***Part 1:***

***Look for potential data that you can use to solve the challenge you identified previously and:***

* ***Identify the types of data that are involved***

[Original] 104 Numeric, 43 non-numeric

[After Transformation fix] 143 numeric, 4 non-numeric

* ***Identify the potential sources for getting this data***

<http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized#>

* ***What structures they may have?***

The data set we have is structured dataset with well-defined rows and columns in it.

* ***What is the unit of analysis?***

Our unit of analysis is at County level which is identified by concatenation of two columns “County Code” and “Community Code”. The hierarchy of unit is following:

* State
  + Community
    - County
* ***What data retrieval/processing/transformation are needed?***

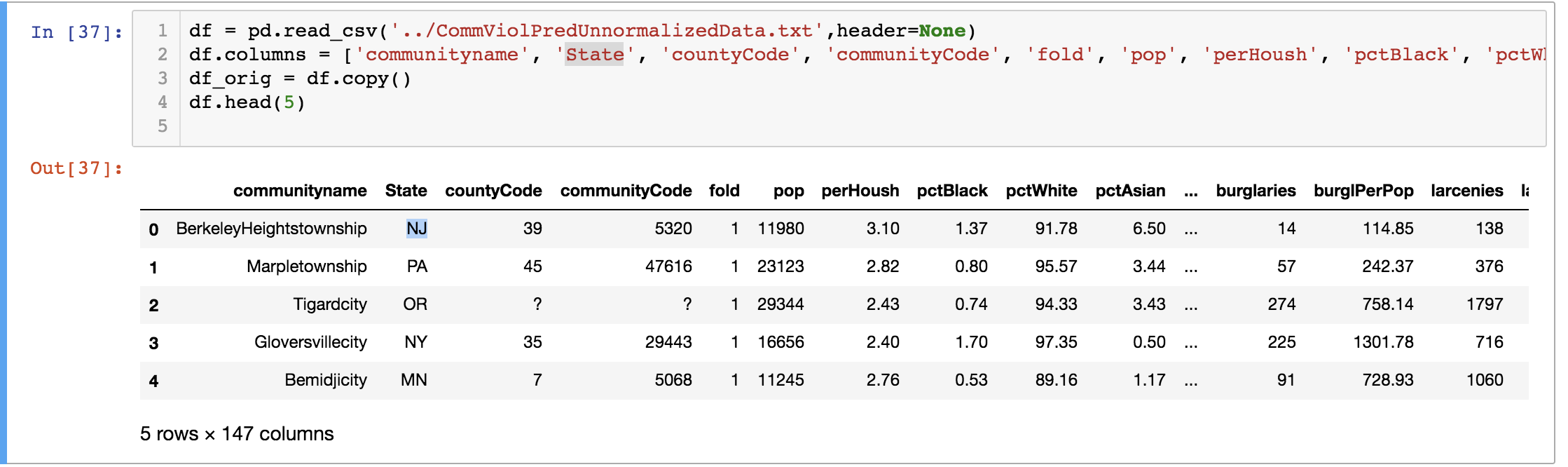
For now we can conclude following from our initial data analysis:

* Missing values in some numeric columns
* There are some numeric columns that are interpreted as “objects” by pandas. We will need to fix this. The reason for this is the dataset marks null values as “?” and not *np.nan*.
* Some numeric columns that are actually ID Codes, these will need to be dealt with before proceeding to visualization.
* Scaling will need to be done as the limit of attributes are not consistent
* Outlier detection will be performed later in the stage
* Distribution is mostly right skewed and normal distributed; we will look into this after outlier fix again
* We do not have any binary variables or categorical variables. We will, however, use “States” column and convert it in to 5 regions. This will enable us to conduct region specific analysis.

***Part 2:***

* ***Use pandas to load the data (csv, json, xlsx, html) into a DataFrame***

Done



* ***Pick one (or more) datasets and run the descriptive analytics to answer:***
  + ***What is the data about?***

The dataset we’ve chosen is about community and county crime. We have some data about housing conditions within each geographical area. We also have some attributes regarding human demographics like age, income etc. We also have some social conditions data like poverty level etc.

* + ***What types of data are involved?***

[Original] 104 Numeric, 43 non-numeric

[After Transformation fix] 143 numeric, 4 non-numeric

* + ***What is the unit of analysis?***

Our unit of analysis is at County level which is identified by concatenation of two columns “County Code” and “Community Code”. The hierarchy of unit is following:

* + State
    - Community
      * County
  + ***Can you check how noisy/clean is the data?***

The data has some data quality issues in it. We have conducted initial analysis of cleanliness of data, results of which are summarized above in “transformations needed” question in part 1.