SPECIALIST LOCI

PURPOSE

To identify and discretize geography on the basis of specialist and venue locations to evaluate the success of the venues.

Data Used: Data from warehouse retrieved on 04-15-2018.
CSV file provided consisting of the following attributes:

- Worker Latitude
- Worker Longitude
- Venue Latitude
- Venue Longitude
- Market_id
- Venue_id
- Worker_id
- Shift_id
- Is_reconciled_needed_primary
- is_reconciled_needed_worked

QUICK SUMMARY

Till date, we were evaluating the success rate of a venue based on its distance from a single latitude and longitude i.e the city center. The main aim of the project was to observe whether specialists closer to a venue are more likely to work a shift as compared to specialists who live further away from a venue.

Outline

A. Methods

We wanted to see if we can capture more information about how specialists at a market are more likely to report to a shift if instead of considering the single lat long for entire market, we could consider clusters of specialists represented by centroids.

I. Cluster Analysis

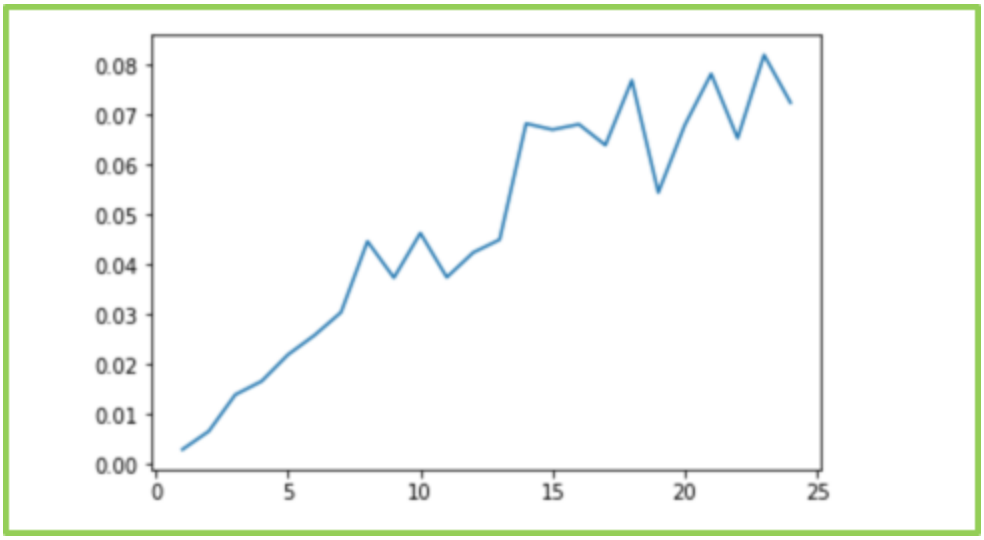
It is a technique of grouping objects in such a way that objects within a group (cluster) are similar in nature and objects in different clusters are different from each other. The technique requires the number of clusters to be specified before the algorithm is run on the group of objects. The following figure shows the geographic concentration of specialists for the market of Chicago. The number of clusters for initial observation purposes was specified as 3.

B. Results

I. Explained Variance

The results were analyzed on the basis of the information content obtained from the number of clusters. The information content assessed here was impact on the success rate of the venue by the change in the number of clusters. The approach was to come up with an optimal number of specialist clusters beyond which further addition of specialist clusters would have minimal impact on the venue success.

It was observed that whether we consider a single cluster (city center) or too many clusters (overfitting) the information gain was almost equal to zero. While assessing abstractness (single lat long) versus granularity (cluster centroids), it was observed that the optimal number of cluster centroids was 13, beyond which addition of more clusters gave very less insights from the clusters. Very point on the map was a specialist point of contact. The cluster centroids were a proxy for the geographic concentration of supply around a venue.



It was observed that beyond 13 clusters of specialists, if more clusters are added even though there is noise, there is consistently less information gain about the success of a venue.

A regression analysis was conducted between the success rate of the venue with respect to different number of clusters. A data science performance metric called the R-squared assessed the information gain with the increase in the number of clusters versus a single lat long.

C. Conclusion

The final outcome was a visualization of the 13 clusters on a geographic map of Chicago along with the plotting of the cluster centroids.

```
array([[-88.0838968 ,  40.835047  ],
       [-87.67409312,  41.97817078],
       [-88.40270838,  42.07016162],
       [-88.03446785,  41.99128367],
       [-87.80400672,  41.89389128],
       [-87.61677289,  41.58508677],
       [-83.5839675 ,  39.8666335 ],
       [-87.64680042,  41.85180678],
       [-87.19240491,  41.59399668],
       [-87.94003817,  42.29945591],
       [-87.68929759,  41.73419531],
       [-87.59706153,  41.75480324],
       [-88.14675298,  41.70056   ],
       [-87.71572841,  41.89418383]])
```

I would opine that considering the explained variance from the number of clusters is a better way rather than considering just the single city center as a representative of the entire market.
These results are still **preliminary** and may contain confounds or different approaches, The approach is not yet productionizable, but does offer a hopeful perspective.

D. Future Directions

- Develop a model to help predict the success rate of a venue in terms of the geographic concentration of specialists for venues for different markets
- Consider the concentration of all starred specialists around a venue
- Try out a different approach other than regression analysis to assess the information content from clusters