# **CSYE 7245 - Big Data Sys and Int Analytics**

## **Lab - GCP-Datalab/Dataflow**

### **Team Information:**

|  |  |
| --- | --- |
| **Name** | **NUID** |
| Tanvi Gurav | 001306848 |
| Keerti Ojha | 001050173 |
| Priyanka Malpekar | 001302741 |

**Lab Completion date** : 29th January’ 21

# **About**

This lab demonstrates GCP services like Datalab, Dataflow and BigQuery for implementing data analysis and preprocessing for machine learning.



Google Cloud Platform (GCP), offered by [Google](https://en.wikipedia.org/wiki/Google), is a suite of [cloud computing](https://en.wikipedia.org/wiki/Cloud_computing) services that runs on the same infrastructure that Google uses internally for its end-user products, such as [Google Search](https://en.wikipedia.org/wiki/Google_Search), [Gmail](https://en.wikipedia.org/wiki/Gmail), [file storage](https://en.wikipedia.org/wiki/Google_Drive), and [YouTube](https://en.wikipedia.org/wiki/YouTube). Alongside a set of management tools, it provides a series of modular cloud services including computing, [data storage](https://en.wikipedia.org/wiki/Computer_data_storage), [data analytics](https://en.wikipedia.org/wiki/Data_analysis) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning).

Google Cloud Platform provides [infrastructure as a service](https://en.wikipedia.org/wiki/Infrastructure_as_a_service), [platform as a service](https://en.wikipedia.org/wiki/Platform_as_a_service), and [serverless computing](https://en.wikipedia.org/wiki/Serverless_computing) environments.

Cloud **Datalab** is a powerful interactive tool created to explore, analyze, transform and visualize data and build machine learning models on Google Cloud Platform. It runs on Google Compute Engine and connects to multiple cloud services easily so you can focus on your data science tasks.

Google Cloud **Dataflow** is a cloud-based data processing service for both batch and real-time data streaming applications. It enables developers to set up processing pipelines for integrating, preparing and analyzing large data sets, such as those found in Web analytics or big data analytics applications.

Storing and querying massive datasets can be time consuming and expensive without the right hardware and infrastructure. **BigQuery** is an [enterprise data warehouse](https://cloud.google.com/solutions/bigquery-data-warehouse) that solves this problem by enabling super-fast SQL queries using the processing power of Google's infrastructure.

# **Dataset**

We used public Natality dataset to create an ML model to predict a baby's weight given a number of factors about the pregnancy and the baby's mother.

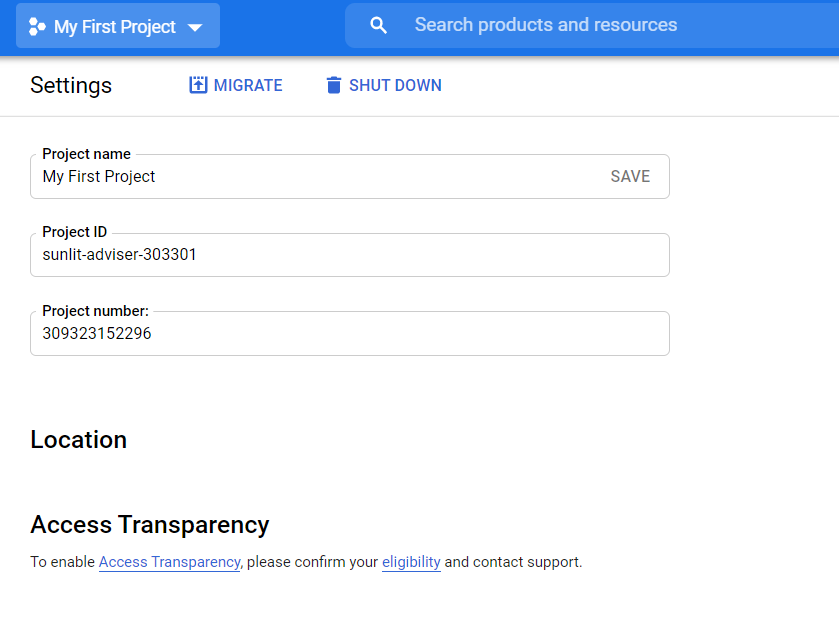
We cloned the [**https://github.com/GoogleCloudPlatform/training-data-analyst**](https://github.com/GoogleCloudPlatform/training-data-analyst) github path and used **training-data-analyst/blogs/babyweight/babyweight.ipynb** notebook for our data processing and model creation.

# 

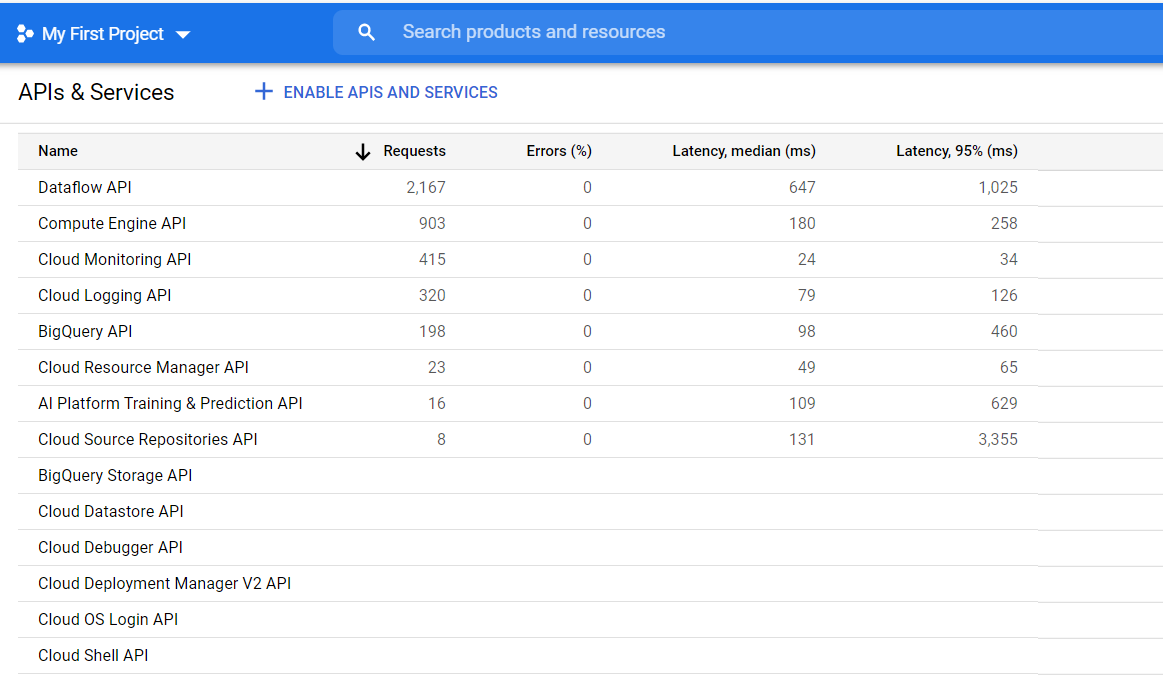
# **Experiment Setup**

## **Prerequisites**

1. In the Google Cloud Console, on the project selector page, select or create a Google Cloud project.



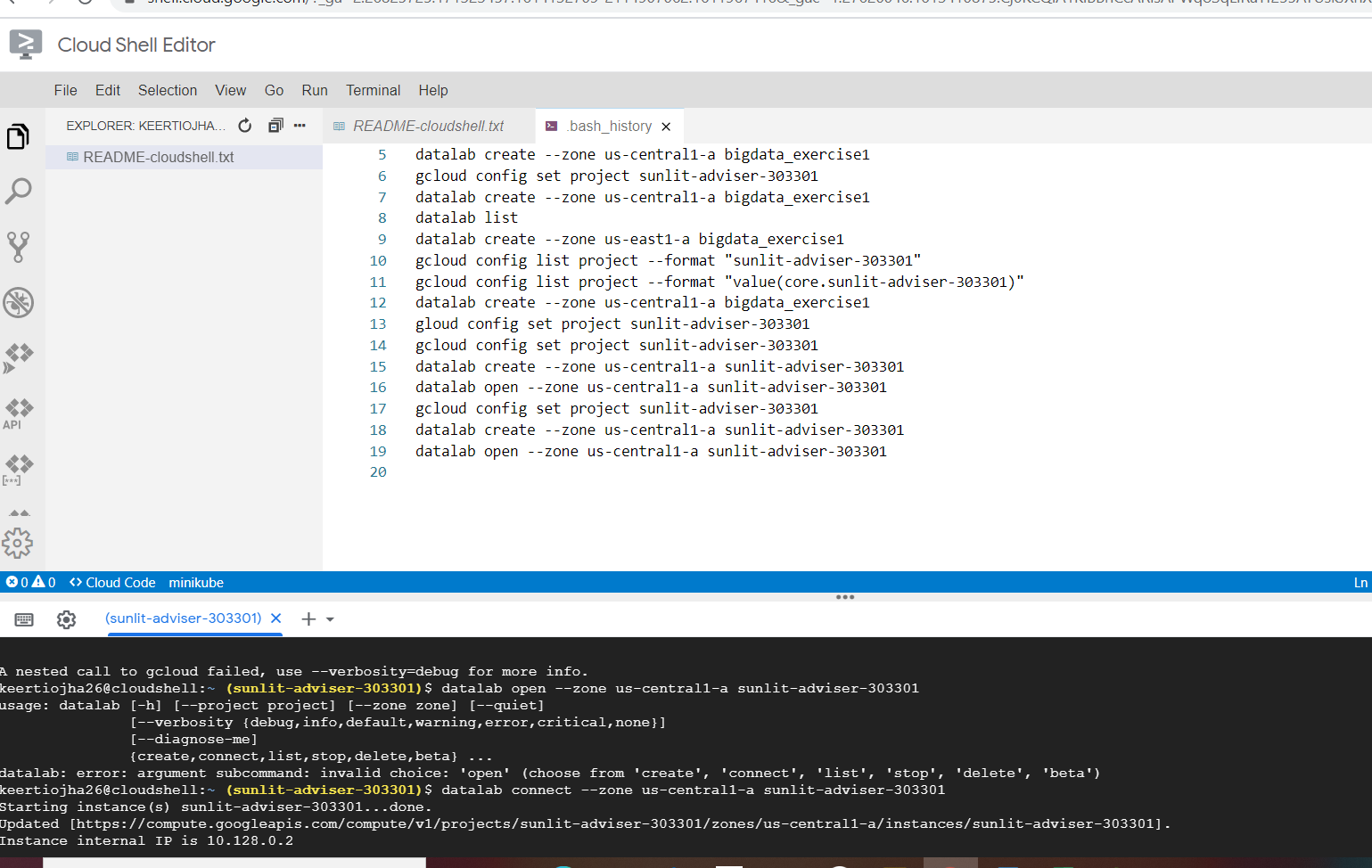
1. Enabled the BigQuery, AI Platform, Cloud Source Repositories, Dataflow, and Datalab APIs.



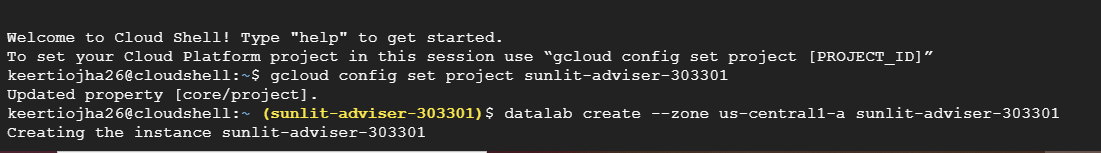
## **Launching Datalab**

Following are the steps to launch Datalab:

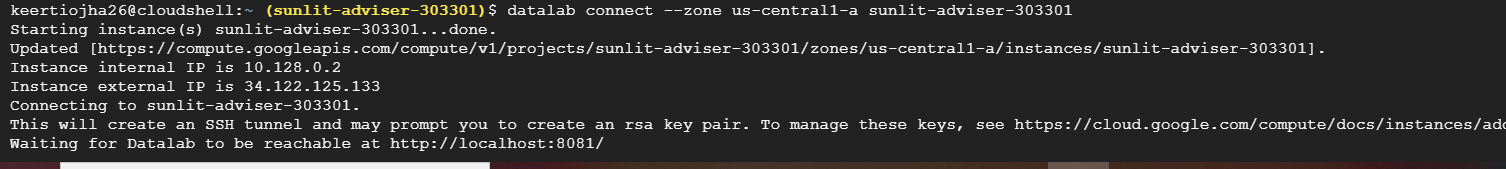
* Open Cloud Shell Editor



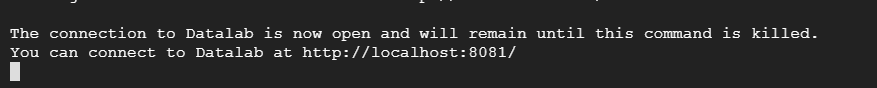
* Retrieved Google Cloud Project id using the below command



* Created a Datalab instance



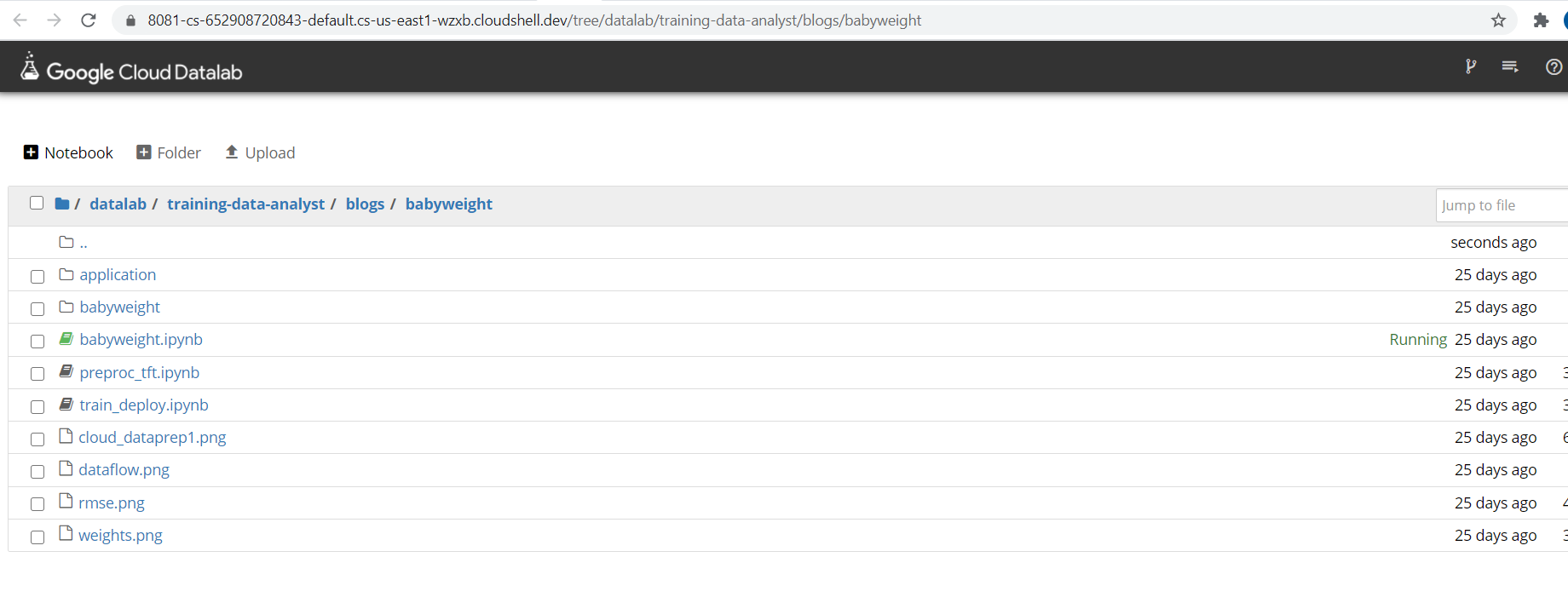
* Connection was established and instance of Datalab created at port 8081



## **Cloning Datalab Notebook**

* Opened a new notebook and copy pasted below command

!git clone <https://github.com/GoogleCloudPlatform/training-data-analyst>

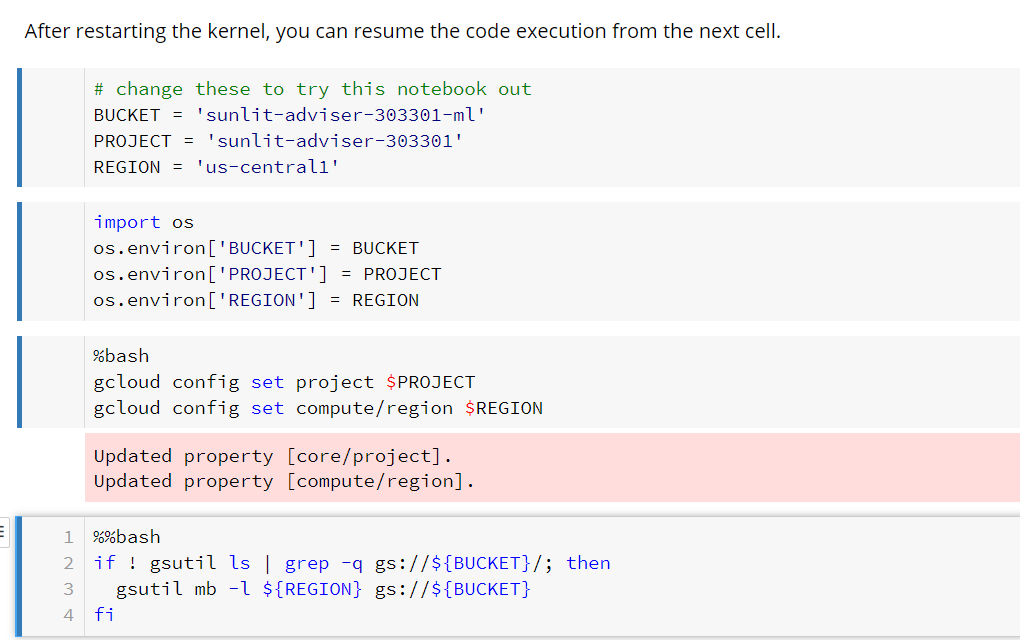


# **Use Cases**

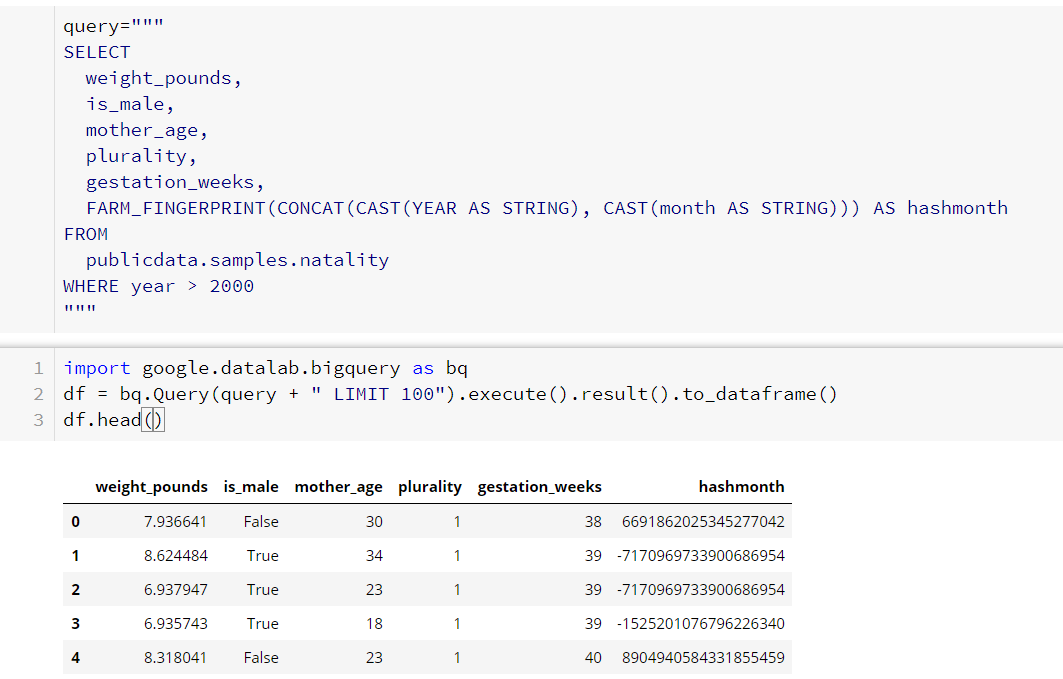
## **Data Exploration, Preprocessing and Visualization in Datalab**

Project ID and Bucket setup in notebook

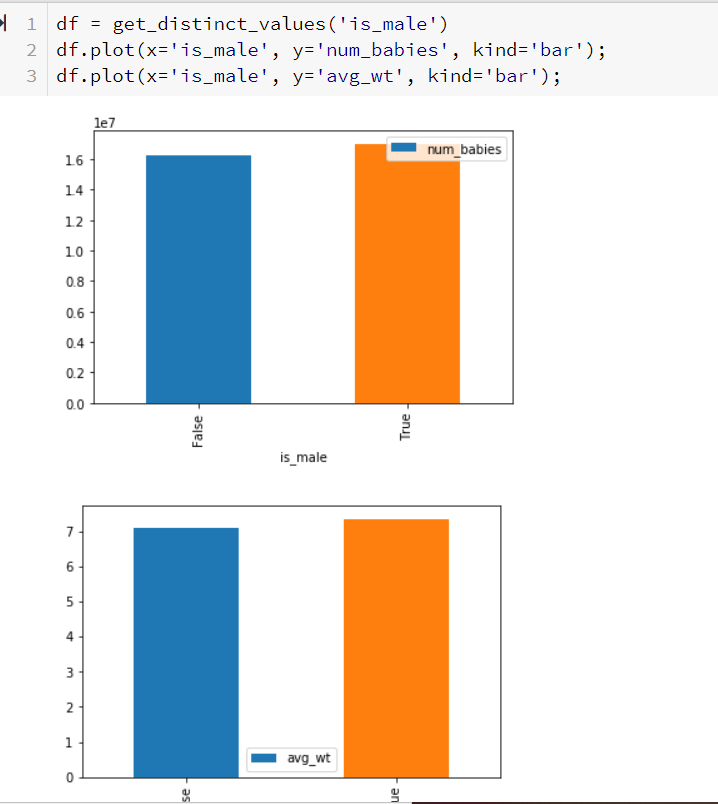
* In the first cell, set the variable PROJECT to your project ID.
* Set the variable BUCKET to your bucket name in the first cell. For your bucket name, use your project ID as a prefix and my-bucket: project-ID-my-bucket
* Leave REGION as us-central1.



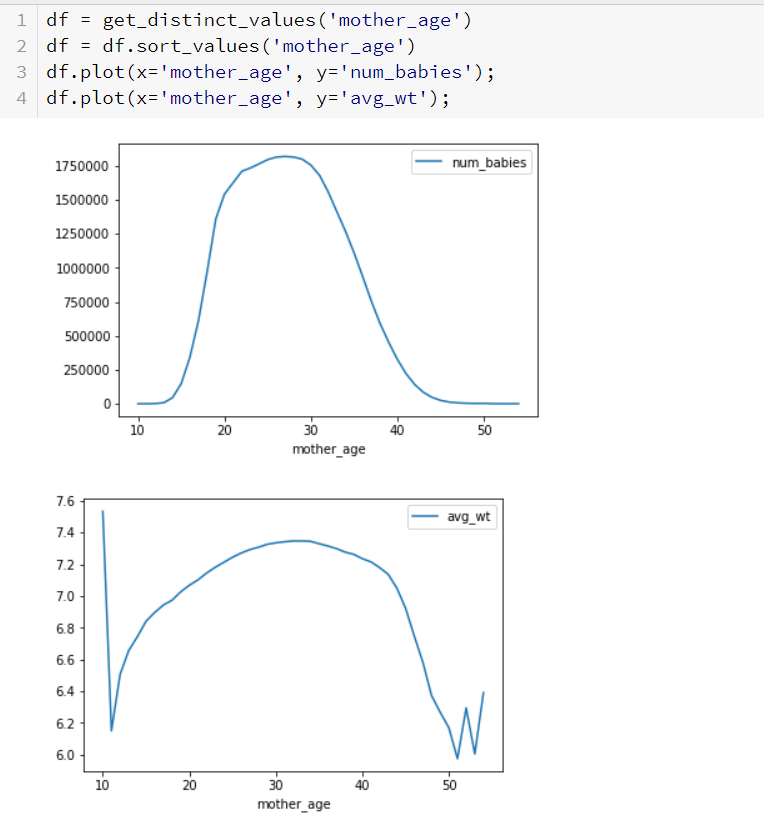
Fetching data in dataframe using BigQuery



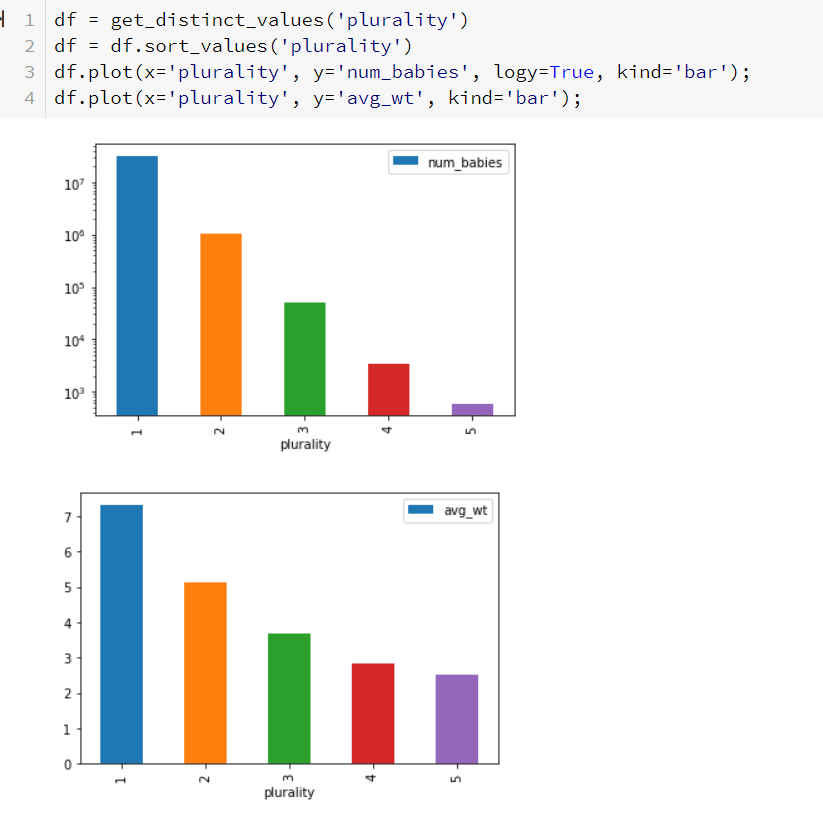
Visualizing count and average of babies



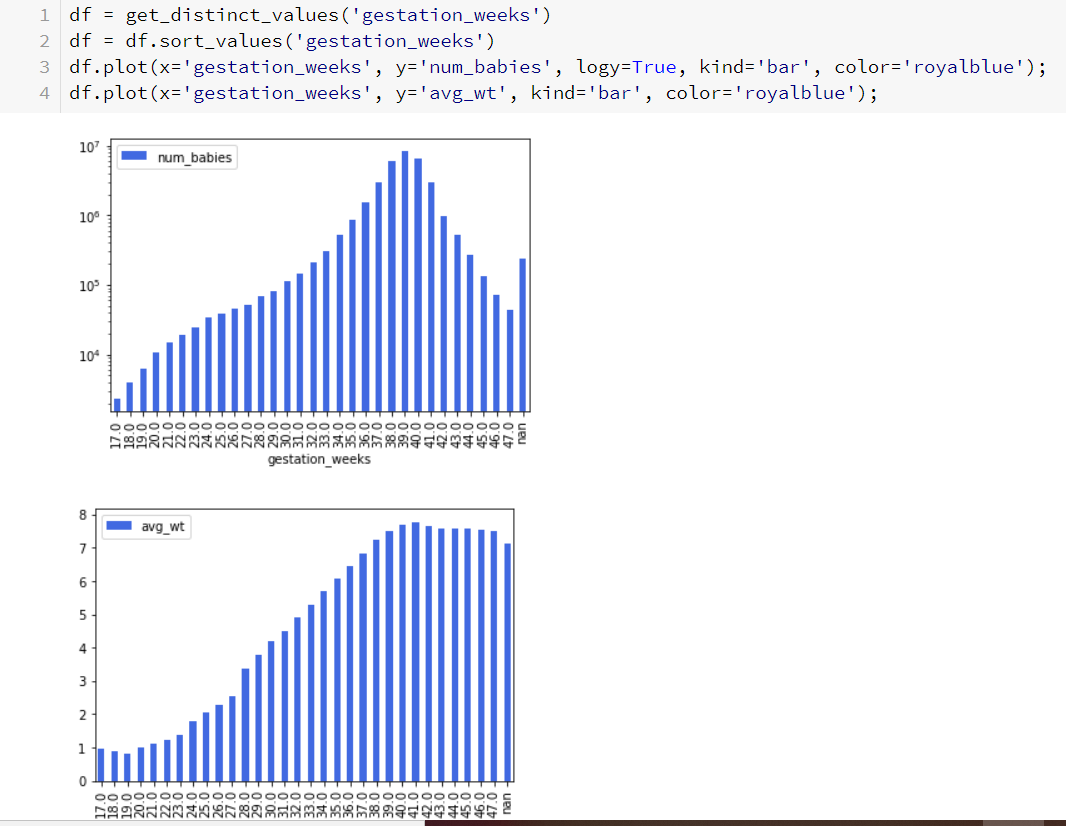
Visualizing correlation between Mother’s age and number of babies and there average weight



Visualizing plurality trends



Correlation between gestation period and babies weight and count

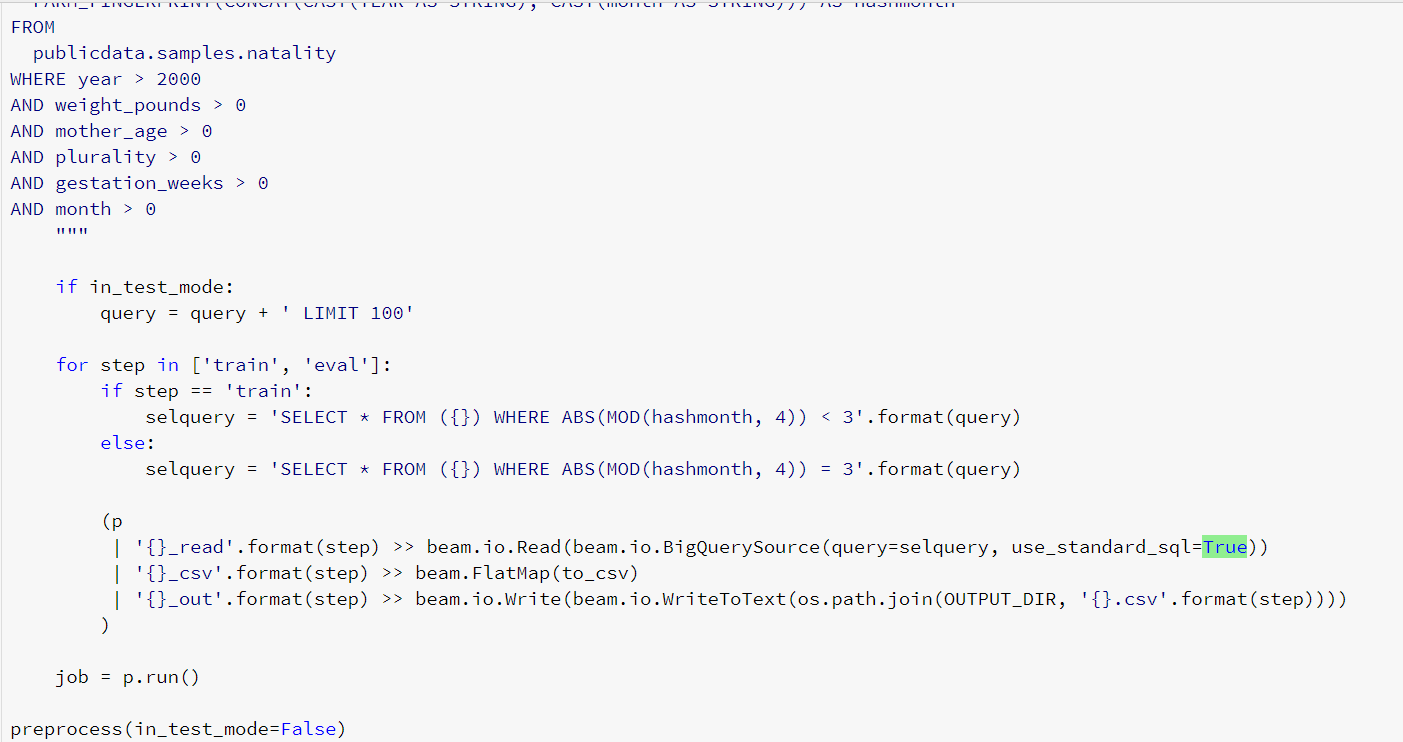


## **Preprocessing using apache beam**

We modified the data such that we can simulate what is known if no ultrasound has been performed. If I didn't need preprocessing, I could have used the web console. Also, I prefer to script it out rather than run queries on the user interface. Therefore, I am using Cloud Dataflow for the preprocessing.

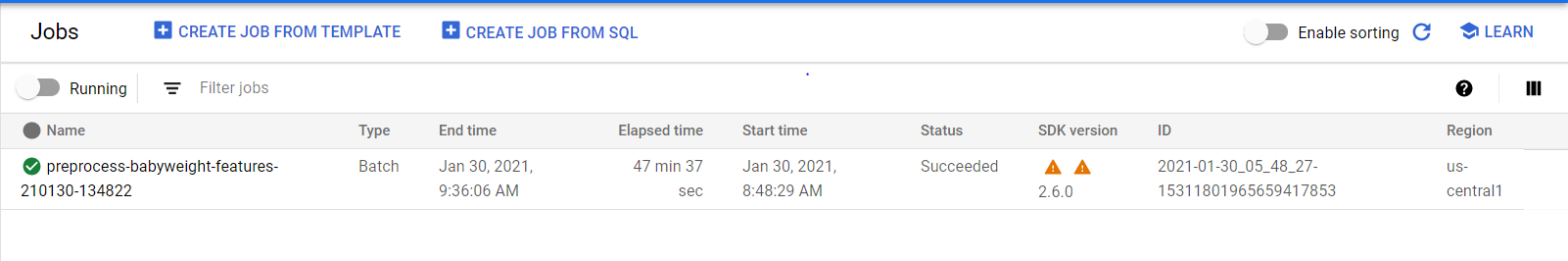


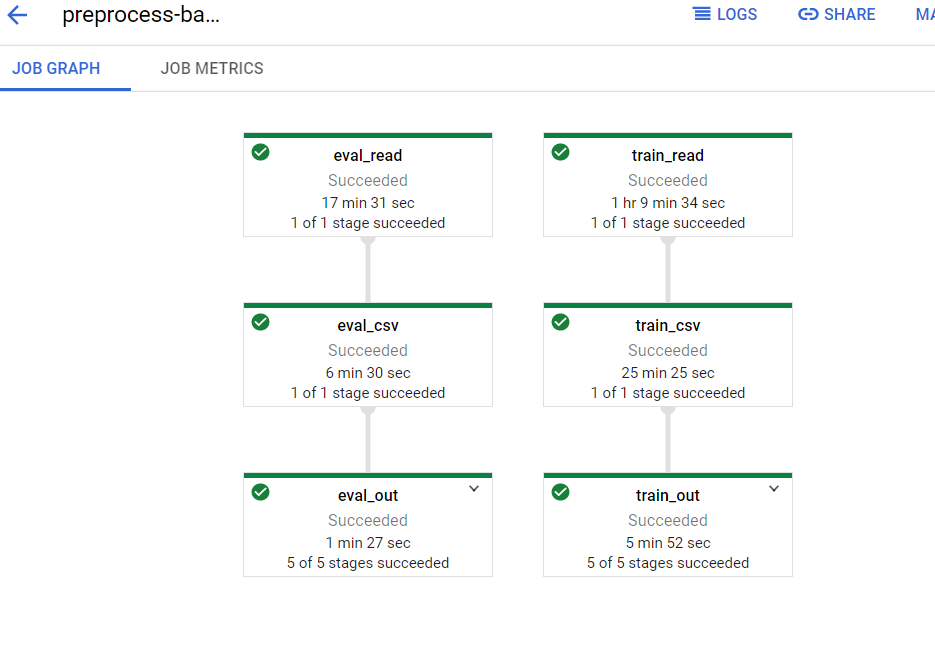




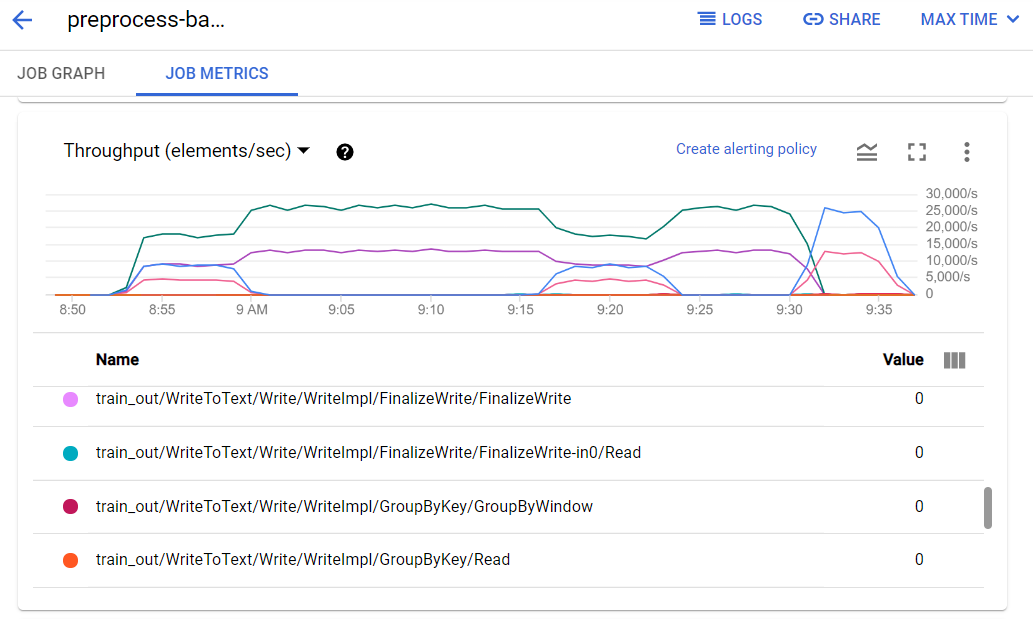
# **Results**

Job took around 48mins to finish

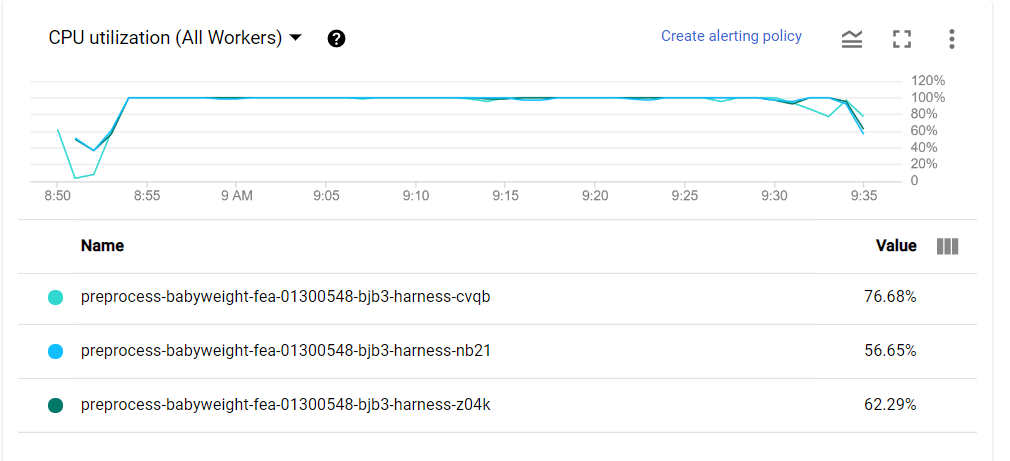




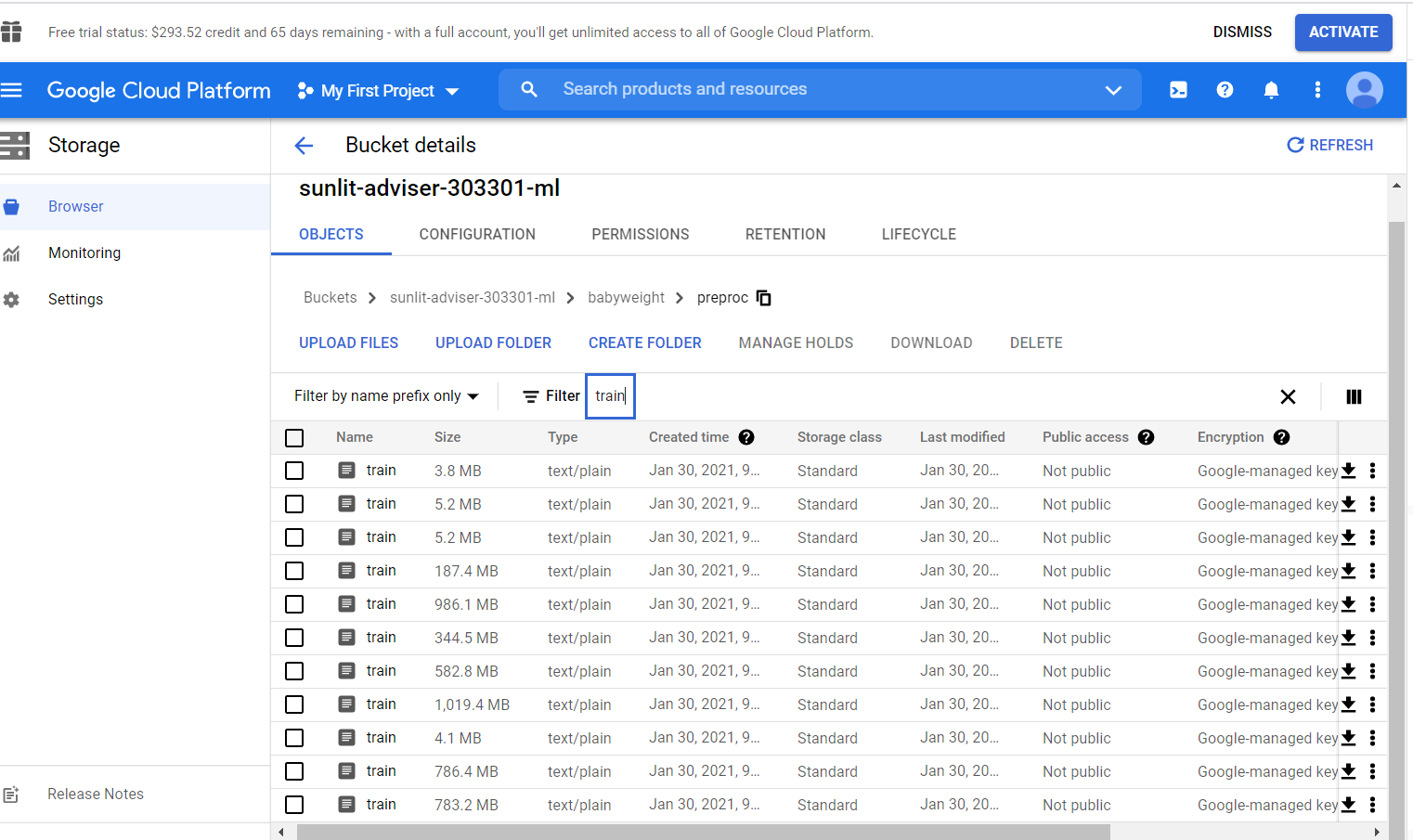
Throughput Metrices



CPU Utilization



## **Data loaded in the GCP bucket in csv form**



# **Lesson Learned**

1. Learned to set up a project in Google Cloud and instantiate DataLab from Cloud shell editor using shell commands.
2. Data exploration in Datalab and writing BigQuery to load data in dataframe and visualization.
3. Preprocessing of data using Dataflow

**References**

[**https://console.cloud.google.com/home/dashboard?project=sunlit-adviser-303301&\_ga=2.197393647.171323457.1614132709-2114907062.1611967416&\_gac=1.91198568.1613410875.Cj0KCQiA1KiBBhCcARIsAPWqoSqLlRaYf255AYUskJXnXDtQpOTIqR6qIKcffq5HM2JDrJjXDpgeKkwaAnrnEALw\_wcB**](https://console.cloud.google.com/home/dashboard?project=sunlit-adviser-303301&_ga=2.197393647.171323457.1614132709-2114907062.1611967416&_gac=1.91198568.1613410875.Cj0KCQiA1KiBBhCcARIsAPWqoSqLlRaYf255AYUskJXnXDtQpOTIqR6qIKcffq5HM2JDrJjXDpgeKkwaAnrnEALw_wcB)

[**https://en.wikipedia.org/wiki/Google\_Cloud\_Platform**](https://en.wikipedia.org/wiki/Google_Cloud_Platform)