**Approach and Methodology**

Our Final deliverable a real time dashboard, visualizing notable locations in New York City for picking up passengers.

Tableau was used to create our dashboard. To start off tableau needs to be linked with BigQuery following which all tables created on BigQuery can be retrieved and displayed on tableau. The tables used for visualization need to linked together, this can be accomplished with the help of SQL joins. Once we have our final dataset the table should be set to update automatically this ensures that any change made to the data table will reflect on the dashboard. When new data from the pipeline is published into the table tableau makes relevant changes to the dashboard.

The dashboard shows prominent pickup locations all over the city and as number of pick ups in that area increases the size of the circle increases along with it. The dashboard also displays the current time, weather and all factors affecting weather.

Predictions were made based on given data and how the present conditions affect rides/min and revenue/min. The results are displayed on the dashboard.

**Self-Reflection**

Arjun

We started our project by diving into GCP and obtaining a certification. Dataflow and Apache beam were alien concepts until this project. The certification was insightful, thorough and more importantly relevant. I believe each of us are well equipped to deal with streaming pipelines and data.

Working on a project with data engineers in the field was even more eye opening, unlike other assignments we had to take into consideration resource consumption, deadlines on prototype, overhead and scaling. Working as part of a team and incorporating everyone's ideas and bringing the best out was remarkable.

We were able to pull through with all deliverables that were promised and more. But, it involved a lot of effort, trial and error like all projects do. Knowing what I know now, if I were to start all over I would make so many changes.

* Starting with our pipeline. Our pipeline was initially set to auto scale and as expected when the size of data multiplied the pipeline allocated more workers this drastically increased our resource consumption. Next time I would just cap the number of workers this way we can be assured that the pipeline works optimally while at the same time we don’t over consume our allotted resource quota.
* Coming to our data, partioning and organizing tables by date would have done wonders to our entire development cycle. We could have saved so much time since,this way we would not be ingesting so much data at the same time.
* Querying data on BigQuery while effective took a whole chunk of daily quota available to us. In reality we did not require all of it, Our aim was to build a real time dashboard so if I were to start all over I would find a way to release data older than 10 mins and perform our required aggregations on the latest data we obtained from the pipeline.
* This further affected Visualizations and the heatmap on Tableau took exceedingly and unnecessarily long because of the poor organization of our data.
* Finally getting to the visualization. Tableau was not our initial choice we were intending to use looker for our visualizations. Looker enables us to perform complex data manipulation. The initial plan was to take advantage of looker’s cluster analysis based on K-means to find noteworthy locations New York City to find passengers. Unfortunately, we were unable to gain access.
* Tableau comes with its fair share of advantages. I was able to link Tableau and BigQuery with ease. Then it was only a matter of performing aggregations on BigQuery and visualizing those results on Tableau.
* After realizing my mistake with portioning the tables, I filtered out all data older than 7mins. This data was then aggregated to retrieve the following details.
  + Total trips.
  + Pickup Locations.
  + Preferred Payment Type.
  + Weather.
* Tableau comes with heatmaps created by mapbox. Once the relevant latitude and longitude is fed as input we are given our heatmap of NYC as the result. Looker provides in built time series analysis, tableau does not. Hence, all time series analysis was performed on BigQuery and insights were visualized on Tableau.