



# Scientific Output

An Investigation



# Our Goals

- We aim to study the various institutional and economic factors which may affect scientific output.
- Produce a top 10 ranking of countries based on their scientific output.

# What is Scientific Output?

- Scientific output can be generally defined as the amount of science that a given individual, institution, or country produces.
- It can be measured quantitatively or qualitatively.
- We used the number of documents per million people as our quantitative measure, with the number of citations per document as our qualitative measure.

# Institutional & Economic Factors

- Our institutional factors include the number of researchers per million, and the education index of a given country.
- Our economic factors include the GDP per capita, and the research and development expenditure.

# Data & Cleaning (1/2)

- We had two main source for our data, the World Bank and SCImago Institutions Rankings.
- The World Bank provided us with the majority of the data we needed, including the GDP, population, researchers per million, and research and development expenditure.
- SCImago in the other hand provided us with the documents and citations data that we needed, including citations per document.

```
19 library(WDI)
20 library(tidyverse)
21 library(dplyr)
22 library(readxl)
23 library(cowplot)
24
25 #-----
26 #           GET THE DATA
27 #-----
28
29
30 #population of each country from 2012 to 2018
31 population = WDI(indicator='SP.POP.TOTL', start=2012, end=2018)
32
33 #GDP of each country from 2012 to 2018
34 gdp = WDI(indicator='NY.GDP.MKTP.CD', start=2012, end=2018)
35
36 #research and development expenditure (% of GDP) from 2012 to 2018
37 RnD_expenditure = WDI(indicator='GB.XPD.RSDV.GD.ZS', start=2012, end=2018)
38
39 #scientific and technical journal articles from 2012 to 2018
40 sci_and_tech_journals = WDI(indicator = 'IP.JRN.ARTC.SC', start=2012, end=2018)
41
42 #researchers per million from 2012 to 2018
43 researchers_million = WDI(indicator = 'SP.POP.SCIE.RD.P6', start =2012, end =2018)
44
```

# Data & Cleaning (2/2)

- When we first downloaded the World Bank and SCImago datasets, there were a lot of countries that were missing data, either for all of the years we chose to study, or a select few.

	symbol	country	Population Population, total	year	gdp GDP (current US\$)	RnD_Expenditure Research and development expenditure (% of GDP)	sci_tech_articles Scientific and technical journal articles	researchers_per_million Researchers in R&D (per million people)
129	BT	Bhutan	736709	2016	2.158972e+09	NA	42.02	NA
130	BT	Bhutan	727876	2015	2.003598e+09	NA	25.18	NA
131	BT	Bhutan	719056	2014	1.907091e+09	NA	28.36	NA
132	BT	Bhutan	710238	2013	1.756216e+09	NA	28.80	NA
133	BT	Bhutan	701583	2012	1.781281e+09	NA	17.22	NA
134	BA	Bosnia and Herzegovina	3323929	2018	2.018351e+10	0.19880	703.79	471.25010
135	BA	Bosnia and Herzegovina	3351527	2017	1.808012e+10	0.20048	724.80	485.42082
136	BA	Bosnia and Herzegovina	3386267	2016	1.691333e+10	0.21594	568.38	419.99063
137	BA	Bosnia and Herzegovina	3429361	2015	1.621154e+10	0.21860	568.22	365.22839
138	BA	Bosnia and Herzegovina	3482104	2014	1.855834e+10	0.25725	482.98	292.38061
139	BA	Bosnia and Herzegovina	3542605	2013	1.817850e+10	0.32132	486.00	233.86738
140	BA	Bosnia and Herzegovina	3604972	2012	1.722685e+10	0.26533	540.99	160.13994
141	BW	Botswana	2254126	2018	1.866327e+10	NA	280.57	NA
142	BW	Botswana	2205128	2017	1.740559e+10	NA	234.51	NA
143	BW	Botswana	2159944	2016	1.564635e+10	NA	252.46	NA
144	BW	Botswana	2120716	2015	1.442059e+10	NA	160.83	NA
145	BW	Botswana	2088614	2014	1.625078e+10	NA	207.35	NA
146	BW	Botswana	2062536	2013	1.490175e+10	0.53728	171.81	185.20781
147	BW	Botswana	2039551	2012	1.438000e+10	0.25880	157.49	172.43949
148	BR	Brazil	209469333	2018	1.885483e+12	NA	60147.96	NA
149	BR	Brazil	207833831	2017	2.062831e+12	1.26326	58022.38	NA
150	BR	Brazil	206163058	2016	1.795700e+12	1.26417	55181.31	NA
151	BR	Brazil	204471769	2015	1.802214e+12	1.34264	53064.56	NA
152	BR	Brazil	202763735	2014	2.455994e+12	1.27131	52632.41	887.67844
153	BR	Brazil	201035903	2013	2.472807e+12	1.19567	50497.90	838.47208
154	BR	Brazil	199287296	2012	2.465189e+12	1.12684	48282.01	788.48979
155	BN	Brunei Darussalam	428962	2018	1.356735e+10	0.27764	293.91	NA
156	BN	Brunei Darussalam	424473	2017	1.212810e+10	NA	214.24	NA



```
131 #-----  
132 #           clean the Super data set  
133 #-----  
134  
135  
136 #now remove all countries with a population less than 1 million  
137 FinalData <- subset(FinalData, Population > 1000000)  
138  
139 #create data frame of countries with NA across all the years and remove them  
140 remove <- read.delim("remove.txt")  
141 FinalData <- anti_join(FinalData, remove, by='country')  
142  
143  
144 #replacing all NA values with the sample mean of the column  
145 FinalData <- group_by(FinalData, country)  
146 FinalData <- mutate(FinalData, mean_rnd = mean(RnD_Expenditure, na.rm = TRUE),  
147                     mean_rpm = mean(researchers_per_million, na.rm = TRUE))  
148  
149 FinalData$RnD_Expenditure <- ifelse(is.na(FinalData$RnD_Expenditure),  
150                                   FinalData$mean_rnd, FinalData$RnD_Expenditure)  
151  
152 FinalData$researchers_per_million <- ifelse(is.na(FinalData$researchers_per_million),  
153                                             FinalData$mean_rpm, FinalData$researchers_per_million)  
154  
155 #removing columns we don't need  
156 FinalData <- FinalData[-c(1, 14, 16:17)]  
157
```

	country	Population	year	gdp	RnD_Expenditure	sci_tech_articles	researchers_per_million	Documents	Citable documents	Citations	Self-citations	Citations per document	education_value	doc_per_mil	gdp_per_capita	RnD_dollar_amount	RnD_per_capita
1	Algeria	42228429	2018	1.754057e+11	0.54243000	5231.44	819.34270	7799	7375	20188	4607	2.59	0.668	184.686009	4153.7340	951452924	2.253110e+01
2	Angola	30809762	2018	1.013532e+11	0.03229000	30.12	18.81599	118	108	249	45	2.11	0.500	3.829955	3289.6467	32726958	1.062227e+00
3	Argentina	44494502	2018	5.176267e+11	0.59494167	8811.13	1215.16384	15108	13797	45240	7830	2.99	0.842	339.547569	11633.4980	3079576919	6.921253e+01
4	Austria	8840521	2018	4.550949e+11	3.17177000	12362.28	5733.07565	27301	24049	125291	20799	4.59	0.866	3088.166410	51478.2853	14434562301	1.632773e+03
5	Bahrain	1569439	2018	3.765250e+10	0.10116000	321.51	368.99126	727	663	3130	218	4.31	0.765	463.222846	23991.0567	38089269	2.426935e+01
6	Belgium	11427054	2018	5.437344e+11	2.62119000	15688.13	5023.26301	35858	31665	169012	27321	4.71	0.900	3137.991647	47583.0749	15339779564	1.342409e+03
7	Bosnia and Herzegovina	3323929	2018	2.018351e+10	0.19880000	703.79	471.25010	1450	1278	4579	430	3.16	0.710	436.230738	6072.1846	40124819	1.207150e+01
8	Botswana	2254126	2018	1.866327e+10	0.39804000	280.57	178.82365	779	671	2469	352	3.17	0.666	345.588490	8279.6017	74287262	3.295613e+01
9	Brazil	209469333	2018	1.885483e+12	1.24396167	60147.96	838.21344	83839	77885	200861	61972	2.40	0.689	400.244746	9001.2342	23455057054	1.119737e+02
10	Bulgaria	7025037	2018	6.623016e+10	0.76803000	3311.27	2342.86988	5686	5345	19680	2987	3.46	0.779	809.390755	9427.7304	508667460	7.240780e+01
11	Burundi	11175378	2018	3.036932e+09	0.21028000	21.12	23.43546	88	82	476	29	5.41	0.412	7.874454	271.7520	6386060	5.714402e-01
12	Cambodia	16249798	2018	2.457175e+10	0.11823000	145.74	30.36704	489	440	1497	272	3.06	0.476	30.092682	1512.1267	29051184	1.787787e+00
13	Canada	37057765	2018	1.716263e+12	1.56625000	59967.79	4501.79747	113652	99272	445264	95612	3.92	0.893	3066.887601	46313.1714	26880963303	7.253800e+02
14	Chad	15477751	2018	1.123917e+10	0.30050000	15.45	57.89175	55	52	214	8	3.89	0.290	3.553488	726.1499	33773697	2.182080e+00
15	Chile	18729160	2018	2.982580e+11	0.37188833	7121.74	432.48076	14971	14005	54971	11571	3.67	0.805	799.341775	15924.7942	1109186777	5.922245e+01
16	China	1392730000	2018	1.389482e+13	2.18568000	528263.25	1307.12128	605616	588976	2161615	1480321	3.57	0.649	434.840924	9976.6771	303696248213	2.180582e+02
17	Colombia	49661056	2018	3.335689e+11	0.23699000	7195.02	72.51704	13282	12342	32786	6463	2.47	0.678	267.453032	6716.9116	790524991	1.591841e+01
18	Costa Rica	4999441	2018	6.055390e+10	0.49772833	507.41	443.64702	1213	1119	4675	444	3.85	0.724	242.627126	12112.1344	301393924	6.028552e+01
19	Croatia	4087843	2018	6.137522e+10	0.97489000	4276.90	1921.13136	7723	7208	24263	3975	3.14	0.803	1889.260424	15014.0850	598340905	1.463708e+02
20	Cyprus	1189265	2018	2.530982e+10	0.55859000	1245.42	1255.85256	3024	2697	13511	1898	4.47	0.827	2542.746991	21281.8994	141378113	1.188786e+02
21	Denmark	5793636	2018	3.568795e+11	3.06408000	13978.80	8065.88729	29888	26495	153733	26292	5.14	0.921	5158.763857	61598.5367	10935073377	1.887428e+03
22	Ecuador	17084357	2018	1.075620e+11	0.38497333	2142.19	343.77254	4565	4369	11738	2425	2.57	0.709	267.203501	6295.9354	414085048	2.423767e+01
23	El Salvador	6420744	2018	2.611740e+10	0.11023167	45.44	64.23833	145	140	221	16	1.52	0.553	22.583053	4067.6594	28789645	4.483849e+00
24	Estonia	1321977	2018	3.063114e+10	1.42515000	1414.72	3755.32912	3466	3126	19694	2730	5.68	0.882	2621.830788	23170.7074	436539723	3.302173e+02
25	Ethiopia	109224559	2018	8.426935e+10	0.43957500	1994.44	67.63007	3563	3351	11062	2278	3.10	0.335	32.620869	771.5238	370426988	3.391426e+00
26	Finland	5515525	2018	2.759466e+11	2.77381000	10598.94	6861.10974	21707	19407	97487	17334	4.49	0.927	3935.618096	50030.8773	7654233126	1.387761e+03
27	France	66965912	2018	2.787864e+12	2.20002000	66352.18	4715.31913	124315	111279	465791	107693	3.75	0.812	1856.392249	41631.0907	61333564668	9.158923e+02
28	Georgia	3726549	2018	1.759970e+10	0.30457000	550.41	1463.77339	1944	1739	12565	1335	6.46	0.851	521.662267	4722.7878	53603407	1.438419e+01
29	Germany	82905782	2018	3.963768e+12	3.09415000	104396.12	5211.87408	184756	165196	712171	214400	3.85	0.943	2228.505607	47810.5077	122644912913	1.479329e+03
30	Greece	10732682	2018	2.181384e+11	1.17732000	10906.99	3482.71652	20420	17547	62118	14780	4.02	0.833	1902.564474	20324.3050	2568186625	2.392821e+02
31	Guatemala	16346950	2018	7.311815e+10	0.03207000	99.89	20.47784	305	276	1115	96	3.66	0.511	18.657915	4472.8924	23448990	1.434457e+00
32	Honduras	9587522	2018	2.402419e+10	0.02748500	45.10	28.53300	188	167	607	53	3.23	0.499	19.608821	2505.7768	6603049	6.887127e-01

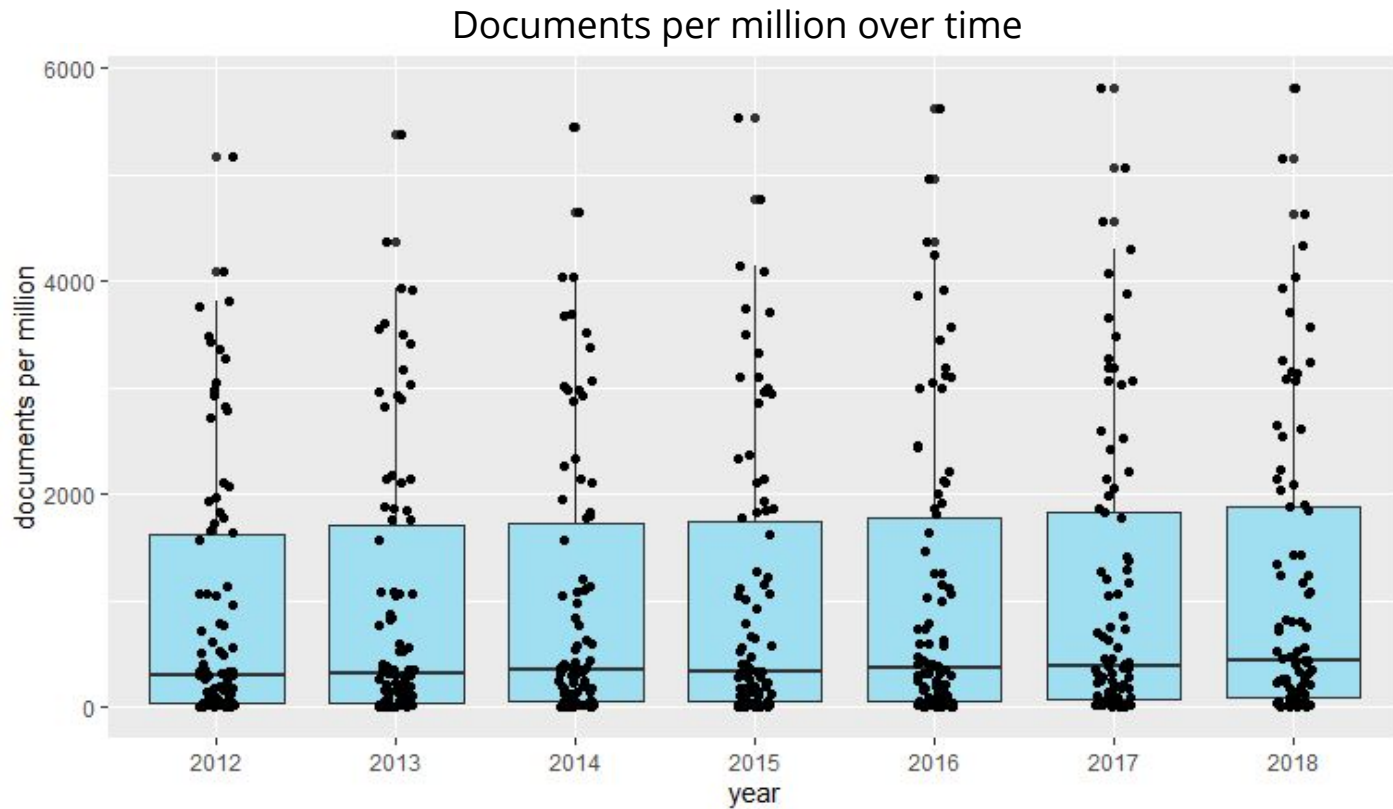
# Methods

- Our main methods were using data visualization techniques coupled with correlations to see the strength of our relationships.
- We also performed regression analysis (linear regression) and determined the statistical significance using confidence intervals.
- When needed, we filtered the data to make the relationship easier to study.

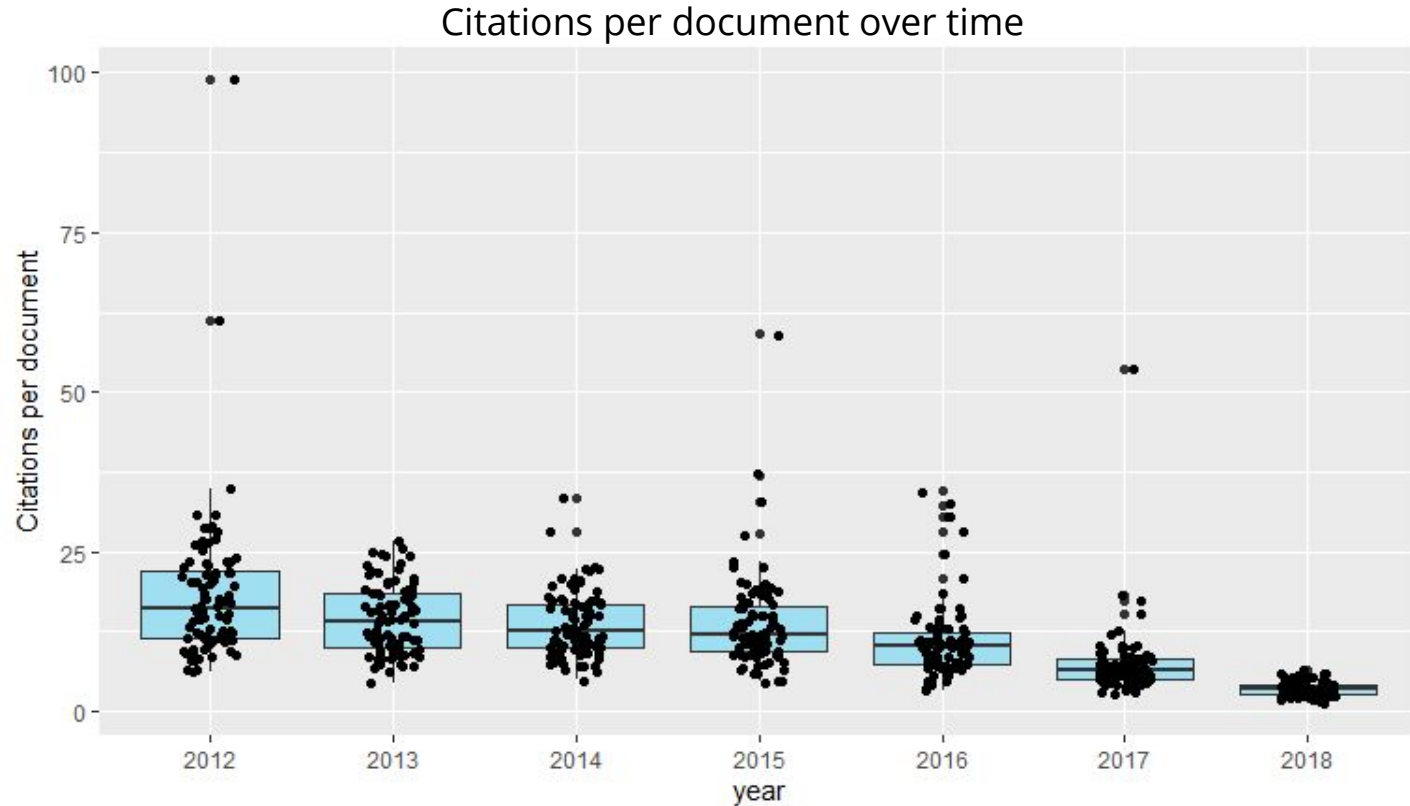
# Results

1. Scientific output
2. Institutional factors
3. Economic factors

# Scientific Output (1/4):

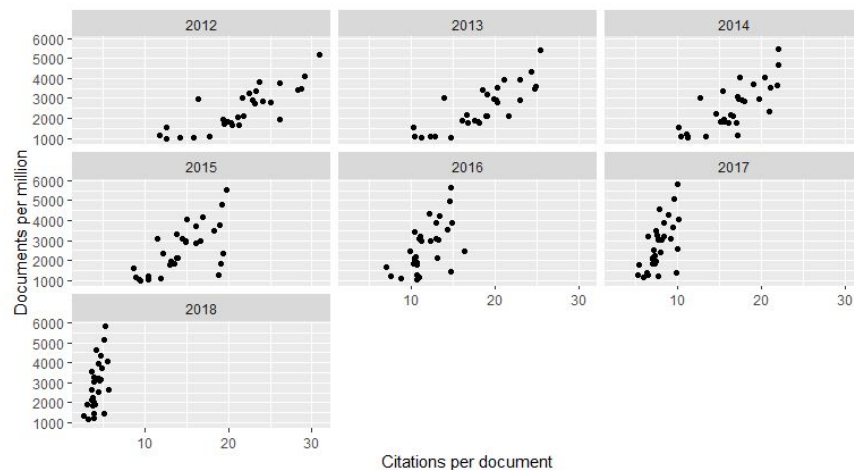
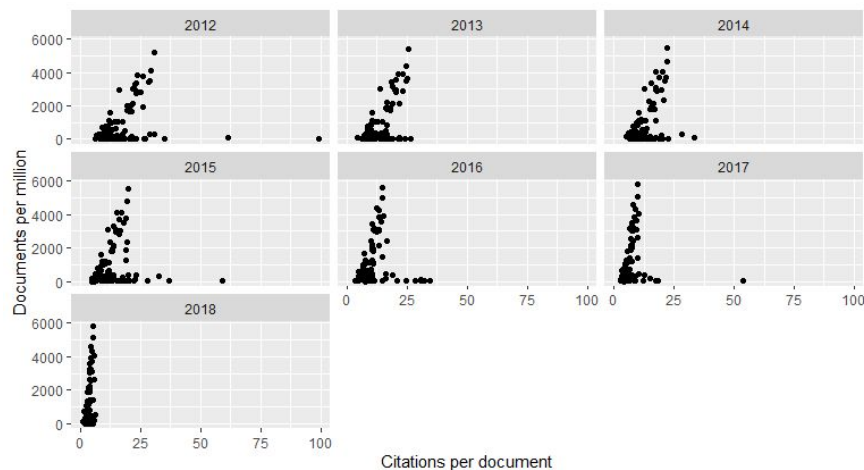


# Scientific Output (2/4):



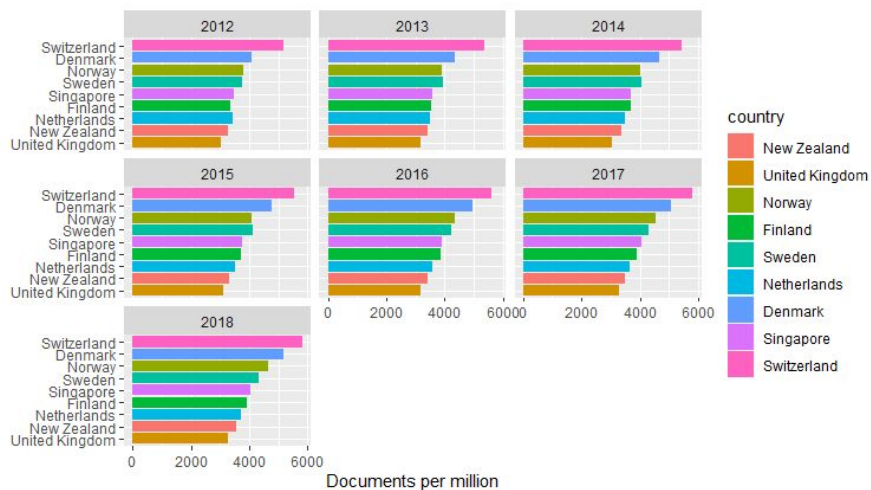
# Scientific Output (3/4):

Documents per Million vs Citations per Document

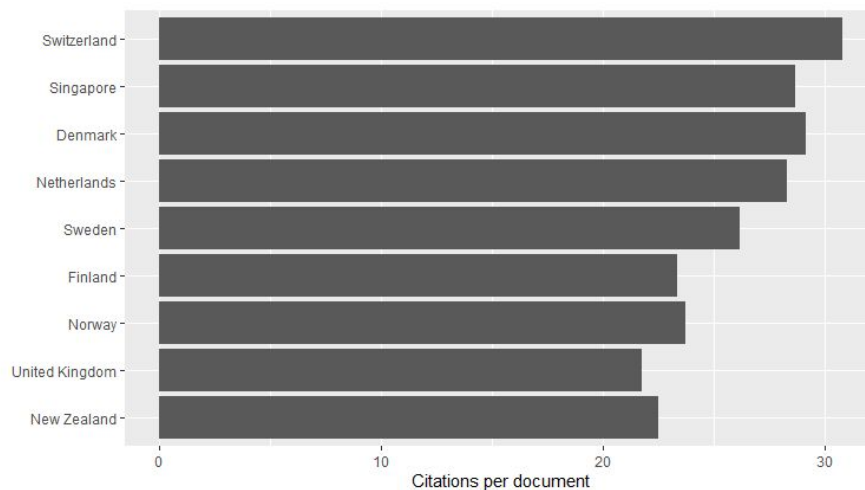


# Scientific Output (4/4):

Top 10 countries by Documents per million



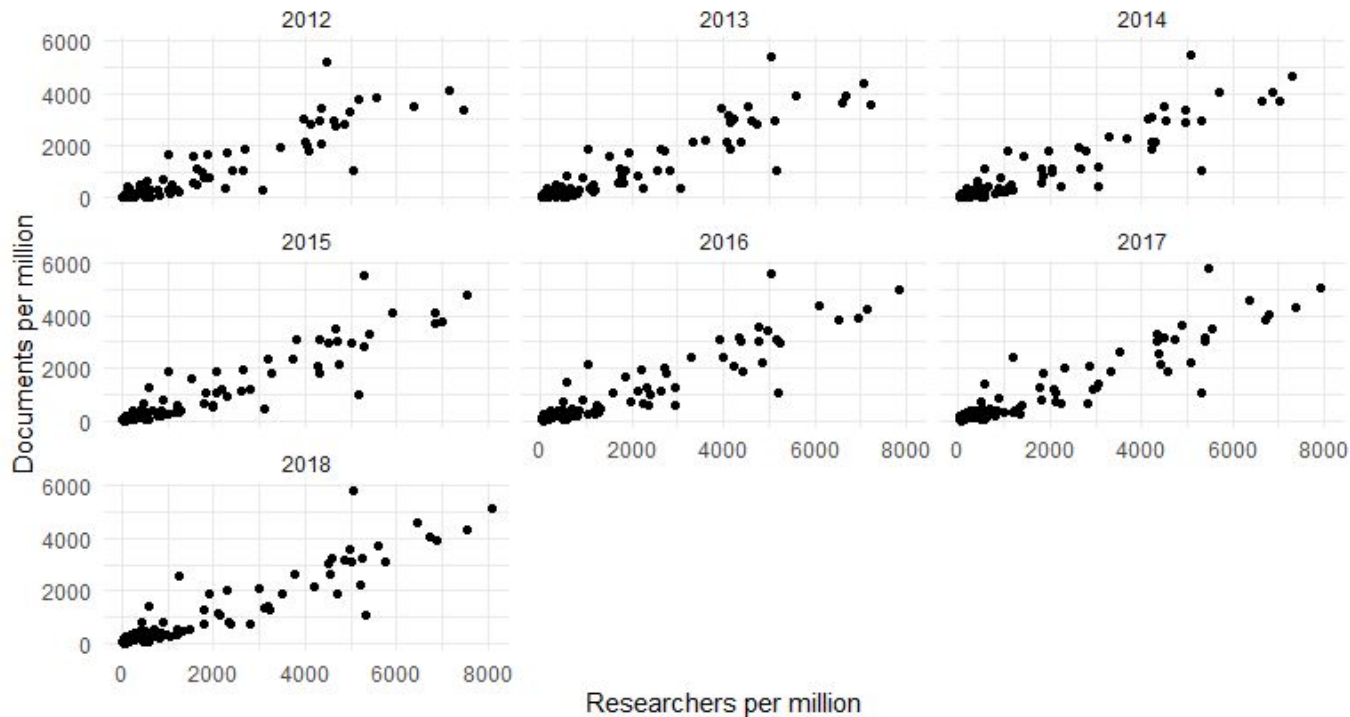
Top 10 countries by Citations per document





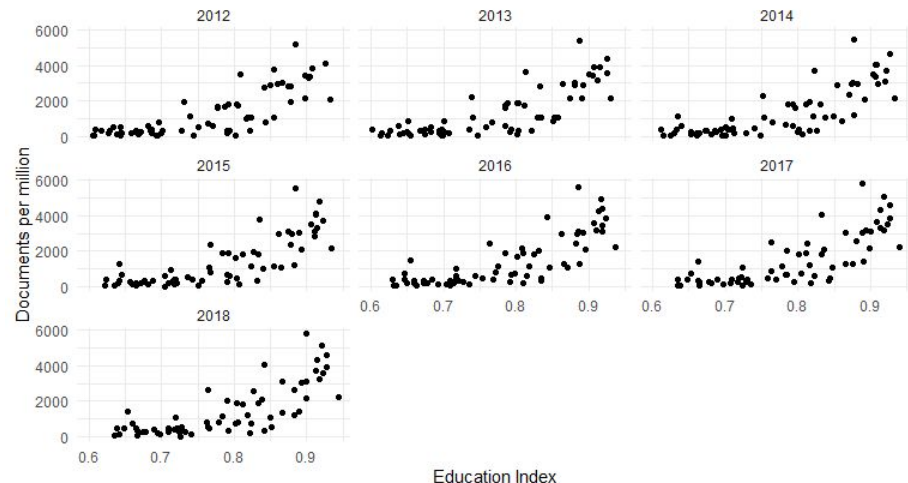
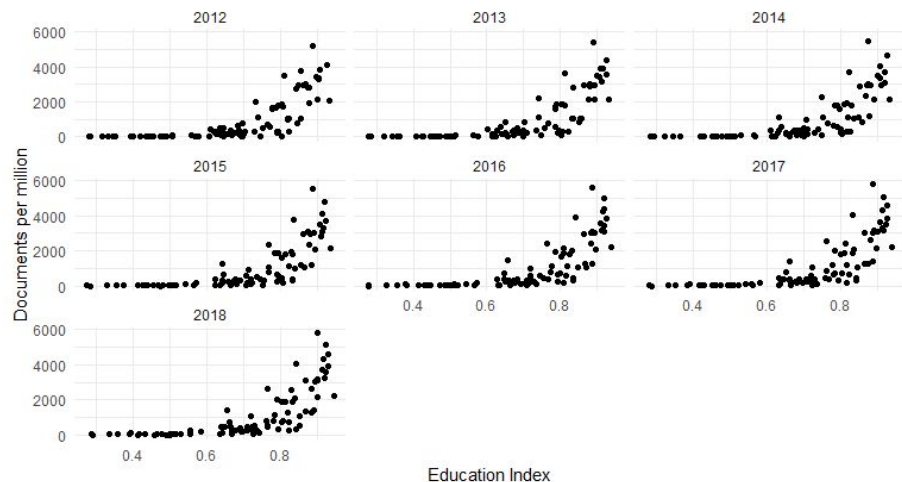
# Institutional (1/2): Researchers per million

Documents per million vs Researchers per million



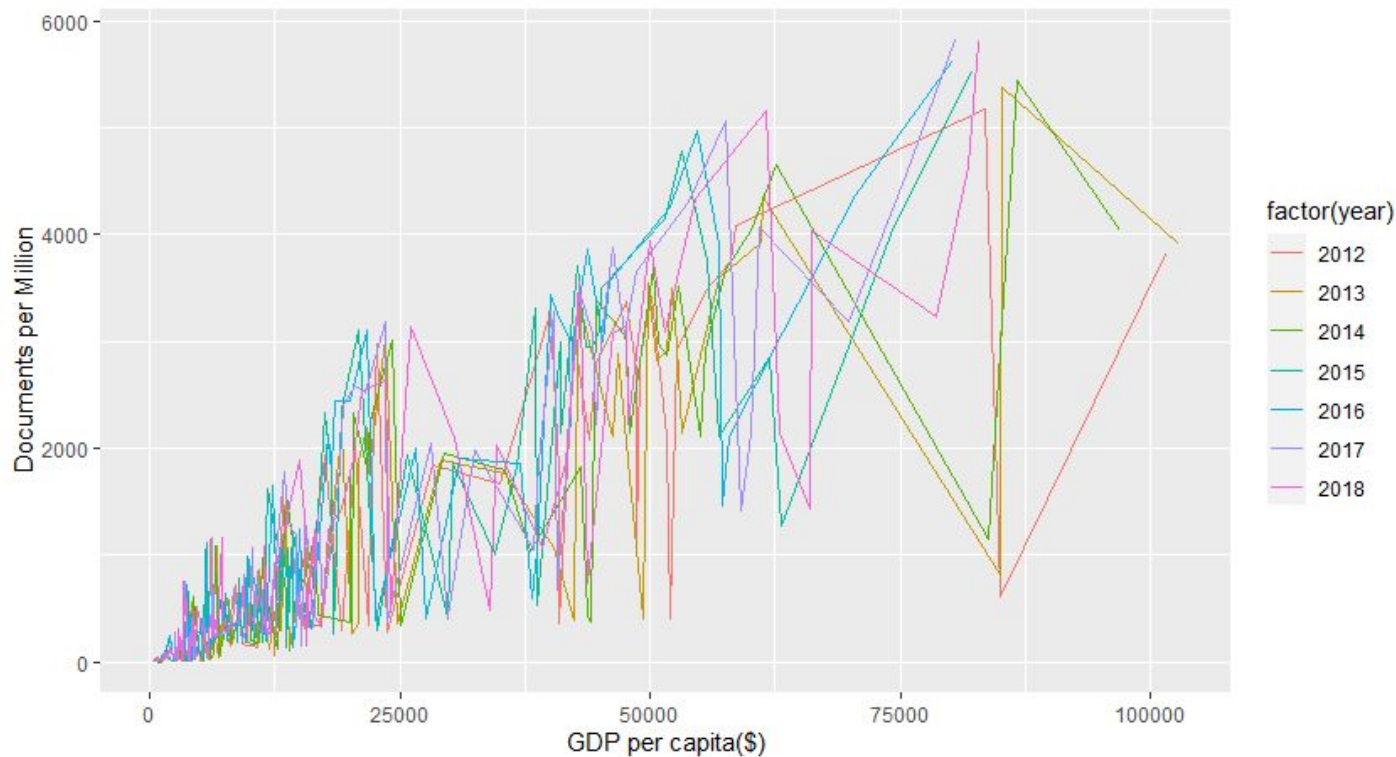
# Institutional (2/2): Education Index

Documents per million vs Education index



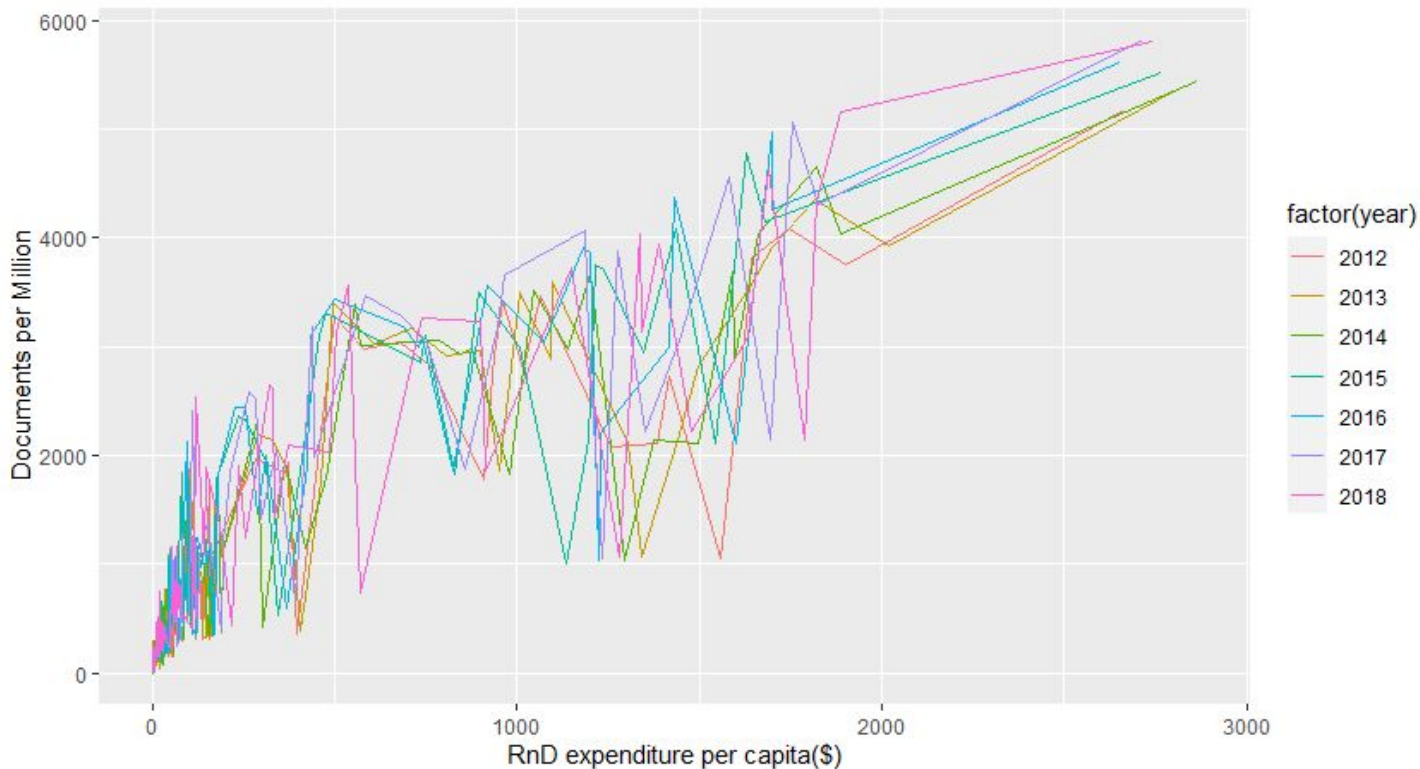
# Economic (1/2): GDP per capita

## Documents per million vs GDP per capita



## Economic (2/2): R&D Expenditure per capita

## Documents per million vs R&D expenditure per capita



# Limitations & Sources of Error

- Lack of available data for many countries
- Questionable datasets
- Sample mean

# Future Work

- We would like to explore the possibility of using fitted values instead of the sample mean to replace the missing data, and compare the results.
- We saw that the Nordic region was overrepresented in the top 10, we would like to perform the same analysis but on regions instead of countries.
- It would be interesting to see how countries perform in different areas of research.

# Discussion

- After concluding our analysis, there were not many surprises.
- Switzerland had the highest scientific output, and the Nordic region was significantly overrepresented in the top 10.
- Surprisingly, the number of documents per million has not increased

Thank you for listening!