International Student Forecasting

August 9, 2020

[83]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

1 Load the files

The file was obtained from The Department of Home Affairs (https://data.gov.au/dataset/ds-dga-324aa4f7-46bb-4d56-bc2d-772333a2317e/details). This is a government website, hence the data is considered to be reliable. The excel worksheet was orginally unreadable through the normal read functions. The data was extracted in readable form via converting it to a text file and getting rid of all the unwanted items.

[91]:
```
df = pd.read_csv('Visas_lodged.csv',header=1, thousands=',')
print(df.shape)
df.head()
```

```
(17, 17)
```

[91]:

	Applicant Type	Sector	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20 to 30 June 2020
0	Primary	Foreign Affairs or Defence Sector	2840	2709	2510	2534	2573	2892	3945	4940	4768	4597	3917	5002	5008	3861	2739
1	Primary	Higher Education Sector	94796	108025	118781	122455	109814	104869	107650	124492	144874	148674	155486	163204	179282	206637	169940
2	Primary	Independent ELICOS Sector	26603	30522	29811	38142	33891	29547	28939	29870	30596	32736					
3	Primary	Non-Award Sector	17749	17823	20975	19244	18064	17250	16314	17716	19413	19333					
4	Primary	Postgraduate Research Sector	3826	4172	4378	5171	5393	5649	6204	6484	6532	6524					

```
2    35944    41502    40596    42284                  33503
3    20443    20592    21104    20466                  12217
4     6527     6663     7096     7797                   7466
```

Due to the anomaly caused by COVID-19, the last column will be dropped and analysed separately month by month

[92]:
```
df.drop(['2019-20 to 30 June 2020'],inplace=True,axis=1)
df.columns
```

[92]:
```
Index(['Applicant Type', 'Sector', '2005-06', '2006-07', '2007-08', '2008-09',
       '2009-10', '2010-11', '2011-12', '2012-13', '2013-14', '2014-15',
       '2015-16', '2016-17', '2017-18', '2018-19'],
      dtype='object')
```

Most of the applicants are primary applicants and there cannot be secondary applicants without primary hence the analysis will be conducted on them.

[139]:
```
primary_df = df[df['Applicant Type']=='Primary']
primary_df.set_index('Sector',inplace=True);
primary_df.drop('Applicant Type',inplace=True,axis=1);
print(primary_df.shape)
primary_df
```

```
(7, 14)
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-
packages/pandas/core/frame.py:3997: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  errors=errors,
```

[139]:

Sector	2005-06	2006-07	2007-08	2008-09	\
Foreign Affairs or Defence Sector	2840	2709	2510	2534	
Higher Education Sector	94796	108025	118781	122455	
Independent ELICOS Sector	26603	30522	29811	38142	
Non-Award Sector	17749	17823	20975	19244	
Postgraduate Research Sector	3826	4172	4378	5171	
Schools Sector	14880	18482	20445	15806	
Vocational Education and Training Sector	32350	46781	66335	101099	

Sector	2009-10	2010-11	2011-12	2012-13	\
Foreign Affairs or Defence Sector	2573	2892	3945	4940	
Higher Education Sector	109814	104869	107650	124492	

```
Independent ELICOS Sector                    33891    29547    28939    29870
Non-Award Sector                             18064    17250    16314    17716
Postgraduate Research Sector                  5393     5649     6204     6484
Schools Sector                               13184    11115    10309     9988
Vocational Education and Training Sector     66891    63634    61861    54355

                                           2013-14  2014-15  2015-16  2016-17  \
Sector
Foreign Affairs or Defence Sector             4768     4597     3917     5002
Higher Education Sector                     144874   148674   155486   163204
Independent ELICOS Sector                    30596    32736    35944    41502
Non-Award Sector                             19413    19333    20443    20592
Postgraduate Research Sector                  6532     6524     6527     6663
Schools Sector                               11446    13195    13653    13857
Vocational Education and Training Sector     54606    61507    76992    77626

                                           2017-18  2018-19
Sector
Foreign Affairs or Defence Sector             5008     3861
Higher Education Sector                     179282   206637
Independent ELICOS Sector                    40596    42284
Non-Award Sector                             21104    20466
Postgraduate Research Sector                  7096     7797
Schools Sector                               12686    12286
Vocational Education and Training Sector     95405   114470
```

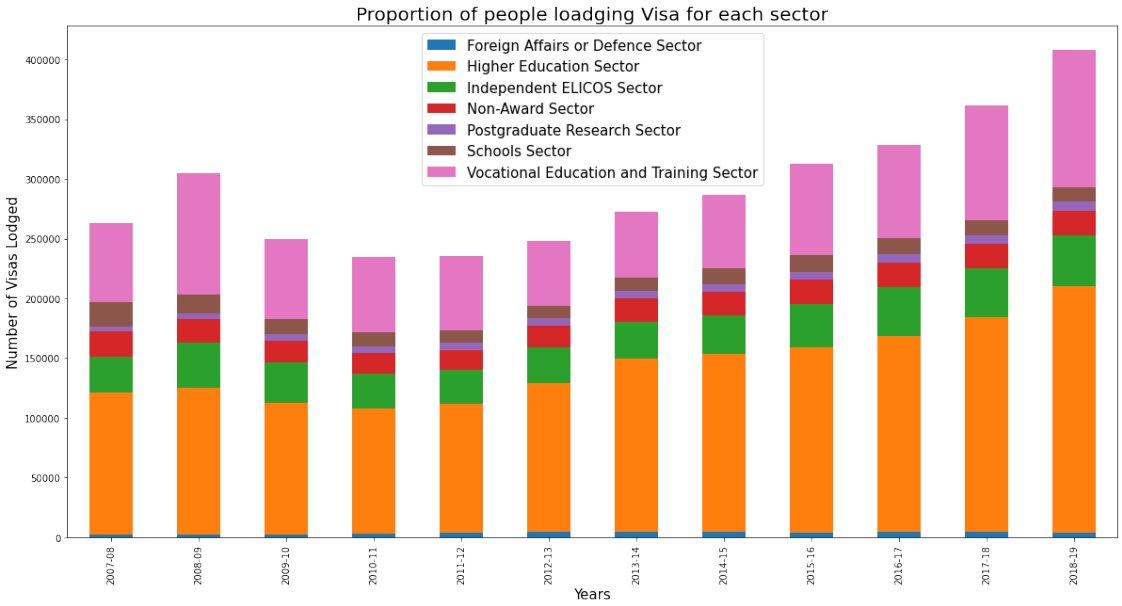
2 Inital Visualization

[94]:
```
cols = np.array(primary_df.columns)
mask = np.ones(len(cols), dtype=bool)
mask[[0,1]] = False
cols = cols[mask]

primary_df[cols].astype('int').transpose().plot(kind='bar',
                                                stacked=True,
                                                figsize=(20,10));


plt.legend(fontsize = 15);
plt.xlabel('Years',fontsize=15);
plt.ylabel('Number of Visas Lodged',fontsize=15);
plt.title('Proportion of people loadging Visa for each sector',fontsize=20);


plt.show()
# plt.savefig('Hmm.png')
```



To see the similarity in trends, all the sectors will be normalised. If all sectors follow similar trends, then the total of all the sectors can be modelled and the predictions can be made.

[95]:
```
norm_primary = primary_df.transpose().copy()
for i in primary_df.index.values:
    norm_primary[i]=((norm_primary[i])/(norm_primary[i].max()))
norm_primary.transpose().head()
```

[95]:

Sector	2005-06	2006-07	2007-08	2008-09
Foreign Affairs or Defence Sector	0.567093	0.540935	0.501198	0.505990
Higher Education Sector	0.458756	0.522777	0.574829	0.592609
Independent ELICOS Sector	0.629151	0.721833	0.705018	0.902043
Non-Award Sector	0.841025	0.844532	0.993887	0.911865
Postgraduate Research Sector	0.490702	0.535078	0.561498	0.663204

Sector	2009-10	2010-11	2011-12	2012-13
Foreign Affairs or Defence Sector	0.513778	0.577476	0.787740	0.986422
Higher Education Sector	0.531434	0.507503	0.520962	0.602467
Independent ELICOS Sector	0.801509	0.698775	0.684396	0.706414
Non-Award Sector	0.855951	0.817381	0.773029	0.839462
Postgraduate Research Sector	0.691676	0.724509	0.795691	0.831602

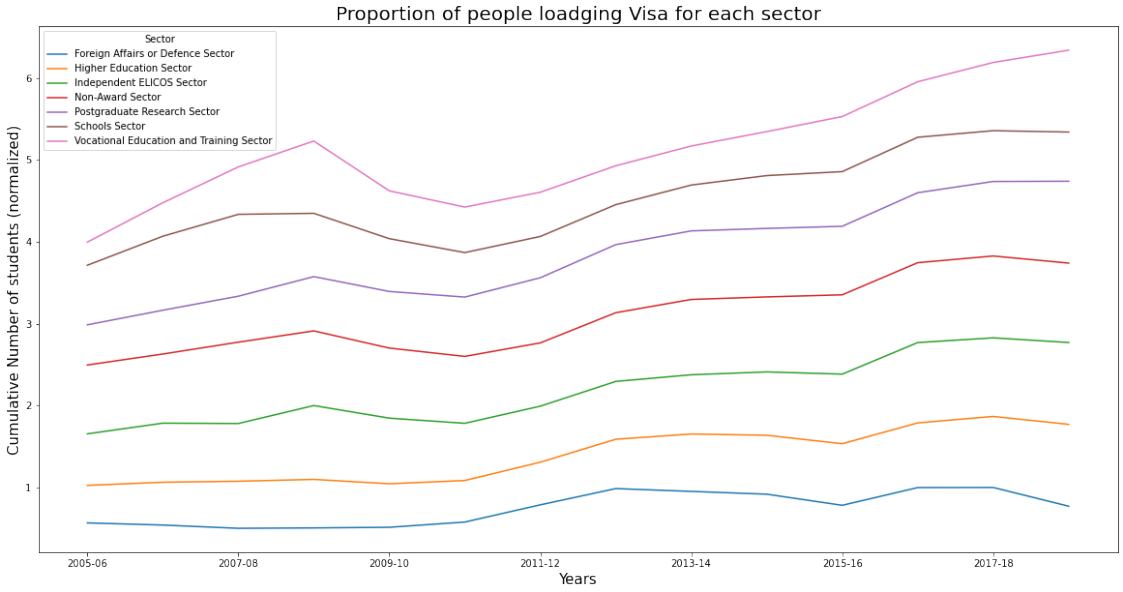
Sector	2013-14	2014-15	2015-16	2016-17
Foreign Affairs or Defence Sector	0.952077	0.917931	0.782149	0.998802

```
Higher Education Sector            0.701104  0.719494  0.752460  0.789810
Independent ELICOS Sector          0.723583  0.774194  0.850061  0.981506
Non-Award Sector                   0.919873  0.916082  0.968679  0.975739
Postgraduate Research Sector       0.837758  0.836732  0.837117  0.854559
```

Sector	2017-18	2018-19
Foreign Affairs or Defence Sector	1.000000	0.770966
Higher Education Sector	0.867618	1.000000
Independent ELICOS Sector	0.960079	1.000000
Non-Award Sector	1.000000	0.969769
Postgraduate Research Sector	0.910094	1.000000

[141]:
```
norm_primary.plot(kind='line',
                  stacked=True,
                 figsize=(20,10)
                 )
plt.xlabel('Years',fontsize=15);
plt.ylabel('Cumulative Number of students (normalized)',fontsize=15);
plt.title('Proportion of people loadging Visa for each sector',fontsize=20);
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-
packages/pandas/plotting/_matplotlib/core.py:1192: UserWarning: FixedFormatter
should only be used together with FixedLocator
  ax.set_xticklabels(xticklabels)
```



Since all the trends are very similar to each other with more differences, the forecasting can be estimated via the grand total insted forcasting each sector individually

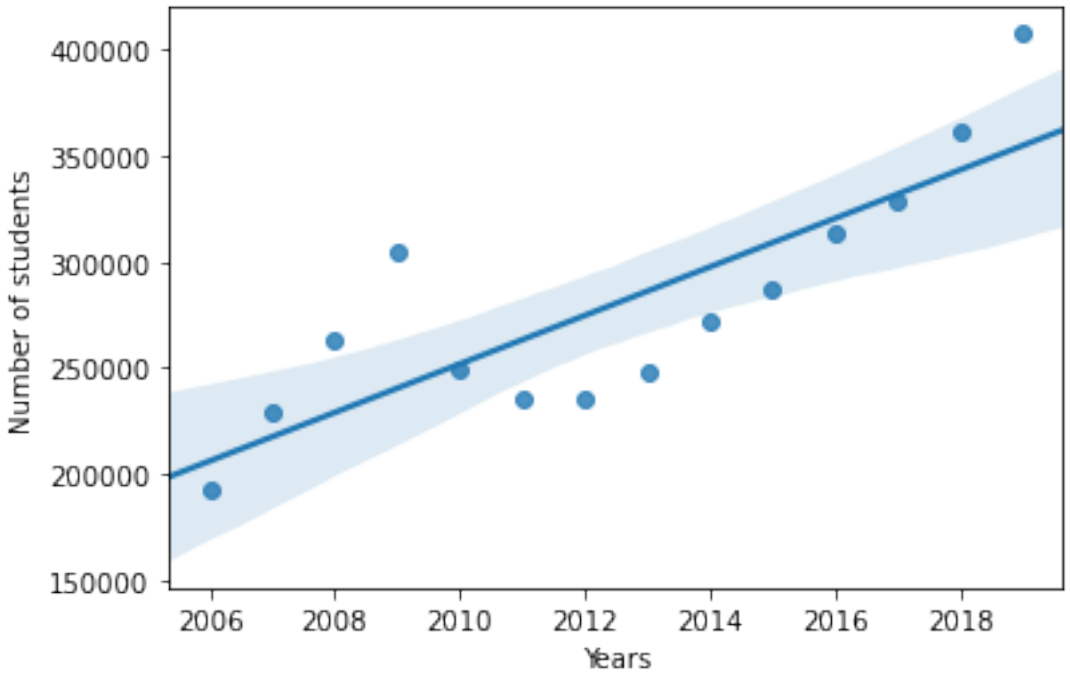
3 Analysis

[97]:
```
total_primary_df = pd.DataFrame(primary_df.sum())
total_primary_df.columns = ['Number of students']
total_primary_df.tail()
```

[97]:

	Number of students
2014-15	286566
2015-16	312962
2016-17	328446
2017-18	361177
2018-19	407801

[157]:
```
sns.regplot(np.linspace(2006,2019,len(total_primary_df)),'Number of␣
 ↪students',data=total_primary_df)
plt.xlabel('Years');
```



We can try fitting in a linear model and check the R^2 value

[99]:
```
from sklearn.linear_model import LinearRegression
```

[100]:
```
lm = LinearRegression()
y = total_primary_df[['Number of students']]
y = np.array(y).reshape(-1,1)
x = np.linspace(2006,2019,len(total_primary_df)).reshape(-1,1)
```

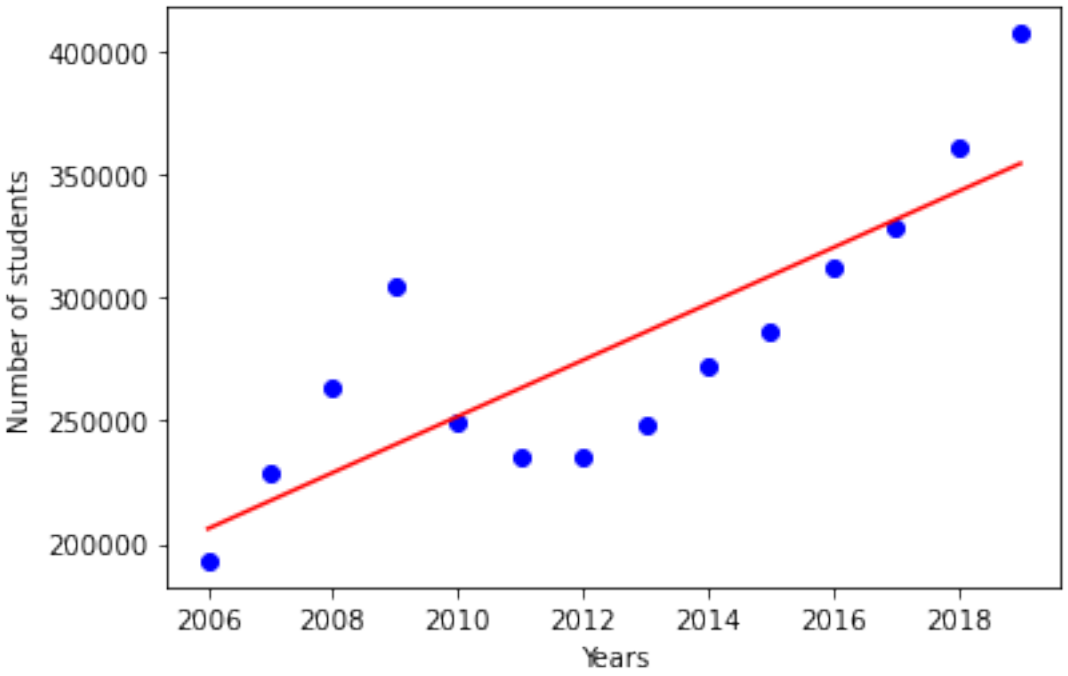
```
lm.fit(x,y)
```

[100]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
```

[101]:
```
print("The slope of the best fit lne in linear regression is ",lm.coef_)
```

```
The slope of the best fit lne in linear regression is  [[11441.42857143]]
```

[158]:
```
plt.scatter(x,y,  color='blue')
plt.plot(x, lm.coef_*x + lm.intercept_, '-r')
plt.xlabel('Years');
plt.ylabel('Number of students');
```



[103]:
```
yhat = lm.predict(x)
yhat[0:4]
```

[103]:
```
array([[206078.14285714],
       [217519.57142857],
       [228961.        ],
       [240402.42857143]])
```

[104]:
```
(yhat-y)[0:5]
```

[104]:
```
array([[ 13034.14285714],
       [-10994.42857143],
       [-34274.        ],
       [-64048.57142857],
       [  2033.85714285]])
```

So we see a $\pm$ devation of about 10000 students which means the ppredictions are not very good

4 Accuracy Check

[105]:
```
from sklearn.metrics import r2_score
```

[106]:
```
r2_score(y,yhat)
```

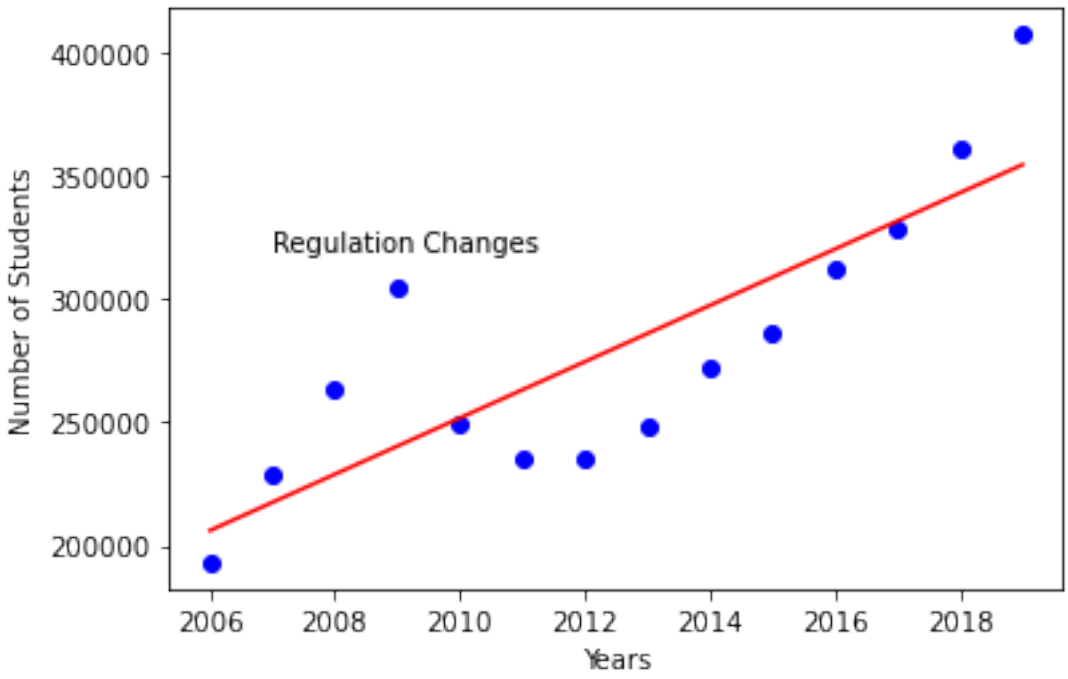
[106]:
```
0.684193933212412
```

This value is considered moderately-strongly effective in forecasting the future scenarios

5 Increasing the accuracy

There is a sharp decrease in the number of students observed between the years 2009 and 2010. It is around that time that we see a sharp decrease in the number of students coming in through the vocational education and training sector as well as a minor decrease in the other sectors. This decrease was the result of a declination of visas from a large number of students due to change in visa regulations and the general skilled residency program (https://www.abs.gov.au/ausstats/abs@.nsf/lookup/4102.0main+features20dec+2011).

[159]:
```
plt.scatter(x,y,  color='blue')
plt.plot(x, lm.coef_*x + lm.intercept_, '-r')
plt.annotate('Regulation Changes',(2007,320000))
plt.xlabel('Years');
plt.ylabel('Number of Students');
plt.show()
```



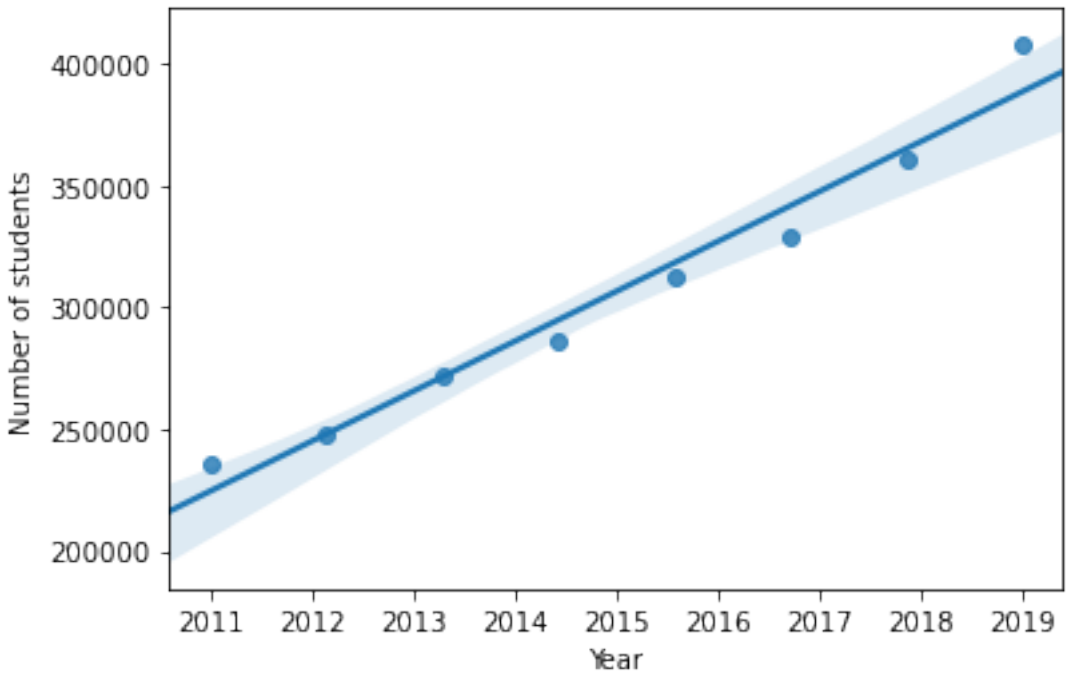
Since an external factor (i.e regulation changes) affected the normal trend of the flux of international students in australia, the data points from before 2010 can be omitted so as to forecast the flux of students with the current policies

[21]:
```
currentreg_df = total_primary_df[6:len(total_primary_df)]
currentreg_df.head()
```

[21]:

	Number of students
2011-12	235222
2012-13	247845
2013-14	272235
2014-15	286566
2015-16	312962

[155]:
```
sns.regplot(np.linspace(2011,2019,len(currentreg_df)),'Number of␣
 ↪students',data=currentreg_df)
plt.xlabel("Year");
```



[23]:
```
reg = LinearRegression()
y = currentreg_df[['Number of students']]
y = np.array(y).reshape(-1,1)
x = np.linspace(2006,2019,len(currentreg_df)).reshape(-1,1)

reg.fit(x,y)
```

[23]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
```

[24]:
```
yhat = reg.predict(x)
yhat[0:4]
```

[24]:
```
array([[224459.16666667],
       [247908.47619048],
       [271357.78571429],
       [294807.0952381 ]])
```

[25]:
```
(yhat-y)
```

[25]:
```
array([[-10762.83333333],
       [    63.47619048],
       [  -877.21428571],
       [  8241.0952381 ],
       [  5294.4047619 ],
```

```
       [ 13259.71428571],
       [  3978.02380952],
       [-19196.66666666]])
```

[26]:
```
r2_score(y,yhat)
```

[26]:
```
0.9676244700172266
```

So we see that this model works very well with the data we have after the regulation change.

If every thing were normal, in the next 3 years, there would be more students following this trend.

6 Predicting the nmber of visas lodged in future years

[49]:
```
threeyr_students_normal = reg.predict(np.array([2020,2021,2022]).reshape(-1,1))
threeyr_students_normal
```

[49]:
```
array([[401230.88461538],
       [413857.43589744],
       [426483.98717949]])
```

[53]:
```
print("Australia would have had about",round(np.
 ↪sum(threeyr_students_normal)),"new visas lodged in the next 3 years if␣
 ↪everything were normal")
```

```
Australia would have had about 1241572 new visas lodged in the next 3 years if
everything were normal
```

However, in the first few months of 2020, the COVID-19 pandemic broke out resulting in travel bans all accross the world and many students hesitating to go overseas to study. Common intuition and perception would be that a sharp decrease in the number of visas lodged will be observed.

7 Effect of COVID-19 on visa appications in the education sector

The month wise data was obtained from (https://internationaleducation.gov.au/research/International-Student-Data/Pages/InternationalStudentData2020.aspx)

[84]:
```
sector_monthly_df = pd.read_csv('Monthy_Visa_17-20.csv')
sector_monthly_df.head()
```

[84]:

	Sector	Month	2017	2018	2019	2020
0	Higher Education	Jan	203099	236097	267097	283966
1	Higher Education	Feb	261580	298736	321514	329819
2	Higher Education	Mar	277392	316975	357634	361130
3	Higher Education	Apr	278608	317776	358631	361685
4	Higher Education	May	279943	319255	360355	362992

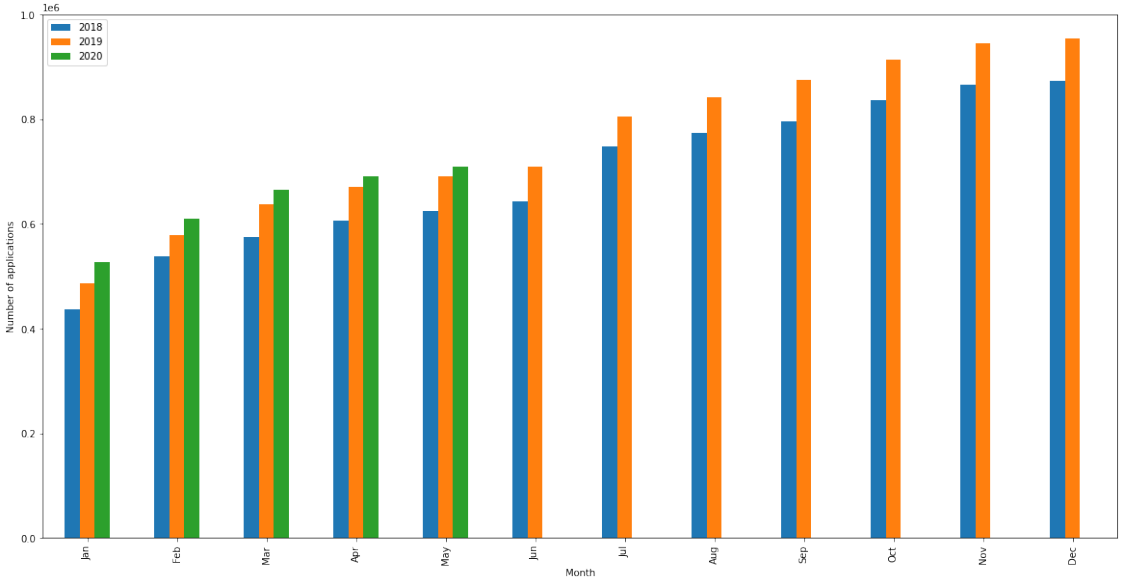
Now, lets take the total nuber of visas lodged per month

[85]:
```
monthly_df = sector_monthly_df.groupby('Month',sort=False).sum()
monthly_df.head()
```

[85]:

Month	2017	2018	2019	2020
Jan	387532	436911	485907	526831
Feb	481209	537594	578884	609828
Mar	517139	575613	636595	664290
Apr	542618	605907	671244	691002
May	564010	624642	691253	708671

[161]:
```
monthly_df[['2018','2019','2020']].plot(kind='bar',figsize=(20,10));
plt.ylabel('Number of applications');
```



The above shows that there is almost no difference in the trend of number of student visa applications up until May which means the above predictions made for the number of visa lodged in 2020-22 are reliable. This data however represents the numbers of visas being lodged. Physically arriving in Australia is a bit more complicated due to all the travel restrictions and the response of various countries towards the pandemic. The australian government has closed their borders since 20th March and will likely stay the same until 2021 as stated by abc news (https://www.abc.net.au/news/2020-06-17/borders-likely-closed-until-next-year-coronavirus-restrictions/12365978). However, there have been reports that international students will be allowed back July onwards (https://www.abc.net.au/news/2020-06-12/morrison-international-students-back-in-july-amid-china-racism/12349422). To get a more accurate model of what is to come, more data as well as how vaarious contries further respond to the pandemic needs to be observed.

[153]:
```
yrs = ['2020','2021','2022']
for i in range(len(yrs)):
    print("Predicted number of visa that will be lodged in {} will be {}".
 ↪format(yrs[i],str(int(np.round(threeyr_students_normal[i])))))
    print("----------------------------------------------------------------------")
```

```
Predicted number of visa that will be lodged in 2020 will be 401231
----------------------------------------------------------------------
Predicted number of visa that will be lodged in 2021 will be 413857
----------------------------------------------------------------------
Predicted number of visa that will be lodged in 2022 will be 426484
----------------------------------------------------------------------
```