x - min) \* factor) + a

**Visualization after scaling:**

I believe I probably did not get what is asked for, but using networkx graph structure, I still could not visualize the edges perfectly because I did not eliminate too many points as outliers and with 9000 requests in POI3, the graph is not clearly visible.



# Conclusion

Since I failed to understand the goal of the mentioned steps, my conclusions would be:

1. **Density calculation can be incremental from min to max distance:** To clearly visualize the requests based on distances around each POIs, the requests can be segregated based on distance ranges and plotted rather than plotting all together. E.g. we can calculate density of requests for each POI based on increasing radius. We start from the minimum distance as the radius, check density with number of requests and circle area. We continue until the radius is the farthest request for that POI. The calculated densities can be used to see the trend of density variation with distance
2. **Density Heatmap as visualization:** Rather than plotting all the points on a single graph, the density circles can be plotted with varying colors for varying values. That gives us a clear idea about the best performing distances.
3. **Regarding outliers**, I believe discarding outliers just based on distance is not enough. Many other aspects can be considered, as I discussed before.

### Why only distance and not include other parameters?

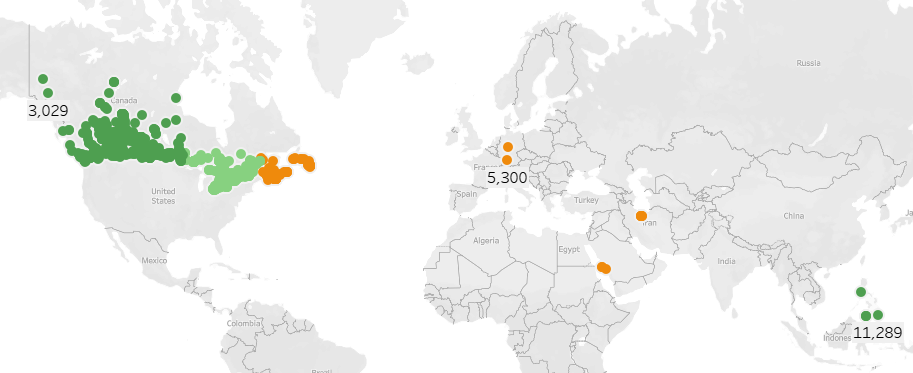
* Analysis based on distance gives us a measure of variance of requests received by POIs. But the variance can also be backed by considering the other factors like states, provinces and cities, which will give us a clear understanding why the variance or why not.
* Visualization based on distance is not very helpful unless plotted on a map.

### Why is scaling required?

I believe distance scaling is required if distance is used as a parameter in some predictive or clustering model. I am unable to understand why scaling is required just to visualize. Surely, it gives us cleaner plots but what information can we get from it?

### Alternative approach

I geo-coded the data grouped by POIs and get the below plot:

1. **Discrepancy check** - Easier to check discrepancies, if any.
2. **Outliers INSIDE Canada –** Province/city/distance wise clustering gives us which requests to consider as outliers, because apart from eliminating all the requests outside Canada, the distant requests inside Canada can also be considered as potential outliers

Inside Canada, the outliers can also be detected by the least performing provinces. The below highlighted province shows only 3 requests, all assigned to POI1. Incidentally, they are also the farthest ones. But this can be another measure of finding outliers other than distance.

