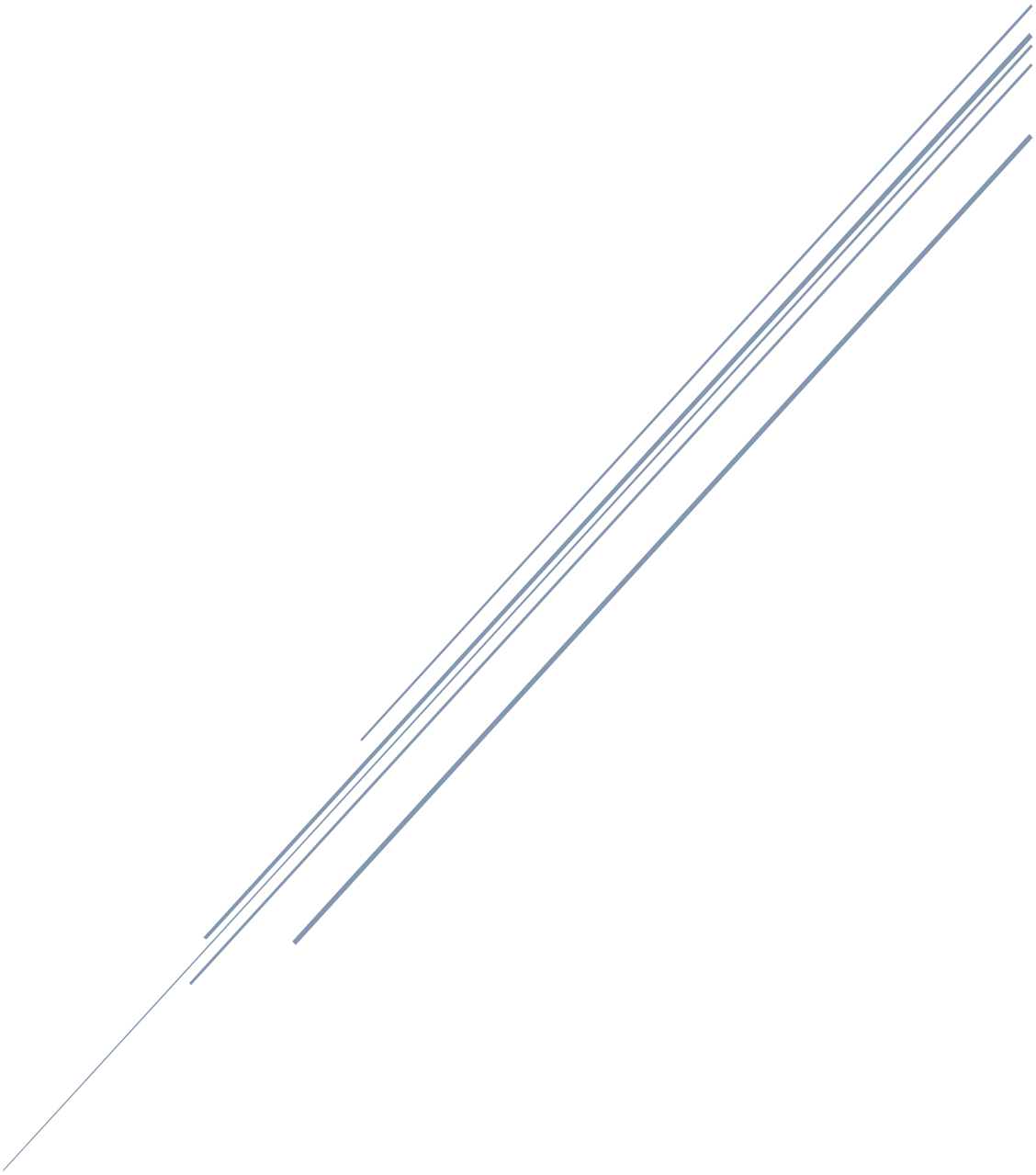
FIT 5145

Assignment 1



Implemented by Mukul Gupta
29873150

TABLE OF CONTENTS

Extra libraries needed

- Seaborn (mainly for boxplots and task C)
- Wordcloud (for task C)

Dataset URL for Task C

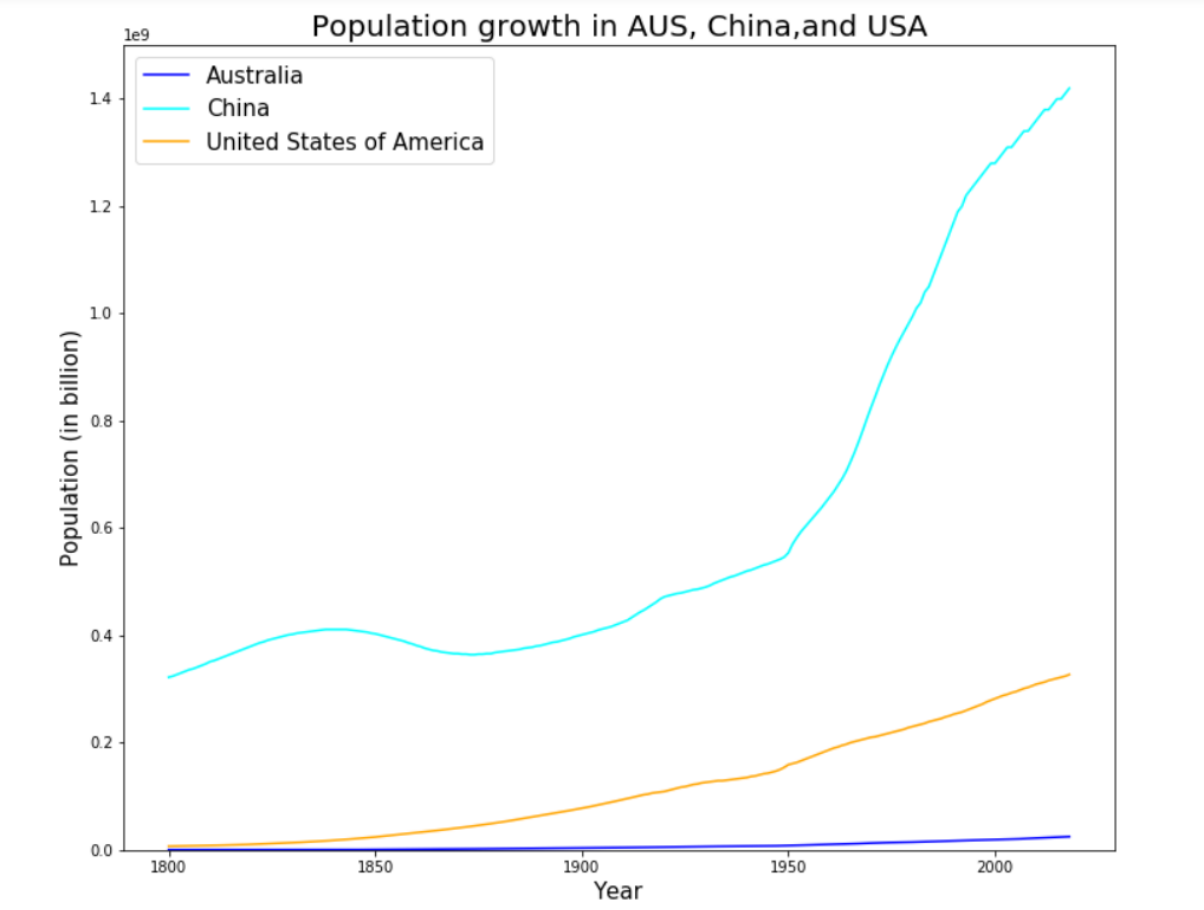
- URL for data: http://data.insideairbnb.com/australia/vic/melbourne/2018-08-08/visualisations/listings.csv

Tasks

Task A: Investigating Population and Gender Equality in Education

A1. Investigating the Population Data

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

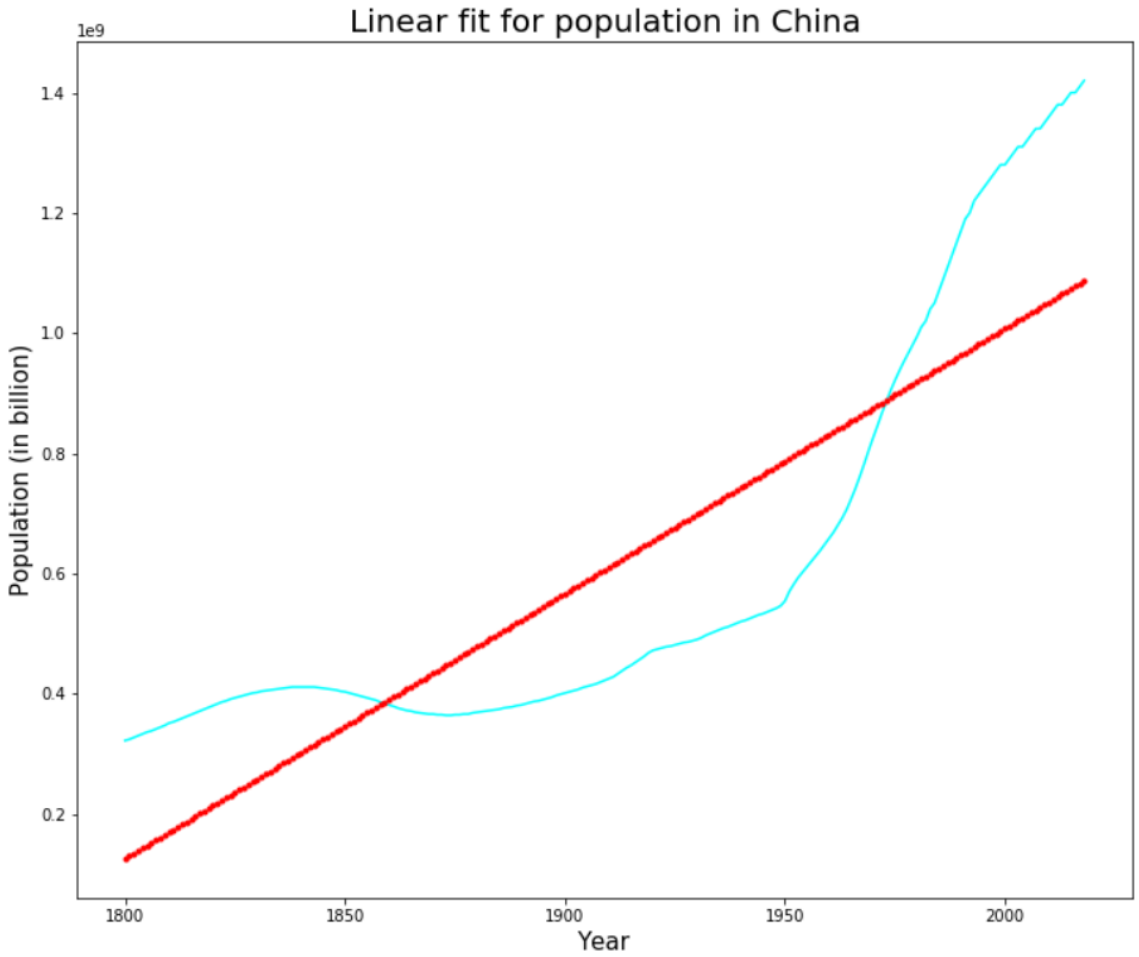


Q: Are the population values increasing or decreasing over time?

Ans: Population is increasing over time.

- For China, population increased till around 1850, after that it decreased till 1875 but since then population is increasing almost exponentially till 1970. After that, it is increasing linearly.
- USA's population is increasing linearly since 1800.
- Australia's population is also increasing since 1800 but still population is quite low as compared to USA and China.

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



Q: Does the linear fit look good?

Ans: No, the linear regression line doesn't fit well. The population of China increases and decreases and then again increases. This linear regression line is not good for predicting the future populations

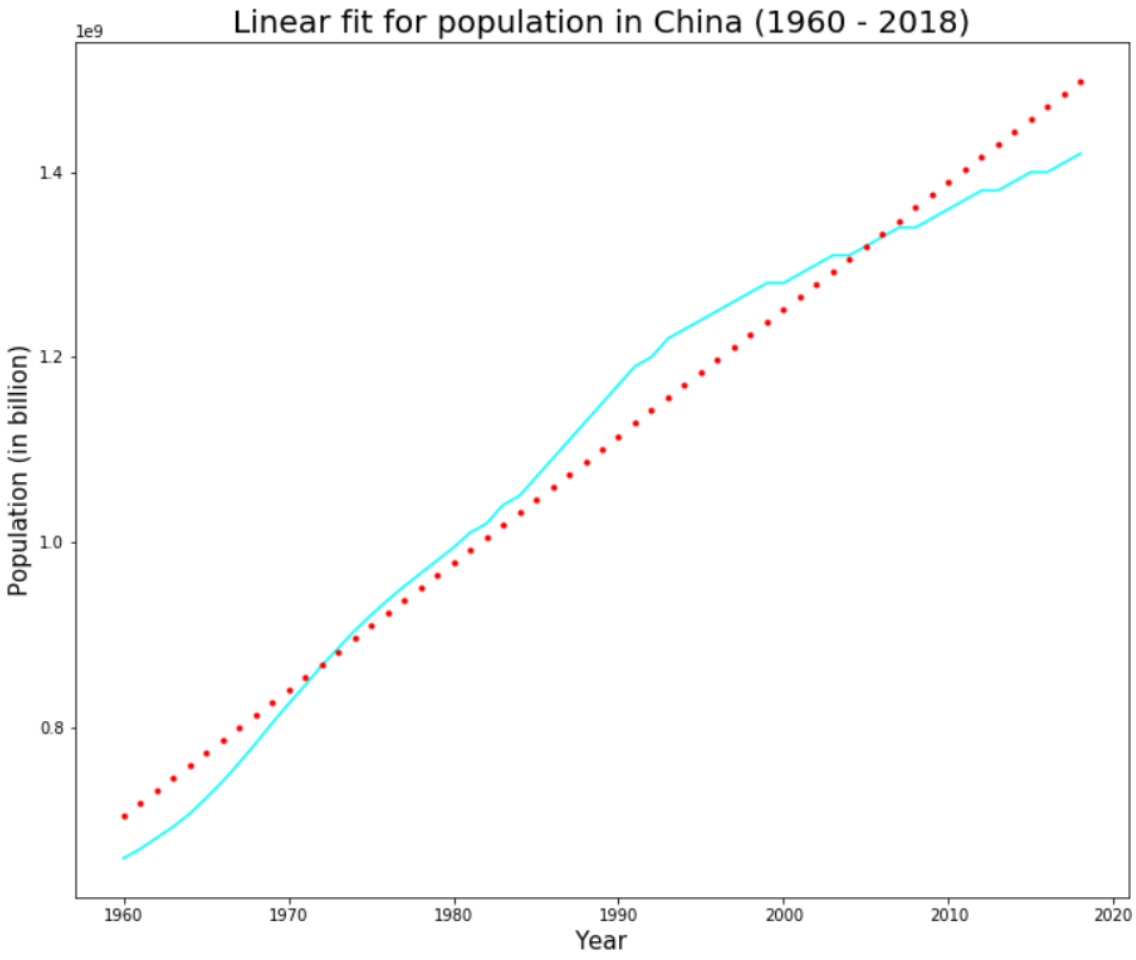
Q: Use the linear fit to predict the resident population in China in 2020 and 2100.

Ans: Predictions of population (in billions) are:

	Year	Population
0	2020	1.095698e+09
1	2100	1.448595e+09

In [48]:

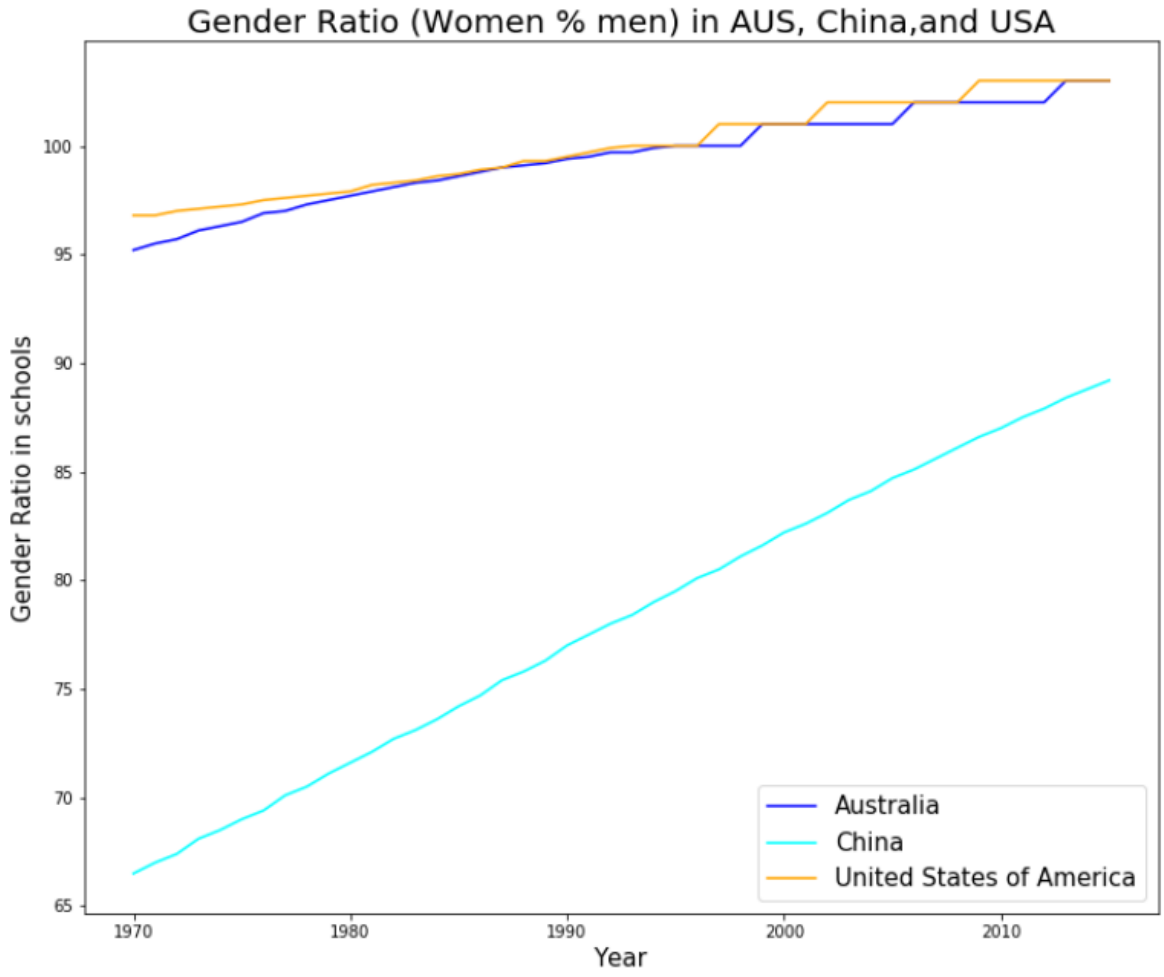
Q: Instead of fitting the linear regression to all of the data, try fitting it to just the most recent data points (say from 1960 onwards). How is the fit? Which model would give better predictions of future population in China do you think?



Ans: This linear regression line (1960 - 2018) fits well for the data in comparison to the previous linear regression line (1800 - 2018). The error of this linear fit is less and is a better model to predict future population of China.

A2. Investigating the Gender Equality Data

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



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

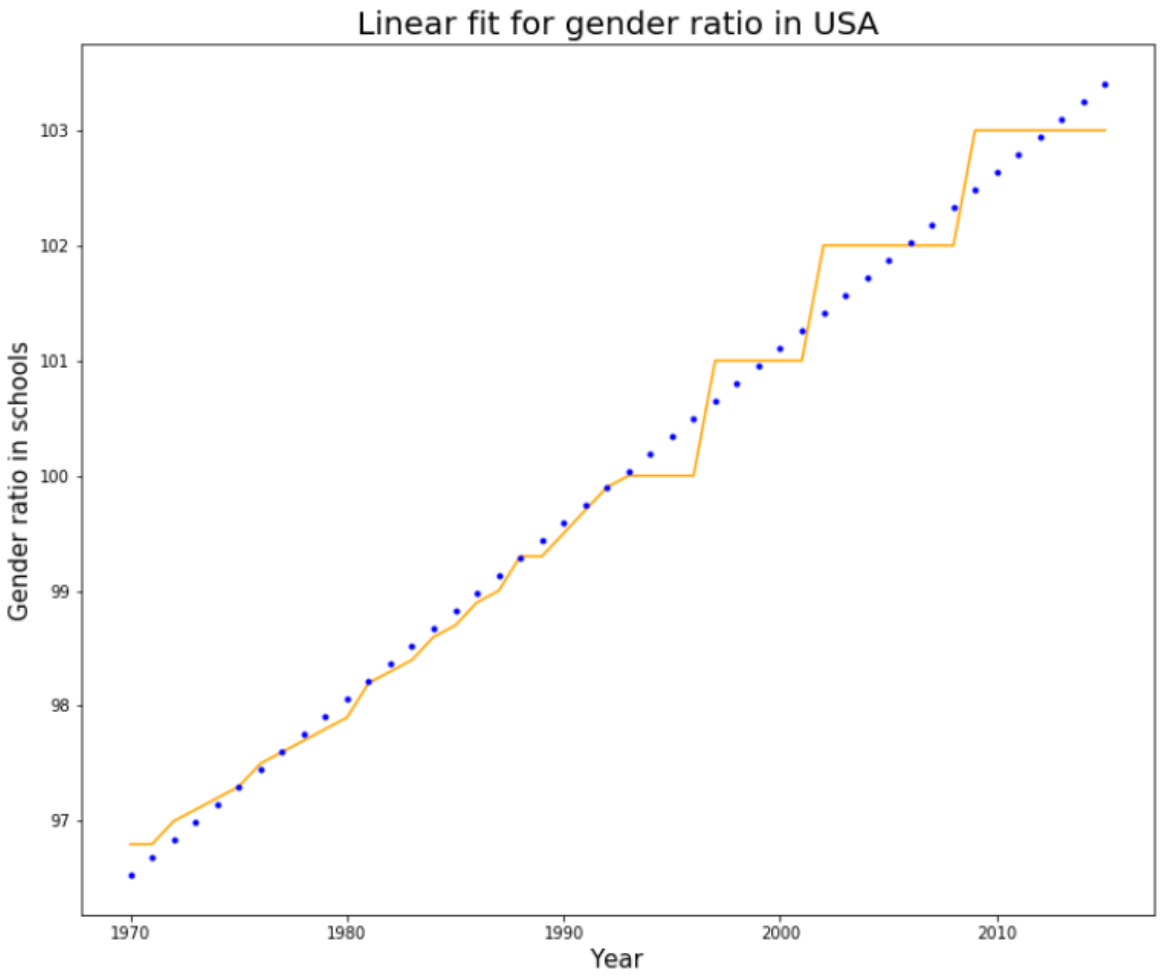
Ans: Maximum gender ratio in Australia: 103.0

Minimum gender ratio in Australia: 95.2

Q: 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?

Ans: In 1970, gender ratio in China was around 66 while in USA and Australia was over 95. Gender ratio has improved in all the three countries. While, the gender ratio in Australia and USA has reached over 100, in China, it is less than 90. Australia and USA have shown similar growth trend over the years.

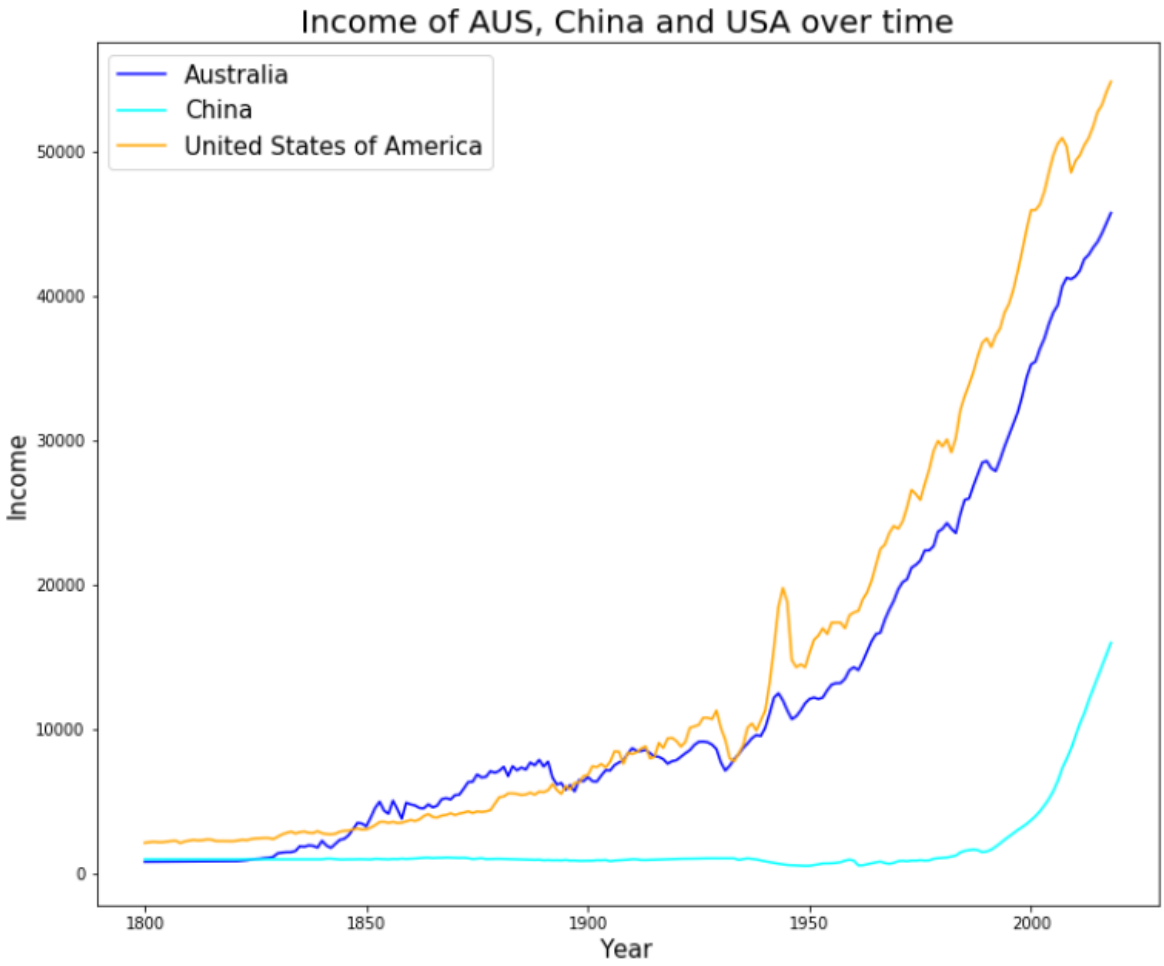
2. Fit a linear regression to the gender ratio in schools in United States and plot it.



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

Ans: The linear fit looks good. There is minimum error. I would not believe that the predictions of the linear model would be correct in the future. This model would predict 200 and 300 as the gender ratio in the future which might not be possible practically.

A3. Investigating the Income Data



Q: 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?

Ans:

Minimum income in China: 530

Income in Australia in the same year when China had their minimum income: 11800

A4. Visualising the Relationship between Gender Equality and Population

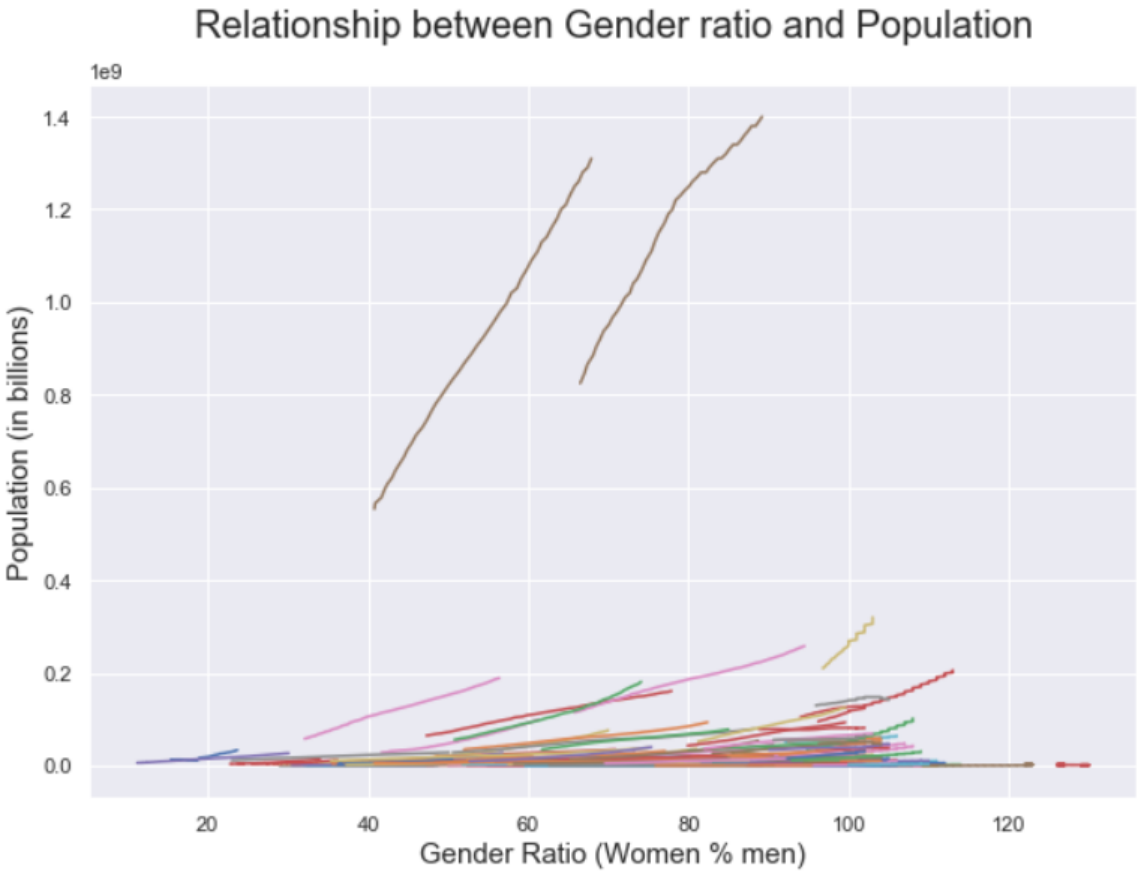
Use Python to combine the data from the different files into a single table. The table should contain population values, income and gender ratio in schools for the different years and different countries

	Year	Country	Population	Gender Ratio	Income
0	1970	Afghanistan	11100000	15.4	1180
1	1971	Afghanistan	11400000	15.8	1100
2	1972	Afghanistan	11700000	15.4	1050
3	1973	Afghanistan	12000000	15.6	1150
4	1974	Afghanistan	12300000	15.9	1180
5	1975	Afghanistan	12600000	16.1	1210
6	1976	Afghanistan	12800000	16.4	1240
7	1977	Afghanistan	13100000	16.6	1130

Q: What is the first year and last year for the combined data?

Ans: First year = 1970, Last year = 2015

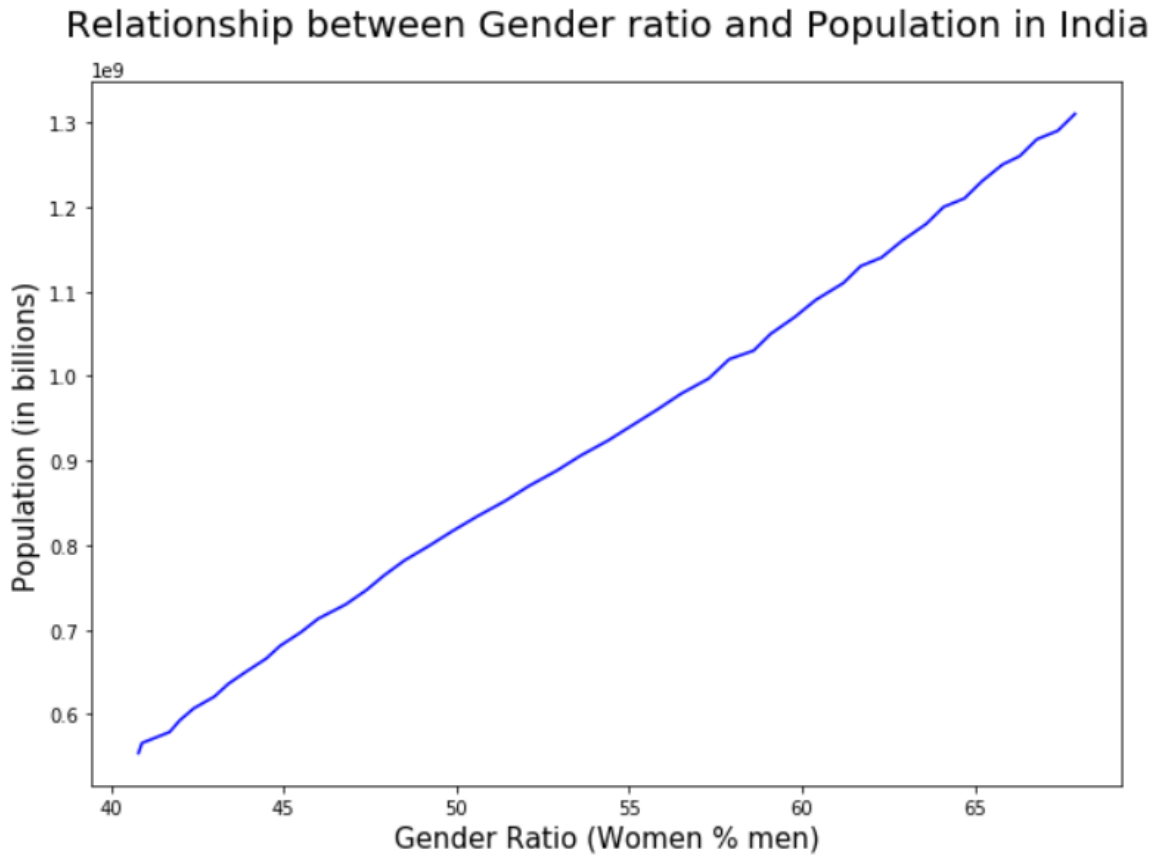
Now that you have the data aggregated, we can see whether there is a relationship between gender ratio in schools and the population. Plot the values against each other.



Q: Can you see a relationship there?

For most countries, gender ratio is improving despite not much change in population. For China and India and some other countries, population is increasing faster than the gender ratio. In most of the other countries, gender ratio is increasing at a faster rate than population. It is not totally clear from the graph as around 187 are plotted in the same graph. It would show a better picture if we plot the graphs separately for each country.

Try selecting and plotting only the data from India.

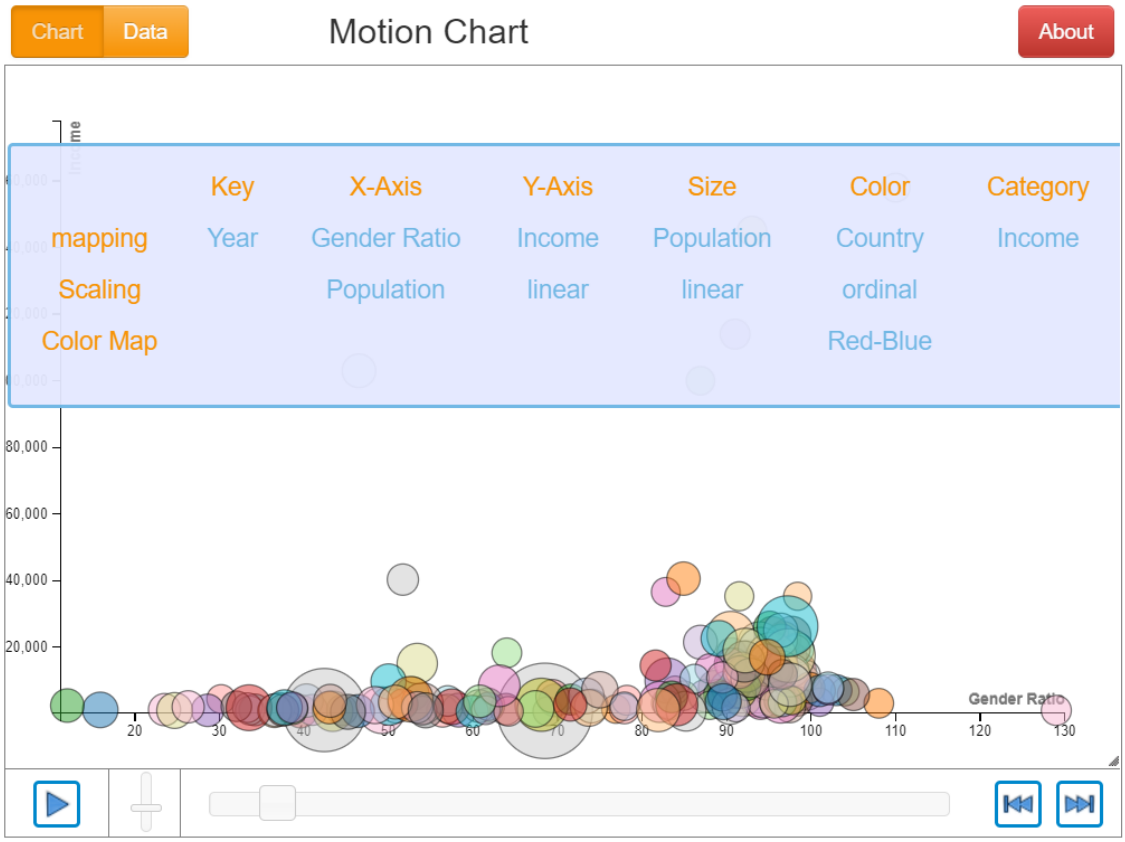


Q: Can you see a relationship now? If so, what relationship is there?

Ans: Yes, there is a strong relationship. As the population increase, gender ratio also increases for India. There is a linear relationship in gender ratio and population.

A5. Visualising the Relationship over Time

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



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

Ans: Afghanistan and Yemen

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

Ans: Lesotho

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

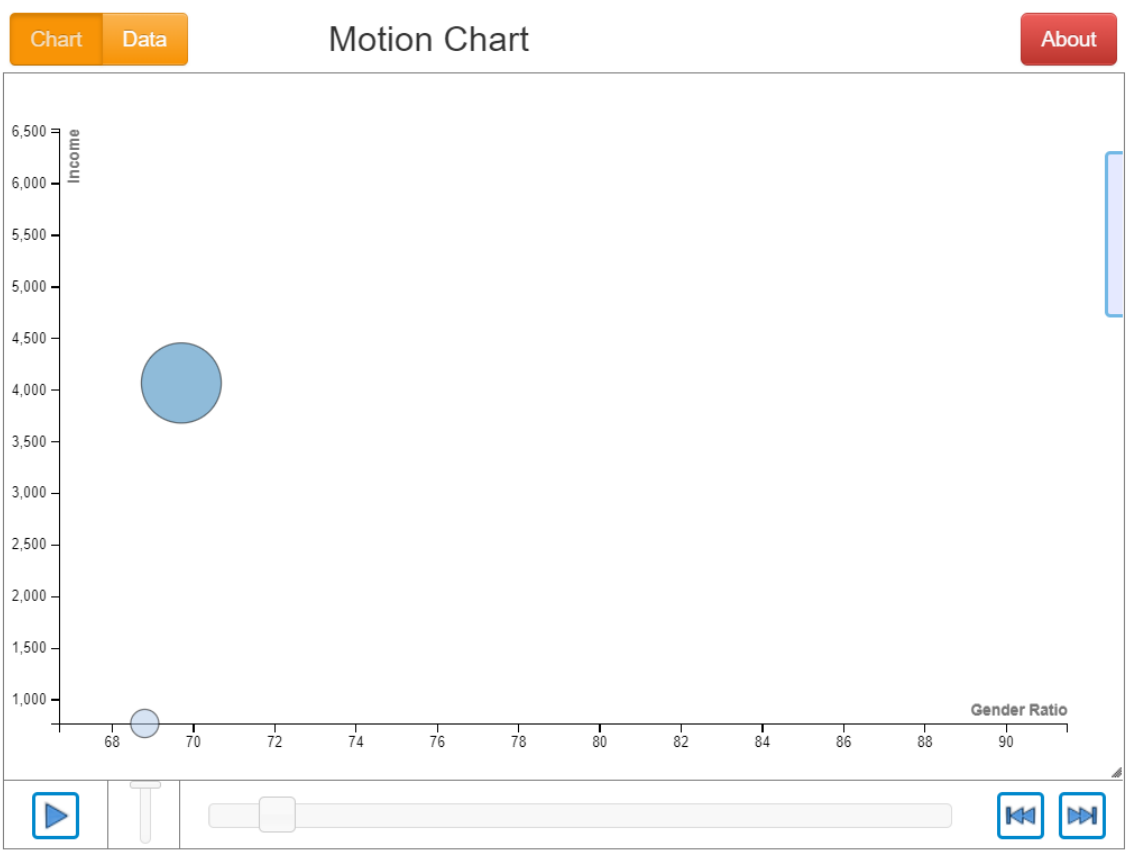
Ans: Gender ratio is generally increasing the while period of time. Income is generally fluctuating. For some countries it is increasing, for some countries like Libya it is decreasing.

Q: 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.

Ans: In the year 2005, ‘Carpe Verde’ started having higher income than ‘Bolivia’. In 2013, ‘Bolivia’ again had more income than ‘Carpe Verde’. ‘Bolivia’ always had higher gender ratio than ‘Carpe Verde’

```
# For countries 'Cape Verde' and 'Bolivia'
data = combined_data[combined_data['Country'].isin(['Cape Verde','Bolivia'])]
mChart = MotionChart(df = data, key='Year', x='Gender Ratio', y='Income', xscale='Population',
                     size='Population', color='Country')

mChart.to_notebook()
```



Q 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.

Ans: Generally, amount of income is increasing as the gender ratio is increasing. There are some countries with the exception though. Income is fluctuating for many countries but most of the countries with higher income have higher gender ratio. But the opposite may not be true. Gender ratio and income have a relationship between them, but gender ratio might not be the only factor that leads to higher income.

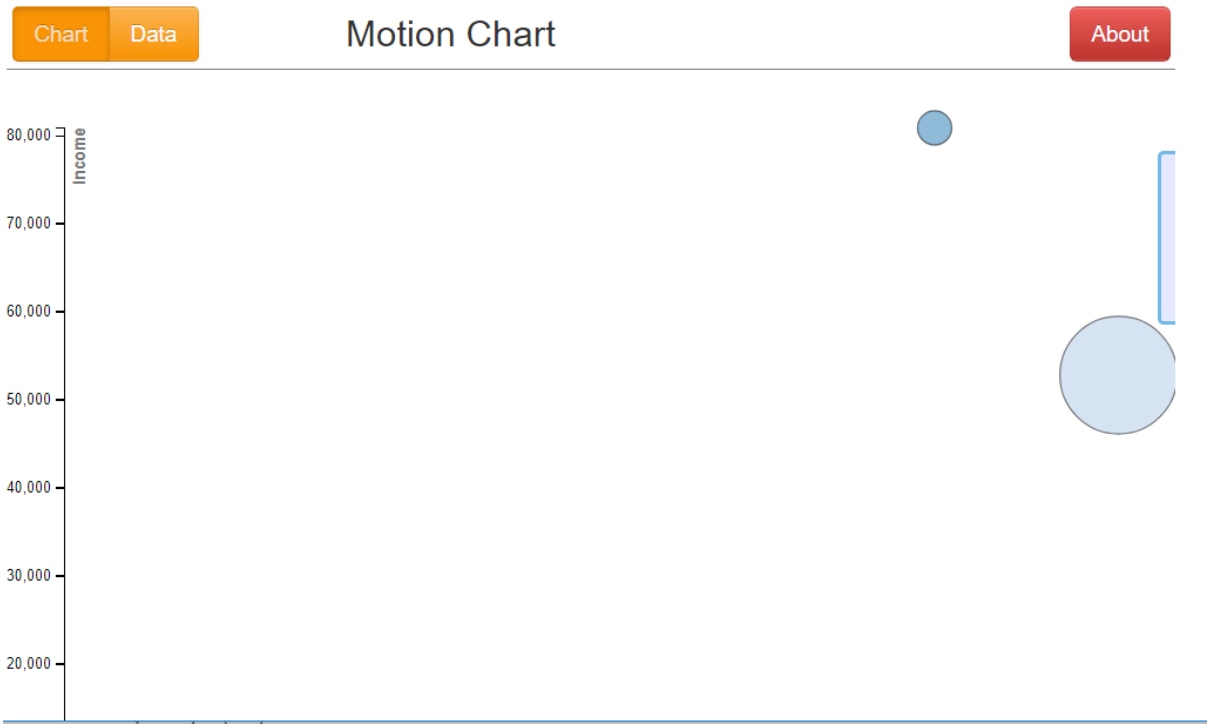
Q Any other interesting things you notice in the data? Please support your answer with relevant python code and/or motion chart

Ans: Most of the countries still have very low income. There are few countries that have very high income in comparison to other countries. This can mean that large number of countries are still developing. Gender ratio is increasing for most countries. There are 2 countries that have the most population. Other countries have comparatively lower population. Generally, countries with higher gender ratio have higher income also.

I compared the countries United States and Singapore. Gender ratio and population have always been higher for United States when compared to Singapore. In terms of income, Singapore went ahead of United States in the year 1993 and since maintained it.

```
# For countries 'Cape Verde' and 'Bolivia'
data = combined_data[combined_data['Country'].isin(['United States','Singapore'])]
mChart = MotionChart(df = data, key='Year', x='Gender Ratio', y='Income', xscale='Population',
                     size='Population', color='Country')

mChart.to_notebook()
```



Task B: Exploratory Analysis on Big Data

B1 Load the InsuranceRates.csv data in Python and answer the following questions:

Q: How many rows and columns are there?

Ans: Rows: 12694445, Columns: 7

```
In [52]: # load the population dataset
         insurance_rates = pd.read_csv('InsuranceRates.csv/InsuranceRates.csv')
```

```
In [8]: insurance_rates.shape
Out[8]: (12694445, 7)
```

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

Ans: 3 (2014, 2015, 2016)

Q: What are the possible values for 'Age'?

Ans: Age values can be between:
['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']

Q: How many states are there?

Ans: 39

Q: How many insurance providers are there?

Ans: 910

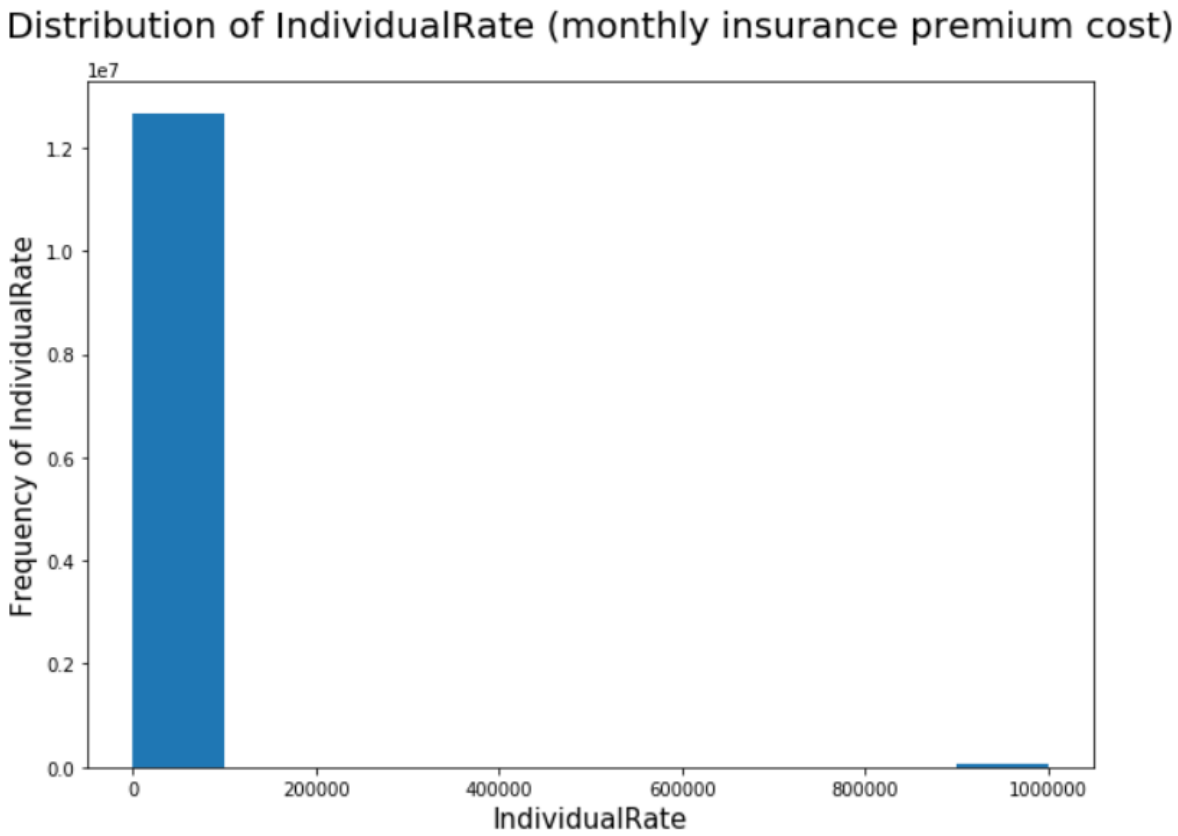
Q: What are the average, maximum and minimum values for the monthly insurance premium cost for an individual? Do those values seem reasonable to you?

Ans: Minimum: 0.0, Maximum: 999999.0, Average: 4098.026458581588

These values seem unreasonable. Maximum monthly insurance premium is too much for an individual. Average monthly insurance is also way too much.

B2. Investigating Individual Insurance Costs

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



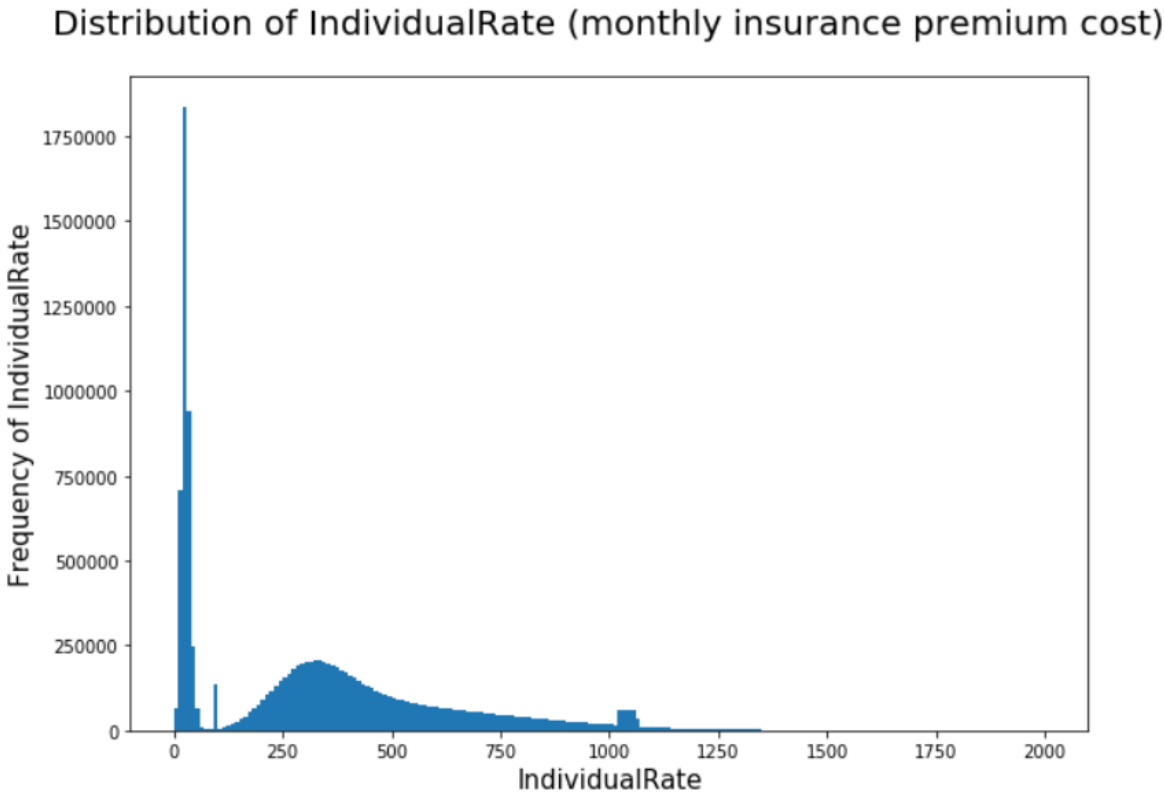
Q: Does the distribution make sense to? What might be going on?

Ans: The histogram doesn't make much sense as most of the data lies in the rate 0 – 100,000. It is because of the outlier range 900,000-1,000,000, we aren't able to gather much information about the data.

2. Remove rows with insurance premiums of 0 (or less) and over 2000. (Use this data from now on.)

```
insurance_rates = insurance_rates[(insurance_rates['IndividualRate'] <= 2000) &
                                  (insurance_rates['IndividualRate'] > 0)]
```

Generate a new histogram with a larger number of bins (say 200).



Q: Does this data look more sensible?

Ans: Yes, the data is looking more sensible now. The maximum IndividualRate is now 2,000.

Q: Describe the data. How many groups can you see?

Ans: The data is described below:

```
insurance_rates['IndividualRate'].describe()
```

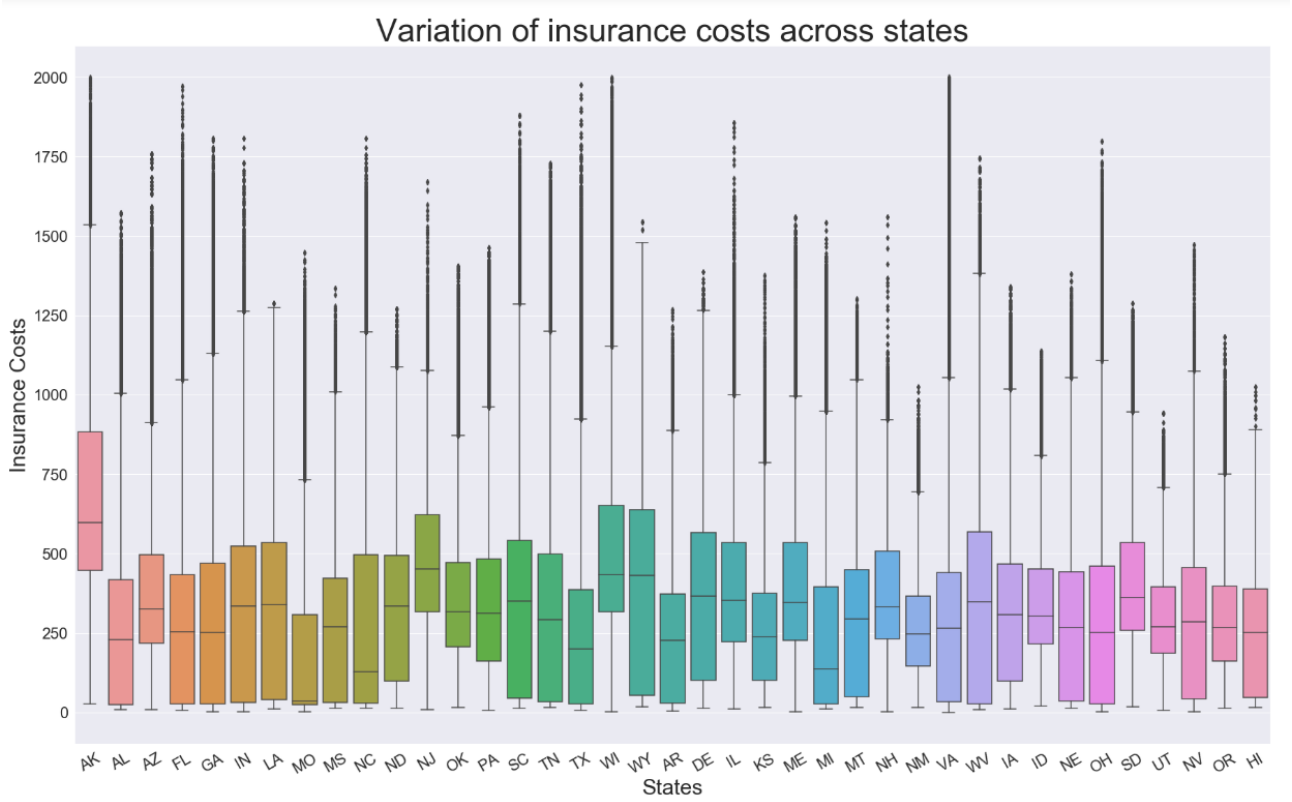
```
count    1.193588e+07
mean     3.329765e+02
std      2.941710e+02
min      1.000000e-02
25%      3.292000e+01
50%      3.069700e+02
75%      4.897800e+02
max      2.000000e+03
Name: IndividualRate, dtype: float64
```

I can see 4 groups in the histogram.

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.



Q: 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.)

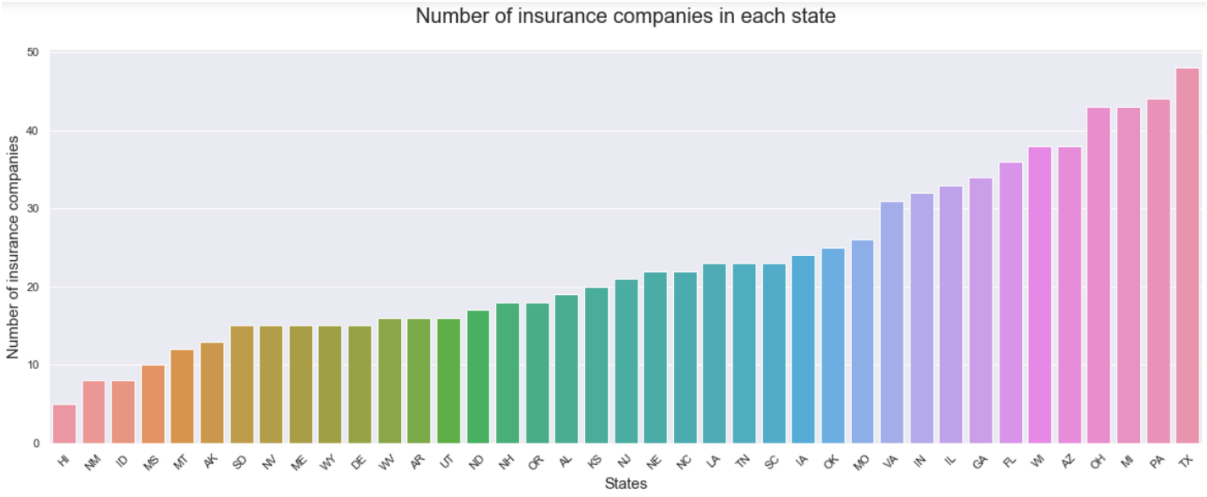
Ans: Highest median: ‘AK’

Lowest median: ‘MO’

Q: Does the number of insurance issuers vary greatly across states?

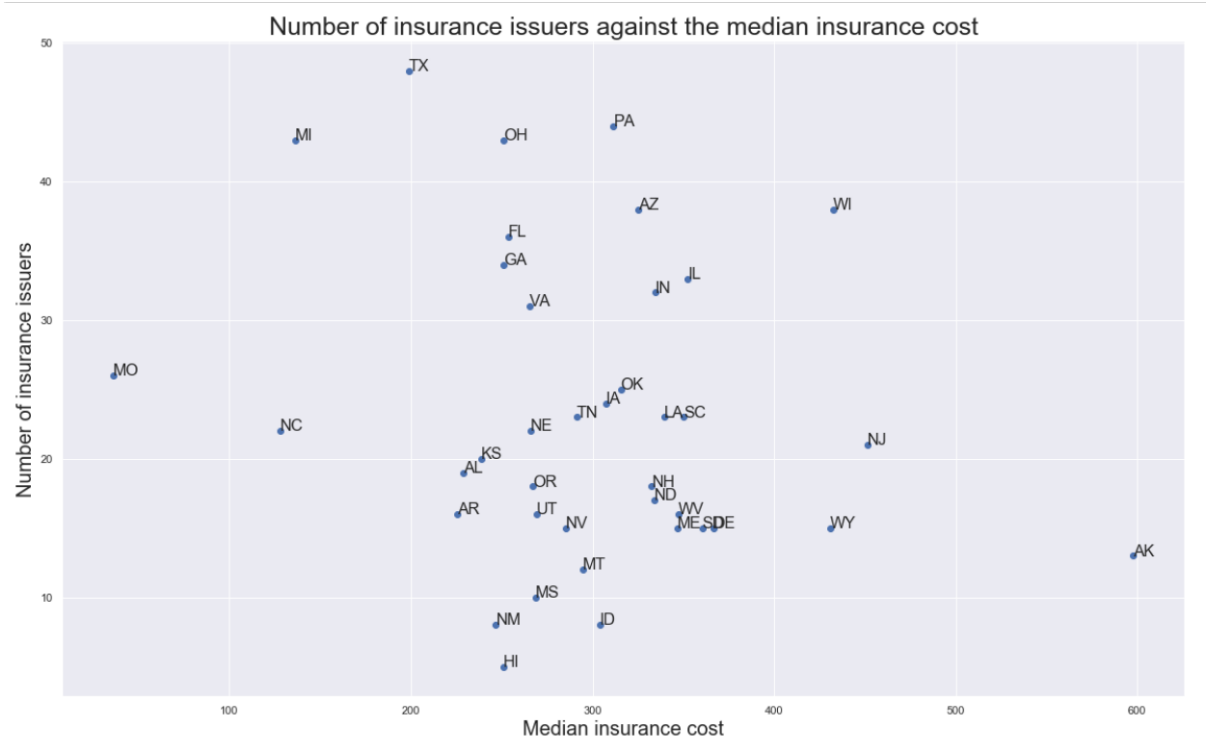
Ans: Yes, number of insurance issuers vary greatly across states as visible from the graph below

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.)



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.

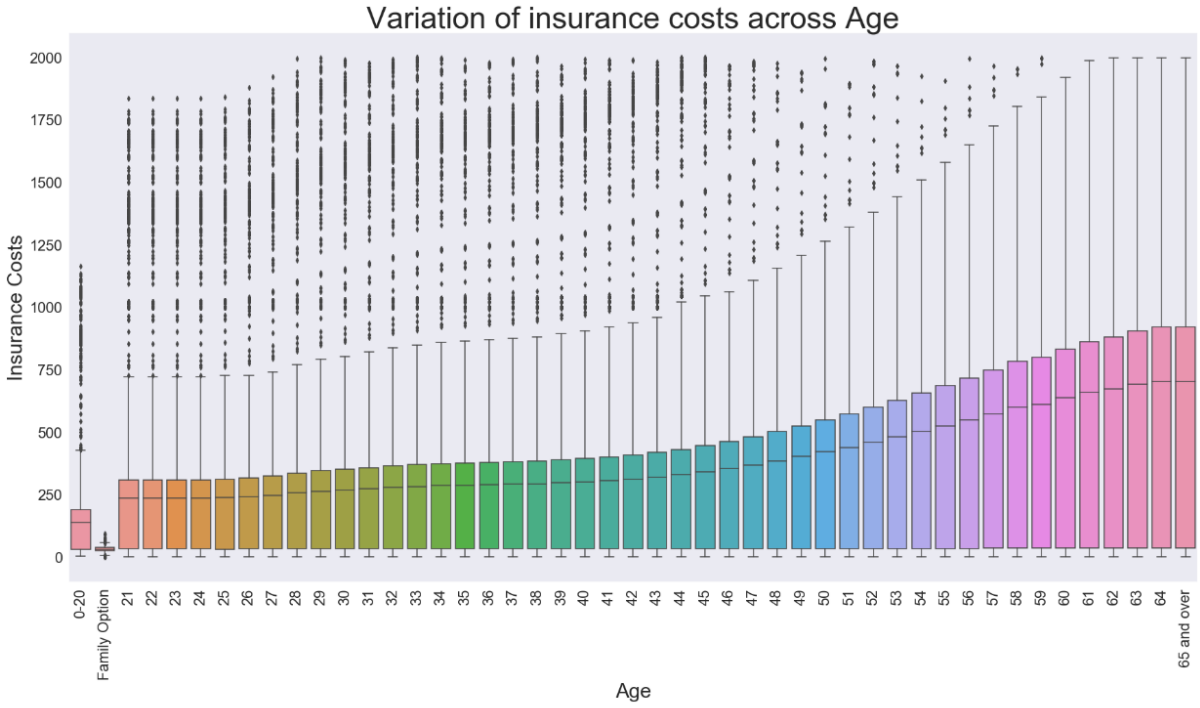


Q: Do you observe a relationship?

Ans: Most of the insurance companies' median insurance cost is around 300. Though, there is no significant relationship between number of insurance issuers and median insurance cost.

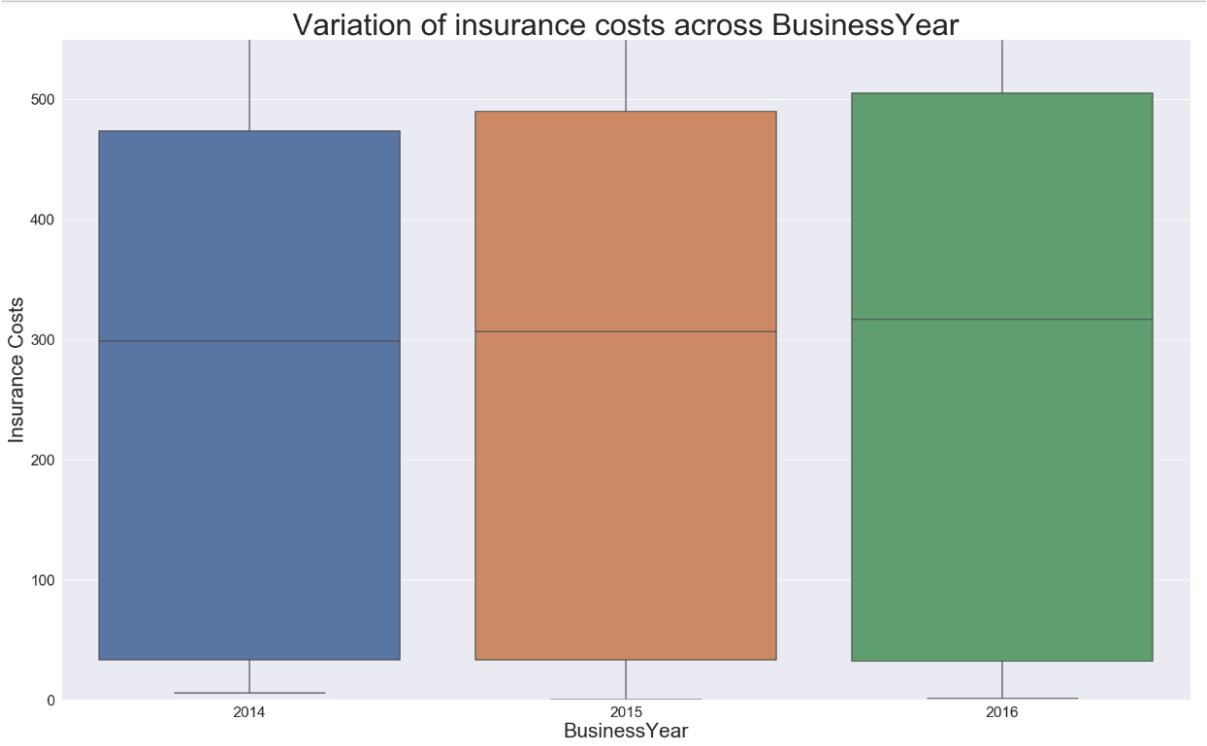
B4. Variation in Costs over Time and with Age

Generate boxplots (or other plots) of insurance costs versus year and age



Q: Are insurance policies becoming cheaper or more expensive over time?

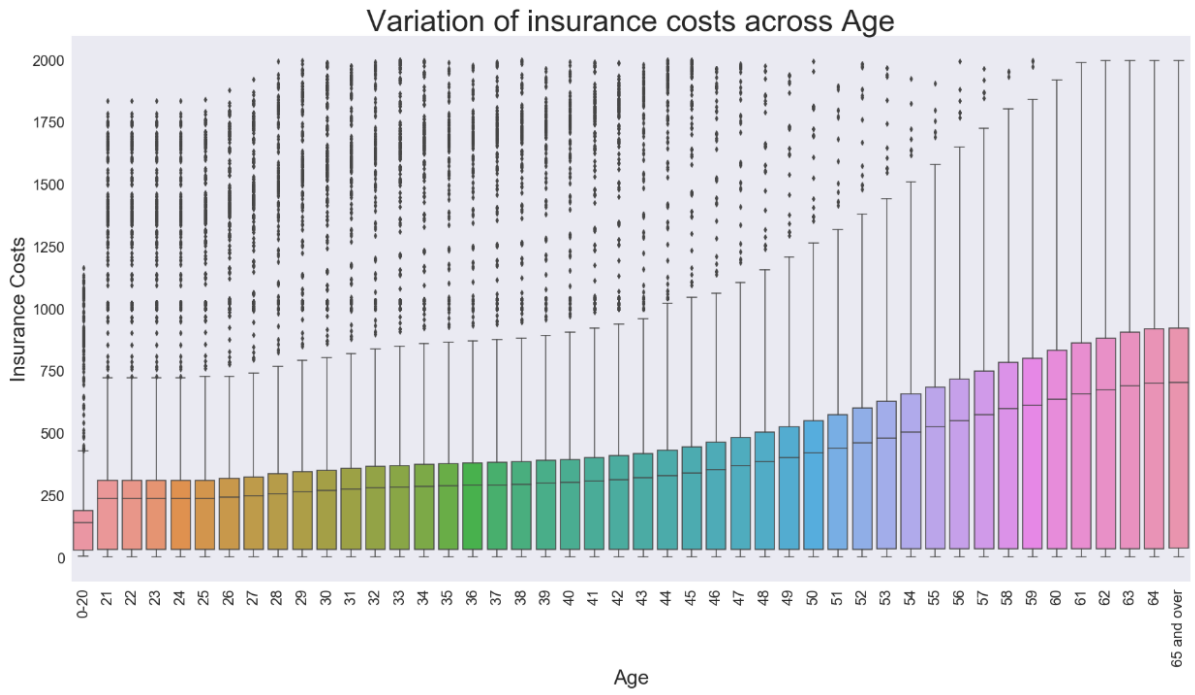
Ans: Insurance policies are becoming expensive over time.



Q: Is the median insurance cost increasing or decreasing?

Ans: We can see from the graph that median of the Insurance increases slightly along with year.

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.)



Q: In terms of median cost, do older people pay more or less for insurance than younger people?

Ans: Older people pay more for insurance than younger people in terms of median cost.

Q: How much more/less to they pay?

Ans:

Median Individual rate for older people: 702.215
Median Individual rate for 0-20 range people: 138.59
On average, older people pay 563.625 more than 0-20 aged people.

Task C: Exploratory Analysis on Other Data

Data describes the detailed Airbnb listings in Melbourne, VIC.

URL for data: http://data.insideairbnb.com/australia/vic/melbourne/2018-08-08/visualisations/listings.csv

Number of rows: 21450

Number of columns: 15 (after dropping 'neighbourhood_group' column which contained only null values)

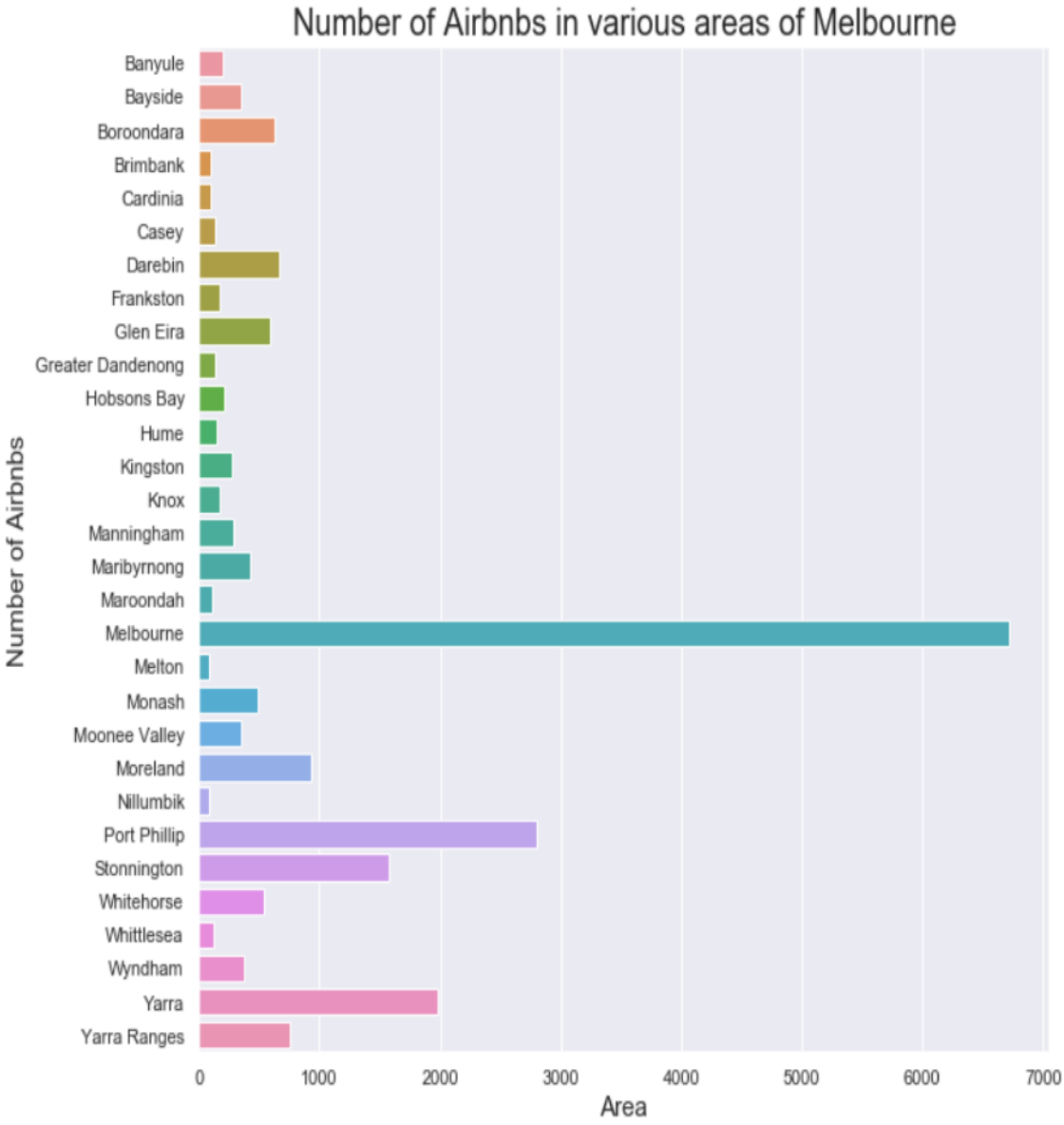
The data looks like:

```
airbnb.head()
```

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review
0	9835	Beautiful Room & House	33057	Manju	Manningham	-37.772684	145.092133	Private room	61	1	4	2015-09-12
1	10803	Room in Cool Deco Apartment in Brunswick	38901	Lindsay	Moreland	-37.766505	144.980736	Private room	35	3	98	2018-06-30
2	12936	Cool Chic Beachside 1 BR Views APT+Garage+WIFI	50121	Frank And Vince	Port Phillip	-37.859755	144.977369	Entire home/apt	159	3	13	2018-07-10
3	15246	Large private room-close to city	59786	Eleni	Darebin	-37.758971	144.989228	Private room	50	2	29	2017-05-15
4	16760	Melbourne BnB near City & Sports	65090	Colin	Port Phillip	-37.864530	144.992238	Private room	69	1	59	2018-03-18

reviews_per_month	calculated_host_listings_count	availability_365
0.05	1	365
1.44	1	211
0.13	17	288
0.31	3	0
0.73	1	322

Find the number of Airbnbs according to various areas in Melbourne

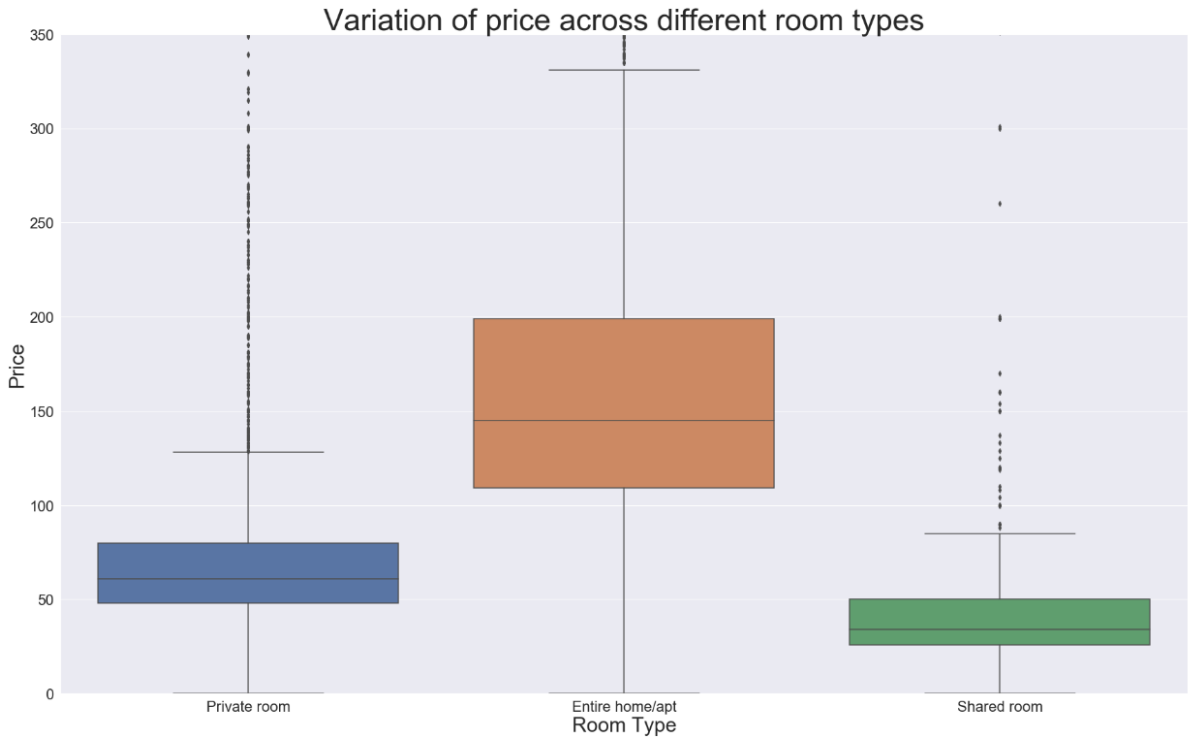


From the graph following inferences can be made:

- Areas which have the maximum Airbnbs are Melbourne (CBD), Port Philip and Yarra
- Areas which have the minimum Airbnbs are Nillumbik, Melton, and Maroondah

Plot prices of different airbnbs according to the room type

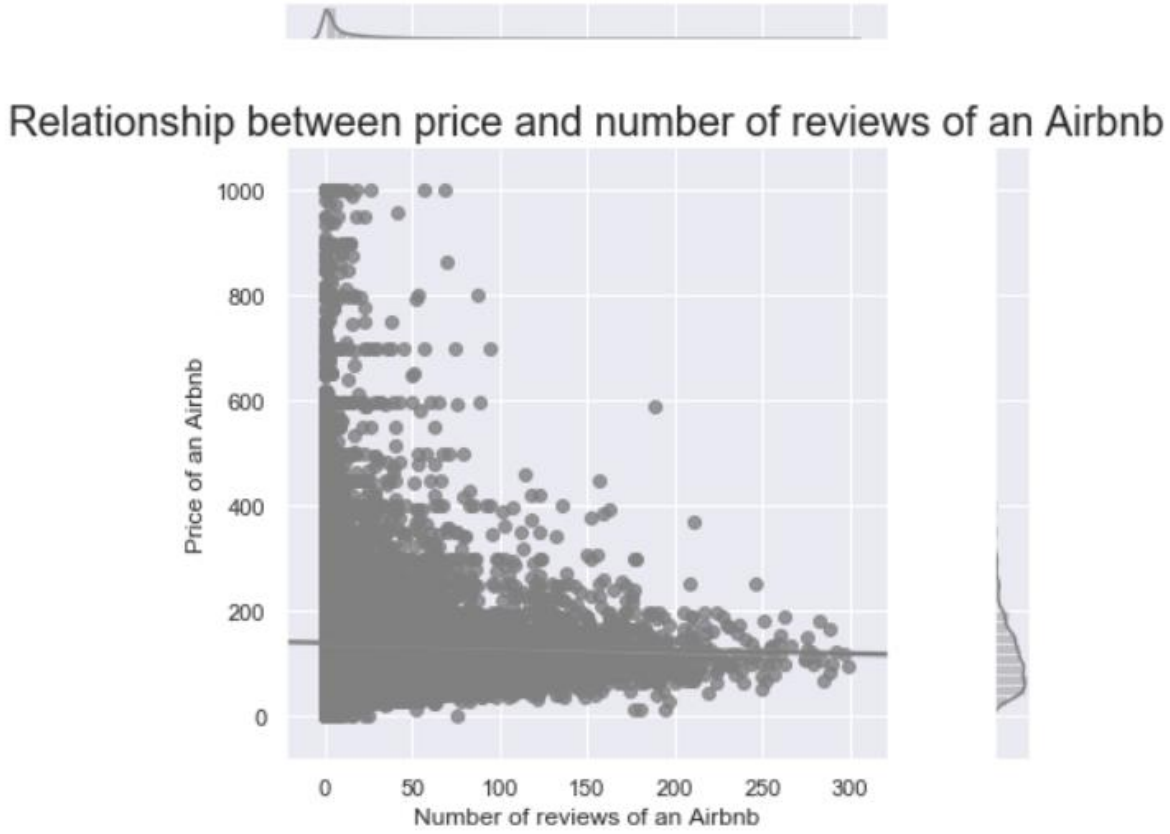
Entire room/apt cost more than private room and shared room



Which words were most commonly used in Airbnb house names

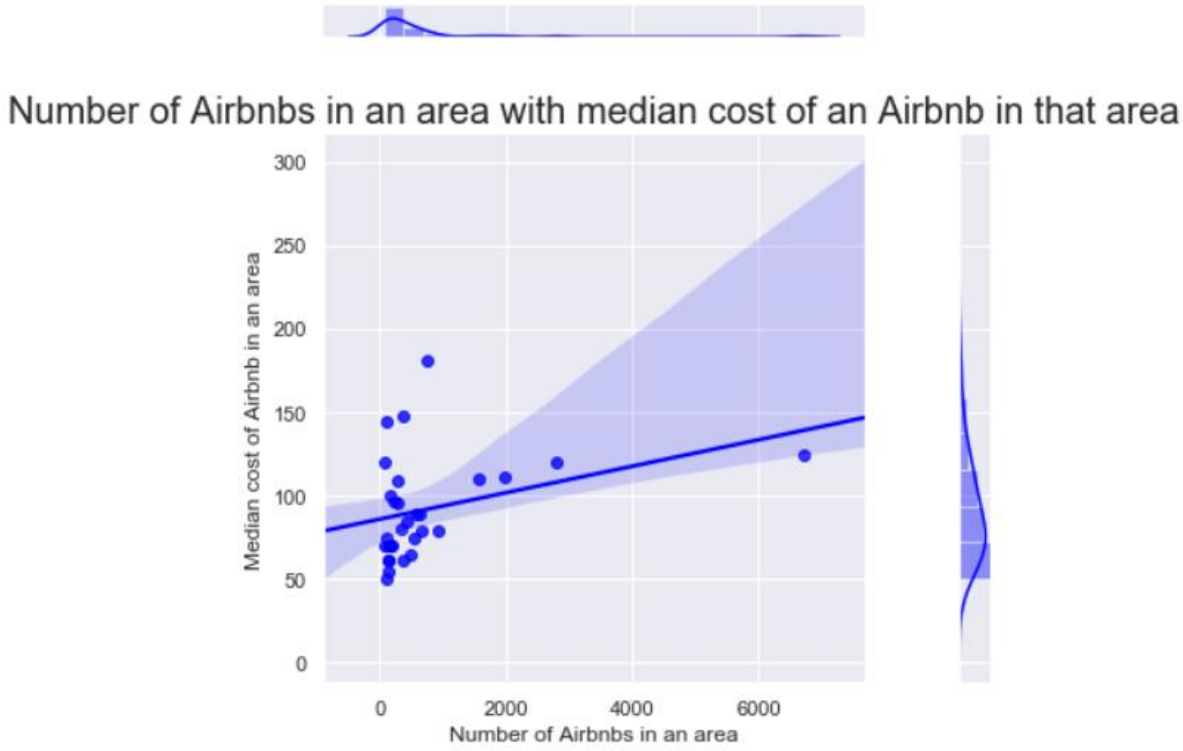
WordCloud to determine which keywords are most commonly used by people for their Airbnbs. Most common words are: Apartment, Private Room, St Kilda, House, Modern, Melbourne CBD, home.

Is there a relationship between price and number of reviews of an Airbnb?



Ans: No relationship. But, from the graph we can see that generally there are more reviews of airbnbs whose price ranges from 50 to 200. This means that people are generally looking for more affordable airbnbs.

Does the number of reviews in an area affect the cost of an Airbnb in an area?



Ans: There might be slight possibility that the number of reviews linearly affect the cost of an Airbnb.

How can we predict price of an Airbnb based on the number of Airbnbs in that area?

The linear fit is not very good as the error is large, but we can still make predictions on the price of the Airbnb using this model for areas which have higher number of Airbnbs.

Predicting price of an Airbnb based on the number of Airbnbs in that area.

	Number of Airbnbs in the area	Price of an Airbnb
0	5000	125.546162
1	10000	165.234410

References

- Pandas documentation, https://pandas.pydata.org/
- Duong Vu, 8 August 2018, https://www.datacamp.com/community/tutorials/wordcloud-python
- Michael Waskom, https://seaborn.pydata.org/
- John Hunter, Darren Dale, Eric Firing, Michael Droettboom and the Matplotlib development team, https://matplotlib.org/contents.html