Farhad Ullah Rezwan
ID: 30270111

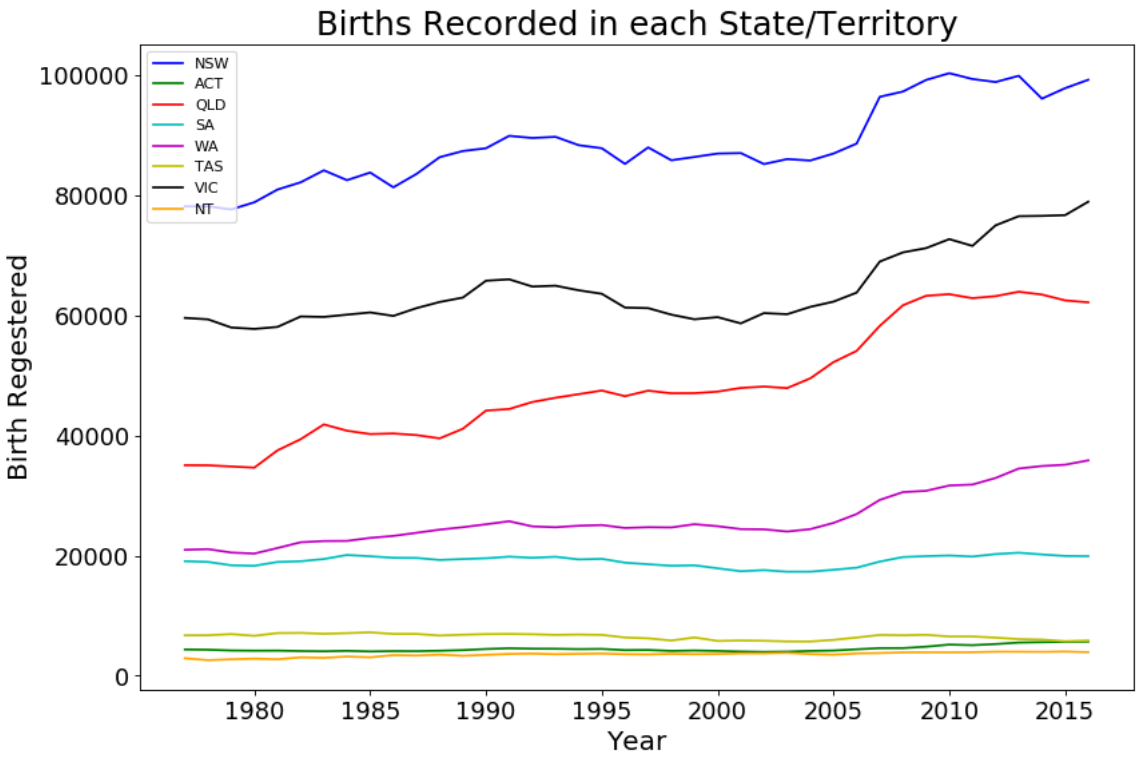
FIT5145 Assignment 1

Farhad Ullah Rezwan | September 9, 2019
Monash ID: 30270111

Task A: Investigating Natural Increase in Australia's population

A1. Investigating the Births, Deaths and TFR Data

1. Plotting the number of births recorded in each state/territory for different Australian states over different years:

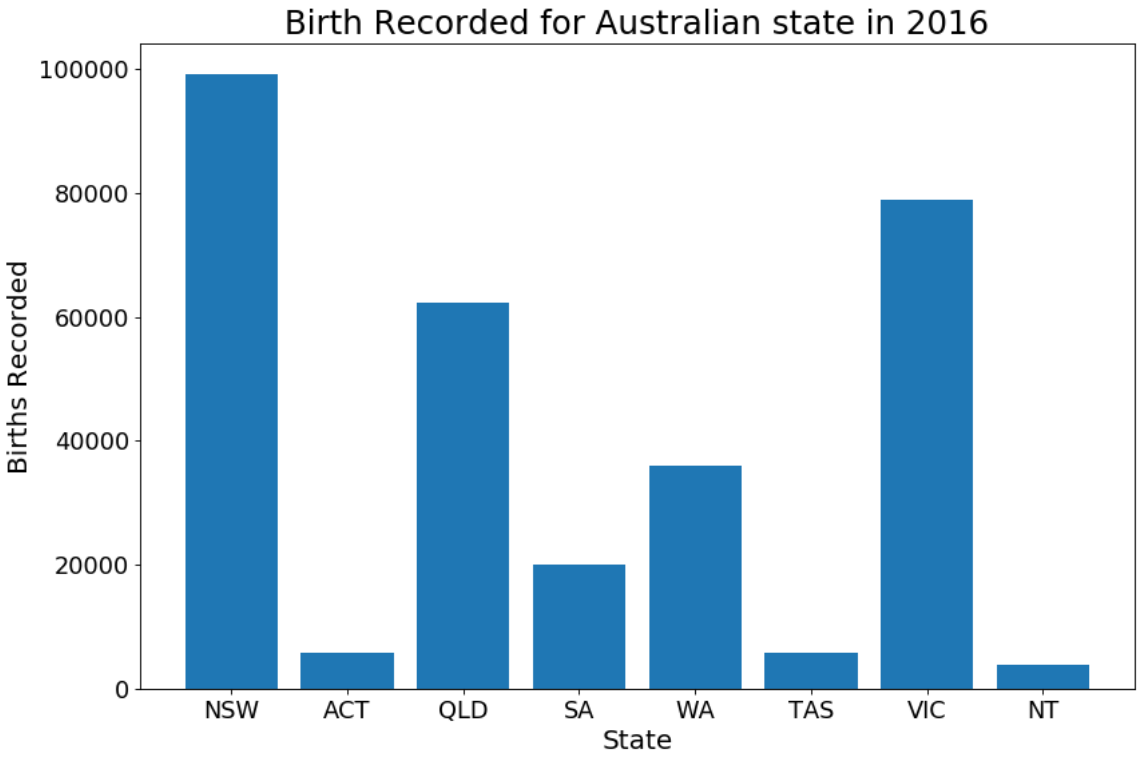


a. Trend in number of births for Queensland and Tasmania (1977 to 2016)?

The bar graph depicts the births recorded in number for each state/territory for different Australian states over years. The graph shows that Queensland(QLD) has faced increased in its total birth recorded however Tasmania's birth recorded was stable during the period of 1977 and 2016.

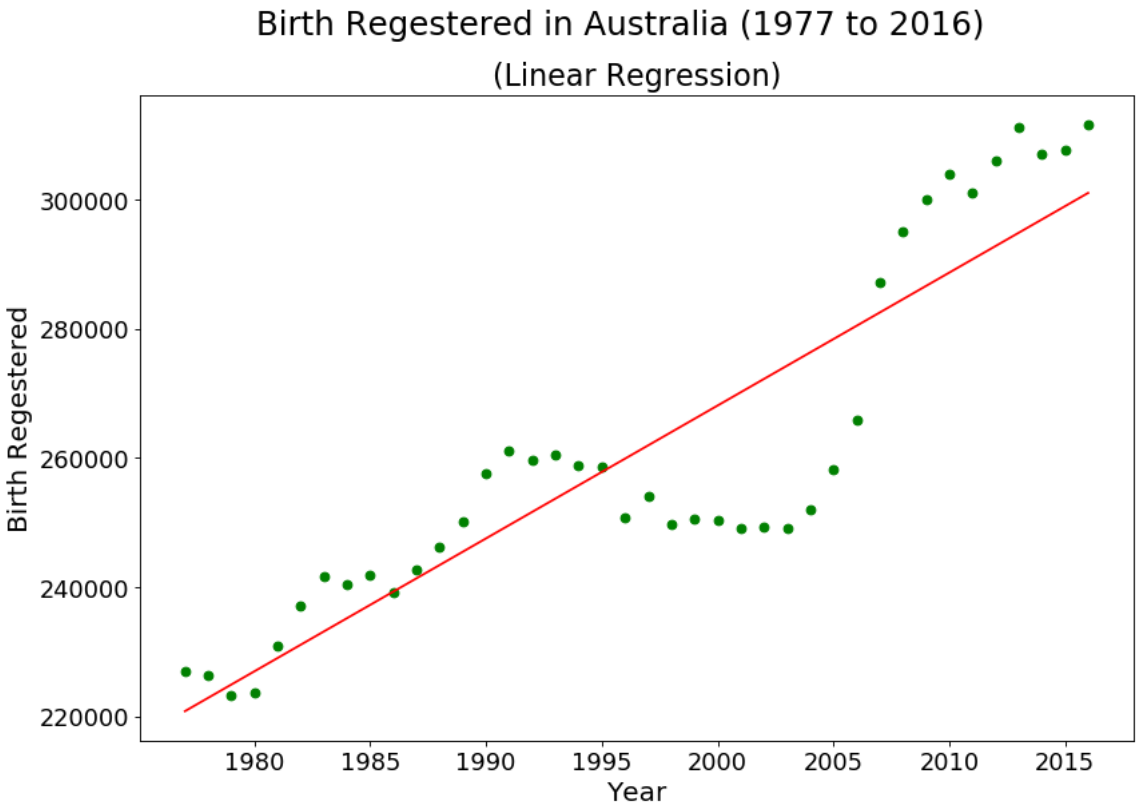
Farhad Ullah Rezwan
ID: 30270111

b. Draw a bar chart to show the number of births in each Australian state in 2016.



2. Investigating the trend in the total number of births over different years.
    1. Fit a linear regression using Python to the above aggregated data and plot the linear fit.

Farhad Ullah Rezwan
ID: 30270111



2. Does it look like a good fit to you? Identify the period time having any unusual trend(s) in your plot.

   This plot using linear regression does not look good fit. The year 2004 prediction for birth registered using linear line is not consistent.

3. Use the linear fit to predict the total births in Australia for the years 2050 and 2100.

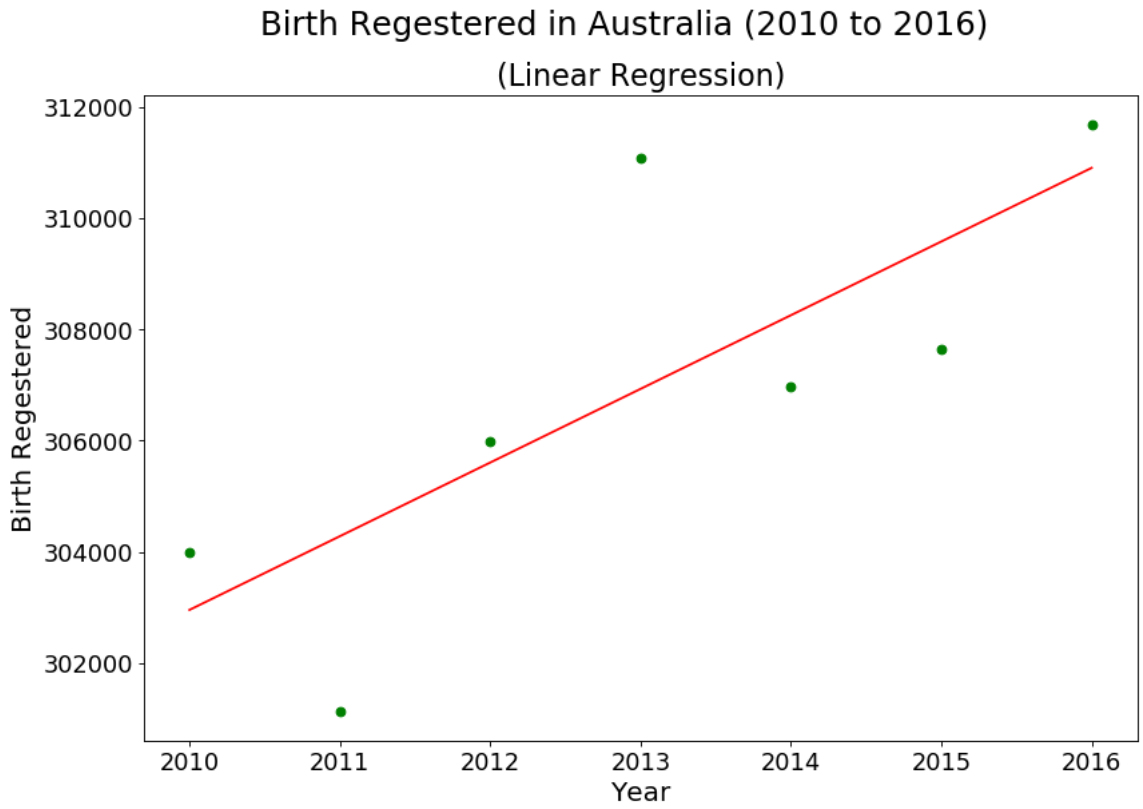
In [38]:
```
# using linear fit to predict total birth in Ausgtralia

# library to import for prediction,
# from sklearn.linear_model import LinearRegression

# fitting the equation, where X is 'Year' and y is 'Total Birth' data
# ln = LinearRegression()
# ln.fit(X, y)

x2050 = int(ln.predict([[2050]]))
x2100 = int(ln.predict([[2100]]))

print('Predicting the Total Birth in Australia(Linear Regression):')
print('\tYear 2050:  ', x2050, 'births')
print('\tYear 2010:  ', x2100, 'births')
```

```
Predicting the Total Birth in Australia(Linear Regression):
        Year 2050:   355966 births
        Year 2010:   422230 births
```

Farhad Ullah Rezwan
ID: 30270111

4. Linear regression to the most recent data points:

   The Birth Registered in Australia between 2010 and 2016 shows that this linear regression equation does not predict as expected. For example for 2013 and 2011 data the equation data does not support. According to the type of data Polynomial Regression will fit the requirements most.

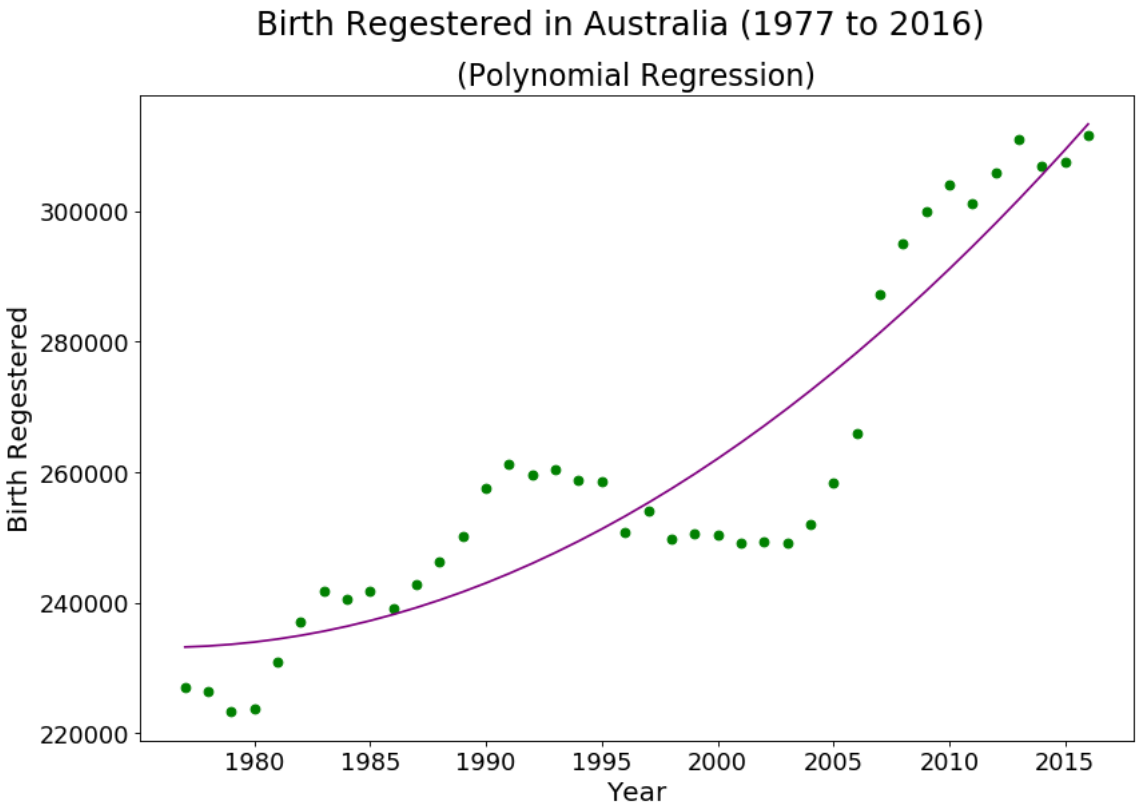


5. **Challenge:** Better model than linear regression:

   Polynomial Regression

   1. Why Polynomial Regression Better: Data with outliers seriously affects the regression result for non-linear regression, however it can predict the values better.

Farhad Ullah Rezwan
ID: 30270111



2. Use your model to predict the total births for the years 2050 and 2100.

```
In [39]: # ln2.predict(pol.fit_transform([[2050]]))

         # using polynomial regression fit to predict total birth in Ausgt

         # library to import for prediction,
         # from sklearn.linear_model import LinearRegression

         # fitting the equation, where X is 'Year' and y is 'Total Birth'
         # pol = PolynomialFeatures(degree = 2)
         # X_pol = pol.fit_transform(X)
         # pol.fit(X_pol, y)
         # ln2 = LinearRegression()
         # ln2.fit(X_pol, y)

         x2050 = int(ln2.predict(pol.fit_transform([[2050]])))
         x2100 = int(ln2.predict(pol.fit_transform([[2100]])))

         print('Predicting the Total Birth in Australia (Polynomial Regres
         print('\tYear 2050:  ', x2050, 'births')
         print('\tYear 2010:  ', x2100, 'births')

Predicting the Total Birth in Australia (Polynomial Regression):
        Year 2050:   507485 births
        Year 2010:   1003052 births
```

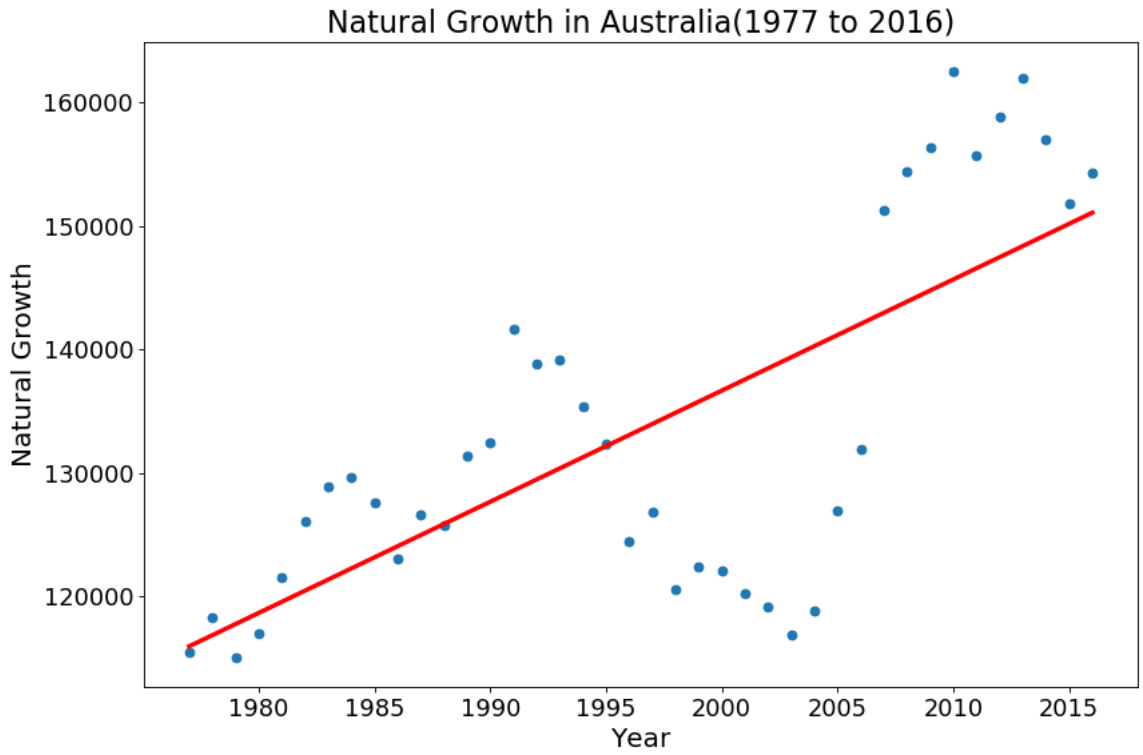
3. Total Fertility Rate (TFR.csv) for Queensland and Northern Territory.

Farhad Ullah Rezwan
ID: 30270111

i. Minimum value for TFR recorded in the dataset for Queensland = 1.8
Occurred on year: 1999
corresponding TFR value for Northern Territory in the same year is: 2.123

```
In [62]: 3.
tfQ = TFR['QLD'].min()
print('Minimum value for TFR recorded for Queensland is: ', tfQ)
TFRLowYearQ = TFR[TFR['QLD'] == tfQ]
QYear = int(TFRLowYearQ['Year'])
NTtfr = float(TFRLowYearQ['NT'])
print('lowest TFR year in the Queensland is: ', QYear)
print('Northern Terretory corresonding TFR when Queensland had least TFR:

Minimum value for TFR recorded for Queensland is:  1.8
lowest TFR year in the Queensland is:  1999
Northern Terretory corresonding TFR when Queensland had least TFR:  2.123
```

4. Natural growth in Australia's population over different years.
   i. Describe the trend in ***natural*** growth in Australian population over time using linear regression?



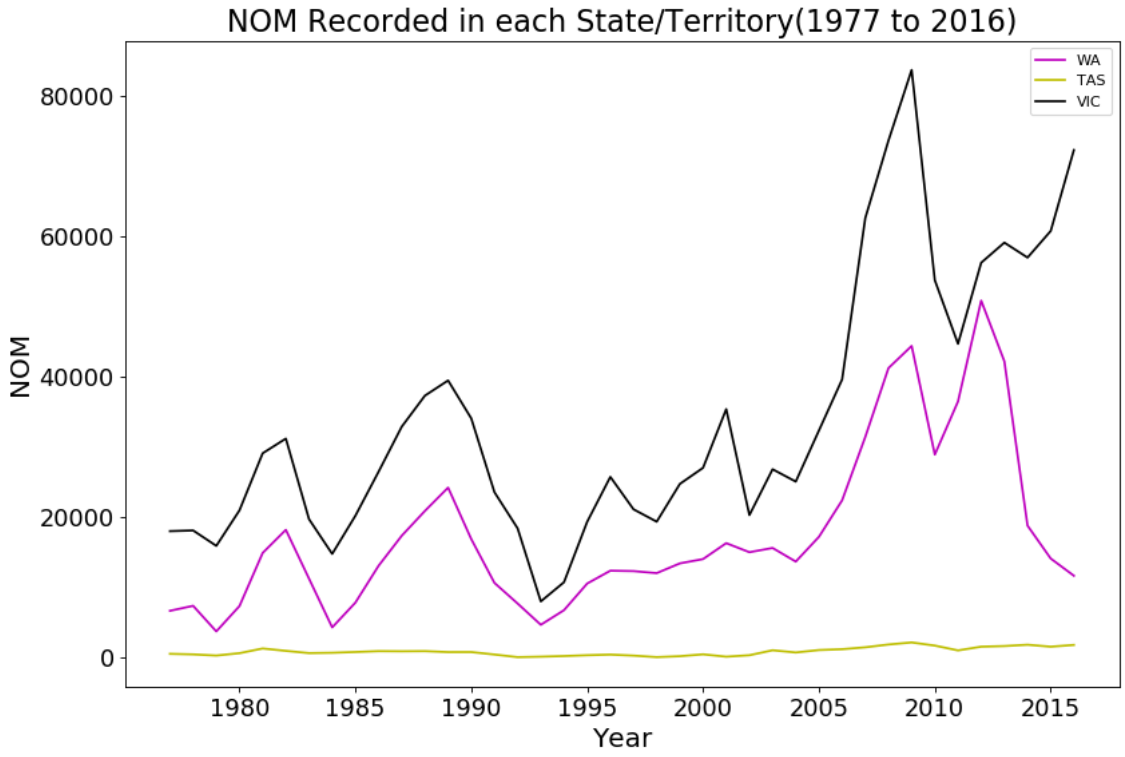
The scatter plot graph and the linear regression shows the trend of natural growth in Australian population over time. It shows that the total population is from 1977 to 2016 faced overall increase though it faced a dip around 2005.

Farhad Ullah Rezwan
ID: 30270111

A2. Investigating the Migration Data (NOM and NIM)

1. Let's look at the Net Overseas Migration (NOM) data in different states over time.

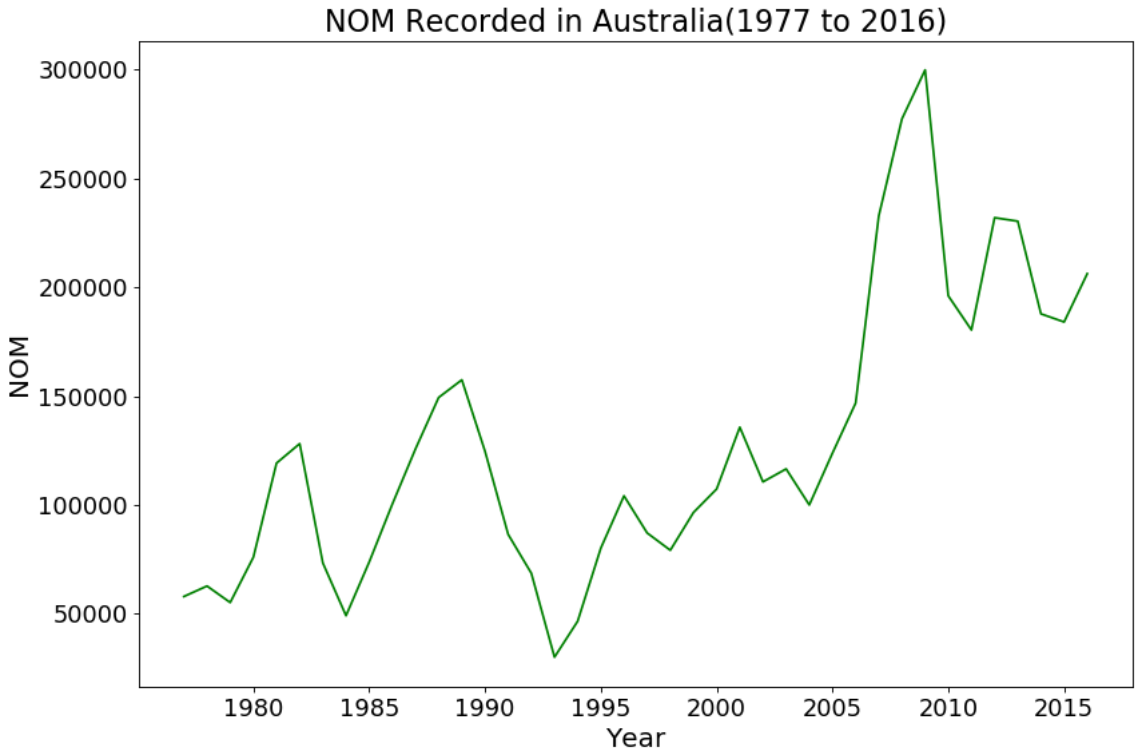
   a. Python to plot the NOM to Victoria, Tasmania and Western Australia over time.



The NOM of TAS was always consistent, however the NOM of WA and VIC was fluctuating. The fluctuation of NOM for WA and VIC was around identical till about 2012.

   b. Plot the Net Overseas Migration (NOM) to Australia over time. Do you find the trend strange?

Farhad Ullah Rezwan
ID: 30270111



2. Now let's look at the relationship between Net Overseas Migration (NOM) and Net Interstate Migration (NIM).

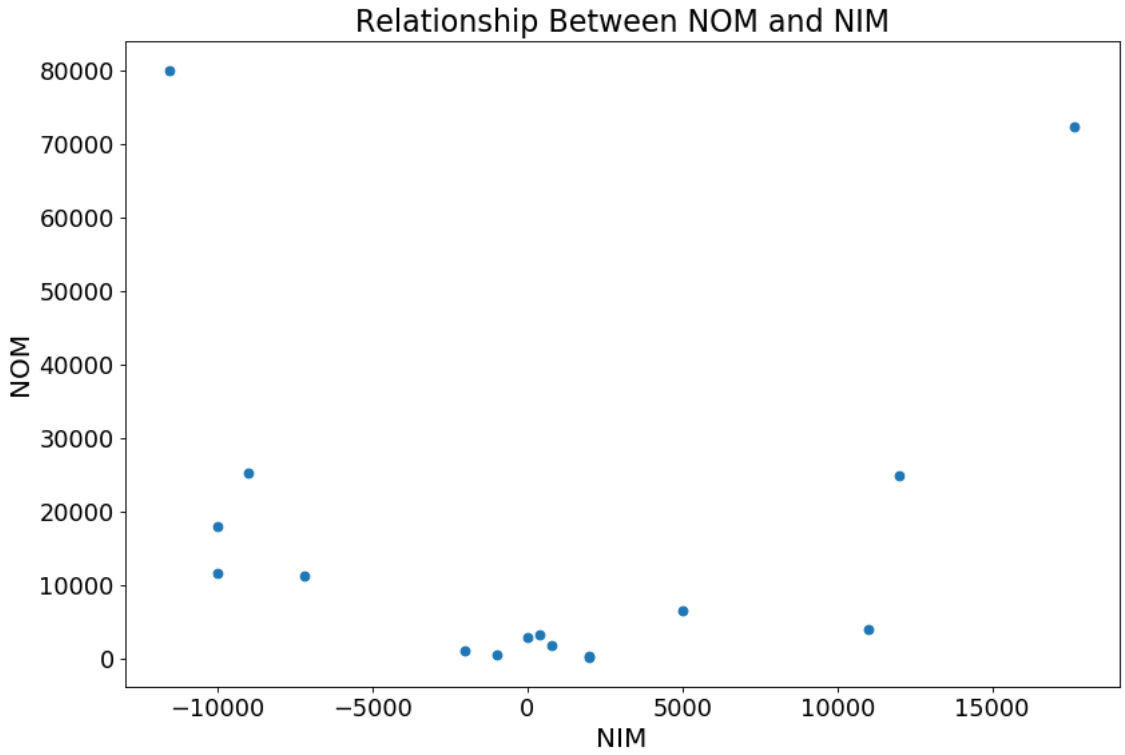
a. NOM and NIM values for each of the states for a given year. First year and last year for the combined data?

Year	NSW_NOM	VIC_NOM	QLD_NOM	SA_NOM	WA_NOM	TAS_NOM	NT_NOM	ACT_NOM
1977	25236	17969	4012	2874	6631	506	408	261
2016	80007	72215	24952	11283	11621	1771	1048	3330

NSW_NIM	VIC_NIM	QLD_NIM	SA_NIM	WA_NIM	TAS_NIM	NT_NIM	ACT_NIM
-9000	-10000	11000	0	5000	-1000	2000	2000
-11539	17639	11986	-7212	-10010	760	-2029	383

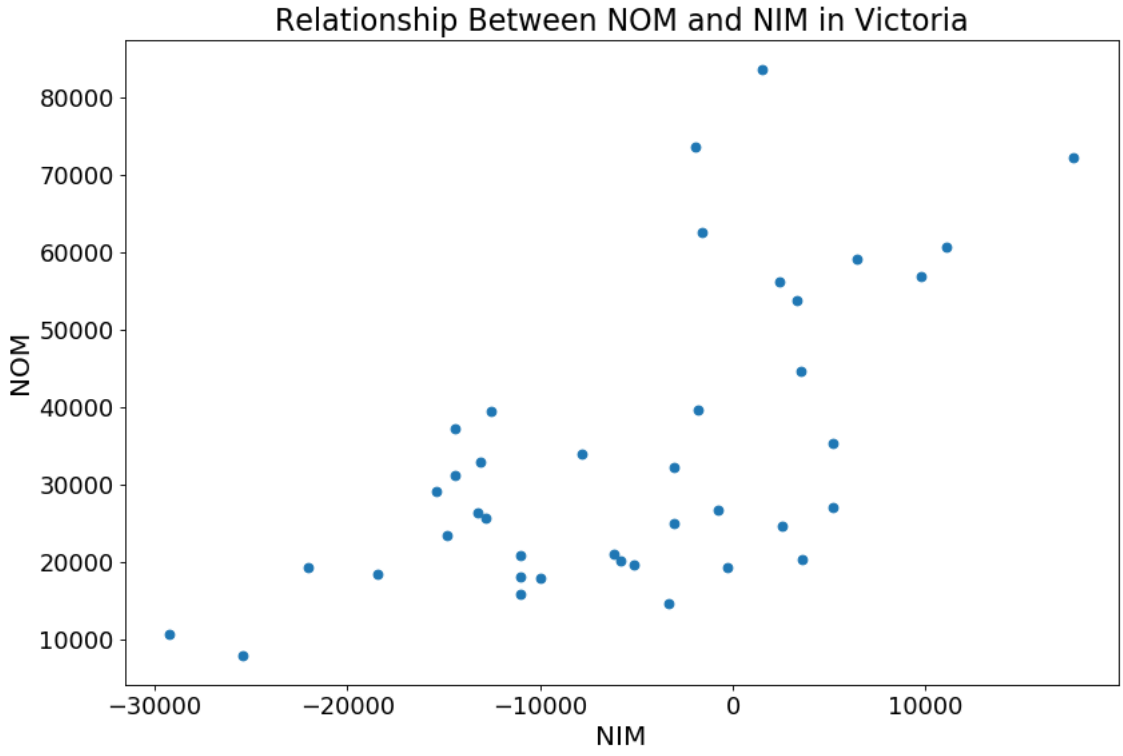
b. Relationship between NOM and NIM.

Farhad Ullah Rezwan
ID: 30270111



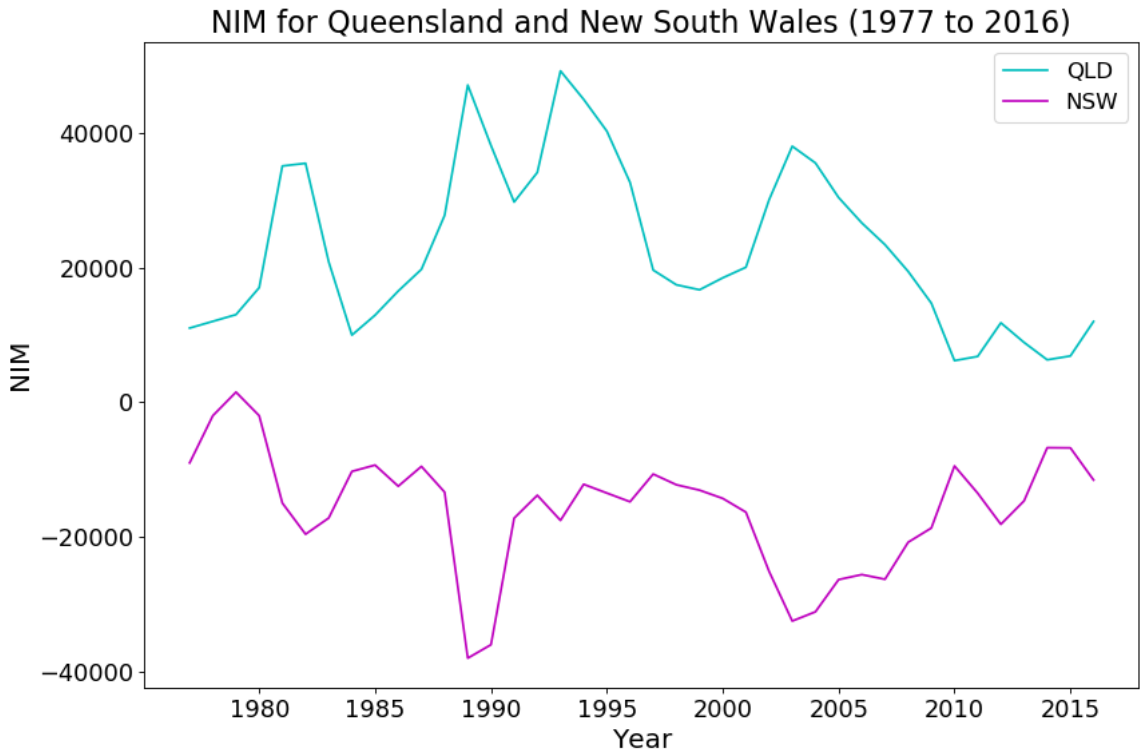
The relationship between NOM and NIM shows that when the NOM increases for a particular state, NIM decreases. There is a inverse relationship between NOM and NIM in Australian population.

c. Try selecting and plotting the data for Victoria only using scatter plot. Can you see a relationship now? If so, explain the relationship.



Here the inverse relationship between NOM and NIM is not that effective for the case of Victoria state.

d. Net Interstate Migration (NIM) for Queensland and New South Wales over different years.

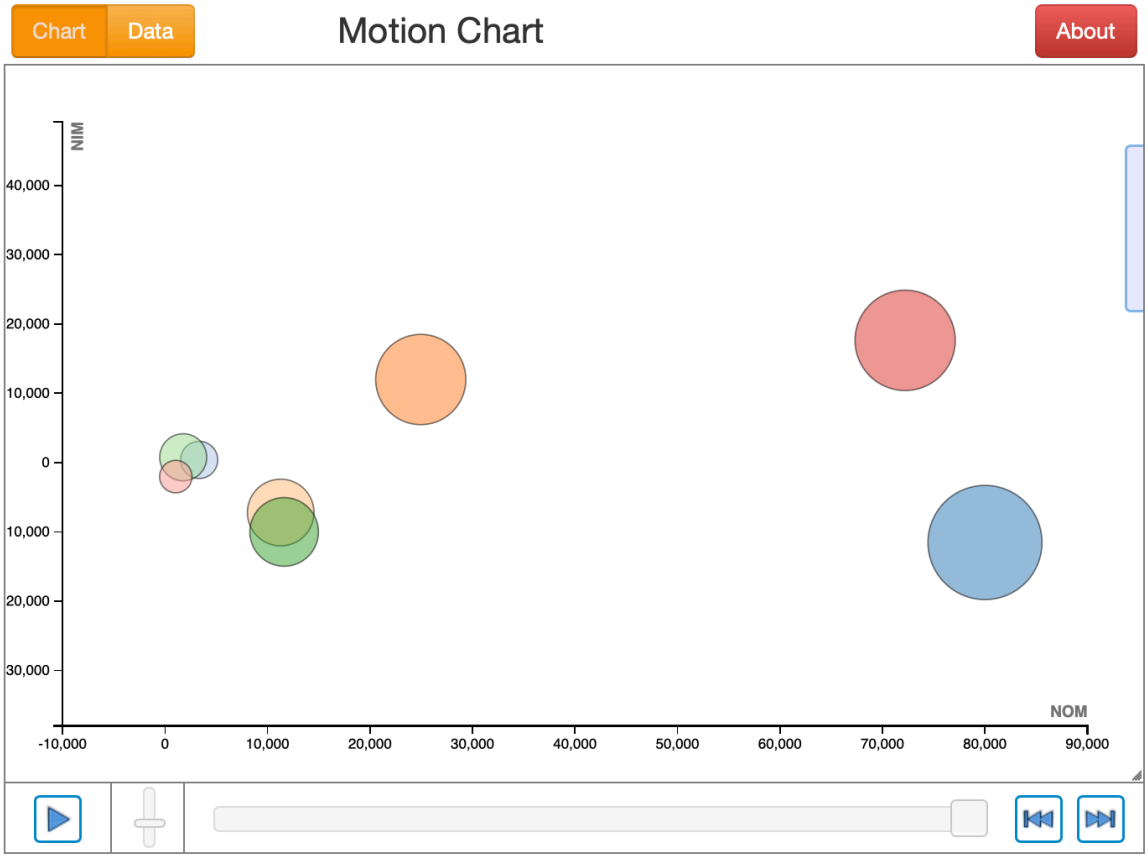


The graph shows that during the period of 1977 and 2016, the Net Interstate Migration was totally opposite for QLD and NSW.

A3. Visualising the Relationship over Time

1. Role of Migration (overseas and interstate) towards population growth in each Australia

Farhad Ullah Rezwan
ID: 30270111



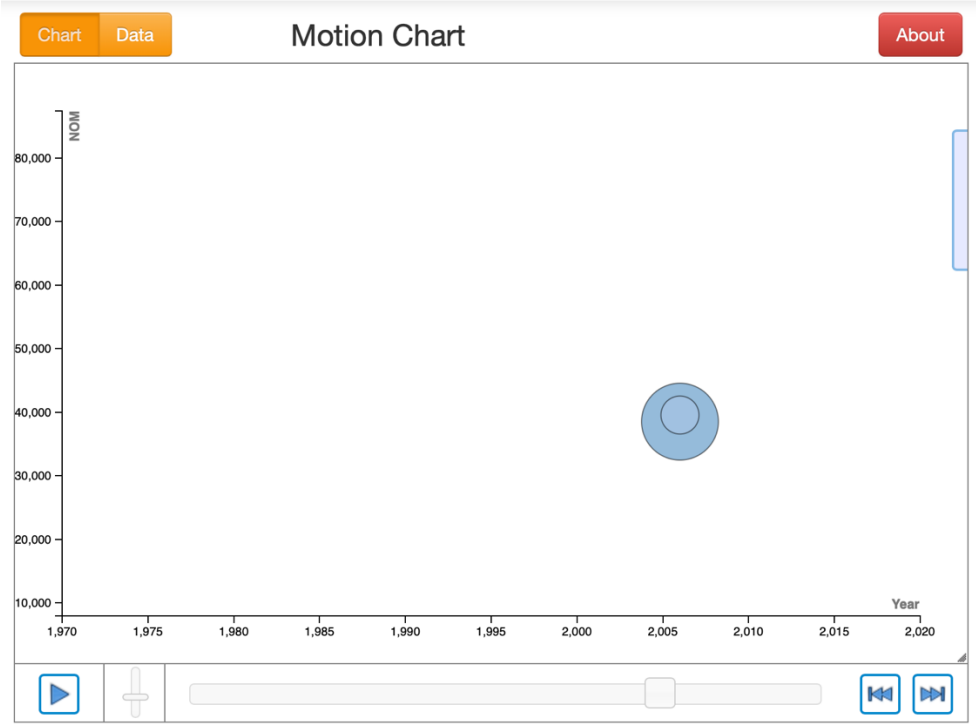
This Motion chart points out that the higher the NOM , the total population growth is significant. For total population growth in Australia, both NIM and NOM has major impact. States where NOM and NIM values are close to zero, has less total population growth.

2. Run the visualisation from start to end. (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:

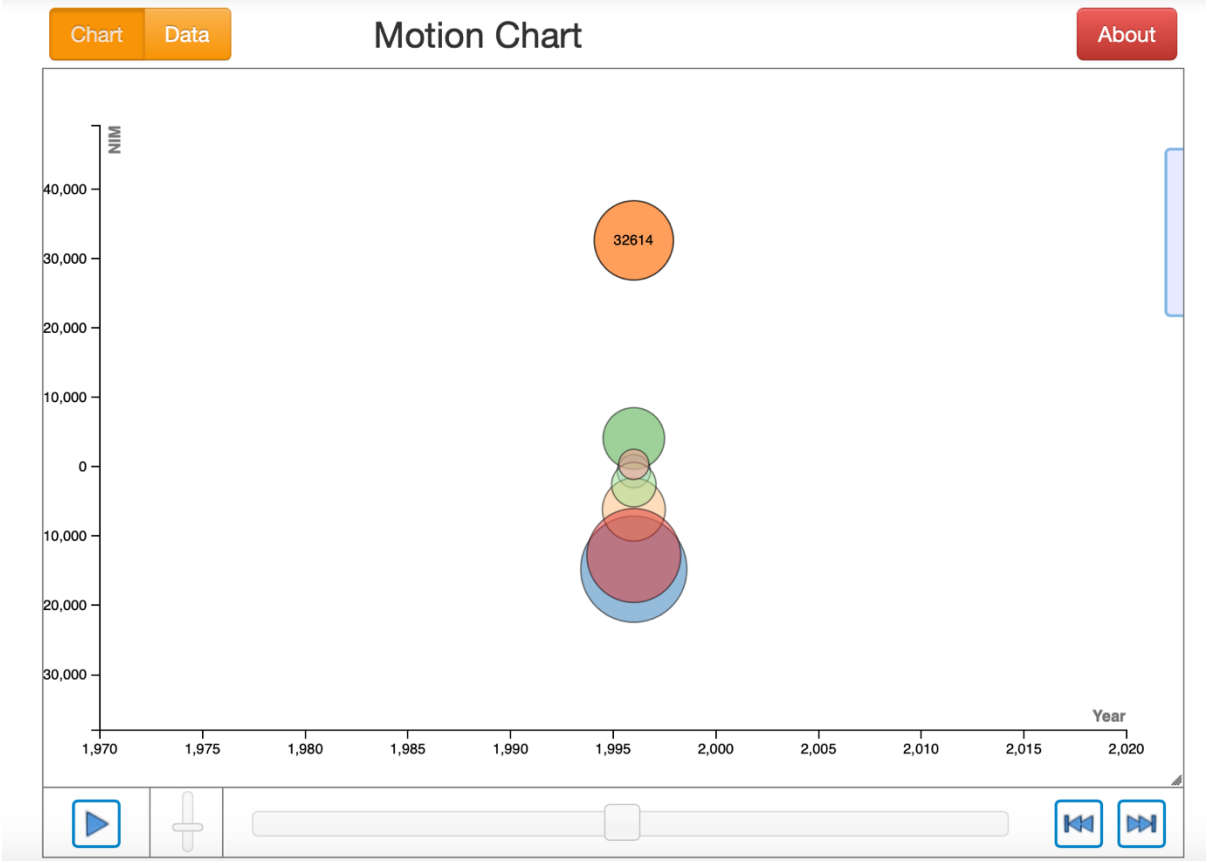
a. Comment generally on the trend you see in Net Overseas Migration (NOM) and Net Interstate Migration (NIM) overtime. Is there any relationship between the two variables?

c. In the year 2006, VIC has higher NOM then NSW, as shown in the motion chart.

Farhad Ullah Rezwan
ID: 30270111



d. According to the Motion chart, QLD has the highest NIM during the period.



QLD

Farhad Ullah Rezwan
ID: 30270111

Task B: Exploratory Analysis of Data

B1. Daily number of crimes

1. For each suburb, number of days that at least 15 crimes have occurred per day.

	Suburb - Incident	Number of days atleast 15 crimes happend
0	ABERFOYLE PARK	0.0
1	ADDRESS UNKNOWN	0.0
2	ADELAIDE	877.0
3	ADELAIDE AIRPORT	0.0
4	AGERY	0.0
5	ALAWOONA	0.0
6	ALBANY	0.0
7	ALBERT PARK	0.0
8	ALBERTON	0.0
9	ALDGATE	0.0
10	ALDINGA	0.0
11	ALDINGA BEACH	0.0
12	ALFORD	0.0
13	ALLENBY GARDENS	0.0
14	ALLENDALE EAST	0.0
15	ALLENDALE NORTH	0.0

Farhad Ullah Rezwan
ID: 30270111

Suburb - Incident	Number of days atleast 15 crimes happend	
16	ALMA	0.0
17	ALTONA	0.0
18	AMATA	0.0
19	AMERICAN RIVER	0.0
20	ANAMA	0.0
21	ANANGU PITJANTJATJARA YANKUNYTJATJARA	0.0
22	ANDAMOOKA	0.0
23	ANDREWS FARM	0.0
24	ANGAS PLAINS	0.0
25	ANGAS VALLEY	0.0
26	ANGASTON	0.0
27	ANGLE PARK	0.0
28	ANGLE VALE	0.0
29	ANNADALE	0.0
...	...	...
1597	WYNN VALE	0.0
1598	WYOMI	0.0
1599	YACKA	0.0

Farhad Ullah Rezwan
ID: 30270111

Suburb - Incident	Number of days atleast 15 crimes happend	
1600	YAHL	0.0
1601	YALATA	0.0
1602	YALLUNDA FLAT	0.0
1603	YAMBA	0.0
1604	YANINEE	0.0
1605	YANKALILLA	0.0
1606	YANKANINNA	0.0
1607	YANYARRIE	0.0
1608	YARANYACKA	0.0
1609	YARDEA	0.0
1610	YATALA VALE	0.0
1611	YATINA	0.0
1612	YATTALUNGA	0.0
1613	YEELANNA	0.0
1614	YELTA	0.0
1615	YINKANIE	0.0
1616	YONGALA	0.0
1617	YONGOLA	0.0

Farhad Ullah Rezwan
ID: 30270111

Suburb - Incident	Number of days atleast 15 crimes happend	
1618	YORKE VALLEY	0.0
1619	YORKETOWN	0.0
1620	YOUNG HUSBAND	0.0
1621	YOUNGHUSBAND	0.0
1622	YOUNGHUSBAND HOLDINGS	0.0
1623	YUMALI	0.0
1624	YUNDI	0.0
1625	YUNTA	0.0
1626	ZADOWS LANDING	0.0

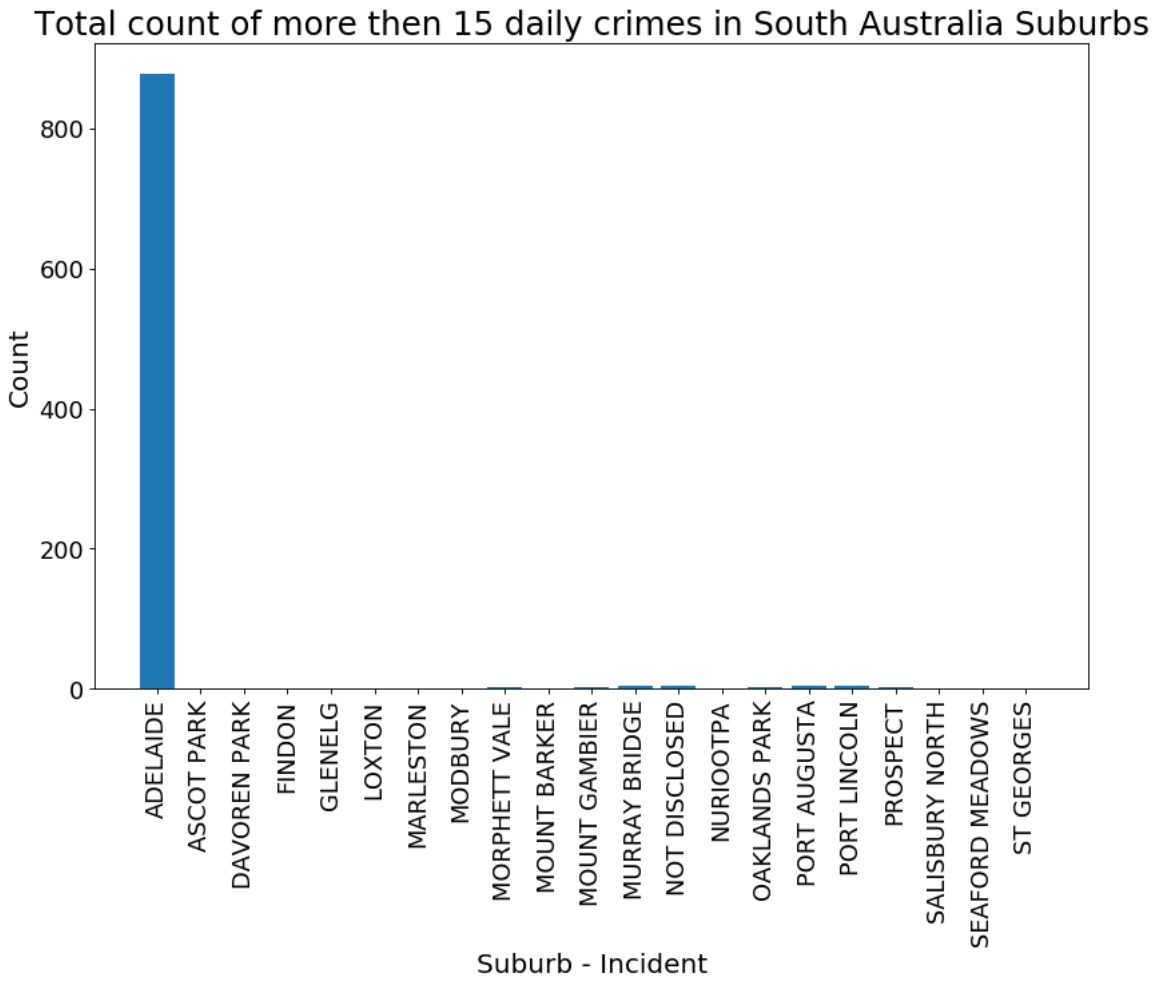
1627 rows × 2 columns

2. Now which suburbs do have at least one day where the daily number of crimes are more than 15.

Farhad Ullah Rezwan
ID: 30270111

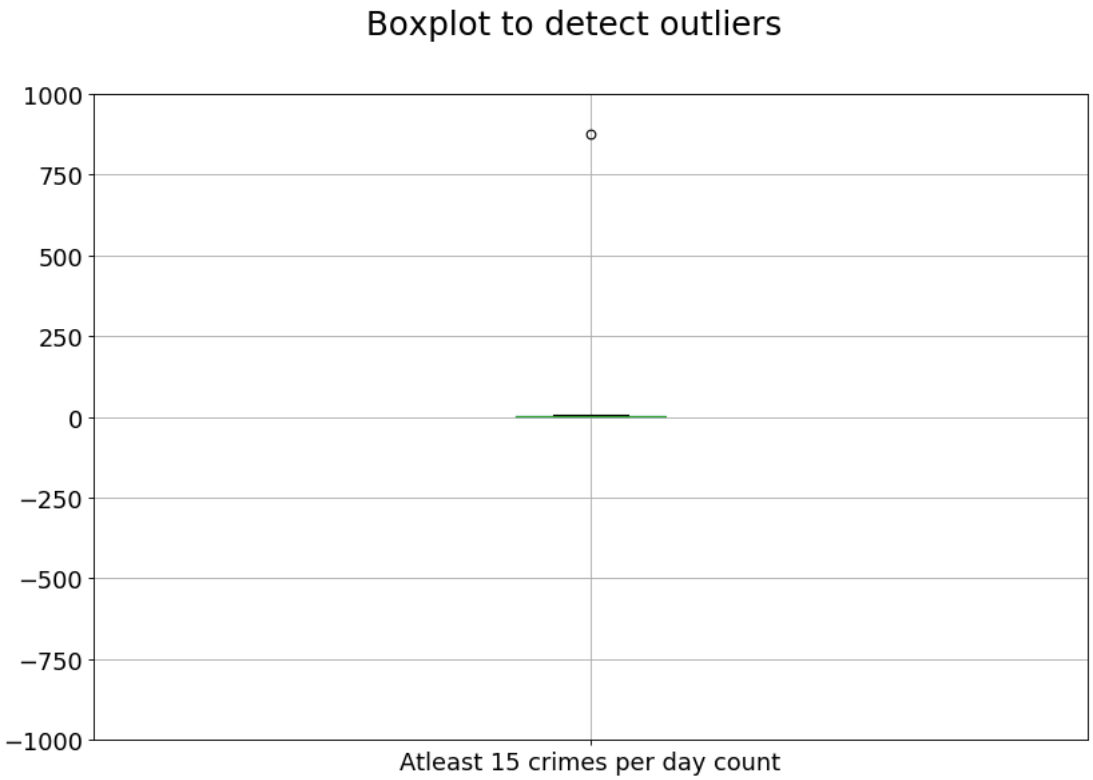
	Suburb - Incident	Atleast 15 crimes per day count
0	ADELAIDE	877.0
1	ASCOT PARK	1.0
2	DAVOREN PARK	1.0
3	FINDON	1.0
4	GLENELG	1.0
5	LOXTON	1.0
6	MARLESTON	1.0
7	MODBURY	1.0
8	MORPHETT VALE	3.0
9	MOUNT BARKER	1.0
10	MOUNT GAMBIER	3.0
11	MURRAY BRIDGE	5.0
12	NOT DISCLOSED	5.0
13	NURIOOTPA	1.0
14	OAKLANDS PARK	3.0
15	PORT AUGUSTA	4.0
16	PORT LINCOLN	5.0
17	PROSPECT	2.0
18	SALISBURY NORTH	1.0
19	SEAFORD MEADOWS	1.0
20	ST GEORGES	1.0

Farhad Ullah Rezwan
ID: 30270111



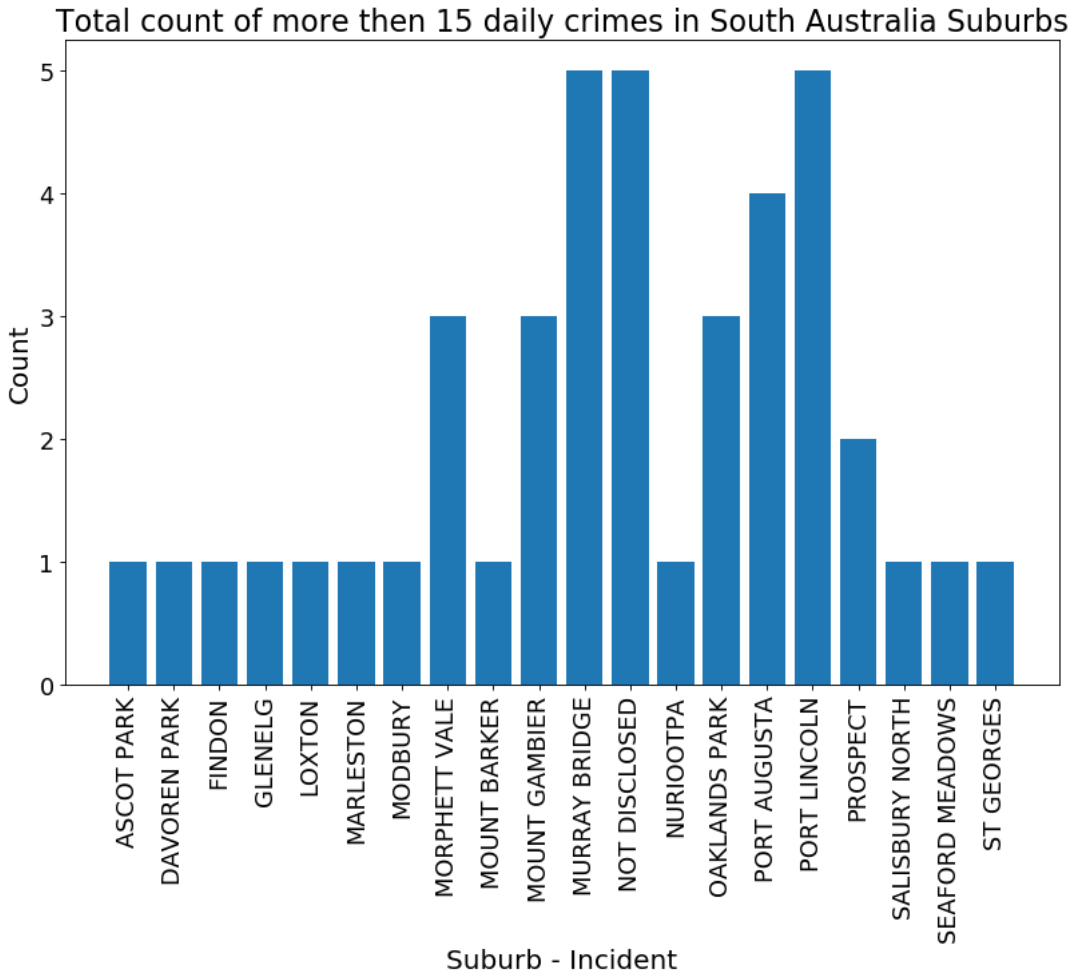
3. Use an appropriate graph to visualize and detect outliers (extreme values) on the data from step 2 and remove them. Then, plot the data again using a bar graph.

Farhad Ullah Rezwan
ID: 30270111

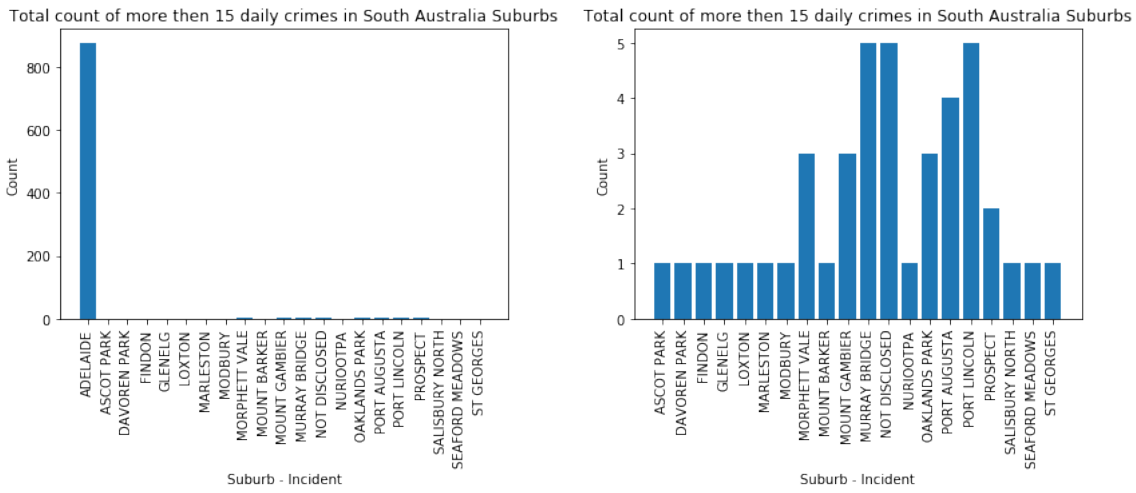


Bar graph with outlier removed dataframe.

Farhad Ullah Rezwan
ID: 30270111



4. Compare the bar graphs in step 2 and 3. Which bar graph is easier to interpret? Why?



Without the outliers of suburb Adilade the bar graph in the left shows the incedent count properly here.

Farhad Ullah Rezwan
ID: 30270111