**DATA**

**Objective 1: Identify data needs and sources**

For our project, we need to get true data sets for getting the right predictions in our data mining process. So we find the open-sourced crime data set from the FBI site which has been stored in 2013. The site of FBI which we plan to get the data set is <https://ucr.fbi.gov/crime-in-the-u.s/2013/crime-in-the-u.s.-2013>. We also need census information about the states and cities in the US. So we plan to get this data set from the official census web site which has open-sourced data sets. The site of official census data set exists is [https://www2.census.gov/ acs2013\_1yr/summaryfile/](https://www2.census.gov/%20acs2013_1yr/summaryfile/). We also need the personal income data set for making a regression with crime data and personal income data to find out a prediction. So, we go forwardly to the Bureau of Economic Analysis web-site for getting the open-sourced data set( <https://www.bea.gov/national/index.htm>).

**Objective 2: Acquire data**

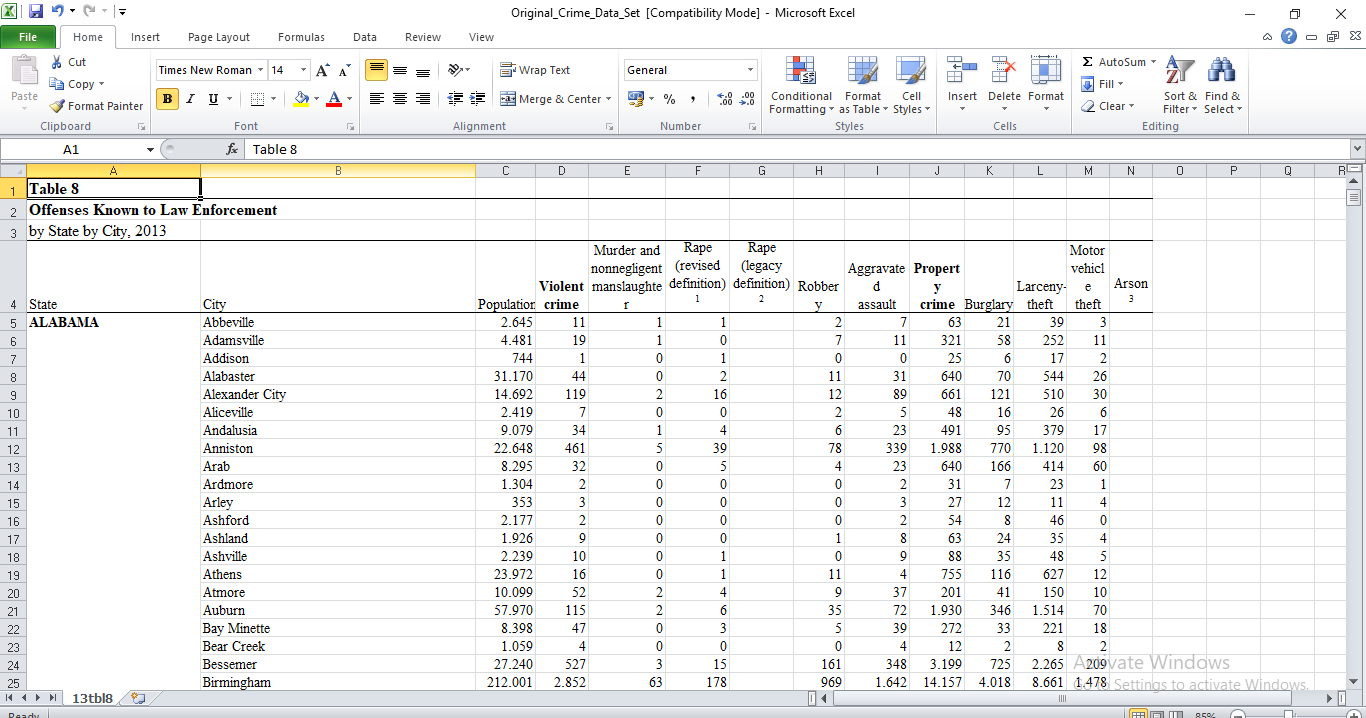
The crime data set has attributes of State, City, Population, Violent Crime, Murder and Non-negligent manslaughter, Rape (revised definition), Rape (legacy definition), Robbery, Aggravated Assault, Property Crime, Burglary, Larceny-theft, Motor vehicle theft, Arson.

The census data set has attributes of Geographic summary level, State FIPS code, Country FIPS code, State name, County Name, Sex, Hispanic Origin, Age group, Imputed race group, Resident population on April 1,2010.

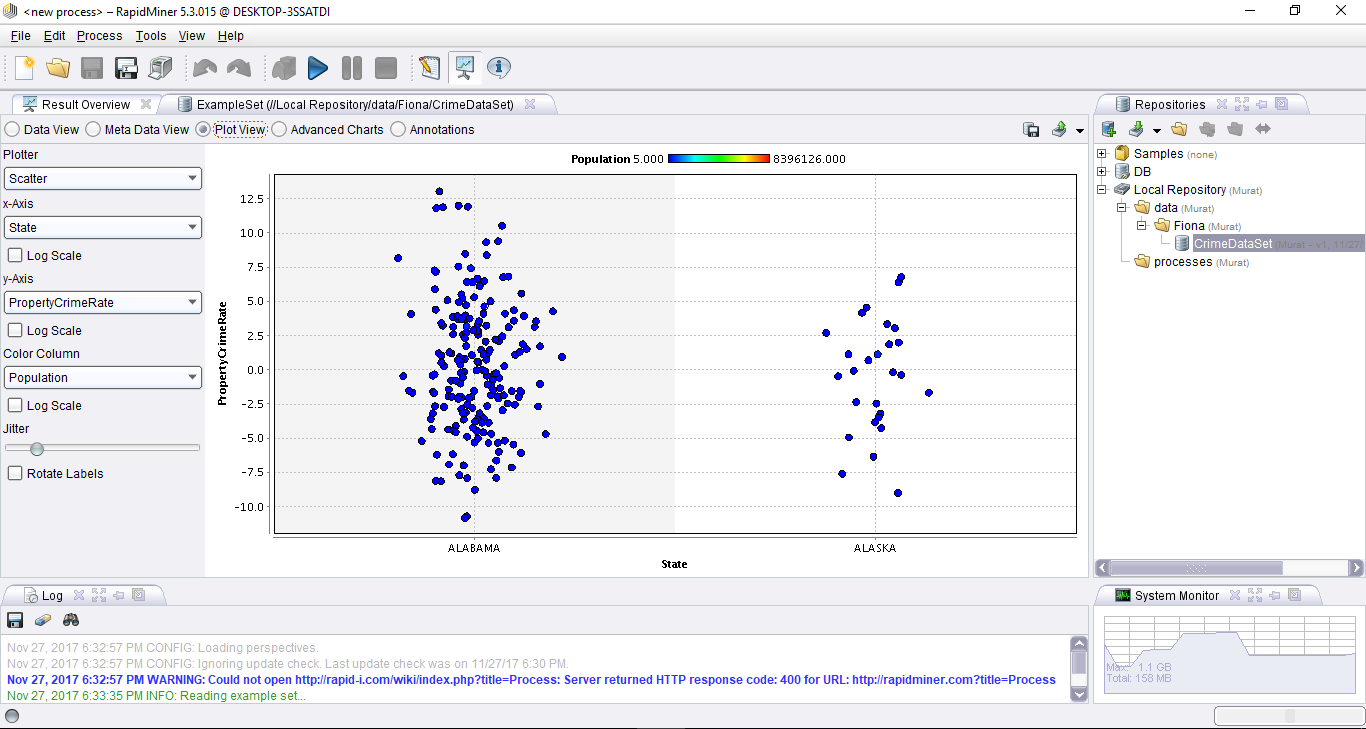
The personal income data set has attributes of GeoFips, GeoName, LineCode, Description, 2013. In GeoName column State name and City Name exist. In 2013 column the personal income value exists.

**Objective 3: Explore data visually**

1. **Crime Data Set**: In Figure-1, you can see the original file we downloaded from the FBI web-site. Figure-2 shows a simple scatter plot taken from this data set.

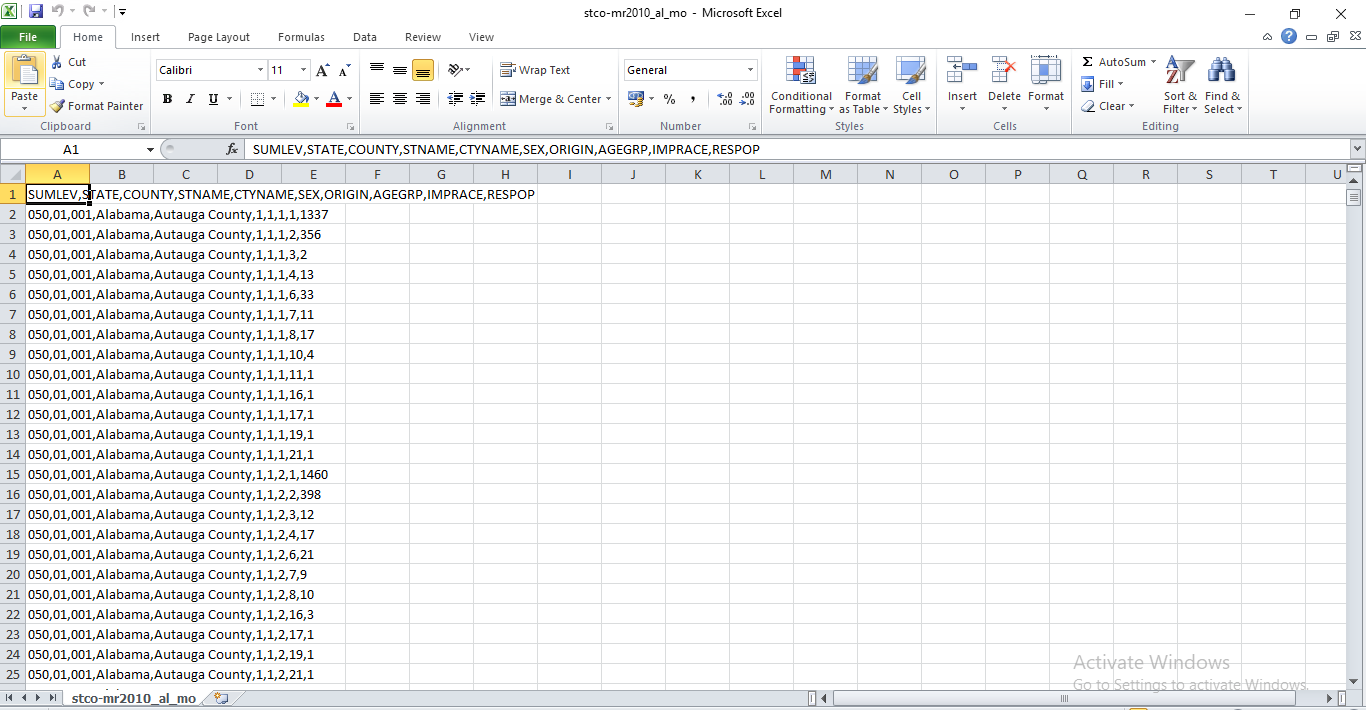


**Figure-1 :** Original Crime Data Set.

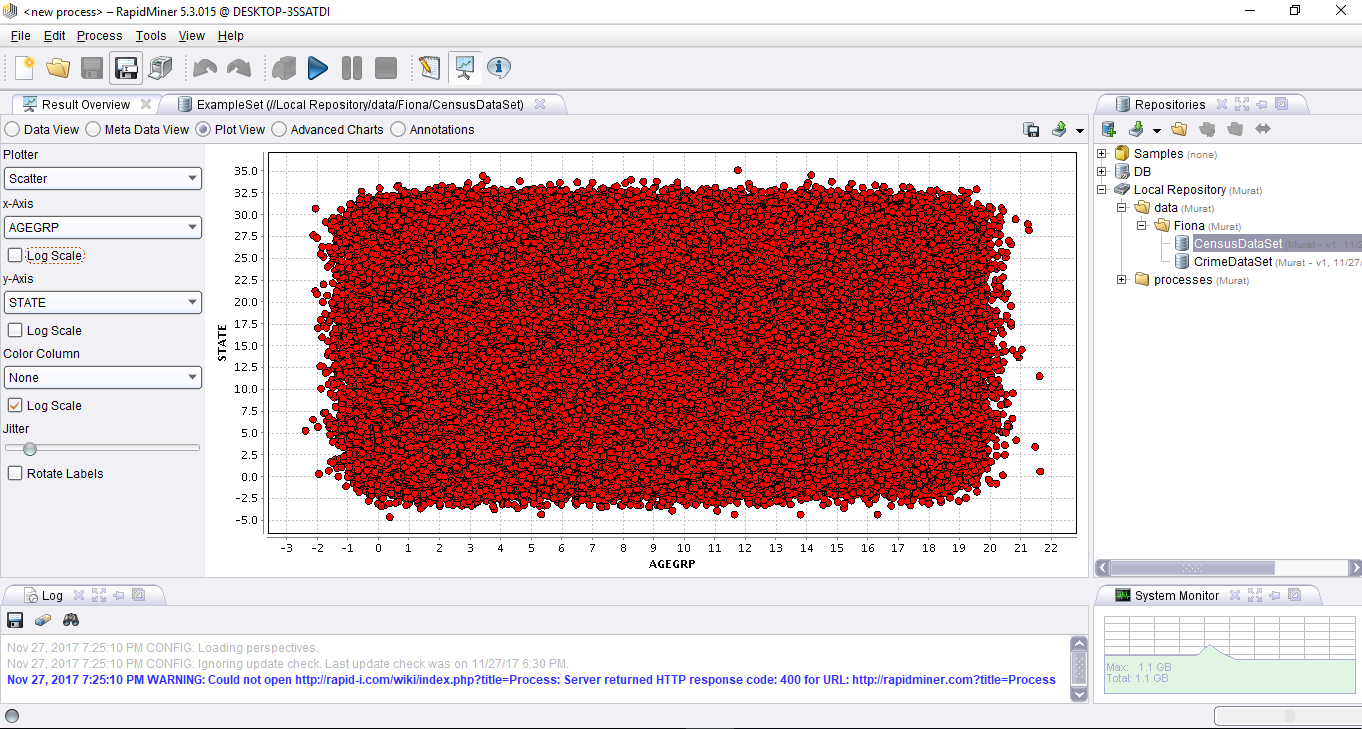


**Figure-2 :** Scatter plot by state to crime rate in Rapid Miner.

1. **Census Data Set:** This data set is in csv file format. In Figure-3, you can see the original version of the file as downloaded from the official census web site of US. In Figure 4 there is a simple scatter plot made in Rapid Miner.

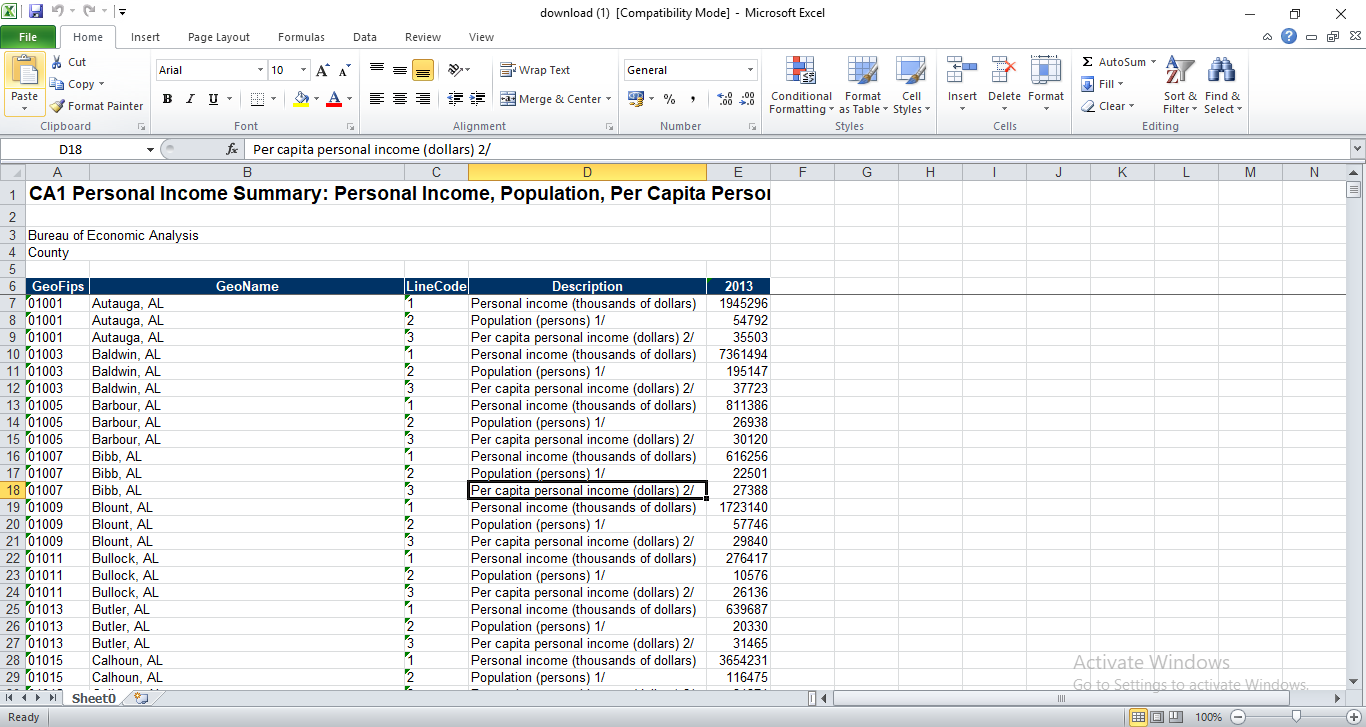


**Figure-3 :** Original Census Data Set.

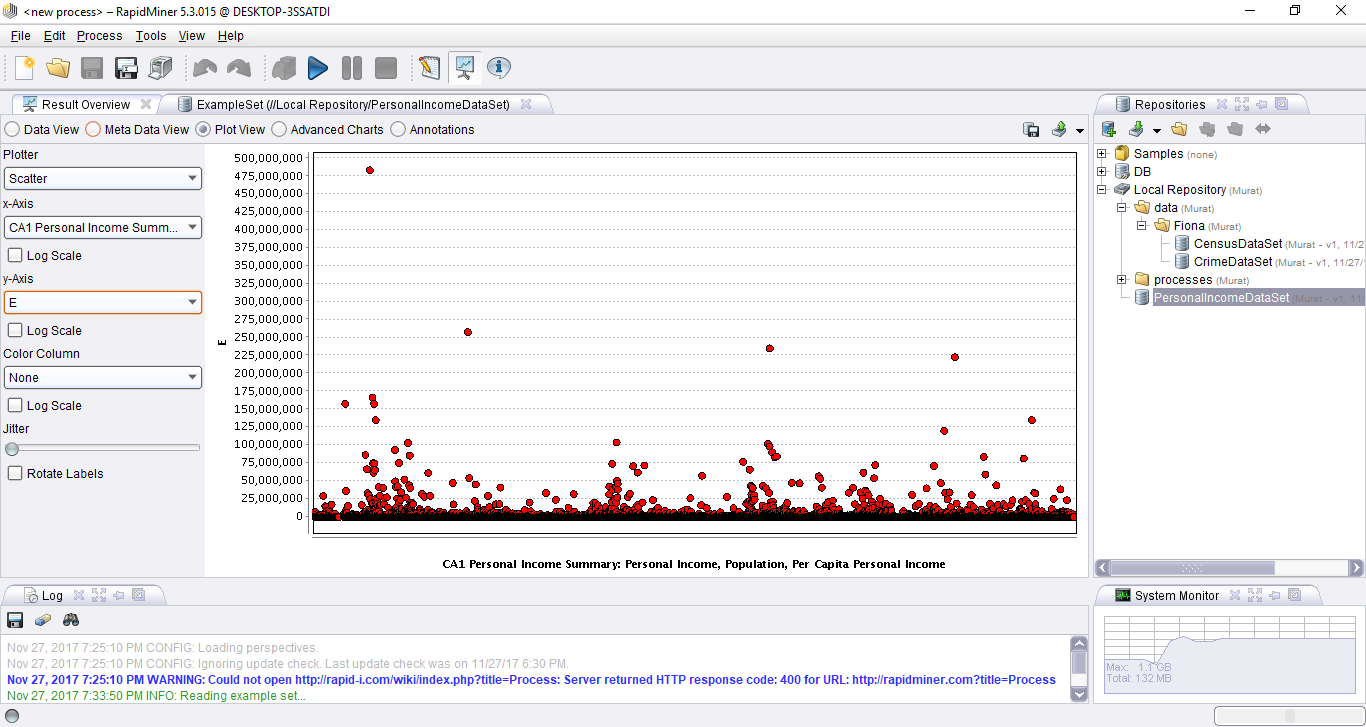


**Figure-4:** Scatter plot of census data set (Age Groups by State).

1. **Personal Income Data Set:** Figure 5 show the original data set downloaded from the Bureau of Economic Analysis. Figure 6 shows the simple scatter plot made in Rapid Miner.



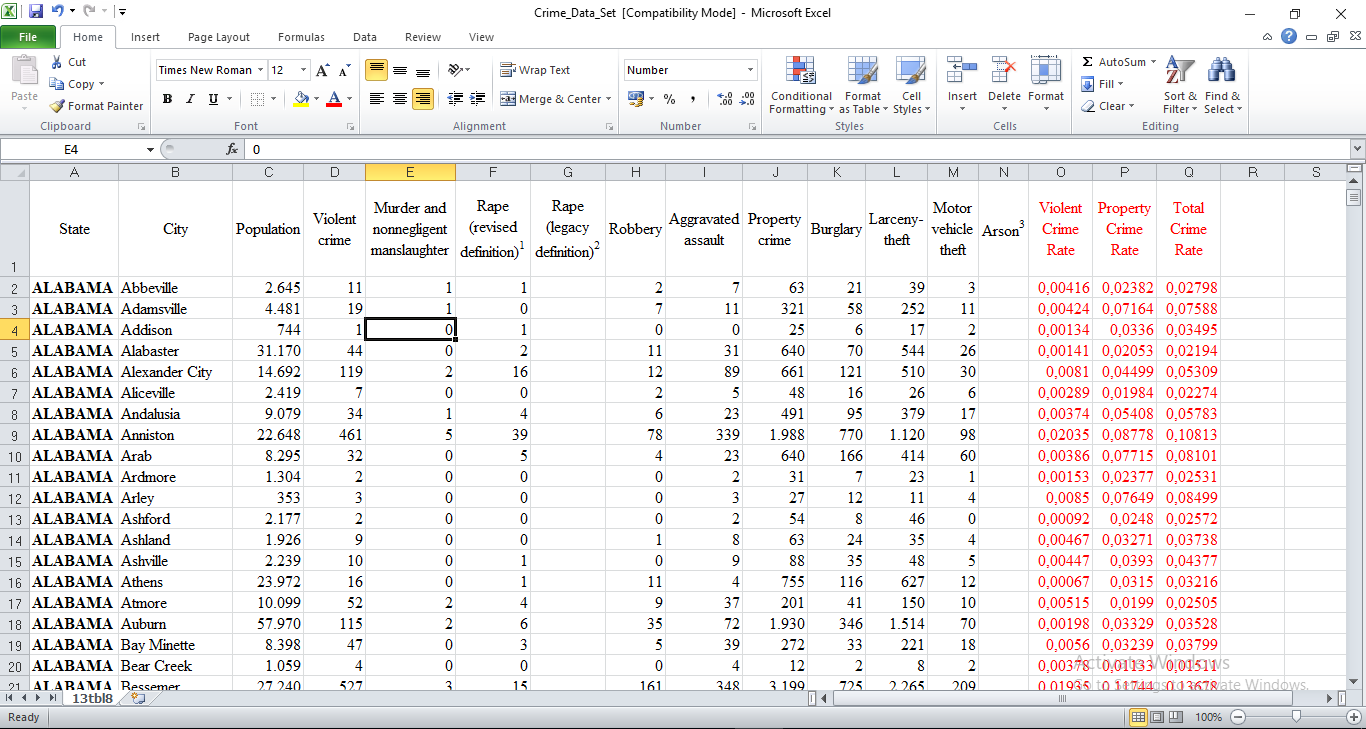
**Figure-5:** Original Personal Income Data Set.



**Figure-6:** Scatter plot of personal income in Rapid Miner.

**Objective 4: Harmonize, rescale and clean data**

1. **Crime Data Set:** In this data set, we make the State column get values row by row. We added three computed columns as “Violent Crime Rate”, “Property Crime Rate”, and “Total Crime Rate”. Violent Crime Rate is computed as “Violent Crime / Population”. Property Crime Rate is computed as “Property Crime / Population”. Total Crime Rate is computed as “(Violent Crime + Property Crime) / Population”.



**Figure-7:** Harmonized, Changed and Cleaned Data Set.

1. **Census Data Set:** This data set is in csv format, and we were able to import It to Rapid Miner by using the option of “Import csv file” and selecting the delimiter as comma “,”. This data is not so clear so, we have the explanation for the data below.

**VARIABLE DESCRIPTION**

SUMLEV Geographic summary level

STATE State FIPS code

COUNTY County FIPS code

STNAME State name

CTYNAME County name

SEX Sex

ORIGIN Hispanic origin

AGEGRP Age group

IMPRACE Imputed race group

RESPOP Resident population on April 1, 2010

**The key for SUMLEV is as follows:**

050 = County and/or statistical equivalent

**The key for SEX is as follows:**

1 = Male

2 = Female

**The key for ORIGIN is as follows:**

1 = Not Hispanic

2 = Hispanic

The key for AGEGRP is as follows:

1 = Age 0 to 4 years

2 = Age 5 to 9 years

3 = Age 10 to 14 years

4 = Age 15 to 19 years

5 = Age 20 to 24 years

6 = Age 25 to 29 years

7 = Age 30 to 34 years

8 = Age 35 to 39 years

9 = Age 40 to 44 years

10 = Age 45 to 49 years

11 = Age 50 to 54 years

12 = Age 55 to 59 years

13 = Age 60 to 64 years

14 = Age 65 to 69 years

15 = Age 70 to 74 years

16 = Age 75 to 79 years

17 = Age 80 to 84 years

18 = Age 85 years or older

**The key for IMPRACE is as follows:**

1 = White alone

2 = Black or African American alone

3 = American Indian and Alaska Native alone

4 = Asian alone

5 = Native Hawaiian and Other Pacific Islander alone

6 = White and Black or African American

7 = White and American Indian and Alaska Native

8 = White and Asian

9 = White and Native Hawaiian and Other Pacific Islander

10 = Black or African American and American Indian and Alaska Native

11 = Black or African American and Asian

12 = Black or African American and Native Hawaiian and Other Pacific Islander

13 = American Indian and Alaska Native and Asian

14 = American Indian and Alaska Native and Native Hawaiian and Other Pacific Islander

15 = Asian and Native Hawaiian and Other Pacific Islander

16 = White and Black or African American and American Indian and Alaska Native

17 = White and Black or African American and Asian

18 = White and Black or African American and Native Hawaiian and Other Pacific Islander

19 = White and American Indian and Alaska Native and Asian

20 = White and American Indian and Alaska Native and Native Hawaiian and Other Pacific

Islander

21 = White and Asian and Native Hawaiian and Other Pacific Islander

22 = Black or African American and American Indian and Alaska Native and Asian

23 = Black or African American and American Indian and Alaska Native and Native

Hawaiian and Other Pacific Islander

24 = Black or African American and Asian and Native Hawaiian and Other Pacific Islander

25 = American Indian and Alaska Native and Asian and Native Hawaiian and Other Pacific

Islander

26 = White and Black or African American and American Indian and Alaska Native and

Asian

27 = White and Black or African American and American Indian and Alaska Native and

Native Hawaiian and Other Pacific Islander

28 = White and Black or African American and Asian and Native Hawaiian and Other

Pacific Islander

29 = White and American Indian and Alaska Native and Asian and Native Hawaiian and

Other Pacific Islander

30 = Black or African American and American Indian and Alaska Native and Asian and

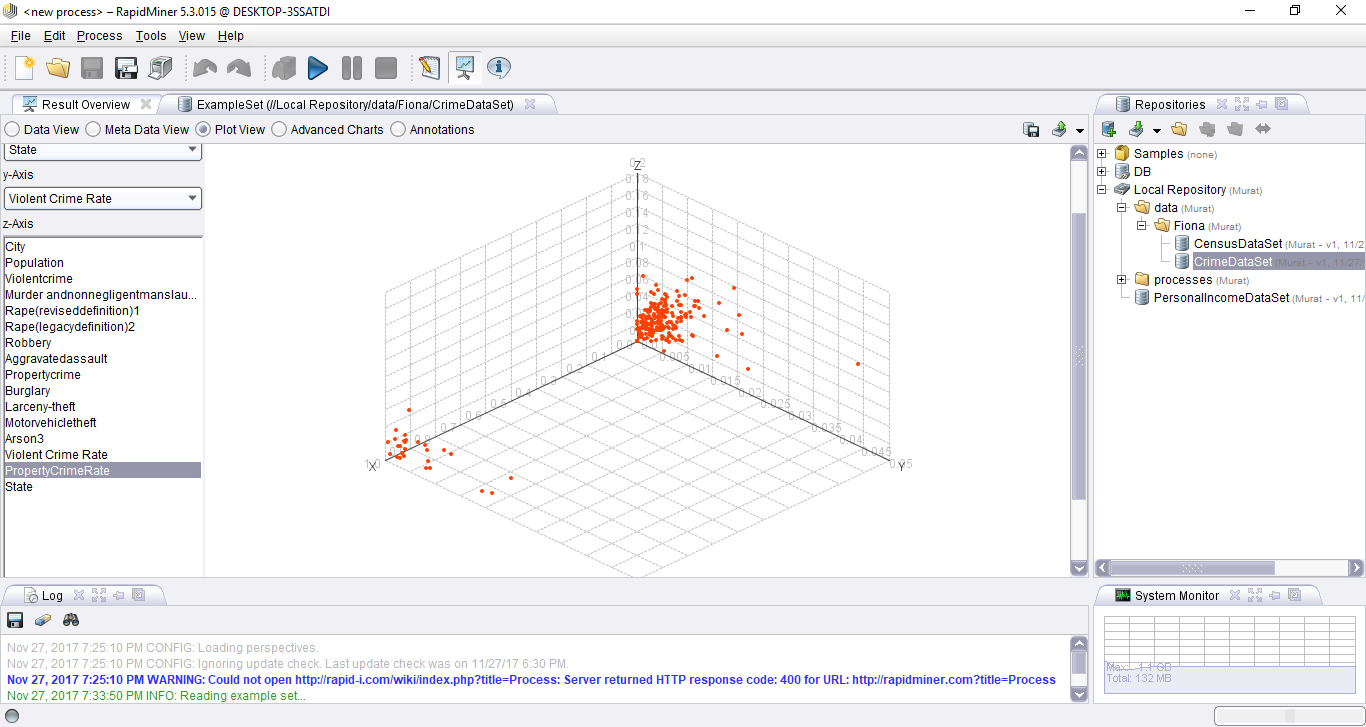
Native Hawaiian and Other Pacific Islander

31 = White and Black or African American and American Indian and Alaska Native and

Asian and Native Hawaiian and Other Pacific Islander

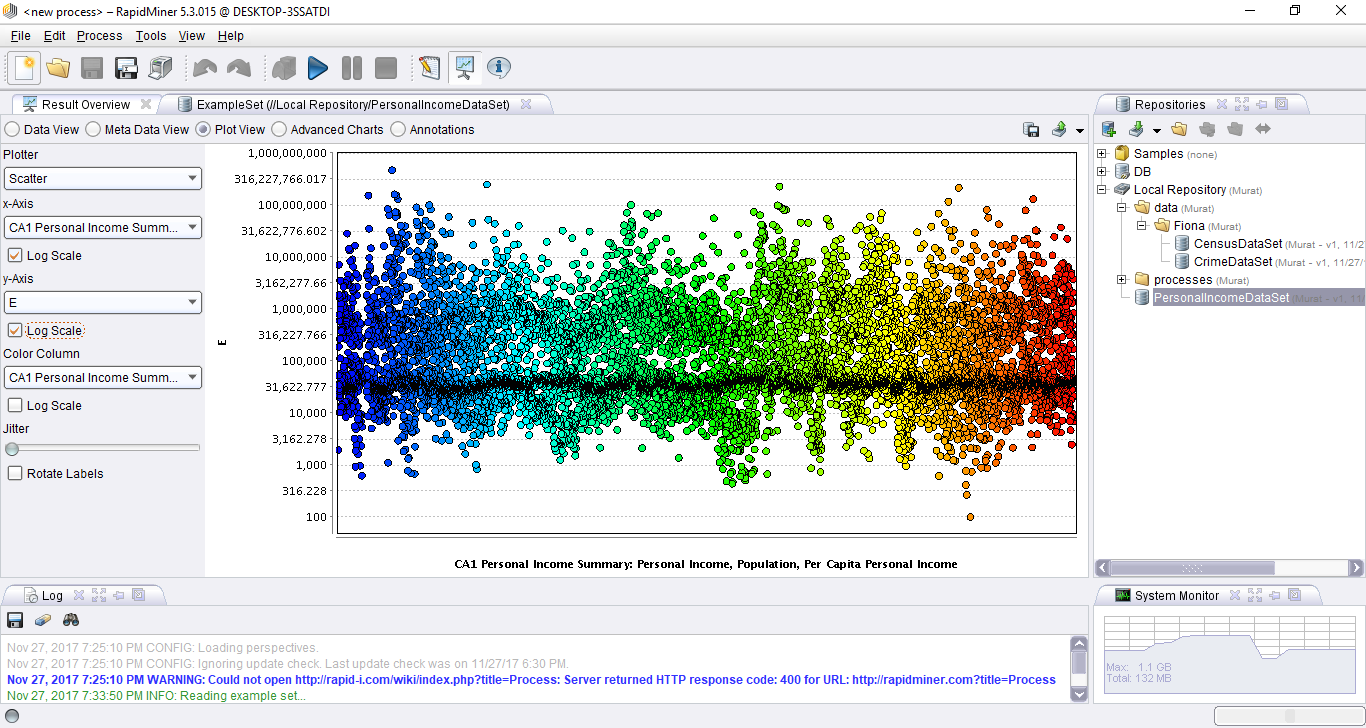
**Objective 5: Document and report findings (e.g. quality report, data insights)**

In Figure 8, we made a 3D plot of dimensions in Violent Crime Rate, Property Crime Rate and State. This graph shows us the Violent Crime Rate and Property Crime Rate that are orange points in the Y and Z dimensions scattering according to the cities’ records in the states in US.



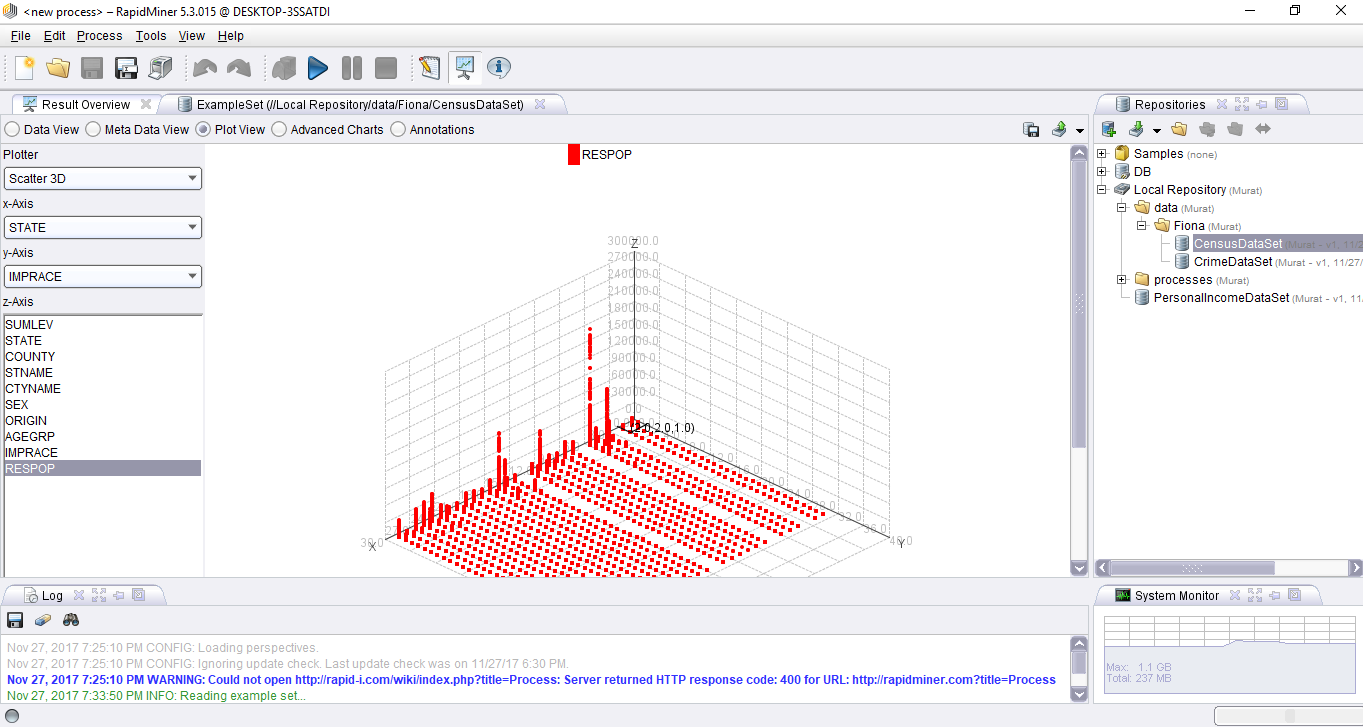
**Figure 8:** Violent Crime and Property Crime by State.

In Figure 9, we made a scatter plot of personal income, population and per capita personal income. This graph shows us the scatter of population data, personal income data and division of both of them that means per capita personal income.



**Figure 9:** Personal Income by State.

In Figure-10, we played with the census data and made a 3D plot of race, population and state dimensions. This graph shows us the resident population(RESPOP) of different races in the counties of the states in US.

**Figure 10:** Race ethnicity and Resident Population by State.

**Objective 6: Refine the business and analytics problem statements**

The problem of high crime rate in some places in US may be solved and determined by data analysis. For coming up a solution for this problem, we will analyze and implement a data mining algorithm in data sets we have taken from open-source data sources. So we will analyze data by using a data-miner tool called Rapid Miner. At the end of the project, we aim to reach out solutions of effects of the crime and safest places in US in terms crime rates.