Introduction

(This part is taken from Project Overview document and contains the Part 1 section from it)

For the first part of the project, I have used the data from 2 different sources.

I have used education spendings data from OECD and I used World Bank for other data.

Then, I merged all the data and eliminated the data with missing entries.

In the end, I have had a dataset with 26 countries with following attributes:

- country codes
- country names
- years (year is not definite, you can find the details in the blog)
- primary to non-tertiary education spendings (% of GDP)
- public tertiary education spendings (% of GDP)
- private tertiary education spendings (% of GDP)
- PISA scores for reading, math and science
- R&D expenditures (% of GDP)
- researchers in R&D (per million people)
- high technology exports (% of manufactured exports)
- resident and nonresident patent numbers (per 100 people)
- GDP per capitas (/ $1000)
- GINI indexes

I have had the following hypothesises and I have used Linear Regression to test them:

- Primary to non-tertiary education spendings explain the variation in PISA scores.
- PISA scores and education spendings explain the variation in R&D expenditures.
- PISA scores and education spendings explain the variation in researchers in R&D.
- PISA scores, education spendings and R&D data explain the variation in high technology exports.
- PISA scores, education spendings and R&D data explain the variation in patent numbers.
- PISA scores, education spendings, R&D data, high technology exports, patent numbers explain the variation in GDP per capitas.
- PISA scores and education spendings explain the variation in GINI indexes.

Data Visualisations

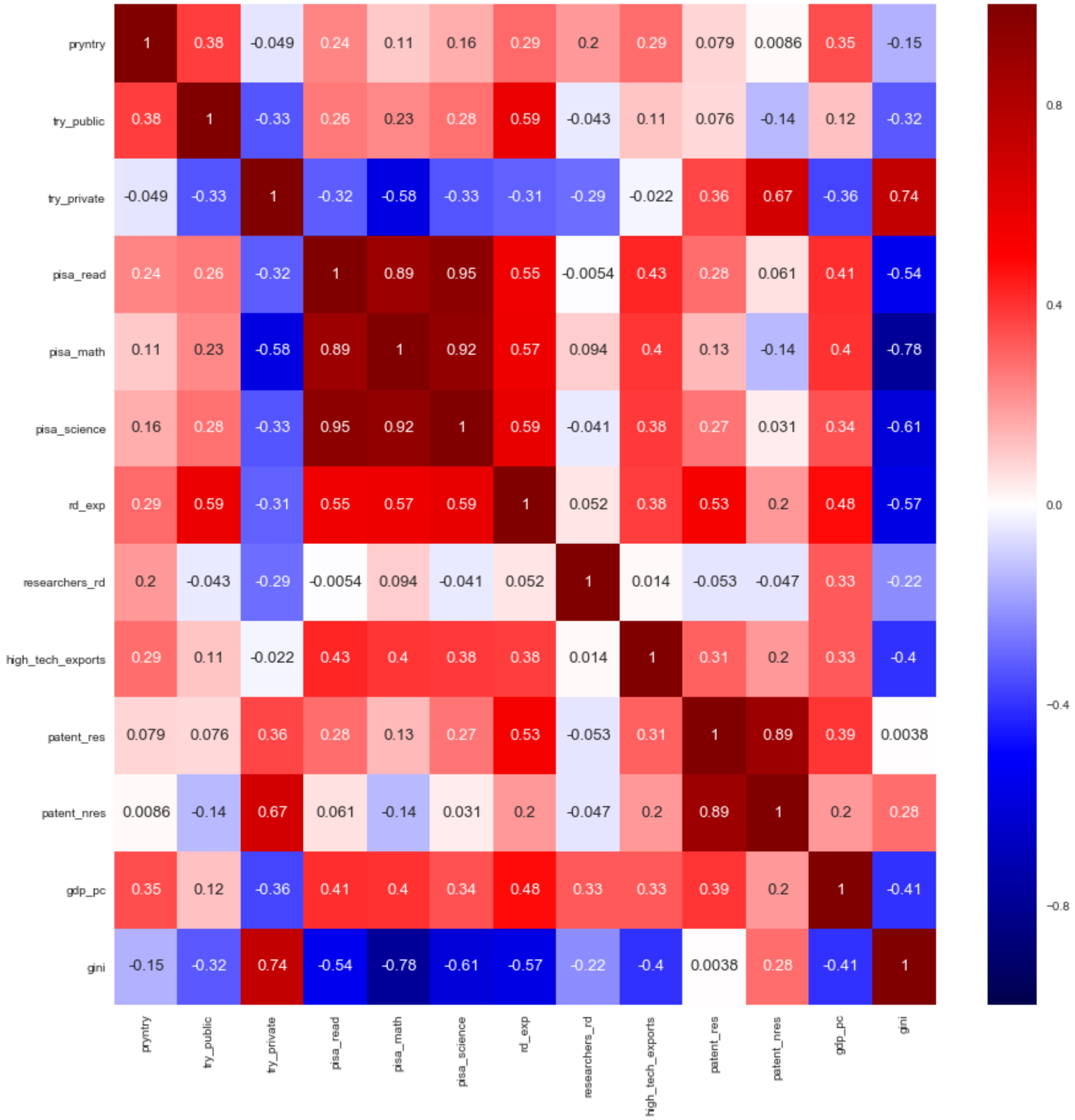
I have plotted histograms, correlation matrix and scatter plot matrix to understand the data.

In next page, you can find all these visualisations.

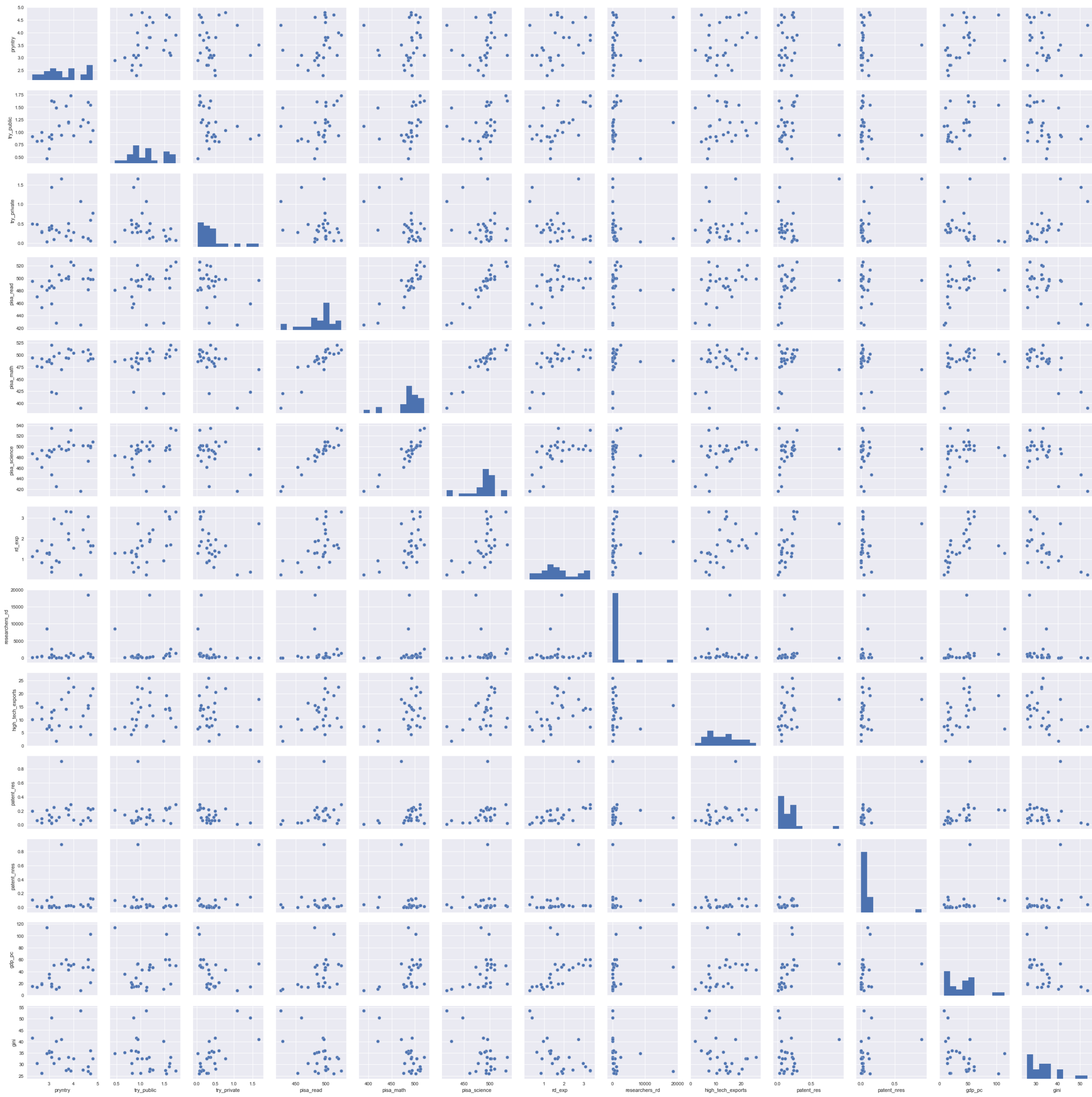
Histograms


Primary to Non-tertiary Education Spending (% of GDP)
2.42 2.68 2.92 3.18 3.42 3.68 3.92 4.18 4.42 4.68
Public Tertiary Education Spending (% of GDP)
0.53 0.66 0.78 0.91 1.04 1.16 1.29 1.42 1.54 1.67
Private Tertiary Education Spending (% of GDP)
0.11 0.27 0.44 0.6 0.76 0.93 1.09 1.25 1.42 1.58
PISA (Read)
430.0 440.2 450.2 460.4 470.4 480.6 490.6 500.8 510.8 521.0
PISA (Math)
396.5 409.5 422.5 435.5 448.5 461.5 474.5 487.5 500.5 513.5
PISA (Science)
421.9 433.7 445.5 457.3 469.1 480.9 492.7 504.5 516.3 528.1
R&D Expenditure (% of GDP)
0.41 0.72 1.02 1.33 1.63 1.94 2.24 2.55 2.85 3.16
Researchers in R&D (per million people)
928.6 2779.3 4630.0 6480.8 8331.5 10182.2 12033.0 13883.7 15734.4 17585.2
High Technology Exports (% of all manufactured exports)
3.08 5.48 7.88 10.29 12.69 15.09 17.49 19.9 22.3 24.7
Resident Patent Numbers (per 100 people)
0.05 0.14 0.24 0.32 0.42 0.5 0.6 0.68 0.77 0.86
Non-resident Patent Numbers (per 100 people)
0.04 0.14 0.22 0.32 0.4 0.5 0.58 0.68 0.76 0.86
GDP per capita (divided by 1000)
13.32 23.88 34.46 45.03 55.6 66.17 76.74 87.3 97.88 108.44
GINI index
27.28 30.04 32.8 35.56 38.32 41.07 43.83 46.59 49.35 52.11

Correlation Matrix



Scatter Plot Matrix



Hypothesises and Hypothesis Testings

Primary to non-tertiary education spendings explain the variation in PISA scores.

Linear Regressions:

Dep. Variable:	pisa_read	**R-squared:**	0.056
Model:	OLS	**Adj. R-squared:**	0.017
Method:	Least Squares	**F-statistic:**	1.431
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.243
Time:	21:38:13	**Log-Likelihood:**	-119.33
No. Observations:	26	**AIC:**	242.7
Df Residuals:	24	**BIC:**	245.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	460.6766	24.136	19.087	0.000	410.862 510.491
pryntry	7.8991	6.602	1.196	0.243	-5.727 21.526

Dep. Variable:	pisa_math	**R-squared:**	0.012
Model:	OLS	**Adj. R-squared:**	-0.030
Method:	Least Squares	**F-statistic:**	0.2818
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.600
Time:	21:38:15	**Log-Likelihood:**	-124.86
No. Observations:	26	**AIC:**	253.7
Df Residuals:	24	**BIC:**	256.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	469.5509	29.862	15.724	0.000	407.919 531.183
pryntry	4.3359	8.169	0.531	0.600	-12.523 21.195

Dep. Variable:	pisa_science	**R-squared:**	0.027
Model:	OLS	**Adj. R-squared:**	-0.014
Method:	Least Squares	**F-statistic:**	0.6572
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.426
Time:	21:38:16	**Log-Likelihood:**	-121.65
No. Observations:	26	**AIC:**	247.3
Df Residuals:	24	**BIC:**	249.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	467.3097	26.395	17.704	0.000	412.833 521.787
pryntry	5.8534	7.220	0.811	0.426	-9.049 20.755

Interpretations:

As seen from the results, p-values for these linear regressions are 0.243, 0.6, 0.426 consecutively. They all are greater than 0.05 which is the max p-value for 95% confidence.

According to these values, we can say that primary to non-tertiary education spendings DO NOT EXPLAIN the variation in PISA scores.

My Comments:

These results were a bit shocking for me and they made me think about the reason of it. I have come up with an explanation such as education spendings are not enough to explain the variation since there is another factor, efficiency, is missing.

Please note that this is only an idea to explain the results and it may be right or wrong.

PISA scores and education spendings explain the variation in R&D expenditures.

Linear Regressions:

Dep. Variable:	rd_exp	**R-squared:**	0.571
Model:	OLS	**Adj. R-squared:**	0.435
Method:	Least Squares	**F-statistic:**	4.213
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.00733
Time:	21:38:17	**Log-Likelihood:**	-21.668
No. Observations:	26	**AIC:**	57.34
Df Residuals:	19	**BIC:**	66.14
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	-7.5145	2.832	-2.653	0.016	-13.442 -1.587
pisa_read	-0.0073	0.018	-0.404	0.690	-0.045 0.030
pisa_math	0.0193	0.019	1.037	0.313	-0.020 0.058
pisa_science	0.0029	0.022	0.128	0.900	-0.044 0.050
pryntry	0.0909	0.198	0.459	0.651	-0.324 0.505
try_public	1.3035	0.494	2.638	0.016	0.269 2.338
try_private	0.4534	0.588	0.772	0.450	-0.776 1.683

Dep. Variable:	rd_exp	**R-squared:**	0.543
Model:	OLS	**Adj. R-squared:**	0.503
Method:	Least Squares	**F-statistic:**	13.66
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.000123
Time:	21:38:17	**Log-Likelihood:**	-22.486
No. Observations:	26	**AIC:**	50.97
Df Residuals:	23	**BIC:**	54.75
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	-6.1143	1.965	-3.112	0.005	-10.178 -2.050
pisa_math	0.0132	0.004	3.177	0.004	0.005 0.022
try_public	1.2484	0.377	3.313	0.003	0.469 2.028

```
(0.30867291748023784, 0.86857973234158825, 4.0)
```

Interpretations:

As seen from the results, p-value of the unrestricted model is 0.00733 which means that PISA scores and education spendings DO EXPLAIN the variation in R&D expenditures.

However, we can see from individual p-values that some of them are not necessary.

Therefore, I have constructed a restricted model. The p-value from F-test is 0.86 which indicates that we CANNOT REJECT the restricted model.

According to the restricted model, PISA math scores and public tertiary education spendings DO EXPLAIN 54.3% of the variation in R&D expenditures.

My comments:

These results were as expected for me. It makes sense PISA math scores and public tertiary education explain the variation in R&D expenditures.

PISA scores and education spendings explain the variation in researchers in R&D.

Linear Regression:

Dep. Variable:	researchers_rd_permillionpeople	**R-squared:**	0.204
Model:	OLS	**Adj. R-squared:**	-0.047
Method:	Least Squares	**F-statistic:**	0.8124
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.573
Time:	21:38:18	**Log-Likelihood:**	-248.08
No. Observations:	26	**AIC:**	510.2
Df Residuals:	19	**BIC:**	519.0
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	8610.5347	1.71e+04	0.502	0.621	-2.73e+04 4.45e+04
pisa_read	3.4926	109.020	0.032	0.975	-224.688 231.673
pisa_math	61.3639	112.971	0.543	0.593	-175.086 297.814
pisa_science	-83.1656	135.550	-0.614	0.547	-366.876 200.544
pryntry	1511.3633	1198.494	1.261	0.223	-997.114 4019.841
try_public	-2161.6652	2990.912	-0.723	0.479	-8421.717 4098.387
try_private	-2272.9246	3556.394	-0.639	0.530	-9716.543 5170.694

Interpretations:

As seen from the results, p-value of the regression is 0.573. It is greater than 0.05 which is the max p-value for 95% confidence level.

According to these values, we can say that education spendings and PISA scores jointly DO NOT EXPLAIN the variation in researchers in R&D.

My comments:

This result was a bit surprising for me and it made me think about the reason of it. However, I couldn't find any compelling explanation for such results.

PISA scores, education spendings and R&D data explain the variation in high tech exports.

Linear Regression:

Dep. Variable:	high_tech_exports	**R-squared:**	0.356
Model:	OLS	**Adj. R-squared:**	0.054
Method:	Least Squares	**F-statistic:**	1.177
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.367
Time:	21:38:20	**Log-Likelihood:**	-77.726
No. Observations:	26	**AIC:**	173.5
Df Residuals:	17	**BIC:**	184.8
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	-42.4479	30.618	-1.386	0.184	-107.047 22.151
pisa_read	0.0810	0.165	0.490	0.630	-0.268 0.430
pisa_math	0.2583	0.176	1.467	0.161	-0.113 0.630
pisa_science	-0.2502	0.207	-1.210	0.243	-0.686 0.186
pryntry	1.8183	1.887	0.963	0.349	-2.164 5.801
try_public	0.0212	5.373	0.004	0.997	-11.314 11.357
try_private	7.8249	5.520	1.417	0.174	-3.822 19.472
rd_exp	1.5125	2.105	0.718	0.482	-2.929 5.954
researchers_rd_permillionpeople	-8.8e-05	0.000	-0.253	0.803	-0.001 0.001

Interpretations:

As seen from the results, p-value of the regression is 0.367. It is greater than 0.05 which is the max p-value for 95% confidence level.

According to these values, we can say that education spendings, PISA scores and R&D data jointly DO NOT EXPLAIN the variation in high technology exports.

My comments:

This result was a bit shocking for me and it made me think about the reason of it. However, I couldn’t find any compelling explanation for such results.

PISA scores, education spendings and R&D data explain the variation in patent numbers.

Linear Regressions:

Dep. Variable:	patent_res_per100people	**R-squared:**	0.655
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	4.042
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.00743
Time:	21:38:21	**Log-Likelihood:**	22.799
No. Observations:	26	**AIC:**	-27.60
Df Residuals:	17	**BIC:**	-16.27
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-0.3862	0.641	-0.603	0.555	-1.739 0.966
pisa_read	0.0031	0.003	0.883	0.389	-0.004 0.010
pisa_math	0.0006	0.004	0.154	0.879	-0.007 0.008
pisa_science	-0.0029	0.004	-0.674	0.509	-0.012 0.006
pryntry	-0.0304	0.040	-0.769	0.453	-0.114 0.053
try_public	-0.0804	0.112	-0.715	0.485	-0.318 0.157
try_private	0.2711	0.116	2.346	0.031	0.027 0.515
rd_exp	0.1654	0.044	3.753	0.002	0.072 0.258
researchers_rd_permillionpeople	3.576e-06	7.28e-06	0.491	0.630	-1.18e-05 1.89e-05

Dep. Variable:	patent_res_per100people	**R-squared:**	0.593
Model:	OLS	**Adj. R-squared:**	0.558
Method:	Least Squares	**F-statistic:**	16.77
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	3.22e-05
Time:	21:38:22	**Log-Likelihood:**	20.641
No. Observations:	26	**AIC:**	-35.28
Df Residuals:	23	**BIC:**	-31.51
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-0.1913	0.065	-2.944	0.007	-0.326 -0.057
try_private	0.2518	0.060	4.184	0.000	0.127 0.376
rd_exp	0.1441	0.028	5.104	0.000	0.086 0.202

```
(0.51157780379266493, 0.79140691116826367, 6.0)
```

Dep. Variable:	patent_nres_per100people	**R-squared:**	0.715
Model:	OLS	**Adj. R-squared:**	0.581
Method:	Least Squares	**F-statistic:**	5.329
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.00185
Time:	21:38:23	**Log-Likelihood:**	25.242
No. Observations:	26	**AIC:**	-32.48
Df Residuals:	17	**BIC:**	-21.16
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-0.5675	0.583	-0.973	0.344	-1.799 0.664
pisa_read	0.0029	0.003	0.905	0.378	-0.004 0.009
pisa_math	0.0019	0.003	0.571	0.576	-0.005 0.009
pisa_science	-0.0039	0.004	-0.983	0.339	-0.012 0.004
pryntry	-0.0317	0.036	-0.881	0.390	-0.108 0.044
try_public	-0.0414	0.102	-0.404	0.691	-0.257 0.175
try_private	0.4178	0.105	3.972	0.001	0.196 0.640
rd_exp	0.1045	0.040	2.603	0.019	0.020 0.189
researchers_rd_permillionpeople	7.923e-06	6.63e-06	1.195	0.248	-6.06e-06 2.19e-05

Dep. Variable:	patent_nres_per100people	**R-squared:**	0.632
Model:	OLS	**Adj. R-squared:**	0.600
Method:	Least Squares	**F-statistic:**	19.79
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	1.00e-05
Time:	21:38:24	**Log-Likelihood:**	21.938
No. Observations:	26	**AIC:**	-37.88
Df Residuals:	23	**BIC:**	-34.10
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	-0.2328	0.062	-3.766	0.001	-0.361 -0.105
try_private	0.3485	0.057	6.087	0.000	0.230 0.467
rd_exp	0.0912	0.027	3.397	0.002	0.036 0.147

```
(0.81993054578322533, 0.56964004401047452, 6.0)
```

Interpretations:

As seen from the results, p-values of the regressions are less than 0.05 in both unrestricted regressions. However, there are some unnecessary variables in both regressions. Therefore, I have modelled restricted regressions and tested them with F-test.

As a result I have come up with models saying that private tertiary education spendings and R&D expenditures DO EXPLAIN 59.3% and 63.2% of the variation in patent numbers consecutively for residents and non-residents.

My comments: These results were as expected for me.

PISA scores, education spendings, R&D data, high technology exports, patent numbers explain the variation in GDP per capitas.

Linear Regressions:

Dep. Variable:	gdp_percapita_dividedby1000	**R-squared:**	0.677
Model:	OLS	**Adj. R-squared:**	0.423
Method:	Least Squares	**F-statistic:**	2.663
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.0438
Time:	21:38:25	**Log-Likelihood:**	-107.13
No. Observations:	26	**AIC:**	238.3
Df Residuals:	14	**BIC:**	253.4
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	93.0838	115.171	0.808	0.432	-153.934 340.101
pisa_read	0.5019	0.582	0.863	0.403	-0.746 1.749
pisa_math	-0.8542	0.662	-1.290	0.218	-2.275 0.566
pisa_science	0.2549	0.764	0.334	0.744	-1.384 1.894
pryntry	9.8251	6.756	1.454	0.168	-4.666 24.316
try_public	-30.7790	18.813	-1.636	0.124	-71.129 9.571
try_private	-83.3863	30.891	-2.699	0.017	-149.641 -17.132
rd_exp	7.4613	10.330	0.722	0.482	-14.695 29.618
researchers_rd_permillionpeople	0.0001	0.001	0.080	0.937	-0.003 0.003
high_tech_exports	0.6059	0.834	0.727	0.479	-1.183 2.394
patent_res_per100people	-32.7776	96.553	-0.339	0.739	-239.864 174.308
patent_nres_per100people	143.1925	105.997	1.351	0.198	-84.149 370.534

Dep. Variable:	gdp_percapita_dividedby1000	**R-squared:**	0.577
Model:	OLS	**Adj. R-squared:**	0.519
Method:	Least Squares	**F-statistic:**	9.983
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	0.000238
Time:	21:38:29	**Log-Likelihood:**	-110.64
No. Observations:	26	**AIC:**	229.3
Df Residuals:	22	**BIC:**	234.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	16.6791	18.833	0.886	0.385	-22.377 55.736
pryntry	10.6455	4.950	2.150	0.043	0.379 20.912
try_private	-57.1964	12.308	-4.647	0.000	-82.721 -31.672
patent_nres_per100people	119.3021	28.553	4.178	0.000	60.088 178.516

Interpretations:

As seen from the results, the unrestricted model has a p-value 0.0438 which is OK for 95% confidence level. However, there were some unnecessary variables for the model, therefore I have computed 4 restricted models and chosen the 3rd one. You can see all the models and F-tests from the code except the resulting regressions.

According to the restricted model, primary to non-tertiary education spendings, private tertiary education spendings and non-resident patent numbers DO EXPLAIN 57.7% of the variation in GDP per capitas.

My comments:

The results were as expected for me.

But, I need to emphasise something about the results. Please note that the coefficient on try_private is negative and we only take non-resident patents in account. This makes me think about brain drain (human capital flight). I think with these results, we can say that the countries getting more human capital from outside has more GDP per capitas.

Please note that these are my ideas and may be right or wrong.

PISA scores and education spendings explain the variation in GINI indexes.

Linear Regressions:

Dep. Variable:	gini_index	**R-squared:**	0.795
Model:	OLS	**Adj. R-squared:**	0.730
Method:	Least Squares	**F-statistic:**	12.27
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	1.17e-05
Time:	21:38:31	**Log-Likelihood:**	-66.836
No. Observations:	26	**AIC:**	147.7
Df Residuals:	19	**BIC:**	156.5
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	86.0094	16.092	5.345	0.000	52.329 119.690
pisa_read	0.1804	0.102	1.763	0.094	-0.034 0.395
pisa_math	-0.2854	0.106	-2.691	0.014	-0.507 -0.063
pisa_science	0.0062	0.127	0.049	0.962	-0.260 0.273
pryntry	-1.2244	1.125	-1.088	0.290	-3.579 1.130
try_public	-2.1268	2.808	-0.757	0.458	-8.004 3.750
try_private	3.6133	3.339	1.082	0.293	-3.374 10.601

Dep. Variable:	gini_index	**R-squared:**	0.725
Model:	OLS	**Adj. R-squared:**	0.701
Method:	Least Squares	**F-statistic:**	30.31
Date:	Fri, 14 Apr 2017	**Prob (F-statistic):**	3.57e-07
Time:	21:38:32	**Log-Likelihood:**	-70.649
No. Observations:	26	**AIC:**	147.3
Df Residuals:	23	**BIC:**	151.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	95.5228	15.417	6.196	0.000	63.630 127.415
pisa_read	0.2079	0.068	3.063	0.006	0.067 0.348
pisa_math	-0.3373	0.056	-6.006	0.000	-0.453 -0.221

```
(1.6188268371641834, 0.21056109536543721, 4.0)
```

Interpretations:

As seen from the results, the unrestricted model has a very low p-value which is OK for 95% confidence level. However, there were some unnecessary variables for the model, therefore I have computed a restricted model and performed F-test for that model.

According to the restricted model, PISA read and math scores DO EXPLAIN 72.5% of the variation in GINI indexes.

My comments:

The results were as expected for me, however 72.5% is a bit more than my expectations.

Berkan Teber
19080