Chicago Business Intelligence and Strategic Planning: Next Steps

I have reviewed the data sources I used for my project. I have outlined what attributes I pulled from these data sets, along with the feature engineering and data cleansing actions I applied to this data. I have described how I created my data lake using PostgreSQL and the relational architecture of this data lake. I have discussed how and why various technologies were incorporated into my project and why others were considered but ultimately omitted. I have given demonstrations of how my full-stack application works and how it can be leveraged by City of Chicago stakeholders to plan accordingly. Now I want to look forward and talk about how this project can be improved.

Containerization was discussed earlier in this report, but I feel now is an appropriate time to revisit this concept. In order to deploy this application, with revision controlled libraries, containerization is needed. I believe “dockerizing” this application would be the next logical next step if the City of Chicago was interested in using this application from their strategic planning. Leveraging Docker and Kubernetes would enable easy deployment of this app while ensuring sustained, wide spread use for anyone gaining important information from the microservices it contains.

Another objective would be to build out the front end of the HTTP API. While the base HTML and Flask application enables the functionality intended, the aesthetics are minimal. I believe there should be continued development on the front end to ensure the user interface is more intuitive and graphically appealing if this application were to move into a production environment.

While the amount of data currently contained within my data lake is substantial, it is by no means exhaustive. I believe data engineers should work to load in more data if given days/weeks to do so. Additionally, the TNP data should be further harmonized with the taxi dataset to give a better representation of the taxi and rideshare usage in the varying Chicago neighborhoods and zip codes. The TNP dataset was used within this project, but because of its ample size relative to the taxi trips dataset, robust development was largely avoided. If given more time to load in data, this would be a worthwhile exercise.

Regarding data, I believe it would also be worth reaching out to the dataset owners at the City of Chicago to see if the data update frequencies can be revised. For instance, the taxi dataset is only updated monthly which could render many time sensitive applications useless. Ideally these datasets would be updated daily at a minimum in order to provide relevant information to Chicago citizens.

Lastly, the tools developed showed poor correlation between the number of taxi trips and the spread of Covid. I believe a Data Scientist should work with medical and public health professionals to better correlate the spread of Covid-19 to various factors. Understanding this will enable city stakeholders to better advise the local community of the current Covid-19 situation and risks involved with continuing to participate in community and social interactions. Doing so will help develop public policy and contribute to taming and overcoming the Covid-19 pandemic.