CS 6635: Visualization of Scientific Data Homework 1

Generating, Using Datasets and Plotting

Yash Gangrade (u1143811), MS First Year, School of Computing 1st February 2017

Contents

1 Part I - Generating your own data 2

1.1 Create an array with 100 elements from 1 to 100 in order: Create a box plot to visualization your data. 2

1.2 Create an array with 10,000 random numbers. Create a histogram of the data using 20 bins. 3

1.3 Write a program to generate 100 random numbers uniformly distributed between 1 and 100. Write the numbers out to a binary file and use a line graph to draw the 100 numbers. 4

1.4 Write a program to read the binary file back, divide the range between 1 and 100 into 7 intervals, and calculate the frequency for each interval: display a histogram of your result. 5

2 Part II - Interesting Datasets 7

2.1 Download the NOAA Land Ocean Temperature Anomalies Data Set: link. Create a bar plot of the data. Include a label called "Year" along the x-axis and a label called Degrees F +/- From Average along the y-axis. Describe trends in the data. 7

2.2 Download the member of Congress by Age data set: link. Create a Star Plot of the data and create a Parallel Coordinates Plot of the data. Describe the trends in the data. . . 8

2.3 Download the U.S. Birth data set: link. What day of the month had the highest number of births? What day of the month had the lowest number of births? Are there any interesting trends in the data, i.e. more births in Summer or Winter? What about births on Friday the 13th? 10

2.4 The U.S. Government maintains a sever with many interesting datasets called Data.Gov: https://www.data.gov/. Choose 3 different data sets to visualize. Visualize the data sets in at least 3 different ways. Describe the trends in the data. 11

1 Part I - Generating your own data

1.1 Create an array with 100 elements from 1 to 100 in order: Create a box plot to visualization your data.

Ans In this question and for rest of the questions, I am using the numpy and matplotlib library, in some cases, I am also using some extra libraries.
For this question, the *arange* function in numpy library is used to create the array. Then, boxplot function defined in *pyplot (matplotlib)* to create the boxplot. The procedure is similar in MATLAB (Please find in the submitted codes). The diagrams from both simulations are attached below. As expected the mean of box plot lies near 50.

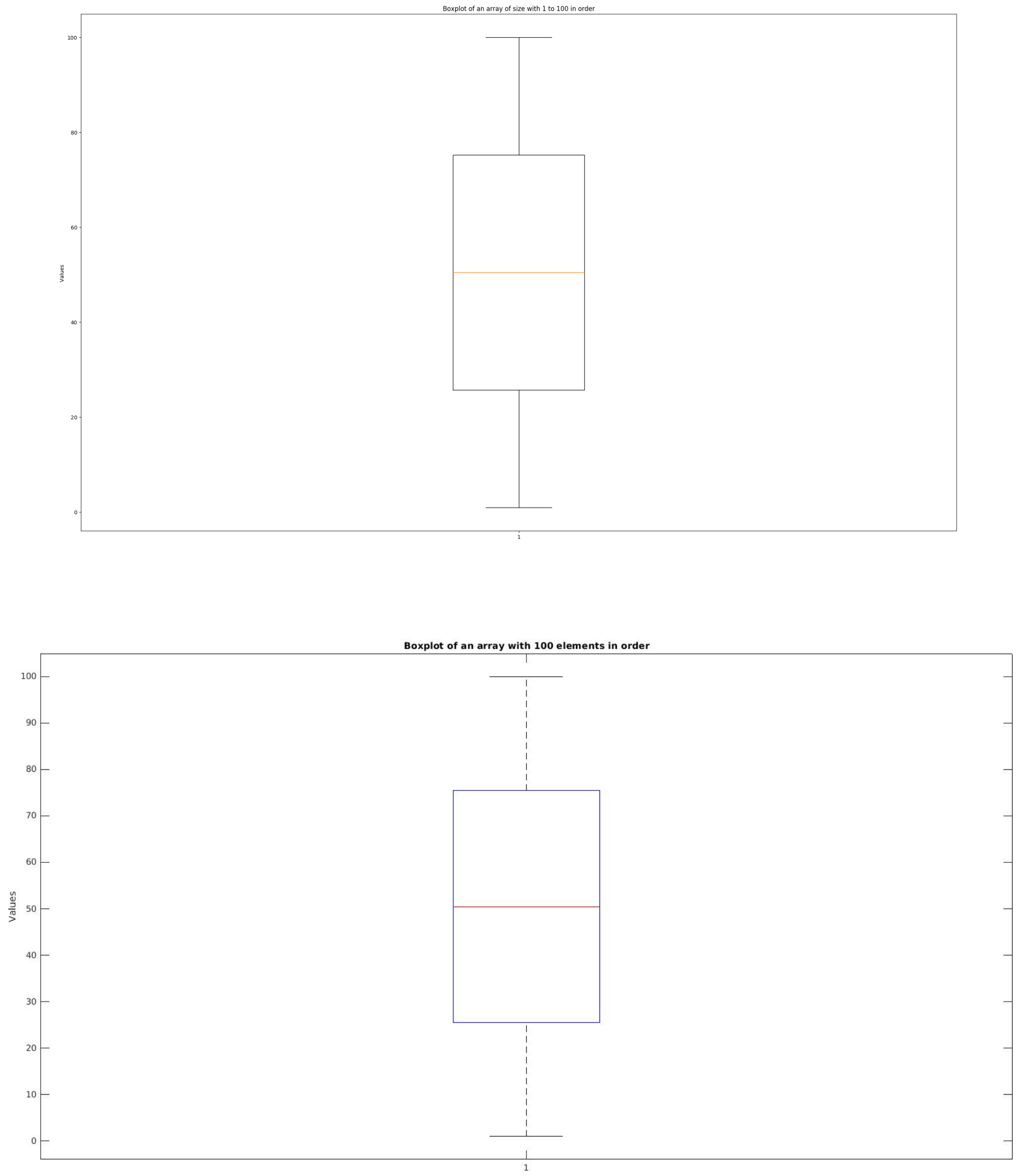


Figure 1: Box plot through Python (top) and MATLAB (bottom)

1.2 Create an array with 10,000 random numbers. Create a histogram of the data using 20 bins.

Ans Here, the rand() function built in numpy and in MATLAB are used to generate an array of 10000 random numbers coming from uniform distribution (*Univ(0,1)*). Similar to the above procedure, we use the *histogram* function defined in matplotlib and MATLAB.

Note: Whenever you run the code, you will get different data thus, different histograms.

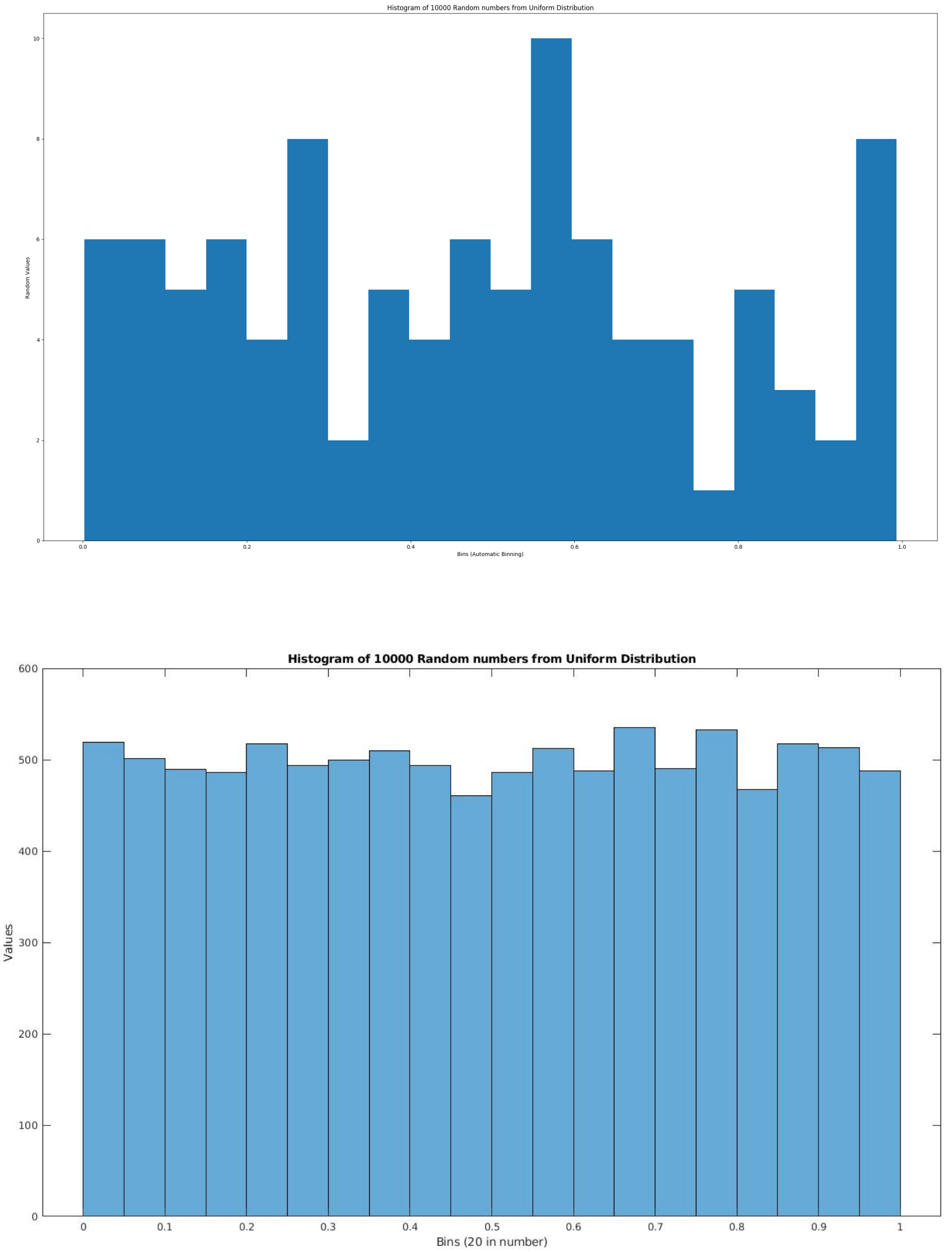
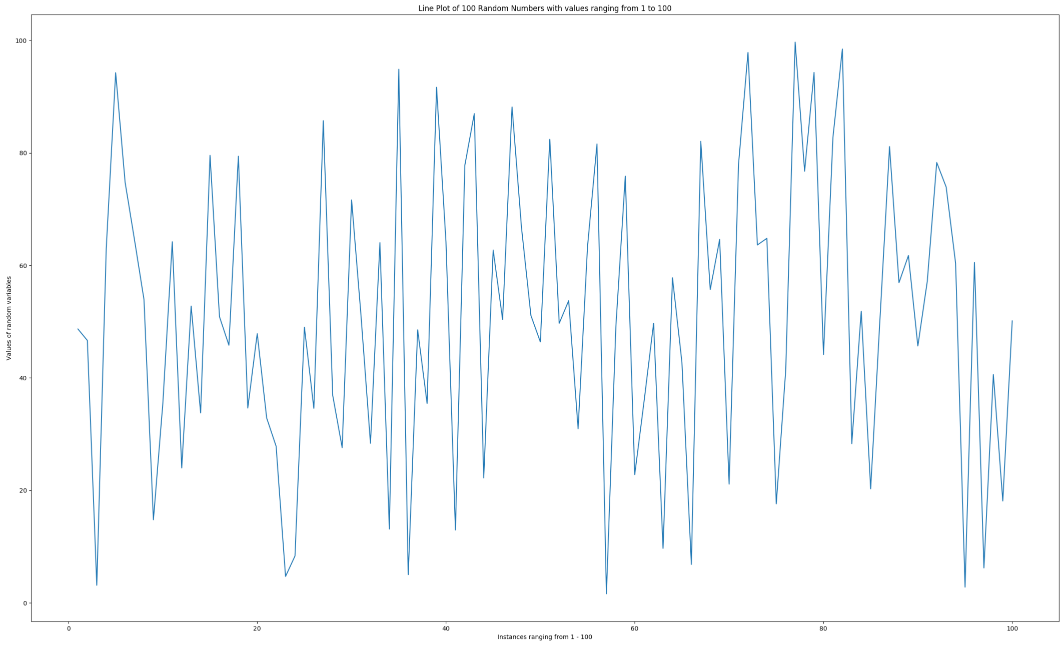


Figure 2: Histogram created through Python (top) and MATLAB (bottom)

1.3 Write a program to generate 100 random numbers uniformly distributed between 1 and 100. Write the numbers out to a binary file and use a line graph to draw the 100 numbers.

Ans Similar to the previous question, we can manipulate a function to produce an array that contains random numbers between 1 to 100. Please refer to the code (in MATLAB and Python) to see the exact procedure. In case of python, I have used a special library/package 'pickle' to make the write and read easier. In MATLAB, I have just used in-build fwrite and fread functions to write and read the binary file. Lastly, matplotlib and MATLAB have a *plot* function defined which does the work for you. The results are attached below:



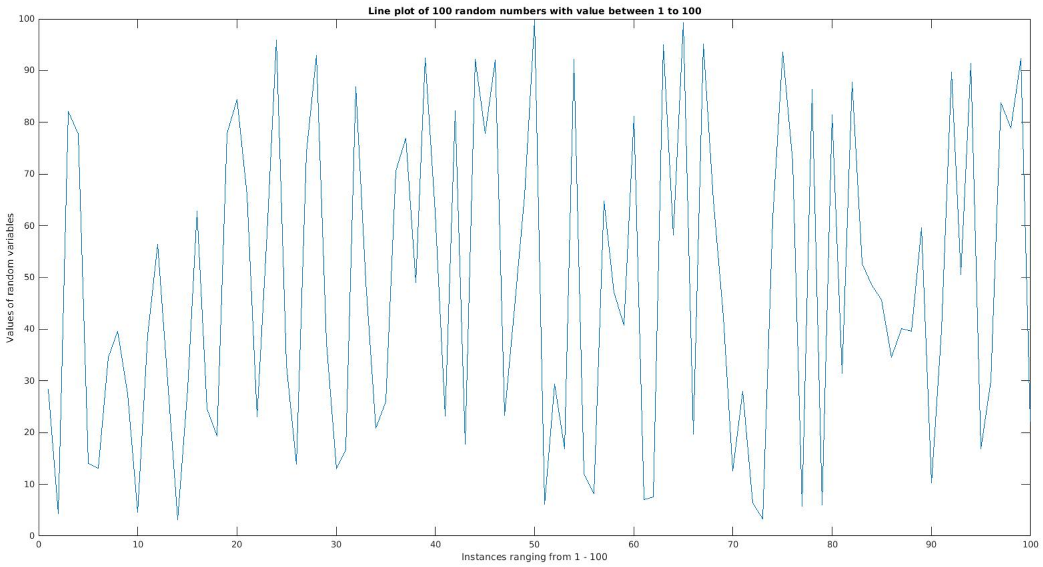
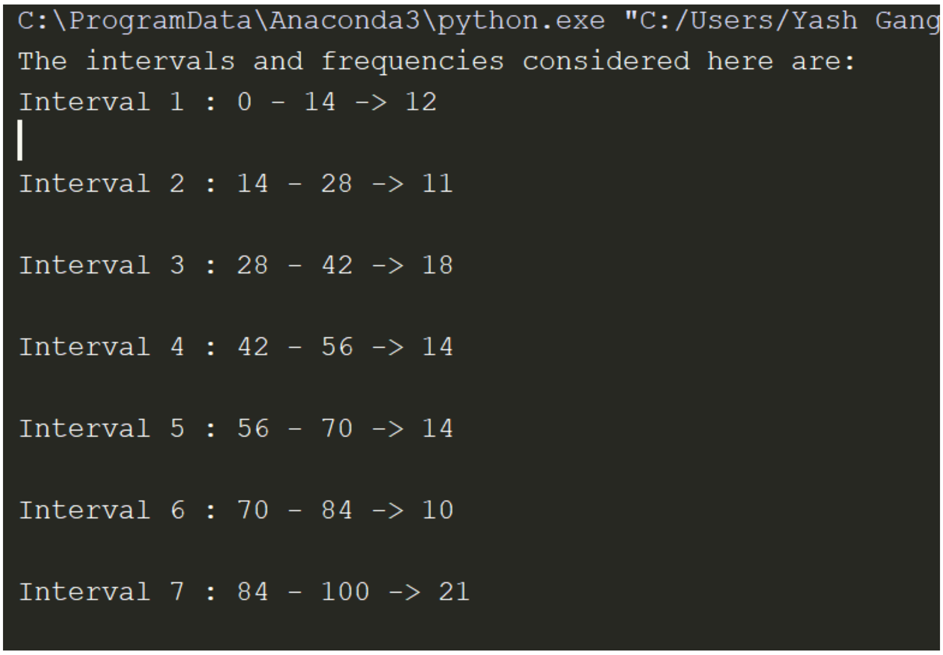


Figure 3: Line graphs created through Python (top) and MATLAB (bottom)

1.4 Write a program to read the binary file back, divide the range between 1 and 100 into 7 intervals, and calculate the frequency for each interval: display a histogram of your result.

Ans In python, I am using the pickle library to read the binary file back and fread function in MATLAB. Then we are specifying the bin array in the histogram method of matplotlib (python) and in MATLAB. The results are attached as follows:



```
C:\ProgramData\Anaconda3\python.exe "C:/Users/Yash Gang
The intervals and frequencies considered here are:
Interval 1 : 0 - 14 -> 12

Interval 2 : 14 - 28 -> 11

Interval 3 : 28 - 42 -> 18

Interval 4 : 42 - 56 -> 14

Interval 5 : 56 - 70 -> 14

Interval 6 : 70 - 84 -> 10

Interval 7 : 84 - 100 -> 21
```

Figure 4: Command Line Output

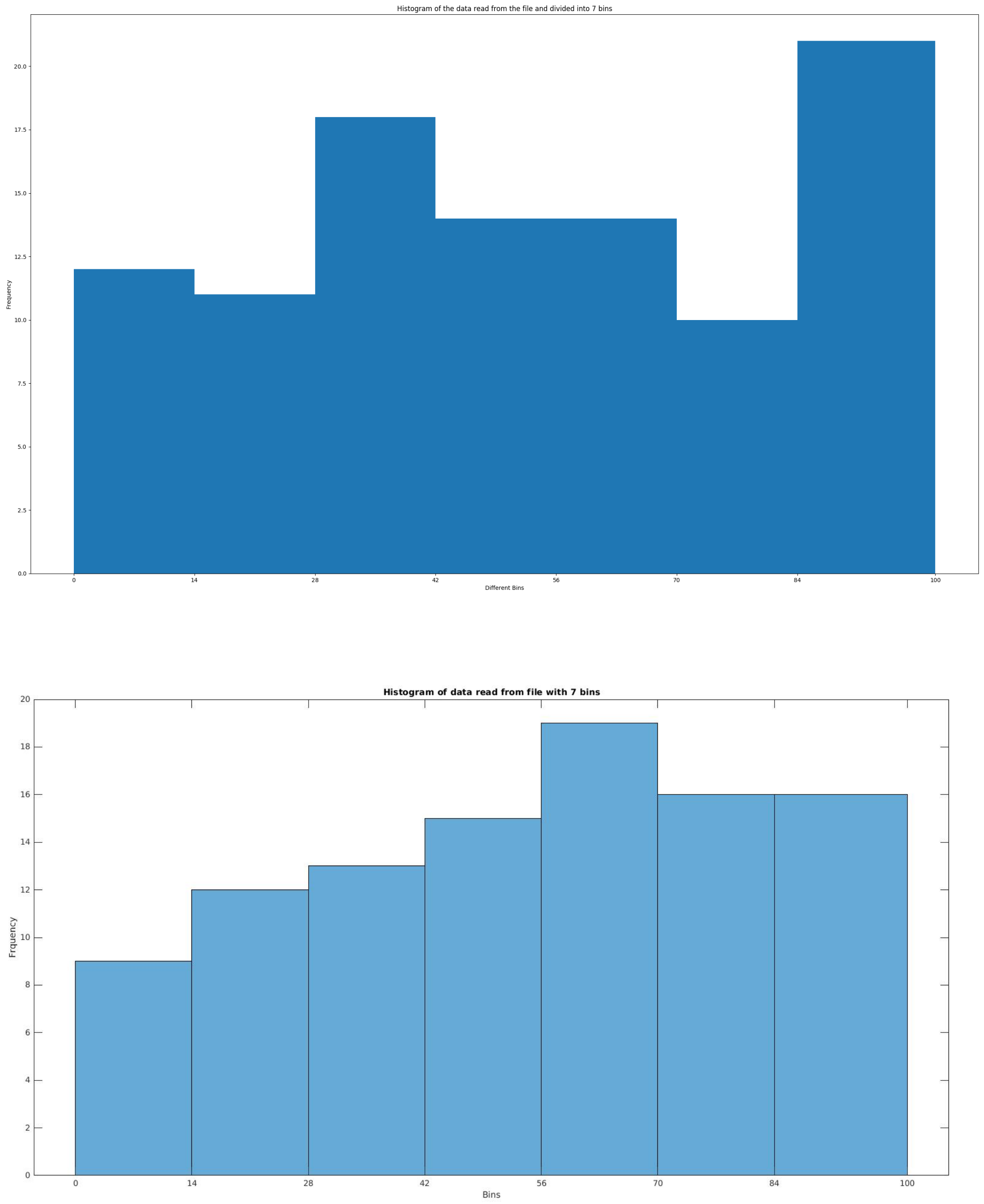
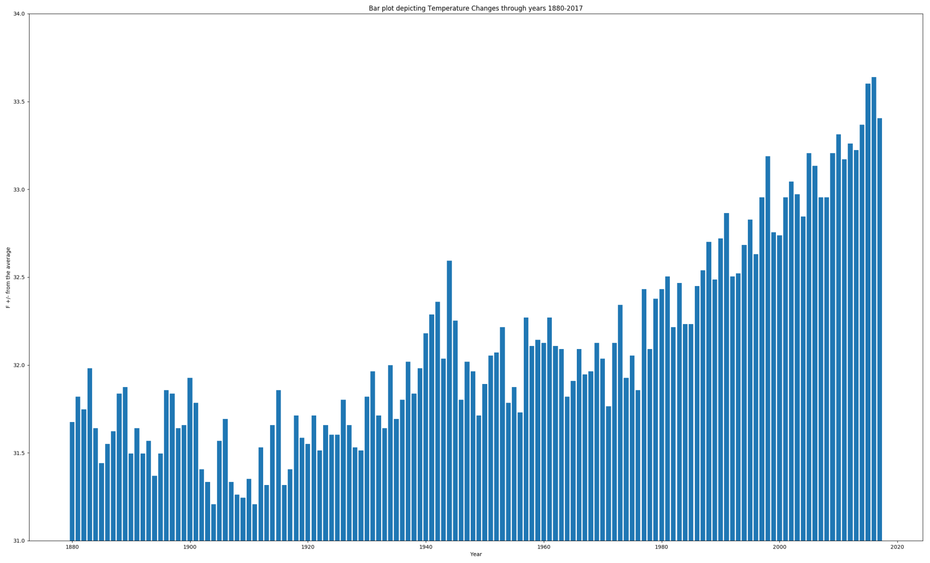


Figure 5: Histograms created through Python (top) and MATLAB (bottom)

2 Part II - Interesting Datasets

2.1 Download the NOAA Land Ocean Temperature Anomalies Data Set: link. Create a bar plot of the data. Include a label called "Year" along the x-axis and a label called Degrees F +/- From Average along the y-axis. Describe trends in the data.

Ans Here we had a temperature vs year data in a csv format. For Python, I am using csv class which has an in-built reader function. I am then storing data into different variables like year and value. Finally, I am converting the values into degree farhenhiet because they are in celsius form in the raw form. Then, the plotting is done through the bar function implemented in matplotlib and MATLAB. The results of the experiments are as follows:



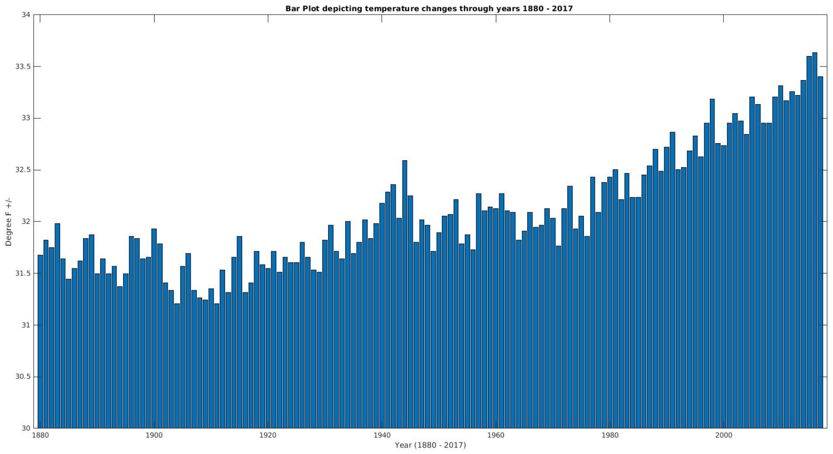


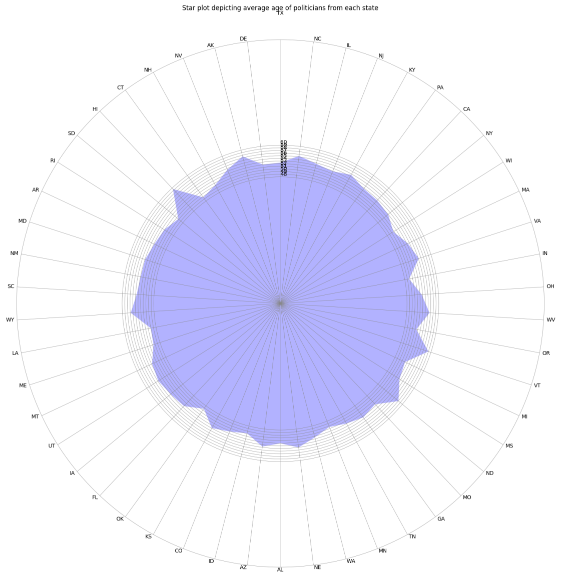
Figure 6: Bar plots created through Python (top) and MATLAB (bottom)

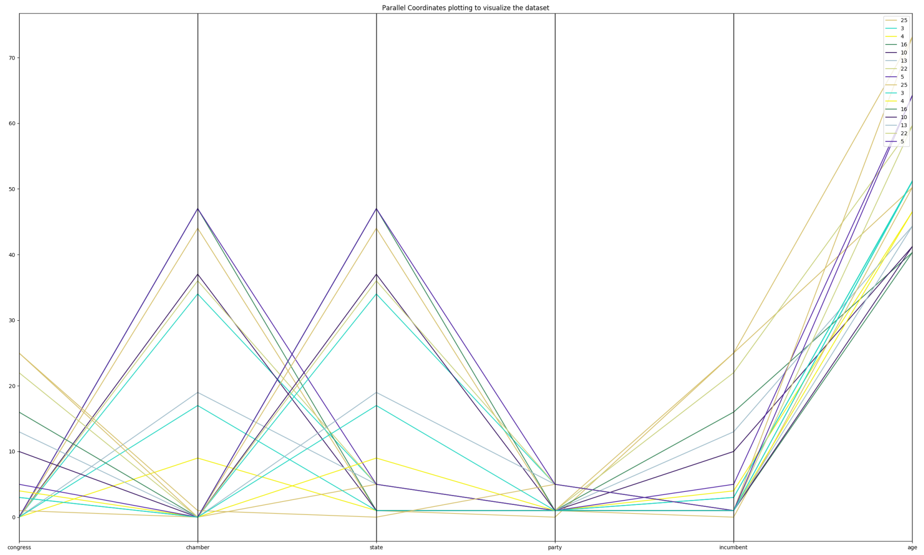
Trends in the data:

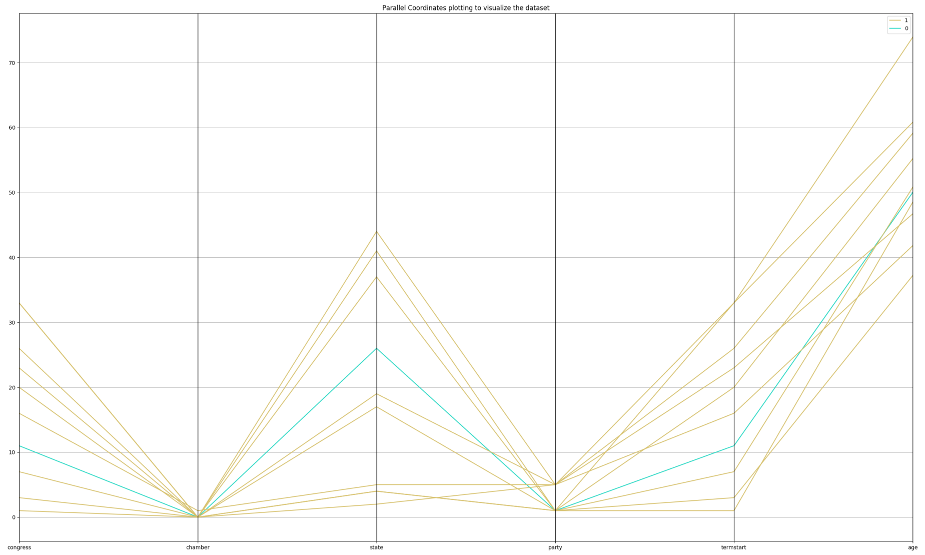
As we can see, the average temperature is very fluctuent and keeps on increasing every year which is visible in the world as well. This rise in temperature might be a consequence of Global Warming. Also, it is clear that this behaviour will follow on forever if we don't start protecting the environment.

2.2 Download the member of Congress by Age data set: link. Create a Star Plot of the data and create a Parallel Coordinates Plot of the data. Describe the trends in the data.

Ans Here, we have a huge database of leaders with different features in table like age, party, state, term served, etc. One of the main challenges here was to identify the important information and analyse it. For Star Plot, one of the things that I did was to find the average age of all politicians from a particular state. This factor is often used in evaluating the parties that's why I thought that it is important. I created a hash table/dictionary to count and store the average age of the politicians. Then for parallel coordinates, I have followed a bit different approach, I used pandas and it's parallel coordinates function to plot it. I used some of the Data Pre-processing by removing the irrelavent columns like First Name, Last Name, Suffix, etc. Then I converted it to fully numerical data by using the *astype('category')* call. The results are attached here. From the star plots and parallel coordinates, there are multiple things we can observe, like the average age of states like Texas is higher as compared to states like Utah, Alabama etc. Then, you can also see a relationship between different factors like age, term start, term served etc. to predict something about a politician. Please find the diagrams on the next page.

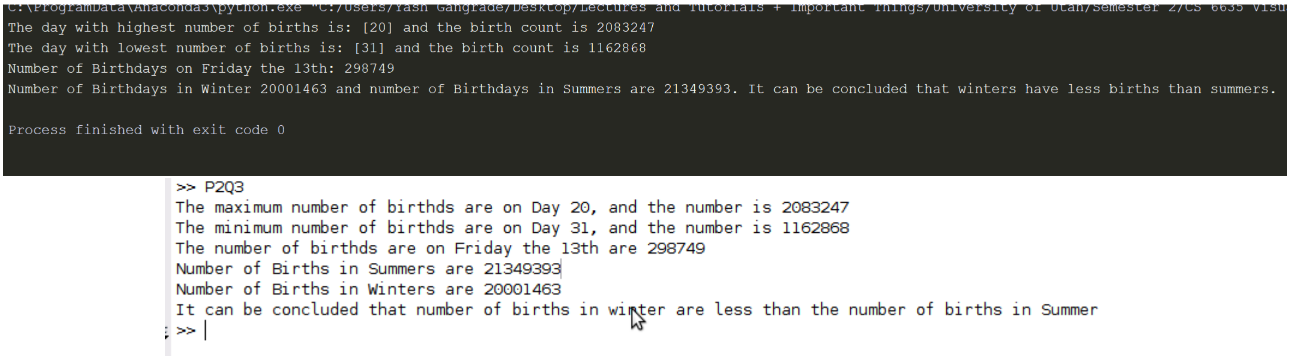

Star plot depicting average age of politicians from each state
TX
NC
IL
NJ
KY
PA
CA
NY
WI
MA
VA
IN
OH
WV
OR
VT
MI
MS
ND
MO
GA
TN
MN
WA
NE
AL
AZ
ID
CO
KS
OK
FL
IA
UT
MT
ME
LA
WY
SC
NM
MD
AR
RI
SD
HI
CT
NH
NV
AK
DE


Parallel Coordinates plotting to visualize the dataset
25
3
4
16
10
13
22
5
25
3
4
16
10
13
22
5
70
60
50
40
30
20
10
0
congress
chamber
state
party
incumbent
age


Parallel Coordinates plotting to visualize the dataset
1
0
70
60
50
40
30
20
10
0
congress
chamber
state
party
termstart
age

2.3 Download the U.S. Birth data set: link. What day of the month had the highest number of births? What day of the month had the lowest number of births? Are there any interesting trends in the data, i.e. more births in Summer or Winter? What about births on Friday the 13th?

Ans Here, we are loading the data of birthdays for every day of every year. Similar to the approach in part 2 ques 1, I parsed the csv file to store the required data in different numpy arrays, dictionaries etc. In MATLAB, I am directly manipulating the data matrix obtained from reading the csv file. Please follow the code for more details. I am attaching the console output and the trends in the data in form of screenshots below:



```
C:\ProgramData\Anaconda3\python.exe "C:/Users/Yash Gangrade/Desktop/Lectures and Tutorials + Important Things/University of Utah/Semester 2/CS 6635 Visu
The day with highest number of births is: [20] and the birth count is 2083247
The day with lowest number of births is: [31] and the birth count is 1162868
Number of Birthdays on Friday the 13th: 298749
Number of Birthdays in Winter 20001463 and number of Birthdays in Summers are 21349393. It can be concluded that winters have less births than summers.

Process finished with exit code 0
```

```
>> P2Q3
The maximum number of birthds are on Day 20, and the number is 2083247
The minimum number of birthds are on Day 31, and the number is 1162868
The number of birthds are on Friday the 13th are 298749
Number of Births in Summers are 21349393
Number of Births in Winters are 20001463
It can be concluded that number of births in winter are less than the number of births in Summer
>> 
```

Figure 7: Outputs from through Python (top) and MATLAB (bottom)

2.4 The U.S. Government maintains a sever with many interesting datasets called Data.Gov: https://www.data.gov/. Choose 3 different data sets to visualize. Visualize the data sets in at least 3 different ways. Describe the trends in the data.

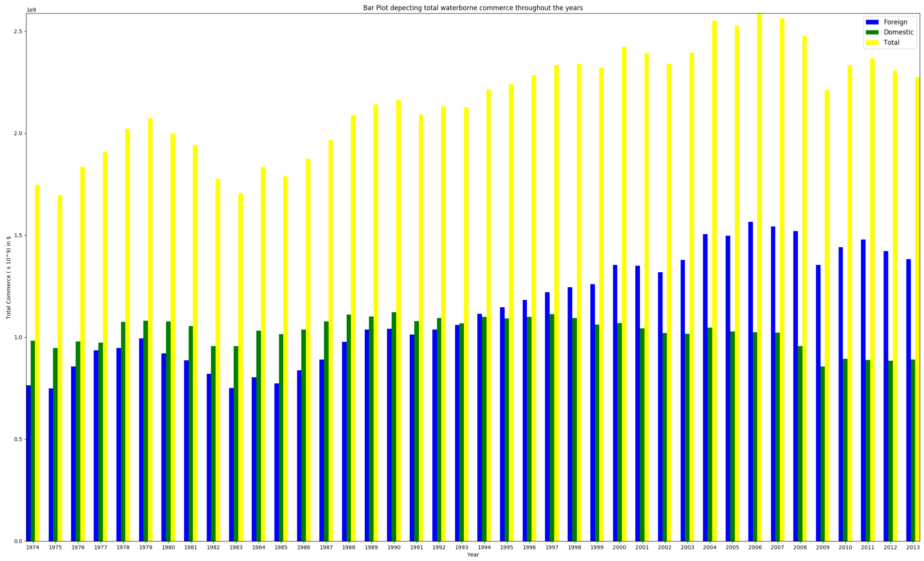
Ans The three datasets that I chose are as follows:

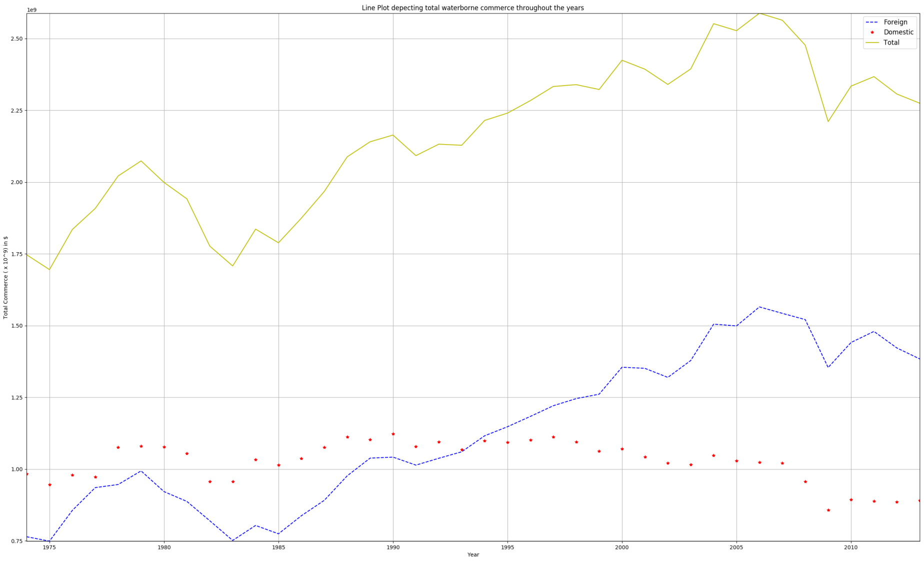
1. Dataset containing details about the total waterborne commerce in and out of US i.e. Domestic and International from years 1974 to 2013
2. Dataset containing the details about the number of pounds of food items and services being recalled for year 2014 - the products are classified in three classes.
3. Dataset depicting the national population estimates for each year from 2000 to 2015. I have stripped the dataset to essential information and deleted extra paragraph of words etc.

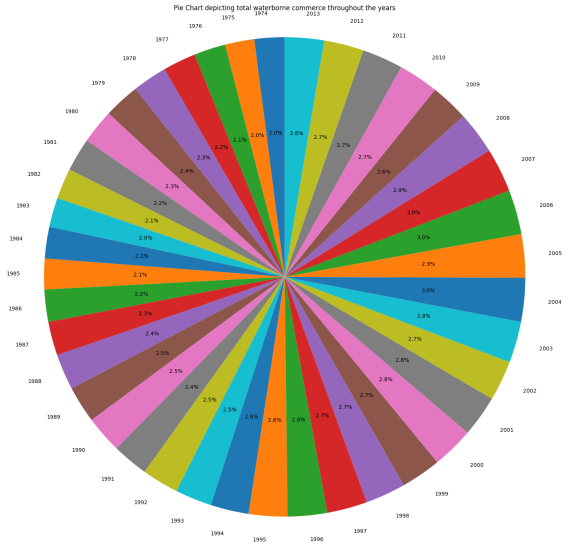
Datset 1 - Total Waterborne Commerce

I found this data on data.gov website under maritime section. Similar to the above questions, I have stored the required information into numpy arrays, dictionaries, etc by parsing the csv files. For this particular dataset, I am using three plots namely, bar plot, line graph and pie chart to visualize the data. The results are attached through diagrams below.

We can observe that the total waterborne commerce had been quite fluctuating as in sometimes low and sometimes high in some years. But we can clearly see from the plots that domestic commerce has been more or less constant throughout the years, but foreign sales have increased a lot in percentage given the year 1974 and the years 2010-13. This may be because of enforcement of new policies.


Bar Plot depecting total waterborne commerce throughout the years
1e9
Total Commerce (x 10^9) in $
Year
Foreign
Domestic
Total
0.0
0.5
1.0
1.5
2.0
2.5
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

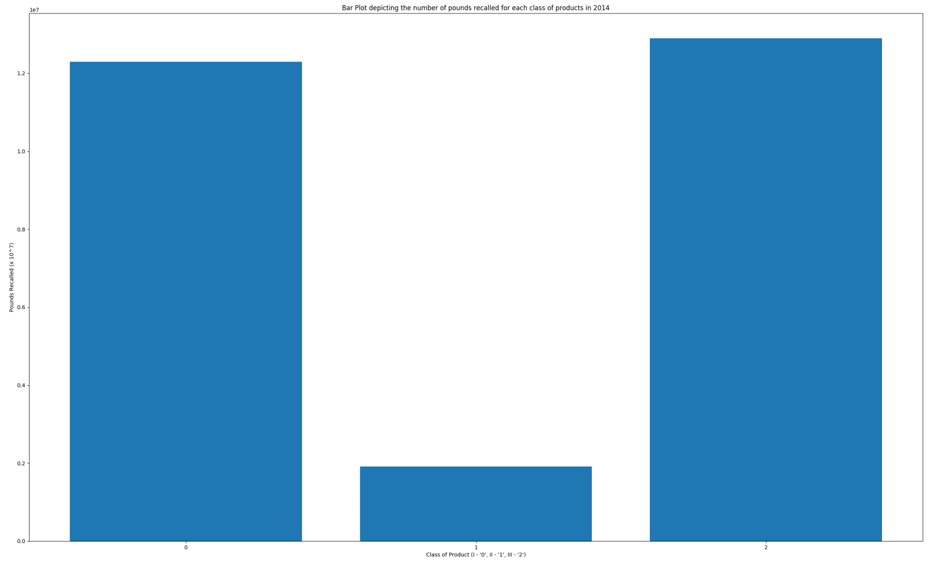

Line Plot depecting total waterborne commerce throughout the years
1e9
Total Commerce (x 10^9) in $
Year
Foreign
Domestic
Total
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
1975
1980
1985
1990
1995
2000
2005
2010

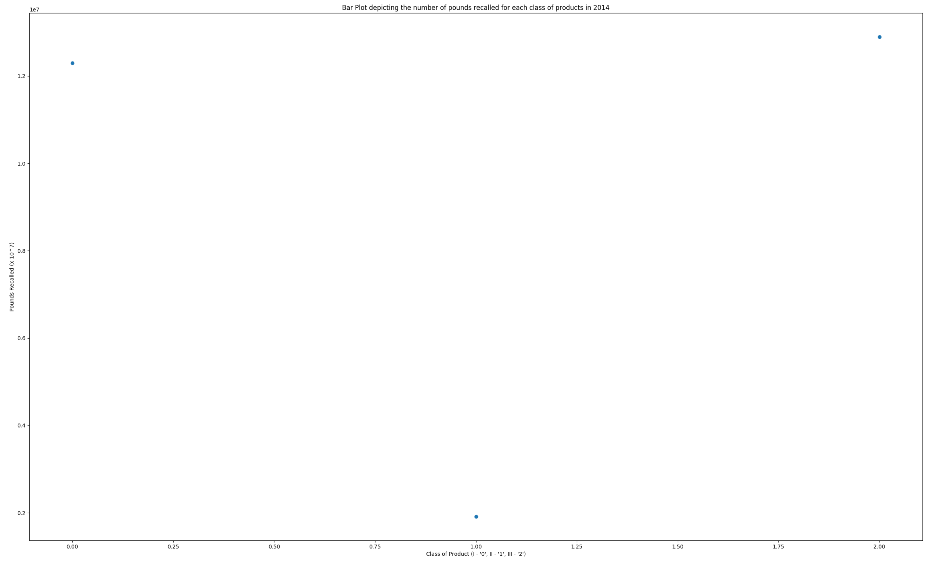

Pie Chart depicting total waterborne commerce throughout the years
1974
2.0%
1975
2.0%
1976
2.1%
1977
2.2%
1978
2.3%
1979
2.4%
1980
2.3%
1981
2.2%
1982
2.1%
1983
2.0%
1984
2.1%
1985
2.1%
1986
2.2%
1987
2.3%
1988
2.4%
1989
2.5%
1990
2.5%
1991
2.4%
1992
2.5%
1993
2.5%
1994
2.6%
1995
2.6%
1996
2.6%
1997
2.7%
1998
2.7%
1999
2.7%
2000
2.8%
2001
2.8%
2002
2.7%
2003
2.8%
2004
3.0%
2005
2.9%
2006
3.0%
2007
3.0%
2008
2.9%
2009
2.6%
2010
2.7%
2011
2.7%
2012
2.7%
2013
2.6%

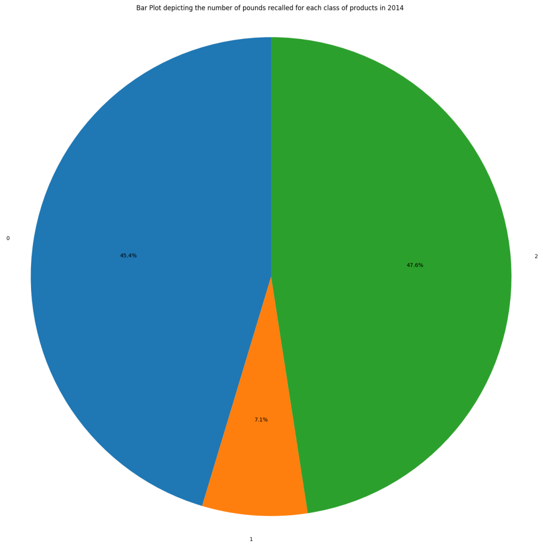
Dataset 2 - FSIS - Recall Summary for 2014

I found this dataset under the Agriculture division on data.gov website. It contains information (like number of pounds) about all the food items like Chicken, bacon etc. which were recalled in the year 2014. I have categorized the data according to the class of food product and then visualized it using Bar plot, Pie chart, and Scatter Plot. The results are attached here.

We can clearly observe that the class 1 products which are mostly beef and pork related products were recalled in the least amount that year. Class three and two were somewhat close with just a minor difference in percentages and it constituted products like Chicken, turkey, cheese etc. This data shows us about the difficulties which might have been caused to people because of all this recall procedure and food going out of the market.


Bar Plot depicting the number of pounds recalled for each class of products in 2014
1e7
Pounds Recalled (x 10^7)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1
2
Class of Product (I - '0', II - '1', III - '2')

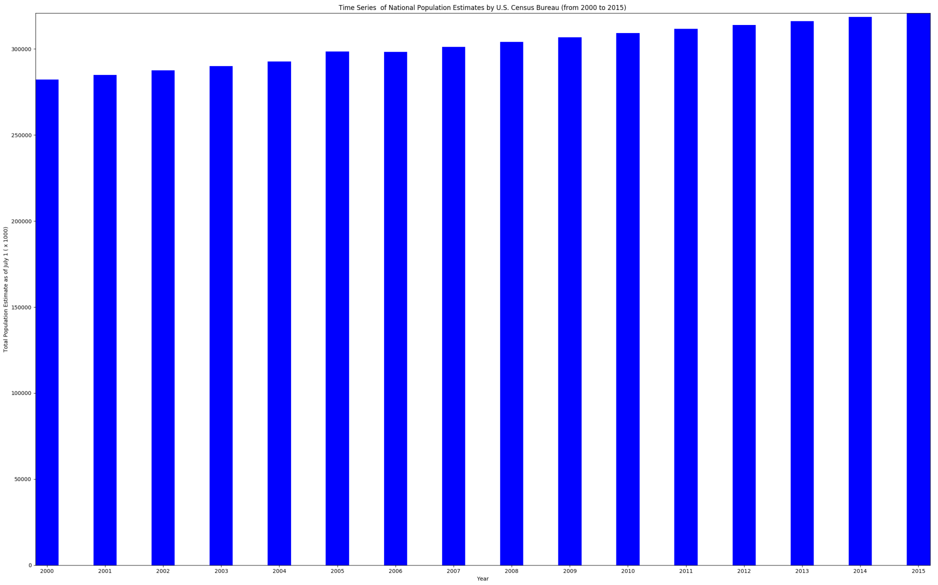

Bar Plot depicting the number of pounds recalled for each class of products in 2014
1e7
Pounds Recalled (x 10^7)
0.2
0.4
0.6
0.8
1.0
1.2
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Class of Product (I - '0', II - '1', III - '2')

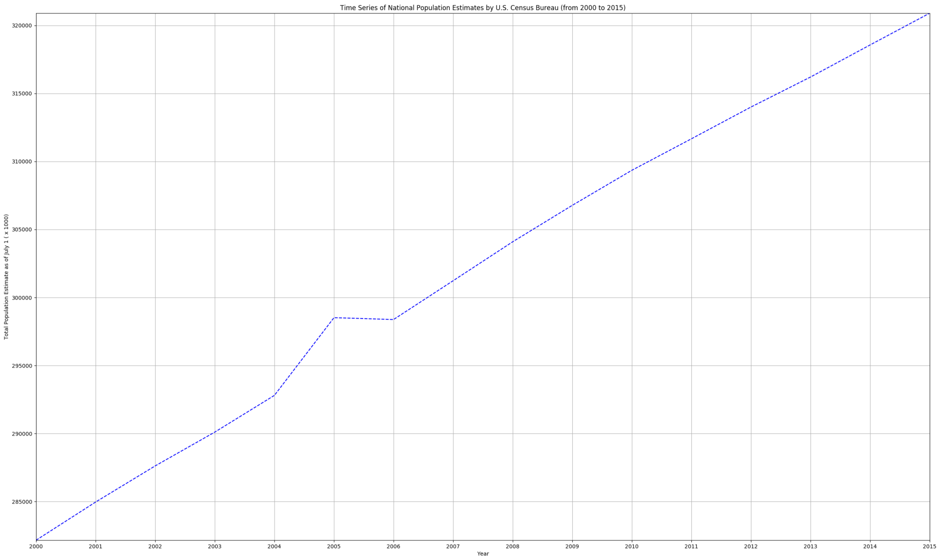

Bar Plot depicting the number of pounds recalled for each class of products in 2014
0
45.4%
2
47.6%
7.1%
1

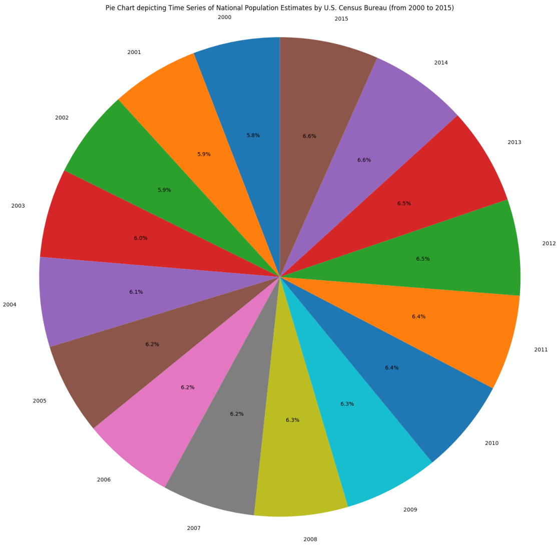
Dataset 3 - National Population Estimates

I found this dataset on the census website of government (linked to data.gov). This data depicts the population estimates in US for each year from 2000 to the year 2015. I have used Bar plots, line graph and pie chart to visualize the data. The results are attached in the next page.

It can be clearly observed that the National Population is increasing every year which is a concern for the country. Although the growth is not completely drastic but this trend if continued might become a cause of worry for the nation. This increasing growth is present in almost every country in the world. This is also one of the reason why many countries are employing rules like 2 child rule to curb this continously increasing growth in the population.


Time Series of National Population Estimates by U.S. Census Bureau (from 2000 to 2015)
Total Population Estimate as of July 1 (x 1000)
0
50000
100000
150000
200000
250000
300000
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Year


Time Series of National Population Estimates by U.S. Census Bureau (from 2000 to 2015)
Total Population Estimate as of July 1 (x 1000)
285000
290000
295000
300000
305000
310000
315000
320000
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Year


Pie Chart depicting Time Series of National Population Estimates by U.S. Census Bureau (from 2000 to 2015)
2000
5.8%
2001
5.9%
2002
5.9%
2003
6.0%
2004
6.1%
2005
6.2%
2006
6.2%
2007
6.2%
2008
6.3%
2009
6.3%
2010
6.4%
2011
6.4%
2012
6.5%
2013
6.5%
2014
6.6%
2015
6.6%