FIT5145 Assignment – 1

Jaimon Thyparambil Thomas
Student ID : 29566428
Email : jthy0001@monash.student.edu
Monash University
September 02, 2018

Task A: Investigating Population and Gender Equality in Education

In the task, you are required to visualise the relationship between the population in different countries, the income in different countries and the gender ratio (women % men, 25 to 34 years) in schools of different countries, and gain insights from how these relations and trends change over time. The data files used in this task were originally downloaded from Gapminder. We have extracted the data from the original files and put into a simpler format. Please download the data from Moodle:

- Population.csv: This file contains yearly data regarding the estimated resident population, grouping by countries around the world, between 1800 and 2018.
- GenderEquality.csv: This data file contains yearly data about the ratio of female to male number of years in school, among 25- to 34-years-olds, including primary, secondary and tertiary education across different countries around the world, for the period between 1970 and 2015.
- Income.csv: This data file contains yearly data of income per person adjusted for differences in purchasing power (in international dollars) across different countries around the world, for the period between 1800 and 2018.

A1. Investigating the Population Data

Have a look at the resident population data. You will see many columns representing different countries.

1.In Python plot the population growth of Australia, China and United States over time

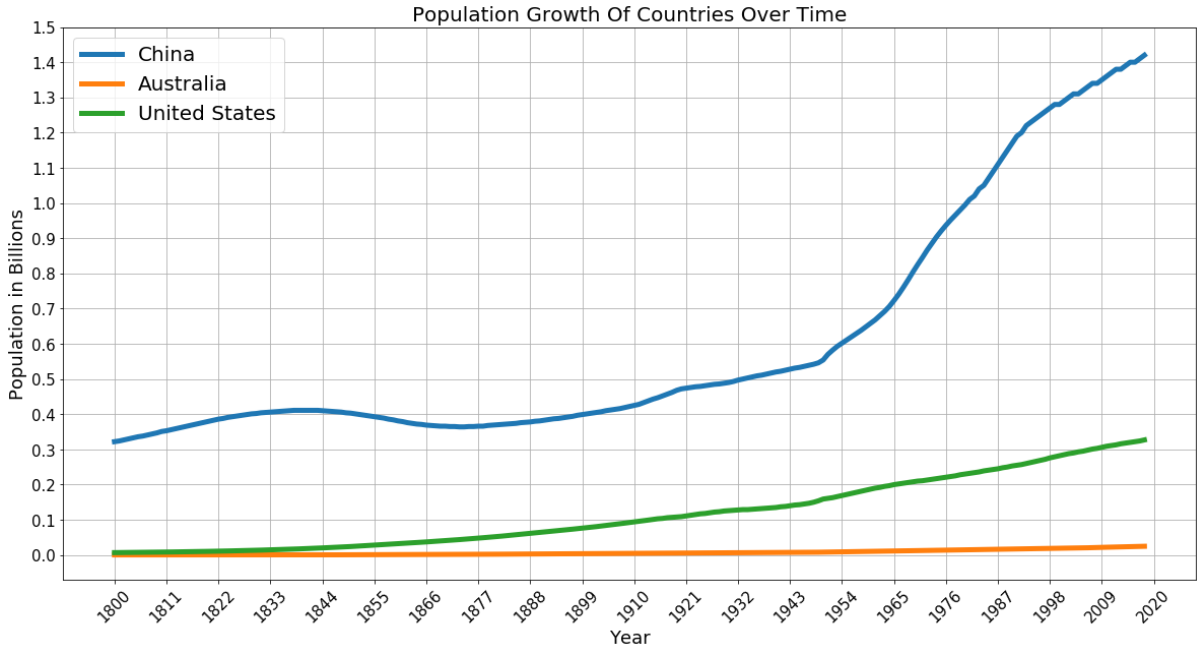


Fig 1

- Are the population values increasing or decreasing over time?

  As we can see from Fig 1 that the population values is increasing over time

2.Fit a linear regression using Python to the Chinese population data and plot the linear fit

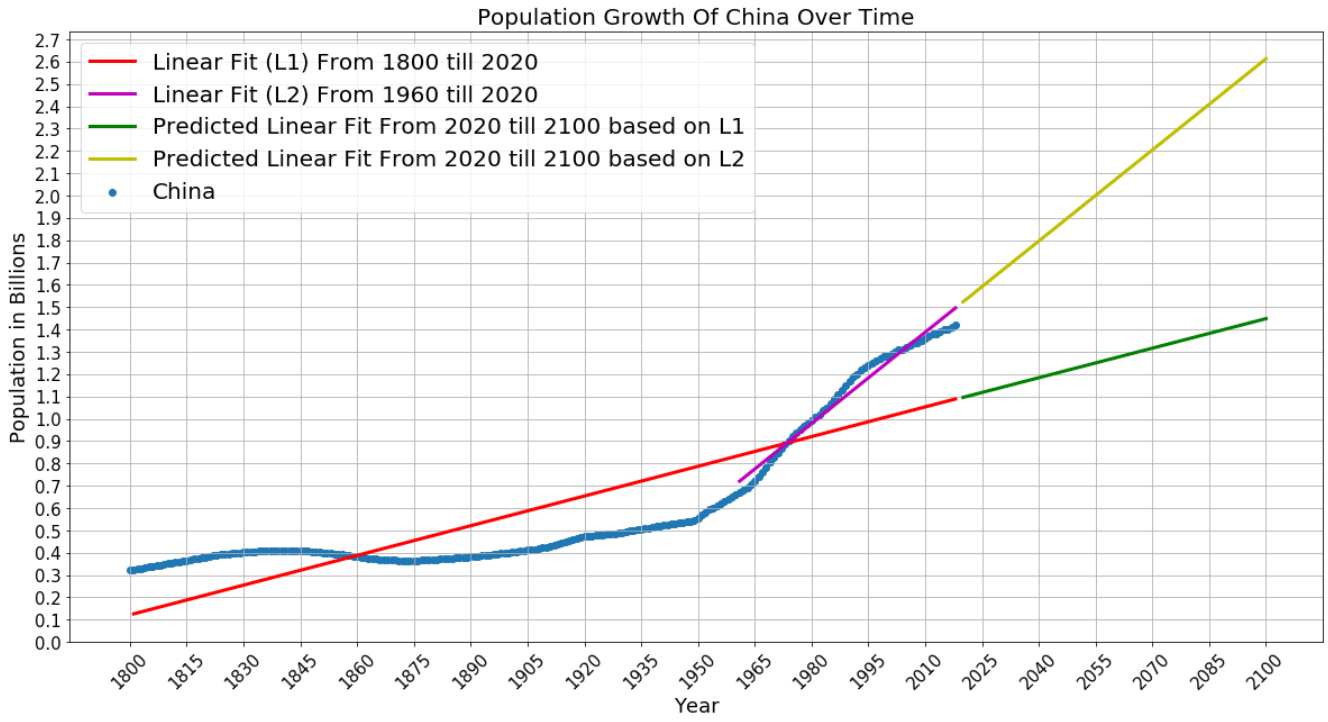


Fig 2

- Does the linear fit look good?

  Here the Linear Fit Line L1 (Red Line) from year 1800 till year 2020 doesn't look good because from the above graph we can see that after 1950 there is a sudden increase in the rate of growth of population of china. Which lies far away from the linear fit L1

- Use the linear fit to predict the resident population in China in 2020 and 2100.

  Here the green line indicates the resident population in china from 2020 till 2100

- instead of fitting the linear regression to all the data, try fitting it to just the most recent data points (say from 1960 onwards). How is the fit? Which model would give better predictions of future population in China do you think?

  Here the Line L2 (Magenta Line) represents the Linear Fit for the Chinese population as per the most recent data points that is from 1960 onwards. This Linear Fit looks more appropriate because when we investigate the data. The data from 1960 lies close to the linear fit of L2. So, I feel the linear Fit based

```
of L2 will provide a better prediction of the future
population of china
```

A2. Investigating the Gender Equality Data

Now have a look at the gender equality data.

1.Use Python to plot the gender ratio (women % men) in schools for Australia, China and United States over time.

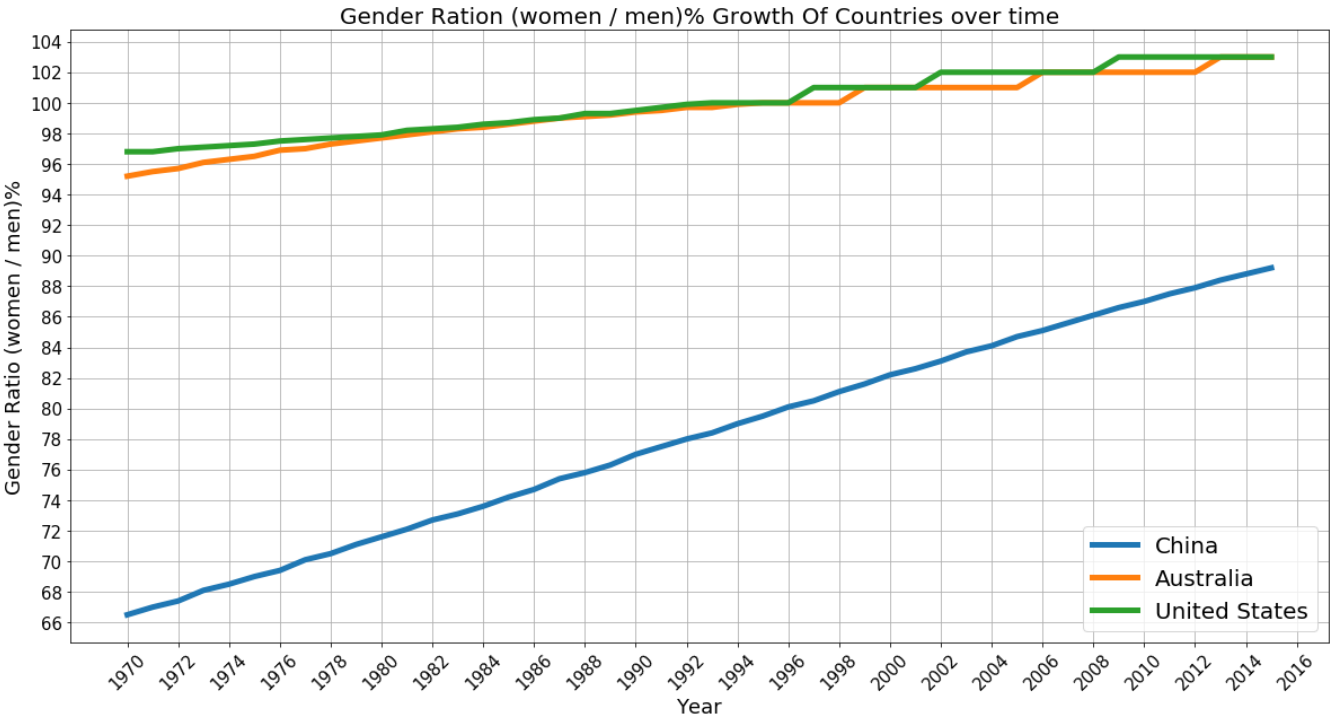


Fig 3

- What are the maximum and minimum values for gender ratio in Australia over the time period?

```
Maximum value for gender ratio in Australia is 103
Minimum value for gender ratio in Australia is 95.2
```

- How do you compare the trend in gender ratio (women % men) in schools for these three countries over the time period? Which two countries have similar growth trend?

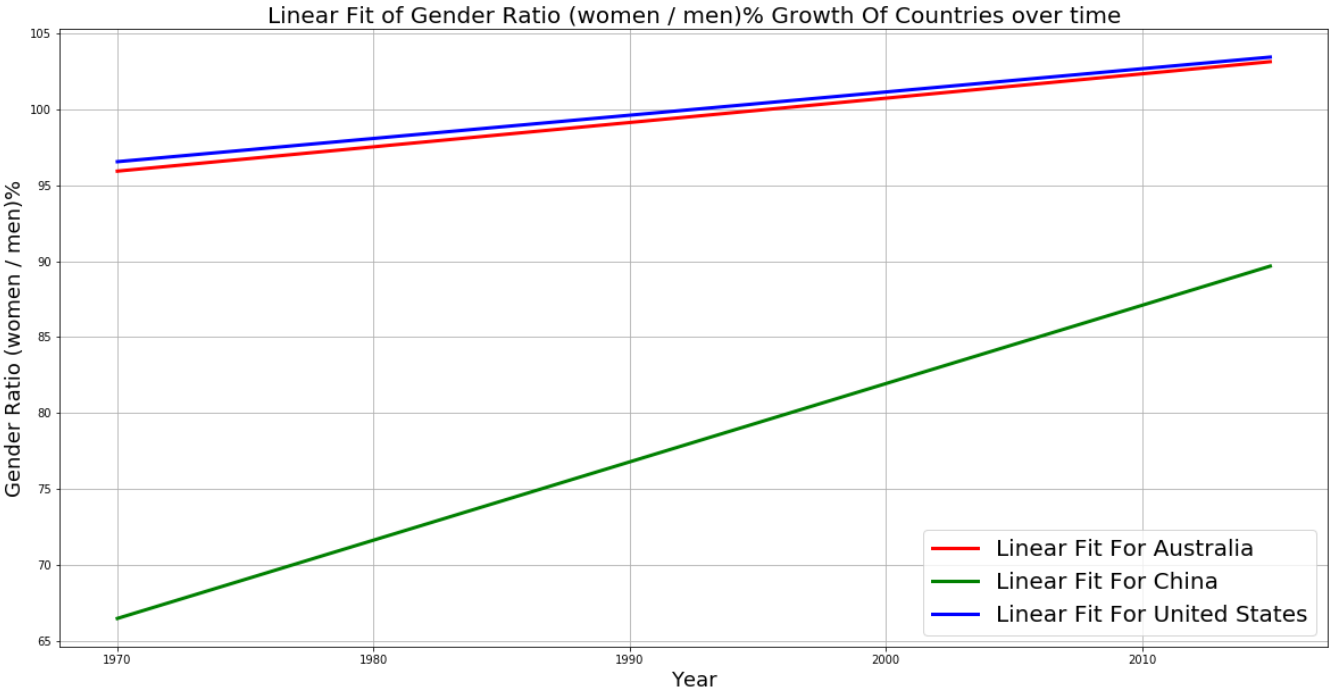


Fig 4

From the linear fit of the three countries in Fig 4 we can say that gender ratio (women/men) % is generally increasing over the time for these countries. We can also see that gender ratio in china is increasing at a faster rate compared to Australia and United States. From the Fig 4 we can also see that the growth trend of Australia and United States are similar

2.Fit a linear regression to the gender ratio in schools in United States and plot it.

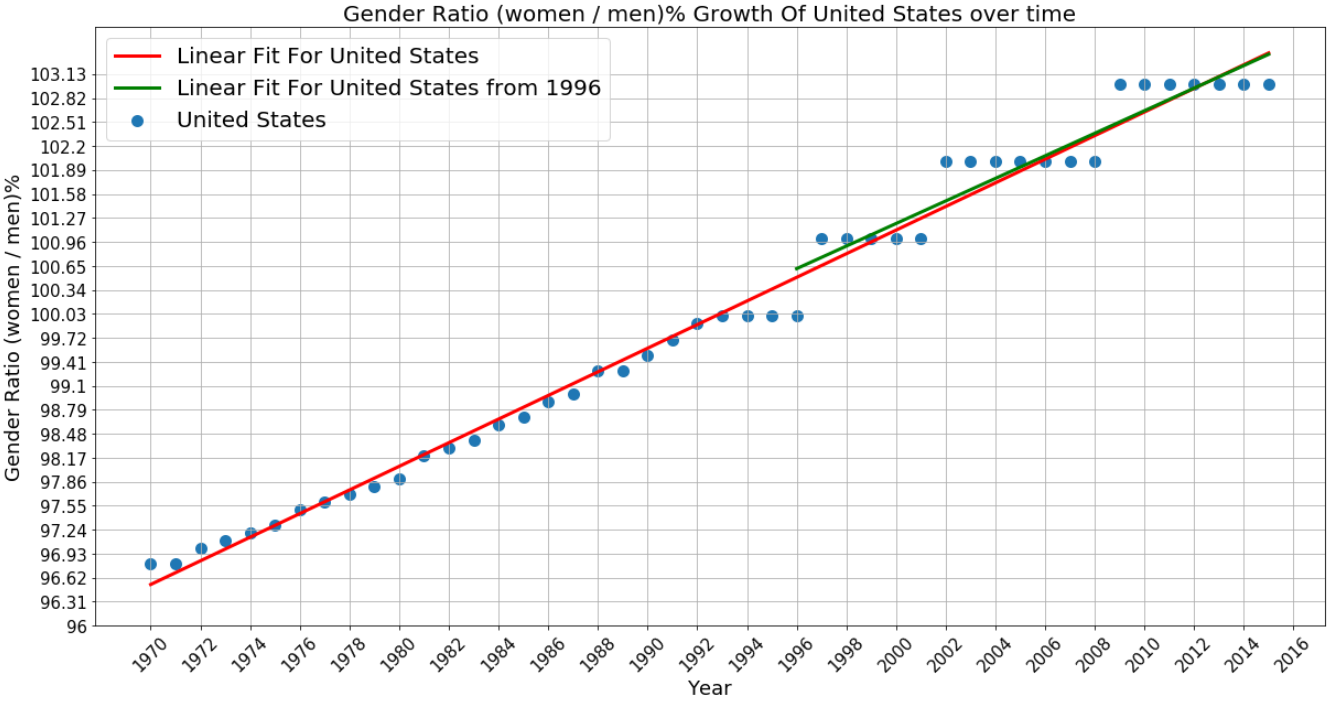


Fig 5

- Does it look like a good fit to you? Would you believe the predictions of the linear model going forward?

  Yes, the Linear Fit in Fig 5 look good for me because all the data lies close to the linear fit. Yes, I would believe the predictions of the linear model going forward because when we look at the linear fit of the last 10 years that is from 1996 against the linear fit from 1970 we could see that there is not much difference and both the linear fits looks almost similar

A3. Investigating the Income Data

Now have a look at the Income data.

1.Use Python to plot the Income of Australia, China and United States over time.

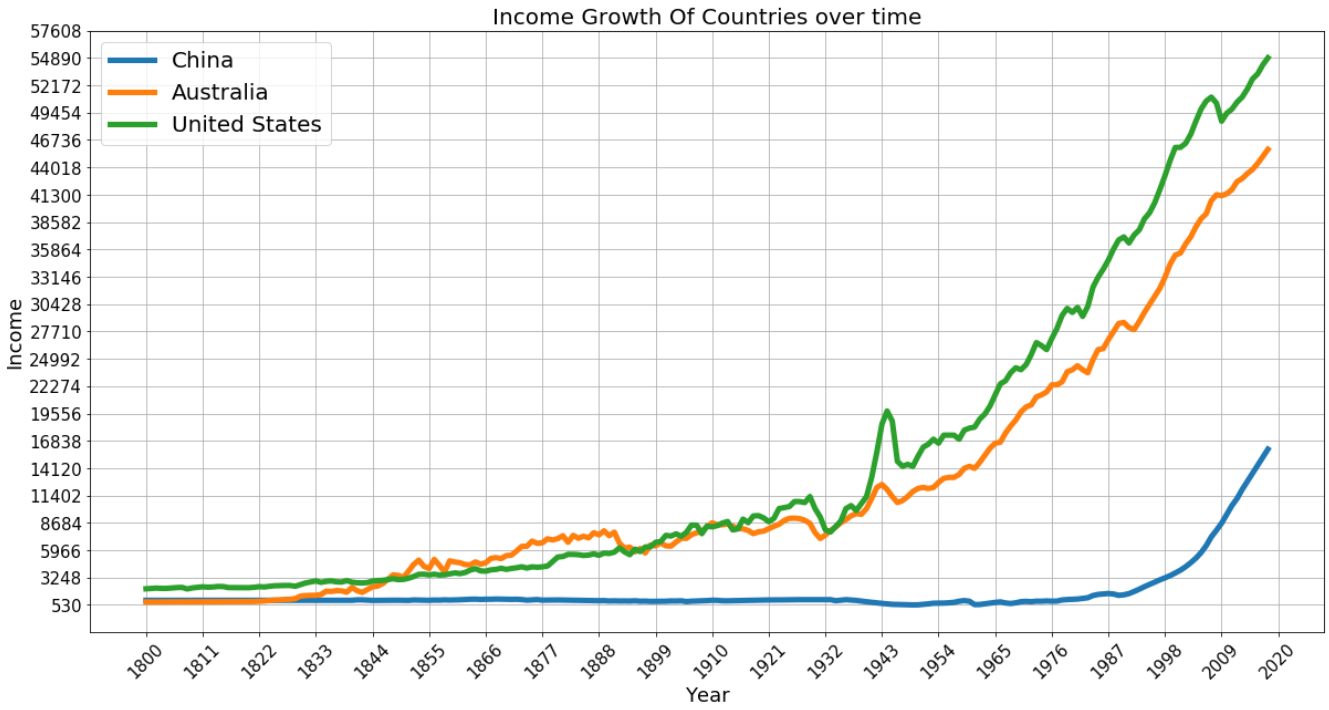


Fig 6

- What was the minimum income in China recorded in the dataset and when did that occur? What was the income in Australia in the same year?

  Minimum income value of China is 530 on 1949 and on the same year income of Australia was 11800

A4. Visualising the Relationship between Gender Equality and Population

Now let's look at the relationship between gender ratio in schools and the population.

1.Use Python to combine the data from the different files into a single table. The table should contain population values, income and gender ratio in schools for the different years and different countries.

- What is the first year and last year for the combined data?

```
First Year of the combined data is 1970
Last year of the combined data is 2015
```

2.Now that you have the data aggregated, we can see whether there is a relationship between gender ratio in schools and the population. Plot the values against each other.

- Can you see a relationship there?

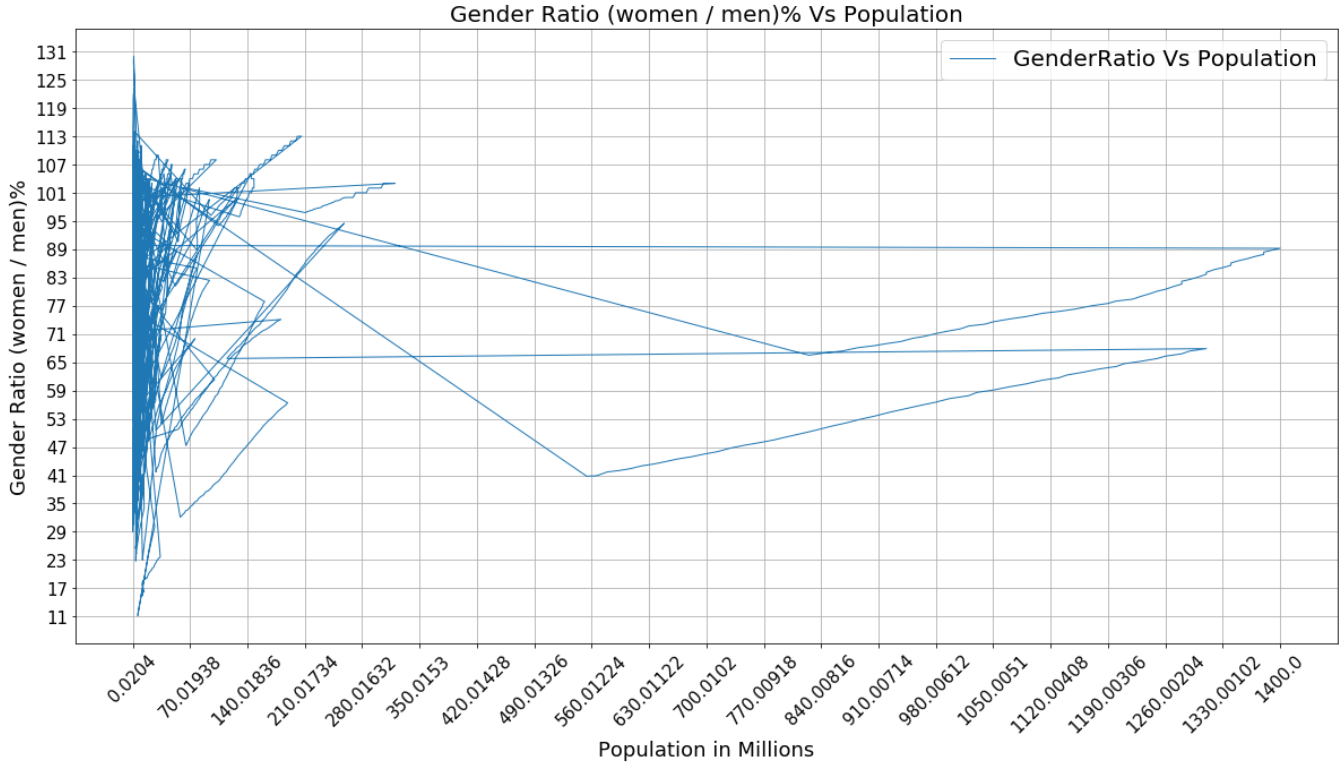


Fig 7

Fig represents the aggregated data of all the countries combined. As we can see from the figure we can't find any relationship here. Now lets try to split the data based on country wise then the graph will look like

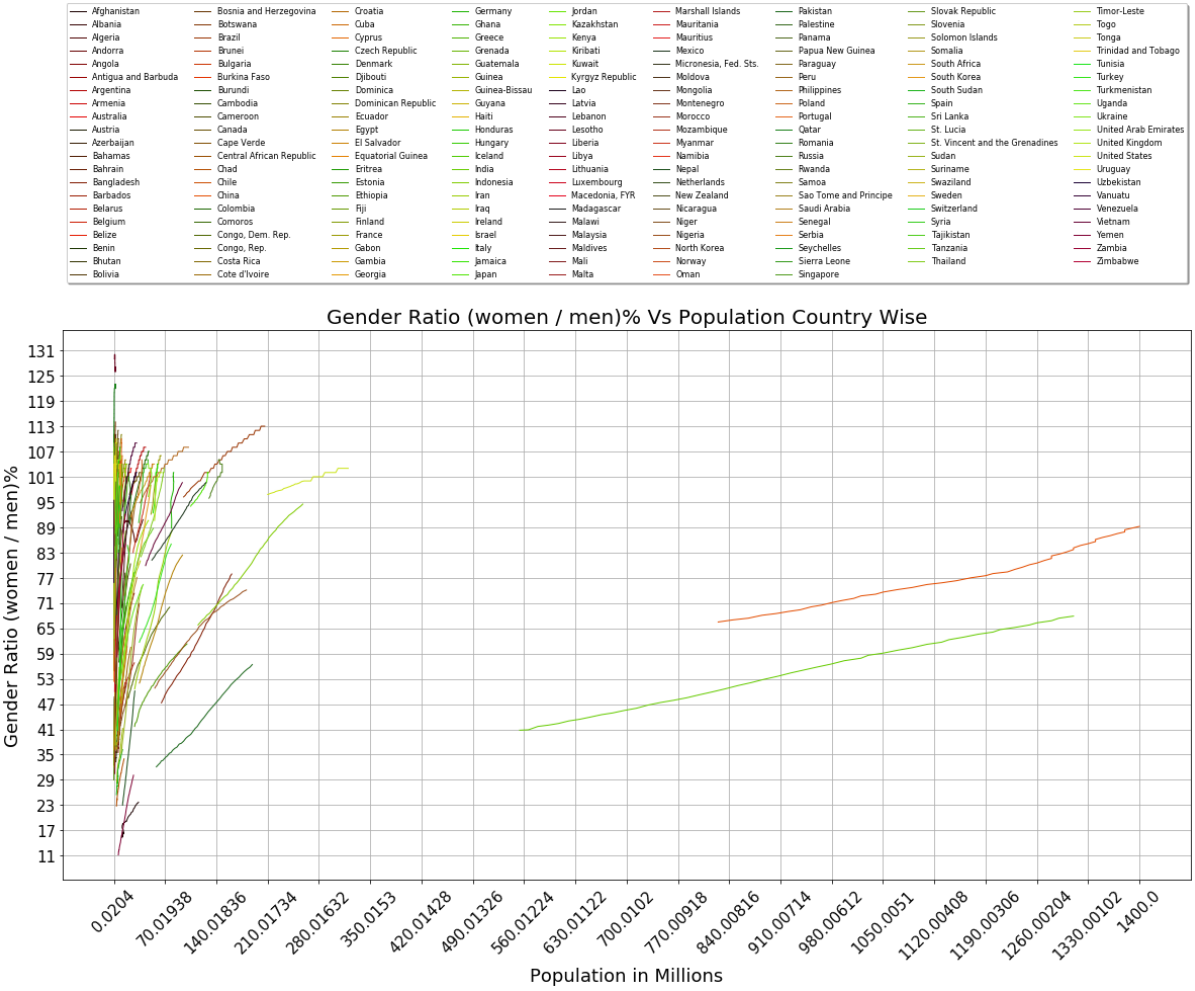


Fig 8

In Figure 8 also since there is lot of information we are not able to clearly find any relation

3.Try selecting and plotting only the data from India.

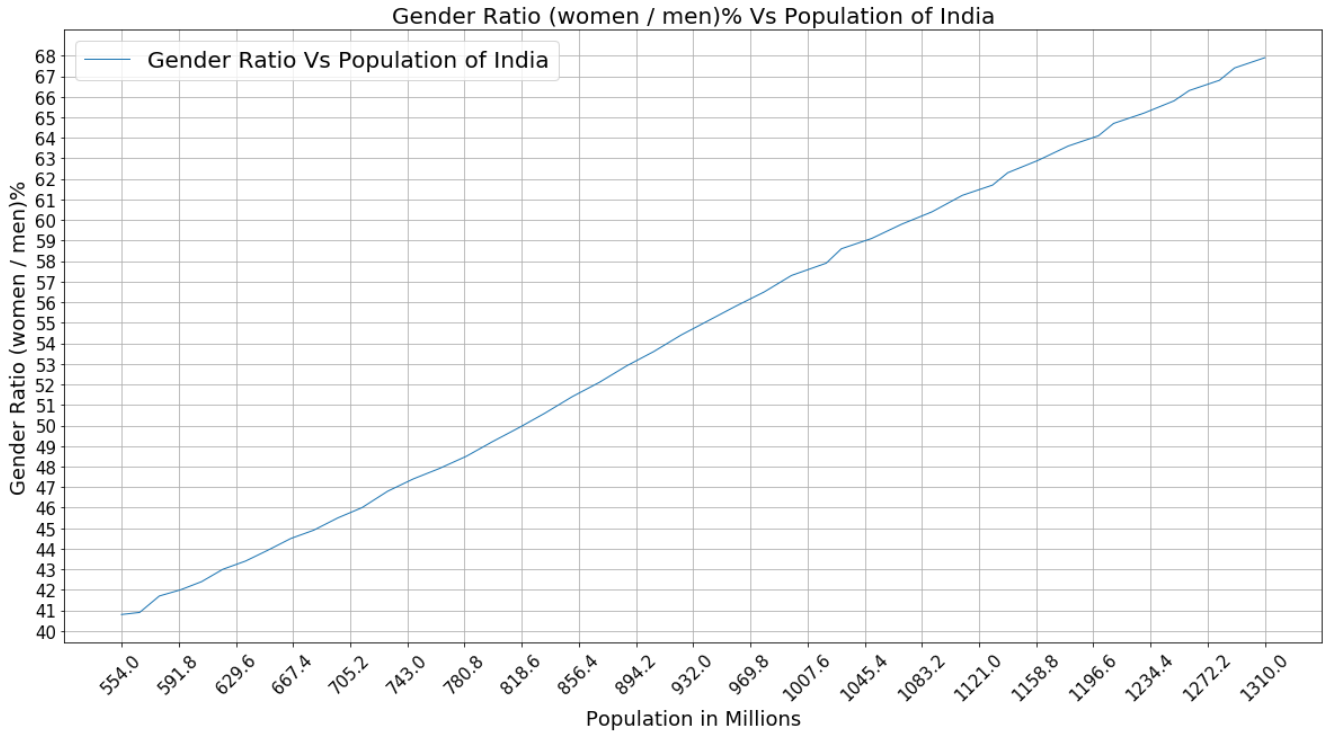


Fig 9

- Can you see a relationship now? If so, what relationship is there?

  Yes, now we can clearly see a relationship. That is as the population increases gender ratio is also increasing.

A5. Visualising the Relationship over Time

Now let's look at the relationship between gender ratio in schools and income over time.

1.Use Python to build a Motion Chart comparing the gender ratio in schools, the income, and the population of each country over time. The motion chart should show the gender ratio in schools on the x-axis, the income on the y-axis, and the bubble size should depend on the population.

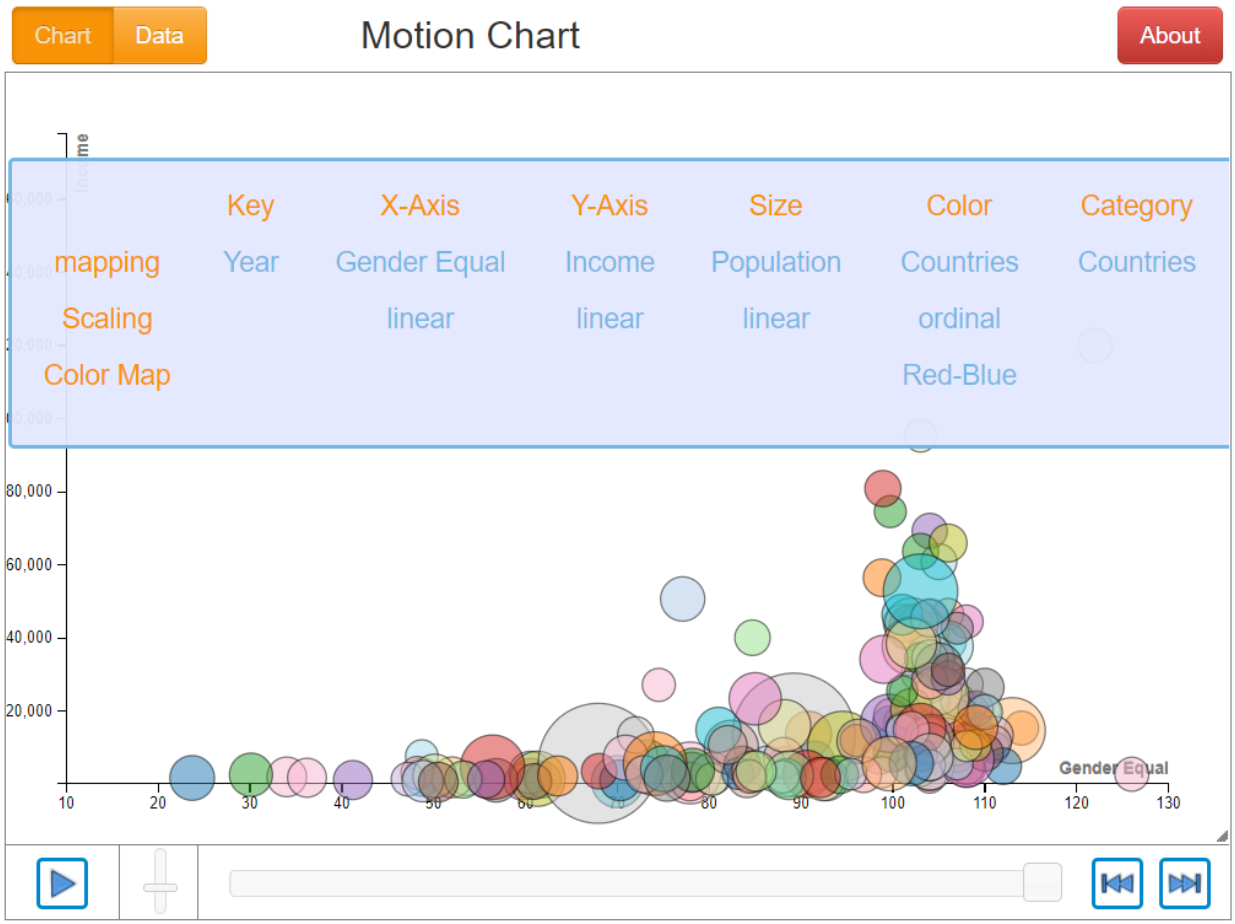


Fig 10

2.Run the visualisation from start to finish. (Hint: In Python, to speed up the animation, set timer bar next to the play/pause button to the minimum value.) And then answer the following questions:

- Which two countries generally have the lowest gender ratio (women % men) in schools?

  Yemen and Afganistan generally have the lowest gender ratio (women % men) in schools. In 2015 yemen had ratio (30.1%) and afganistan has (23.7%). Whereas in 1970 Yemen had a gender ratio of (11.2%) and Afganistan has a ratio of (15.4%)

- Which country has the highest gender ratio during the whole period of time?

  Lesotho has the highest gender ratio during the whole period with a max value of 130

- Is the gender ratio generally increasing or decreasing during the whole period of time? How about income? Explain your answer.

  Gender ratio is generally increasing during the whole period. Whereas income if we consider overall image we can say that it is slightly increasing. For many countries with less

population like UAE income keeps on fluctuating like sometime increasing whereas some time decreasing.

- Select Cape Verde and Bolivia for this question: From which year onwards does Cape Verde start to have a higher gender ratio and a higher income from Bolivia. Please support your answer with a relevant python code and motion chart.

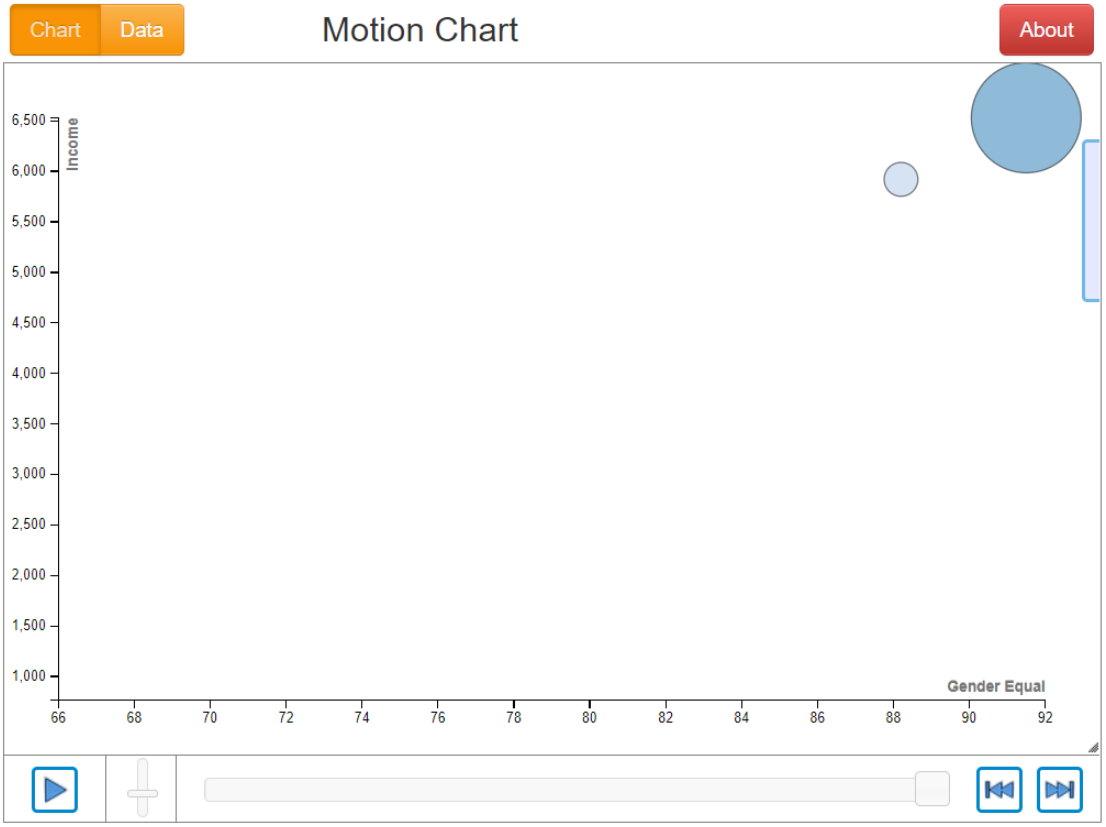


Fig 11

From year 2006 onwards till 2013 Cape Verde has a higher income than Bolivia. As per the data given Cape Verde never had gender ratio higher than Bolivia

- Is there generally a relationship between the amount of income and gender ratio (women %men) in schools in all countries during the whole period of time? What kind of relationship? Explain your answer.

Yes, there is generally a relationship between amount of income and gender ratio in schools in all countries. That is generally as income increases gender ratio (women % men) also increases.

- Any other interesting things you notice in the data? Please support your answer with relevant python code and/or motion chart

As population increases income also increases.

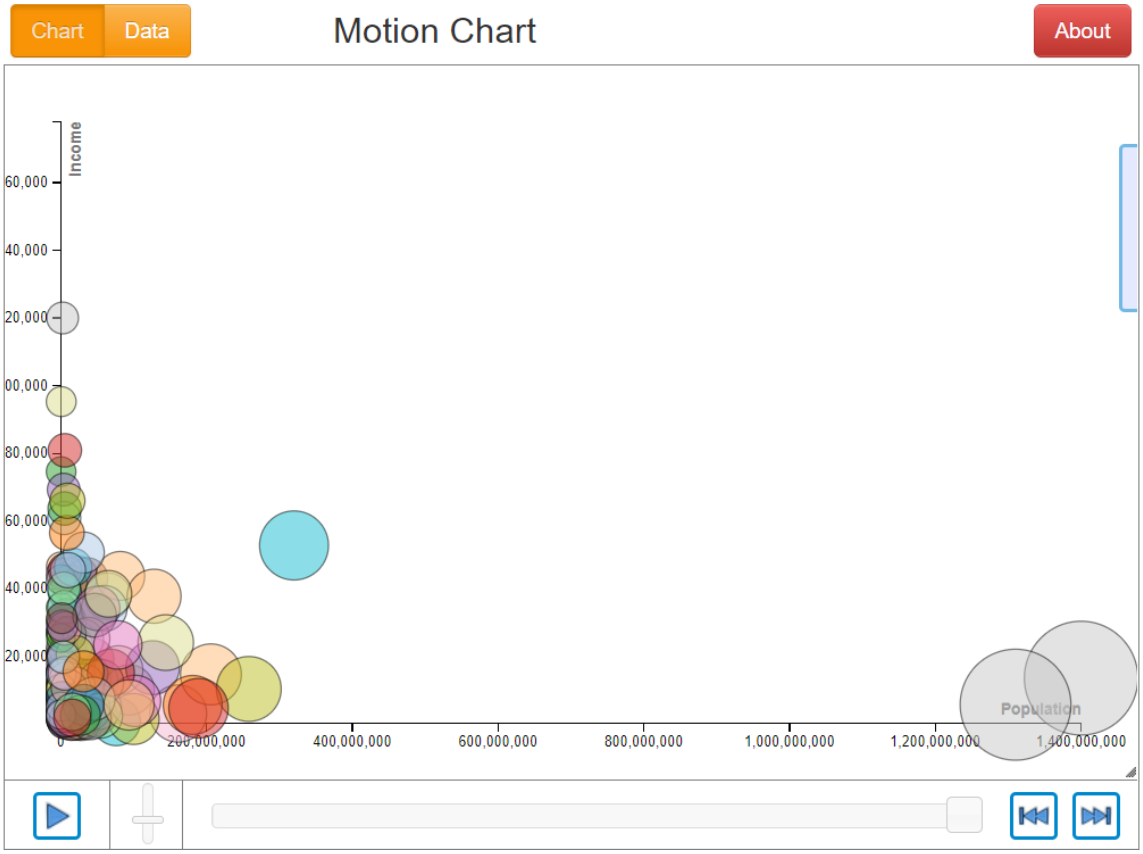


Fig 12

Task B: Exploratory Analysis on Big Data

In this part, you are required to do some exploratory analysis on the health insurance marketplace data. The file InsuranceRates.csv.zip contains data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace. This data was originally prepared and released by the Centers for Medicare & Medicaid Services (CMS). The data was then published on Kaggle. The file we provide is an extract from the data on Kaggle. Unzipped, the file is over 500MB and contains the following fields:

COLUMN	DESCRIPTION
BusinessYear	Year for which plan provides coverage to enrollees.
StateCode	Two-character state abbreviation indicating the state where the plan is offered
IssuerId	Five-digit numeric code that identifies the issuer organization in the Health Insurance Oversight System (HIOS)
PlanId	Fourteen-character alpha-numeric code that identifies an insurance plan within HIOS

Age	Categorical indicator of whether a subscriber's age is used to determine rate eligibility for the insurance plan.
IndividualRate	Dollar value for the monthly insurance premium cost applicable to a non-tobacco user for the insurance plan in a rating area, or to a general subscriber if there is no tobacco preference.
IndividualTobaccoRate	Dollar value for the monthly insurance premium cost applicable to a tobacco user for the insurance plan in a rating area

Load the InsuranceRates.csv data in Python and answer the following questions:

1.How many rows and columns are there?

```
Total No of Rows : 12694445
Total No of Columns : 7
```

2.How many years does the data cover? (Hint: pandas provides functionality to see 'unique' values.)

```
It cover around 3 years of data. That is 2014, 2015 and 2016
```

3.What are the possible values for 'Age'?

```
The possible values of age are 0-20, 21, 22, 23, 24, 25, 26, 27, 28, 29
, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 4
7, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64 a
nd 65 and over, Family Option

Which can be further categorised into 4 categories ie (0-20), (entries
for ages between 21 - 64), (65 and over) and Family Option
```

4.How many states are there?

```
There are around 39 states in this data.
```

5.How many insurance providers are there?

```
There are around 910 insurance providers in this data.
```

6.What are the average, maximum and minimum values for the monthly insurance premium cost for an individual? Do those values seem reasonable to you?

```
Insurance plan rates for individual without considering tobacco user's
stats are
Minimum Rate: 0.0
Maximum Rate: 999999.0
Average Rate: 4098.026458581588
both values that is minimum and maximum rates doesn't seem reasonable
as both values are far away from the overall average
```

B2. Investigating Individual Insurance Costs

Now let's look more in detail at the individual insurance costs.

1.Show the distribution of ‘IndividualRate’ values using a histogram.

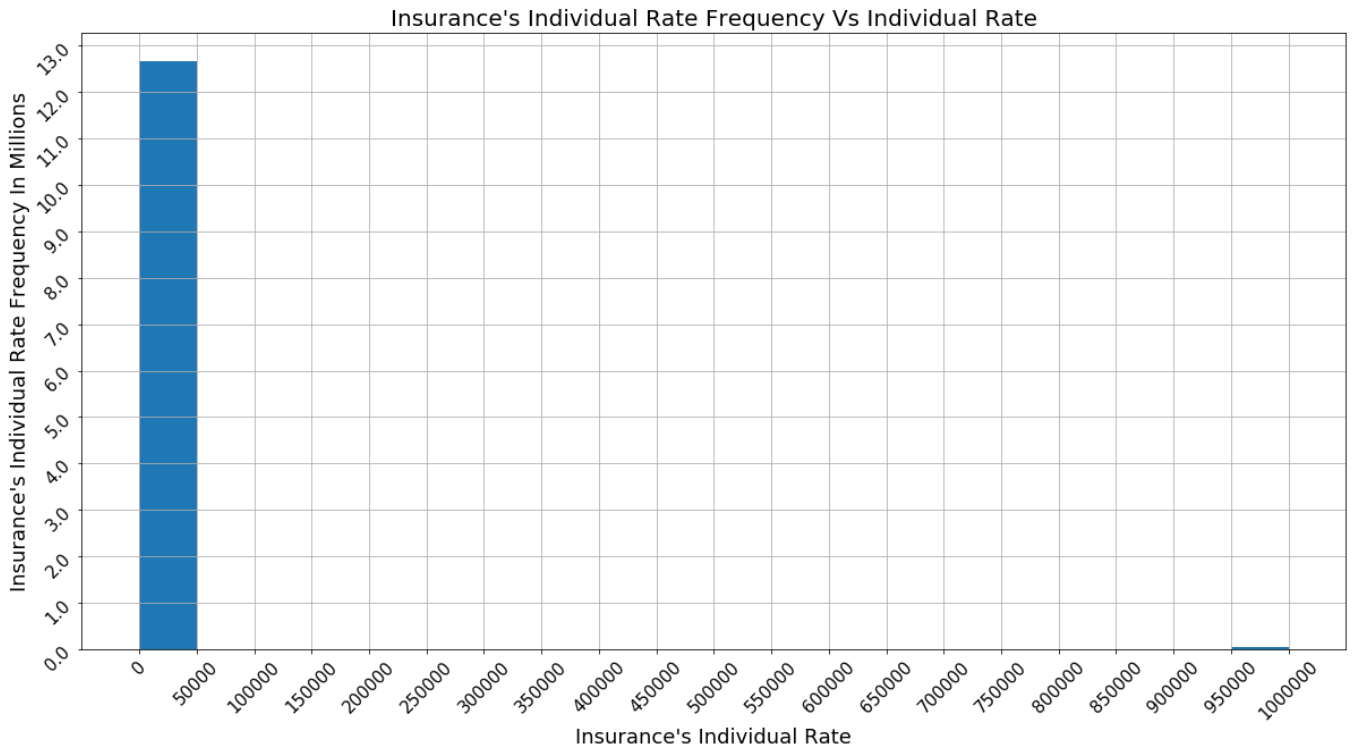


Fig 13

- Does the distribution make sense to? What might be going on?

  This data doesn’t make much sense as it is not been distributed properly. Also, as per this most of the users who have bought insurance. Their individual insurance rates lie between section 0 to 50000. Some users have bought plans between 950000 and 1000000. But the issue in this data is that we don’t know how the rates is been distributed within the range of 0 to 50000. Here the values between 950000 and 1000000 looks like outliers as there is only less frequency for it. Moreover, there doesn’t seem to be having any values between 50000 and 1000000

2.Remove rows with insurance premiums of 0 (or less) and over 2000. (Use this data from now on.) Generate a new histogram with a larger number of bins (say 200).

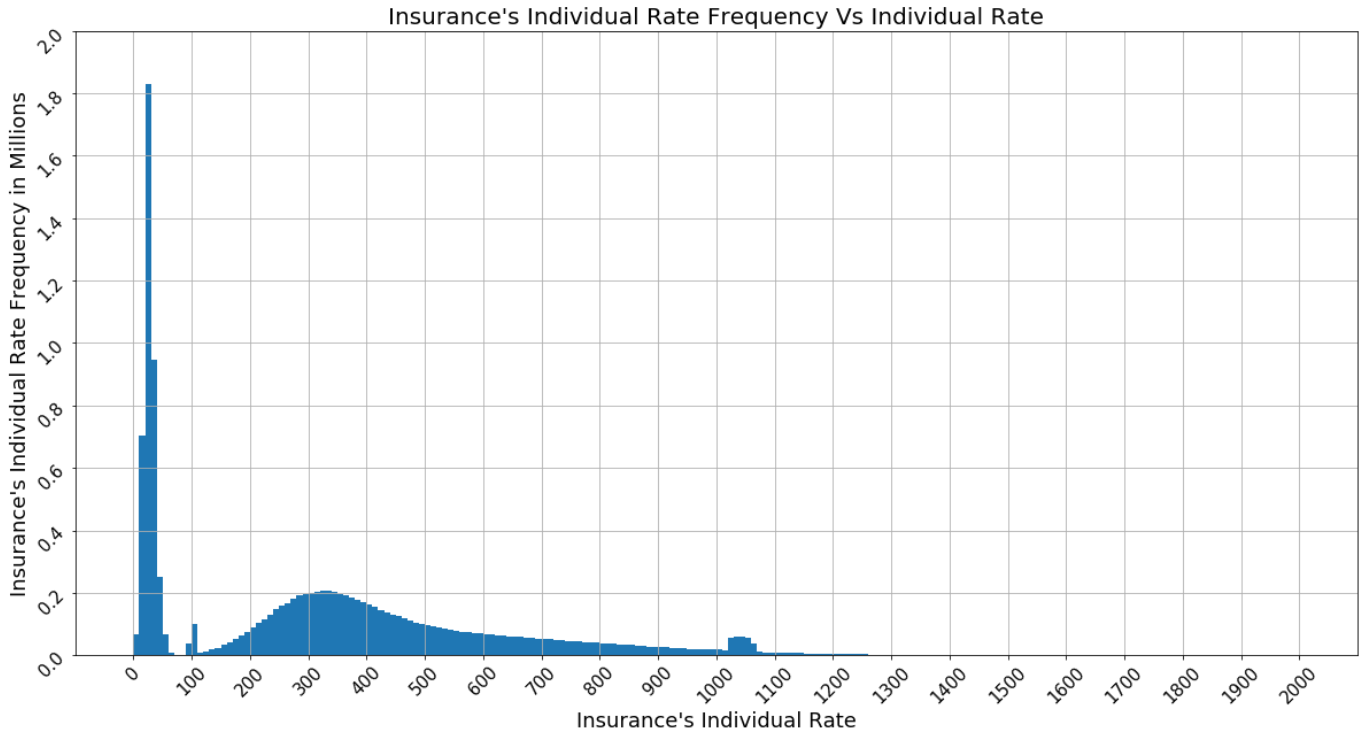


Fig 14

- Does this data look more sensible?

  Yes, this data looks more sensible as the values are more distributed and there doesn't seem to be having any outliers.

- Describe the data. How many groups can you see?

  From the above data we can see mainly 4 groups. They are
  - First group is between 0 -100 and most of the users lies in that range.
  - Next a more distributed group between 100 – 1000 with most no of users between 300 – 400.
  - Next is a group between 1000 – 1100 as it has more no of users compared to 900 – 1000.
  - Last group is above 1100.

B3. Variation in Costs across States

How do insurance costs vary across states?

1.Generate a graph containing boxplots summarising the distribution of values for each state.

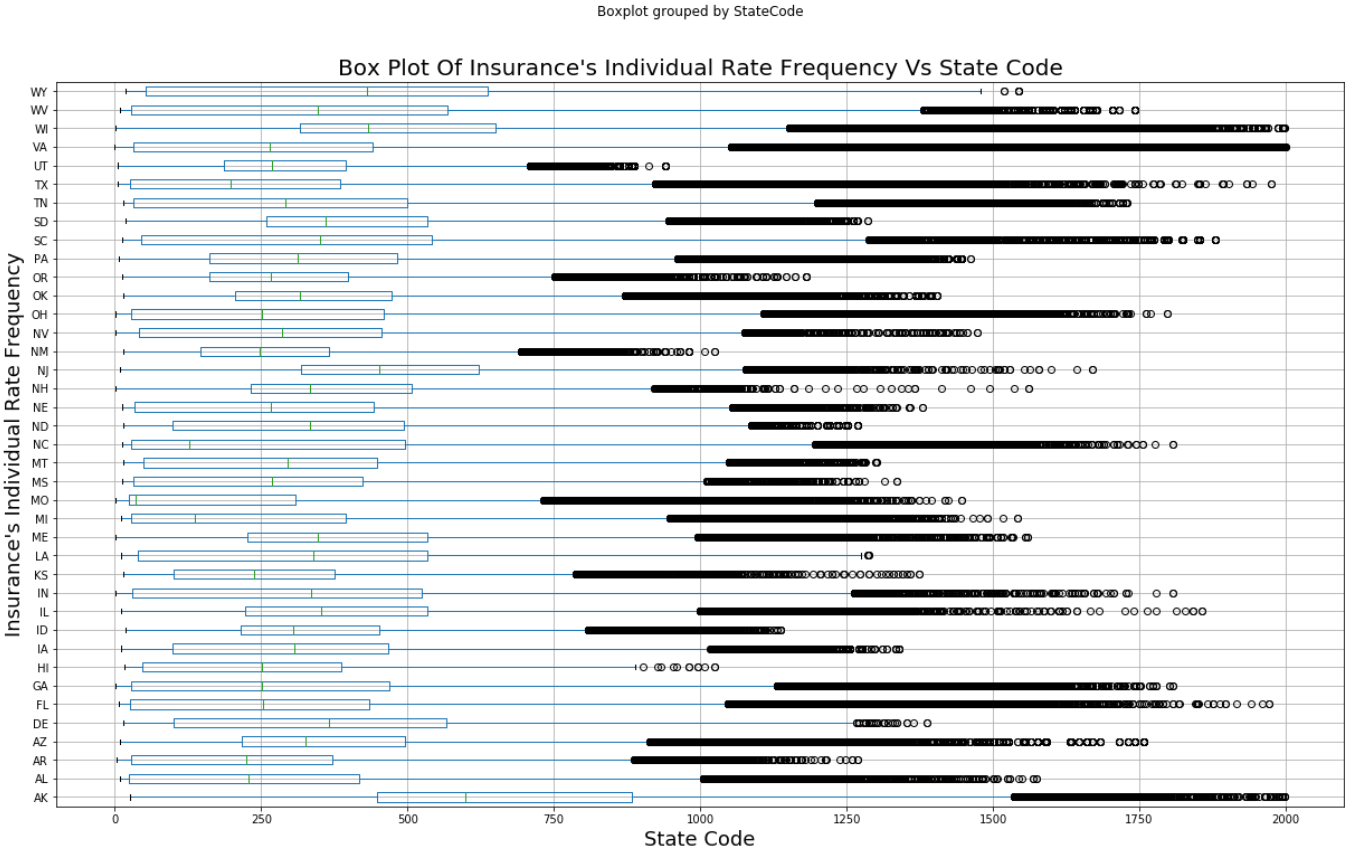


Fig 15

- Which state has the lowest median insurance rates, and which one has the highest? (Hint: you may need to rotate the state labels to be able to read the plot.)

```
State AK has the highest median insurance rate
State MO has the lowest median insurance rate
```

2.Does the number of insurance issuers vary greatly across states?

- Create a bar chart of the number of insurance companies in each state to see. (Hint: you will need to aggregate the data by state to do this.)

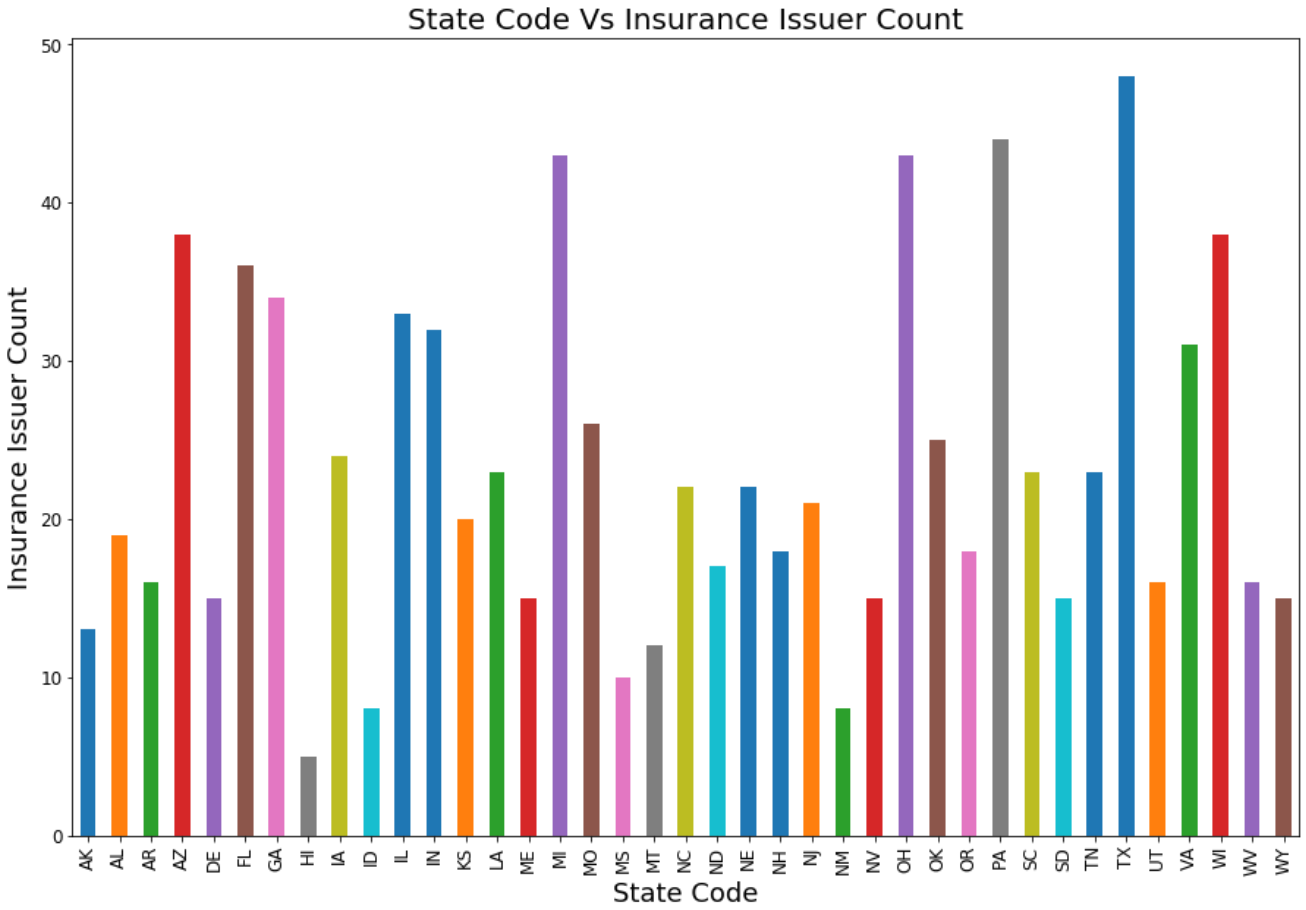


Fig 16

From the above graph we can clearly see that the number of issuers vary greatly across the states

3.Could competition explain the difference in insurance premiums across states?

- Use a scatterplot to plot the number of insurance issuers against the median insurance cost for each state.

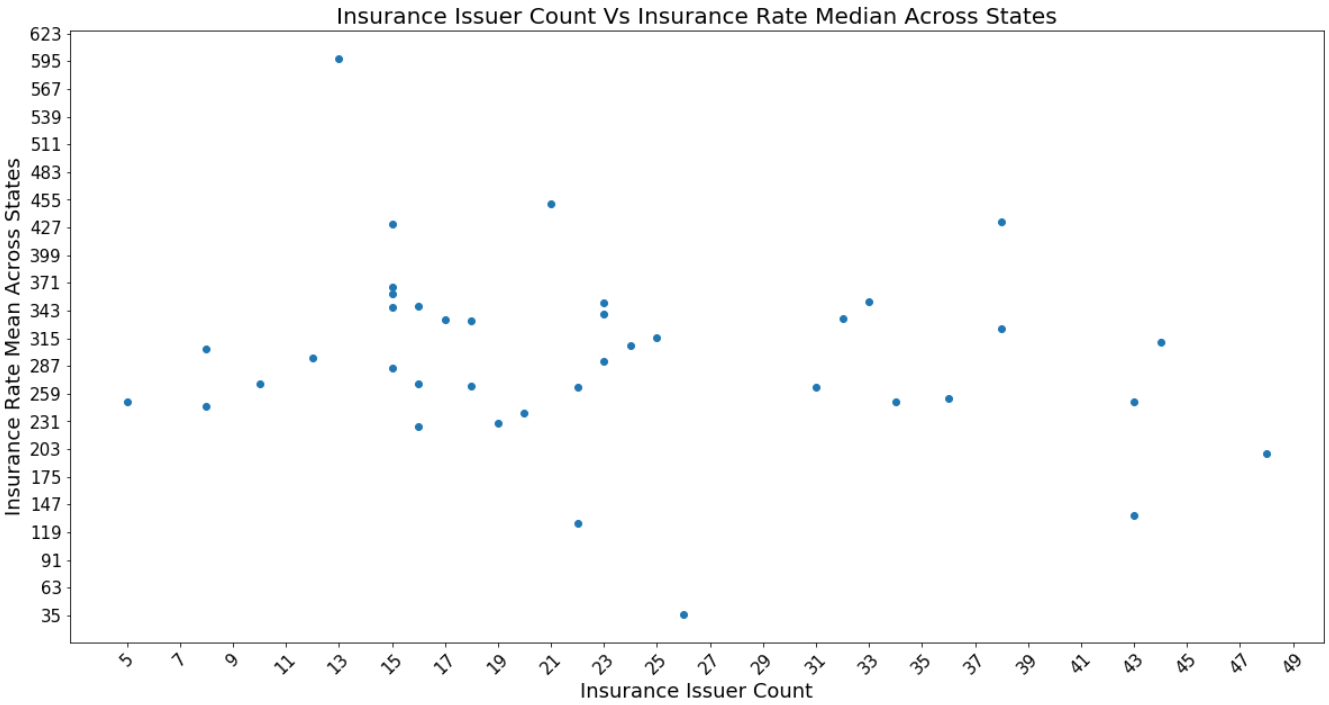


Fig 17

- Do you observe a relationship?

  Yes, competition explains the difference in insurance premiums across states. We can see in the above graph that as the no of Issuers count increases the median across various states comes more closer towards the overall median (ie 306.97) of the data. It is also the reason why most of the data states cost is approximately near the overall median of the entire data.

B4. Variation in Costs over Time and with Age

Generate boxplots (or other plots) of insurance costs versus year and age to answer the following questions:

1.Are insurance policies becoming cheaper or more expensive over time?

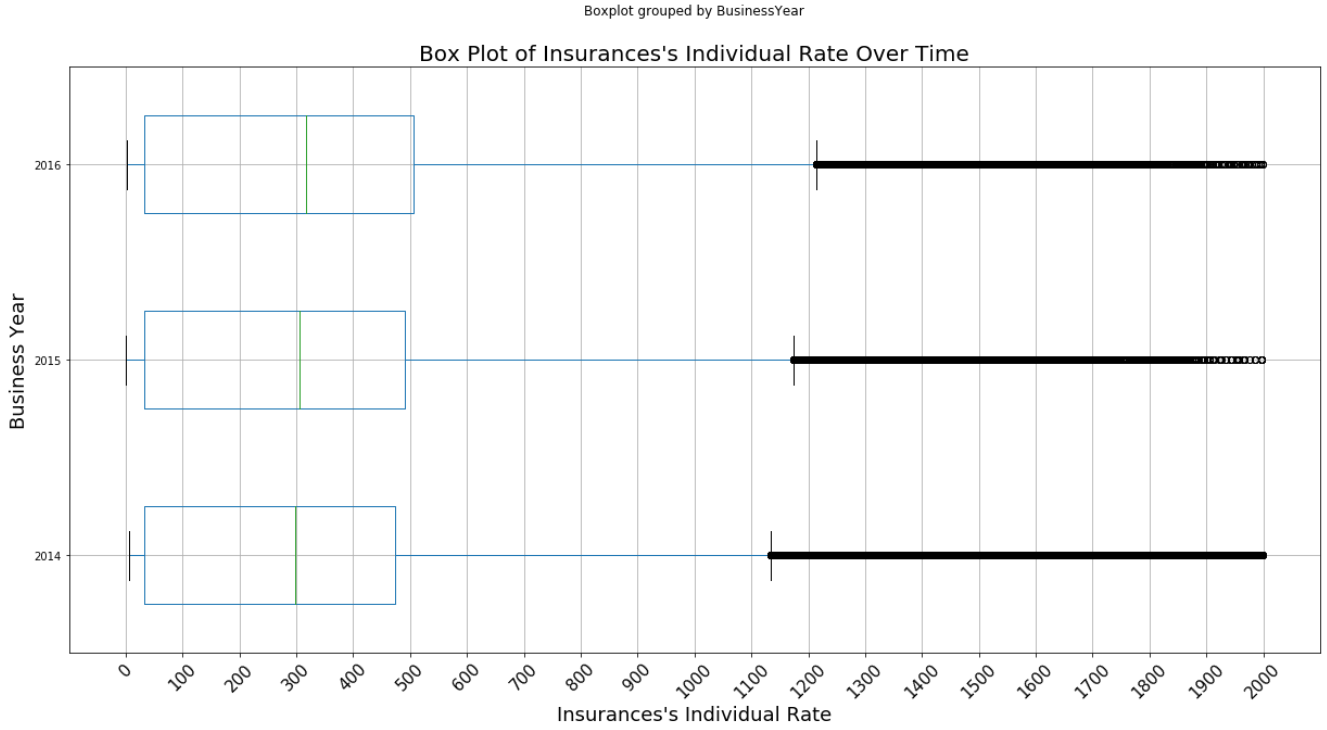


Fig 18

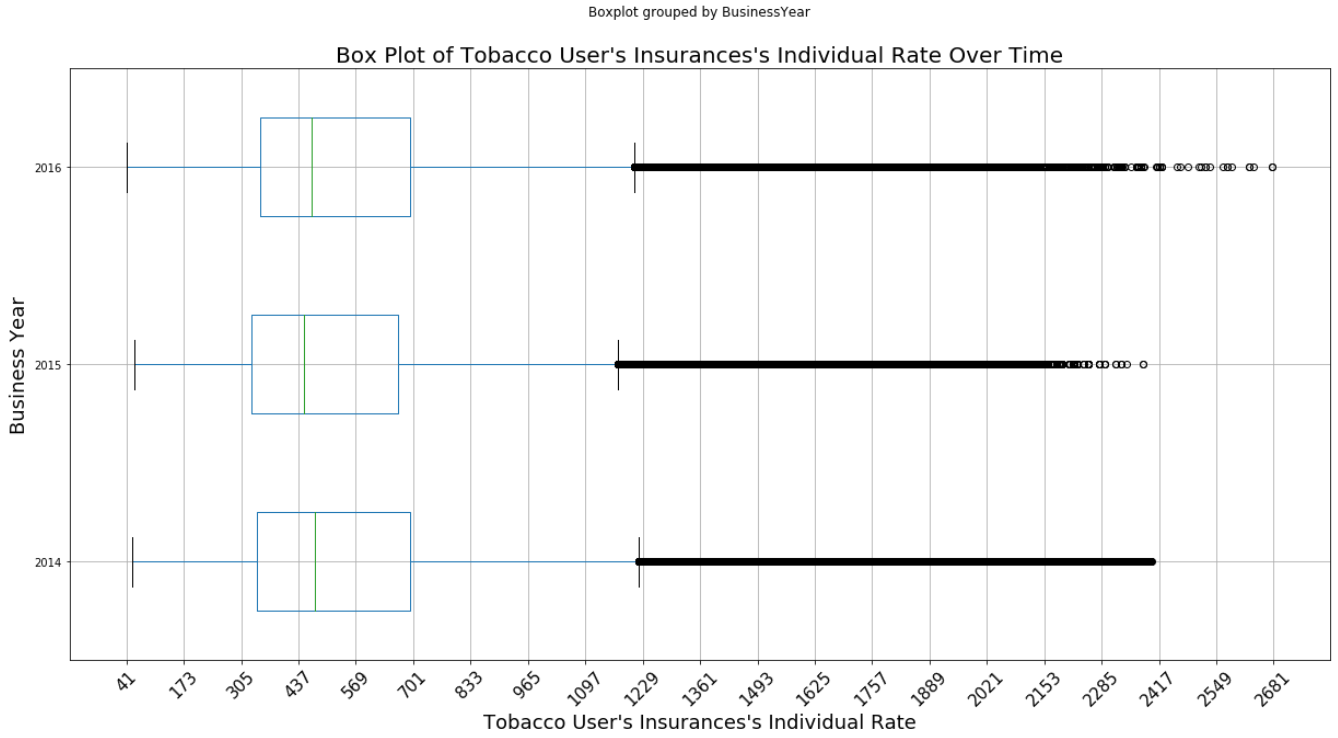


Fig 19

- Is the median insurance cost increasing or decreasing?

  From Fig 18 we can clearly see that the median of individual insurance rates of users who doesn't use tobacco keeps on increasing over time.

From Fig 19 we can see that the median of individual insurance rates of the users who uses tobacco initially decreases and then increases over time.

2.How does insurance costs vary with the age of the person being insured? (Hint: filter out the value 'Family Option' before plotting the data.)

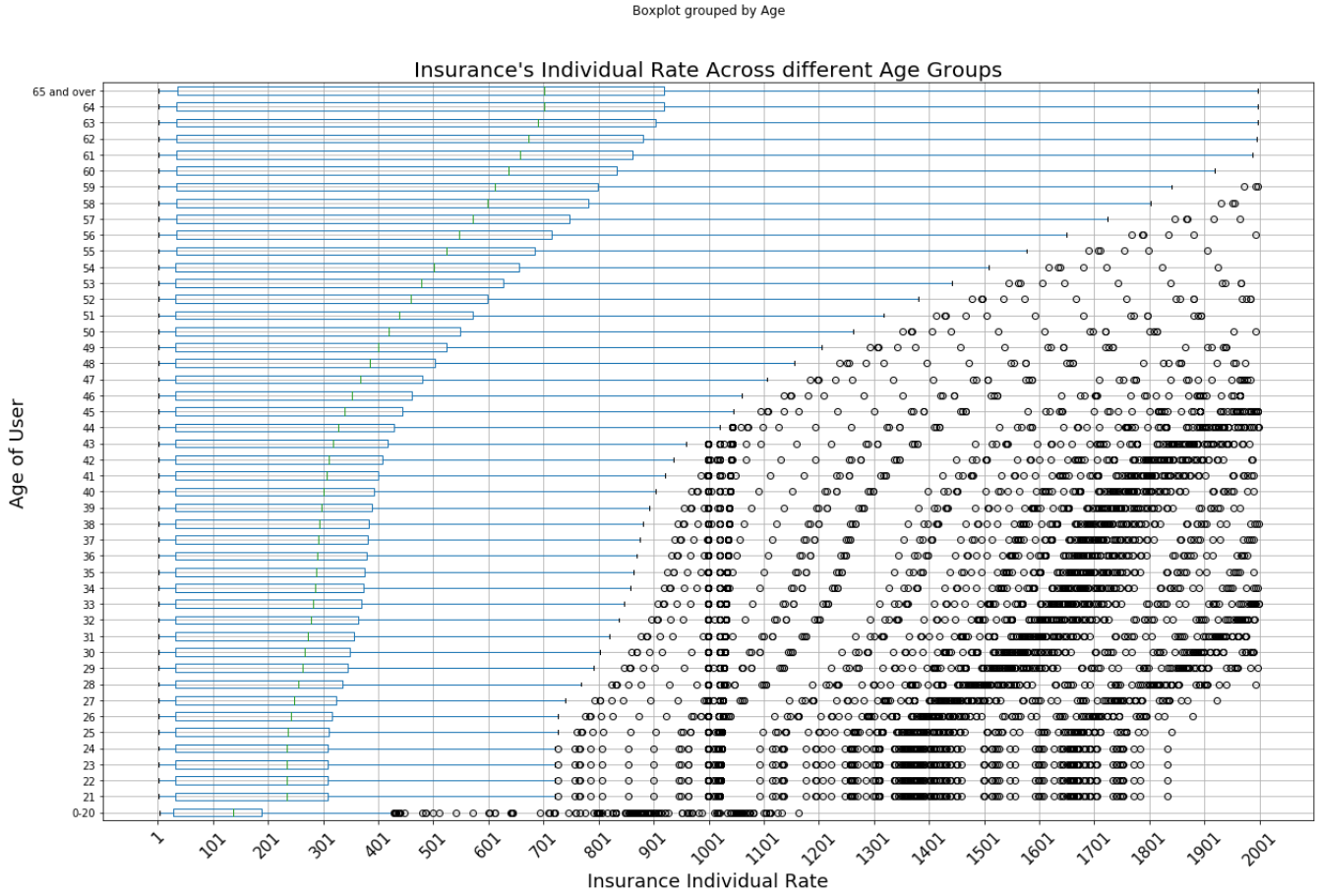


Fig 20

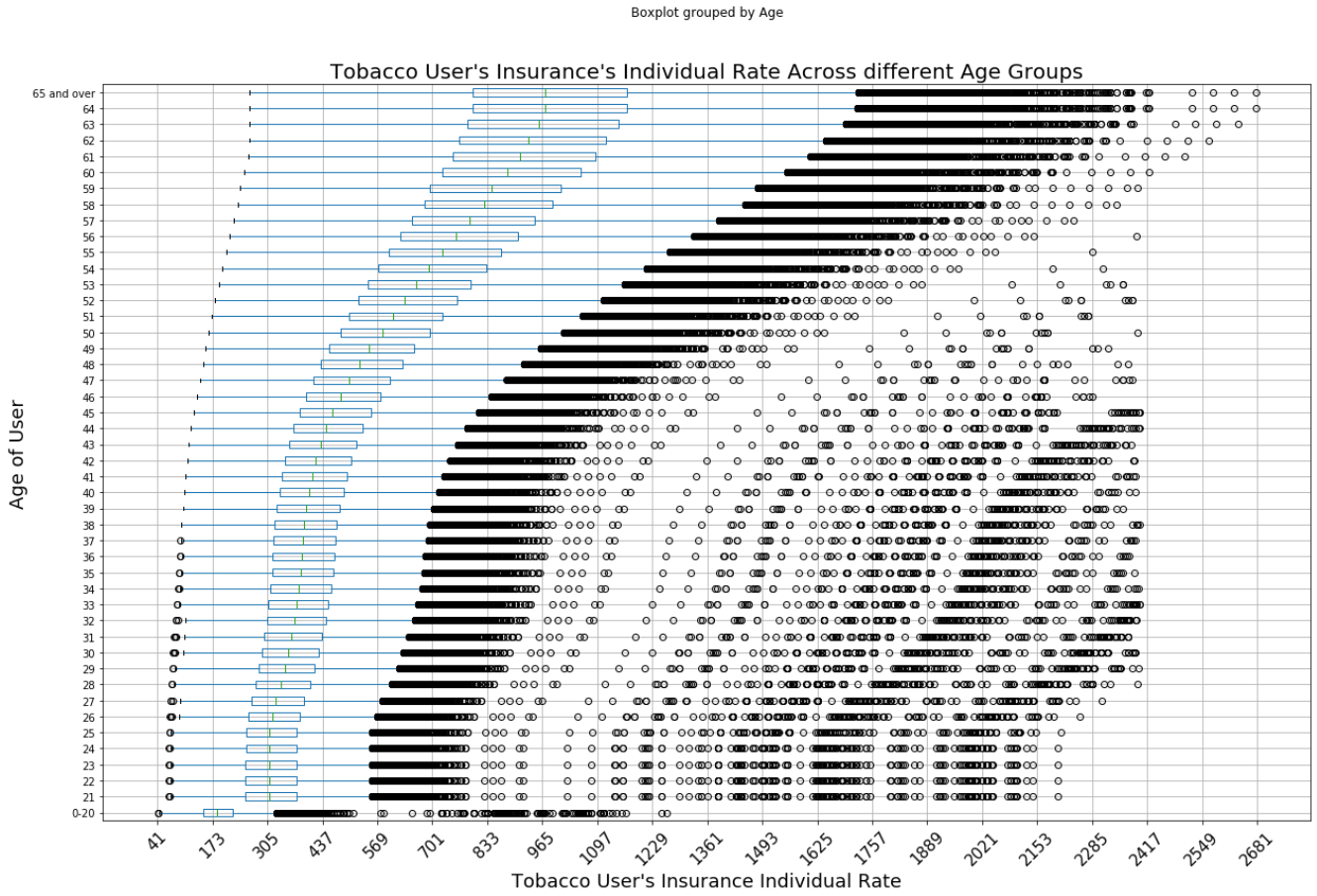


Fig 21

- In terms of median cost, do older people pay more or less for insurance than younger people? How much more/less to they pay?

  In terms of median cost older people pay more for insurance than younger people.
  Note that here the old users is assumed to be in category of 65 and over. Also, young people is assumed of the category (0-20). As per this the values are

  - Median of old user's Insurance Individual Rate is 702.21
  - Median of young user's Insurance Individual Rate of users is 138.59
  - Difference of Median of old users to that of young users Insurance Individual Rate is
    - Value: 563.625
    - In Percentage: 506.69%

  - Median of old user's Insurance Individual Rate having to bacco is 972.51
  - Median of young user's Insurance Individual Rate of users having tobacco is 184.36

- Difference of Median of old users to that of young users Insurance Individual Rate having tobacco is
  - Value: 788.155
  - In Percentage: 527.51%

Task C: Exploratory Analysis on Other Data

(Note: This additional task is for those students wishing to get higher grades for their assessment. It is not required to pass the assignment, but it is required to get higher credit. Please refer to the marking rubrics for more details)
Find some publicly available data and repeat some of the analysis performed in Tasks A and B above. Good sources of data are government websites, such as: data.gov.au, data.gov, data.gov.in, data.gov.uk, ...
Please note that your analysis should at least contain visualisation, interpretation of y our visualisation and a prediction task.

Notifiable Infectious Diseases Reports in UK From 2016 - 2018 July

Data Description: Notifiable Infectious Diseases Reports in UK for past Three years

Published by: OpenDataNI

Last updated: 27 July 2018

Topic: Health Licence: Open Government Licence

Summary

Cumulative data showing the breakdown of the notifiable infectious diseases reported to Public Health. The report is based on figures recorded by the Duty Room and Surveillance and is generated on a weekly basis.Source Link : https://data.gov.uk/dataset/6bf61328-a250-44fd-a787-481503f02865/notifiable-infectious-diseases-reports

Final Data After Merging will look like

	Disease	Week Number	Weekly Report	Year	Yearly Report
101	Tuberculosis (Non Pulmonary)	Week 01	0	2018	0
102	Typhoid	Week 01	0	2018	0
103	Typhus	Week 01	0	2018	0
104	Viral Haemorraghic Fever	Week 01	0	2018	0
105	Whooping Cough	Week 01	0	2018	2

Fig 22

Aggregated Data Information

Here in this data the Disease column contains the names of the various infectious diseases

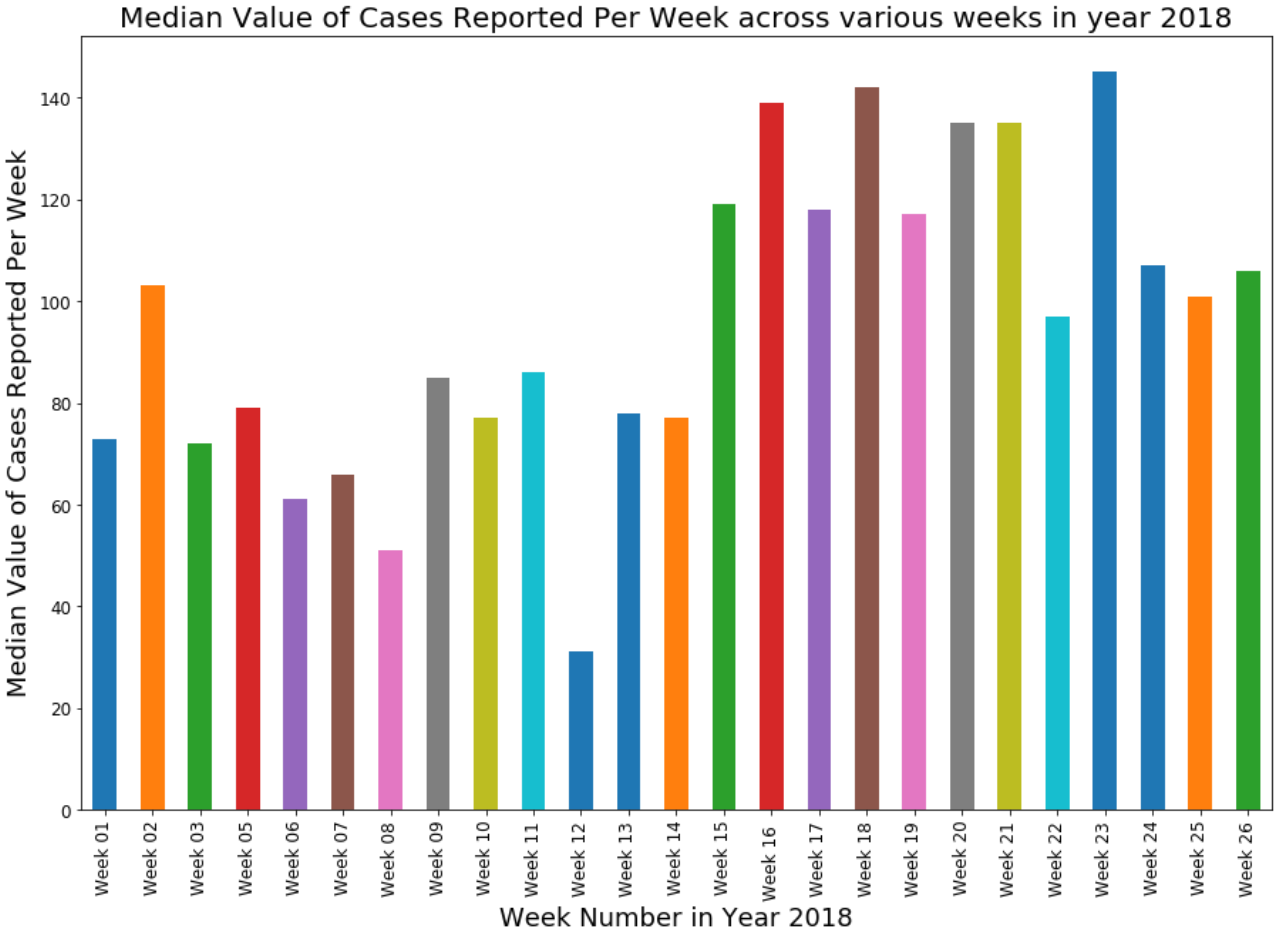
Week number represents the data is of which week in the year 2018

Weekly Report denotes the no of cases reported in the current week

Year represents the current year that is been considered for the Yearly report

Yearly report represents the no of cases reported in till the current week in the Year present in the year column

Lets First look at the general over view of the no of cases reported per week



From the above graph we can see that the data is approximately distributed all over the various weeks

Now let’s consider the simple plot of cases reported based on the various infectious diseases found

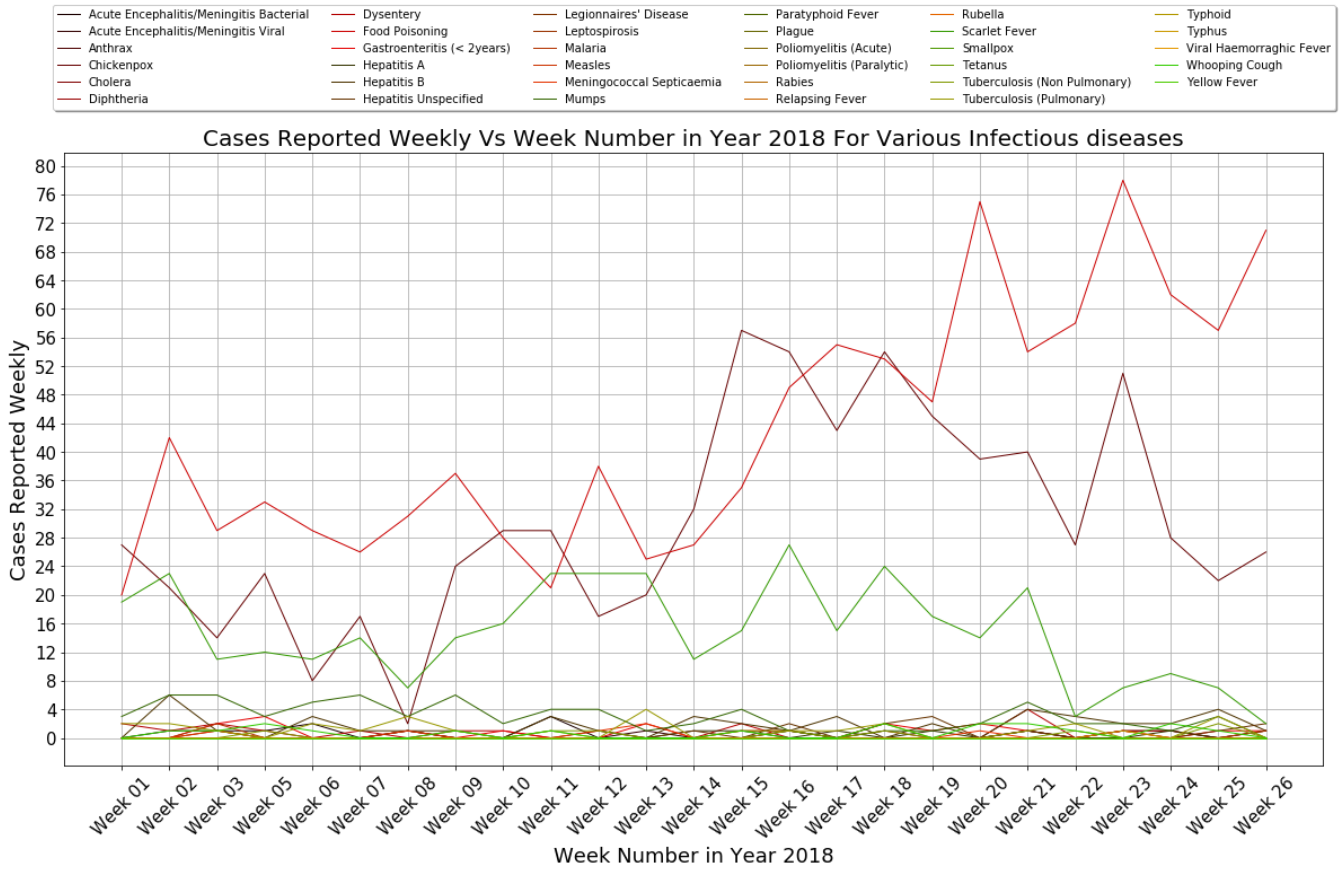


Fig 23

Here the Figure 23 represents the weekly reported cases for various infectious diseases in the year 2018

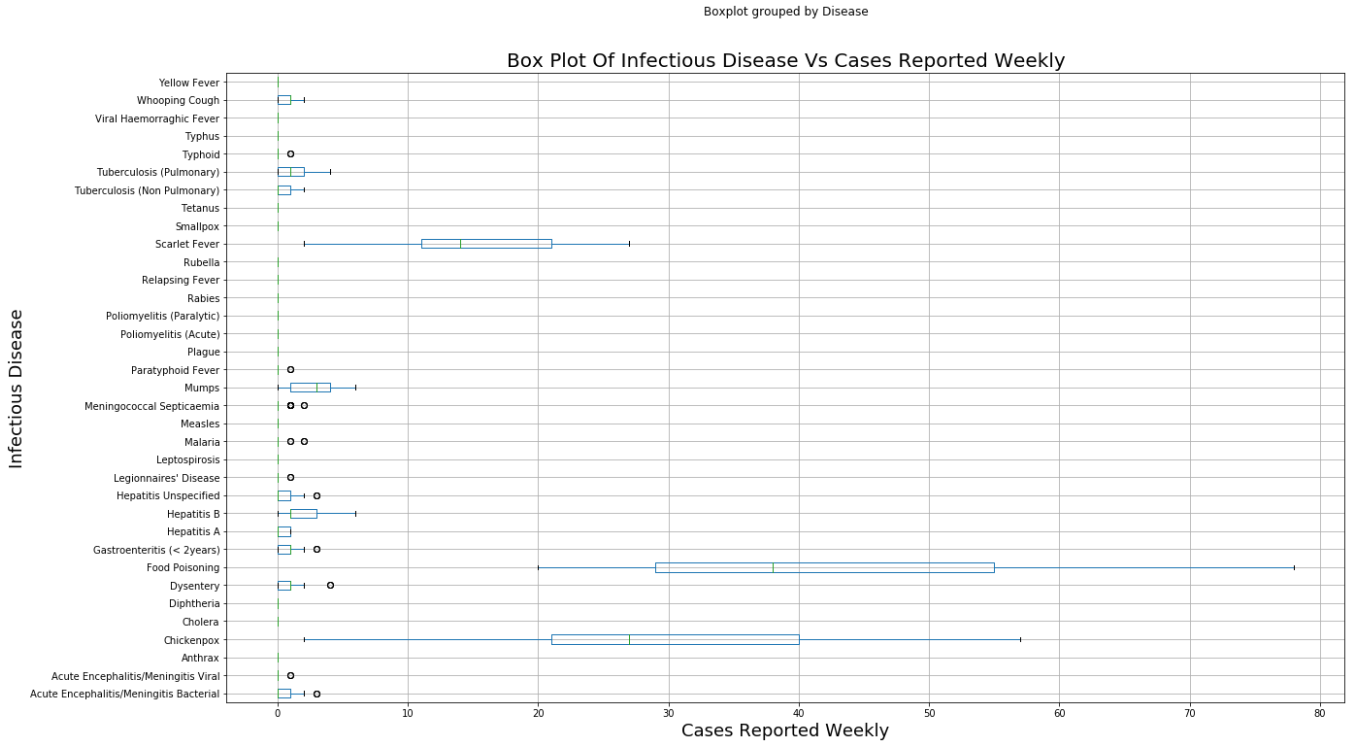


Fig 24

From this Figure 24 it is evident that for the year currently the major infectious disease that acts as a treat are chickenpox, Food Poisoning and Scarlet Fever

So Now lets just consider these three diseases for further processing. Then the plot will look like

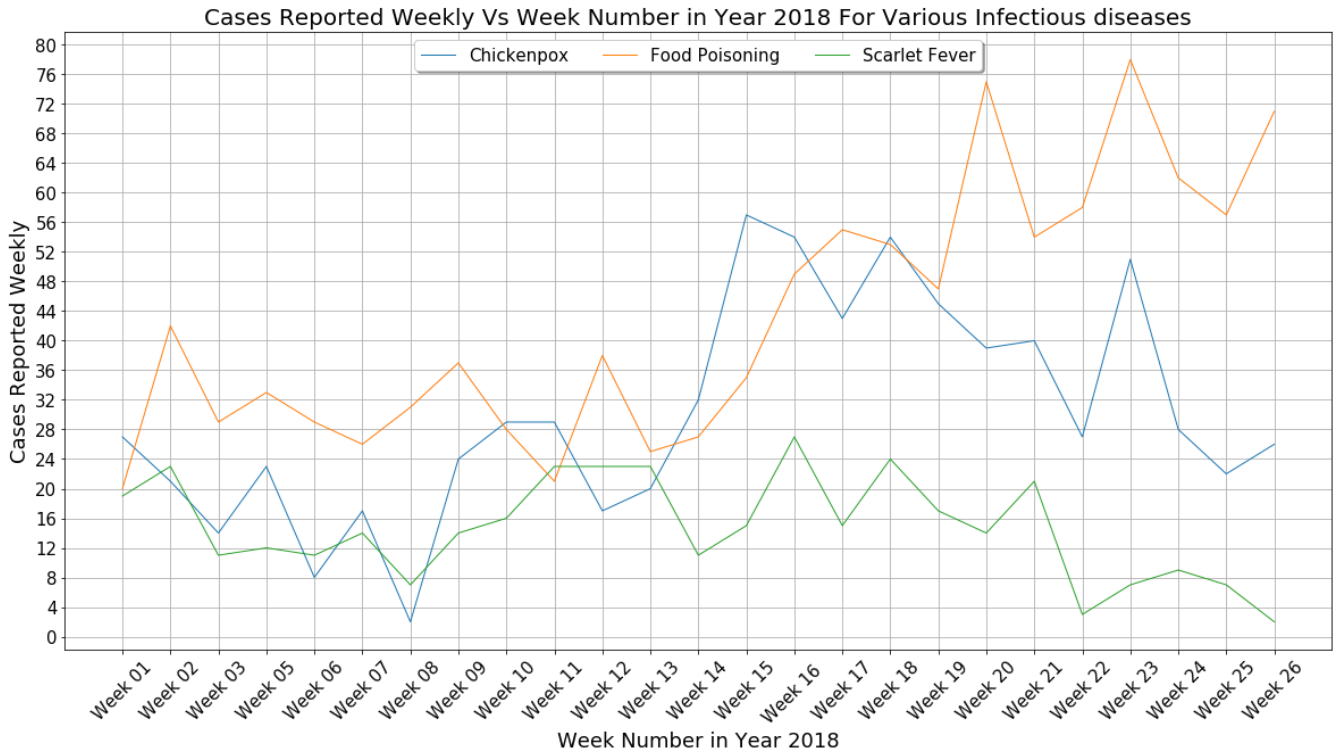


Fig 25

Now lets try to plot a Linear Fit for the above three diseases and also try to predict the possible future behaviour of these disease b ased on the linear fit we found out earlier

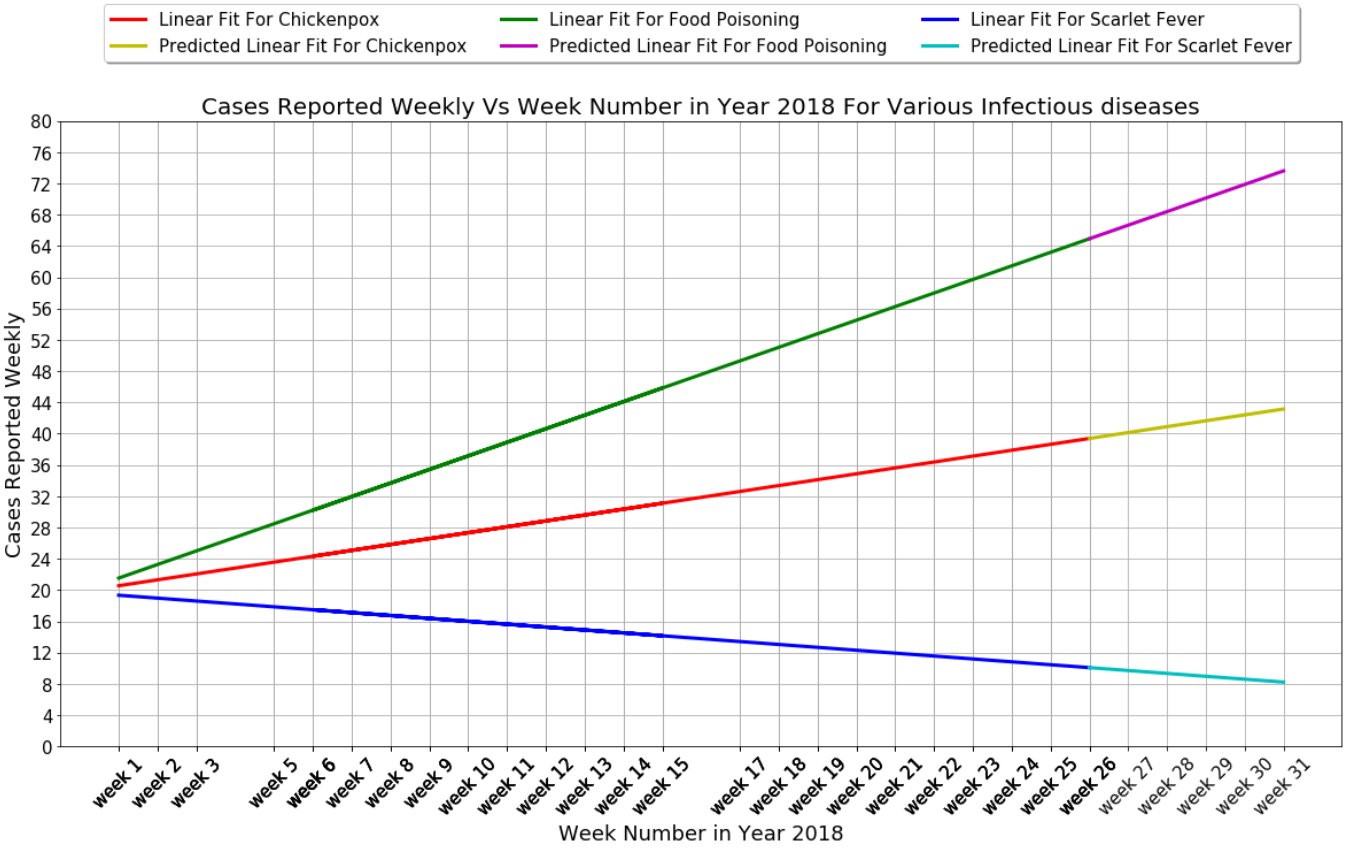


Fig 26

From the above figure the chances of cases of Food poisoning and chicken pox is going to rise in the next 5 weeks. Whereas the cases of scarlet Fever are going to decrease.
From this Info we must be more prepared for diseases such as food poisoning and chicken pox compared to other diseases

The data representation of the data in Motion Chart will look like

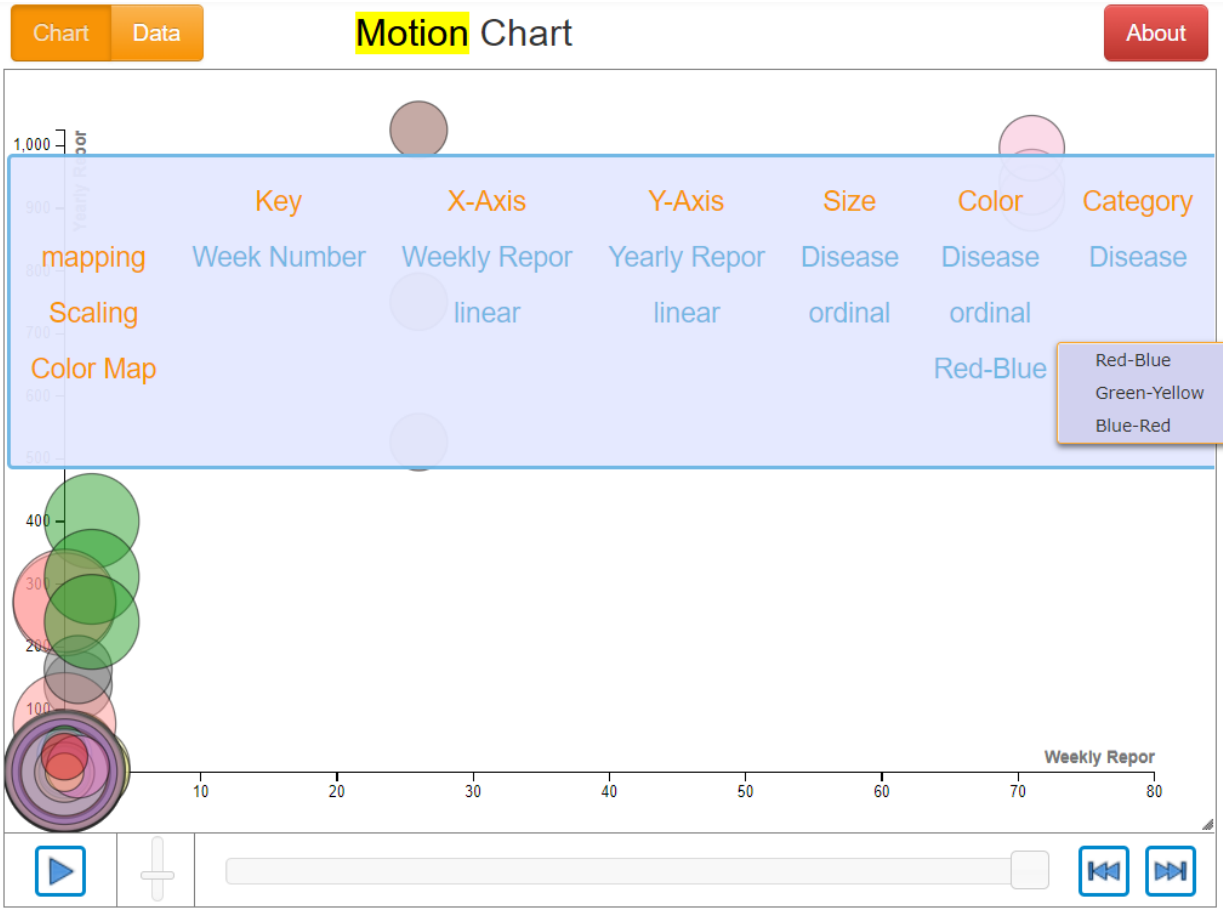


Chart
Data
Motion Chart
About
Key
X-Axis
Y-Axis
Size
Color
Category
mapping
Week Number
Weekly Repor
Yearly Repor
Disease
Disease
Disease
Scaling
linear
linear
ordinal
ordinal
Color Map
Red-Blue
Red-Blue
Green-Yellow
Blue-Red
1,000
400
100
Weekly Repor
10
20
30
40
50
60
70
80