

COURSE PROJECT – BIG DATA CONCEPTS

“Analysis on Chicago Taxi Trips”

NAME: RISHIKA SAMALA

EMAIL: rsamala@iu.edu

Course: INFO-I535

Course Name: MGMT ACCESS USE BIG DATA

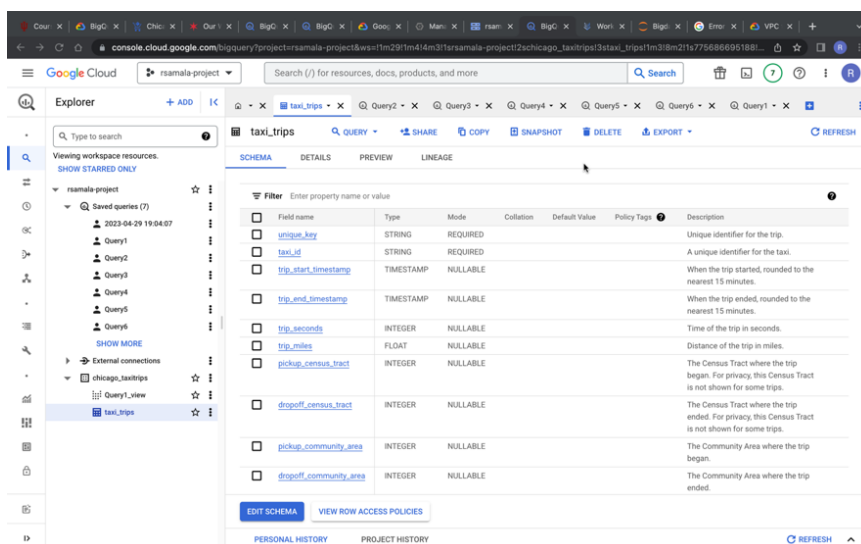
INTRODUCTION:

The “Analysis on Chicago Taxi Trips” project aims to create a data pipeline, visualizations, and live interactive dashboard to analyze the Chicago Taxi Trips data. The project aims to utilize the learnings from Google Cloud Platform assignments and try various other resources from GCP. This project achieves a project pipeline from Chicago taxi trips dataset. The pipeline addresses data ingestion, data transfer, data transformation, analyzing of data and summarization and finally creating a real time dashboard using Tableau tool.

The dataset is chosen from Google Cloud Marketplace which includes taxi trips from 2013 to the present, reported to the City of Chicago in its role as a regulatory agency. The data is updated daily and has more than 200 million rows of trips data with 23 attributes represented as columns. The current data is about 74.89GB.

Link of data: <https://console.cloud.google.com/marketplace/product/city-of-chicago-public-data/chicago-taxi-trips?project=rsamala-project>

Data Schema:



The screenshot displays the Google Cloud BigQuery console interface. On the left, the 'Explorer' pane shows the project structure with 'rsamala-project' selected, containing 'Saved queries (7)' and 'External connections'. The 'taxi_trips' table is highlighted under 'External connections'. The main pane shows the 'taxi_trips' table schema with columns: unique_key, taxi_id, trip_start_timestamp, trip_end_timestamp, trip_seconds, trip_miles, pickup_census_tract, dropoff_census_tract, pickup_community_area, and dropoff_community_area. The schema details are as follows:

Field name	Type	Mode	Collation	Default Value	Policy Tag	Description
unique_key	STRING	REQUIRED				Unique identifier for the trip.
taxi_id	STRING	REQUIRED				A unique identifier for the taxi.
trip_start_timestamp	TIMESTAMP	NULLABLE				When the trip started, rounded to the nearest 15 minutes.
trip_end_timestamp	TIMESTAMP	NULLABLE				When the trip ended, rounded to the nearest 15 minutes.
trip_seconds	INTEGER	NULLABLE				Time of the trip in seconds.
trip_miles	FLOAT	NULLABLE				Distance of the trip in miles.
pickup_census_tract	INTEGER	NULLABLE				The Census Tract where the trip began. For privacy, this Census Tract is not shown for some trips.
dropoff_census_tract	INTEGER	NULLABLE				The Census Tract where the trip ended. For privacy, this Census Tract is not shown for some trips.
pickup_community_area	INTEGER	NULLABLE				The Community Area where the trip began.
dropoff_community_area	INTEGER	NULLABLE				The Community Area where the trip ended.

BACKGROUND:

Chicago population is around 2.7 million as per 2021 and it is the 3rd most populous city in United States. The transport options available for people are, taxis (including Uber, Lyft, and public taxis), public transport and own vehicles. People in Chicago prefer taking taxis in downtown and crowded areas where parking their vehicles is not possible. Most of the city during peak hours preferred taxis to reduce the burden of driving in traffic. There was an article in Chicago Tribune that gives us information about how number of Chicago taxi drivers hits 10 years low as ride share companies take off. The percentage of people using taxis was decreasing gradually. So, Chicago released the data set containing data of taxi trips from 2013 to present to public.

When I visited Chicago, people suggested me not to take own vehicles to travel to downtown and certain areas as the parking and traffic is impossible in such conditions. People preferred taxis and I was thinking to understand how taxi trips, time taken during a trip and such information could tell us about traffic conditions, busiest routes, taxis availability and planning for alternate options, know about economic changes in the city based on cash, card and other means of payment and to even get insights about where to live when you visit, based on taxis availability to locations you want to visit if you don't have own vehicle. But reports suggest that taxis and drivers of taxis are not preferring taxis now a days recently. Even covid effected this situation more. I wanted to analyze this data to get insights about the situation with taxis, which helps in understanding more about the travel and other conditions of Chicago! The dashboard created will be helpful for public for gathering such insights and for their own analysis too.

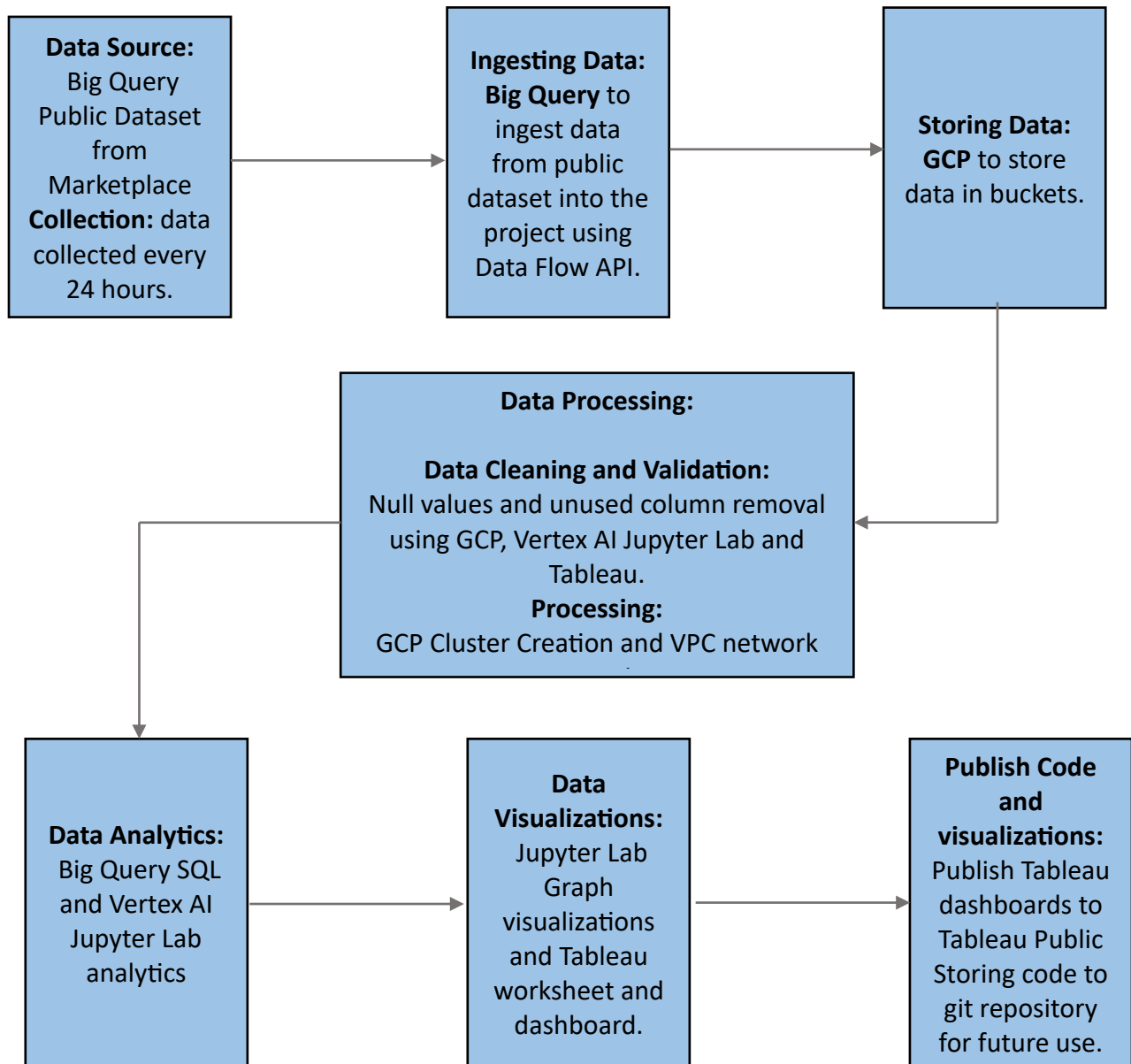
METHODOLOGY:

In this project I addressed all the 5 V's of Big Data:

1. **Volume:** The data has a large volume with 200 million records and 23 attributes, which makes it a Big Data.
2. **Velocity:** The data is continuously generated with daily number of trips and updated and refreshed every 24 hours.
3. **Variety:** The data is obtained from various sources. Regarding information about taxis, it is from list of public passenger vehicles which includes licensed taxicabs too. Regarding information about start and end time of trips it is from an application. Regarding information about geographical locations of the pickup and drop locations it is tracked by other applications. All these variety of data is collected and maintained by Chicago Digital which a collective individual groups from various city departments who created the dataset.
4. **Veracity:** The data is consistent with proper attributes and is accurate and defined by a proper schema and trustable as the source of the dataset is from City of Chicago.
5. **Value:** The data is very valuable in terms of different insights it provides on the condition of taxi riding and many other helpful insights on traffic.

For this project I have used Google Cloud Platform and Tableau to implement the whole pipeline. The following figure shows the pipeline methodology of Big Data I used in my project.

PIPELINE METHODOLOGY:

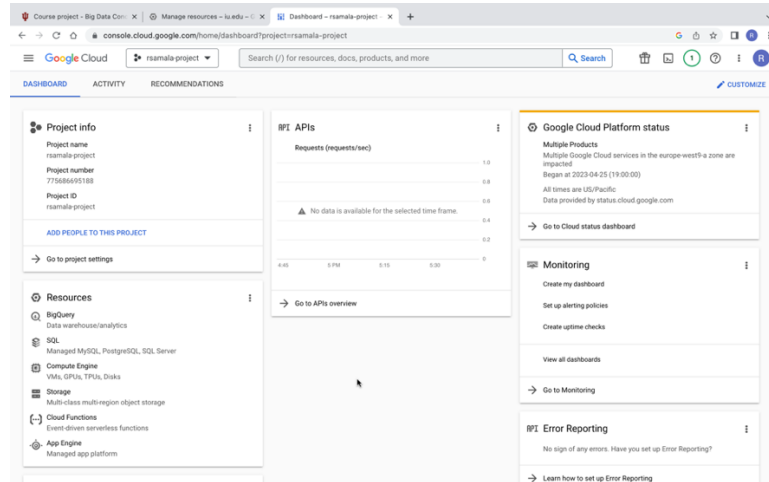


TECHNOLOGICAL SETUP:

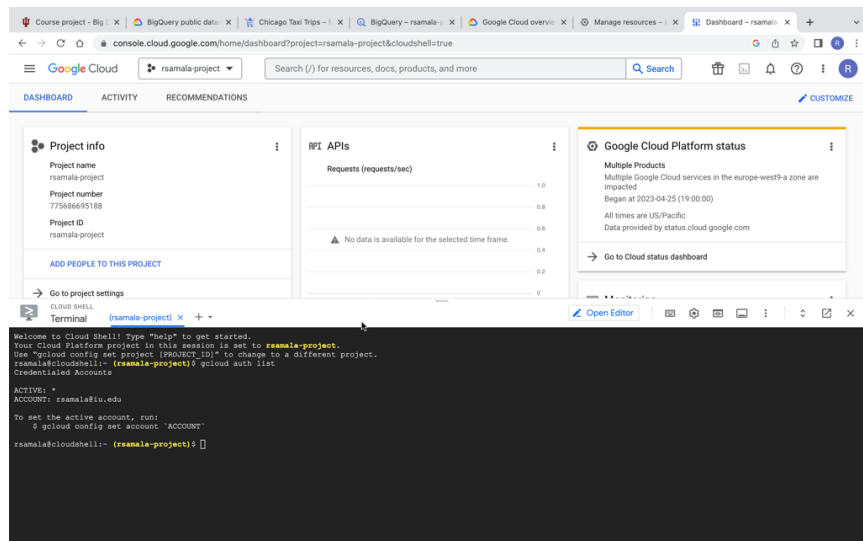
1. **Data Source and Collection:** The data is generated by Chicago City and is continuously collected every 24 hours into Google Cloud Marketplace. The data is got into this pipeline through Big Query.

2. **Ingesting Data:** The following steps are involved in this process:

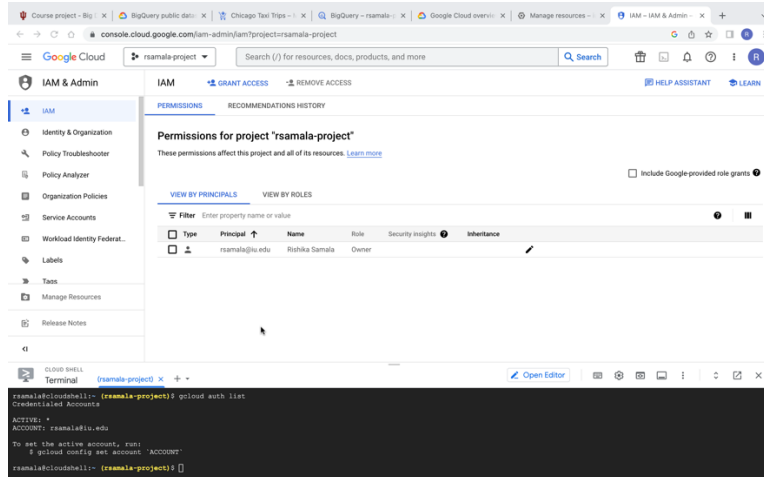
- a. **Creating a new project in GCP:** I have created a new project named “rsamala-project” in Google Cloud Platform as shown below.



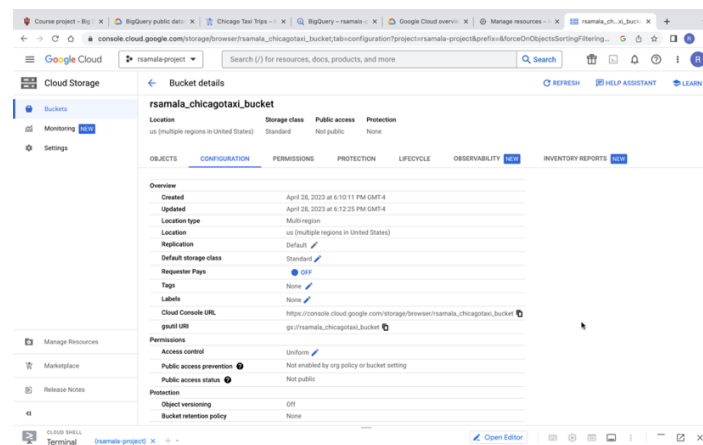
- b. **gcloud Authorization:** I have activated my project using gcloud command as shown below.



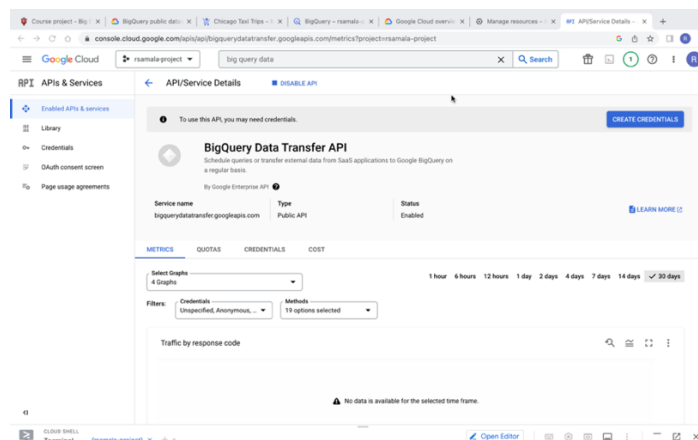
- c. **IAM:** I have configured necessary IAM instance management actions. Since I have created the project, I have not granted any additional permissions and made sure all actions are configured appropriately.



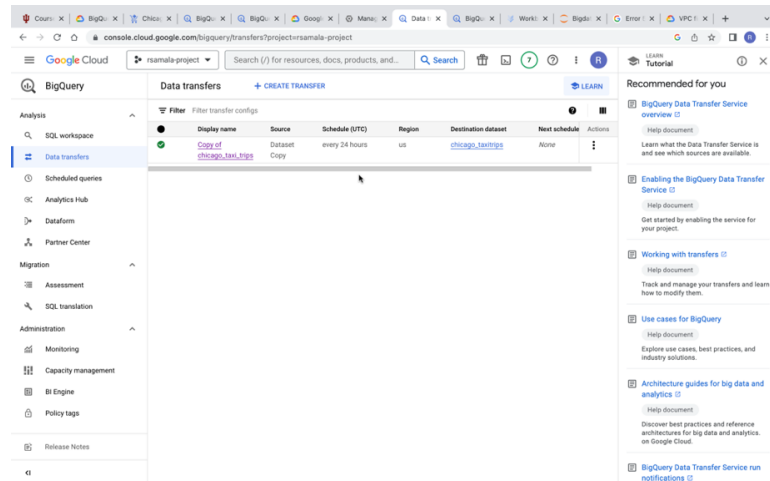
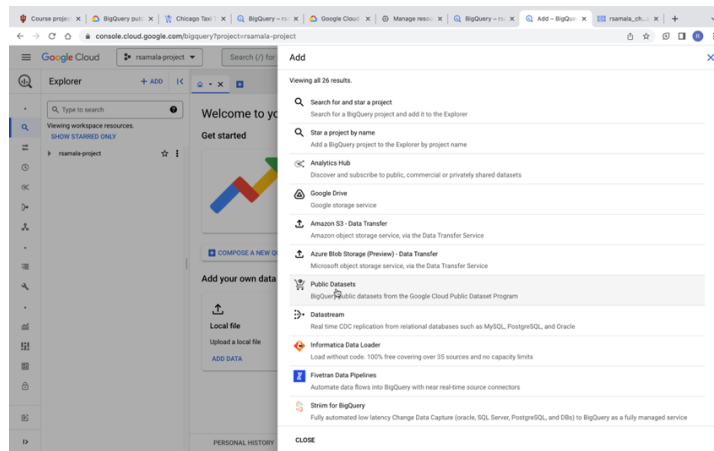
- d. **Bucket Creation and configuration:** I have created a bucket named “rsamala_chicagotaxi_bucket” as shown below to store the data and its related information.



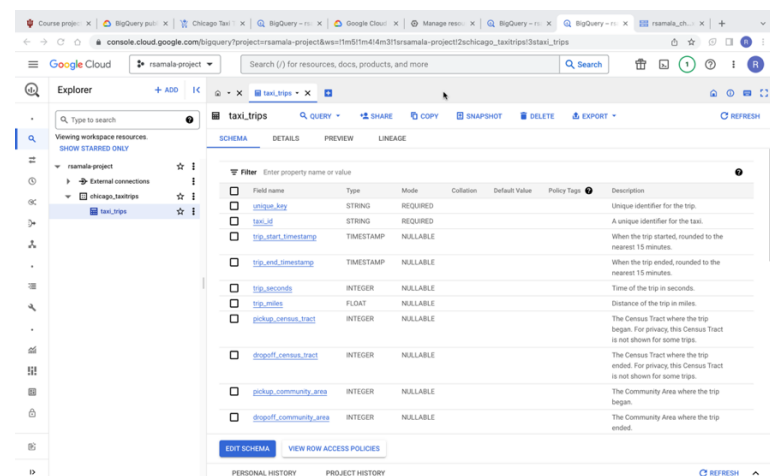
- e. **Big Query Data Transfer API:** I have enabled the BigQuery Data Transfer API to be able to ingest the dataset.



- f. **Ingesting Data:** I have added data to my project from Big Query Public Datasets as shown below and created a data transfer.

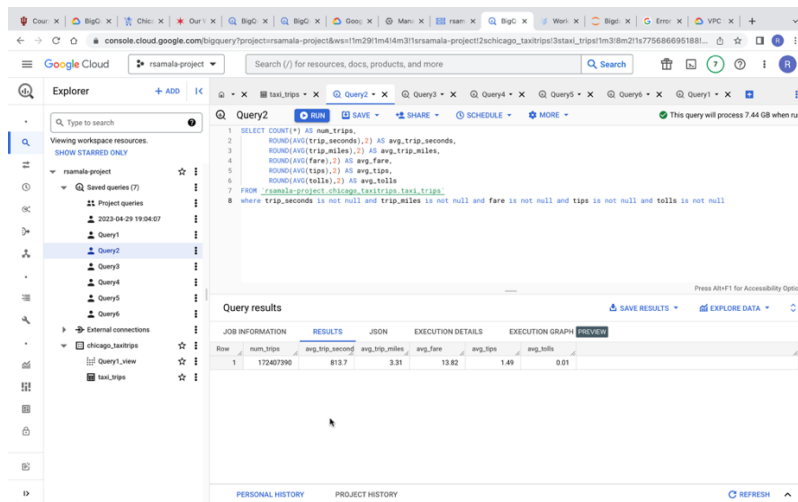


- g. **Dataset:** I have added and created a table of dataset named “chicago_taxitrips” under my project. The dataset in my project is as shown below.

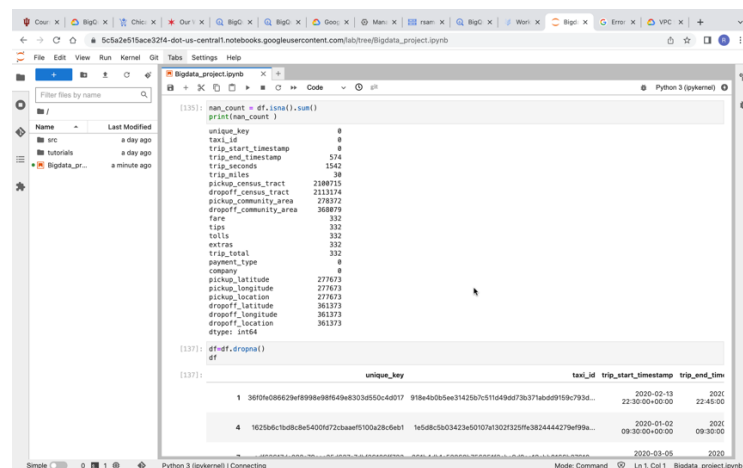


3. **Storing Data into GCP:** The table data is exported, and this is stored into the GCP bucket created by me in previous step into the folder with a file name.
4. **Data Processing:** Performed Data Processing as shown in below steps while performing analytics and using respective tools.

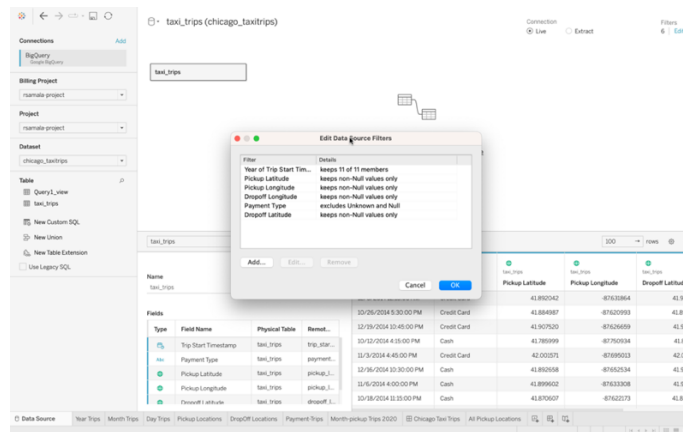
- a. Performed Null values checking while querying the data in BigQuery.



- b. Performed removal of null values in the Vertex AI notebook which is shown below.

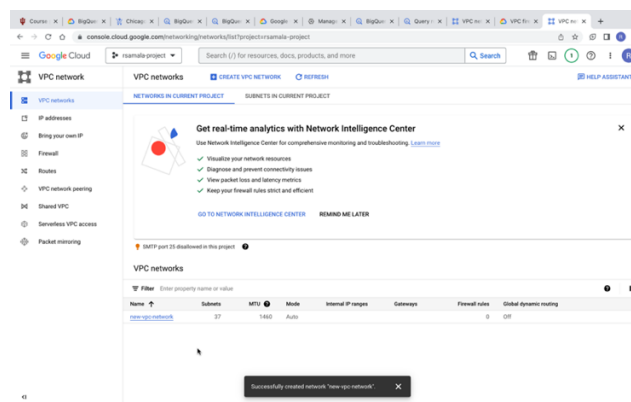


- c. Performed removal of null values in Tableau.

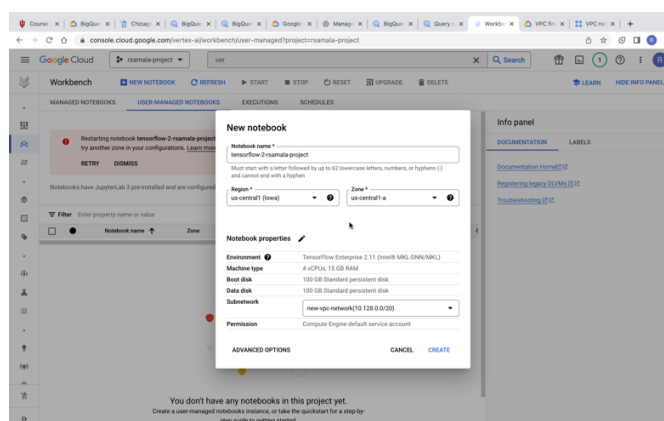


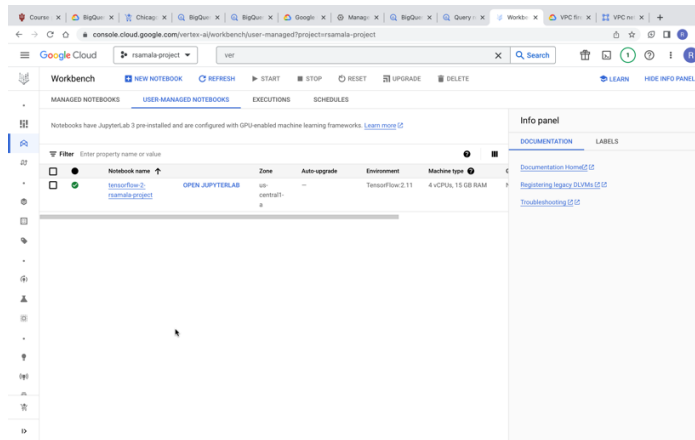
d. Data Processing Steps required for Data Analytics:

- **Creating a new VPC network:** I have created a new vpc network named “new-vpc-network” as shown below.



- **Create Vertex AI notebook and link the VPC network:** The new vpc network with subnetwork 10.128.0.0/20 is used to create a connection for hosting Jupyter Lab. A new TensorFlow notebook has been created with no GPU, named “tensorflow-2-rsamala-project”. The US region is chosen whichever has low traffic.

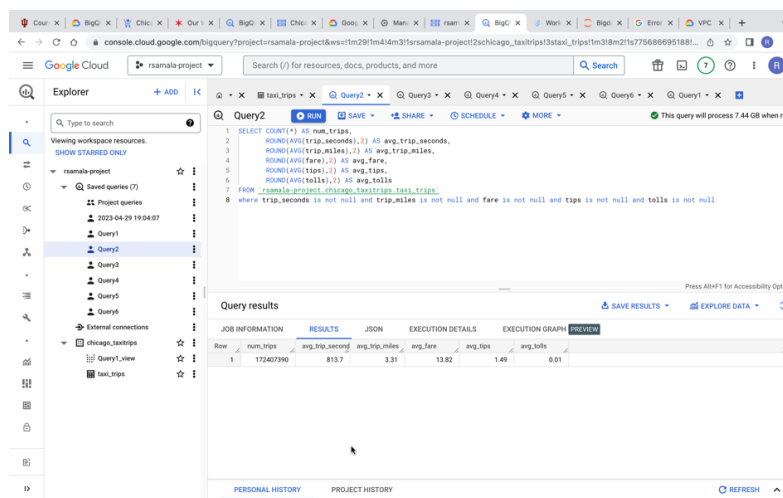




5. **Data Analytics:** I have performed data analytics through Big Query SQL queries and Jupyter notebook created through Vertex AI as shown above.

a. **Big Query SQL Queries:** Few sample sql queries from the project are shown below. Remaining can be found in the project in GCP.

Query2: To analyze the average number of trips, average time of trips, average miles, average fare, average tips, and average tolls in the data.



Query5: To analyze the total data during year 2020, in which covid was declared as outbreak.

The screenshot shows the Google Cloud BigQuery Explorer interface. On the left, the 'Explorer' pane displays the project structure, including 'Saved queries (7)' and 'External connections'. The main pane shows 'Query5' with the following SQL query:

```
SELECT *
FROM `rsamala-project:chicago_taxi_trips.taxi_trips`
WHERE trip_start_timestamp > '2020-01-01' AND trip_start_timestamp < '2020-12-31'
```

The 'Query results' pane displays the results in a table format. The table has columns: unique_key, taxi_id, trip_start_timestamp, trip_end_timestamp, trip_seconds, and trip_miles. The results show two rows of data for taxi trips in 2020.

Row	unique_key	taxi_id	trip_start_timestamp	trip_end_timestamp	trip_seconds	trip_miles
1	ef641418f89a7724286c035a...	2f2aa64117eb041f0ac81d550a4e9b02c7f02f6762ab154d4f4a6d7f0ac0b6f153a70ad04e5a4036a1a0e4e83172a488c0fca	2020-07-24 12:00:00 UTC	2020-07-24 12:30:00 UTC	1380	5.5
2	fa658d8a0743f2f4f58f1611...	8f968db0e4760990890f16b12c9fa849326f96b3862f1a6e0238a79f3217781761868b0a0f90a0c195	2020-07-11 18:15:00 UTC	2020-07-11 18:45:00 UTC	1860	20.0

- b. **Vertex AI Jupyter Notebook Analysis:** I performed Exploratory Data Analysis on the data of year 2020 to understand the impact of covid particularly. Few screenshots of EDA are attached below. The whole notebook contains the whole EDA and is uploaded in GIT (link provided in results section).

Import the 2020 data from GCP bigquery and converting to data frame.

The screenshot shows a Jupyter Notebook with the following code cells:

```
[126]: #Importing required libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

[127]: #Importing big query from google cloud
from google.cloud import bigquery

[128]: #Getting all the trips data from year 2020
data = bigquery.Client().query("""
SELECT *
FROM `rsamala-project:chicago_taxi_trips.taxi_trips`
WHERE trip_start_timestamp > '2020-01-01' AND trip_start_timestamp < '2020-12-31' """)

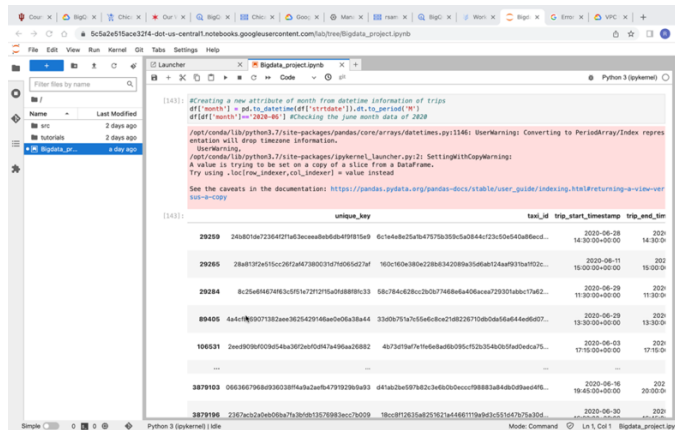
[129]: #Converting the data into pandas dataframe
df = data.to_dataframe()

[130]: #Checking the first 5 trips details
df.head()
```

The output of the last cell shows the first 5 rows of the data frame:

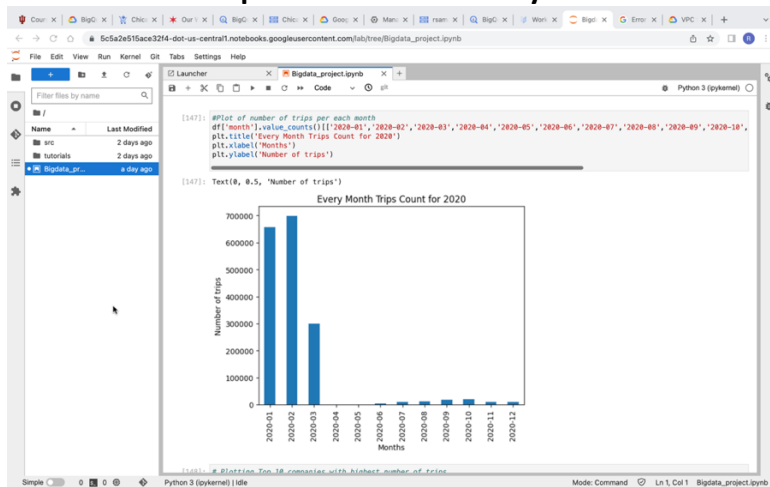
	unique_key	taxi_id	trip_start_timestamp	trip_end_timestamp
0	34e3574aba31871a072854fa0950f8d8fbd9f5	6c6cc4c918a70492231337ac2a7726a405aa0ad33ee5...	2020-02-14 21:00:00+00:00	2020-02-14 21:00:00+00:00
1	36f0f068629ef899e88f649b303d550c4d517	918b4b0b5ea3142b0751f4a8a073b377abdd9f59c793d...	2020-02-13 22:30:00+00:00	2020-02-13 22:45:00+00:00
2	3546239a0bac075bc7763276a8244d72475bc87	f38ba84dbec77476c301449f329c430486a09db4929b...	2020-02-25 14:45:00+00:00	2020-02-25 15:00:00+00:00
3	23ac07faa7aa23ccab719bcd9a8d8a2863b13f7f	3cc07933f92b6167ba8ad2e7ff9652c2a427e45ead57...	2020-01-10 08:30:00+00:00	2020-01-10 08:30:00+00:00
4	1625b6c1bdc8e4005f72c3aaaf5100a28c6e1	1e5db8c603423a50107a1302f325f93824444279e99a...	2020-01-02 09:30:00+00:00	2020-01-02 09:30:00+00:00

Creating new attribute of month from the start and end dates to continue further analysis and visualizations.

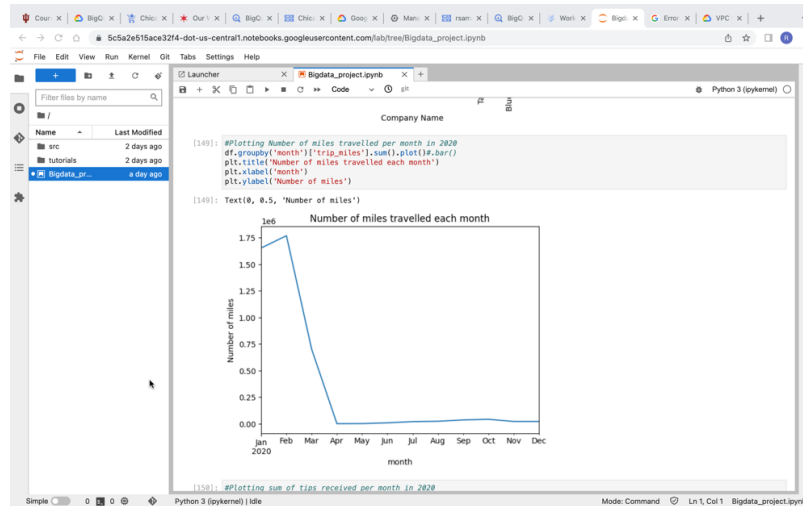


RESULTS:

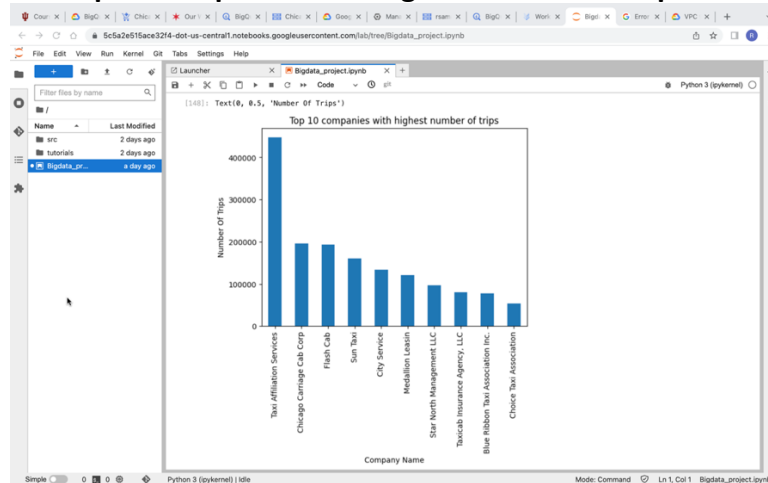
Link to the notebook: <https://github.iu.edu/rsamala/INFO-I535-Final-Project>



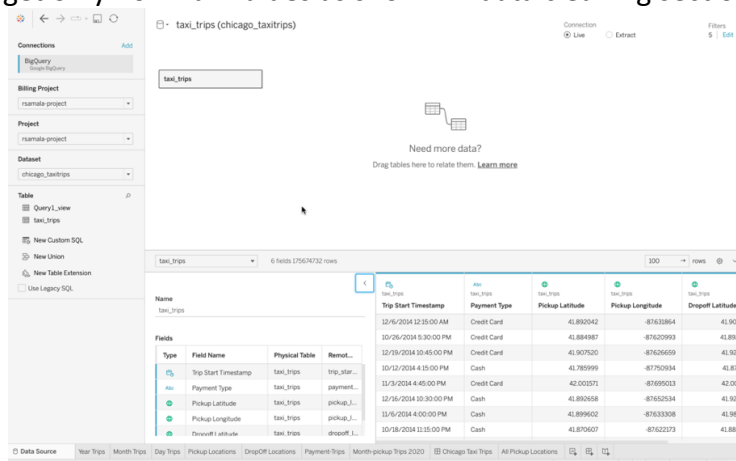
Number of miles travelled in 2020 each month.



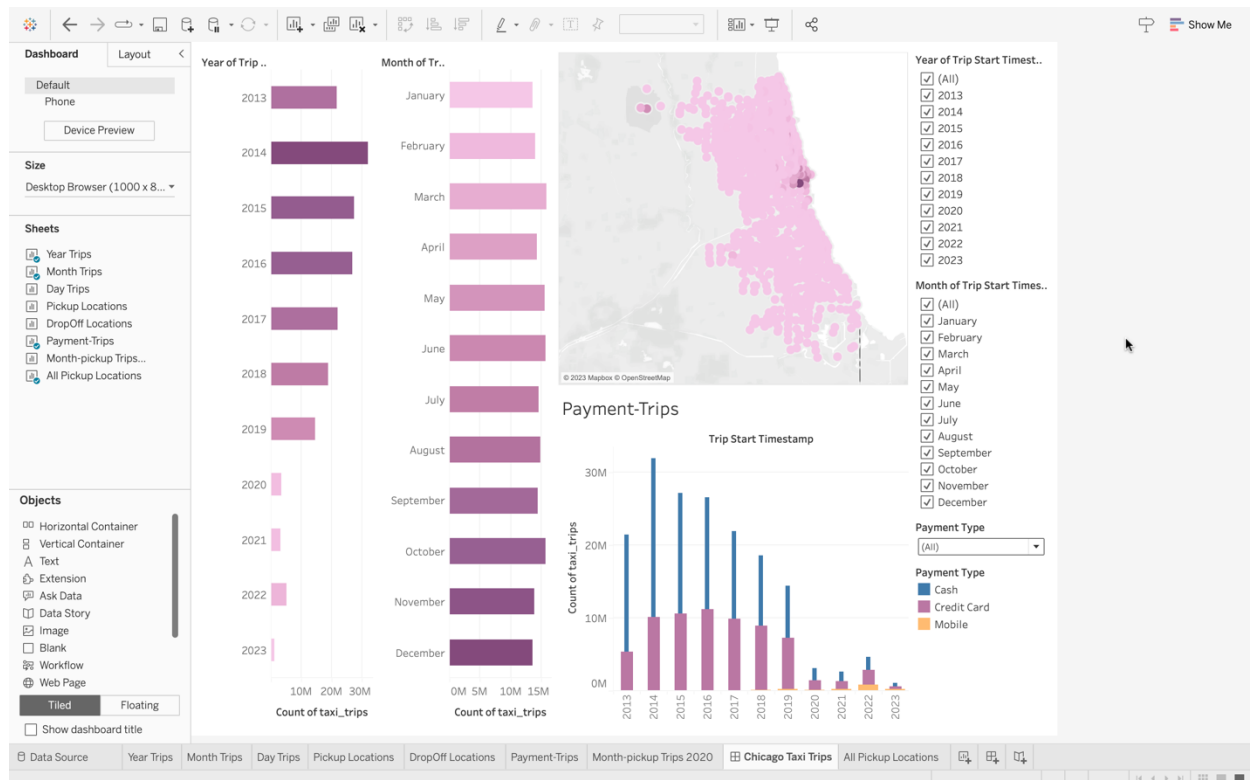
The top 10 companies with highest number of trips



- Tableau Visualizations:** I have connected my **Tableau Desktop** to Google BigQuery and got my data from the project I am working on in GCP as shown below. I applied filters to get only non-null values as shown in data cleaning section in processing section above.



I had learned Tableau and have built a dashboard from the whole 176M rows data (after removing null values from all attributes used in building the dashboard) which gives insights on the number of trips per year per month and respective pickup locations and information of payments made through cash or card or mobile. The below is the screenshot of the dashboard created from whole data on **Tableau Desktop**.



In this dashboard I have provided flexibility for users for their analysis. Users can drill down specific year, and drill to months in them and the locations of taxi trips in those months and years and whether the payment is made through cash or card. For this I added filters and user can also drill down based on payments type or even on locations separately to analyze how many taxi trips occurred in that location.

Transfer of Dashboard to Tableau public: The dashboard I created is on **Tableau Desktop** to connect to GCP (BigQuery can be integrated to Tableau only from Tableau Desktop), while for public viewing, I had to extract the tableau hyper files and transfer to **Tableau Server** and then publish dashboard to **Tableau Public**. Tableau Public has limit on number of rows in the data to be 15M. So, I had filtered the data to be only from years 2018-2022 and used only 15M rows in publishing the dashboard. So, the dashboard you see on the **Tableau Public** link below consists of that filtered data (the bottom filtered rows contained data from 2021 mostly, so you observe the dashboard having very low trips data from 2021).

Link to live hosted dashboard on GCP:

https://public.tableau.com/app/profile/rishika.samala/viz/BigData_Project_2023/ChicagoTaxiTrips?publish=yes

Project Name in Tableau: BigData_Project_2023

DISCUSSION:

INTERPRETATION OF RESULTS:

The first two visualizations from Jupyter Notebook EDA which are regarding number of trips and number of miles travelled. They are drastically reduced in the months of April, May and started to increase slowly from June till October and again decreased in November and December. These results show in April and May the impact of covid being declared as pandemic in May and its peak 1st wave. Back in November and December there was a 2nd wave. So, people reduced travelling and even work from home options were provided. Hence, the number of trips is low during those months.

Another visualization from Jupyter Notebook EDA shows the top 10 taxi companies with highest number of trips. This gives us the insight of why the Taxi Affiliation Services is ranked top in cab services than others in Chicago. It also implies the good service provided by them and we can prefer them over other cab companies as a visitor and can rate them based on the service.

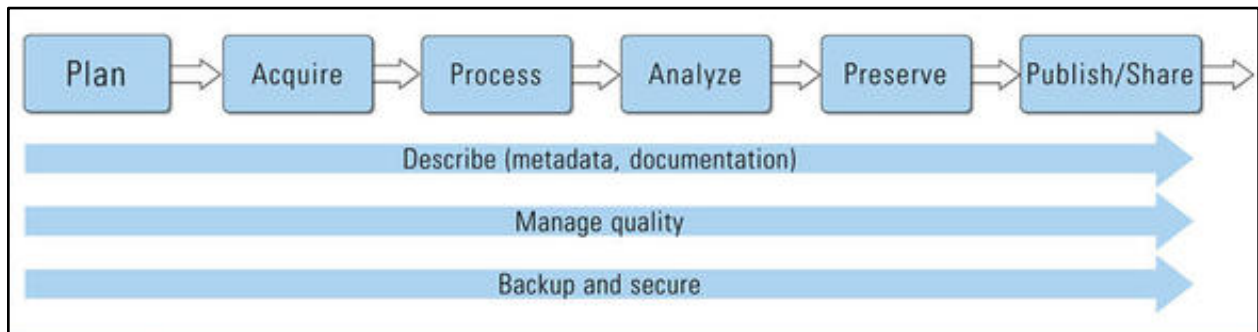
Tableau Dashboard gives us insights about how the number of taxi trips are high in 2014 and are reduced further on continuously. This was due to the impact of Uber and Lyft which have hit the market from late 2014. Because of their new methods and easy sharing of cabs, they have attracted both users and drivers. Users feel it low priced as Uber and Lyft do not require high maintenance cost burden on drivers which is why the cost are high for normal taxis. Also, the taxi drivers if working under Uber and Lyft are not required to obtain the fingerprints and their renewals from Chicago state by themselves which is also a hectic process, and taxes of taxis are high on drivers. These all changed the situation in the taxis market. Many such insights can be drawn if we could dig deep into the data.

USE OF TECHNOLOGY / SKILLS FROM THIS COURSE:

The data I have used is structured data and batch data. In this project I have used Google Cloud and BigQuery Tool to perform Data Ingestion, Data Cleaning, Data Processing, Data manipulation, Data Storage. All these were part of the Cloud Computing module described in the lectures. I have decided to use Cloud as the data I am dealing has 200M rows which is very huge. Choosing cloud has helped me in easy and fast analysis of the data which rather would be impossible in local desktops.

BigQuery data ingestion and storage were specifically part of the Ingest and Storage module of the class where we learned them in a Qwiklabs session. Ingestion and Storage methods helped in my lifecycle pipeline creation of ingesting data into cloud and storing data for further use. In this project I have particularly used the technology and class on Lifecycles and pipelines. I was really interested in topic of building a whole pipeline of a data lifecycle about how and where the data goes through different phases during its existence. This inspired me to take up this project of implementing a pipeline from BigQuery Data to analyze the big data and visualize the insights from the analysis.

I used USGS lifecycle by US Geological Survey which was introduced to me in the lifecycles and pipelines module. The following figure describes the USGS Lifecycle.



The USGS lifecycle consists of 6 main phases.

Plan: In this phase, I planned on what data and tools to use. I decided working on Chicago Taxi Trips dataset and use BigQuery and Tableau tools for analysis and visualization.

Acquire: I acquired the dataset from Google Cloud Market Place and transferred it into my project to BigQuery.

Process: In this stage, I pre-processed data by cleaning and removing null values and have made the data and environment ready for the further analysis in GCP.

Analyze: I have used BigQuery, AI Vertex Jupyter Instance and Tableau to analyze my data and created visualizations and dashboard as my results.

Preserve: I preserved the data in GCP bucket and the code, SQL queries and visualizations in GCP project and repository for future use. I preserved Tableau Sheet visualizations and the dashboard in Tableau Server.

Publish/Share: I published my dashboard in TableauPublic and the link is provided in Results section above. The visualizations and code are shared in a repository link provided in Results section.

Describe: In every phase, I ensured the meta data is properly described and maintained in BigQuery in GCP.

Manage Quality: I maintained the quality of data at and after every phase after performing necessary actions.

Backup and Secure: I saved all the data, visualizations safely and securing in respective servers, repositories and in GCP for future use.

BARRIERS OR FAILURES ENCOUNTERED:

In this section I would like to address the challenges I encountered during this project.

1. The data transfer from public data set into the Big Query of my project. This is where most of my time is consumed. Though there are many resources available in GCP but they were confusing and it took me a lot of time to figure a proper option out.
2. Trying to load whole data in jupyter notebook even in Vertex AI notebook was crashing. So, I decided to analyze only data from 2020 in the notebook.
3. One more most time-consuming part was in Tableau. I created the dashboard in Tableau Desktop. But I could not find the option to publish it directly in TableauPublic for viewing by all others. Then after multiple attempts of extract creation I found optimized extract creation method which finally was able to create extract after applying filters and with only 15M rows data, I was able to put my data dashboard into server initially which then provided me option to publish in TableauPublic. It would be nice if TableauPublic would allow for more data visualization.

CONCLUSION:

I have implemented a Big Data pipeline to analyze and visualize the data from BigQuery. This pipeline has several steps like data storage, data ingestion, metadata documentation etc., which are very helpful for analyzing and visualization of big data projects as in whole for quality insights and outputs. Further publishing the insights helps others to understand the analysis and decide what is the best choice for them like which Taxi service is better for a particular location. I have analyzed the trend of people using taxis got reduced from 2014 and this became even worse during and after covid times. So, by these insights State could help individual taxi owners and drivers by addressing low people travelling in taxis by relaxing few regulations which are very fair for companies like Uber and Lyft so they can reduce costs. Further many other insights can be drawn from such a huge amount of data to analyze the traffic conditions, modify tolls rates and even choose a locality to stay which has high frequency of taxi services if you are daily commuting person. Such insights from a data can change many situations in the real world. This course has changed my point of view of Data Science as in a whole in the case when we deal with Big Data.

REFERENCES:

1. <https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew>
2. <https://www.chicagotribune.com/news/ct-chicago-taxi-driver-decline-met-20161214-story.html>
3. <https://medium.com/@yennhi95zz/exploring-the-chicago-taxi-trips-dataset-visualizations-on-tableau-74b4b7fffb44>
4. <https://cloud.google.com/vpc/docs/vpc#default-network>
5. <https://cloud.google.com/vertex-ai/docs/workbench/notebooks>

6. https://help.tableau.com/current/pro/desktop/en-us/extracting_data.htm
7. <https://data.ucsf.edu/ssa/step-72-best-practices-building-tableau-extracts#:~:text=We%20strongly%20recommend%20that%20the,to%20the%20server%20and%20performance.>
8. <https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-share.htm>
9. https://help.tableau.com/current/server/en-us/perf_optimize_extracts.htm#:~:text=Speed%20up%20specific%20extracts,-Use%20the%20Background&text=You%20can%20help%20improve%20server,Aggregate%20data%20for%20visible%20dimensions.
10. https://help.tableau.com/current/pro/desktop/en-us/publish_workbooks_tableaupublic.htm#:~:text=With%20your%20workbook%20open%20in,to%20create%20a%20new%20one.
11. <https://kb.tableau.com/articles/howto/sharing-workbooks-without-tableau-desktop>
12. <https://www.cloudskillsboost.google/quests/23>
13. <https://www.cloudskillsboost.google/focuses/3692?parent=catalog>
14. <https://www.cloudskillsboost.google/focuses/1846?parent=catalog>