ECS648U/ECS784P - DATA ANALYTICS - GROUP COURSEWORK

Group Size : 4

|  |  |
| --- | --- |
| Student Name | Effort |
| Anket |  |
| Jaikiran |  |
| Pranav |  |
| Suruchi Rastogi |  |

CONTENTS

Introduction 4

Background Information 4

Problem Statement 5

Project Aims 5

Dataset Description 5

Learnings from exploring data 5

Features used 6

Details of modelling process 6

Challenges 6

Success 7

Possible extension of project. 8

Conclusion 7

**Introduction:**

With the widespread production of data everyday the idea is to make sense of the data, analyze it and use it for high-level business productivity and efficiency. Data analysis involves the extraction of patterns and trends from the data so that they can be successfully used to predict business insights and patterns. Data analysis is the process of cleaning, processing and modelling data to achieve important information and decision-making conclusions.

One of the promising areas where it can be applied in order to make a change is healthcare. Data analytics in this field has the potential to reduce costs of treatment, avoid predictable diseases and outbreaks as well as improving the quality of life around us in general. Today’s treatment delivery methods are challenged in accordance with the growth in an average human lifespan. But in accordance with this there is massive amounts of data generated with the healthcare professionals which gives us the opportunity to overcome these challenges using data analytics and machine learning techniques.

In this project we are going to use the dataset from Health Inventory Data Platform to analyze different patterns and try to predict diseases. Depending on the race, sex, place, year we will show statistics of population of different cities and race suffering from Cancer, Chronic diseases, HIV/ Aids, infectious diseases, injury violence etc. and we will predict the same for the current year depending on the previous year’s data.

**Background Information:**

External libraries used:

Pandas: It is a prominent python open-source library used for data manipulation and analysis. Pandas is an high-performance, easy-to-use data structures and data analysis tool. It also has a variety of methods that can be utilized when working on data science and machine learning problems in python .The main Pandas functionality that will be using in this coursework are the DataFrame objects, which makes data suitable for data analysis, also the other features like renaming the columns, reading the csv, dropping null values using dropna function, filtering the rows as per our needs. Pandas make the data cleaning process more easy and provide the sufficient data with minimum number of line codes. The pandas library is also easy to learn and to implement. There are also many developer communities on which we can get different examples of code to be used and to understand.

NumPy: NumPy is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features a few important ones being a powerful N-dimensional array object, sophisticated functions, tools for integrating C/C++ code etc. NumPy can also be used as an efficient multi-dimensional container of generic data. It is a library for python programming language which adds support for large, multi-dimensional arrays and matrices along with a massive collection of high-level mathematical functions to operate on these multi-dimensional arrays.

Scikit-Learn: Scikit learn is a machine learning library for python programming. It provides a range of supervised and unsupervised algorithms. It is a library focused on modelling data not loading, manipulating and summarizing the data. The scikit learn can be very helpful library in implementing supervised learning, unsupervised learning algorithms and functions, it has inbuild regression and hypothesis function which can be used to do predictions.

Matplotlib: Matplotlib is a plotting library which is a numerical extension of the Numpy library used for Python programming language. It provides an object-oriented API for embedding plots into applications using GUI toolkits. It produces publication quality figures in a variety of card copy formats and interactive environments across platforms.

Supervised-Learning: Within machine learning there are two major categories, unsupervised learning and supervised learning. Supervised learning in machine learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. A typical supervised learning problem is solved using these following steps, firstly determining the training example. The initial basic step before doing anything else the user should determine what kind of data will be used as the training set. The next step would be to gather that training set with inputs and outputs because we need a representation of the real-world function. Thirdly the determination of the input feature of the function. The accuracy of the training set heavily depends on how the object is represented. Next step would be to determine the structure of the learned function and corresponding learning algorithm. Then the design is complete, and we will have to run the algorithm on the training set. Sometimes certain control parameters are added to optimize the performance on a sub set of the training set via cross-validation. Lastly, we evaluate the accuracy of the learned function. After adjusting the parameters and learning performance of the function the function should be tested on a different test set from the training set to measure it

Regression: Regression belongs to the class of supervised learning tasks where the datasets that are used for predictive/statistical modelling contain continuous labels. Regression attempts to predict one dependent variable and a series of changing variables. In regression we have two types namely, simple linear regression and multiple linear regression. Linear regression is used as a way of predicting a response based on a single predictor variable. An assumption is made that there is a linear relationship between the predictor response and the predictor variable. Linear regression is a statistical model that examines the linear relationship between two or more variables, namely a dependent variable and independent variable. The basic concept of linear relationship is that when one variable increases the dependent also increases with it correspondingly. In a simple linear regression model the slope and y-intercept are derived from the data and we build a model based on it. Simple linear regression models also include some errors in the data. In cases where there is more than one independent variable, we will be using multiple linear regression model. There are two main ways to perform linear regression in python using statsmodel and scikit-learn

**Problem Statement:**

The Healthcare and regulatory sector is so big due to the large number of population all around the world and it makes it difficult for the researchers to perform their research since they can’t select the right number of population to do the analysis or in short they don’t get the proper and generalized data on which they can be reliable. The data generated is huge but is not efficiently harnessed. The data needs to be cleaned, preprocessed and analyzed before it can be useful to any research.

With the evolution of data science and machine learning techniques now this is achievable. In this project we will be cleaning, processing and analyzing the data set for different categories like cancer, HIV/Aids, infectious diseases, chronic diseases, injury and violence and life death expectancy. The data set is huge and with many missing values, so it needs to be processed properly before analysis.

**Project Aims:**

1. To process the data in order to get separate data in each category and to further clean the data to ensure all columns have values present.
2. Try to separate each category based on an indicator like race, place, year and to also calculate the number of cases for each sub indicator within the data.
3. Aim to plot the graphs in each category for visual representation.
4. Aim to the generated data in each category in order to predict the modules.
5. Aim to present a statistical representation which will help in a simpler understanding to the analysis and future use of that data.

**Dataset Description:**

Healthcare is one of the biggest beneficiaries of big data & analytics. The Health Inventory Data Platform is an open data platform that allows users to access and analyse health data from 26 cities, for 34 health indicators, and across six demographic indicators. It is the sixth edition of a report initially developed by the Chicago Department of Public Health to present epidemiologic data specific to large cities. The last BCHI was published in 2007. This edition is the first to be produced and issued by the Big Cities Health Coalition.

We have Referred to data from this link <https://www.datasciencecentral.com/profiles/blogs/10-great-healthcare-data-sets>:

**Learnings From exploring data:**

Features Used: We have used some features of pandas, NumPy , and matplotlib .Such as dropping a particular column because of many null values and encoding the categorical data into numerical data to be used in machine learning.

Suruchi pls explain more

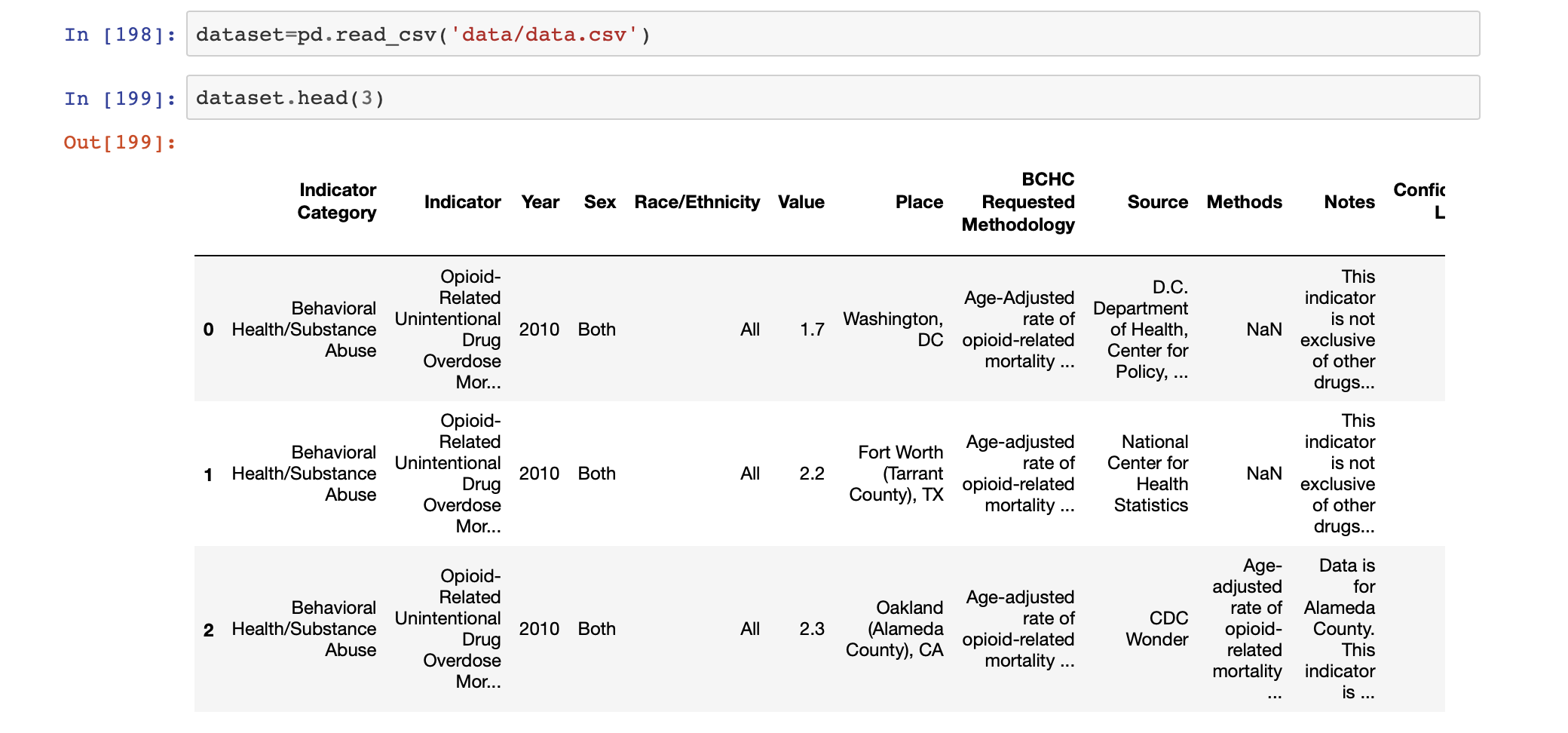
**Details of process:**

Following are the steps we followed for data analysis:

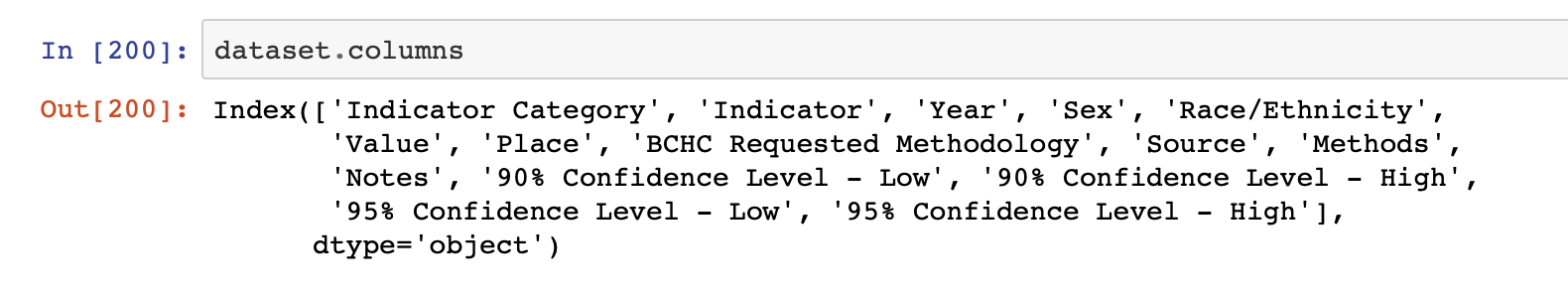
1. We imported the libraries NumPy, pandas and matplotlib which we will be using for data preprocessing and cleaning. We also use these to create arrays and read csv files. Using these libraries, we can plot graphs as well.



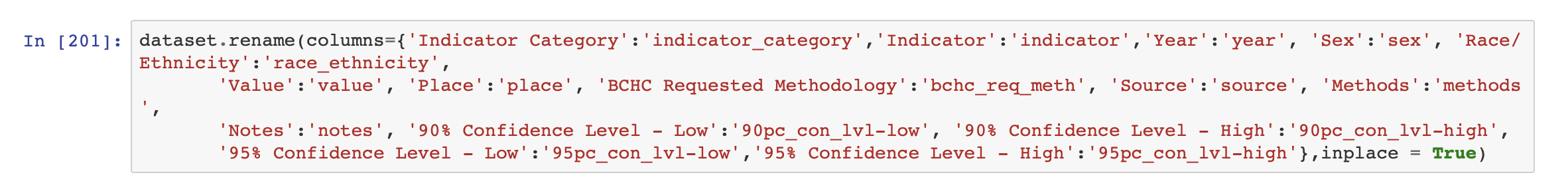
2. We loaded our CSV file into the dataset on which we will be performing the analysis and Data science techniques to provide the solution to our problems. The CSV file used is “data.csv”



Once the CSV is loaded into the Dataset we analyzed it starting with the column names and it appears that the column names contains white spaces which in longer run can create problems for us so we renamed the column names but before that we confirmed the column names by using pandas column function to list all the column.

****

Now we renamed the column names using the pandas rename function.

****

3. Our next step in data analysis is to filter out the values in “Indicator\_category” so that we get the actual data and the unused rows will be eliminated. First we will start working with the category named “Cancer”.

We filtered data using

Cancer\_ds = dataset.loc [dataset [“indicator\_category”] == “cancer”]

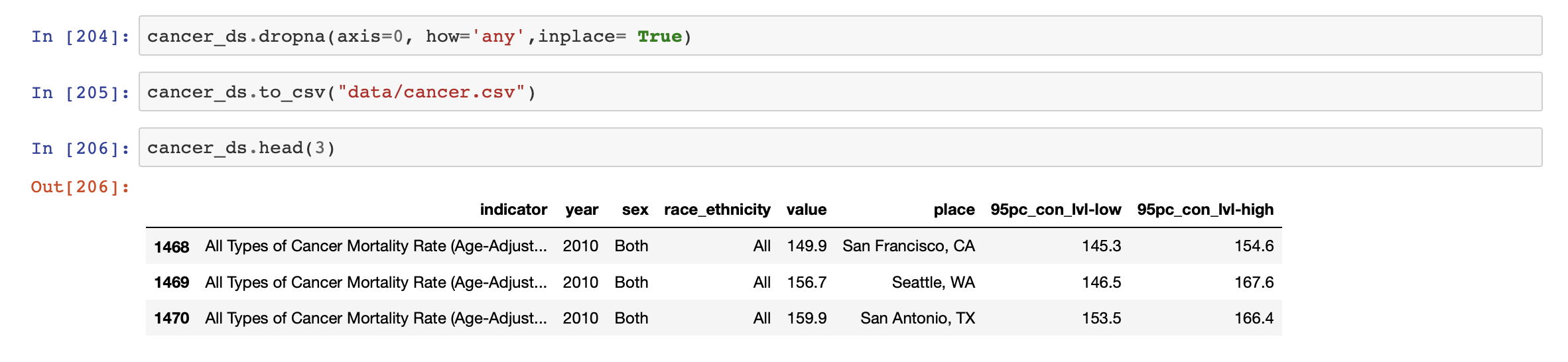
****

4. After filtering the data with the indicator category there were some rows and columns with no values, and it was unnecessary for us. Which should be removed from the data so that we get the proper data to perform our analyses. Hence, the next step is to remove empty columns and unnecessary columns.

We removed the empty rows and columns from the filtered dataset by using the:

****

5. Also there were some NaN or NA values in the data which should be removed. So we remove all rows having Nan and NA values which is achieved by :



And then we saved the data to a CSV file as a backup and to refer that later while plotting graphs.



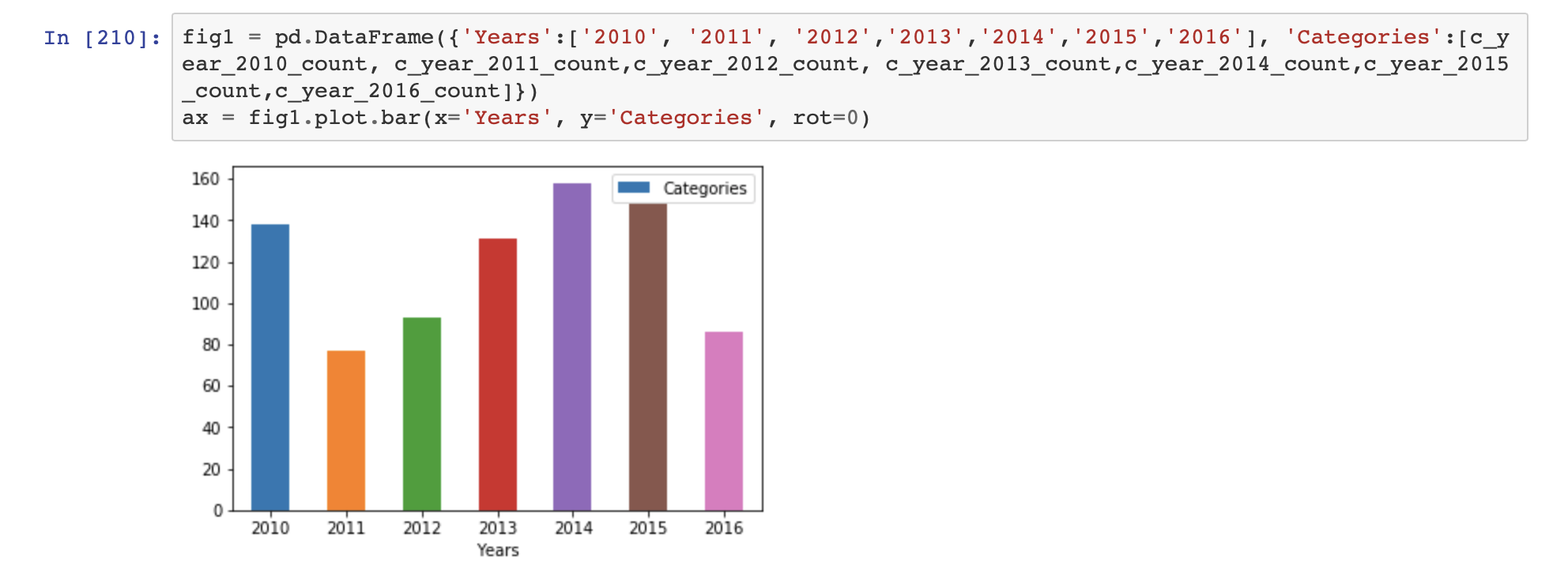
Then we Performed the above 4 steps for indicator categories which we will be using to do our analysis and prediction. The indicator categories which we will be using are “Cancer”, “Chronic Disease”,” HIV/AIDS”,” Infectious Disease”,” Injury/Violence”,” Life Expectancy and Death Rate (Overall)”

5. Once we have separated the data and followed the 4 steps for all categories we started to plot the graphs according to the indicator categories separately.

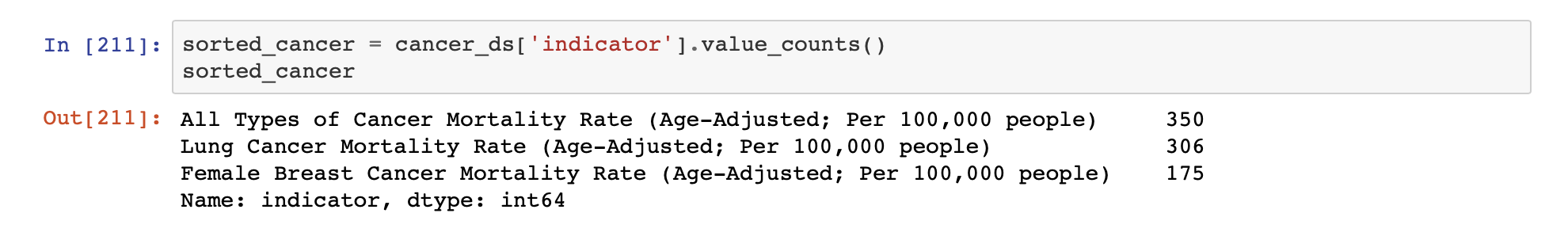
First we plotted the graph for Cancer as every year how many cancer patients are registered.

Using the following command.





Now we calculate the number of cases for each type of cancer. In order to do that we will group according to the indicator and take the count.

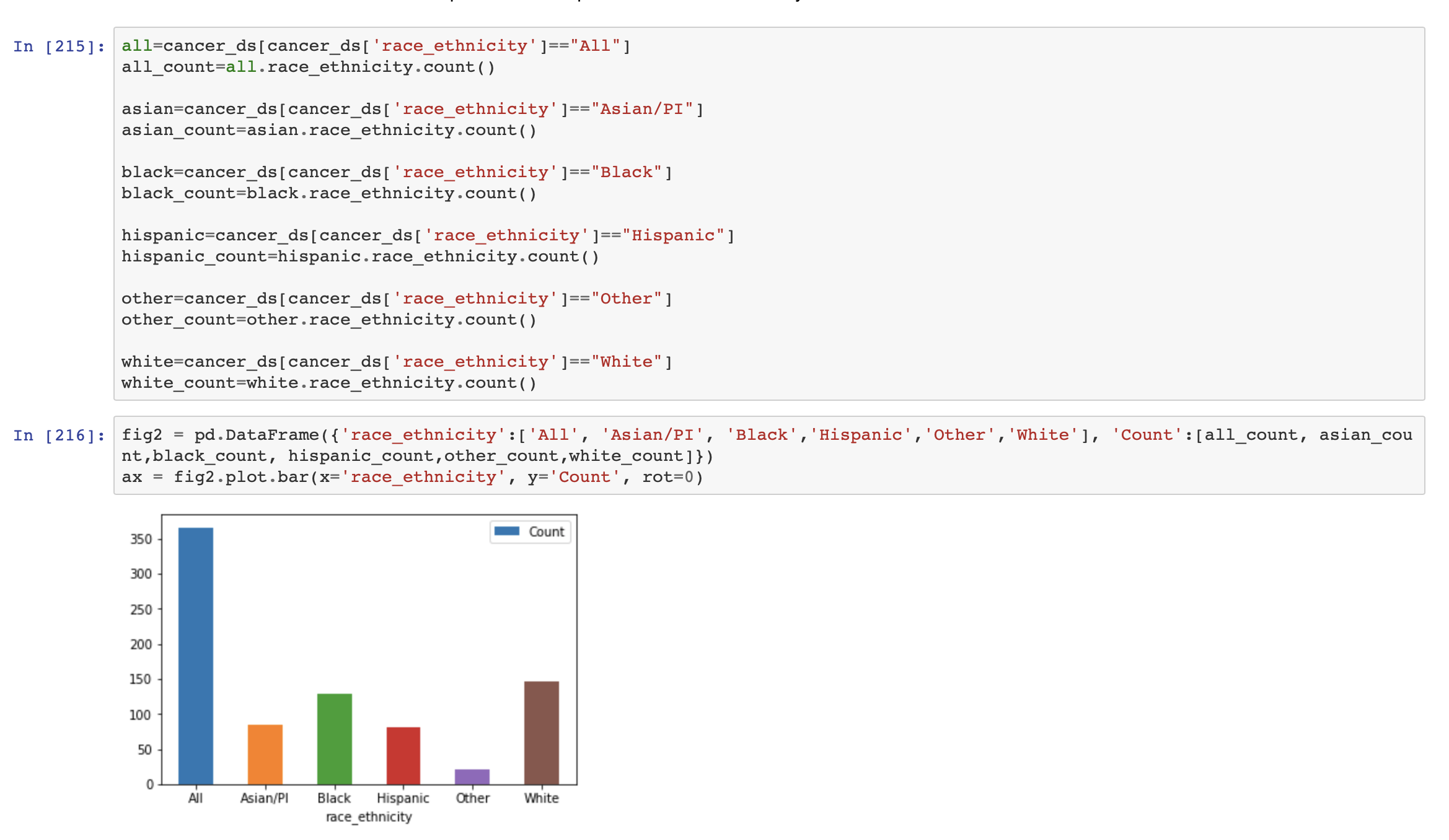


Then we plot a histogram to see :



Now we find out the distribution of cancer patients with respect to the race and ethnicity.

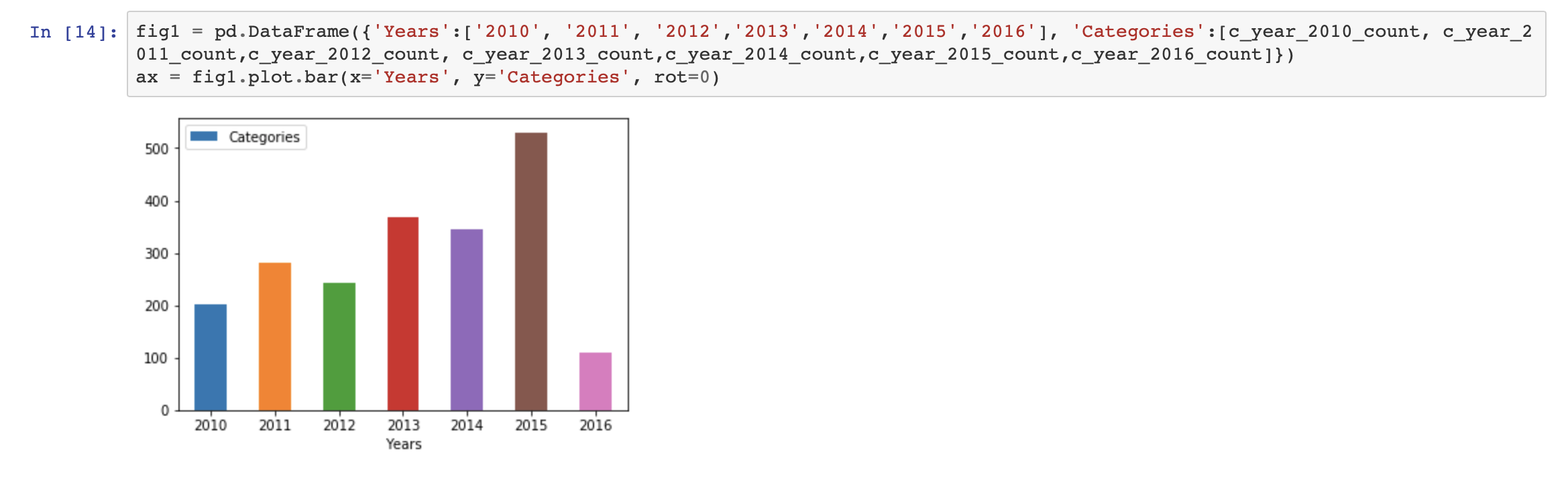
Below is the graph generated:



Similar process is followed for the remaining categories and below are the graphs generated:

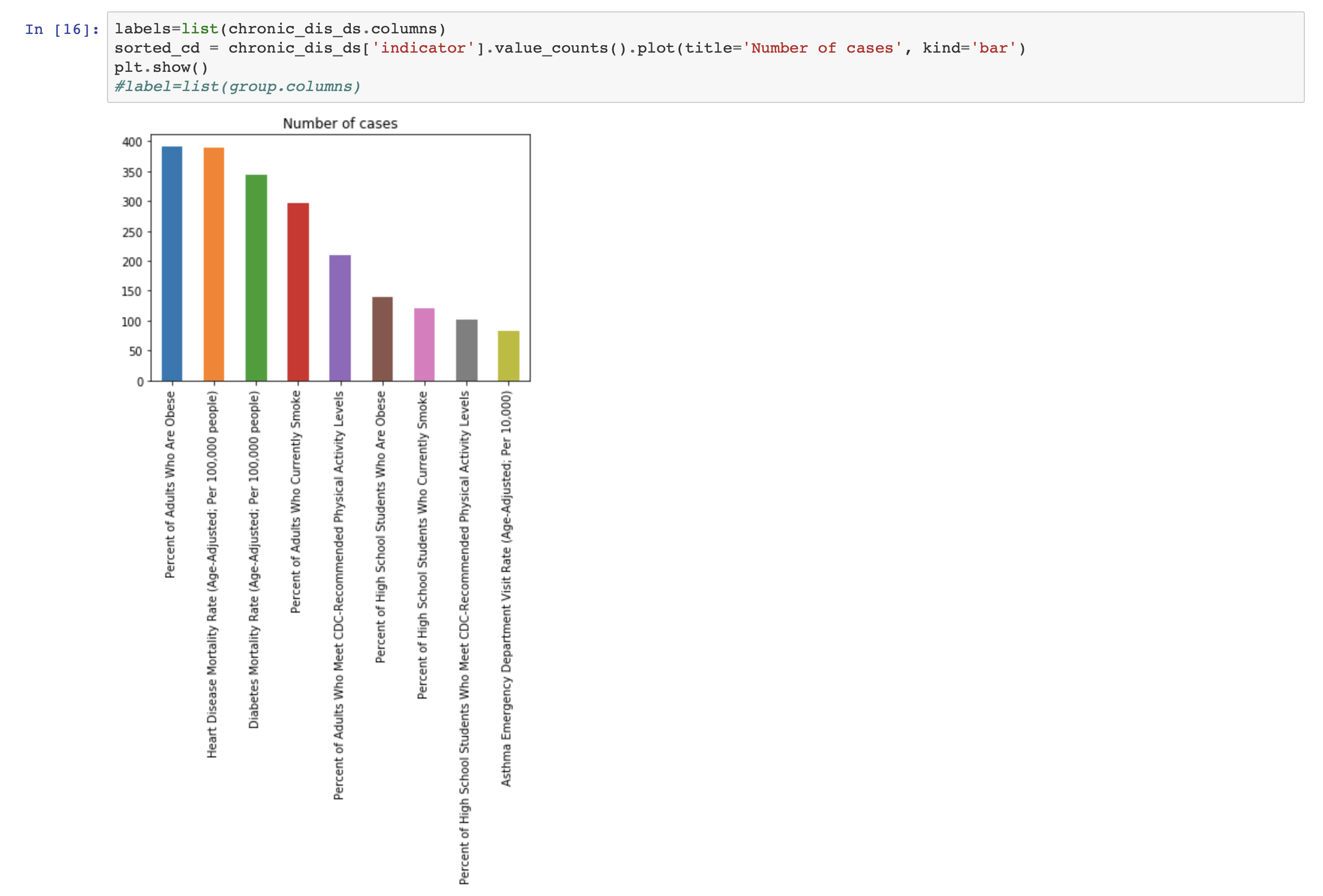
Chronic disease:

The below graph is generated for chronic disease patients during the year 2010, 2011, 2012, 2013, 2014, 2015 and 2016.



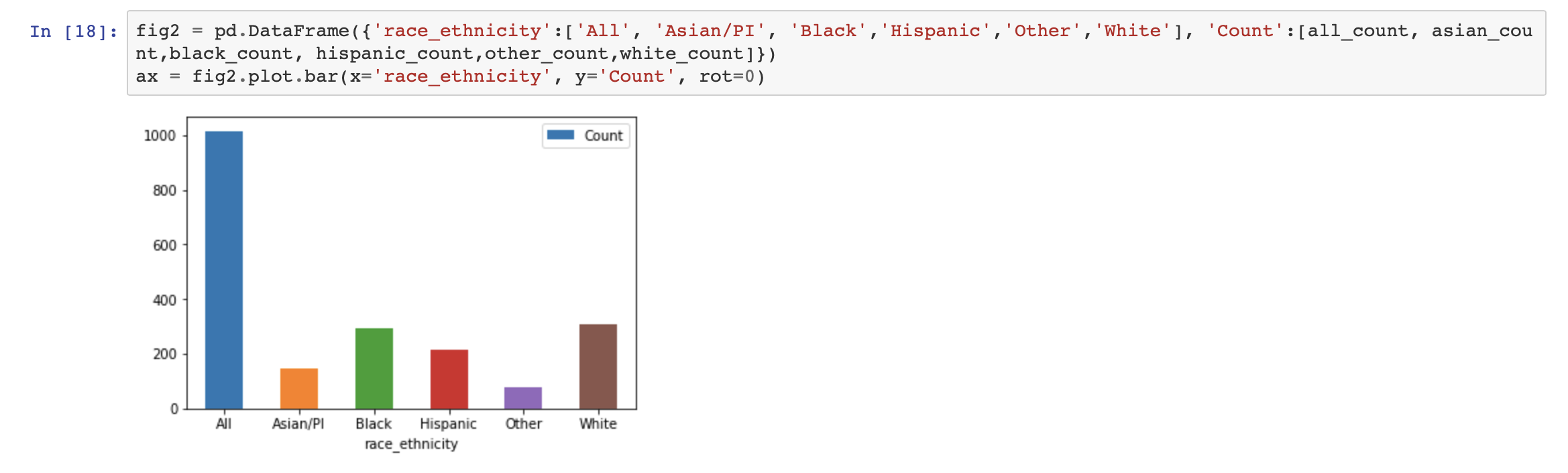
Now we calculate the number of cases for each type. In order to do that we will group according to the indicator and take the count.

Below is the histogram plotted:



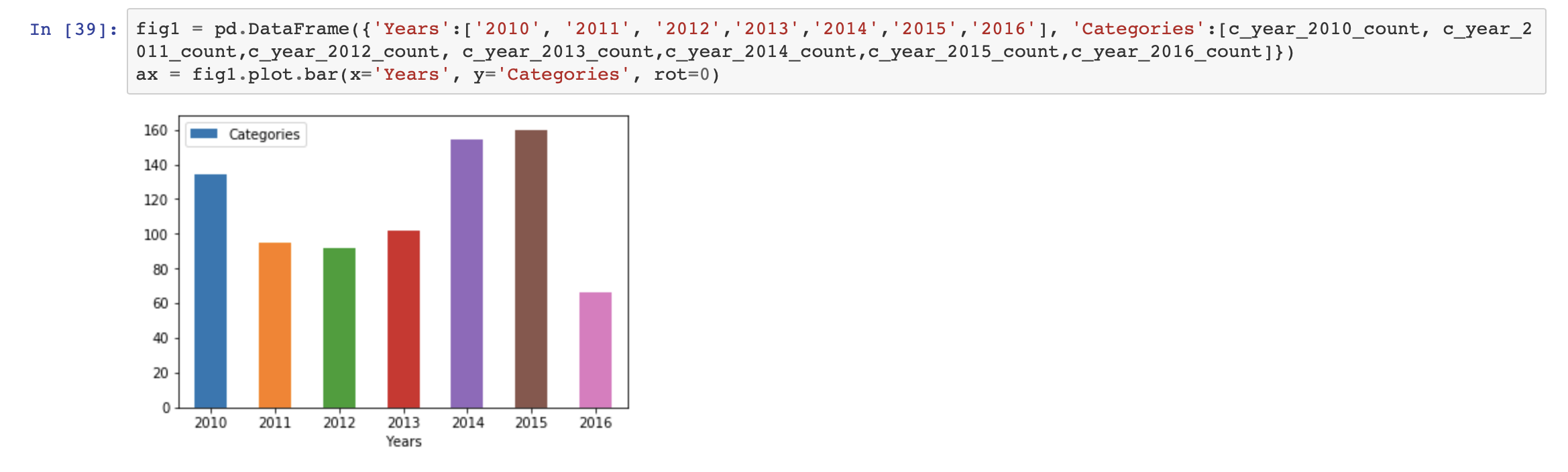
Now we find out the distribution of chronic disease patients with respect to the race and ethnicity.

Below is the graph generated:

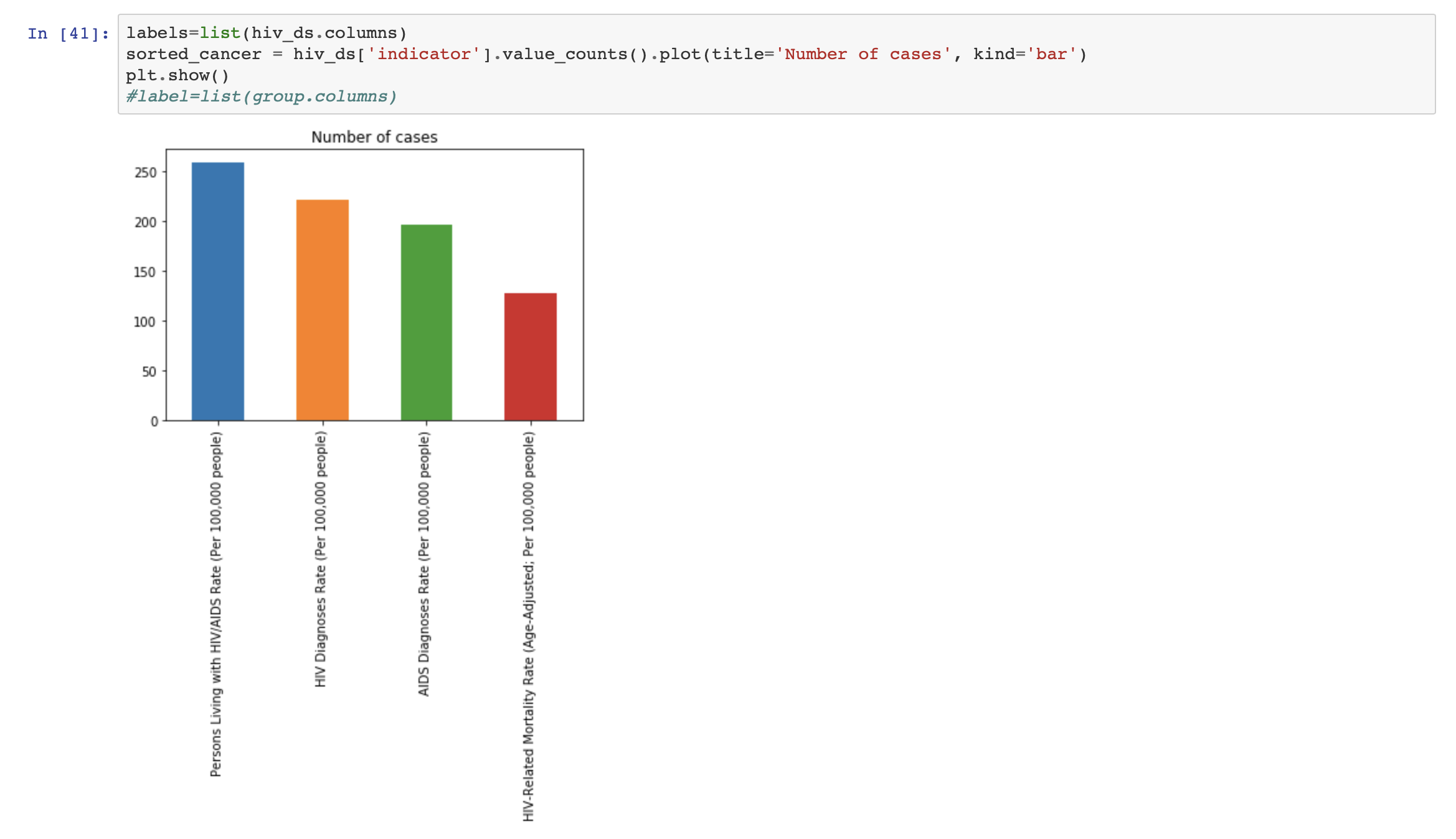


HIV/Aids:

The below graph is generated for HIV/Aids patients during the year 2010, 2011, 2012, 2013, 2014, 2015 and 2016.

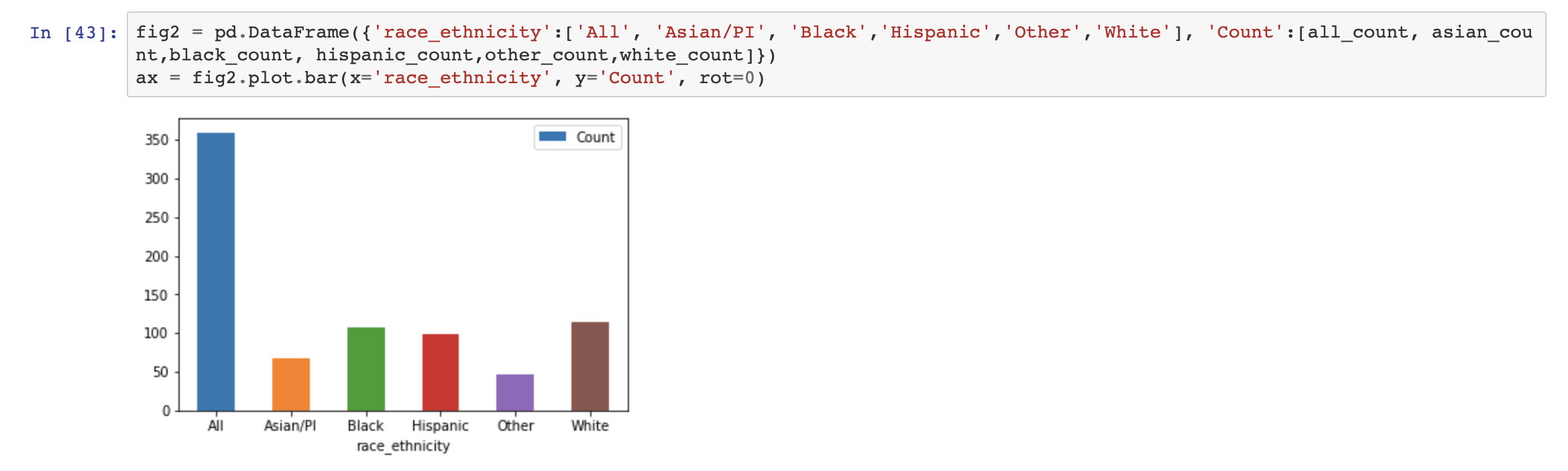


Now we calculate the number of cases for each type. In order to do that we will group according to the indicator and take the count.

Below is the histogram plotted:

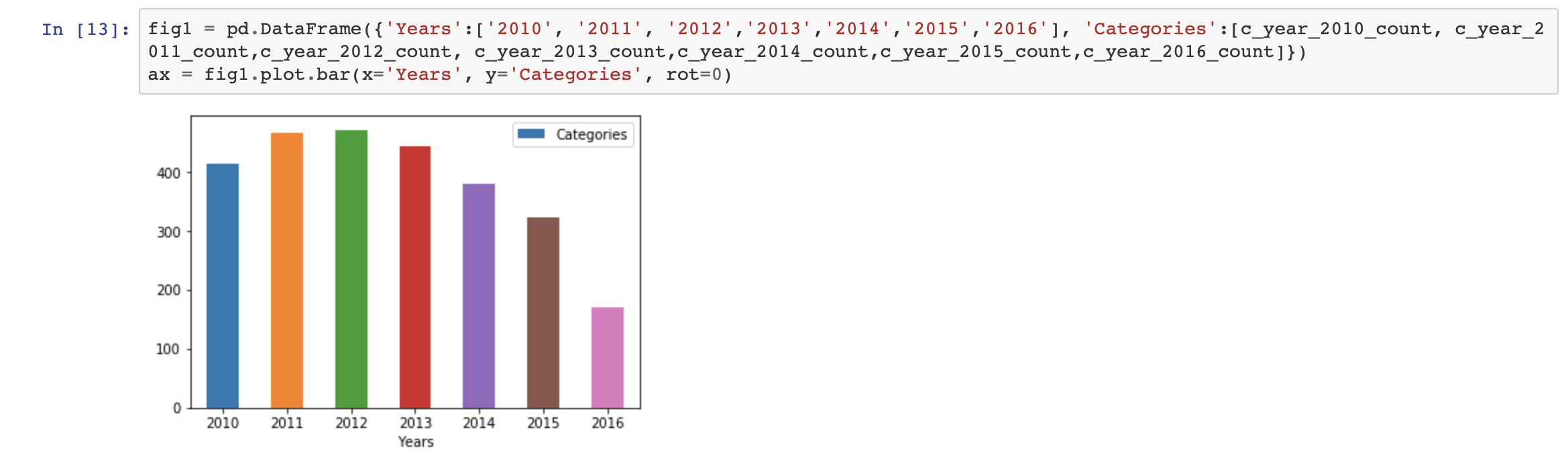
Now we find out the distribution of chronic disease patients with respect to the race and ethnicity.

Below is the graph generated:

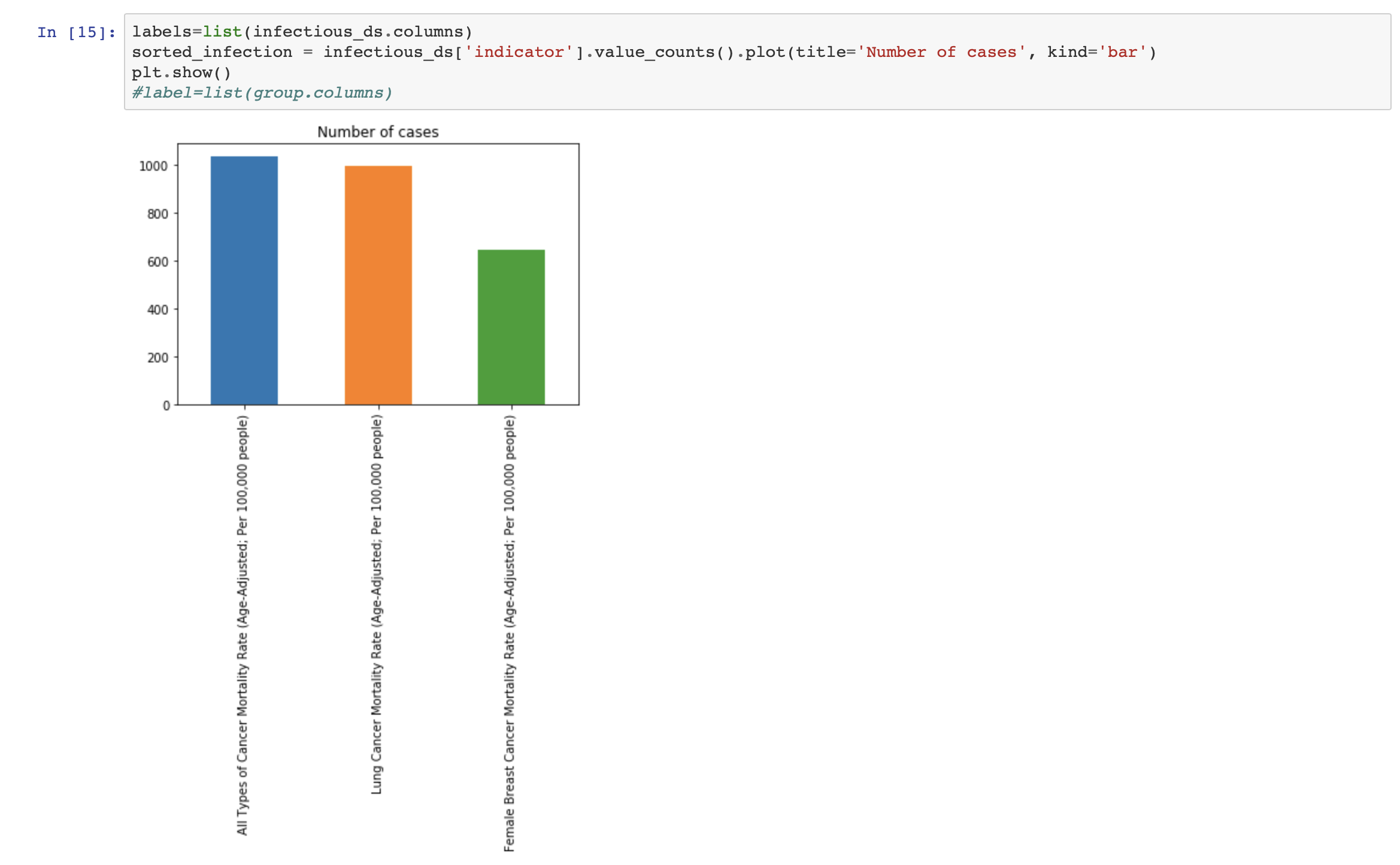


Infectious disease:

The below graph is generated for infectious disease patients during the year 2010, 2011, 2012, 2013, 2014, 2015 and 2016.

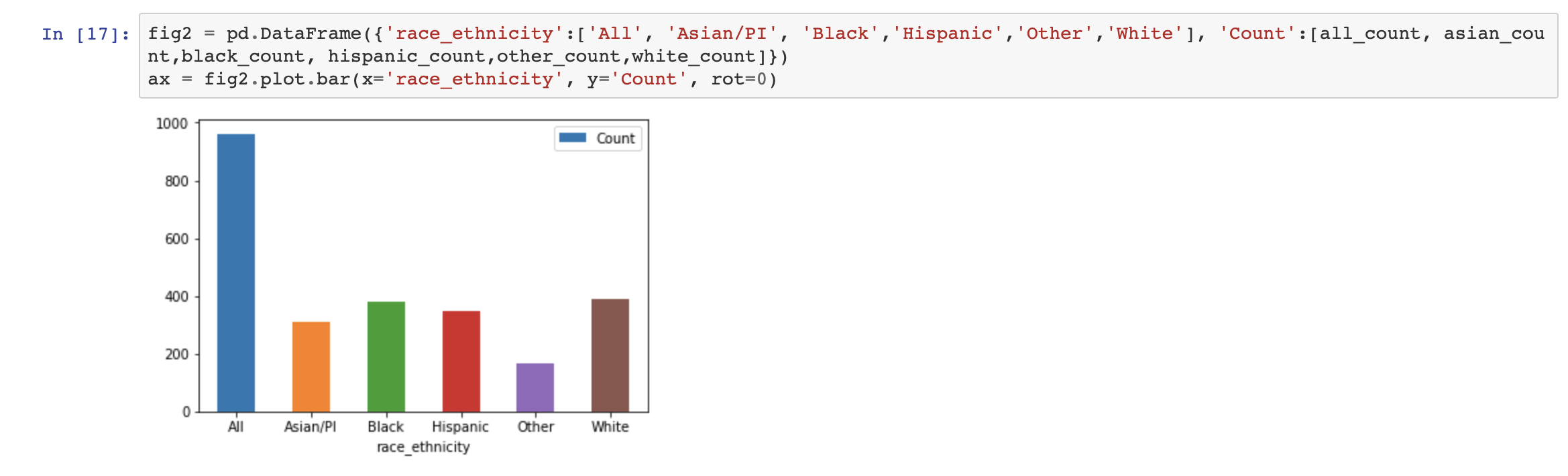


Now we calculate the number of cases for each type. In order to that we will group according to the indicator and take the count.

Below is the histogram plotted 

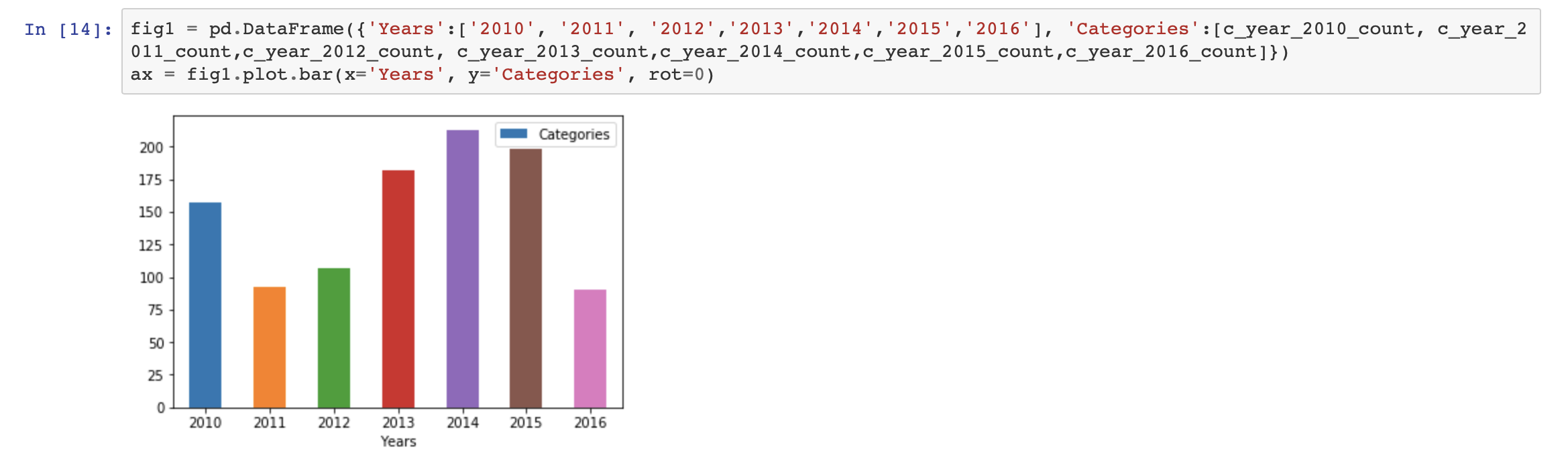
Now we find out the distribution of infectious disease patients with respect to the race and ethnicity.

Below is the graph generated:

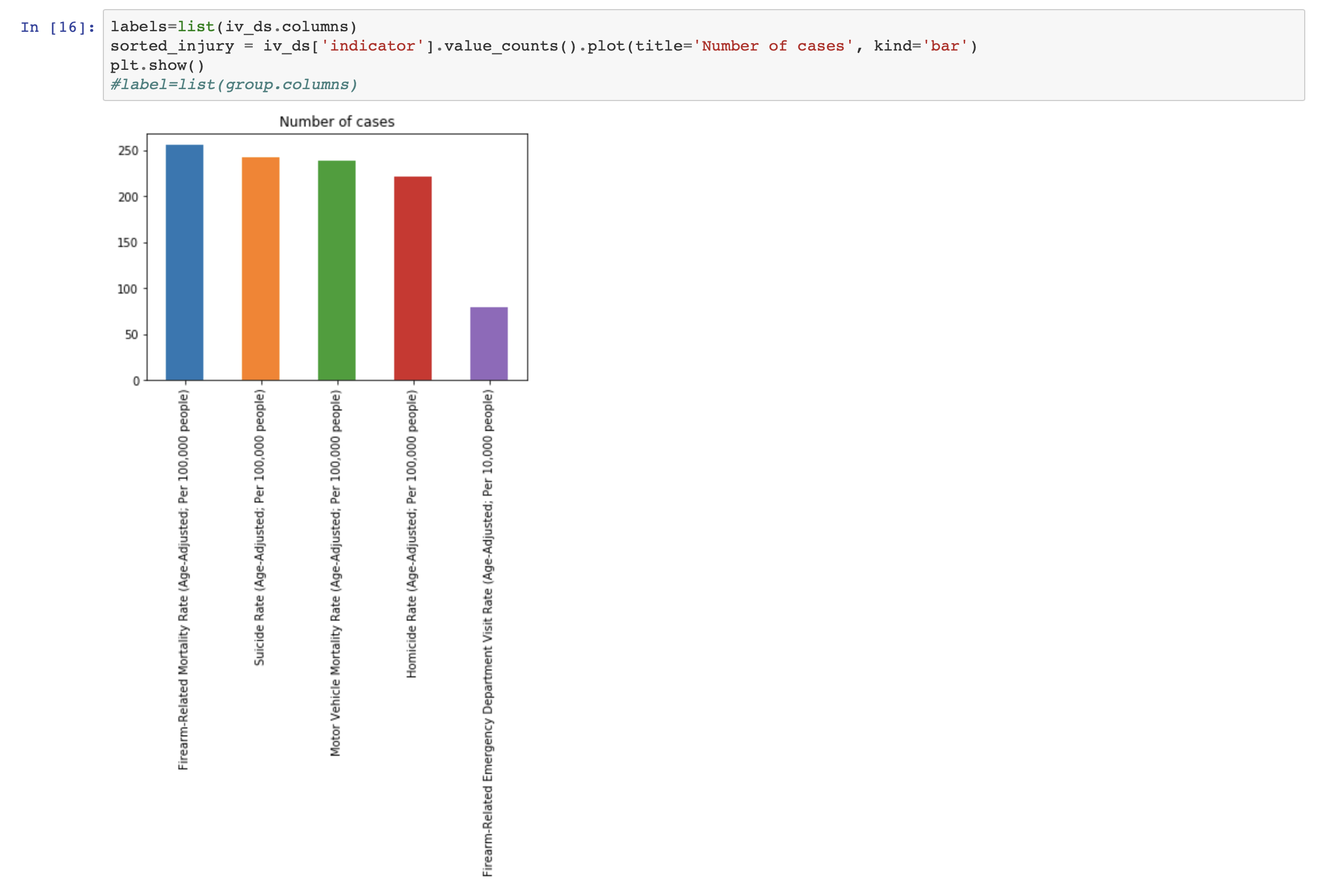


**Injury and Violence:**

The below graph is generated for injury and violence patients during the year 2010, 2011, 2012, 2013, 2014, 2015 and 2016.

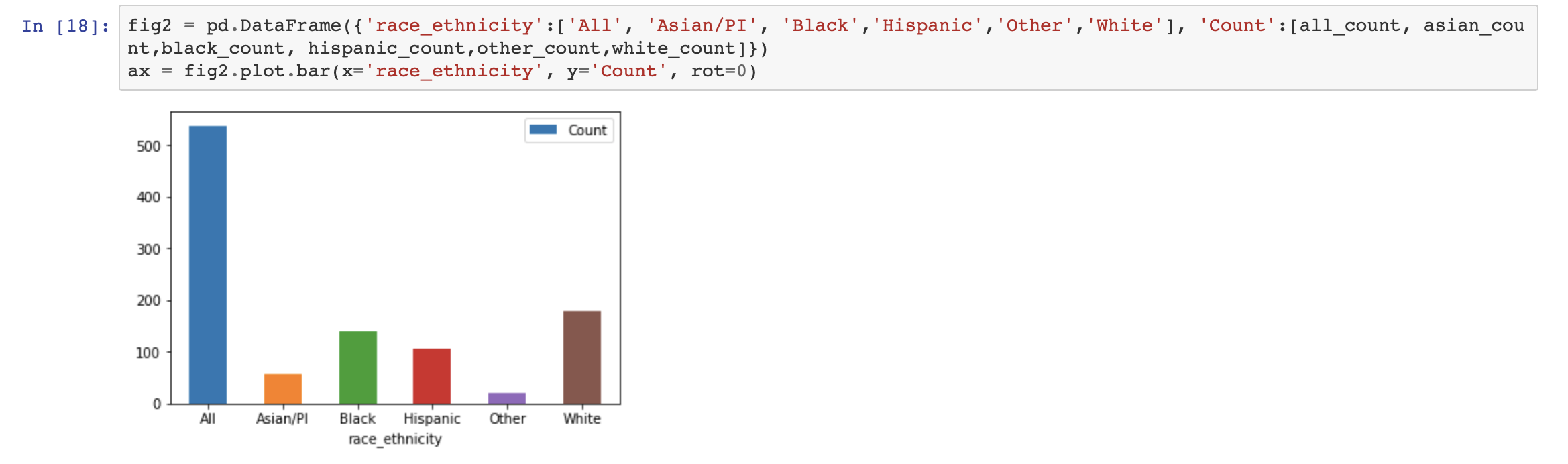
****

Now we calculate the number of cases for each type. In order to do that we will group according to the indicator and take the count.

Below is the histogram plotted 

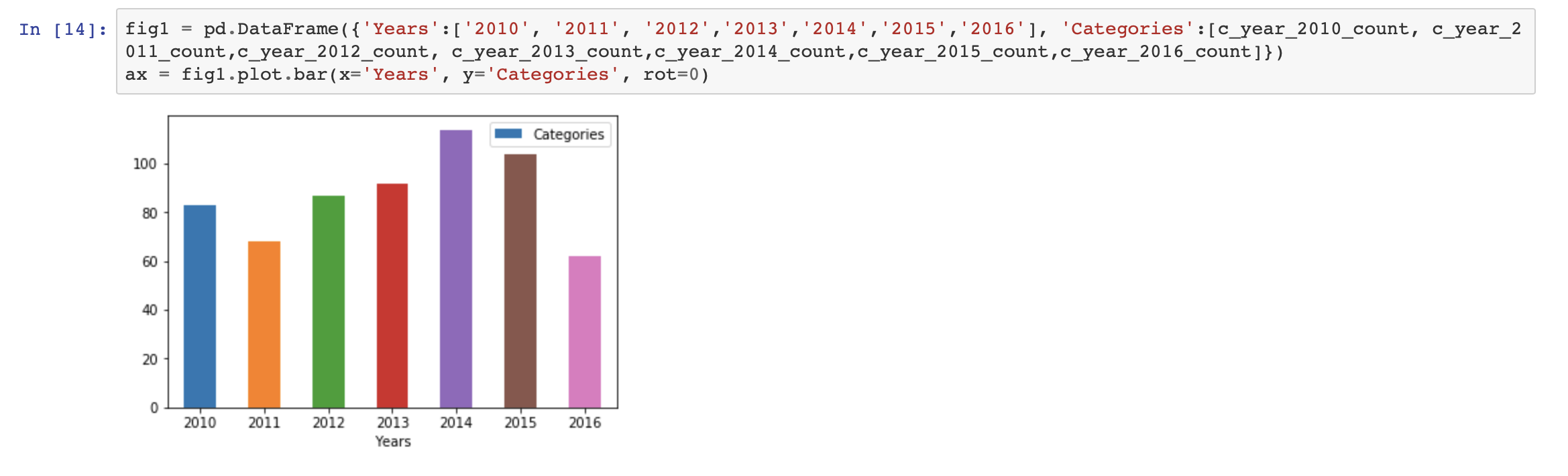
Now we find out the distribution of injury and violence disease patients with respect to the race and ethnicity.

Below is the graph generated:

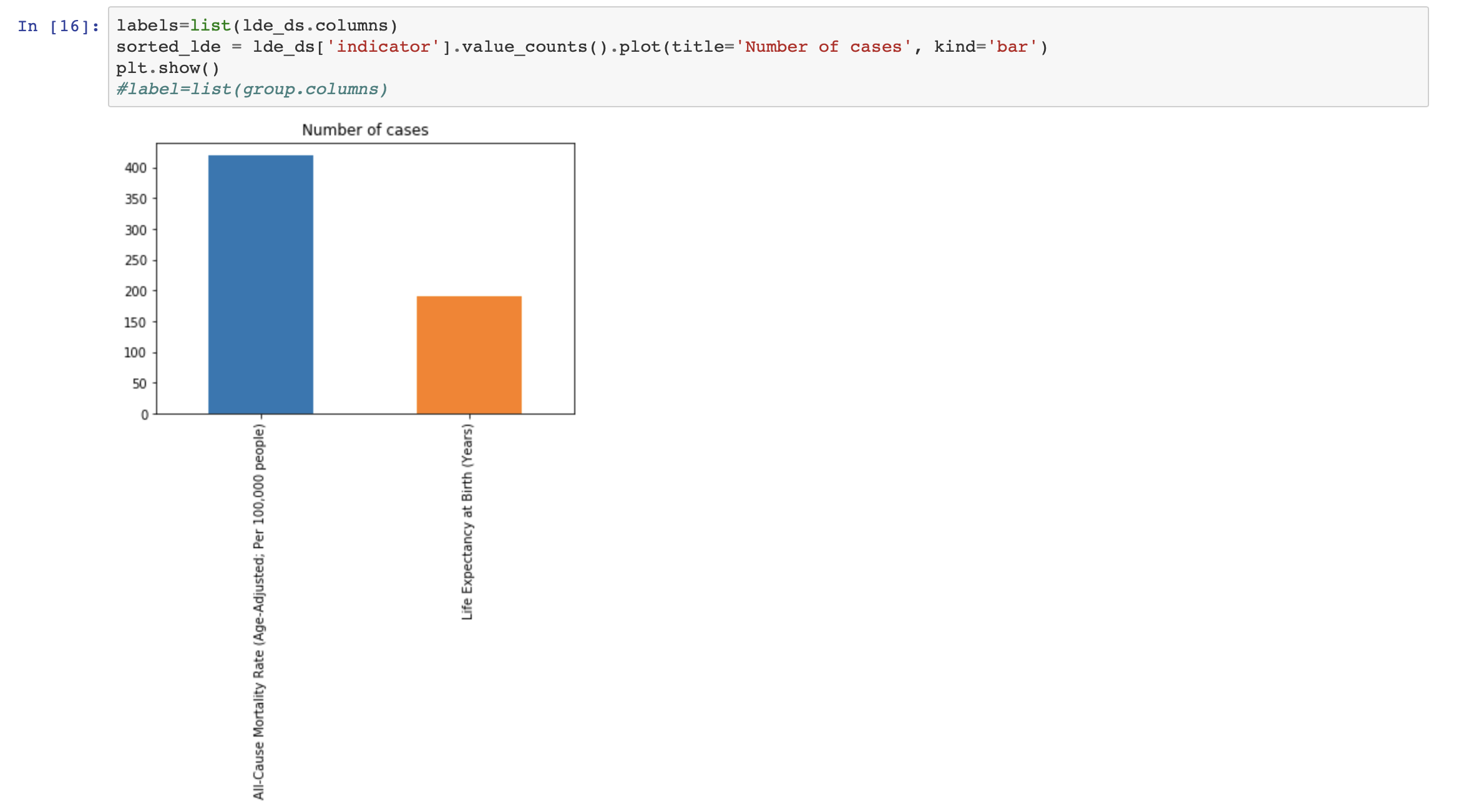


Life death expectancy:

The below graph is generated for life death expectancy patients during the year 2010, 2011, 2012, 2013, 2014, 2015 and 2016.

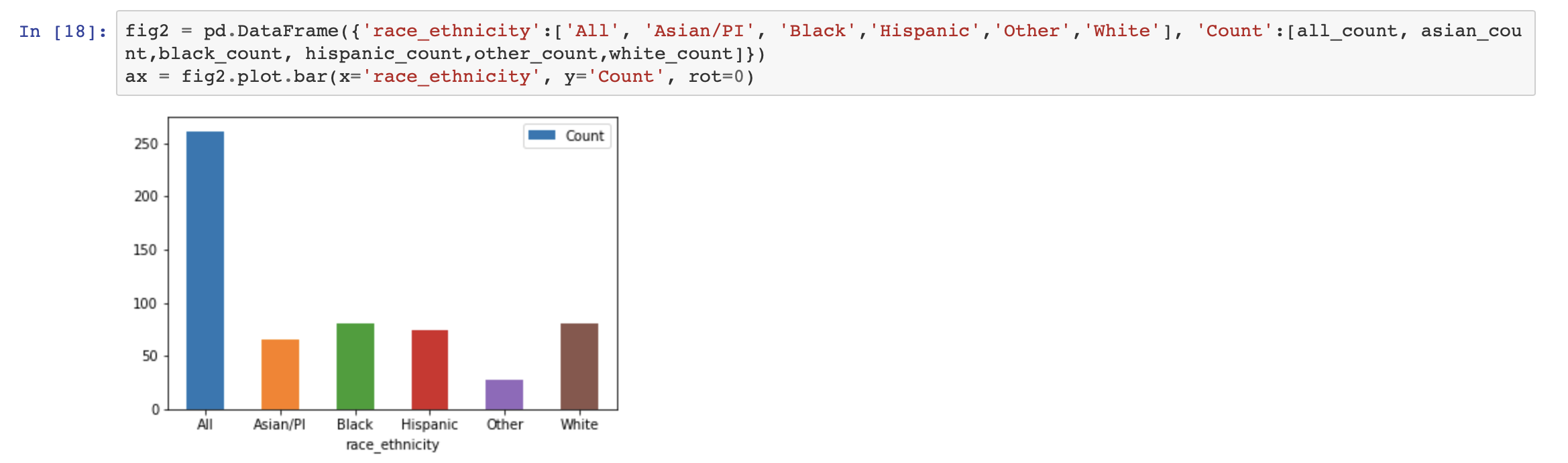


Now we calculate the number of cases for each type. In order to do that we will group according to the indicator and take the count.

Below is the histogram plotted 

Now we find out the distribution of life death expectancy patients with respect to the race and ethnicity.

Below is the graph generated:



**Challenges:**

1. The first challenge was to separate the data from the generalized data set.
2. The second challenge was to clean the data successfully so that it can be further used for data analysis.
3. Third challenge was to provide the findings that we had discovered within our data after the clean data is generated. So, we plotted some graphs for the same.
4. Another challenge was to show the relationship between these features.
5. The main challenge was the application of any machine learning algorithm to the clean and processed data set in order to train the algorithm.

**Success:**

1. We successfully cleaned and processed the data to get separate data for each category with proper values in each column.
2. After successfully cleaning the data we analyzed each indicator category rate with respect to race, place, year and we also calculated the number of cases for each sub indicator within the data. Then we also considered the sub indicator category for all indicators to analyze the count of number of cases within the sub indicator.
3. We successfully plotted the graphs in each category for visual representation.
4. We used the generated data in each category in order to predict the modules.
5. We successfully presented the analysis in a statistical format so that it enables a research to quickly navigate through the massive amount of data and easily understand it hence saving up important time.
6. Since it is now easy for researchers to understand the clean and analyzed data the cure of such disease can be available and developed more easily and at a faster rate than before.

**Possible extension of project:**

In future in this project with the help of this data we can perform the prediction of successful methods which were used for treatment and which were more successful.

With the help of machine learning and data science we can develop and train more efficient module using this data to help researchers to perform more deep analysis and predictions.

**Conclusion:**