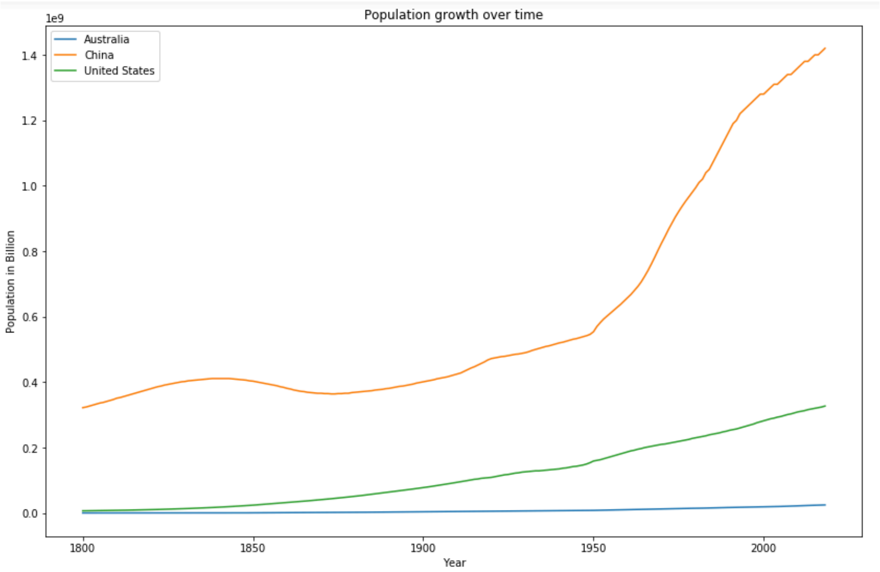
Table of Contents

Task A
Investigating Population and Gender Equality in Education

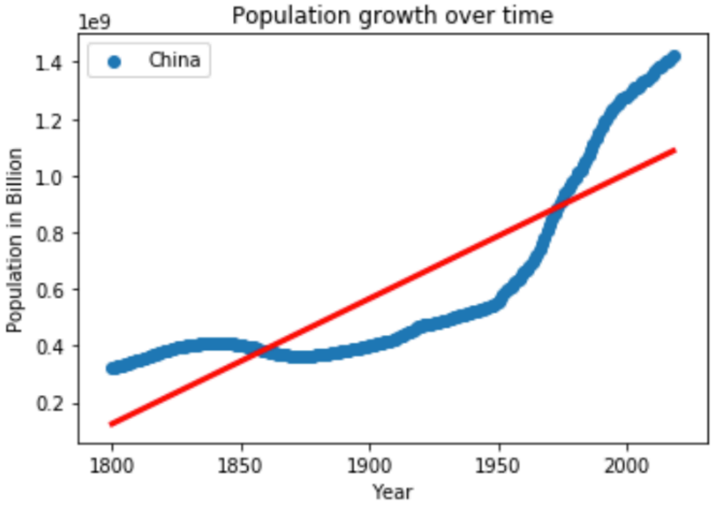
A1. Investigating the Population Data

1. Below is the plot of population growth of Australia, China and United States over time.

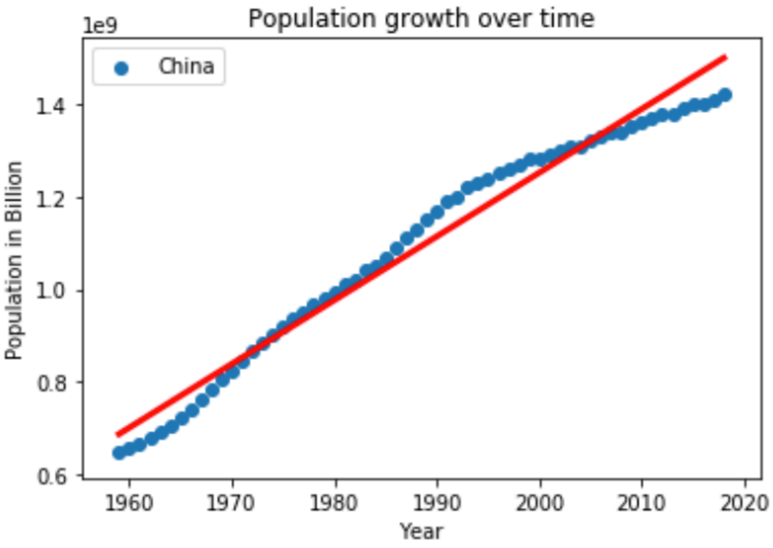


The population values of all the counties are **increasing** over time. However, population of china is on a steeper growth compared to **United States and Australia.**

2. Below is the chinese population plot and its linear fit.



The liner fit doesn't align great for the existing data set.
Using linear fit, prediction of the resident population in China during **2020 is** $\mathbf{1.0957 * 10^9}$ **and 2100 is** $\mathbf{1.4486 * 10^9}$**.**
Chinese population plot and its linear fit for data from the year 1960 onwards.

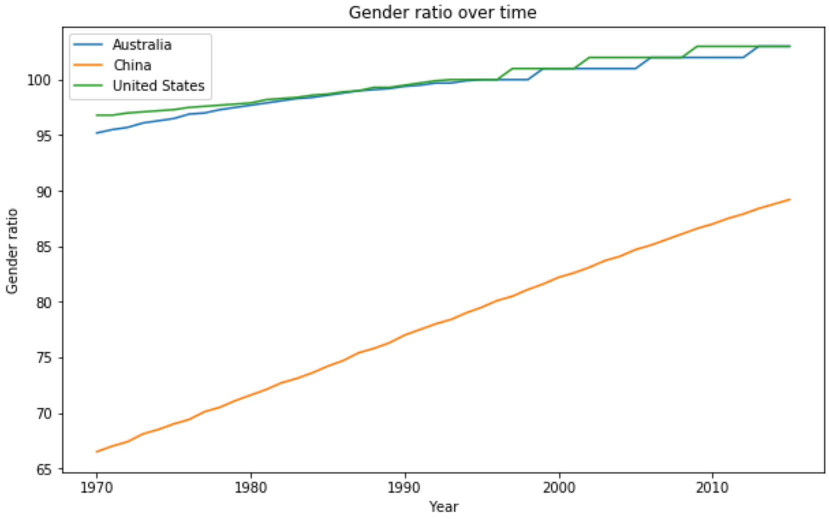


The liner fit is almost perfect for the recent data set.
Using linear fit, prediction of the resident population in China for the recent years during **2020 is $\mathbf{1.5269 * 10^9}$ and 2100 is $\mathbf{2.6283 * 10^9}$.**

The recent popultion data set of China gives a better prediction of the future as the steep increase of the population has occurred post 1950.

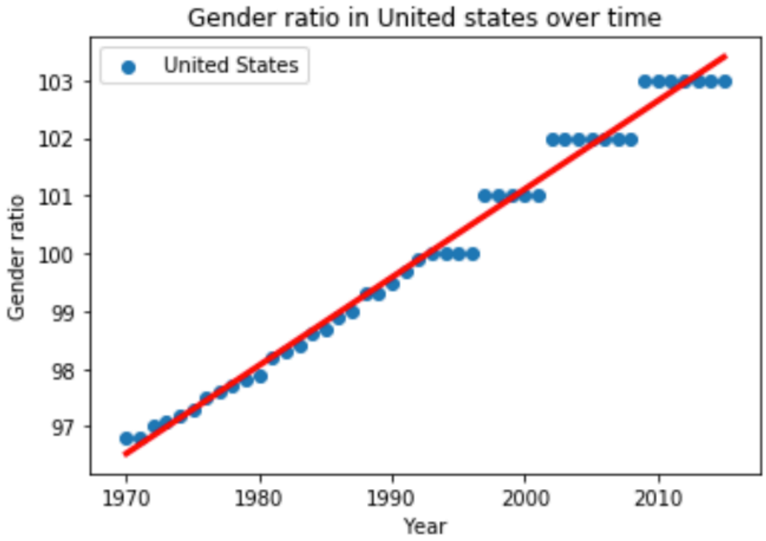
A2. Investigating the Gender Equality Data

1. Plot of gender ratio (women % men) in schools for Australia, China and United States over time.



103.0 is the maximum and **95.2 is the minimum** values for gender ratio (women % men) in Australia over the time period.
Gender ratio of all the three counties have **increased** over the period of time.
Australia and United states have shown similar growth trends over the years.

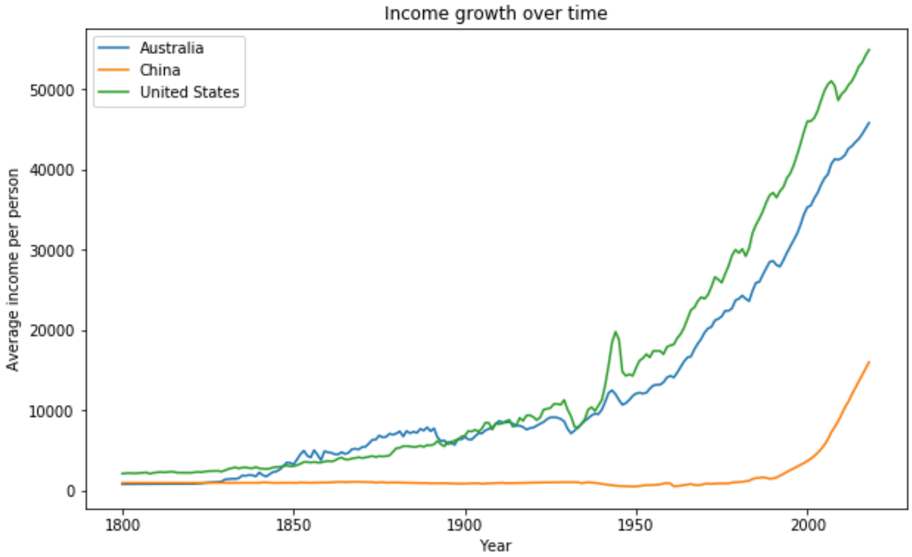
2. Below is the plot for gender ratio in schools in United States and its linear fit.



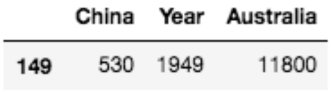
The liner fit to the gender ratio in schools in United States has a steep slope which depicts an obstructive growth of the ratio and cannot offer a reliable prediction over the period of time.

A3. Investigating the Income Data

1. Below is the plot for Income of Australia, China and United States over time.



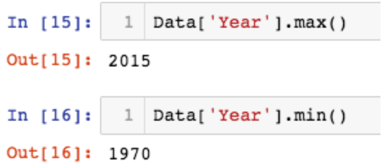
Minimum income recorded in **China is 530 during 1949**. However, income recorded in **Australia during 1949 is 11800**.



	China	Year	Australia
149	530	1949	11800

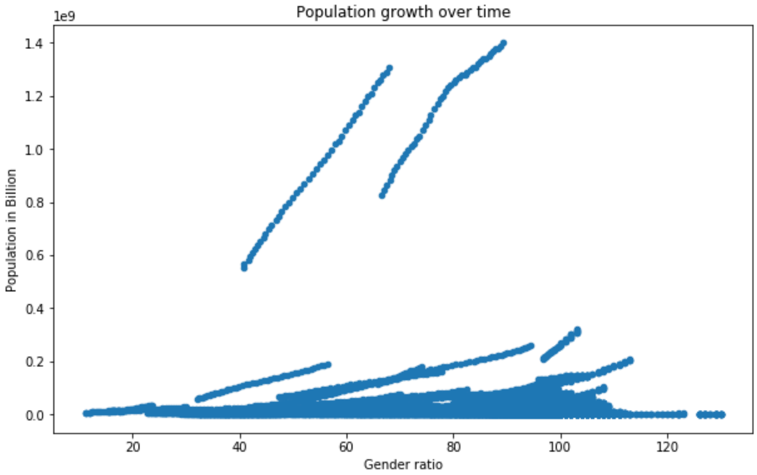
A4. Visualising the Relationship between Gender Equality and Population

1. Data from the different files are **melted down** into a single table containing **population** values, **income** and **gender ratio** in schools for the different **years** and different **countries**.
   First year of the combined data is **1970** and the last year is **2015**.



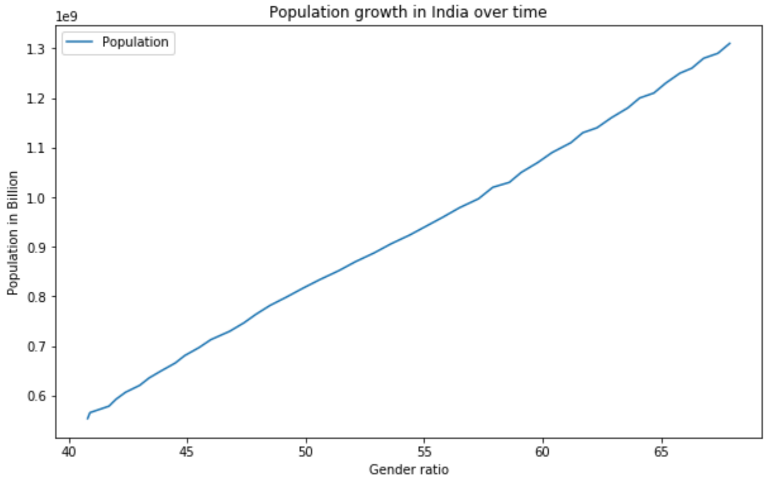
```
In [15]: Data['Year'].max()
Out[15]: 2015

In [16]: Data['Year'].min()
Out[16]: 1970
```

2. Below plot is between **gender ratio** in schools and **population**.



   We observe an increase trend with population and gender ratio over the years. However, don't see any significant relationship between gender ratio and population as the data contains information of all the counties and such a plot doesn't portray a strong relationship.

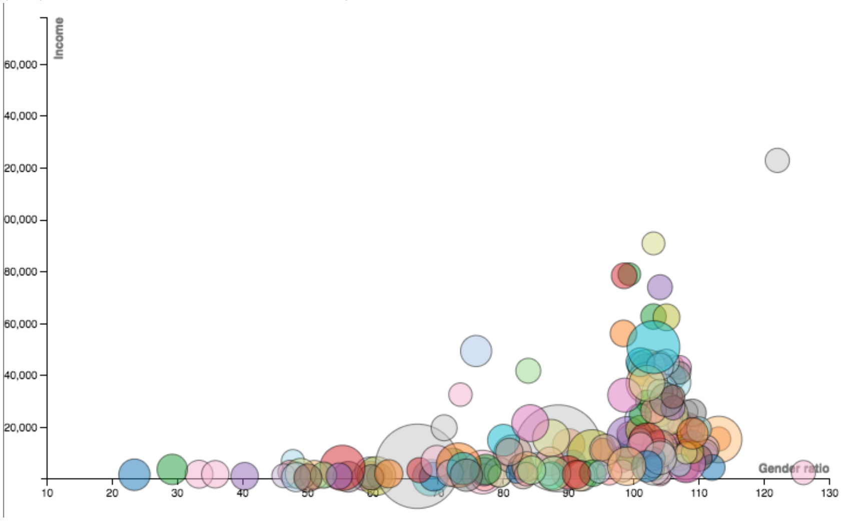
3. The relationship between gender ratio in schools and the population in India is plotted below.



The relationship between gender ratio in schools and the population in India is observed to be a linear **direct or positive relation** between gender ratio and population.

A5. Visualising the Relationship over Time

1. Motion Chart to compare the gender ratio in schools, the income, and the population of each country over time.



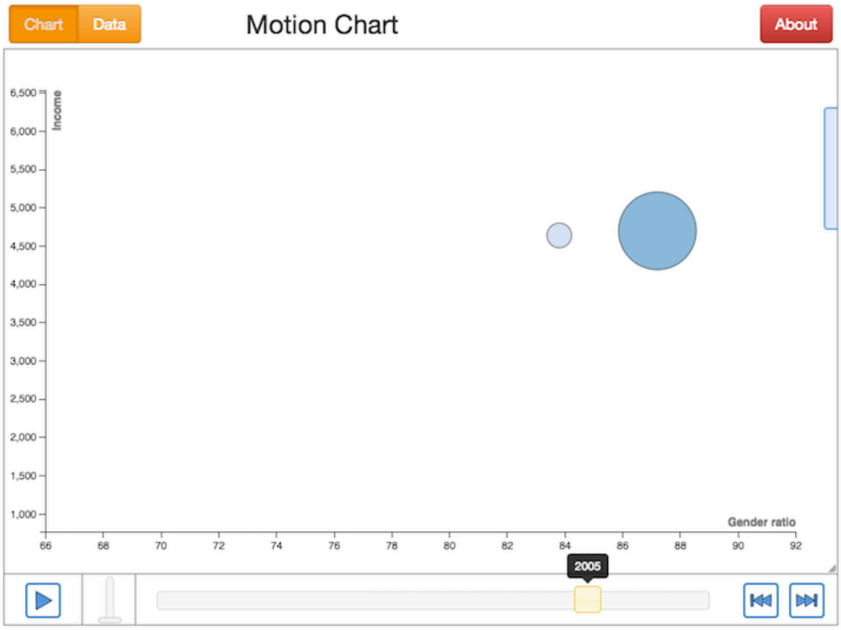
2.

Yemen and **Afghanistan** generally have the lowest gender ratio (women % men) in schools over the years.

Lesotho has the highest gender ratio during the whole period of time.

The gender ratio across all the counties is **generally increasing** during the whole time. On the other hand, income has an overall collective increase. However, Income of certain counties like United Arab Emirates and Kuwait have increased steeply and plummeted which shows that the individual counties incomes have been **bobbing (movement up and down)** during the whole period of time.

Below is the motion chart for the gender ratio in schools on the x-axis, the income on the y-axis, and the bubble size should depend on the population of **Bolivia and Cape Verde**.
We observe that Cape Verde will start to have a higher income than Bolivia from the year **2005**. However, Cape Verde fails to surpass Bolivia's gender ratio. According to the data, Bolivia still has a higher gender ratio than Cape Verde.



The amount of income and gender ratio (women % men) in schools in all countries during the whole period of time is **generally increasing** which leads to a **positive relationship** on a whole.This could be because of the awareness of gender equality among the society of most of the countries.

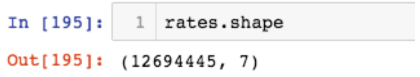
The general **population has increased** in most of the counties over the period of time.
The **income of gulf countries** were significantly high compared to rest of the world over the period of 1972 to late 2009. However, the difference has collapsed for most of the gulf countries.

Task B

Exploratory Analysis on Big Data

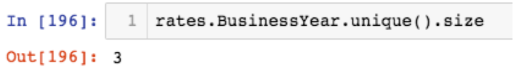
B1. Load InsuranceRates data.

1. There are **12694445 rows** and **7 columns** in the given InsuranceRate data.



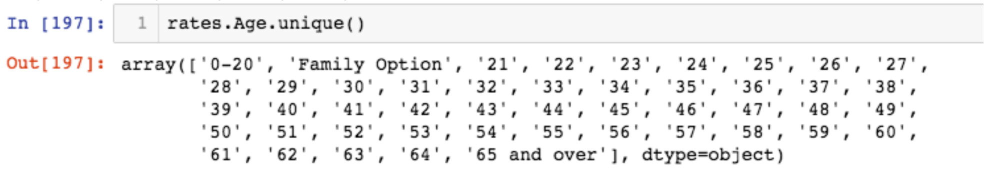
```
In [195]: rates.shape
Out[195]: (12694445, 7)
```

2. Data covers **3 unique years**.



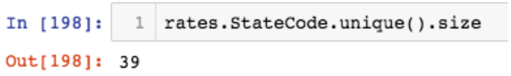
```
In [196]: rates.BusinessYear.unique().size
Out[196]: 3
```

3. Possible values for the “Age” column are **'0-20', 'Family Option', '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', '47', '48', '49','50', '51', '52', '53', '54', '55', '56', '57', '58', '59', '60','61', '62', '63', '64', '65 and over'.**



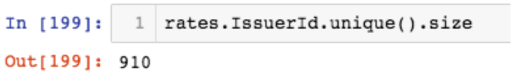
```
In [197]: rates.Age.unique()
Out[197]: array(['0-20', 'Family Option', '21', '22', '23', '24', '25', '26', '27',
                 '28', '29', '30', '31', '32', '33', '34', '35', '36', '37', '38',
                 '39', '40', '41', '42', '43', '44', '45', '46', '47', '48', '49',
                 '50', '51', '52', '53', '54', '55', '56', '57', '58', '59', '60',
                 '61', '62', '63', '64', '65 and over'], dtype=object)
```

4. There are **39 states** covered in the data.



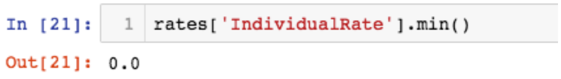
```
In [198]: rates.StateCode.unique().size
Out[198]: 39
```

5. There are **910 unique insurance providers** in the dataset.

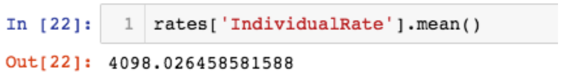


```
In [199]: rates.IssuerId.unique().size
Out[199]: 910
```

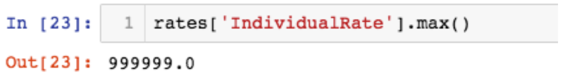
6. The average, maximum and minimum values for the monthly insurance premium cost for an individuals are **4098.0264, 999999.0 and 0.0** respectively.
The values don’t seem reasonable as the range is huge and it doesn’t make sense for few individuals to pay 0.0 towards the insurance and the others pay a huge amount for the same.



```
In [21]: rates['IndividualRate'].min()
Out[21]: 0.0
```



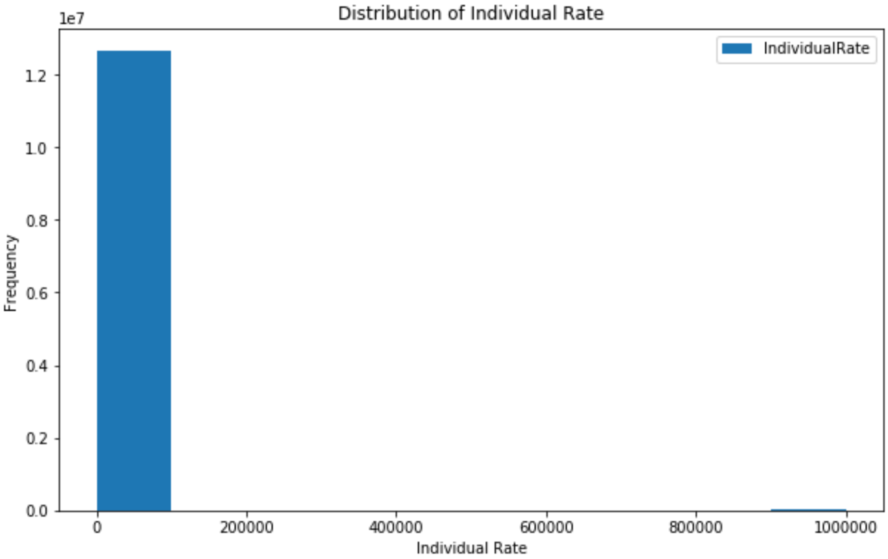
```
In [22]: rates['IndividualRate'].mean()
Out[22]: 4098.026458581588
```



```
In [23]: rates['IndividualRate'].max()
Out[23]: 999999.0
```

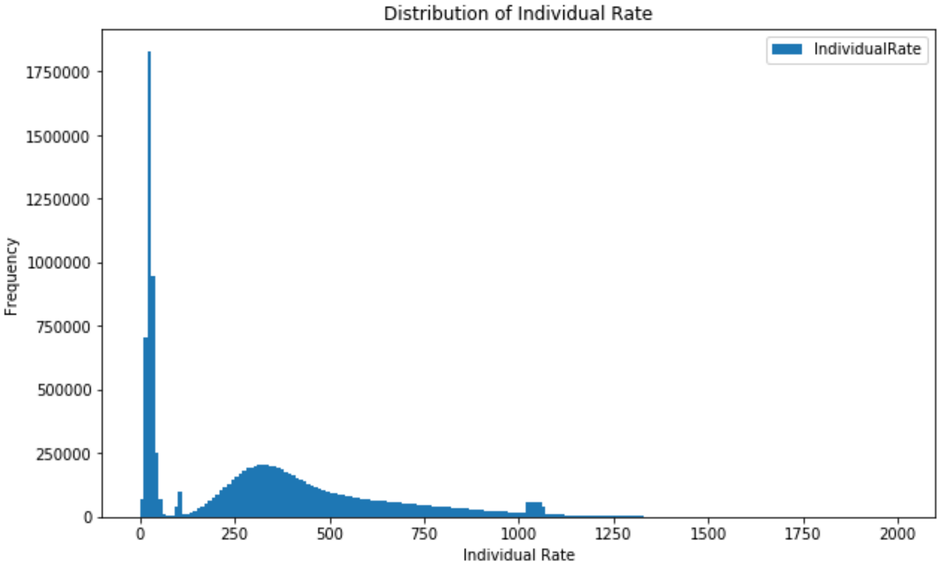
B2. Investigating Individual Insurance Costs.

1. Below plot shows the distribution of 'IndividualRate' values using a histogram



There is a vast difference in the distribution and it has outliers. There are few instances where the individual rates are way higher and close to 1000000 which shows the values could be a genuine set of high insurance payers or a measurement error.

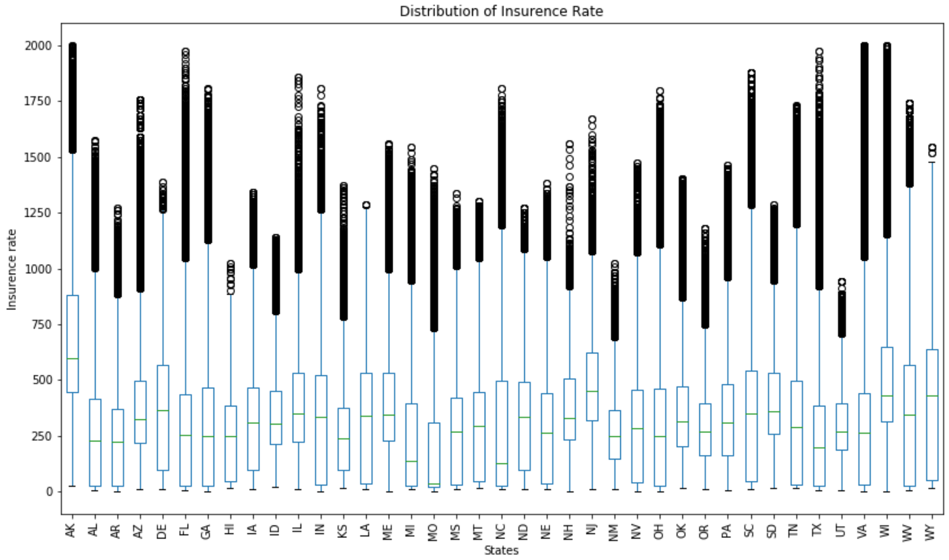
2. Below is the histogram plot for the insurance premiums between 0 and 2000.



Yes, the above plot of data looks **more sensible** with realistic distribution.
We notice majorly **two groups** of data with over 1750000 individuals paying between 0 and 60 for insurance. The other segment of individuals scattered between 100 and 1250 of individual rate bracket.

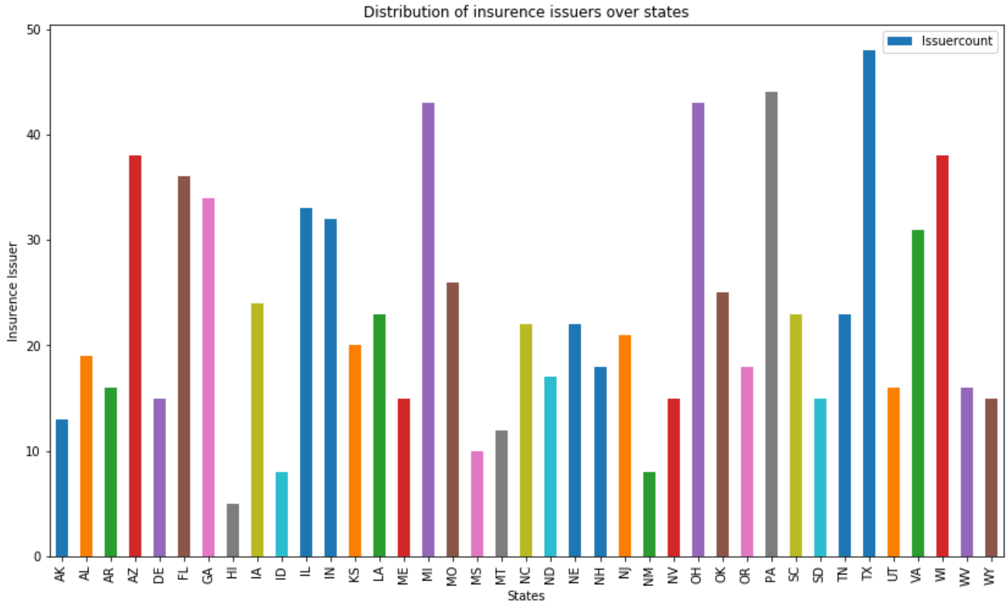
B3. Variation in Costs across States.

1. Below is the boxplot summarising the distribution of values for each state in the data set.

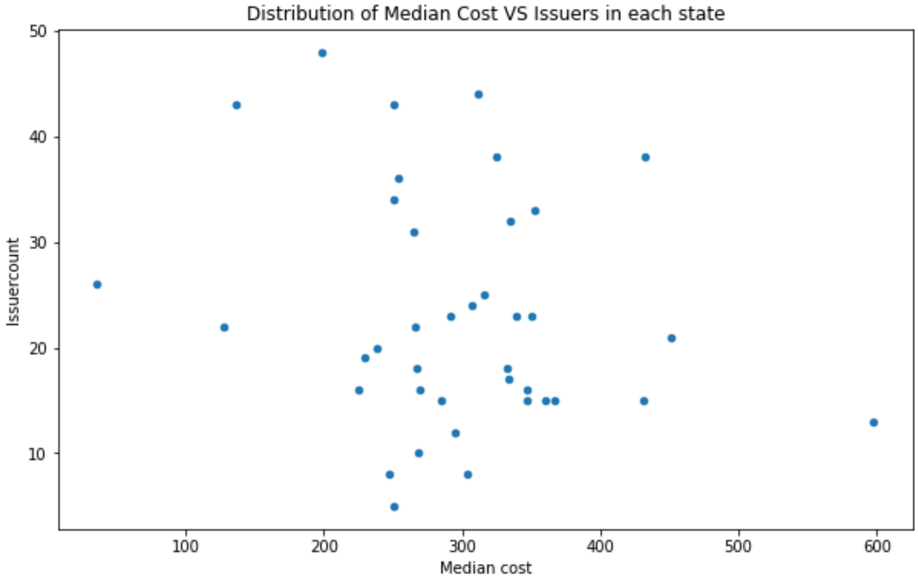


The state **‘AK’ has the highest** median insurance rate and state **‘MO’ has the lowest** median insurance rate.

2. Yes, the number of insurance issuers vary greatly across states. Below is the bar graph to represent the same.



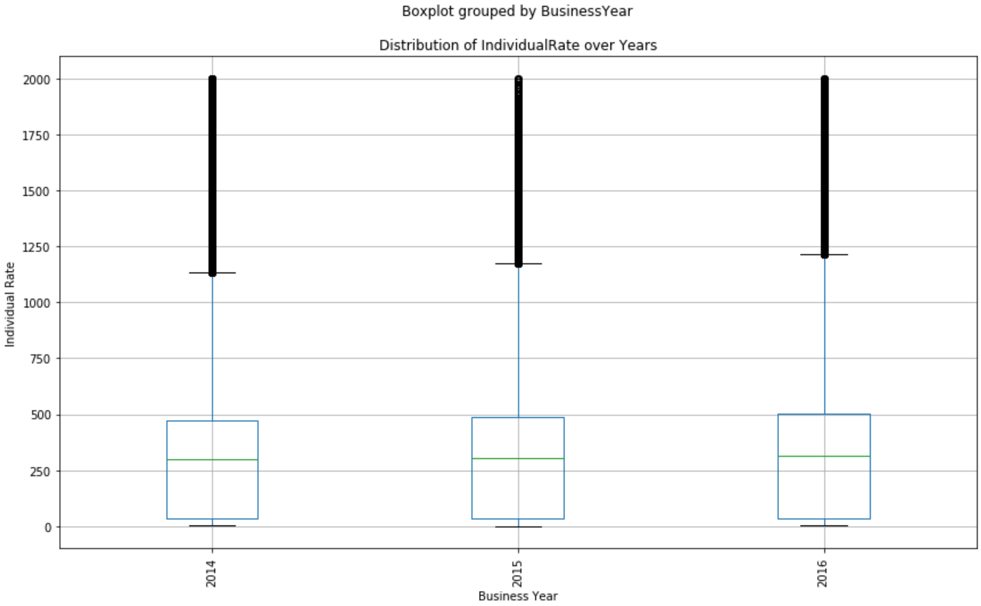
3. Below is the scatterplot for the number of insurance issuers against the median insurance cost for each state.



Most of the issuers have their median costs between **200-400** and this close margin pricing could be a result of competition. This shows **no strong correlation**, yet it shows us the insight that the issuers have gathered their pricing between a certain range.

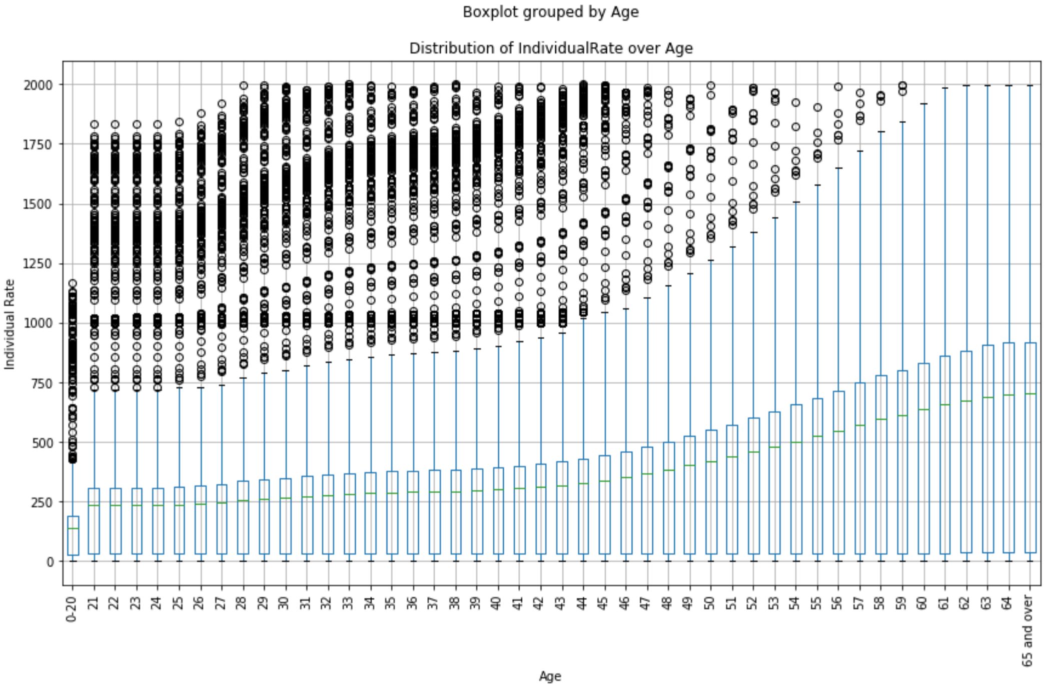
B4. Variation in Costs over Time and with Age.

1. Below is the boxplot of insurance costs over year.



The insurance policies have roughly remained the same throughout the years and the median insurance cost have remained constant.

2. Below plot insurance costs vary with the age of the person being insured.



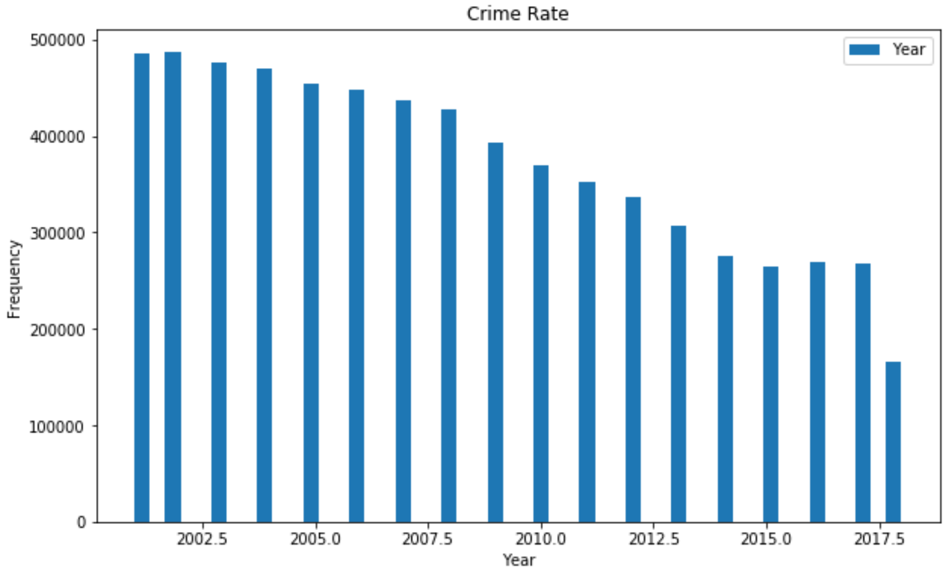
The insurance costs increase over age. In terms of median cost, older people pay more for the insurance than younger people.

Task C
Exploratory Analysis on

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. (PSITAdministration@ChicagoPolice.org, 2001-2018)

The data consists of about **6681498** criminal cases registered over **18 years**. The data covers **35 types** of criminal cases, **180 locations** in Chicago.

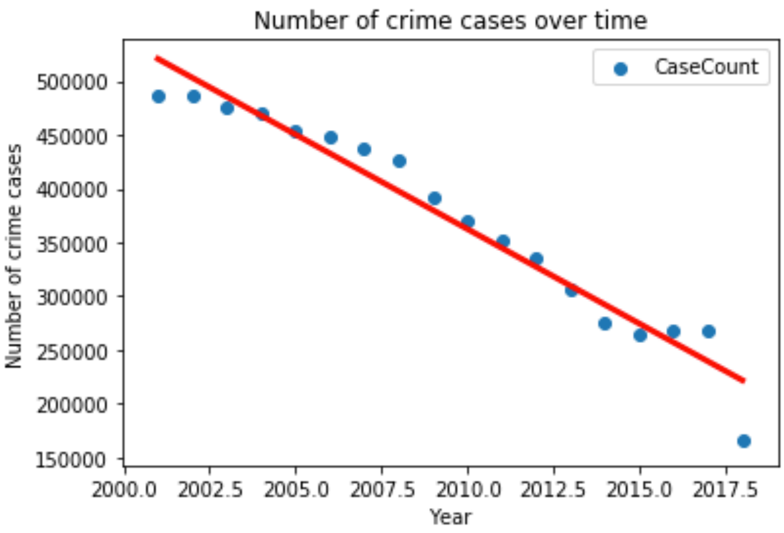
The below histogram plot shows the frequency of crime rate over years.



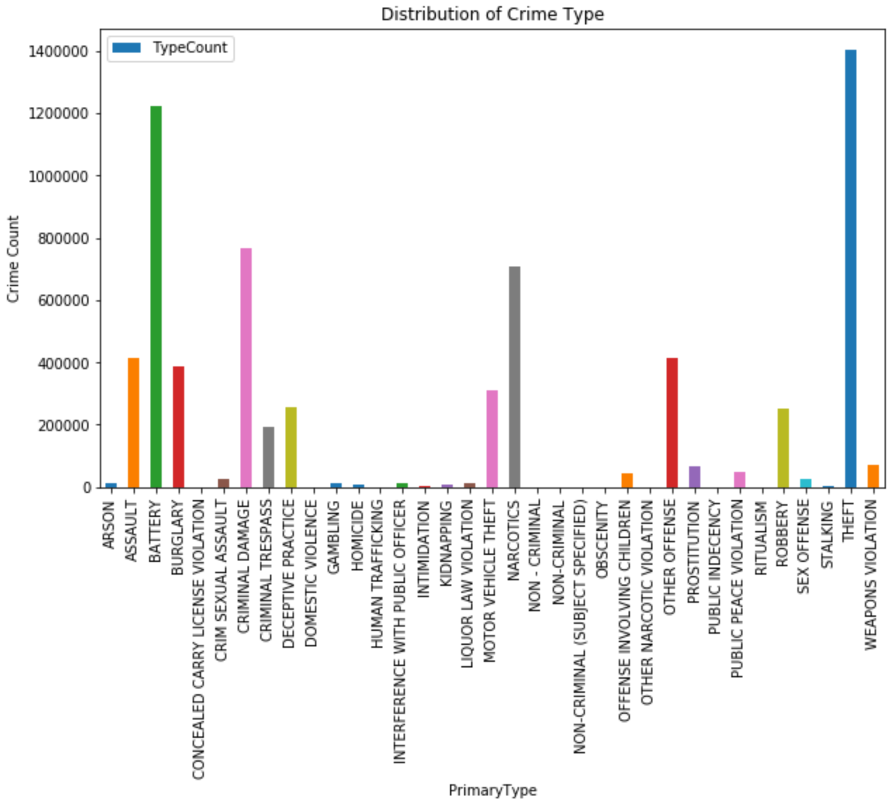
Below Plot is the linear fit for the number of cases lodged over the years to view the trend.

It is predicted that the Chicago will just have about **5604.56** and **2337.98** criminal cases lodges in the year **2020** and **2025** respectively.

The plot shows a negative relationship of number of criminal cases over years.

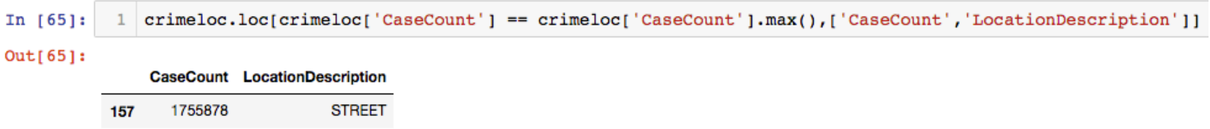


Analysis of the primary types of cases lodges revealed the below plot.



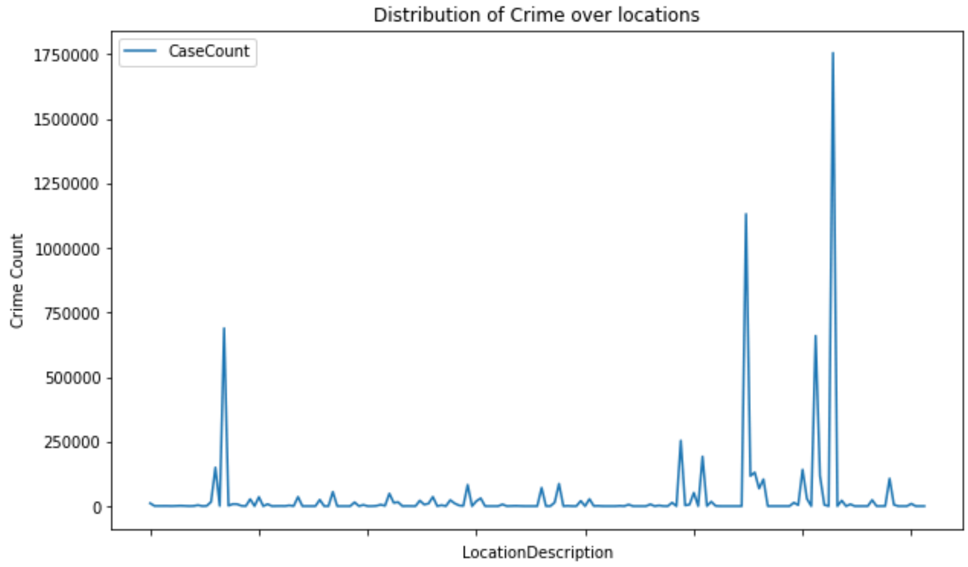
Theft is recorded as the highest type of case recorded over 18 years in Chicago.

Analysis on the location of the cases shows a plot as shown below:
The highest number of cases recorded in one location description is **'Street'** with **1755878** cases lodged over the span of 18 years.

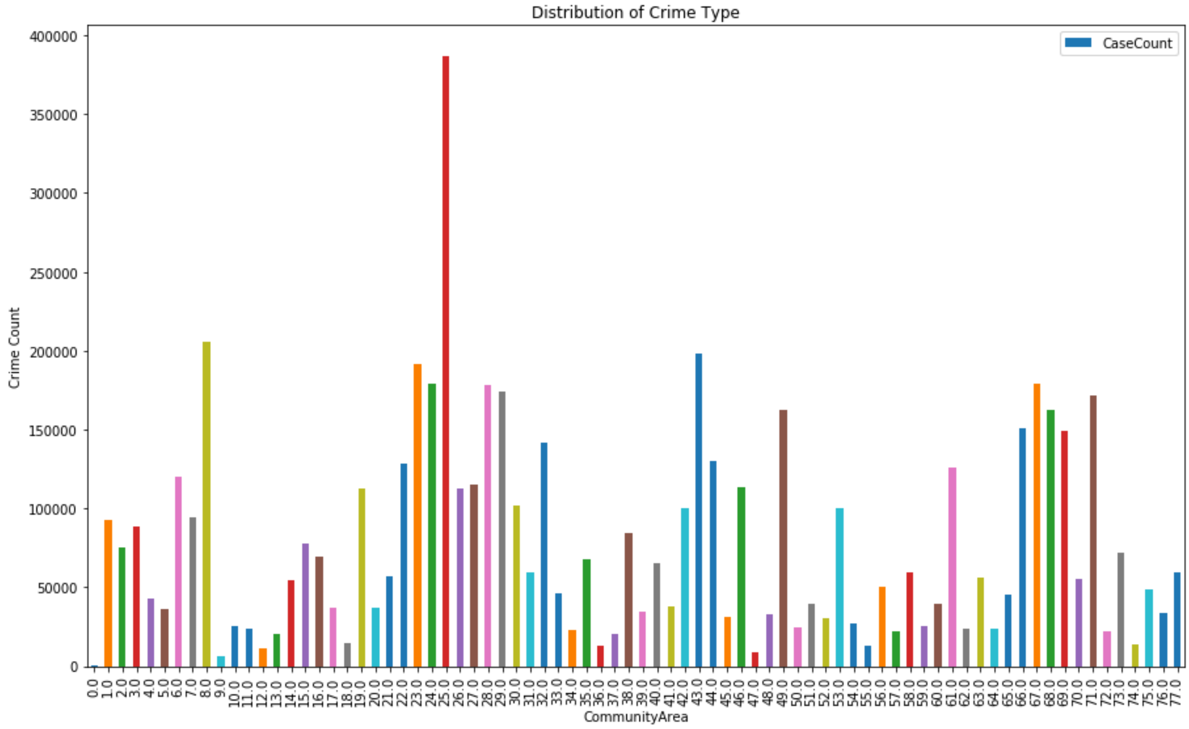


```
In [65]: crimeloc.loc[crimeloc['CaseCount'] == crimeloc['CaseCount'].max(),['CaseCount','LocationDescription']]
Out[65]:
```

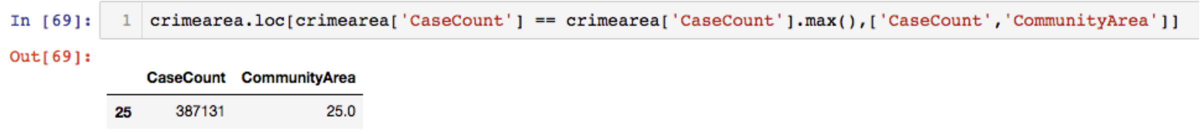
	CaseCount	LocationDescription
157	1755878	STREET



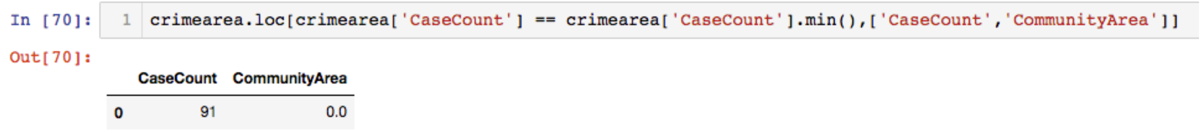
A similar analysis on the crime rate across the areas in Chicago shows the below plot.



This shos that the area with the code **25.0** has the **highest number** of cases with **387131** registered over the years. However, area with the code **0.0** has the **lowest number** of cases with **91** registered over 18 years.

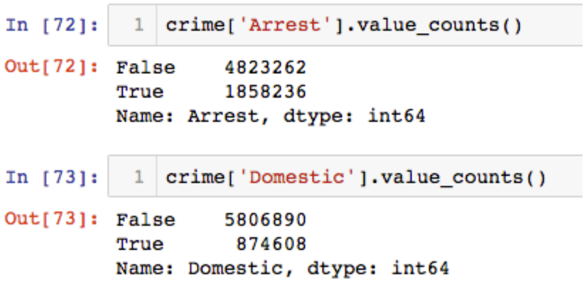


```
In [69]:  crimearea.loc[crimearea['CaseCount'] == crimearea['CaseCount'].max(),['CaseCount','CommunityArea']]
Out[69]:
        CaseCount  CommunityArea
    25     387131           25.0
```



```
In [70]:  crimearea.loc[crimearea['CaseCount'] == crimearea['CaseCount'].min(),['CaseCount','CommunityArea']]
Out[70]:
        CaseCount  CommunityArea
    0          91            0.0
```

Analysis on number of successful arrests and number of domestic crimes have revealed about **1858236** have been **arrested** of **6681498 cases** which were registered, of which **874608** were **domestic** cases.



```
In [72]:  crime['Arrest'].value_counts()

Out[72]:  False    4823262
          True     1858236
          Name: Arrest, dtype: int64

In [73]:  crime['Domestic'].value_counts()

Out[73]:  False    5806890
          True      874608
          Name: Domestic, dtype: int64
```

References

PSITAdministration@ChicagoPolice.org. (2001-2018). *Crimes - 2001 to present.* Chicago: data.cityofchicago.org.