CRIME RATE PREDICTOR

CLOUD COMPUTING

AISHVWARYA IYER
STUDENT ID: 16227781

GAYATHREE IYER
STUDENT ID: 16227784

1. INTRODUCTION

1.1 MOTIVATION

According to The New York Times, America has definitely more violence when compared to other rich countries. Murder rates are also much higher in the United States than in countries like Japan, Europe, or Canada. We also have more assaults, and robberies than other rich countries. When compared to other affluent nations, crime rates have always been much higher in America. Hence predicting and analyzing crime data is extremely crucial.

Crime has also been the hot topic for the 'Smart City' conference.

1.2 GOAL

Crime Rate Predictor is a Cloud Computing project that focusses on applying machine learning algorithm to build models that will not only predict crime rates in various states but also predict the police funding and educational attainment required in the future to mitigate crime.

2. TEAM MEMBER'S CONTRIBUTION AREA

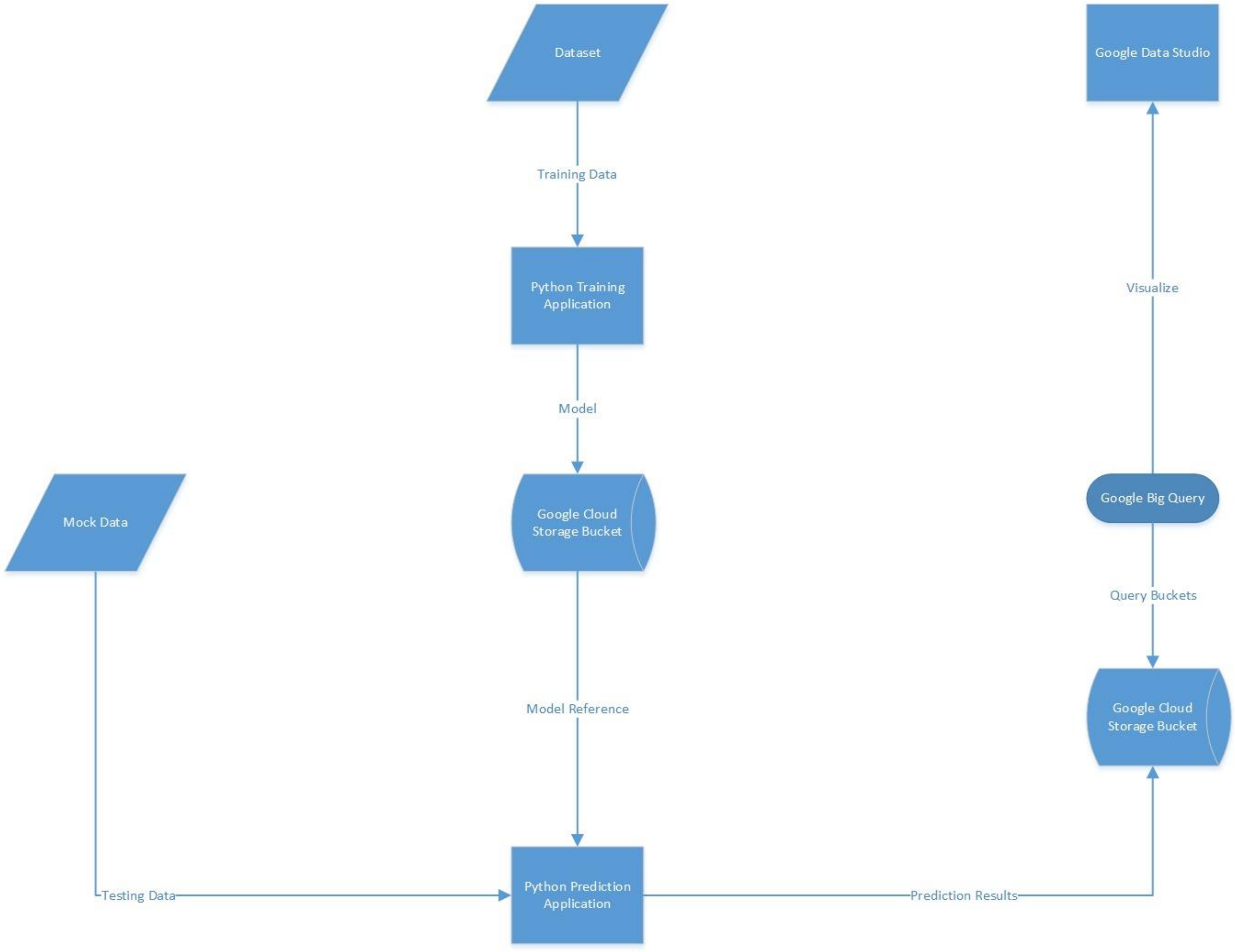
***Aishvwarya Iyer*:**

- Google Cloud Platform Infrastructure setup
- Python Training Program
- Google Big Query

Gayathree Iyer:

- Python Prediction Program
- PowerShell Automation Scripts
- Google Data Studio

3. SYSTEM ARCHITECTURE



System Architecture

4. MACHINE LEARNING ALGORITHM

We use **Multi variate linear regression model** for the prediction of crime. Multiple correlated dependent variables are predicted using this model. Regression analysis enables to describe the relationship between response variable and one or more predictor variables.

$$Yi=\alpha+\beta_1x_i+\beta_2x_i+....+\beta_nx_i$$

Y_i is the estimate of ith component of dependent variable y, where we have **n** independent variables and x^j_i denotes the ith component of the jth independent variable/feature.

5. DATA SET DESCRIPTION

The data set contains a case study of education, crime, and police funding for small cities in ten southeastern and eastern states. The states are New York, South Carolina, Florida, Connecticut, Georgia, Rhode Island, New Hampshire, Maine, North Carolina and Virginia.

The data collected are for a sample of 50 small cities in these states.

X1 = Total Overall Reported Crime Rate per 1 Million Residents

X2 = Reported Violent Crime Rate per 100,000 Residents

X3 = Annual Police Funding (Dollars) per Resident

X4 = Percent of People 25 Years and Older That Have Had 4 years of High School

X5 = Percent of 16- to 19-Year-Olds Not in High School and Not High School Graduates

X6 = Percent of 18- to 24-Year-Olds Enrolled in College

X7 = Percent of People 25 Years and Older with at Least 4 Years of College

X8 = States

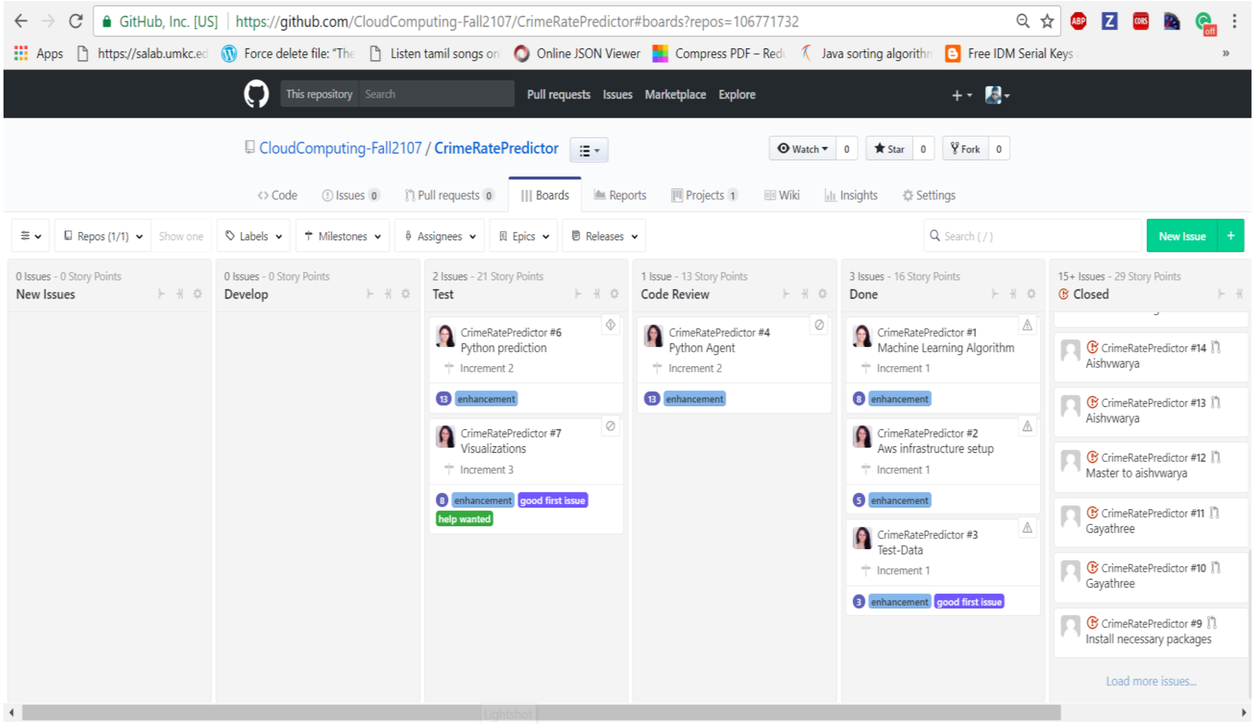
6. TECHNOLOGIES

- Programming language: Python.
- Google Cloud infrastructure is built using terraform scripts.
- Automation is done using PowerShell scripts.
- Source control management: Github

Github link: https://github.com/CloudComputing-Fall2107/CrimeRatePredictor

7. PROJECT MANAGEMENT

The work progress is tracked using Zenhub, which is an agile project management for Github.



0 Issues - 0 Story Points
New Issues

0 Issues - 0 Story Points
Develop

2 Issues - 21 Story Points
Test

- CrimeRatePredictor #6 Python prediction, Increment 2, 13 enhancement
- CrimeRatePredictor #7 Visualizations, Increment 3, 8 enhancement, good first issue, help wanted

1 Issue - 13 Story Points
Code Review

- CrimeRatePredictor #4 Python Agent, Increment 2, 13 enhancement

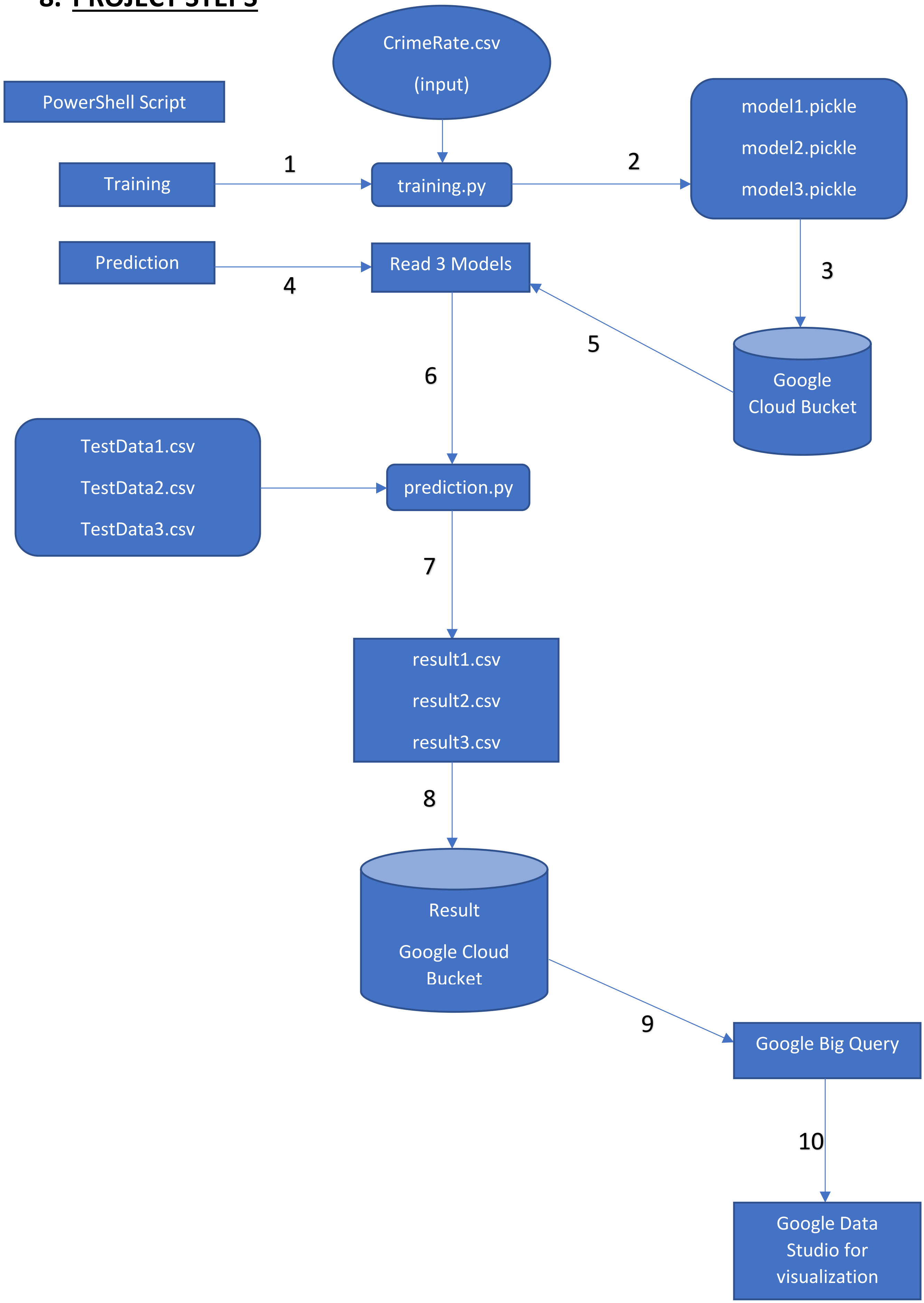
3 Issues - 16 Story Points
Done

- CrimeRatePredictor #1 Machine Learning Algorithm, Increment 1, 8 enhancement
- CrimeRatePredictor #2 Aws infrastructure setup, Increment 1, 5 enhancement
- CrimeRatePredictor #3 Test-Data, Increment 1, 3 enhancement, good first issue

15+ Issues - 29 Story Points
Closed

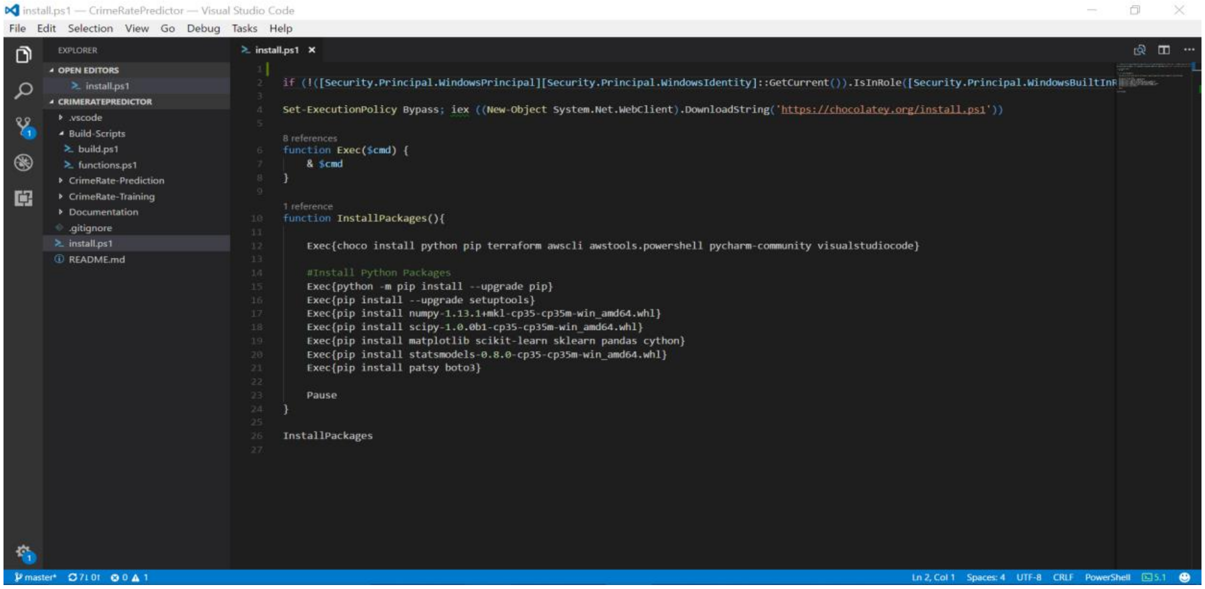
- CrimeRatePredictor #14 Aishvwarya
- CrimeRatePredictor #13 Aishvwarya
- CrimeRatePredictor #12 Master to aishvwarya
- CrimeRatePredictor #11 Gayathree
- CrimeRatePredictor #10 Gayathree
- CrimeRatePredictor #9 Install necessary packages

8. PROJECT STEPS



8.1 INSTALLATION OF PACKAGES

Installation of necessary packages for the project using Choclatey, the package manager for windows.

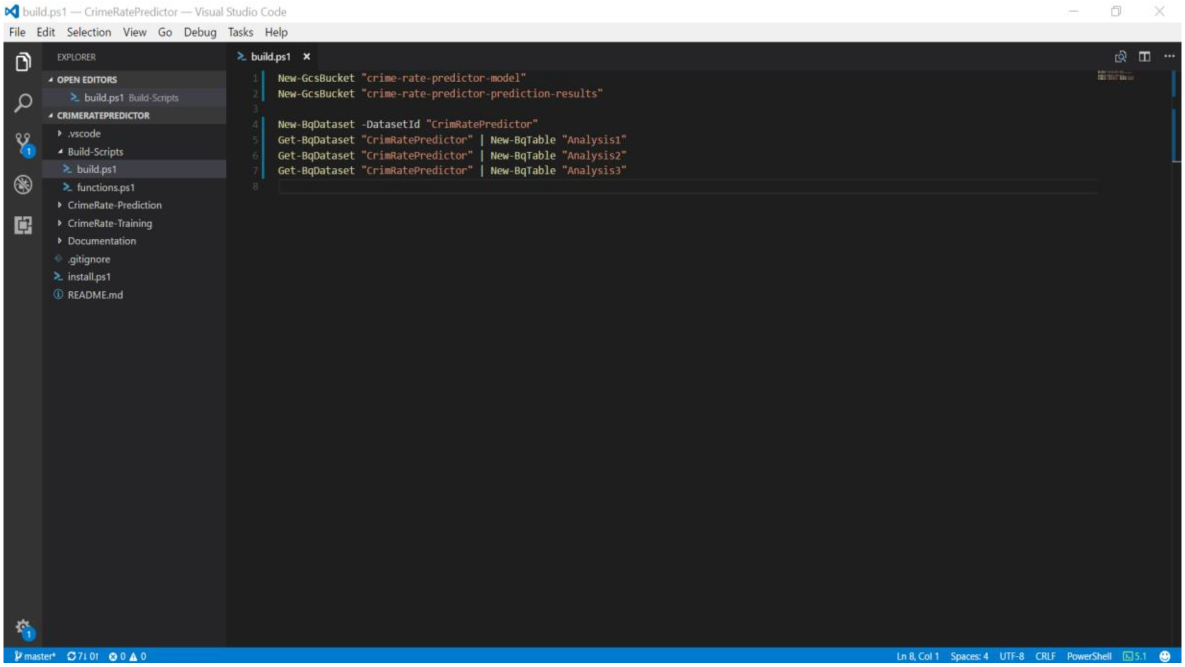


```
if (!([Security.Principal.WindowsPrincipal][Security.Principal.WindowsIdentity]::GetCurrent()).IsInRole([Security.Principal.WindowsBuiltInR

Set-ExecutionPolicy Bypass; iex ((New-Object System.Net.WebClient).DownloadString('https://chocolatey.org/install.ps1'))

8 references
function Exec($cmd) {
    & $cmd
}

1 reference
function InstallPackages(){

    Exec{choco install python pip terraform awscli awstools.powershell pycharm-community visualstudiocode}

    #Install Python Packages
    Exec{python -m pip install --upgrade pip}
    Exec{pip install --upgrade setuptools}
    Exec{pip install numpy-1.13.1+mkl-cp35-cp35m-win_amd64.whl}
    Exec{pip install scipy-1.0.0b1-cp35-cp35m-win_amd64.whl}
    Exec{pip install matplotlib scikit-learn sklearn pandas cython}
    Exec{pip install statsmodels-0.8.0-cp35-cp35m-win_amd64.whl}
    Exec{pip install patsy boto3}

    Pause
}

InstallPackages
```

Install_Packages_PowerShell

8.2 GOOGLE CLOUD PLATFORM INFRASTRUCTURE SETUP

Using Google Cloud PowerShell Tools.



```
New-GcsBucket "crime-rate-predictor-model"
New-GcsBucket "crime-rate-predictor-prediction-results"

New-BqDataset -DatasetId "CrimRatePredictor"
Get-BqDataset "CrimRatePredictor" | New-BqTable "Analysis1"
Get-BqDataset "CrimRatePredictor" | New-BqTable "Analysis2"
Get-BqDataset "CrimRatePredictor" | New-BqTable "Analysis3"
```

GCS Infratructure Setup

8.3 TRAINING PROGRAM

The prediction strategy that we execute is to first train the features from training data and create training model using Pickle python library.

- **Predicting Total Overall Reported Crime:**

  Here, we predict *the total overall reported crime rate per 1 million residents using:*

  X2 = Reported Violent Crime Rate per 100,000 Residents

  X3 = Annual Police Funding in Dollars per Resident

  X4 = Percent of People 25 Years and Older That Have Had 4 years of high school

  X5 = Percent of 16- to 19-Year-Olds Not in High School and not high school graduates

  X6 = Percent of 18- to 24-Year-Olds Enrolled in college

  X7 = Percent of People 25 Years and older with at Least 4 Years of college

  X8 = States

- **Predicting Annual police Funding in Dollars:**

  Here, we predict *the annual police funding in dollars per resident* using:

  X1 = Total Overall Reported Crime Rate per 1 Million Residents

  X2 = Reported Violent Crime Rate per 100,000 Residents

  X8 = States

  - Government can decide the annual police funding using this model efficiently to mitigate the crime rate.
  - According to FBI statistics, annual police funding affects the crime rate to a great extent. Hence, it is very crucial to help the government know the annual police funding in advance.

- **Predicting the Age Group and Educational Attainments** :

  Here, we predict the *age group and educational attainments* of the person committing the crime using:

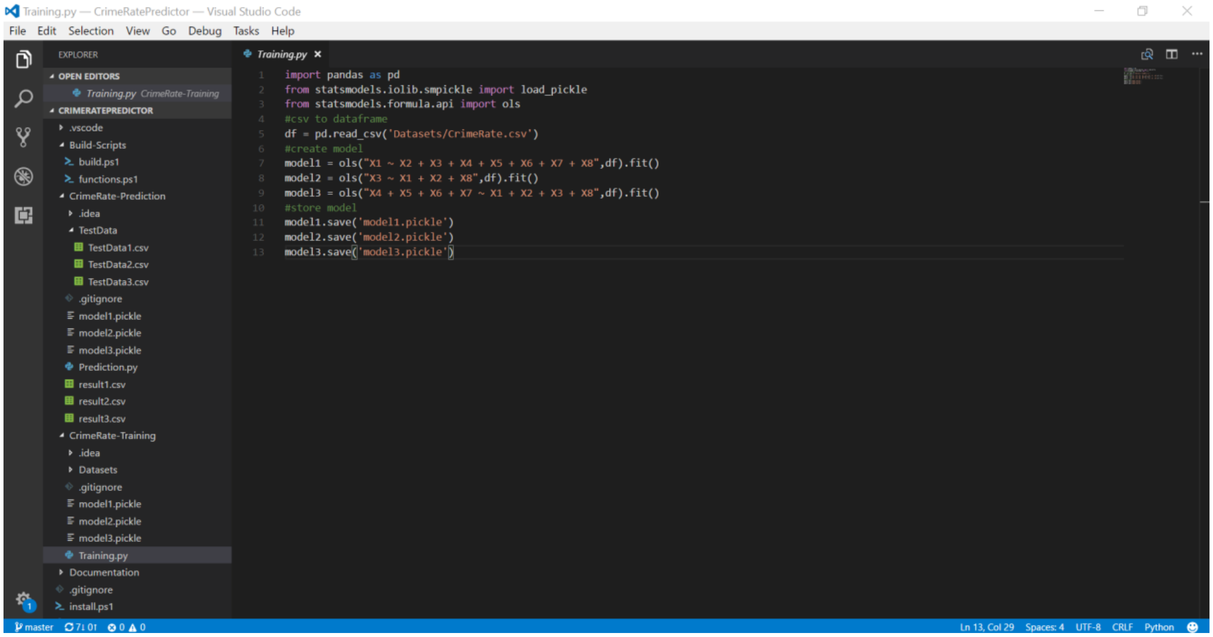
  X1 = Total Overall Reported Crime Rate per 1 Million Residents

  X2 = Reported Violent Crime Rate per 100,000 Residents

  X3 = Annual Police Funding in Dollars per Resident

  X8 = States

- According to Alliance report findings, there is an indirect correlation between educational attainment and arrest rates.
- According to the most recent data from the U.S. Bureau of Justice, 56 percent of federal inmates, 67 percent of inmates in state prisons, and 69 percent of inmates in local jails did not complete high school.
- Plus, when examining total crime savings, the report also forecasts the number of individual crimes that could be avoided by expanding the high school graduation rate by 5 percent, and finds that such an increase would decrease overall annual incidences of larceny by more than 37k; assault by nearly 60k; burglaries by more than 17k and motor vehicle theft by more than 31k. It would also avert nearly 1,300 murders, more than 1,500 robberies and more than 3,800 occurrences of rape.

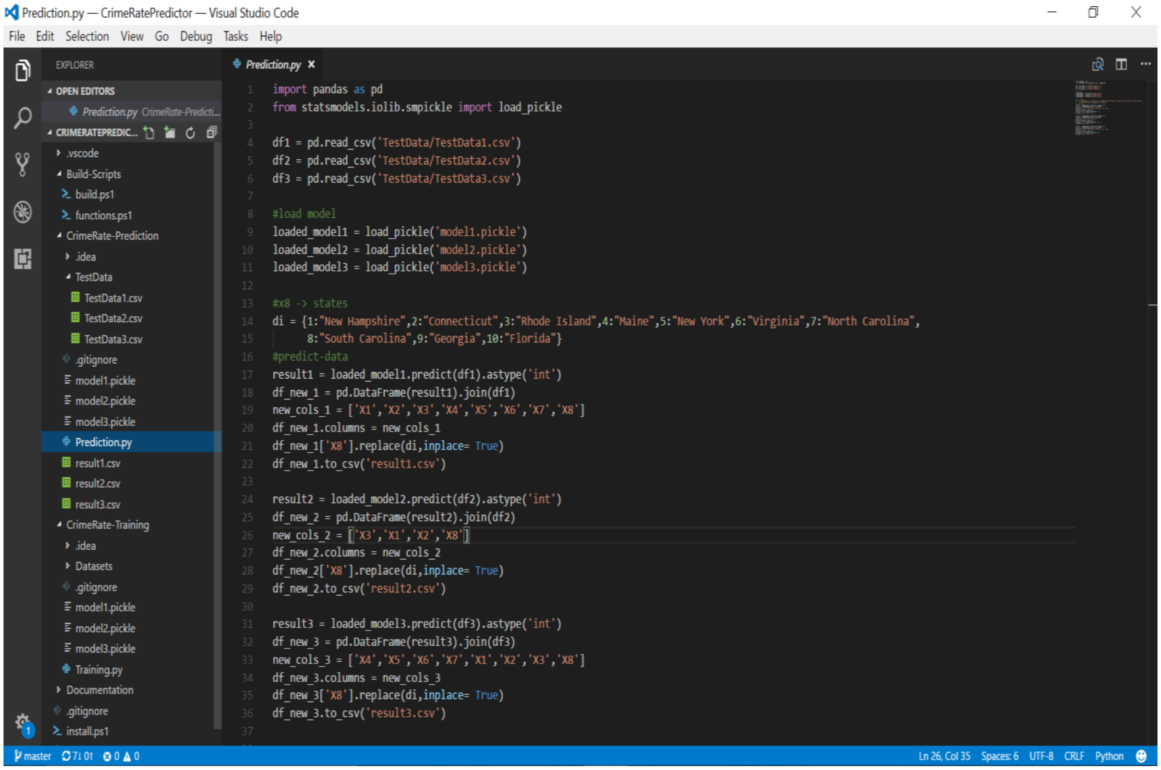


```python
import pandas as pd
from statsmodels.iolib.smpickle import load_pickle
from statsmodels.formula.api import ols
#csv to dataframe
df = pd.read_csv('Datasets/CrimeRate.csv')
#create model
model1 = ols("X1 ~ X2 + X3 + X4 + X5 + X6 + X7 + X8",df).fit()
model2 = ols("X3 ~ X1 + X2 + X8",df).fit()
model3 = ols("X4 + X5 + X6 + X7 ~ X1 + X2 + X3 + X8",df).fit()
#store model
model1.save('model1.pickle')
model2.save('model2.pickle')
model3.save('model3.pickle')
```

Training_Python

8.4 PREDICTION PROGRAM

Using the model we predict necessary features.

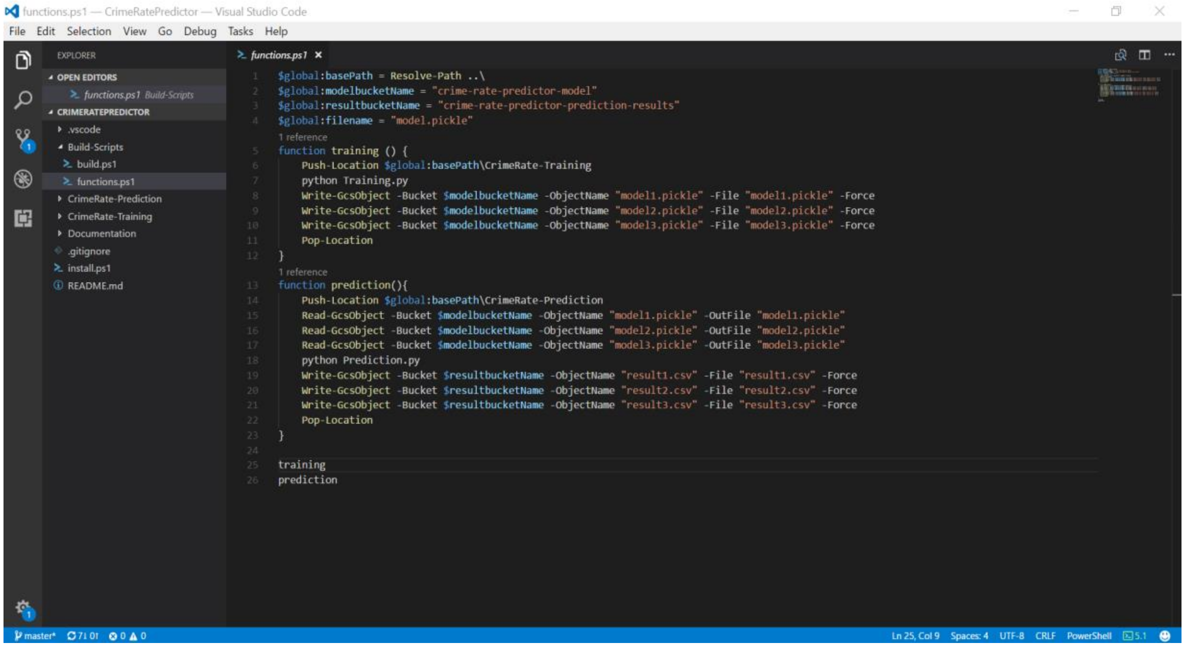


```python
import pandas as pd
from statsmodels.iolib.smpickle import load_pickle

df1 = pd.read_csv('TestData/TestData1.csv')
df2 = pd.read_csv('TestData/TestData2.csv')
df3 = pd.read_csv('TestData/TestData3.csv')

#load model
loaded_model1 = load_pickle('model1.pickle')
loaded_model2 = load_pickle('model2.pickle')
loaded_model3 = load_pickle('model3.pickle')

#x8 -> states
di = {1:"New Hampshire",2:"Connecticut",3:"Rhode Island",4:"Maine",5:"New York",6:"Virginia",7:"North Carolina",
      8:"South Carolina",9:"Georgia",10:"Florida"}
#predict-data
result1 = loaded_model1.predict(df1).astype('int')
df_new_1 = pd.DataFrame(result1).join(df1)
new_cols_1 = ['X1','X2','X3','X4','X5','X6','X7','X8']
df_new_1.columns = new_cols_1
df_new_1['X8'].replace(di,inplace= True)
df_new_1.to_csv('result1.csv')

result2 = loaded_model2.predict(df2).astype('int')
df_new_2 = pd.DataFrame(result2).join(df2)
new_cols_2 = ['X3','X1','X2','X8']
df_new_2.columns = new_cols_2
df_new_2['X8'].replace(di,inplace= True)
df_new_2.to_csv('result2.csv')

result3 = loaded_model3.predict(df3).astype('int')
df_new_3 = pd.DataFrame(result3).join(df3)
new_cols_3 = ['X4','X5','X6','X7','X1','X2','X3','X8']
df_new_3.columns = new_cols_3
df_new_3['X8'].replace(di,inplace= True)
df_new_3.to_csv('result3.csv')
```

Prediction_Python

8.5 PROJECT AUTOMATION SCRIPTS

Automation is done using PowerShell scripts.

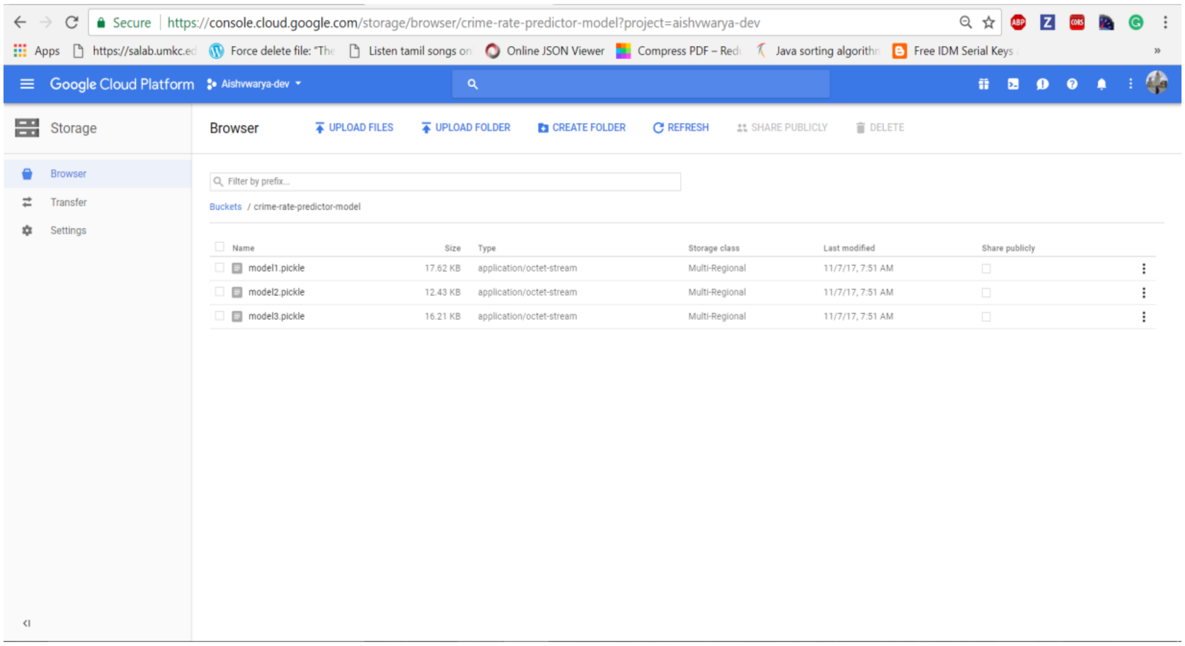


```powershell
$global:basePath = Resolve-Path ..\
$global:modelbucketName = "crime-rate-predictor-model"
$global:resultbucketName = "crime-rate-predictor-prediction-results"
$global:filename = "model.pickle"
function training () {
    Push-Location $global:basePath\CrimeRate-Training
    python Training.py
    Write-GcsObject -Bucket $modelbucketName -ObjectName "model1.pickle" -File "model1.pickle" -Force
    Write-GcsObject -Bucket $modelbucketName -ObjectName "model2.pickle" -File "model2.pickle" -Force
    Write-GcsObject -Bucket $modelbucketName -ObjectName "model3.pickle" -File "model3.pickle" -Force
    Pop-Location
}
function prediction(){
    Push-Location $global:basePath\CrimeRate-Prediction
    Read-GcsObject -Bucket $modelbucketName -ObjectName "model1.pickle" -OutFile "model1.pickle"
    Read-GcsObject -Bucket $modelbucketName -ObjectName "model2.pickle" -OutFile "model2.pickle"
    Read-GcsObject -Bucket $modelbucketName -ObjectName "model3.pickle" -OutFile "model3.pickle"
    python Prediction.py
    Write-GcsObject -Bucket $resultbucketName -ObjectName "result1.csv" -File "result1.csv" -Force
    Write-GcsObject -Bucket $resultbucketName -ObjectName "result2.csv" -File "result2.csv" -Force
    Write-GcsObject -Bucket $resultbucketName -ObjectName "result3.csv" -File "result3.csv" -Force
    Pop-Location
}

training
prediction
```

Automation_PowerShell

8.6 GOOGLE STORAGE BUCKET

We store the necessary model and the prediction results in appropriate buckets.

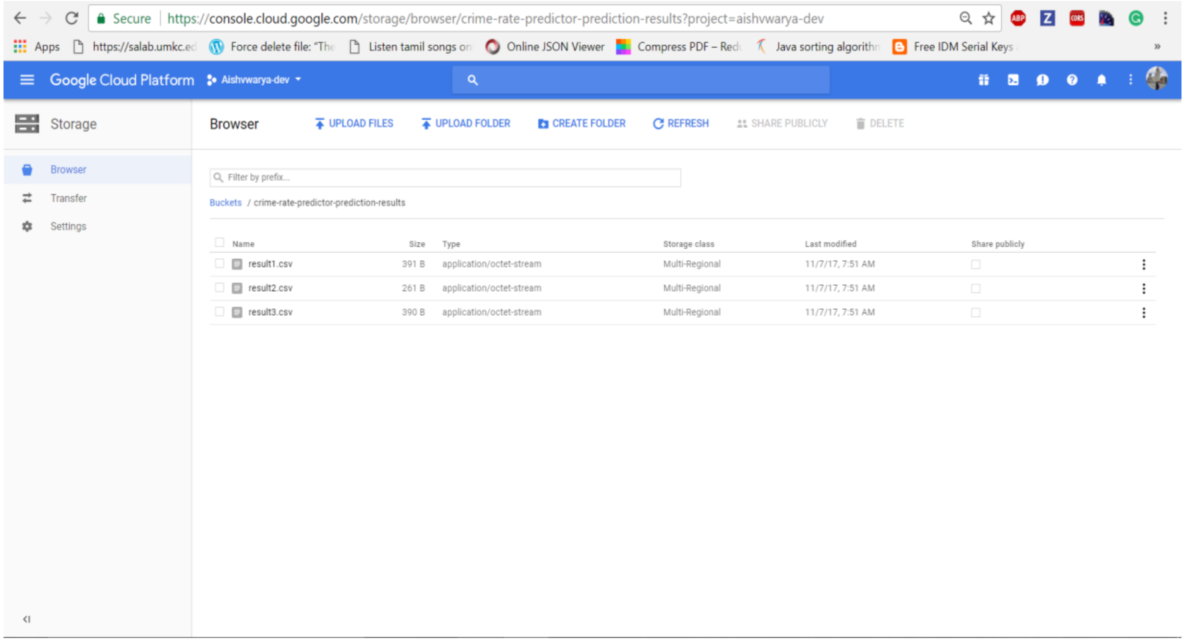


https://console.cloud.google.com/storage/browser/crime-rate-predictor-model?project=aishvwarya-dev

Buckets / crime-rate-predictor-model

Name	Size	Type	Storage class	Last modified	Share publicly
model1.pickle	17.62 KB	application/octet-stream	Multi-Regional	11/7/17, 7:51 AM	□
model2.pickle	12.43 KB	application/octet-stream	Multi-Regional	11/7/17, 7:51 AM	□
model3.pickle	16.21 KB	application/octet-stream	Multi-Regional	11/7/17, 7:51 AM	□

GCS- Bucket Modules



https://console.cloud.google.com/storage/browser/crime-rate-predictor-prediction-results?project=aishvwarya-dev

Buckets / crime-rate-predictor-prediction-results

Name	Size	Type	Storage class	Last modified	Share publicly
result1.csv	391 B	application/octet-stream	Multi-Regional	11/7/17, 7:51 AM	□
result2.csv	261 B	application/octet-stream	Multi-Regional	11/7/17, 7:51 AM	□
result3.csv	390 B	application/octet-stream	Multi-Regional	11/7/17, 7:51 AM	□

GCS- Bucket Results

8.7 GOOGLE BIG QUERY

We created Google Big Query tables from source which is google cloud storage bucket files. After the table is created, we can see schema of csv files and its configurations. We use Google Big Query to query the CSV files stored in GCS buckets.

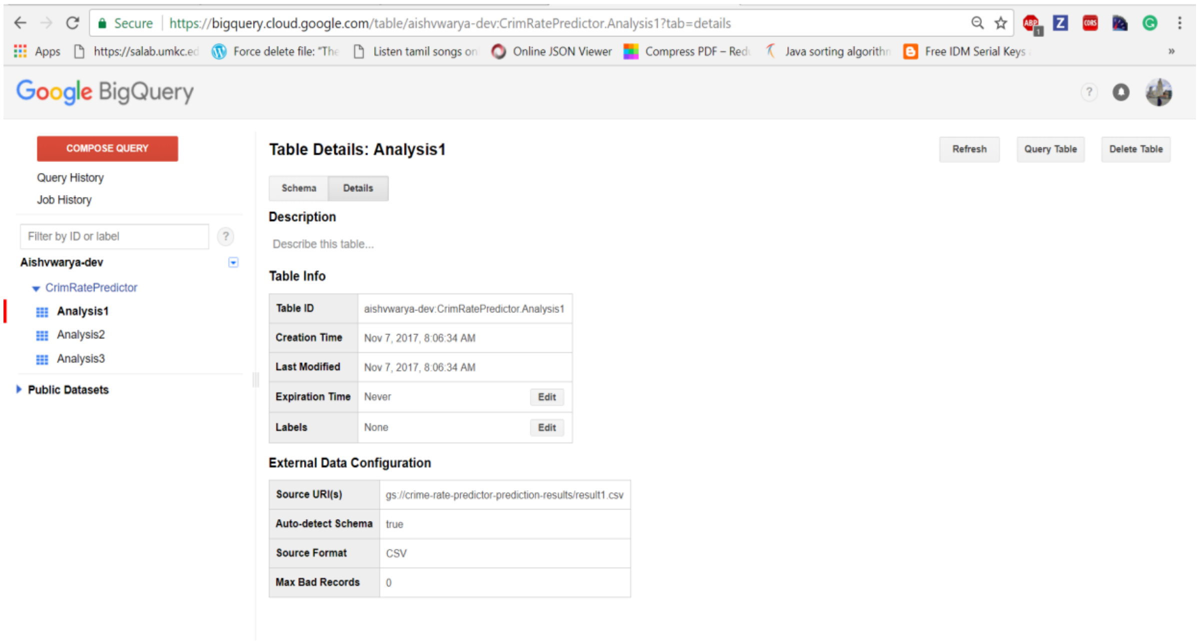


Table Details: Analysis1

Description

Describe this table...

Table Info

Table ID	aishvwarya-dev:CrimRatePredictor.Analysis1
Creation Time	Nov 7, 2017, 8:06:34 AM
Last Modified	Nov 7, 2017, 8:06:34 AM
Expiration Time	Never Edit
Labels	None Edit

External Data Configuration

Source URI(s)	gs://crime-rate-predictor-prediction-results/result1.csv
Auto-detect Schema	true
Source Format	CSV
Max Bad Records	0

GCS Big Query Details

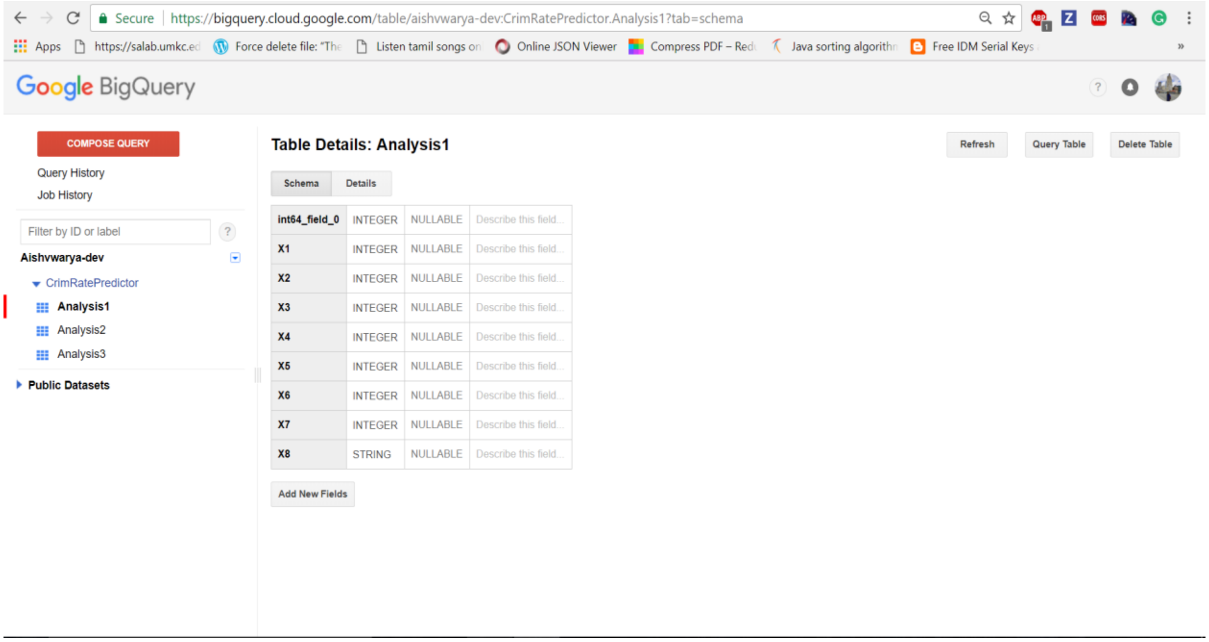


Table Details: Analysis1

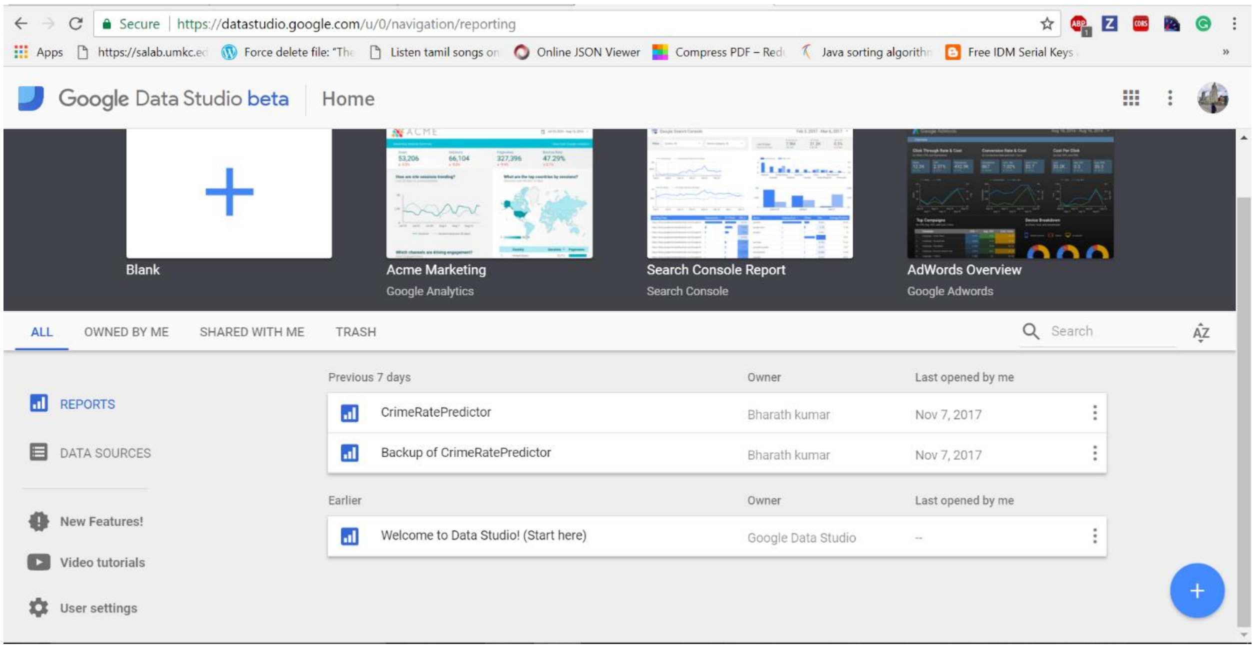
int64_field_0	INTEGER	NULLABLE	Describe this field...
X1	INTEGER	NULLABLE	Describe this field...
X2	INTEGER	NULLABLE	Describe this field...
X3	INTEGER	NULLABLE	Describe this field...
X4	INTEGER	NULLABLE	Describe this field...
X5	INTEGER	NULLABLE	Describe this field...
X6	INTEGER	NULLABLE	Describe this field...
X7	INTEGER	NULLABLE	Describe this field...
X8	STRING	NULLABLE	Describe this field...

GCS Big Query Schema

8.8 GOOGLE DATA STUDIO

Visualize big query results in data studio by choosing data source as BigQuery table.

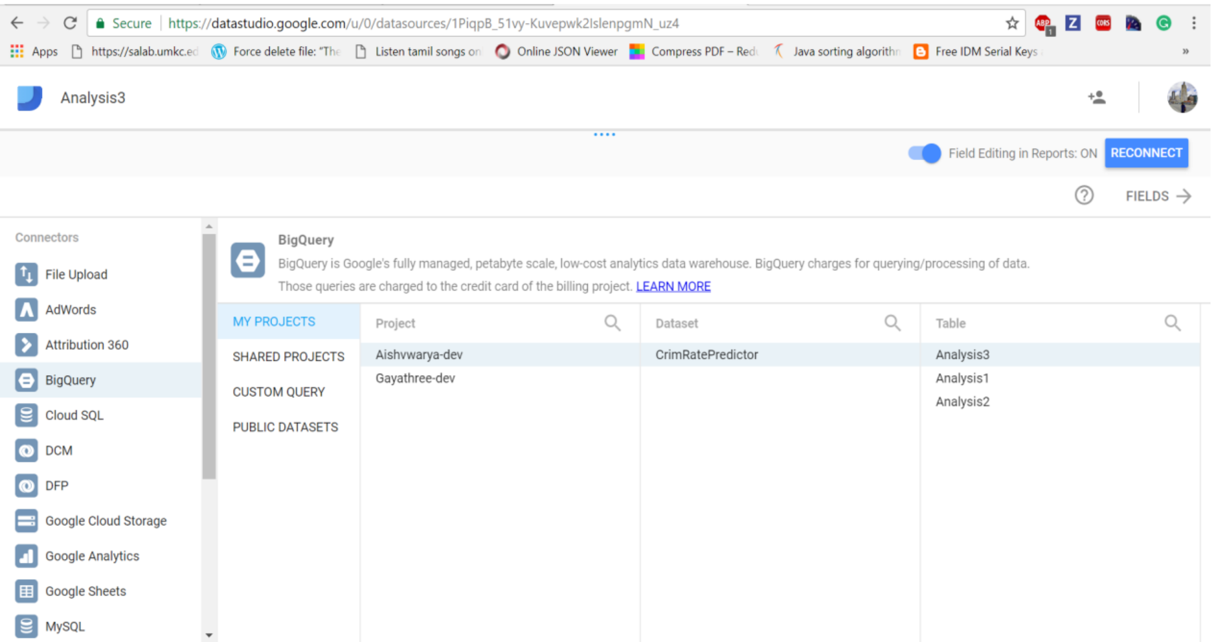
- We use Google data studio which is a business analytics tool to visualize the prediction results.
- First, we create data source from big query and choose appropriate table.
- After creating data source, we build metrics like Overall expected crimes by calculating sum of all expected crimes.
- Then, we create pie chart, geo locations and bar chart for visualizing prediction results.



Previous 7 days	Owner	Last opened by me
CrimeRatePredictor	Bharath kumar	Nov 7, 2017
Backup of CrimeRatePredictor	Bharath kumar	Nov 7, 2017

Earlier	Owner	Last opened by me
Welcome to Data Studio! (Start here)	Google Data Studio	--

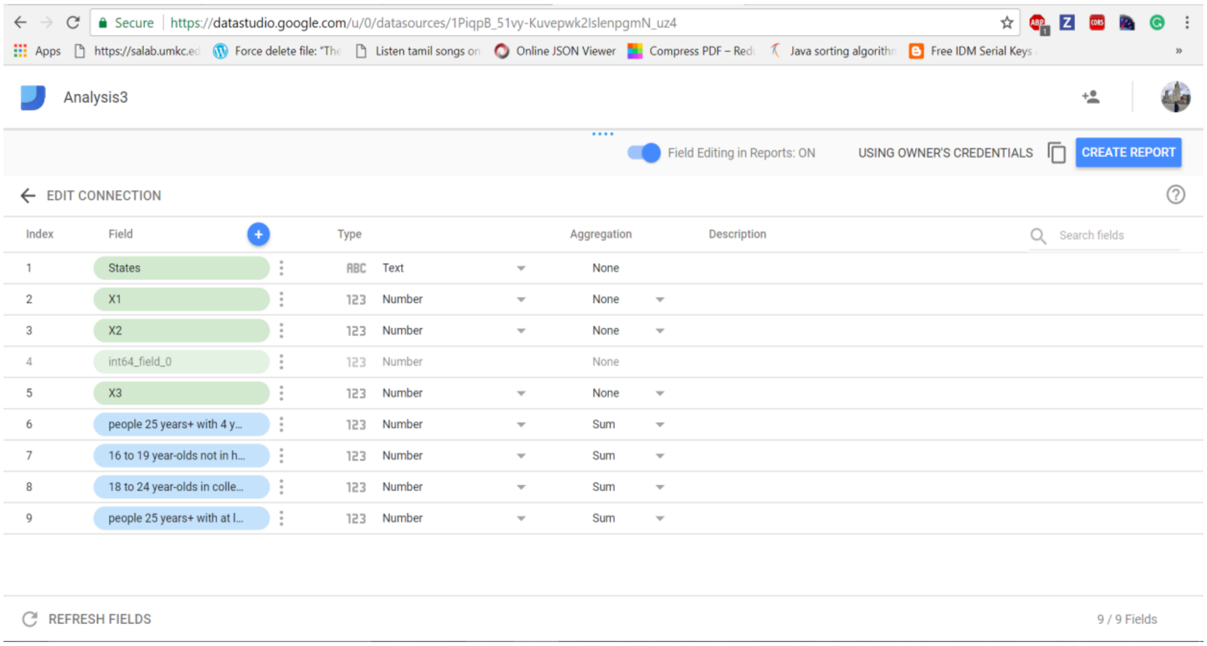
Data Studio HomePage



BigQuery

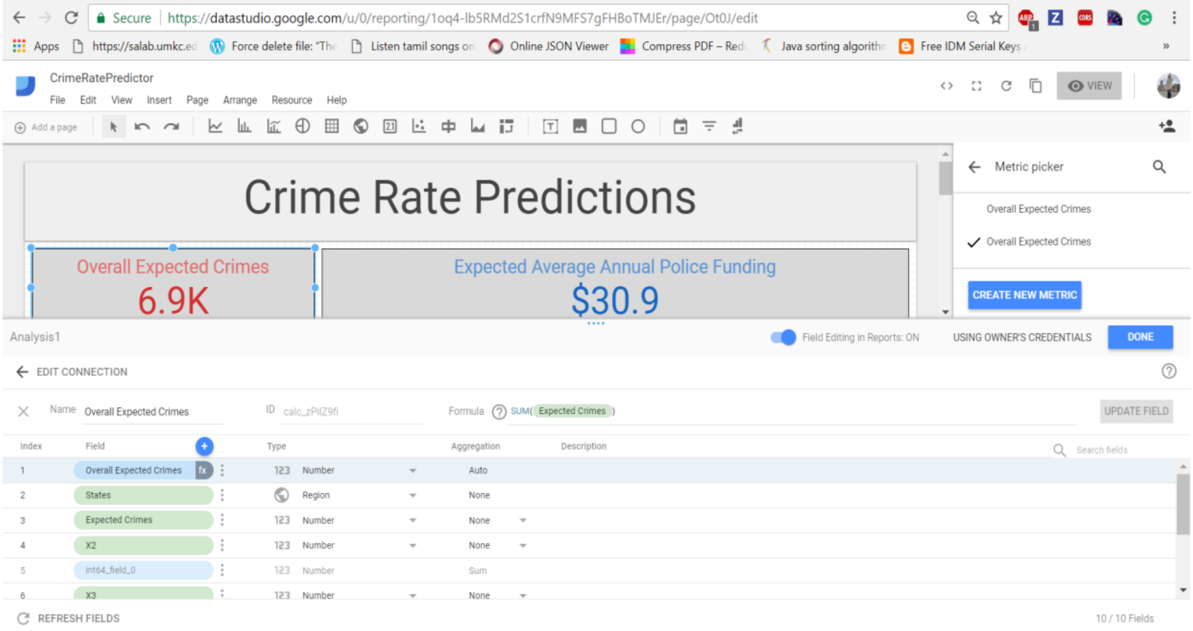
BigQuery is Google's fully managed, petabyte scale, low-cost analytics data warehouse. BigQuery charges for querying/processing of data. Those queries are charged to the credit card of the billing project. LEARN MORE

Data Studio Choosing Data Source


Secure
https://datastudio.google.com/u/0/datasources/1PiqpB_51vy-Kuvepwk2IslenpgmN_uz4
Apps
Analysis3
Field Editing in Reports: ON
USING OWNER'S CREDENTIALS
CREATE REPORT
EDIT CONNECTION
Search fields
REFRESH FIELDS
9 / 9 Fields

Index	Field	Type	Aggregation	Description
1	States	ABC Text	None	
2	X1	123 Number	None	
3	X2	123 Number	None	
4	int64_field_0	123 Number	None	
5	X3	123 Number	None	
6	people 25 years+ with 4 y...	123 Number	Sum	
7	16 to 19 year-olds not in h...	123 Number	Sum	
8	18 to 24 year-olds in colle...	123 Number	Sum	
9	people 25 years+ with at l...	123 Number	Sum	

Data Studio Data Source Schema



Crime Rate Predictions

Overall Expected Crimes 6.9K

Expected Average Annual Police Funding $30.9

Analysis1 — Field Editing in Reports: ON — USING OWNER'S CREDENTIALS — DONE

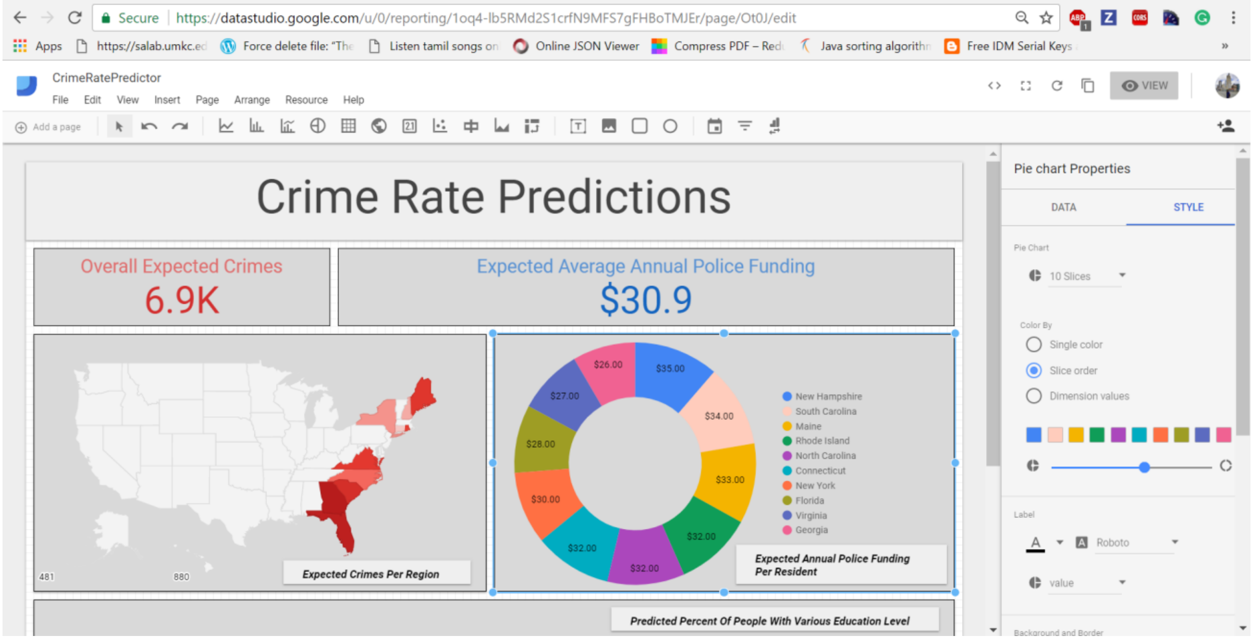
EDIT CONNECTION

Name: Overall Expected Crimes   ID: calc_zPilZ9fi   Formula: SUM(Expected Crimes)   UPDATE FIELD

Index	Field	Type	Aggregation	Description
1	Overall Expected Crimes	123 Number	Auto	
2	States	Region	None	
3	Expected Crimes	123 Number	None	
4	X2	123 Number	None	
5	int64_field_0	123 Number	Sum	
6	X3	123 Number	None	

REFRESH FIELDS   10 / 10 Fields

Data Studio Charts Sum



Crime Rate Predictions

Overall Expected Crimes 6.9K

Expected Average Annual Police Funding $30.9

Expected Crimes Per Region

Expected Annual Police Funding Per Resident

Predicted Percent Of People With Various Education Level

Pie chart Properties — DATA | STYLE

Pie Chart: 10 Slices

Color By: ○ Single color ● Slice order ○ Dimension values

Label: Roboto — value

Background and Border

Data Studio Charts Edit

8.9 RESULTS

The geo-location chart result:

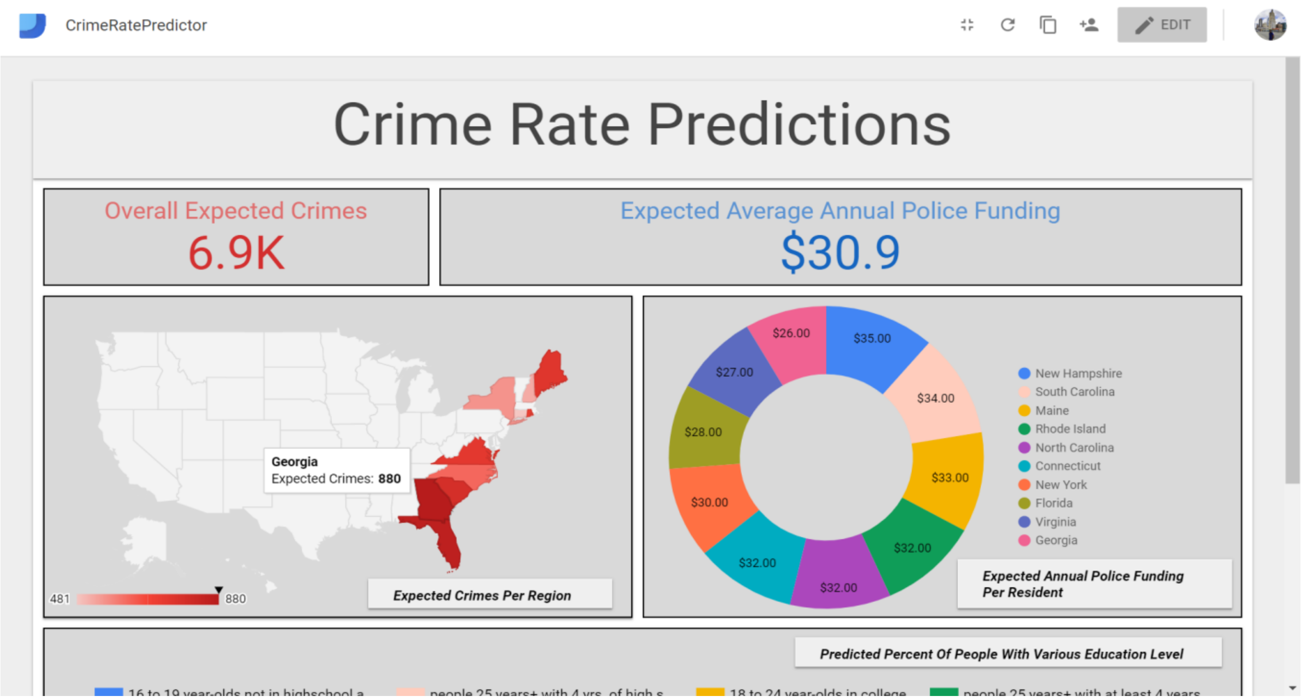
- Here, we visualize the expected crime per region. More the intensity of color red, more the crime.
- The result shows '**Georgia**' with expected crime of 880 and overall expected crime of all regions is 6,900 approximately.

The pie chart result:

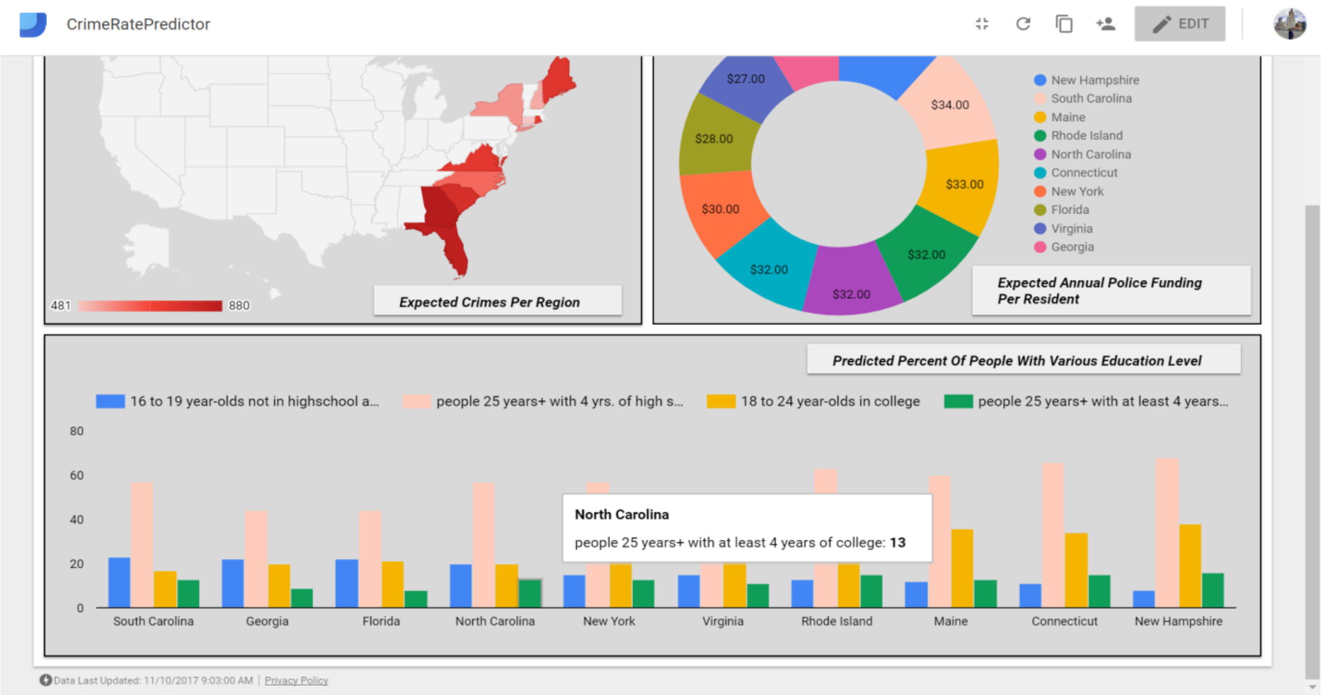
- Here, we visualize expected average annual police funding.
- The region '**New Hampshire**' requires most of the police funding with 11.3% of total funding.

The bar chart result:

- Here, we visualize the predicted percentage of people with various educational level.
- We see that **Georgia** and **Florida** have less literacy rate compared to other regions.
- This also reflects in **Expected Crime per region** where the intensity of red (crime) is maximum.
- Thus, our prediction algorithm works with maximum accuracy.


CrimeRatePredictor
EDIT
Crime Rate Predictions
Overall Expected Crimes
6.9K
Expected Average Annual Police Funding
$30.9
Georgia
Expected Crimes: 880
481
880
Expected Crimes Per Region
$26.00
$35.00
$27.00
$34.00
$28.00
$33.00
$30.00
$32.00
$32.00
$32.00
New Hampshire
South Carolina
Maine
Rhode Island
North Carolina
Connecticut
New York
Florida
Virginia
Georgia
Expected Annual Police Funding Per Resident
Predicted Percent Of People With Various Education Level
16 to 19 year-olds not in highschool a...
people 25 years+ with 4 yrs. of high s...
18 to 24 year-olds in college
people 25 years+ with at least 4 years...

Data Studio Result-1


CrimeRatePredictor
EDIT
481
880
Expected Crimes Per Region
$27.00
$34.00
$28.00
$33.00
$30.00
$32.00
$32.00
$32.00
New Hampshire
South Carolina
Maine
Rhode Island
North Carolina
Connecticut
New York
Florida
Virginia
Georgia
Expected Annual Police Funding Per Resident
Predicted Percent Of People With Various Education Level
16 to 19 year-olds not in highschool a...
people 25 years+ with 4 yrs. of high s...
18 to 24 year-olds in college
people 25 years+ with at least 4 years...
80
60
40
20
0
North Carolina
people 25 years+ with at least 4 years of college: 13
South Carolina
Georgia
Florida
North Carolina
New York
Virginia
Rhode Island
Maine
Connecticut
New Hampshire
Data Last Updated: 11/10/2017 9:03:00 AM | Privacy Policy

Data Studio Result-2

Link - https://datastudio.google.com/open/1oq4-lb5RMd2S1crfN9MFS7gFHBoTMJEr

9. REFERENCE

- http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html
- https://www.brookings.edu/research/more-cops/
- https://all4ed.org/press/crime-rates-linked-to-educational-attainment-new-alliance-report-finds/
- https://www.hackerearth.com/practice/machine-learning/linear-regression/multivariate-linear-regression-1/tutorial/
- http://ptl.sys.virginia.edu/ptl/sites/default/files/Area-Specific%20Crime%20Prediction%20Models.pdf