

**BDAT 1008 – DATA COLLECTION AND CURATION**

**FINAL PROJECT**

**TECHNICAL DESIGN DOCUMENT**

***Group Number: 10***

***Team Members:***

**Abdul-Baatinul-Haad Yusuf (200497977)**

**Athul Verghese**

**Milind Satishkumar Tank**

**Deepak Thalian**

**Table of Contents**

[**Programming Languages Used** 2](#_Toc111411452)

[**Infrastructure of System / Tech Stack used** 3](#_Toc111411453)

[**Dataset Structure** 4](#_Toc111411454)

[**Screenshot of Environment Set-up and Data Manipulation in Google Colab** 5](#_Toc111411455)

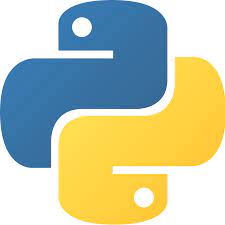
[**Screenshot of Dashboard in Power BI** 8](#_Toc111411456)

[**Screenshots of Code to Create Web App Using Dash by Plotly** 9](#_Toc111411457)

[**Screenshots of Web App** 11](#_Toc111411458)

# **Programming Languages Used**

This project was undertaken using the pyspark environment and python language with its components.



# **Infrastructure of System / Tech Stack used**

For environment setup, we have used Google Colab’s notebook, with spark environment. The component of spark used is the graph analysis. We have also used python’s pandas and plotly components to manipulate and visualize our dataset easily.

We have also used mongo DB to create a database, insert our dataset into the database with Google cloud services, and extracted data directly from the database to manipulate.

Then for the creation of the web app, we used visual studio code, python, and dash by plotly.

Dash by plotly is essentially the react for python. It can be used to create frontend dashboards with HTML components integrated in it.

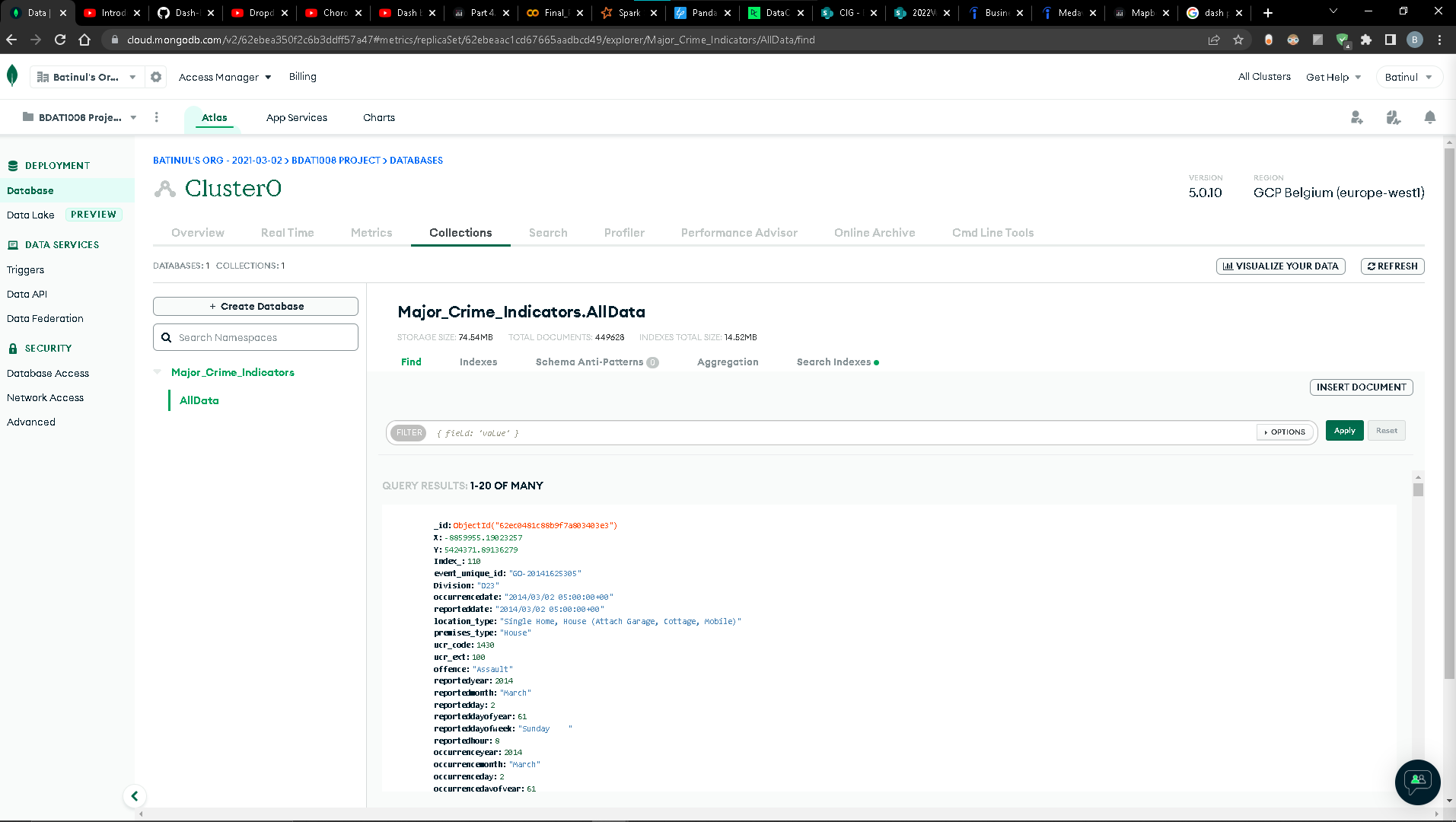
We have also created a dashboard with power BI, to showcase our familiarity with a wide range of tools. 



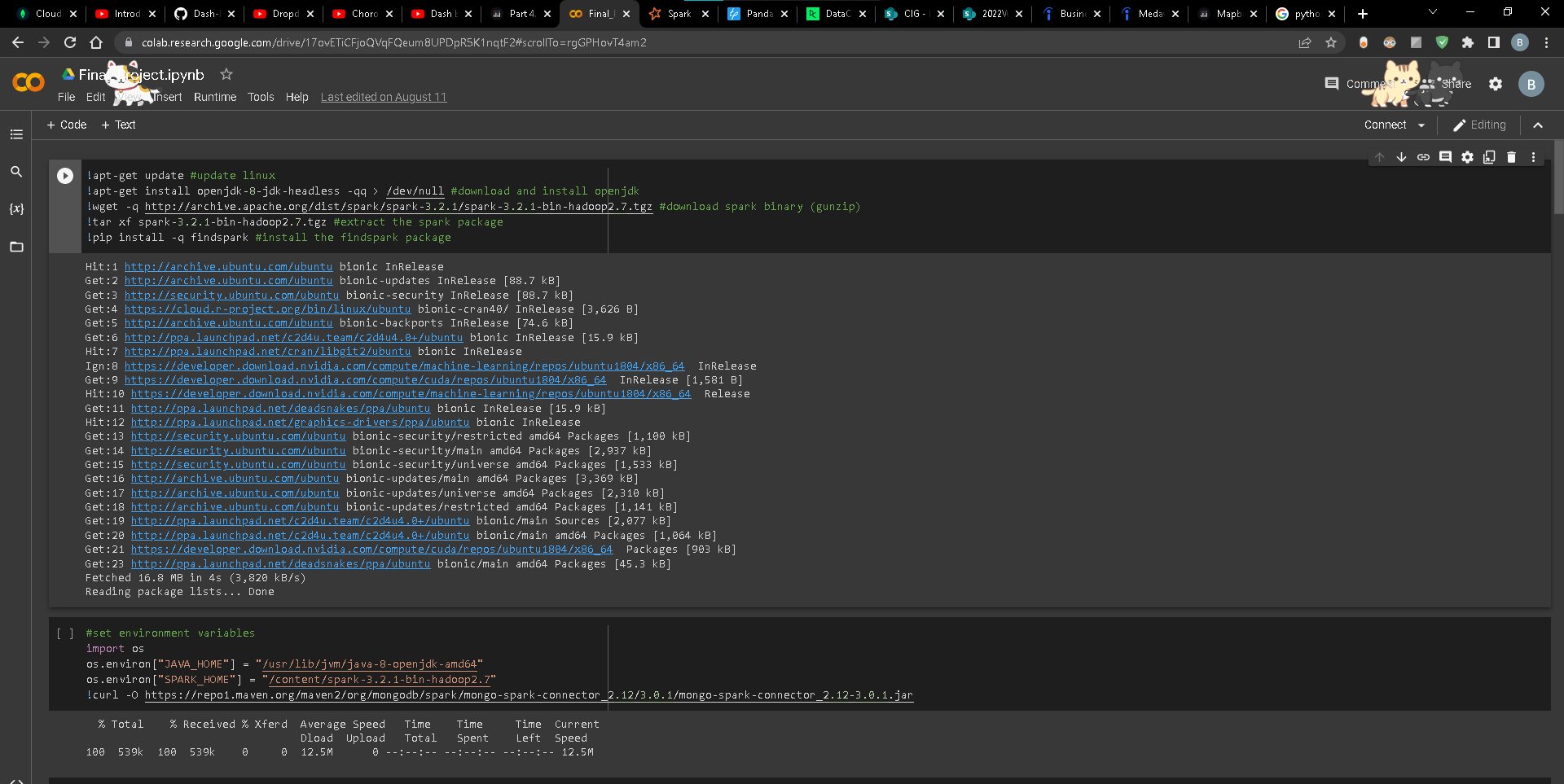


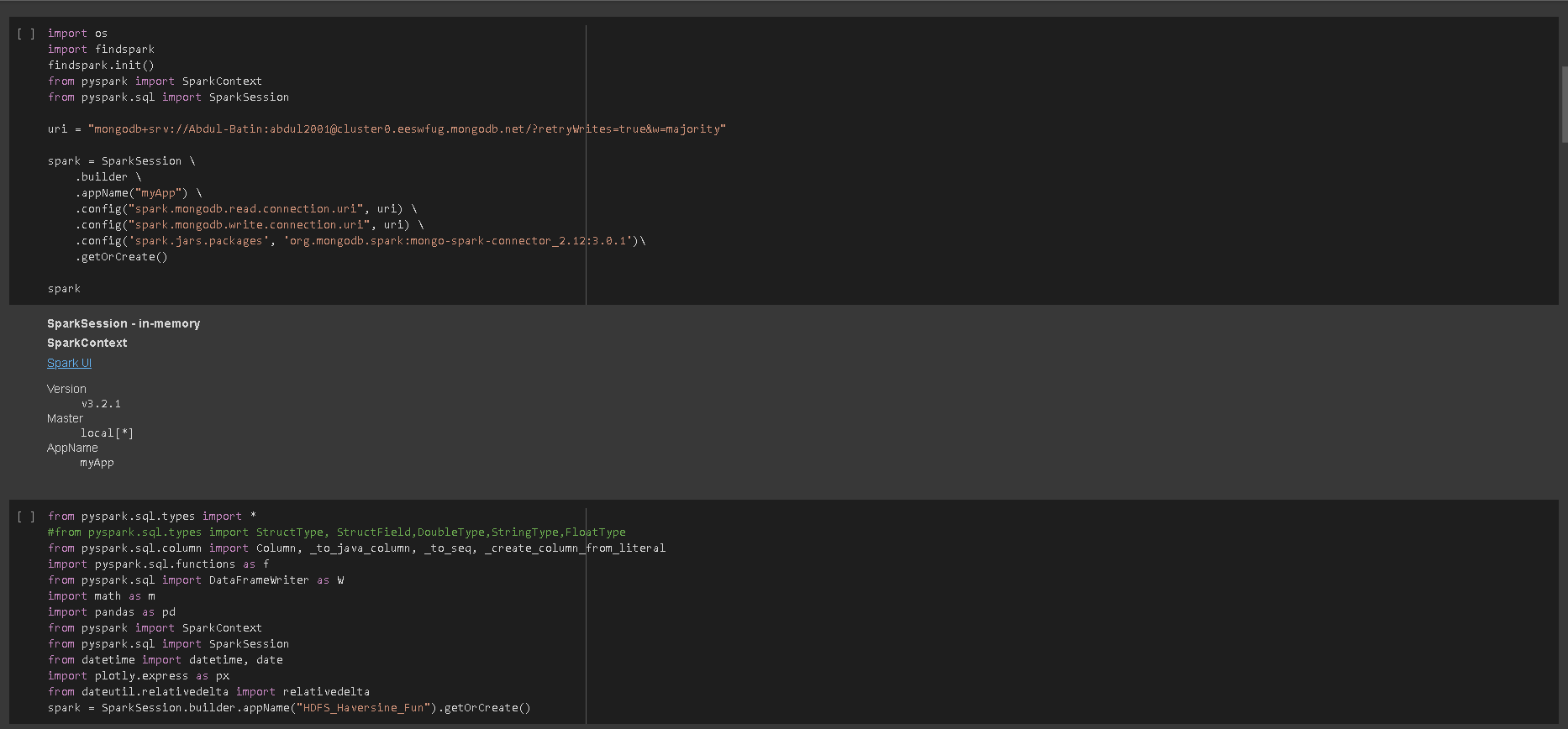
# **Dataset Structure**

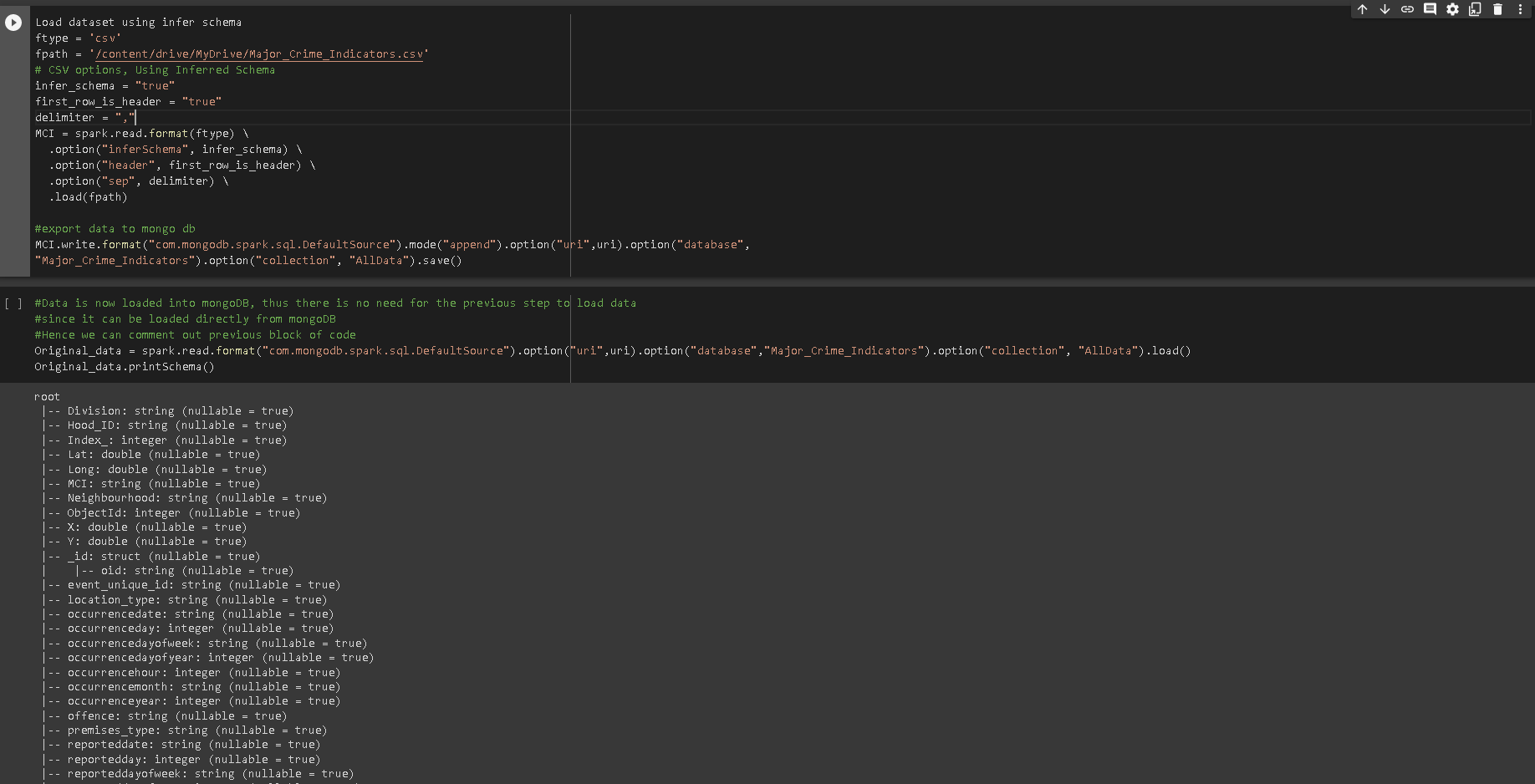
Below is a screenshot of the database in mongo DB. The compressed storage size of the dataset is 74.54 MB, whereas its uncompressed size is 316.27 MB. The dataset has 281,692 rows with 31 columns. This is considered a relatively large amount of data.

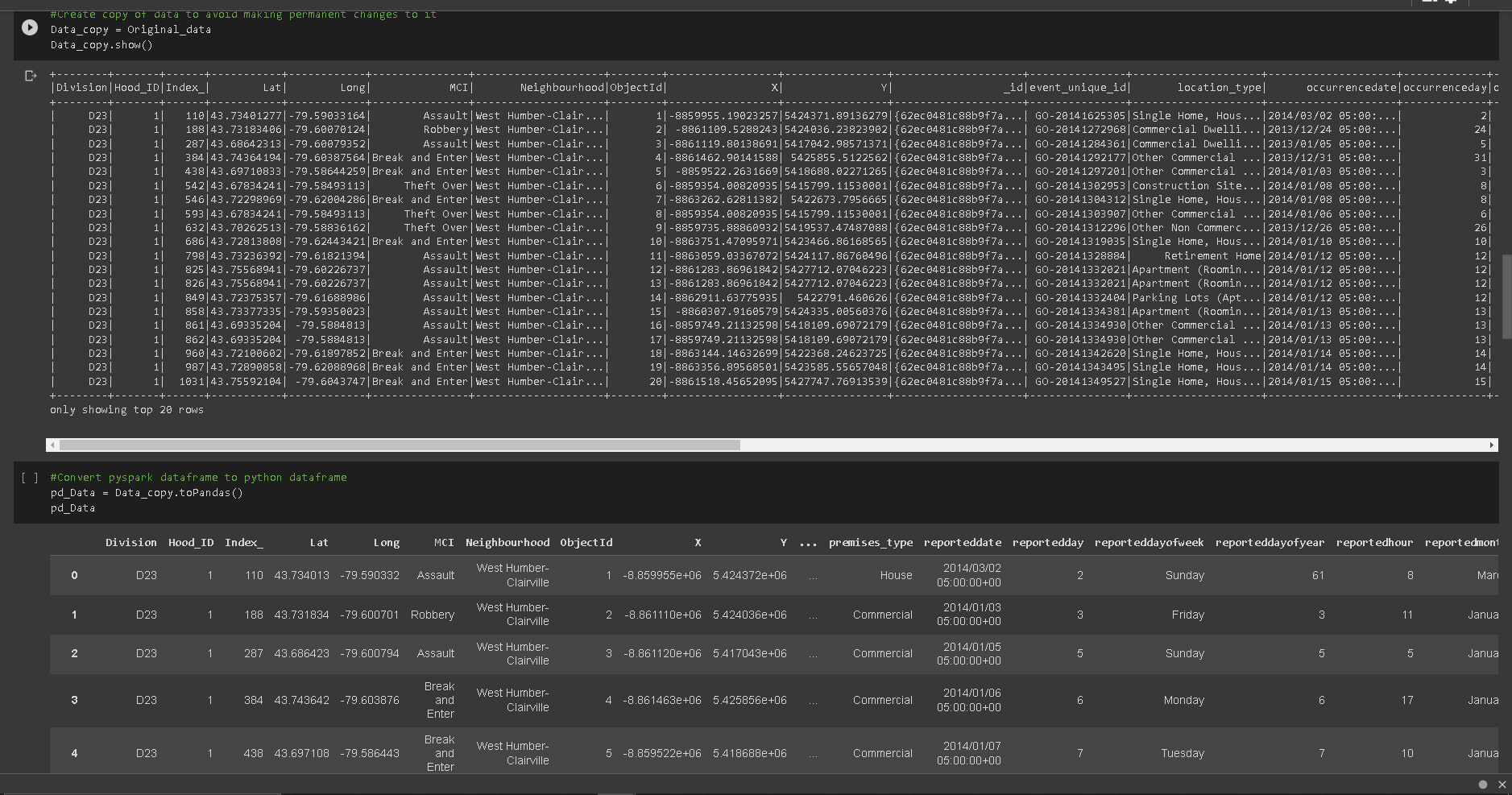


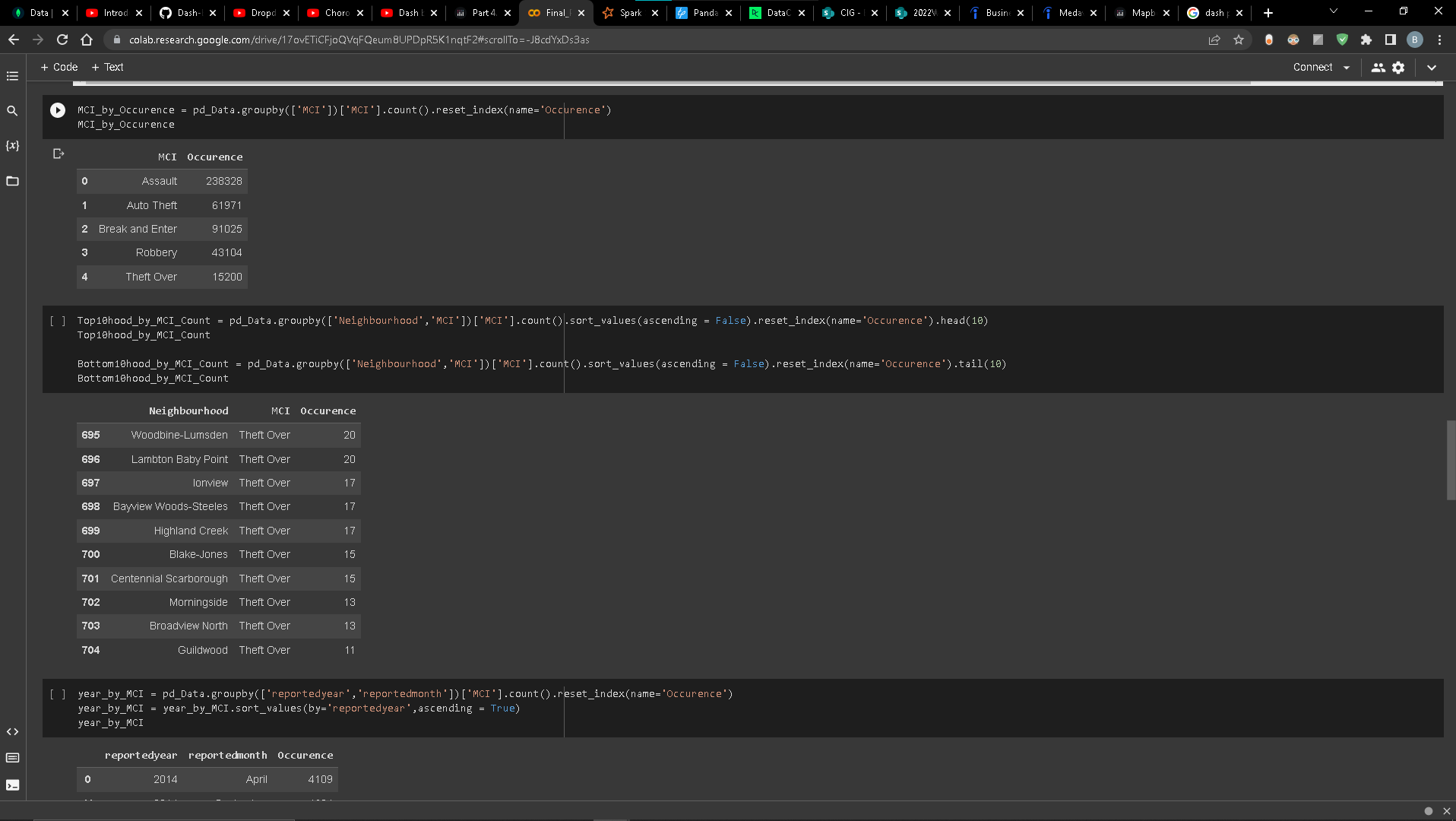
# **Screenshot of Environment Set-up and Data Manipulation in Google Colab**

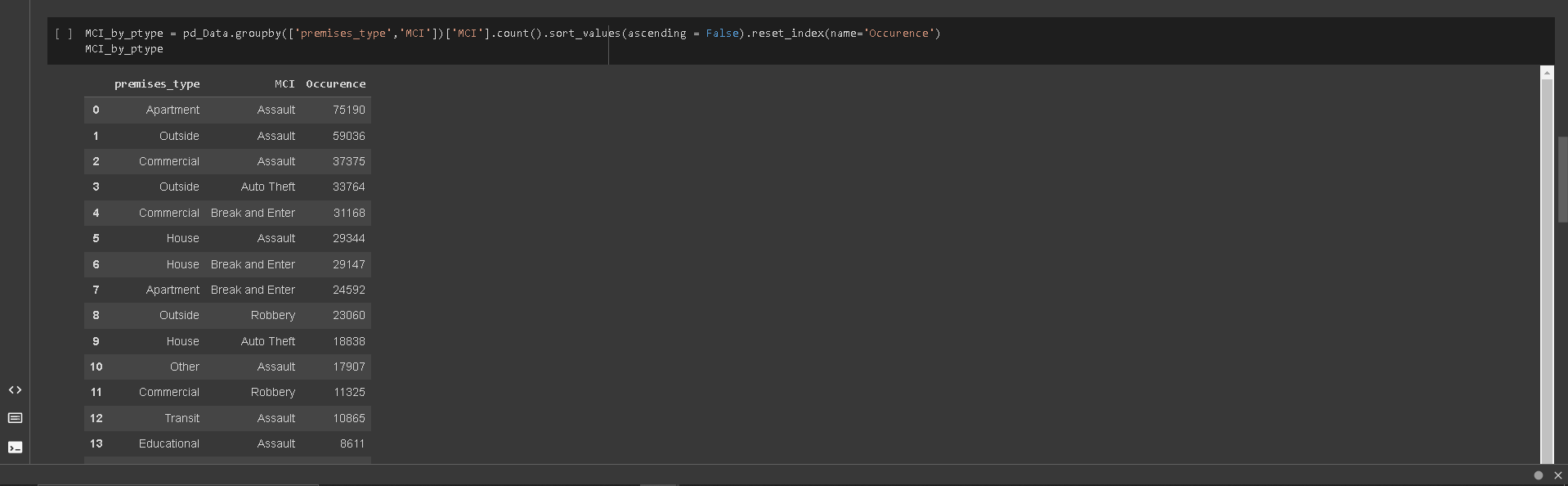


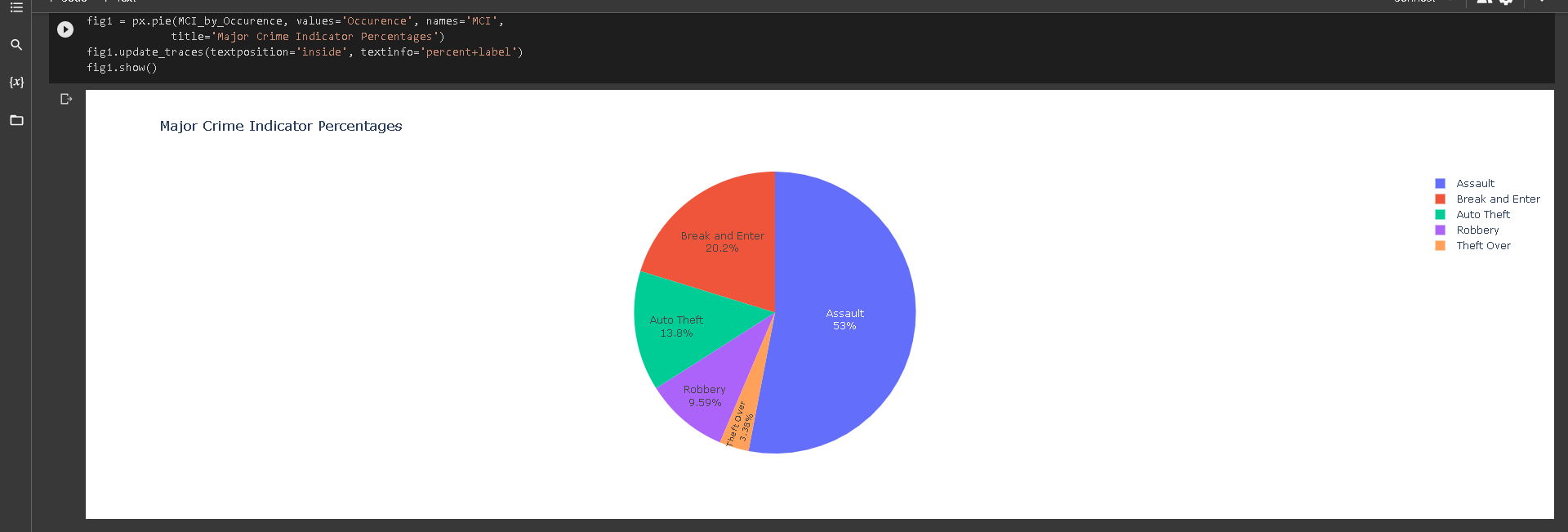


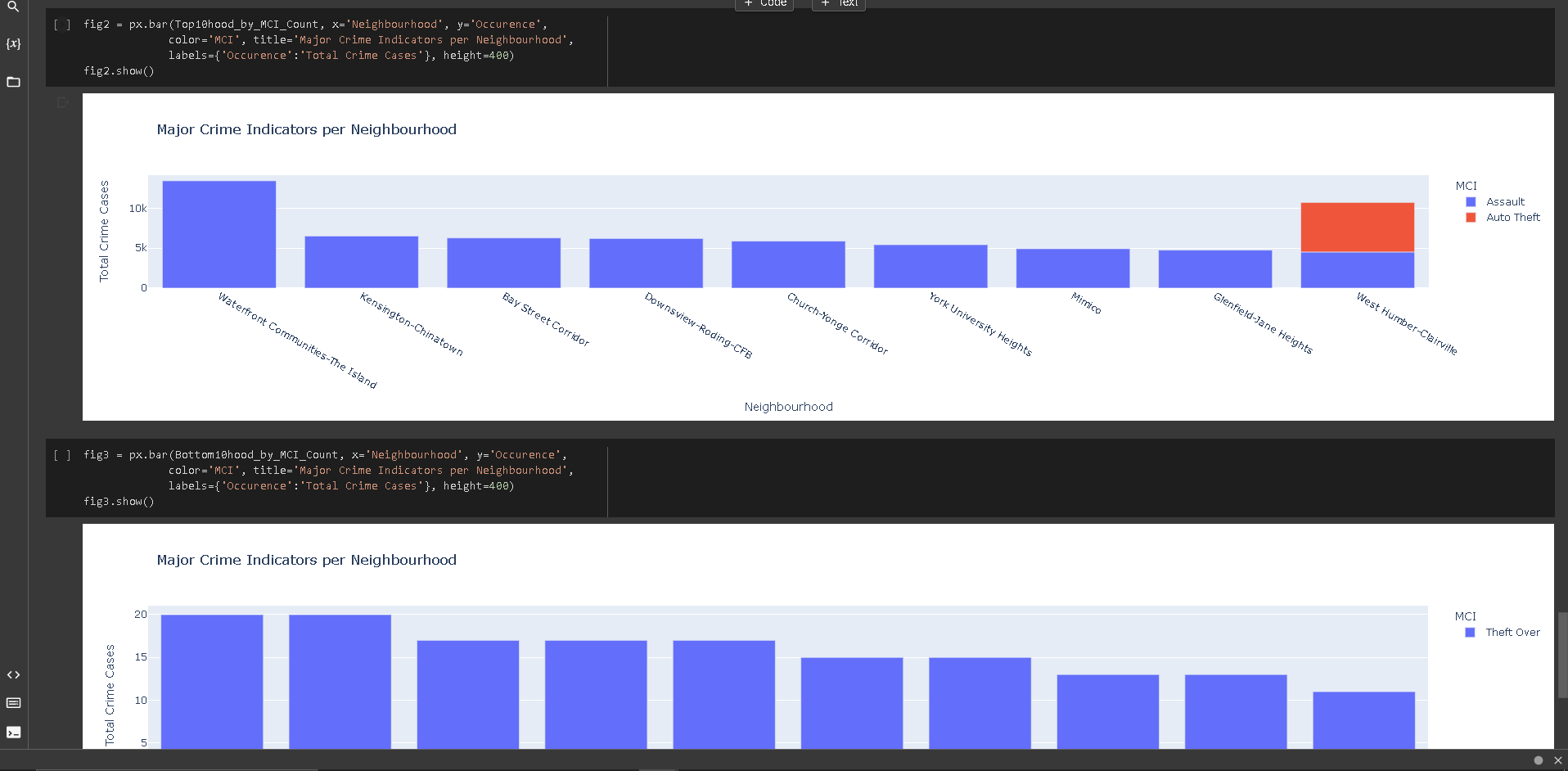




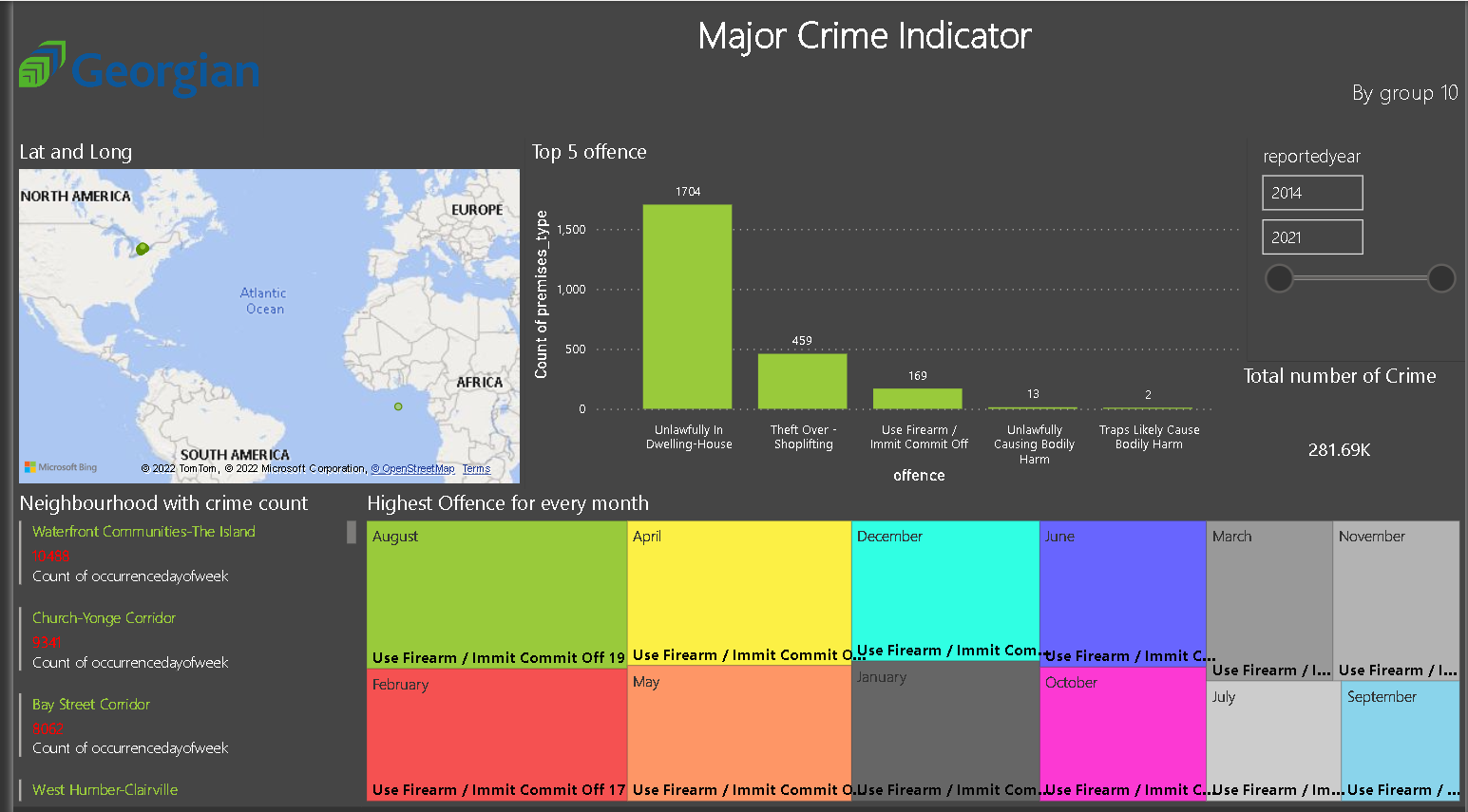


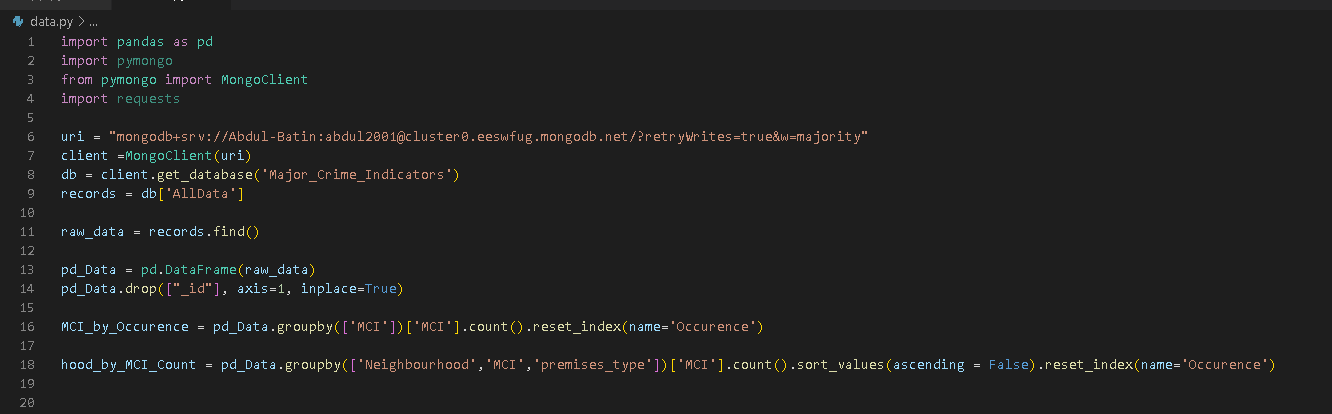






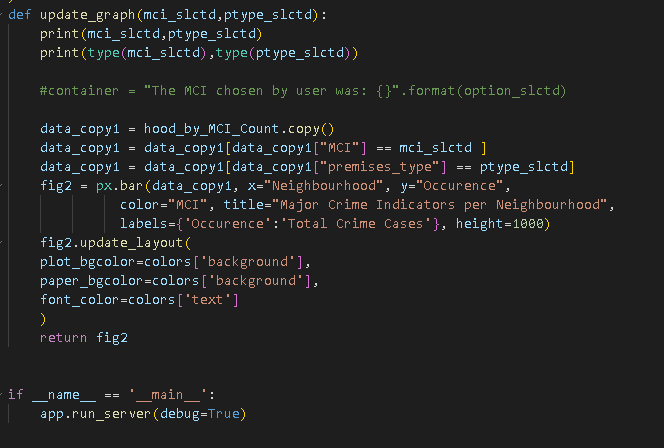
# **Screenshot of Dashboard in Power BI**



**Screenshots of Code to Create Web App Using Dash by Plotly**







# **Screenshots of Web App**