Using Google Cloud Platform for Econometrics

GOOGLE CLOUD PLATFORM

Prerequisites

- Need a Google Cloud Platform account
  - Go to https://cloud.google.com/ and set up a free trial account.
- Need a background in econometrics / multi-variate statistics.
- Some prior background in Python / R is helpful.

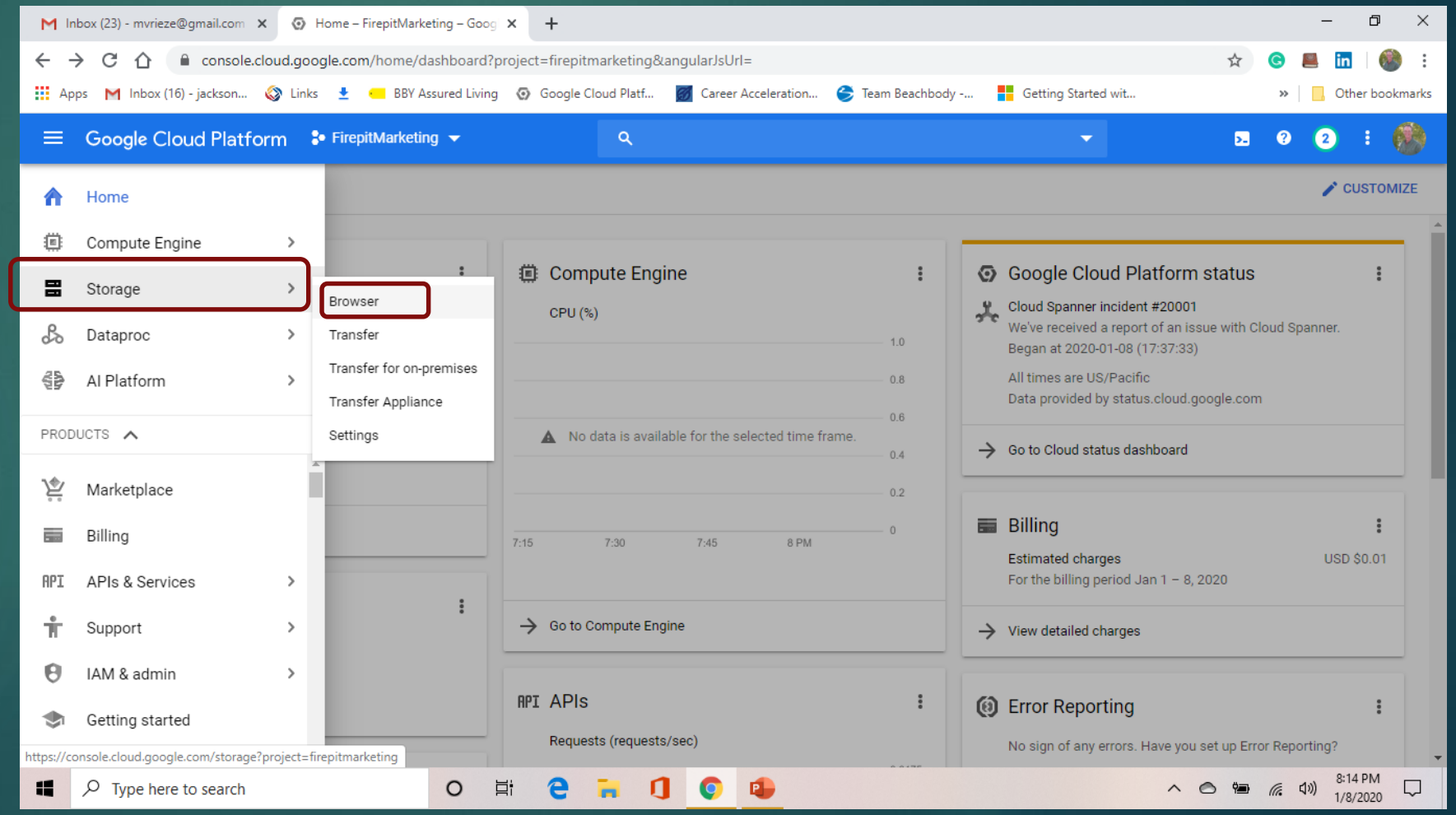
© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

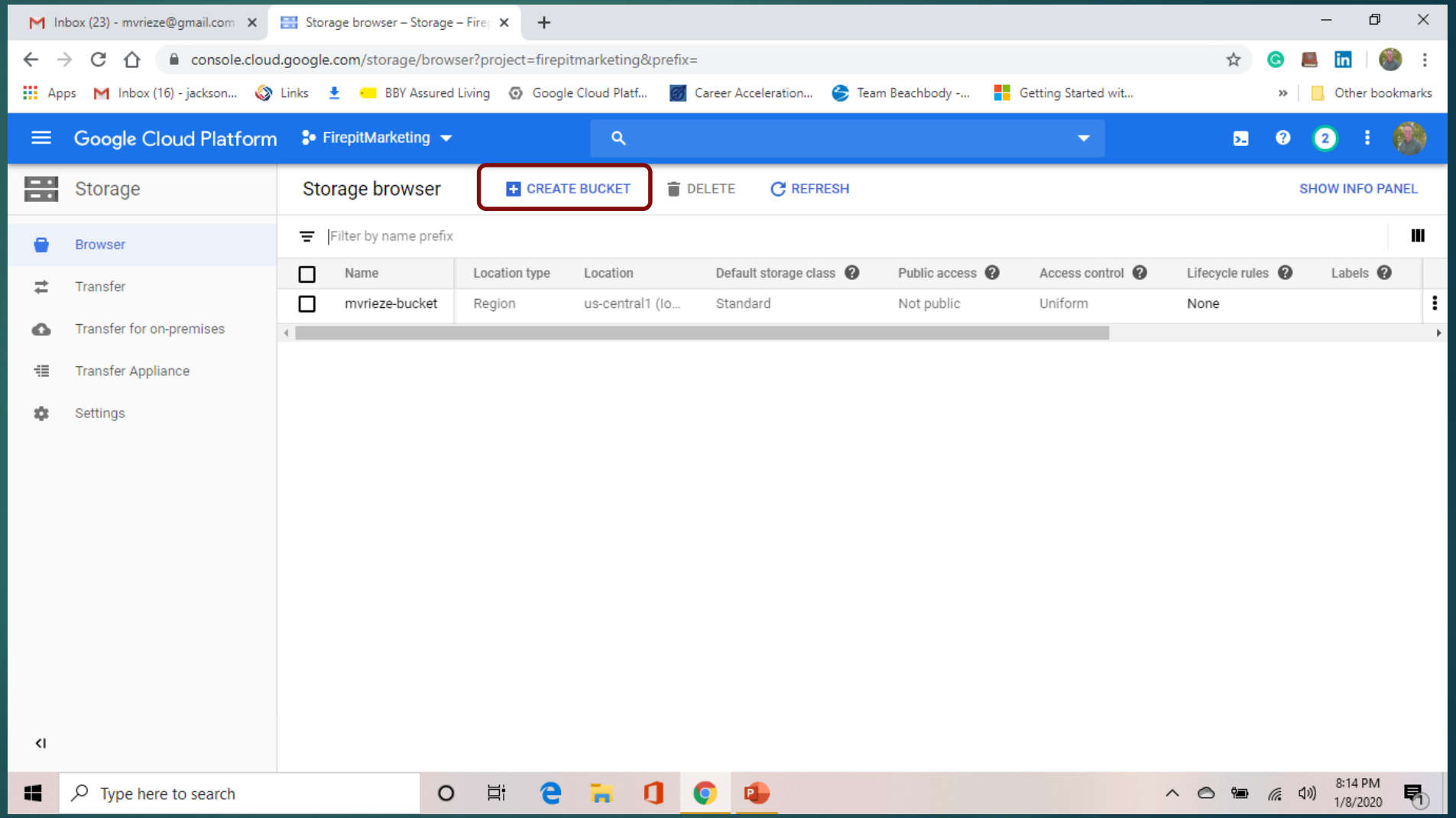
- Navigate to "Storage"
- Select "Browser"



© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

- Select "Create Bucket"



© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

- Name your bucket
- Choose a location type - "Region"

console.cloud.google.com/storage/create-bucket?project=firepitmarketing

Google Cloud Platform | FirepitMarketing

Storage

Browser
Transfer
Transfer for on-premises
Transfer Appliance
Settings

Create a bucket

Name your bucket

Pick a **globally unique**, permanent name. Naming guidelines

mvrieze-bucket2

Tip: Don't include any sensitive information

CONTINUE

Choose where to store your data

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. Learn more

Location type

- Region — Lowest latency within a single region
- Multi-region — Highest availability across largest area
- Dual-region — High availability and low latency across 2 regions

Location

Monthly cost estimate

Enter values below to check this bucket's monthly cost. For guidance only. Pricing details

Storage and retrieval

Storage size — GB
$0.020 per GB-month

Data retrieval size — GB
Free

Operations

Class A operations — per-month
$0.005 per 1,000 ops

Class B operations — per-month
$0.0004 per 1,000 ops

© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

- Choose a Location – In this case, "us-west1 (Oregon)"
- Choose the "Standard" storage class

Inbox (23) - mvrieze@gmail.com | Storage – FirepitMarketing – Goo

console.cloud.google.com/storage/create-bucket?project=firepitmarketing

Apps | Inbox (16) - jackson... | Links | BBY Assured Living | Google Cloud Platf... | Career Acceleration... | Team Beachbody -... | Getting Started wit... | Other bookmarks

Google Cloud Platform | FirepitMarketing

Storage

- Browser
- Transfer
- Transfer for on-premises
- Transfer Appliance
- Settings

Create a bucket

Location

us-west1 (Oregon)

CONTINUE

Choose a default storage class for your data

A storage class sets costs for storage, retrieval, and operations. Pick a default storage class based on how long you plan to store your data and how often it will be accessed. Learn more

- Standard — Best for short-term storage and frequently accessed data
- Nearline — Best for backups and data accessed less than once a month
- Coldline — Best for disaster recovery and data accessed less than once a quarter
- Archive — Best for long-term digital preservation of data accessed less than once a year

CONTINUE

Monthly cost estimate

Enter values below to check this bucket's monthly cost. For guidance only. Pricing details

Storage and retrieval

Storage size — GB
$0.020 per GB-month

Data retrieval size — GB
Free

Operations

Class A operations — per-month
$0.005 per 1,000 ops

Class B operations — per-month
$0.0004 per 1,000 ops

Type here to search | 8:19 PM 1/8/2020

© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

- Accept the default access control (Google-managed key)
- Choose "Create" (at the bottom) to complete setup.

Inbox (23) - mvrieze@gmail.com | Storage – FirepitMarketing – Goo

console.cloud.google.com/storage/create-bucket?project=firepitmarketing

Apps | Inbox (16) - jackson... | Links | BBY Assured Living | Google Cloud Platf... | Career Acceleration... | Team Beachbody -... | Getting Started wit... | Other bookmarks

Google Cloud Platform | FirepitMarketing

Storage

- Browser
- Transfer
- Transfer for on-premises
- Transfer Appliance
- Settings

Create a bucket

Choose how to control access to objects

Advanced settings (optional)

Encryption

- Google-managed key
  No configuration required
- Customer-managed key
  Manage via Google Cloud Key Management Service

Retention policy

Set a retention policy to specify the minimum duration that this bucket's objects must be protected from deletion or modification after they're uploaded. You might set a policy to address industry-specific retention challenges. Learn more

☐ Set a retention policy

Labels

Labels are key:value pairs that allow you to group related buckets together or with other Cloud Platform resources. Learn more

+ ADD LABEL

Monthly cost estimate

Enter values below to check this bucket's monthly cost. For guidance only. Pricing details

Storage and retrieval

Storage size — GB
$0.020 per GB-month

Data retrieval size — GB
Free

Operations

Class A operations — per-month
$0.005 per 1,000 ops

Class B operations — per-month
$0.0004 per 1,000 ops

Type here to search

8:19 PM
1/8/2020

© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

- Select the storage bucket you created

Inbox (23) - mvrieze@gmail.com | Storage browser – Storage – Fire

console.cloud.google.com/storage/browser?project=firepitmarketing&prefix=

Apps | Inbox (16) - jackson... | Links | BBY Assured Living | Google Cloud Platf... | Career Acceleration... | Team Beachbody -... | Getting Started wit... | Other bookmarks

Google Cloud Platform | FirepitMarketing

Storage

- Browser
- Transfer
- Transfer for on-premises
- Transfer Appliance
- Settings

Storage browser | CREATE BUCKET | DELETE | REFRESH | SHOW INFO PANEL

Filter by name prefix

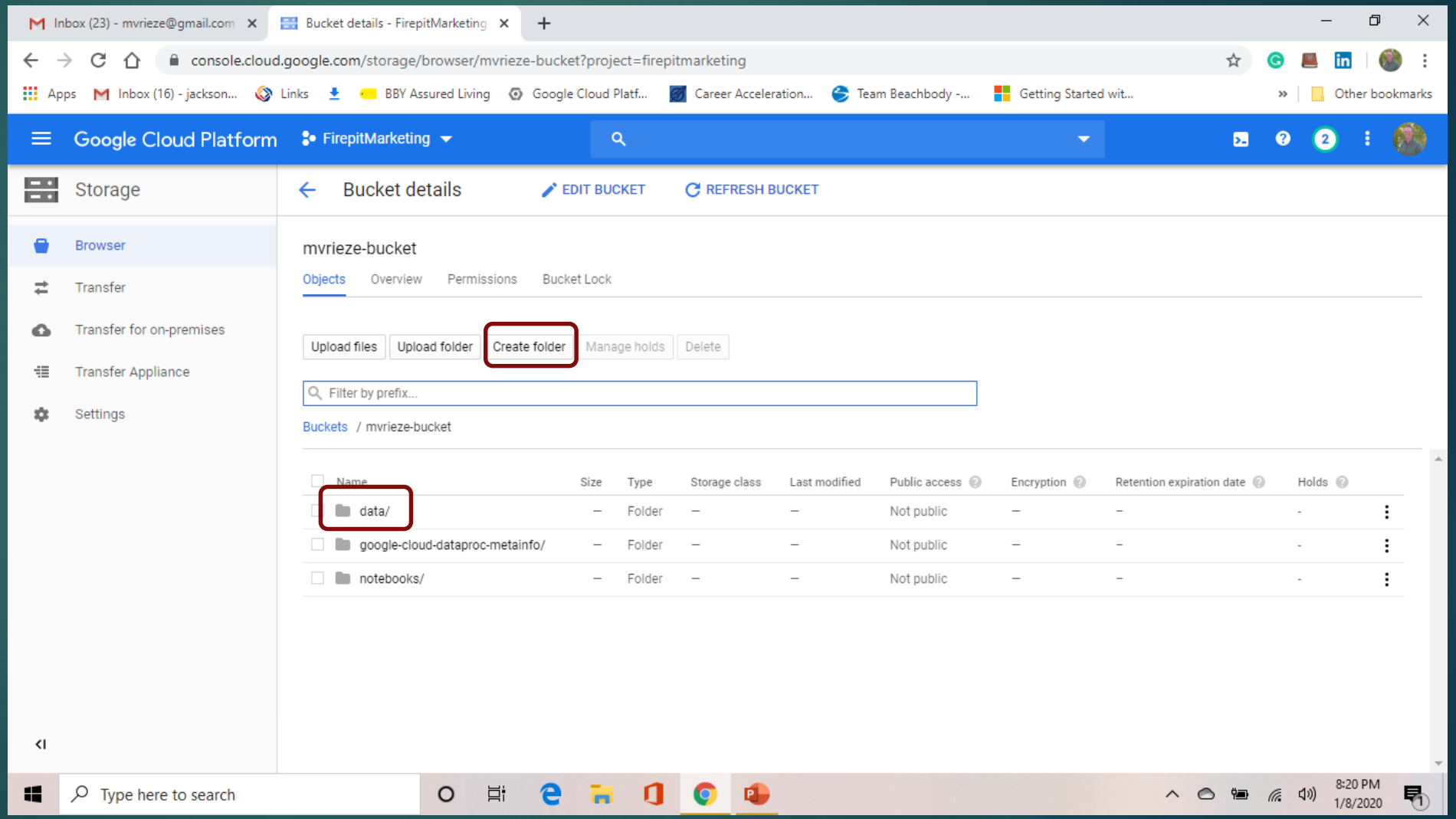
Name	Location type	Location	Default storage class	Public access	Access control	Lifecycle rules	Labels
mvrieze-bucket	Region	us-central1 (lo...	Standard	Not public	Uniform	None	

Type here to search | 8:14 PM 1/8/2020

© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

- Select "Create Folder" to set up a sub-folder inside your bucket and call it "data".



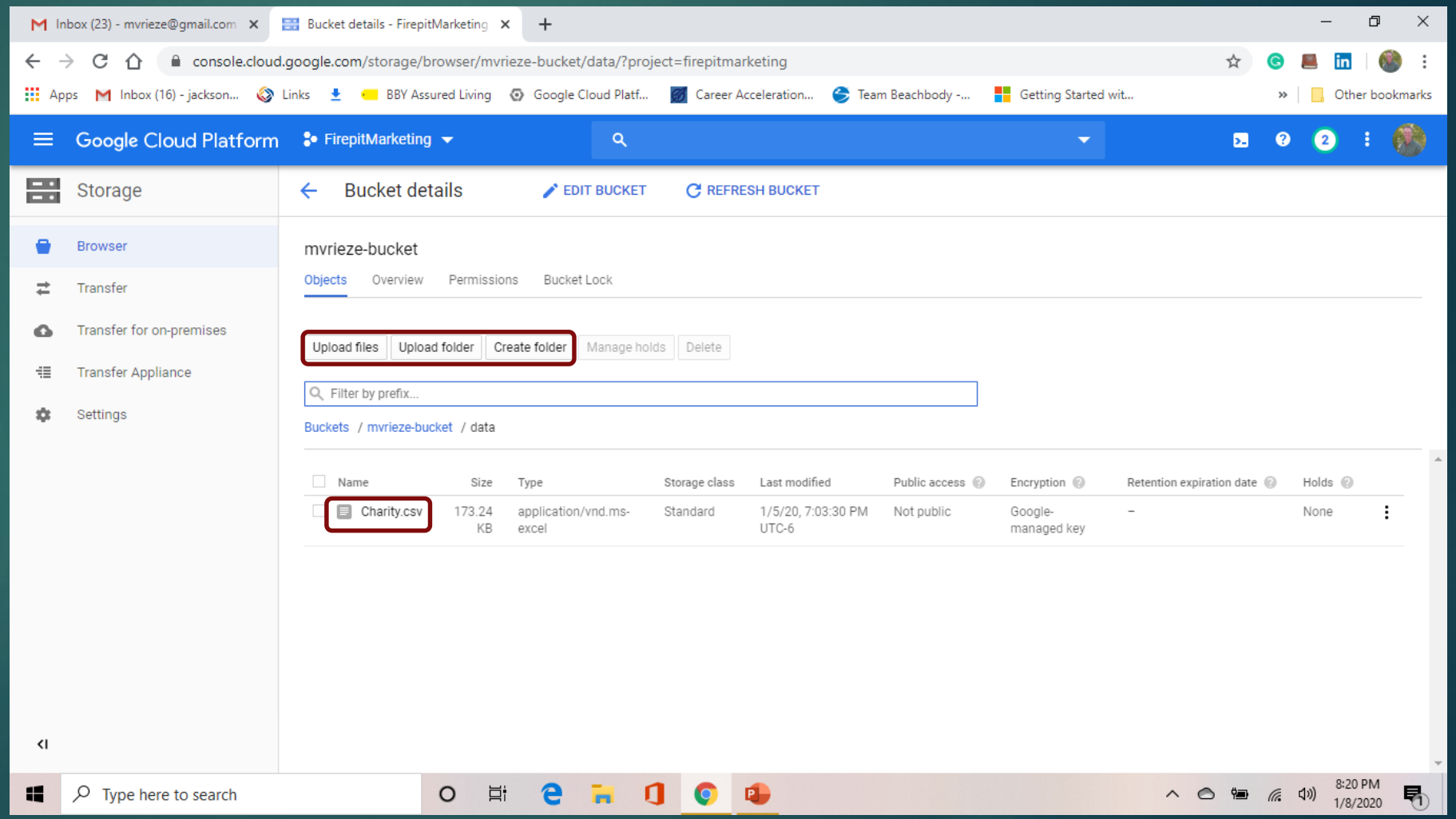
Note that GCP will also use your storage bucket for other objects besides data. The screenshot also contains folders for Notebooks as well as Dataproc metadata.

© Copyright 2020 - Martin Vrieze

Storage Bucket Setup

- Use the "Upload files" or "Upload folders" or "Create folder" buttons.
- Note we have already uploaded the "Charity.csv" dataset to our "/data" folder.

Note that upload time depends on the speed of your internet connection, the size of the file and how you connect to GCP. For larger data sets, you might consider using the Google Cloud SDK tool.



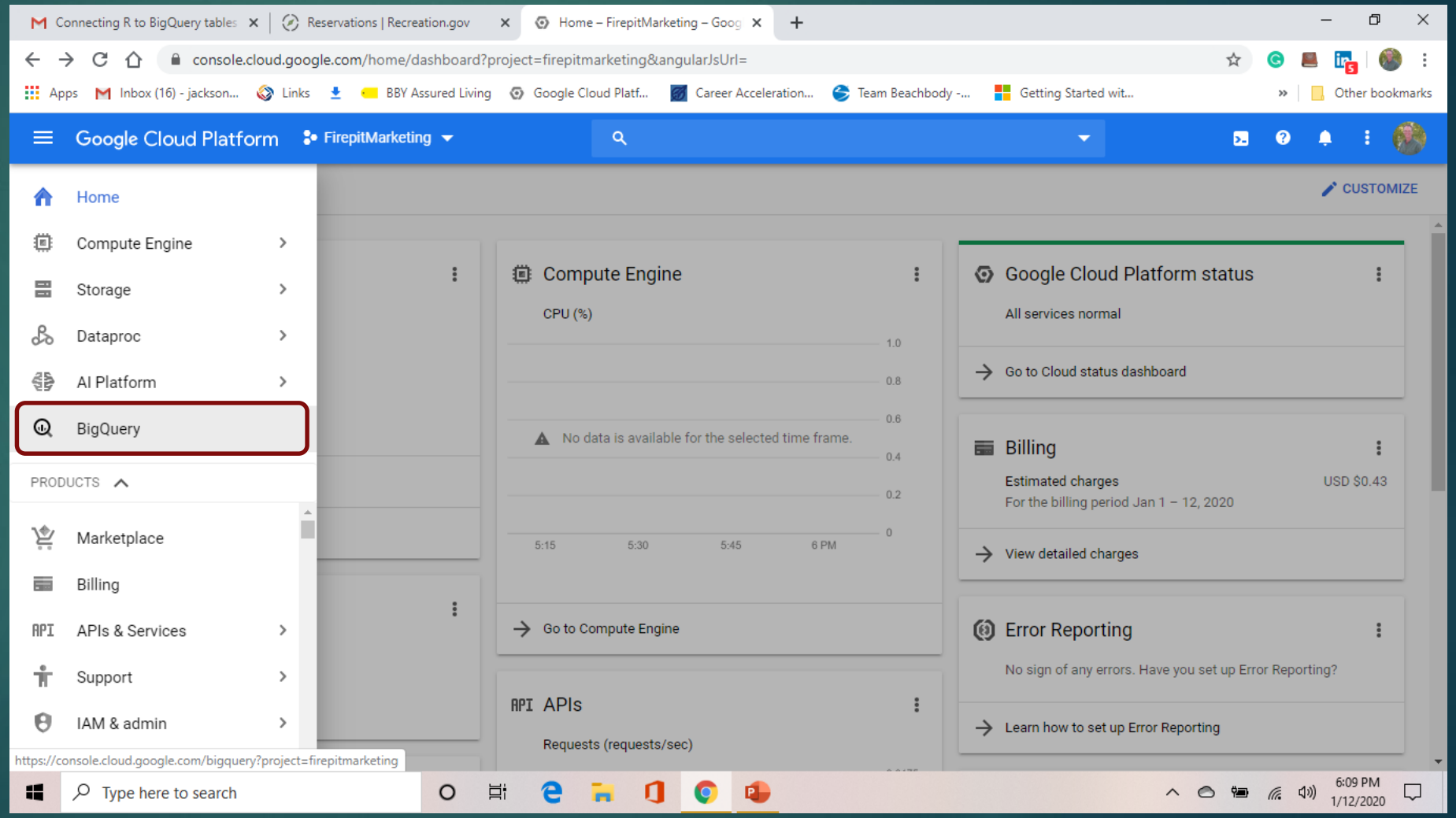
© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

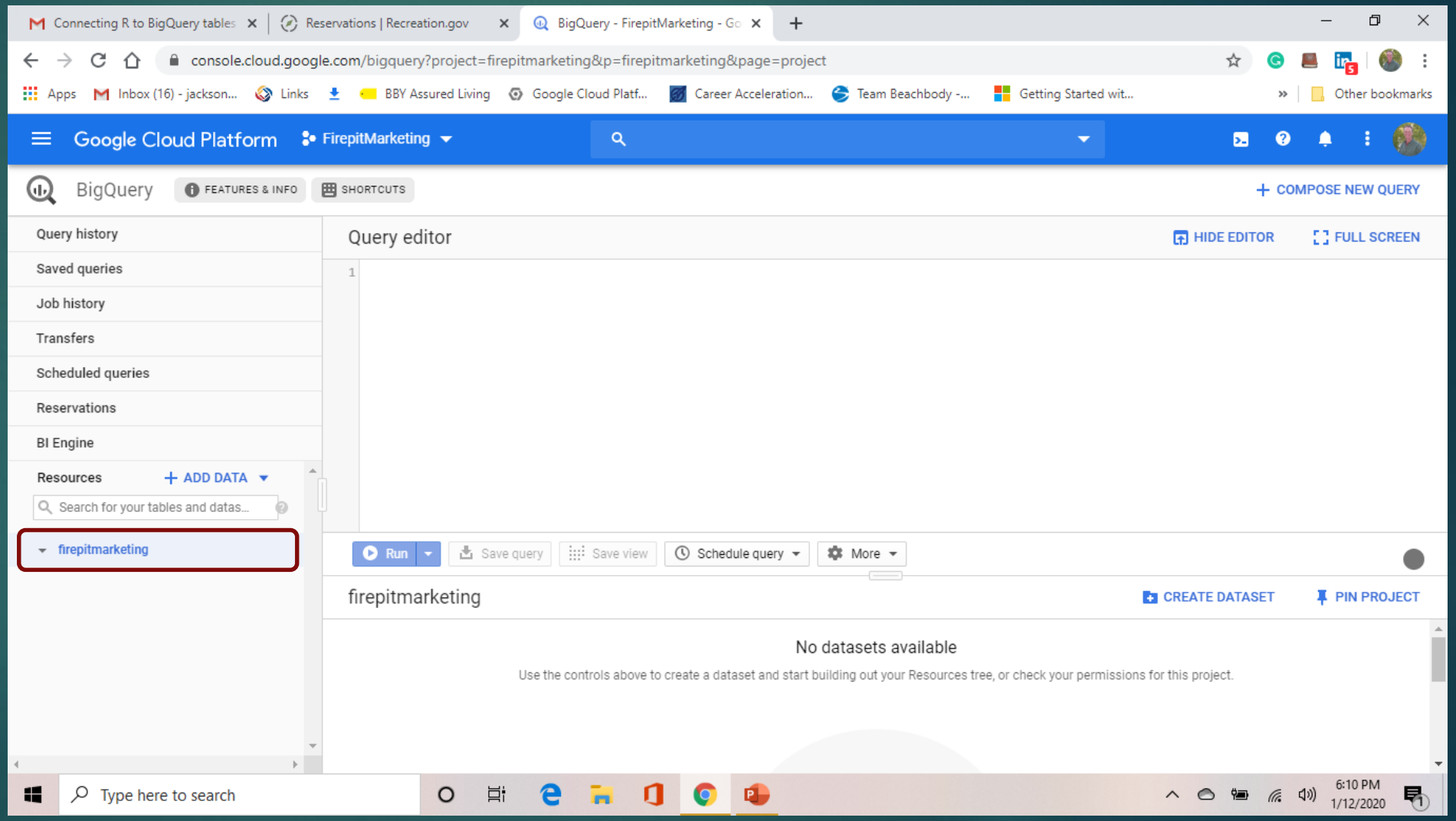
- Select "BigQuery.



© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

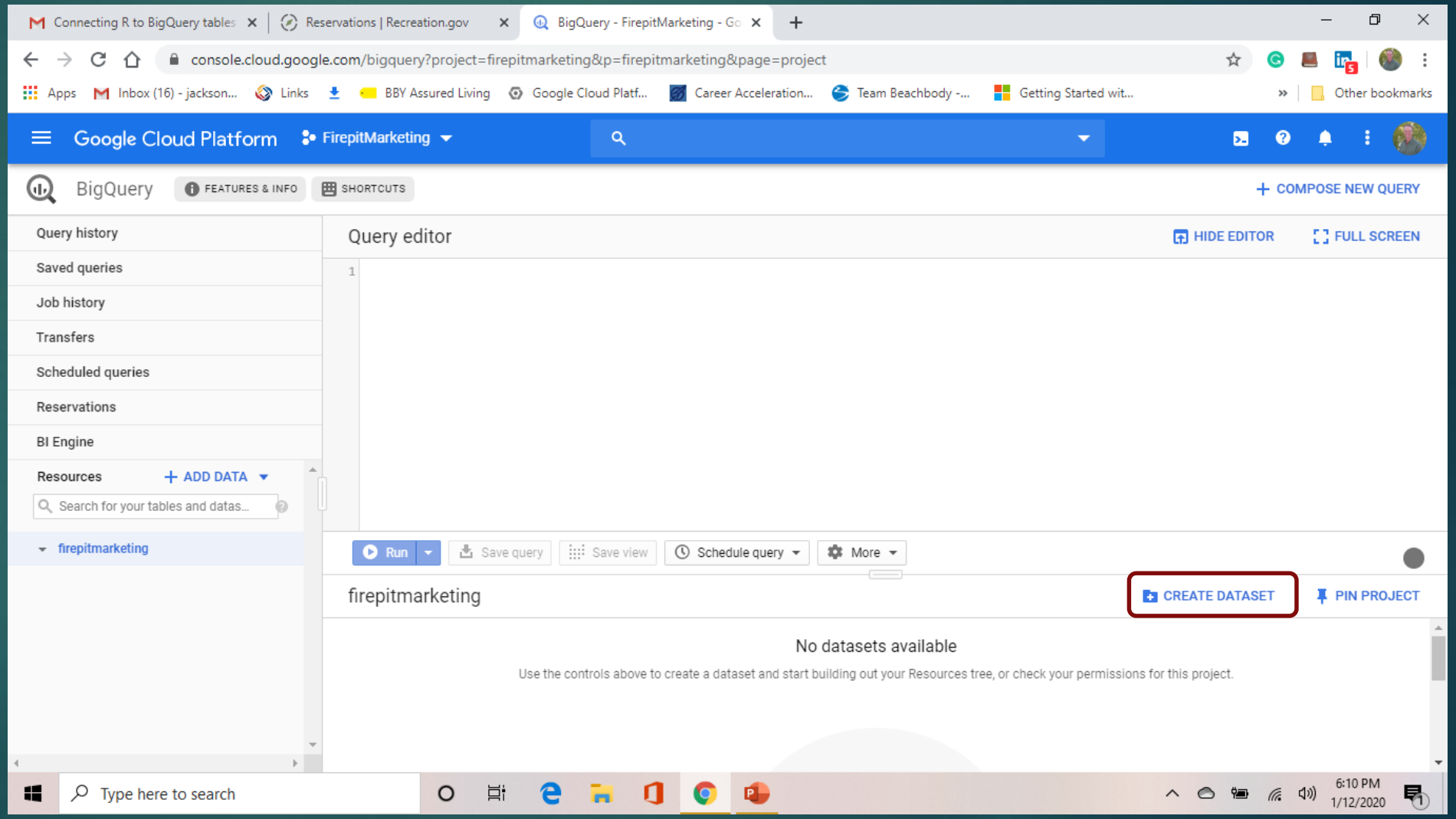
- Click on your project name.



© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

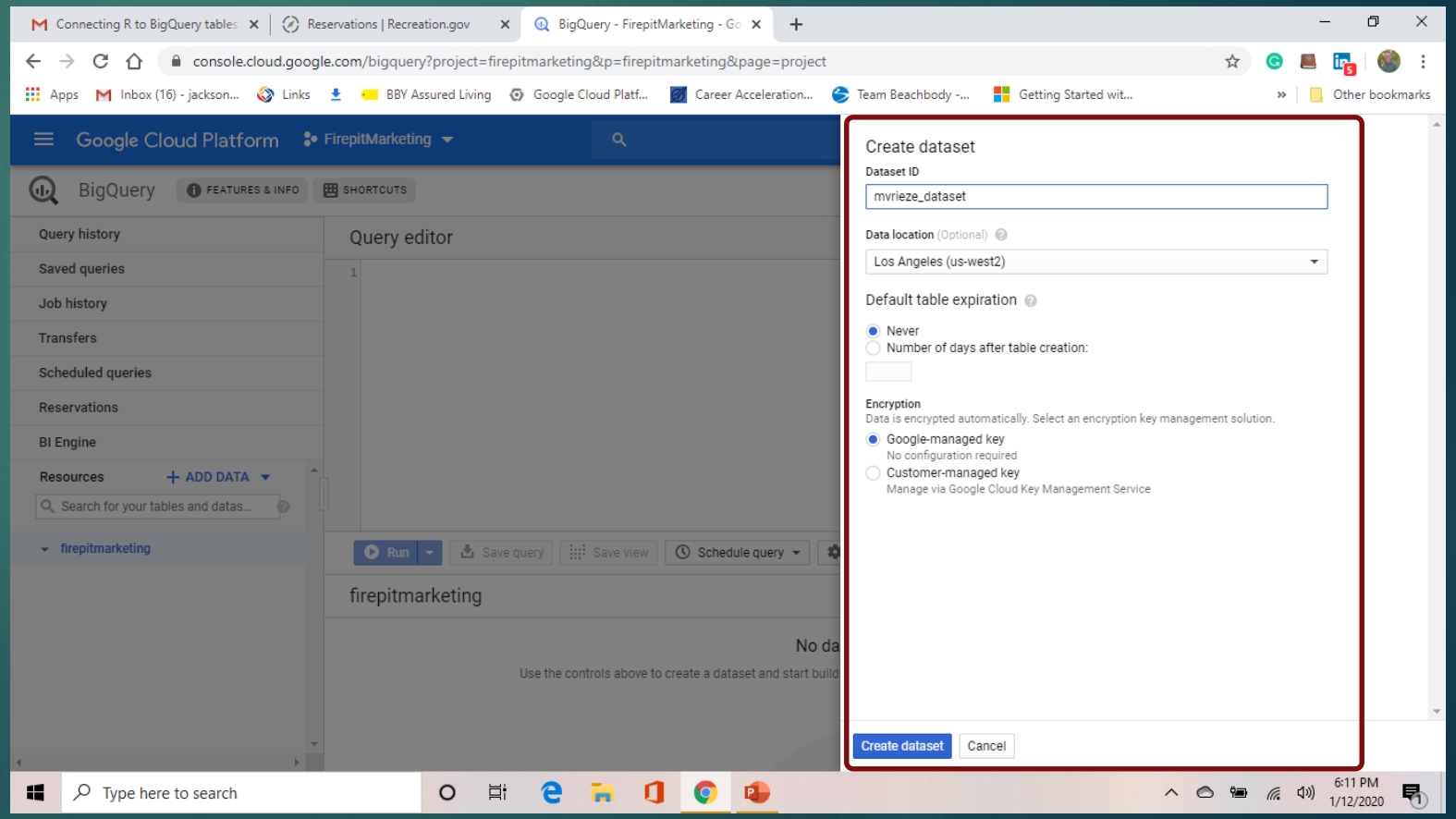
- Select "Create Dataset"



© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

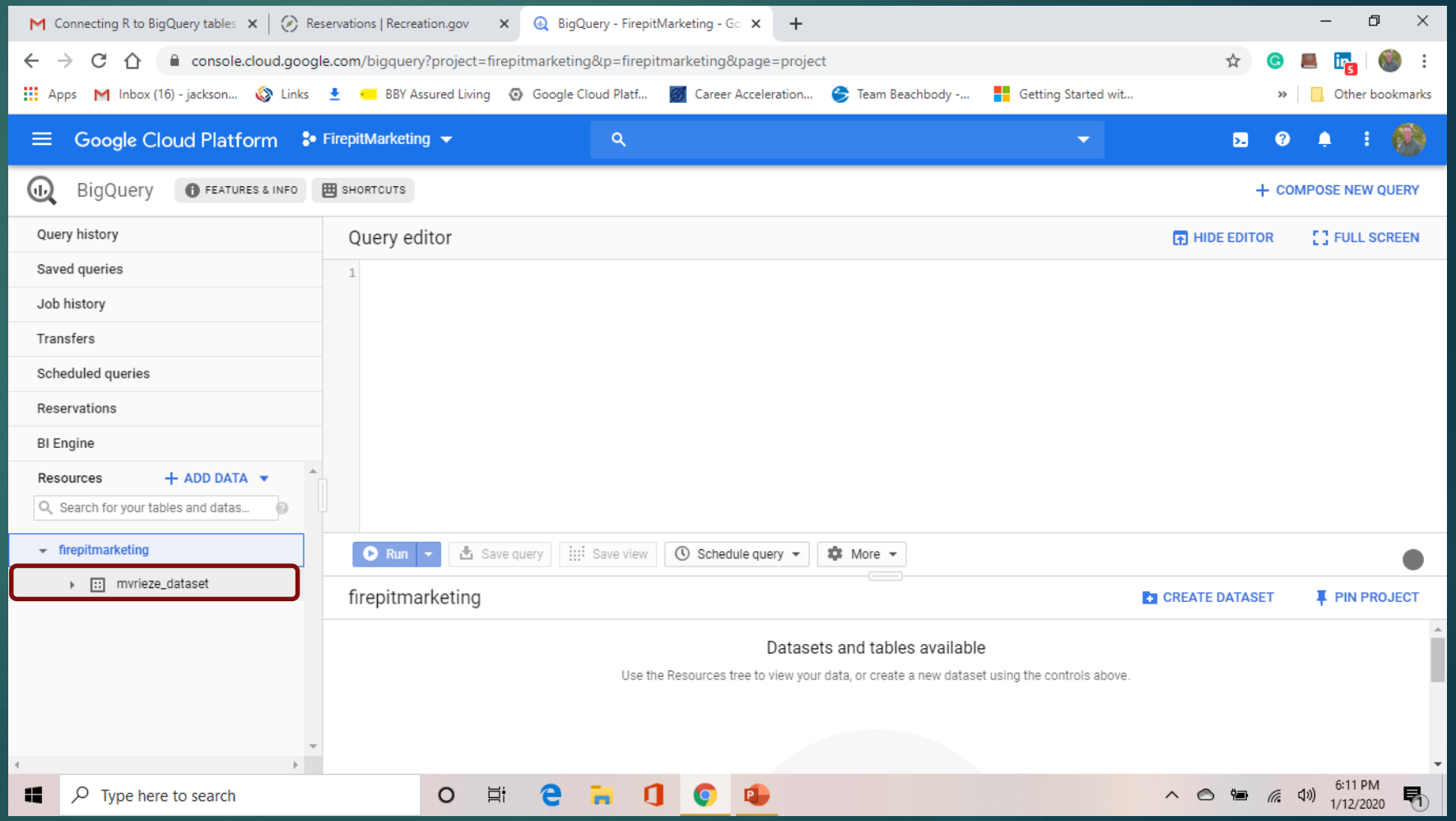
- Set a name for the dataset
- Choose a region – We recommend the generic "United States" region
- Leave the default table expiration and encryption defaults and select "Create Dataset.



© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

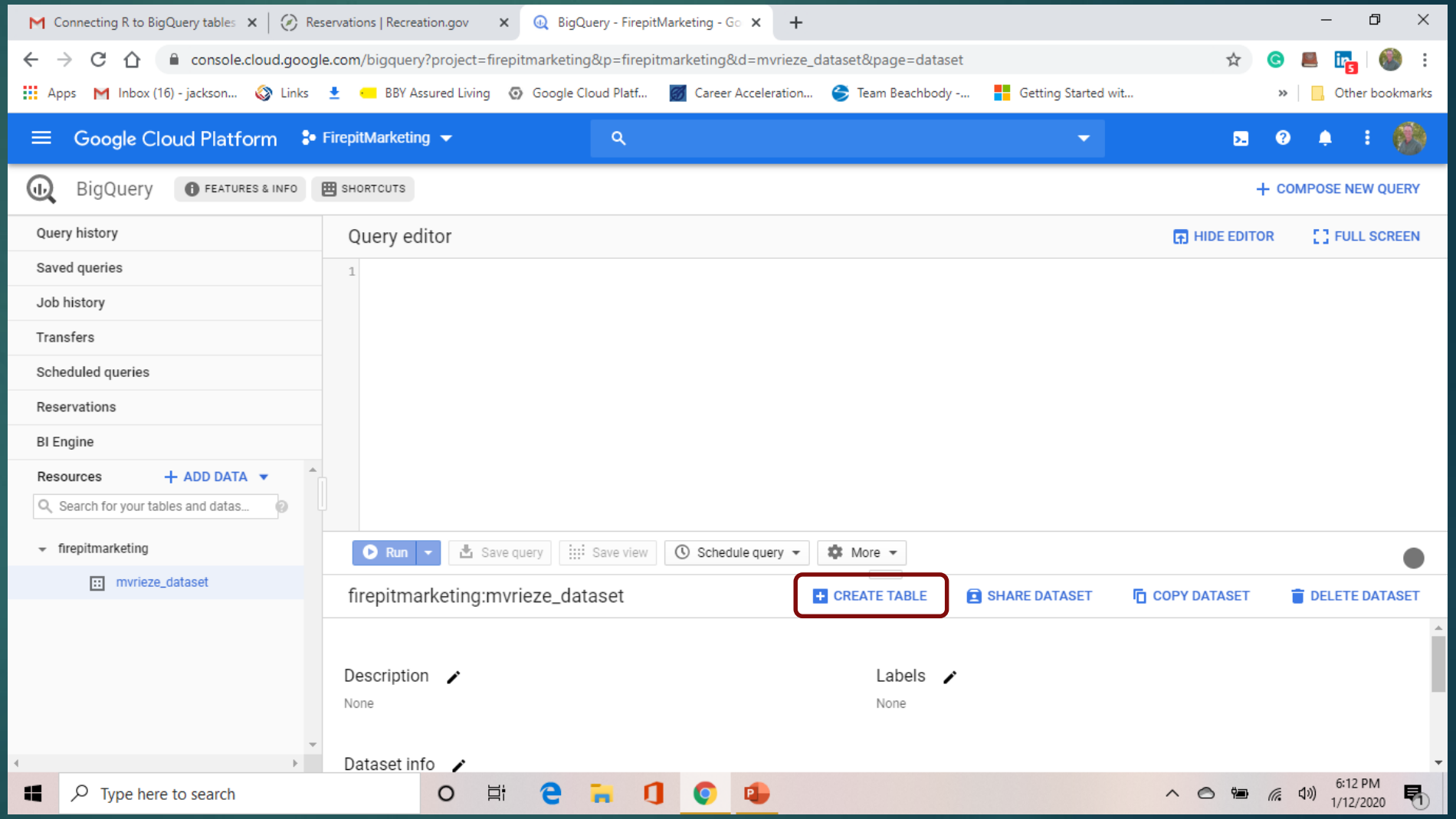
- Your dataset now appears below your project name.
- Click on your new dataset.



© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

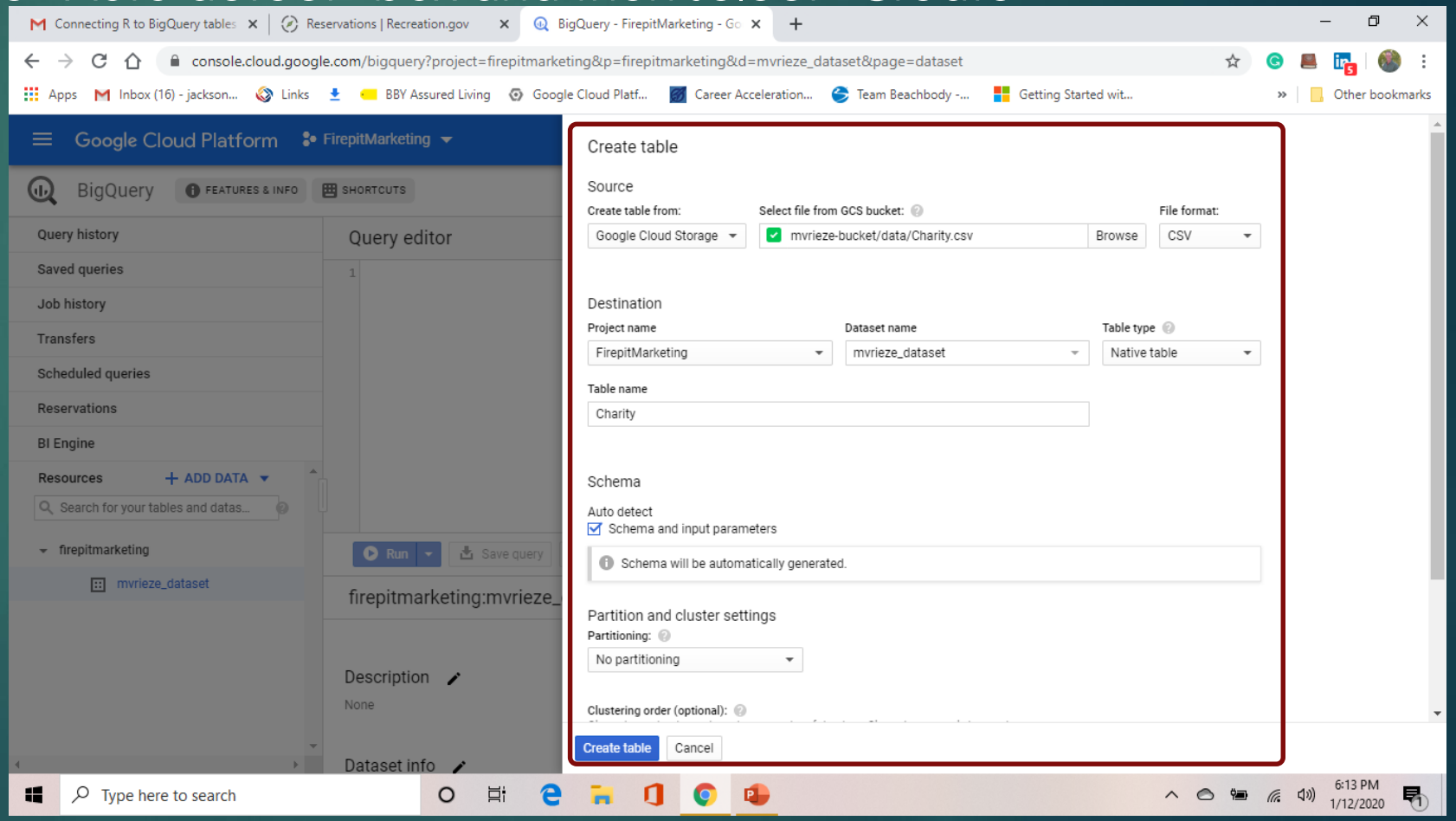
- Click on "Create Table".



© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

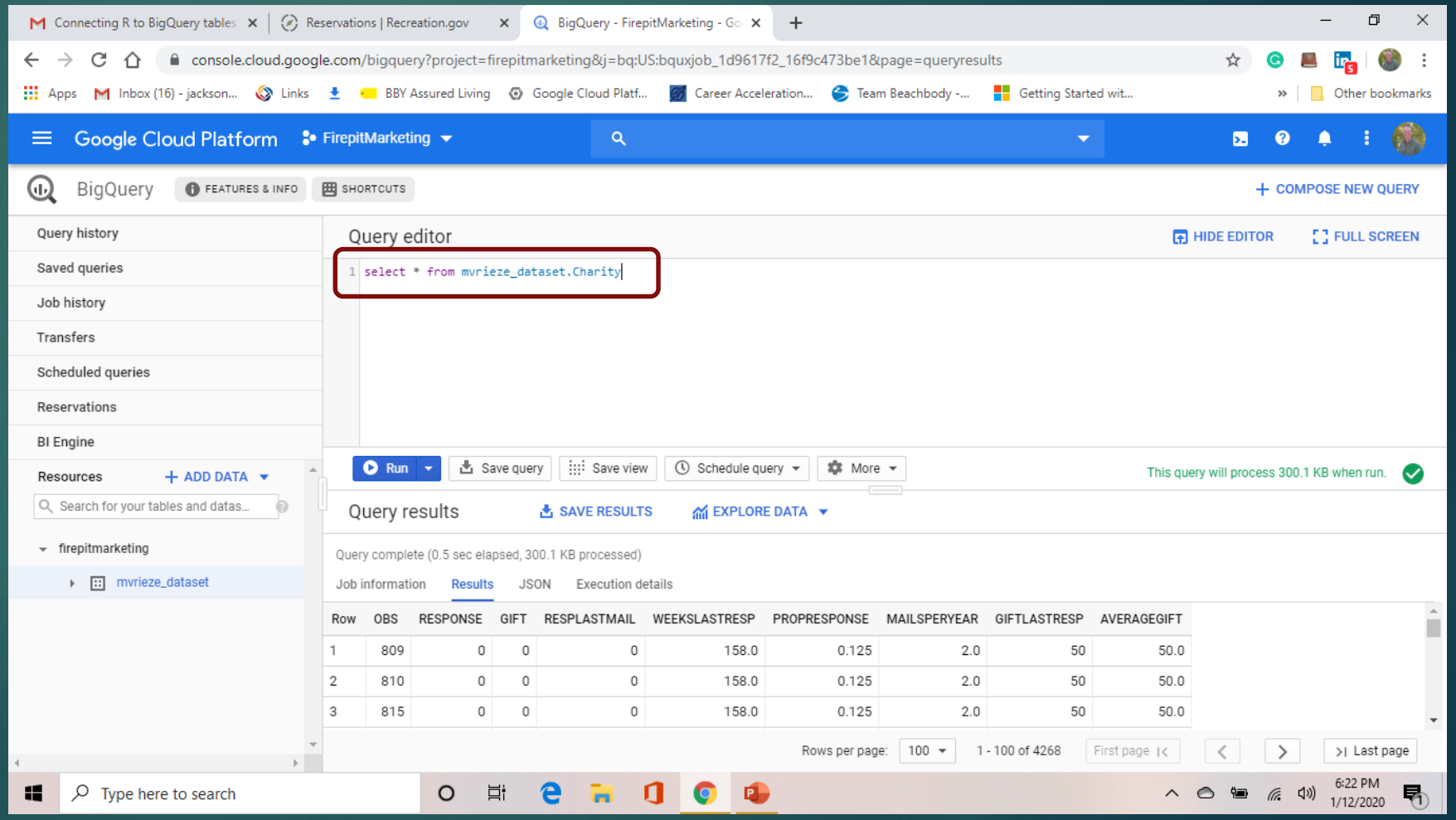
- Select Google Cloud Storage and browse/select the Charity.csv file we uploaded earlier.
- Be sure to check the "Auto detect" box and then select "Create Table"



© Copyright 2020 - Martin Vrieze

Load Data to BigQuery

- In the query editor, write a simple select query and select "Run"
- Results appear in bottom box.



Row	OBS	RESPONSE	GIFT	RESPLASTMAIL	WEEKSLASTRESP	PROPRESPONSE	MAILSPERYEAR	GIFTLASTRESP	AVERAGEGIFT
1	809	0	0	0	158.0	0.125	2.0	50	50.0
2	810	0	0	0	158.0	0.125	2.0	50	50.0
3	815	0	0	0	158.0	0.125	2.0	50	50.0

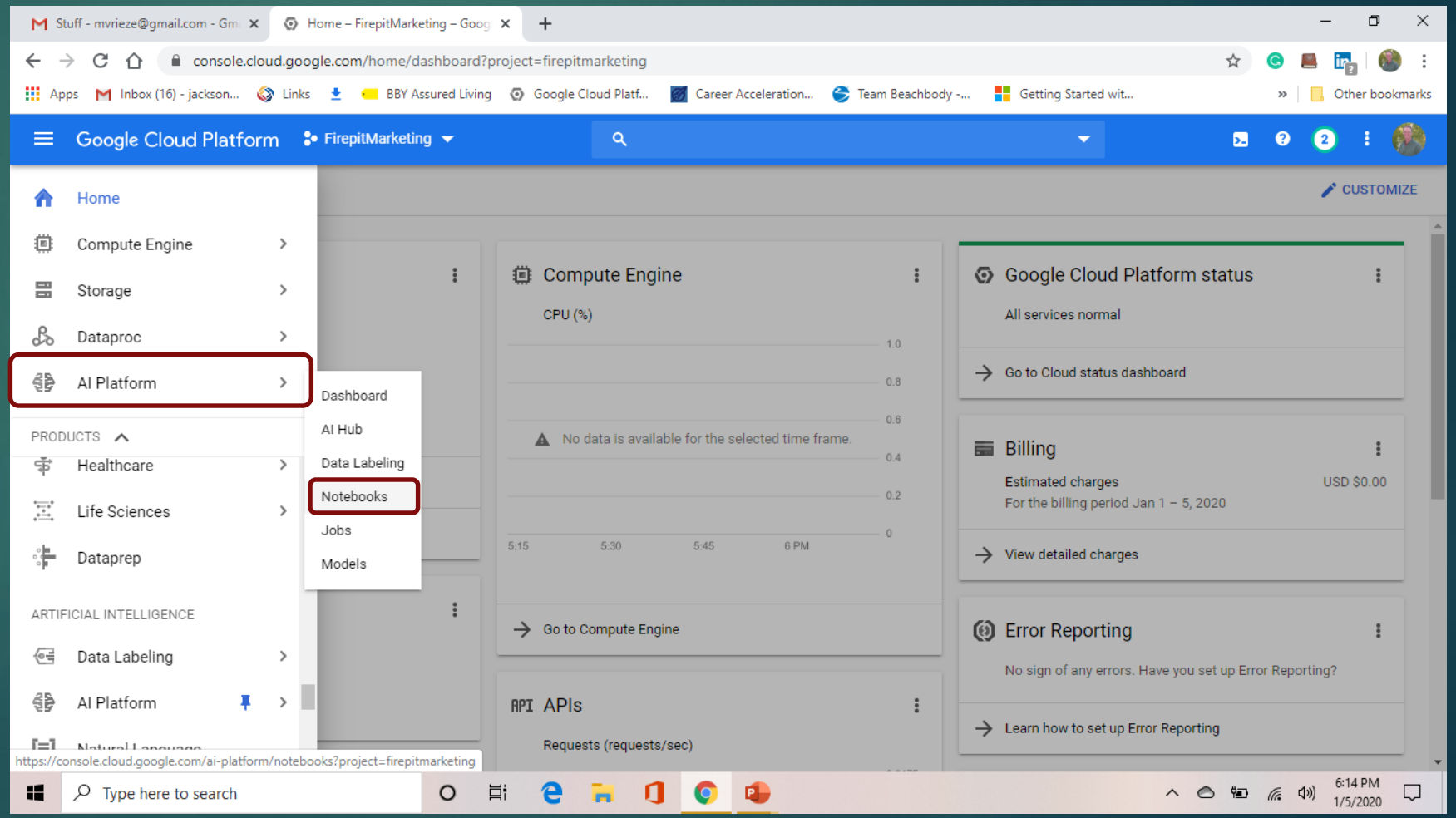
© Copyright 2020 - Martin Vrieze

Virtual Machine Setup

© Copyright 2020 - Martin Vrieze

Virtual Machine Setup

- Navigate to "AI Platform"
- Select "Notebooks"



© Copyright 2020 - Martin Vrieze

Virtual Machine Setup

- Select "+ New Instance"

Stuff - mvrieze@gmail.com - Gm
Notebook instances – AI Platform
console.cloud.google.com/ai-platform/notebooks/instances?project=firepitmarketing
Apps | Inbox (16) - jackson... | Links | BBY Assured Living | Google Cloud Platf... | Career Acceleration... | Team Beachbody -... | Getting Started wit... | Other bookmarks

Google Cloud Platform | FirepitMarketing

AI Platform

- Dashboard
- AI Hub
- Data Labeling
- Notebooks
- Jobs
- Models

Notebook instances BETA | NEW INSTANCE | REFRESH | START | STOP | RESET | DELETE | SHOW INFO PANEL

Create and use Jupyter Notebooks with a notebook instance. Notebook instances have JupyterLab pre-installed and are configured with GPU-enabled machine learning frameworks. Learn more

Filter table

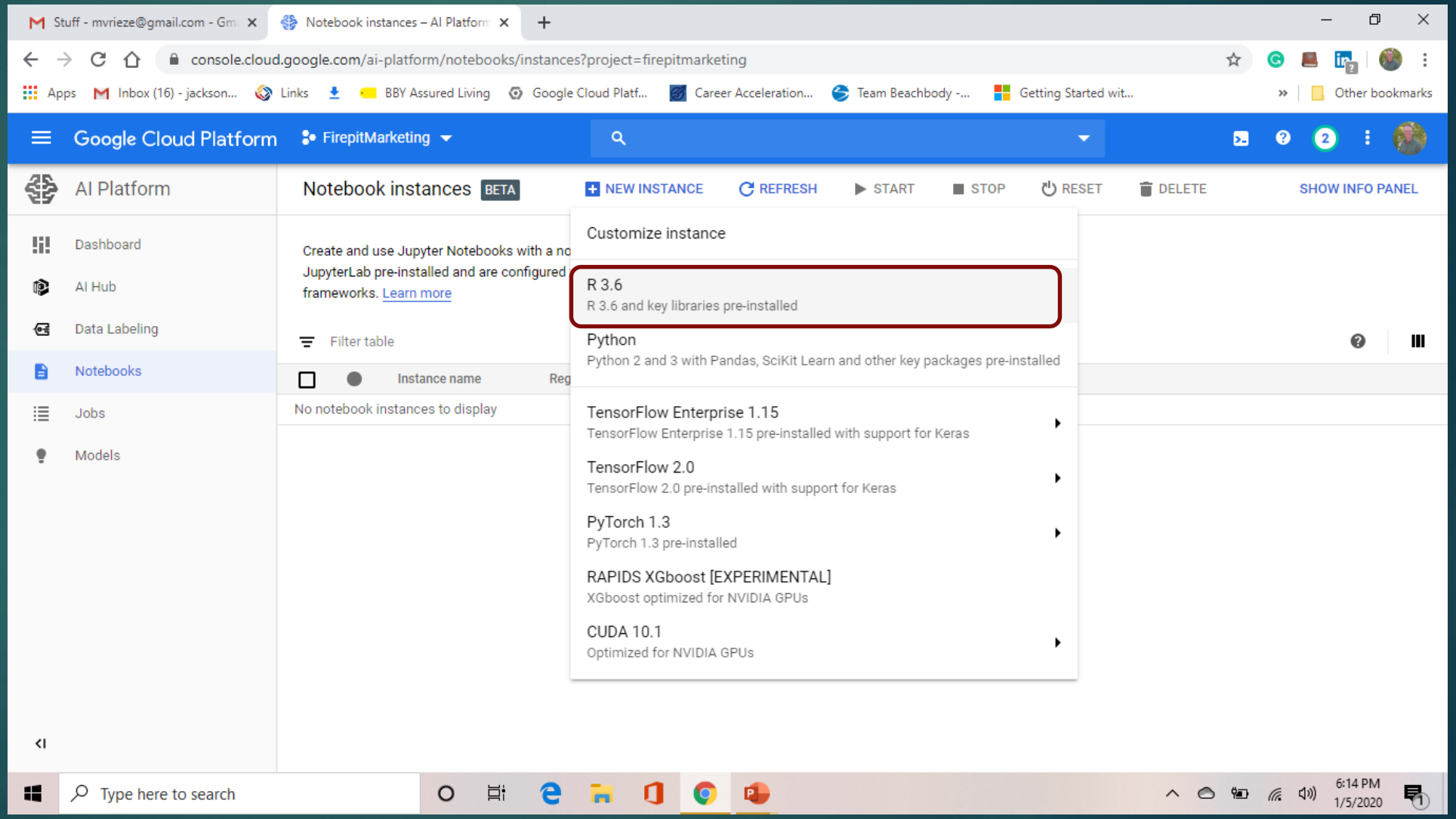
Instance name	Region	Environment	Machine type	GPUs	Permission	Labels
No notebook instances to display						

Type here to search | 6:14 PM 1/5/2020

© Copyright 2020 - Martin Vrieze

Virtual Machine Setup

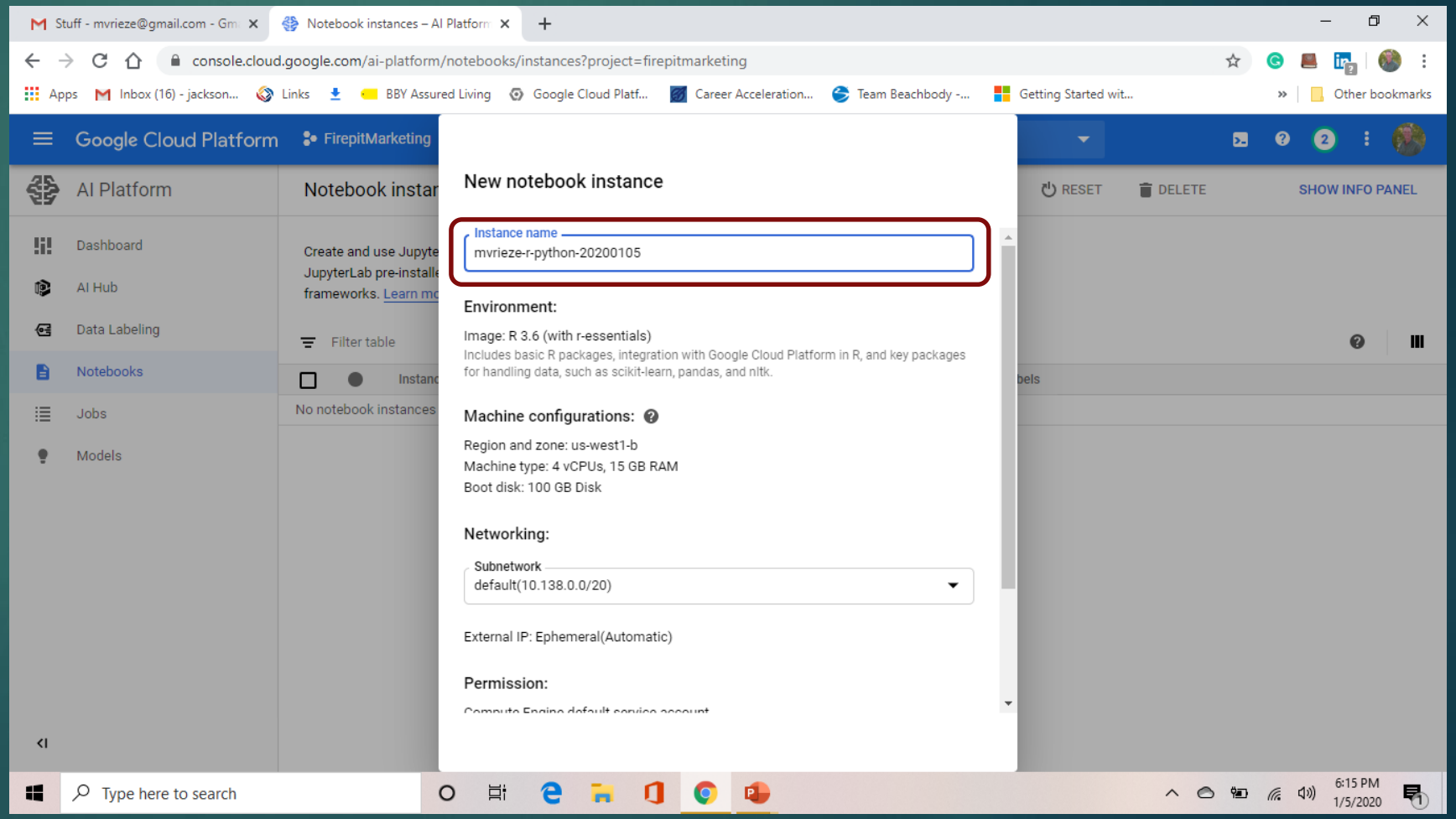
- Select "R 3.x" (If you wanted Python, don't worry – Python comes along for the ride)



© Copyright 2020 - Martin Vrieze

Virtual Machine Setup

- Name your instance
- Choose your configuration
  - Rule of thumb: You need 3x-5x the size of your data in memory to handle your project.



© Copyright 2020 - Martin Vrieze

Virtual Machine Setup

- Select "Open Jupyterlab"

Stuff - mvrieze@gmail.com - Gm
Notebook instances – AI Platform
console.cloud.google.com/ai-platform/notebooks/instances?project=firepitmarketing

Apps | Inbox (16) - jackson... | Links | BBY Assured Living | Google Cloud Platf... | Career Acceleration... | Team Beachbody -... | Getting Started wit... | Other bookmarks

Google Cloud Platform | FirepitMarketing

AI Platform

- Dashboard
- AI Hub
- Data Labeling
- Notebooks
- Jobs
- Models

Notebook instances BETA | NEW INSTANCE | REFRESH | START | STOP | RESET | DELETE | SHOW INFO PANEL

Filter table

Instance name		Region	Environment	Machine type	GPUs	Permission	Labels
mvrieze-r-python-20200105	OPEN JUPYTERLAB	us-west1-b	R:3.6	4 vCPUs, 15 GB RAM	None	Service account	No

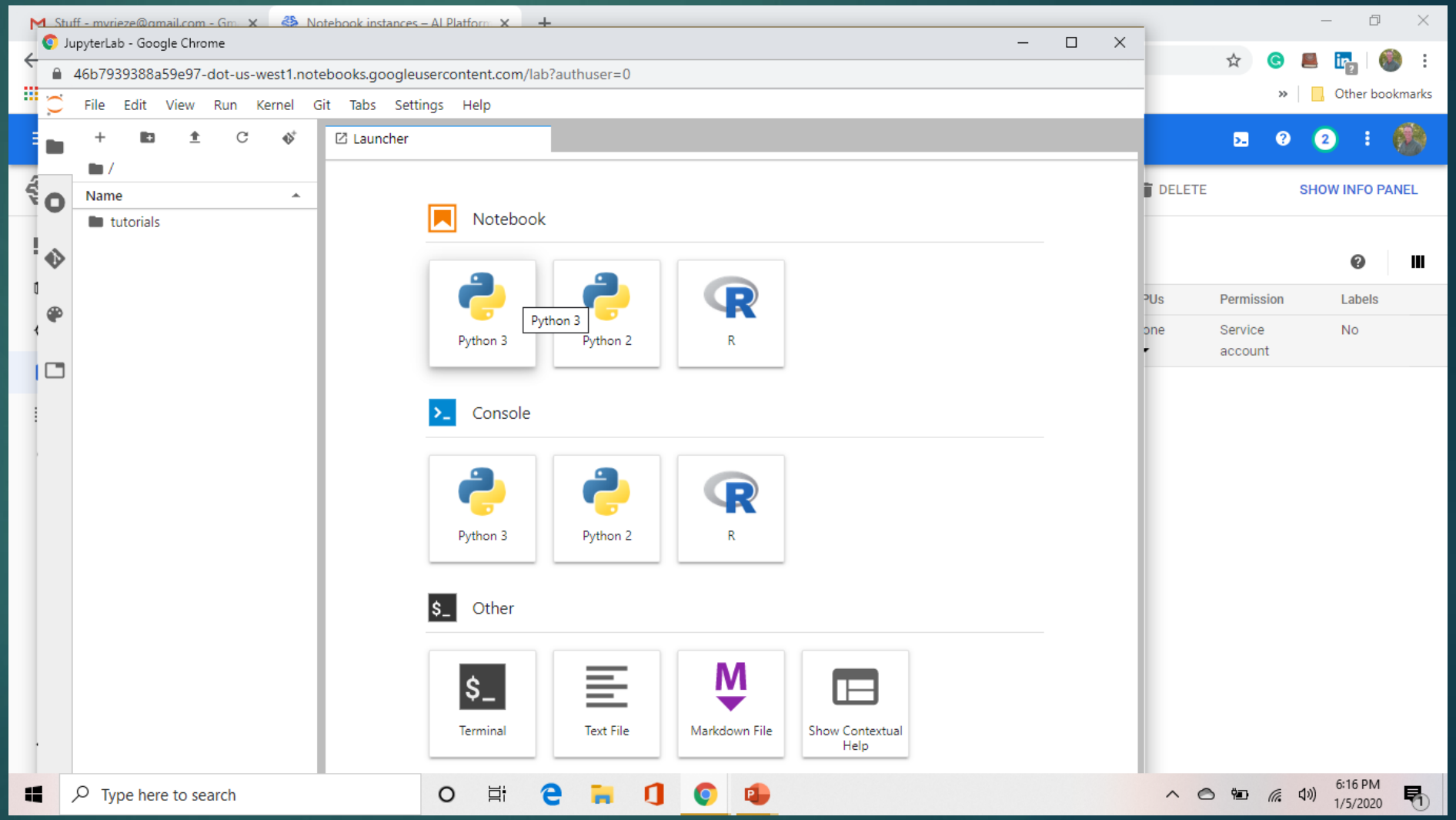
Type here to search

6:16 PM
1/5/2020

© Copyright 2020 - Martin Vrieze

Virtual Machine Setup

- Select whichever tool you wish to use (R or Python – notebook or console)



© Copyright 2020 - Martin Vrieze

Connect Python to GCP Storage Bucket

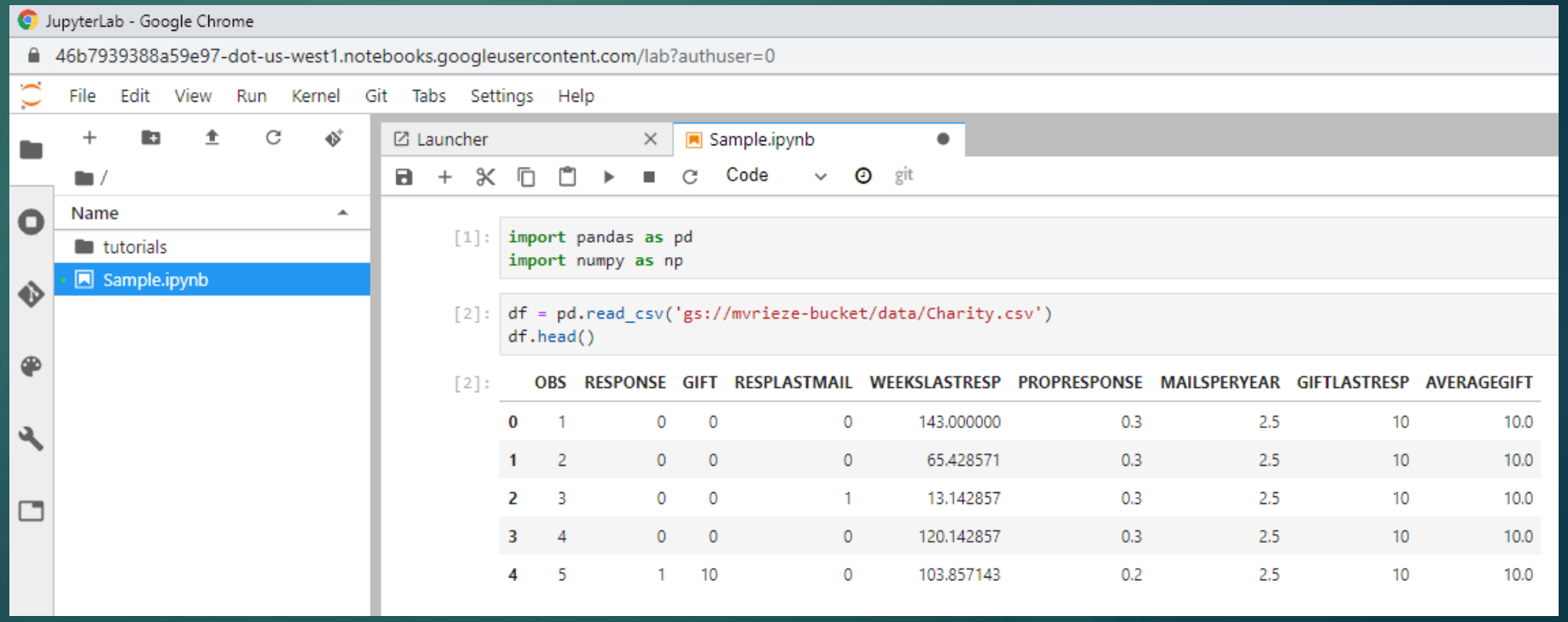
© Copyright 2020 - Martin Vrieze

Connect Python to GCP Storage Bucket

https://stackoverflow.com/questions/49357352/read-csv-from-google-cloud-storage-to-pandas-dataframe

Code :

```
# Note how easy it is to connect to a csv file in This is really easy
import pandas as pd
import numpy as np
df = pd.read_csv('gs://bucket-name/folder-name/file-name.csv')
```



	OBS	RESPONSE	GIFT	RESPLASTMAIL	WEEKSLASTRESP	PROPRESPONSE	MAILSPERYEAR	GIFTLASTRESP	AVERAGEGIFT
0	1	0	0	0	143.000000	0.3	2.5	10	10.0
1	2	0	0	0	65.428571	0.3	2.5	10	10.0
2	3	0	0	1	13.142857	0.3	2.5	10	10.0
3	4	0	0	0	120.142857	0.3	2.5	10	10.0
4	5	1	10	0	103.857143	0.2	2.5	10	10.0

© Copyright 2020 - Martin Vrieze

Connect Python to Google BigQuery

© Copyright 2020 - Martin Vrieze

Connect Python to Google BigQuery

https://pypi.org/project/google-cloud-bigquery/
https://cloud.google.com/bigquery/docs/datalab-migration

```
Code:
import numpy as np
import pandas as pd

# Import the library
from google.cloud import bigquery

# Instantiate the client
client = bigquery.Client()

#Write the query - note the ticks surrounding project.dataset.table
sql = """
SELECT * from `firepitmarketing.mvrieze_dataset.Charity`
"""

#Execute the query and set results to Pandas dataframe
df1 = client.query(sql).to_dataframe()
print(df1.head())
```


© Copyright 2020 - Martin Vrieze

Connect Python to Google BigQuery

JupyterLab - Google Chrome

46b7939388a59e97-dot-us-west1.notebooks.googleusercontent.com/lab?authuser=0

File Edit View Run Kernel Git Tabs Settings Help

Name

tutorials

Sample.ipynb

Launcher | Sample.ipynb

Code | git | Python 3

[6]:
```python
import numpy as np
import pandas as pd

# Import the library
from google.cloud import bigquery

# Instantiate the client
client = bigquery.Client()

#Write the query - note the ticks surrounding project.dataset.table
sql = """
SELECT * from `firepitmarketing.mvrieze_dataset.Charity`
"""

#Execute the query and set results to Pandas dataframe
df1 = client.query(sql).to_dataframe()
print(df1.head())
```

```
   OBS  RESPONSE  GIFT  RESPLASTMAIL  WEEKSLASTRESP  PROPRESPONSE  \
0  809         0     0             0          158.0         0.125
1  810         0     0             0          158.0         0.125
2  815         0     0             0          158.0         0.125
3  820         0     0             0          143.0         0.125
4  893         0     0             0          158.0         0.125

   MAILSPERYEAR  GIFTLASTRESP  AVERAGEGIFT
0           2.0            50         50.0
1           2.0            50         50.0
2           2.0            50         50.0
3           2.0            50         50.0
4           2.0            71         71.0
```

0 1 Python 3 | Idle Mode: Command Ln 1, Col 1 Sample.ipynb

Type here to search 2:44 PM 1/19/2020

© Copyright 2020 - Martin Vrieze

Connect R to GCP Storage Bucket

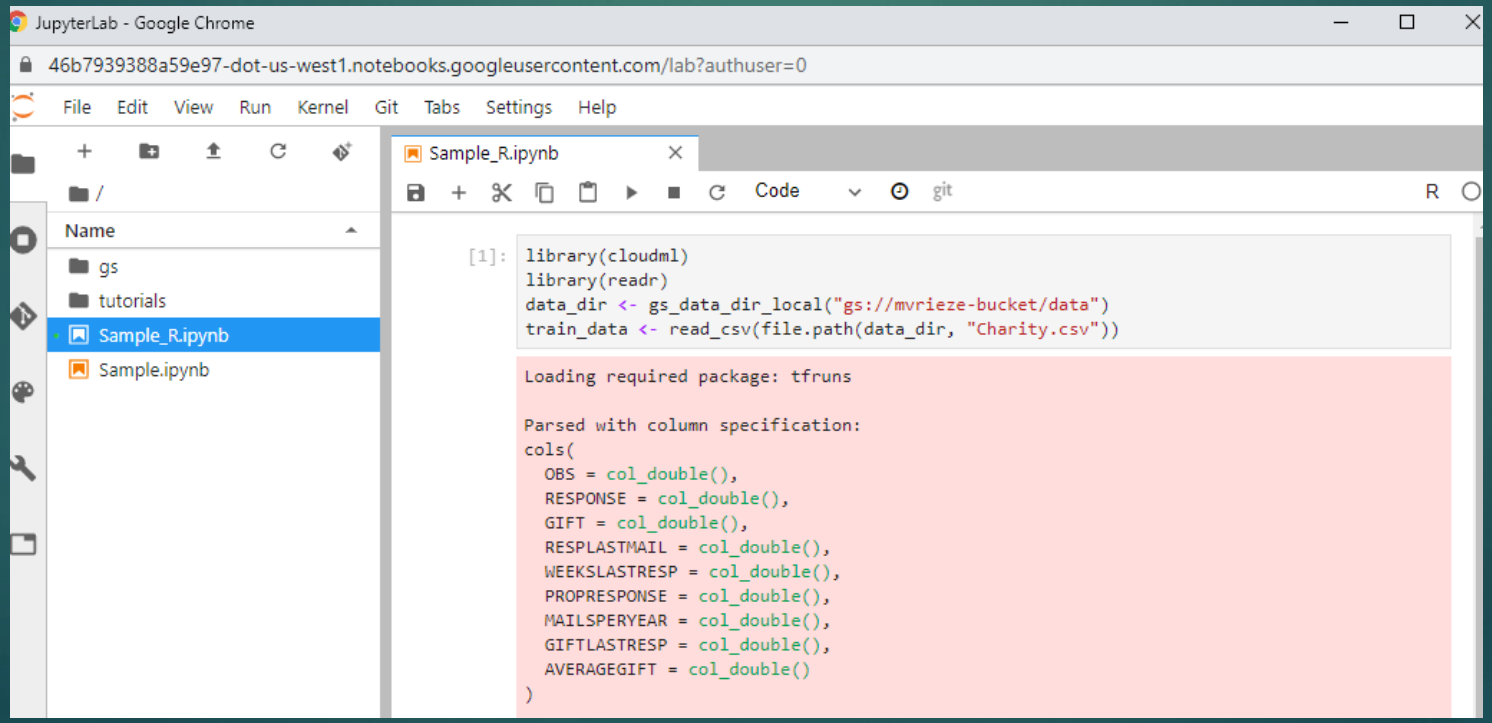
© Copyright 2020 - Martin Vrieze

Connect R to GCP Storage Bucket

https://cran.r-project.org/web/packages/cloudml/vignettes/storage.html

Code:

```
library(cloudml)
library(readr)
data_dir <- gs_data_dir_local("gs://bucket/path_in_bucket")
train_data <- read_csv(file.path(data_dir, "filename.csv"))
```



© Copyright 2020 - Martin Vrieze

Connect R to Google BigQuery

© Copyright 2020 - Martin Vrieze

Connect R to Google BigQuery

https://cloud.google.com/ai-platform/notebooks/docs/use-r-bigquery
https://bigrquery.r-dbi.org/

Code:

```
# Load the package
library(bigrquery)

# Add another notebook block and paste in the following block
# Provide authentication
bq_auth(use_oob = TRUE)

# Add another notebook block
# Store the project id
projectid = "firepitmarketing"

# Set your query - note the ticks around the <project name>.<data set name>.<table name>
sql <- "SELECT * from `firepitmarketing.mvrieze_dataset.Charity`"

# Run the query and store the data in a dataframe
df <- query_exec(sql, projectid, use_legacy_sql = FALSE)

# Print the query result
df
```


© Copyright 2020 - Martin Vrieze

Connect R to Google BigQuery

JupyterLab - Google Chrome

46b7939388a59e97-dot-us-west1.notebooks.googleusercontent.com/lab?authuser=0

File Edit View Run Kernel Git Tabs Settings Help

Name

- gs
- tutorials
- Sample_R.ipynb
- Sample.ipynb

Sample_R.ipynb

Code git R

[2]:
```
# Load the package
library(bigrquery)

# Add another notebook block and paste in the following block
# Provide authentication
bq_auth(use_oob = TRUE)

# Add another notebook block
# Store the project id
projectid = "firepitmarketing"

# Set your query - note the ticks around the <project name>.<data set name>.<table name>
sql <- "SELECT * from `firepitmarketing.mvrieze_dataset.Charity`"

# Run the query and store the data in a dataframe
df <- query_exec(sql, projectid, use_legacy_sql = FALSE)

# Print the query result
df
```

```
10.0 megabytes processed
```

A data.frame: 4268 × 9

OBS	RESPONSE	GIFT	RESPLASTMAIL	WEEKSLASTRESP	PROPRESPONSE	MAILSPERYEAR	GIFTLASTRESP	AVERAGEGIFT
<int>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>
809	0	0	0	158.00000	0.125	2	50	50
810	0	0	0	158.00000	0.125	2	50	50
815	0	0	0	158.00000	0.125	2	50	50

0 1 R | Idle Mode: Command Ln 1, Col 1 Sample_R.ipynb

Type here to search 3:42 PM 1/19/2020

© Copyright 2020 - Martin Vrieze

Next Steps

1. Go to: https://github.com/mvrieze/gcp_econometrics
2. Pull down Sample_Python.ipynb notebook and import into your GCP notebook environment.
3. Go ahead and run the code blocks:
   - GCP Storage Bucket: Change the <bucket name>.<folder name>.<file name>.csv pointers in the code that pulls in the data.
   - BigQuery: Change the <project>.<dataset name>.<table name> pointers in the code that pulls in the data.

© Copyright 2020 - Martin Vrieze