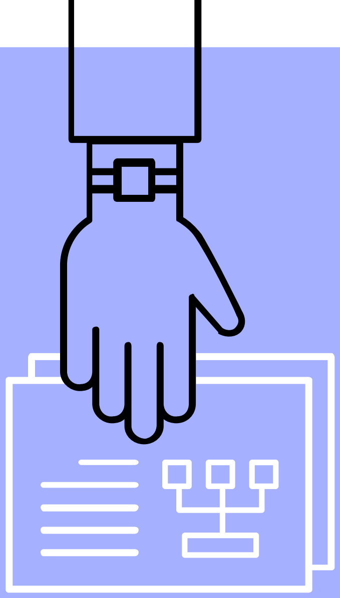
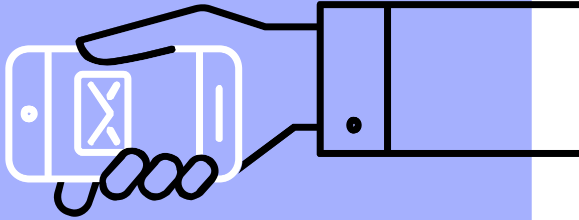
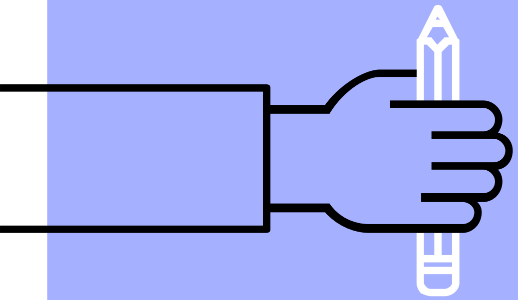
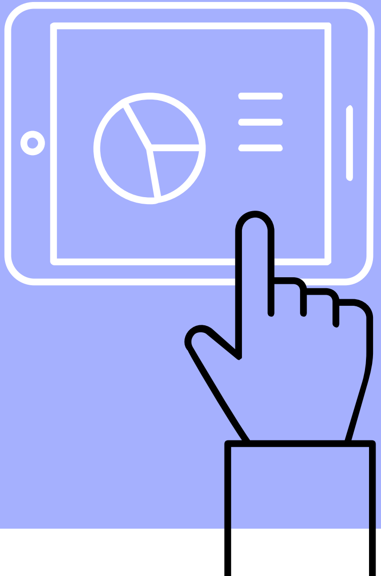




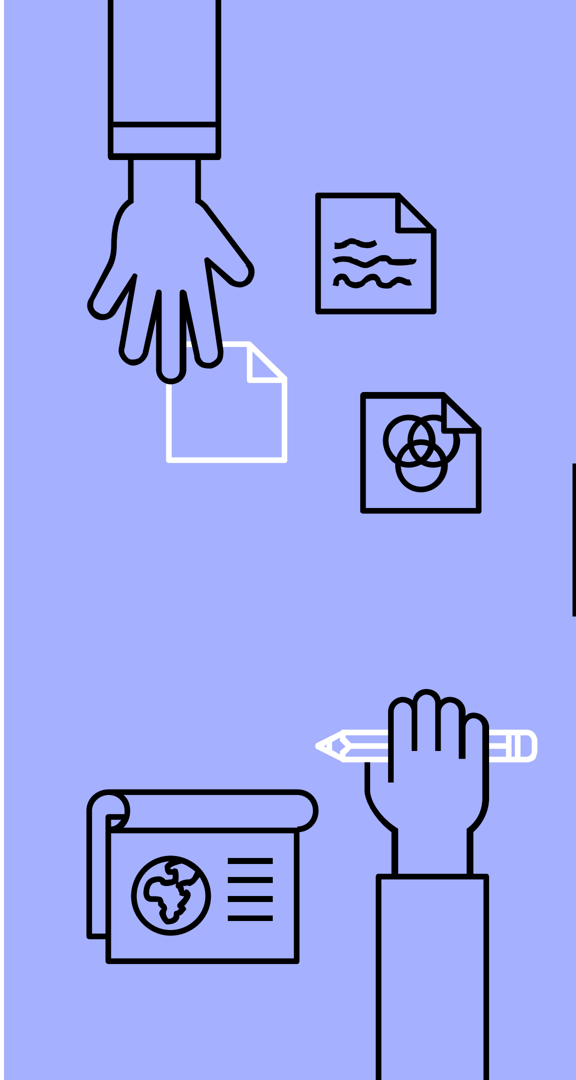
Cogs 108 project presentation video





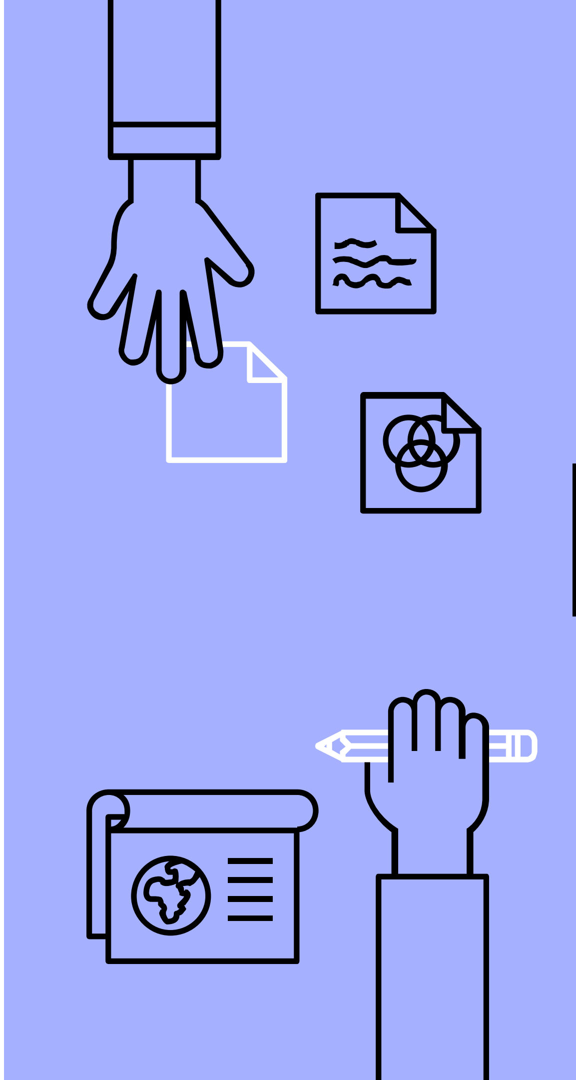
Overview

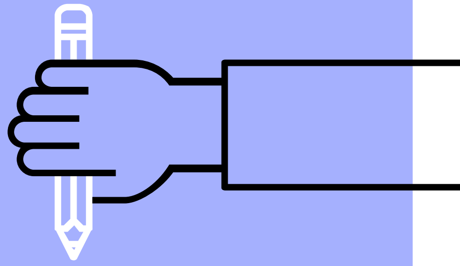
We all know that terrorism is bad since it uses people as a means to an end, which disrespect people’s lives and bodies. Because of that, our group wants to do some data science to try to predict terrorist attacks frequency. Thus, we want to ask: Is the frequency of terrorist attacks related to countries' GDP/location/life expectancy/CO2 emission? Do terrorist attacks happen more often toward some certain targets/countries/regions?



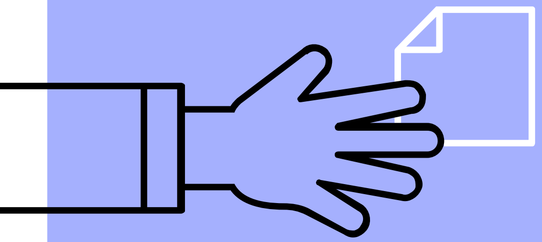
Hypothesis

We hypothesize that lower GDP and countries with high disease rate are more likely to encounter terrorist attacks. Also, we hypothesize that we can make a rough prediction of attacks' frequency using certain inputs.





1. Let’s see some data



Global Terrorism Database

Dataset 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	eventid	iyear	imonth	iday	approxdate	extended	resolution	country	country_txt	region	region_txt	provstate	city	latitude
2	1.97E+11	1970	7	2		0		58	Dominican R	2	Central America & Caribbe		Santo Domin	18.4567
3	1.97E+11	1970	0	0		0		130	Mexico	1	North Ameri	Federal	Mexico city	19.3718
4	1.97E+11	1970	1	0		0		160	Philippines	5	Southeast As	Tarlac	Unknown	15.4785
5	1.97E+11	1970	1	0		0		78	Greece	8	Western Eur	Attica	Athens	37.997
6	1.97E+11	1970	1	0		0		101	Japan	4	East Asia	Fukouka	Fukouka	33.5804
7	1.97E+11	1970	1	1		0		217	United State	1	North Ameri	Illinois	Cairo	37.0051
8	1.97E+11	1970	1	2		0		218	Uruguay	3	South Ameri	Montevideo	Montevideo	-34.8911
9	1.97E+11	1970	1	2		0		217	United State	1	North Ameri	California	Oakland	37.7919
10	1.97E+11	1970	1	2		0		217	United State	1	North Ameri	Wisconsin	Madison	43.0765
11	1.97E+11	1970	1	3		0		217	United State	1	North Ameri	Wisconsin	Madison	43.072
12	1.97E+11	1970	1	1		0		217	United State	1	North Ameri	Wisconsin	Baraboo	43.46
13	1.97E+11	1970	1	6		0		217	United State	1	North Ameri	Colorado	Denver	39.7589
14	1.97E+11	1970	1	8		0		98	Italy	8	Western Eur	Lazio	Rome	41.8909
15	1.97E+11	1970	1	9		0		217	United State	1	North Ameri	Michigan	Detroit	42.3316
16	1.97E+11	1970	1	9		0		217	United State	1	North Ameri	Puerto Rico	Rio Piedras	18.3869
17	1.97E+11	1970	1	10		0		499	East Germar	9	Eastern Euro	Berlin	Berlin	52.501
18	1.97E+11	1970	1	11		0		65	Ethiopia	11	Sub-Saharan	Unknown	Unknown	
19	1.97E+11	1970	1	12		0		217	United State	1	North Ameri	New York	New York Cit	40.6971
20	1.97E+11	1970	1	12		0		217	United State	1	North Ameri	Puerto Rico	Rio Grande	18.3799
21	1.97E+11	1970	1	13		0		217	United State	1	North Ameri	Washington	Seattle	47.6107
22	1.97E+11	1970	1	14		0		217	United State	1	North Ameri	Illinois	Champaign	40.1167
23	1.97E+11	1970	1	15		0		218	Uruguay	3	South Ameri	Montevideo	Montevideo	-34.8911
24	1.97E+11	1970	1	19		0		217	United State	1	North Ameri	Washington	Seattle	47.6107
25	1.97E+11	1970	1	19		0		217	United State	1	North Ameri	Washington	Seattle	47.6107
26	1.97E+11	1970	1	19	January 19-2	0		217	United State	1	North Ameri	New Jersey	Jersey City	40.7178
27	1.97E+11	1970	1	20		0		83	Guatemala	2	Central Ame	Guatemala	Guatemala (	14.6228
28	1.97E+11	1970	1	21		0		160	Philippines	5	Southeast As	Metropolitar	Quezon City	14.674
29	1.97E+11	1970	1	22		0		222	Venezuela	3	South Ameri	Caracas	Caracas	10.4828
30	1.97E+11	1970	1	22		0		217	United State	1	North Ameri	Nebraska	South Sioux (	42.470
31	1.97E+11	1970	1	25		0		217	United State	1	North Ameri	Mississippi	West Point	33.606
32	1.97E+11	1970	1	25		0		217	United State	1	North Ameri	New York	New York Cit	40.6971
33	1.97E+11	1970	1	26		0		217	United State	1	North Ameri	Mississippi	West Point	33.606
34	1.97E+11	1970	1	26		0		217	United State	1	North Ameri	New York	New York Cit	40.6971
35	1.97E+11	1970	1	27		0		217	United State	1	North Ameri	Ohio	Norwalk	41.2419
36	1.97E+11	1970	1	28		0		499	East Germar	9	Eastern Euro	Berlin	Berlin	52.501
37	1.97E+11	1970	1	28		0		217	United State	1	North Ameri	Washington	Seattle	47.6107
38	1.97E+11	1970	1	30		0		217	United State	1	North Ameri	Nebraska	South Sioux (	42.4799

co2-emission-dataset

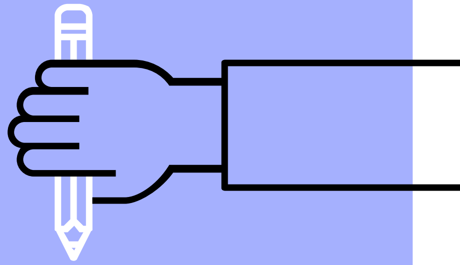
Dataset 2

	A	B	C	D	E
1	Entity	Code	Year	Annual CO2 emissions	
2	Afghanistan	AFG	1750	0	
3	Afghanistan	AFG	1751	0	
4	Afghanistan	AFG	1752	0	
5	Afghanistan	AFG	1753	0	
6	Afghanistan	AFG	1754	0	
7	Afghanistan	AFG	1755	0	
8	Afghanistan	AFG	1756	0	
9	Afghanistan	AFG	1757	0	
10	Afghanistan	AFG	1758	0	
11	Afghanistan	AFG	1759	0	
12	Afghanistan	AFG	1760	0	
13	Afghanistan	AFG	1761	0	
14	Afghanistan	AFG	1762	0	
15	Afghanistan	AFG	1763	0	
16	Afghanistan	AFG	1764	0	
17	Afghanistan	AFG	1765	0	
18	Afghanistan	AFG	1766	0	
19	Afghanistan	AFG	1767	0	
20	Afghanistan	AFG	1768	0	
21	Afghanistan	AFG	1769	0	
22	Afghanistan	AFG	1770	0	
23	Afghanistan	AFG	1771	0	
24	Afghanistan	AFG	1772	0	
25	Afghanistan	AFG	1773	0	
26	Afghanistan	AFG	1774	0	
27	Afghanistan	AFG	1775	0	
28	Afghanistan	AFG	1776	0	
29	Afghanistan	AFG	1777	0	
30	Afghanistan	AFG	1778	0	
31	Afghanistan	AFG	1779	0	
32	Afghanistan	AFG	1780	0	
33	Afghanistan	AFG	1781	0	
34	Afghanistan	AFG	1782	0	
35	Afghanistan	AFG	1783	0	
36	Afghanistan	AFG	1784	0	
37	Afghanistan	AFG	1785	0	
38	Afghanistan	AFG	1786	0	

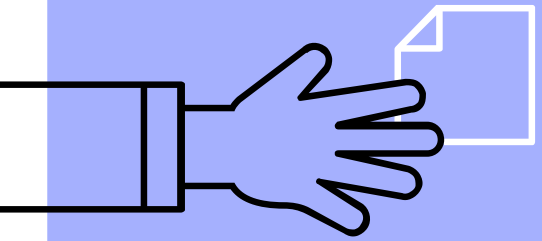
Life Expectancy (WHO)

Dataset 3

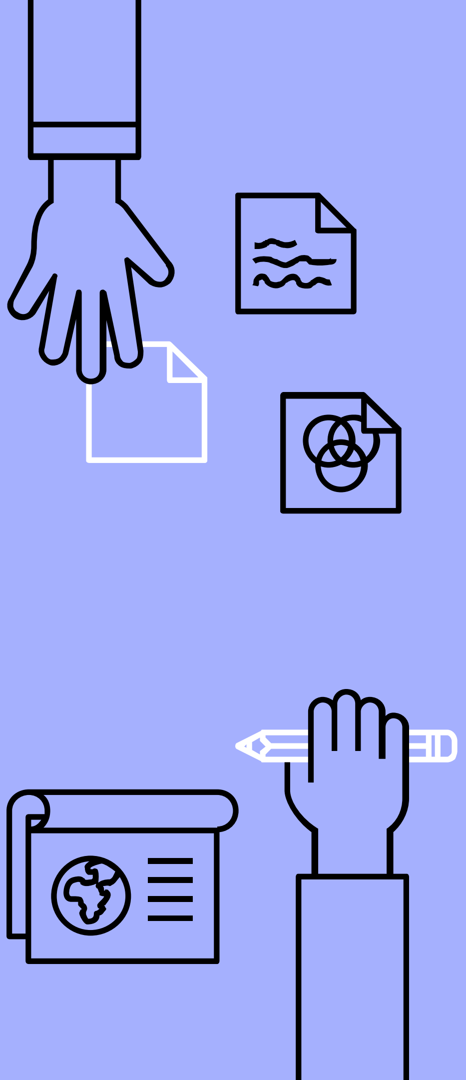
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Country	Year	Status	Life expectar	Adult Mortal	infant death:	Alcohol	percentage e	Hepatitis B	Measles	BMI	under-five de	Polio	Total expend
Afghanistan	2015	Developing	65	263	62	0.01	71.2796236	65	1154	19.1	83	6	8.16
Afghanistan	2014	Developing	59.9	271	64	0.01	73.5235817	62	492	18.6	86	58	8.18
Afghanistan	2013	Developing	59.9	268	66	0.01	73.2192427	64	430	18.1	89	62	8.13
Afghanistan	2012	Developing	59.5	272	69	0.01	78.1842153	67	2787	17.6	93	67	8.52
Afghanistan	2011	Developing	59.2	275	71	0.01	7.0971087	68	3013	17.2	97	68	7.87
Afghanistan	2010	Developing	58.8	279	74	0.01	79.6793674	66	1989	16.7	102	66	9.2
Afghanistan	2009	Developing	58.6	281	77	0.01	56.7622168	63	2861	16.2	106	63	9.42
Afghanistan	2008	Developing	58.1	287	80	0.03	25.8739254	64	1599	15.7	110	64	8.33
Afghanistan	2007	Developing	57.5	295	82	0.02	10.910156	63	1141	15.2	113	63	6.73
Afghanistan	2006	Developing	57.3	295	84	0.03	17.1715175	64	1990	14.7	116	58	7.43
Afghanistan	2005	Developing	57.3	291	85	0.02	1.38864773	66	1296	14.2	118	58	8.7
Afghanistan	2004	Developing	57	293	87	0.02	15.2960664	67	466	13.8	120	5	8.79
Afghanistan	2003	Developing	56.7	295	87	0.01	11.0890527	65	798	13.4	122	41	8.82
Afghanistan	2002	Developing	56.2	3	88	0.01	16.8873509	64	2486	13	122	36	7.76
Afghanistan	2001	Developing	55.3	316	88	0.01	10.5747282	63	8762	12.6	122	35	7.8
Afghanistan	2000	Developing	54.8	321	88	0.01	10.42496	62	6532	12.2	122	24	8.2
Albania	2015	Developing	77.8	74	0	4.6	364.975229	99	0	58	0	99	6
Albania	2014	Developing	77.5	8	0	4.51	428.749067	98	0	57.2	1	98	5.88
Albania	2013	Developing	77.2	84	0	4.76	430.876979	99	0	56.5	1	99	5.66
Albania	2012	Developing	76.9	86	0	5.14	412.443356	99	9	55.8	1	99	5.59
Albania	2011	Developing	76.6	88	0	5.37	437.0621	99	28	55.1	1	99	5.71
Albania	2010	Developing	76.2	91	1	5.28	41.8227572	99	10	54.3	1	99	5.34
Albania	2009	Developing	76.1	91	1	5.79	348.055952	98	0	53.5	1	98	5.79
Albania	2008	Developing	75.3	1	1	5.61	36.6220685	99	0	52.6	1	99	5.87
Albania	2007	Developing	75.9	9	1	5.58	32.2465523	98	22	51.7	1	99	6.1
Albania	2006	Developing	74.2	99	1	5.31	3.3021542	98	68	5.8	1	97	5.86
Albania	2005	Developing	73.5	15	1	5.16	26.9931214	98	6	49.9	1	97	6.12
Albania	2004	Developing	73	17	1	4.54	221.8428	99	7	48.9	1	98	6.38
Albania	2003	Developing	72.8	18	1	4.29	14.7192888	97	8	47.9	1	97	6.27
Albania	2002	Developing	73.3	15	1	3.73	104.516916	96	16	46.9	1	98	6.3
Albania	2001	Developing	73.6	14	1	4.25	96.2055708	96	18	46	1	97	6
Albania	2000	Developing	72.6	11	1	3.66	91.7115405	96	662	45	1	97	6.26
Algeria	2015	Developing	75.6	19	21		0	95	63	59.5	24	95	
Algeria	2014	Developing	75.4	11	21	0.01	54.2373183	95	0	58.4	24	95	7.21
Algeria	2013	Developing	75.3	112	21	0.53	544.450743	95	25	57.2	24	95	7.12
Algeria	2012	Developing	75.1	113	21	0.66	555.926083	95	18	56.1	24	95	6.14
Algeria	2011	Developing	74.9	116	21	0.56	509.002041	95	112	55	24	95	5.29

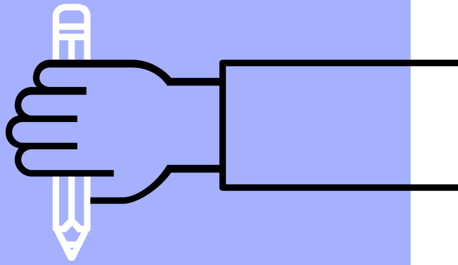


2.
Data cleaning

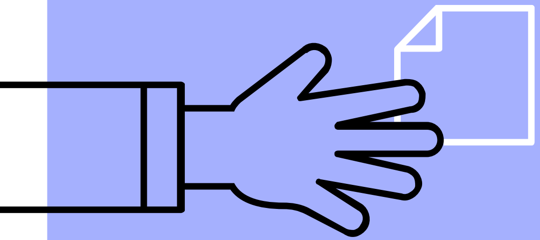


We merged all three datasets together because it is convenient for us to make analysis and find correlation in the future. We choose to merge on "Country" and "Year" becasue these are shared column of all three datasets. Since the "Year" column of all three datasets cover different years and dataset 3 only cover year 2000 - 2005, we limit the year to 2000 - 2015 so there won't be many missing data in the final dataset.

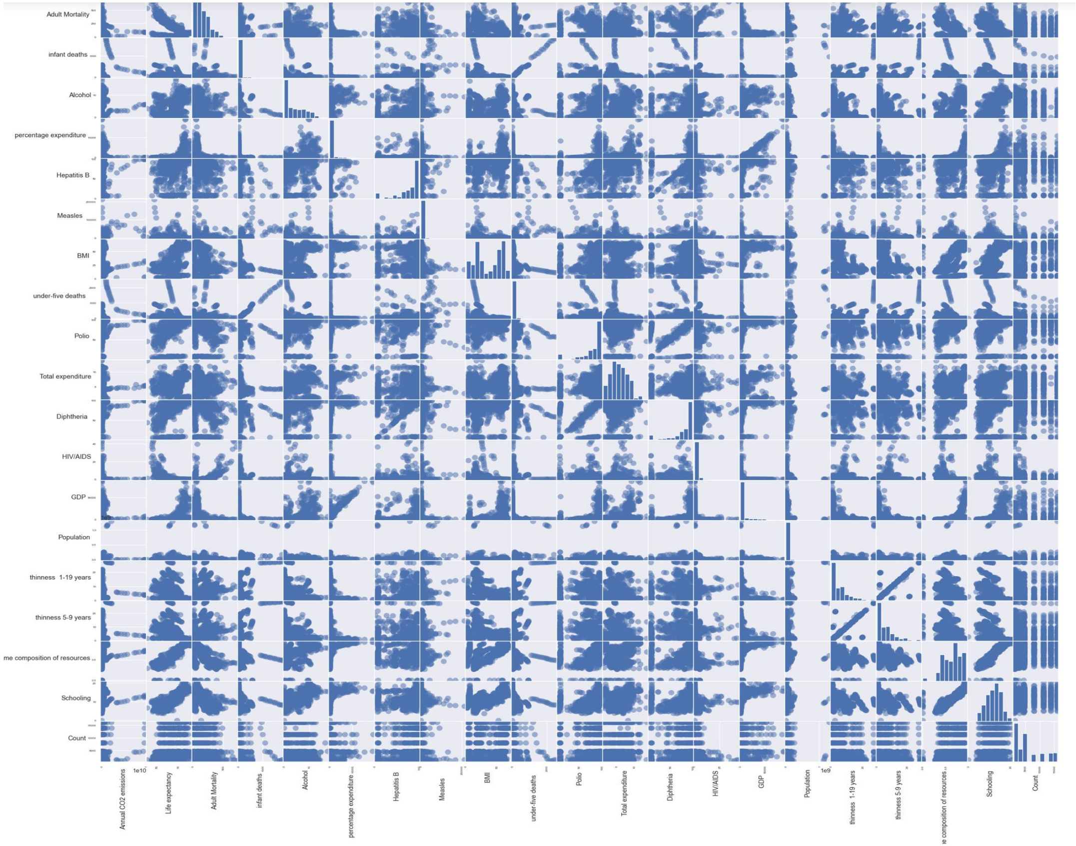




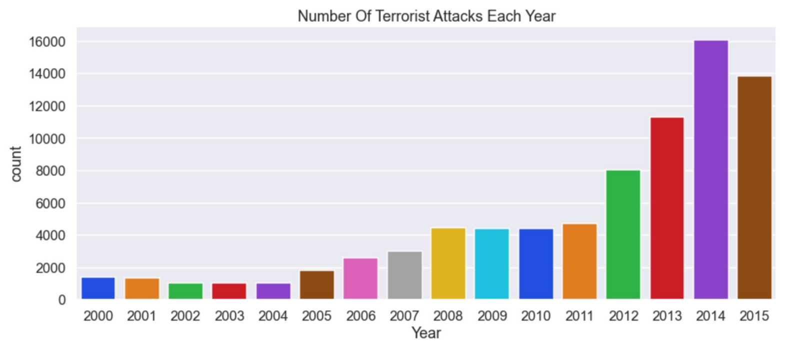
2.
EDA

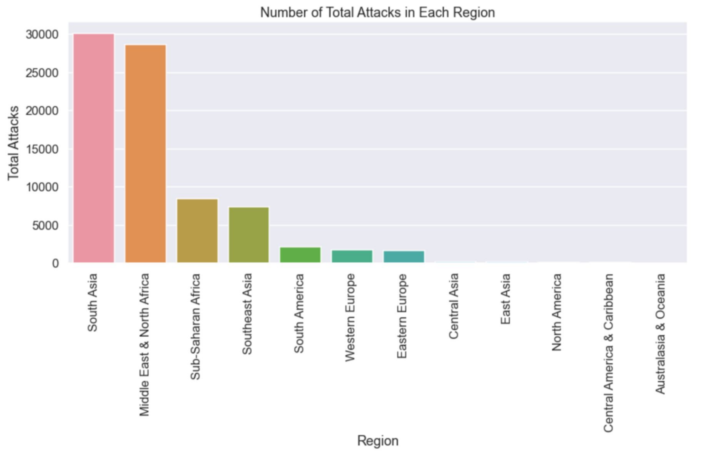


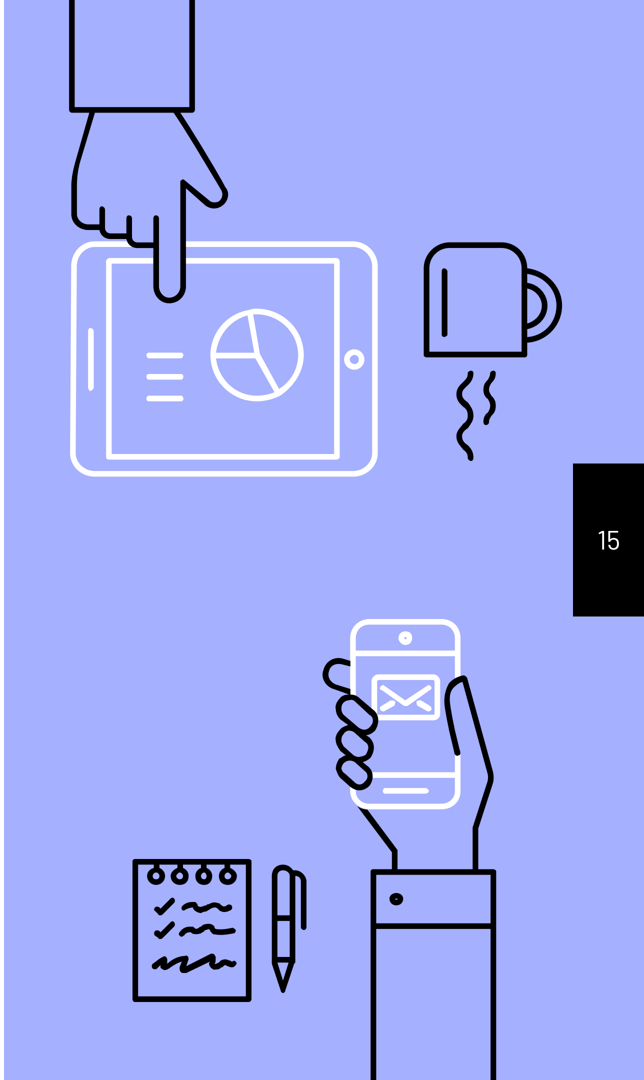
Visualizations



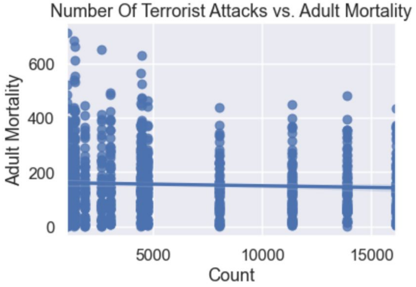
Visualizations

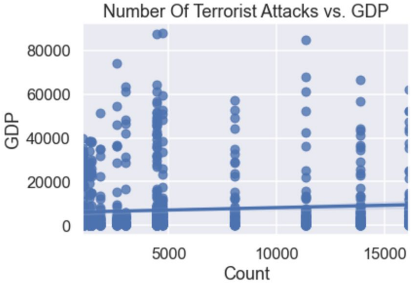


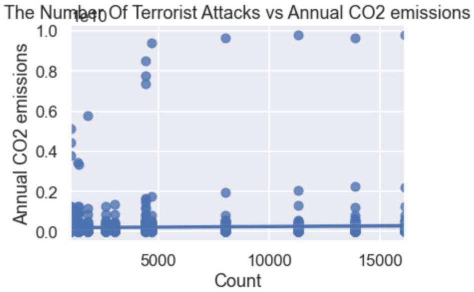


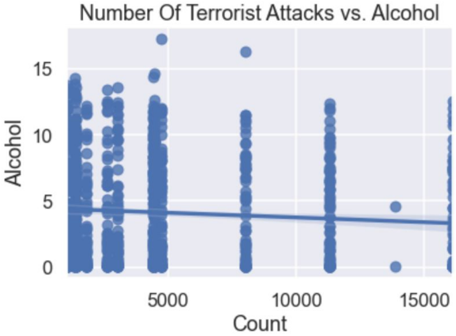


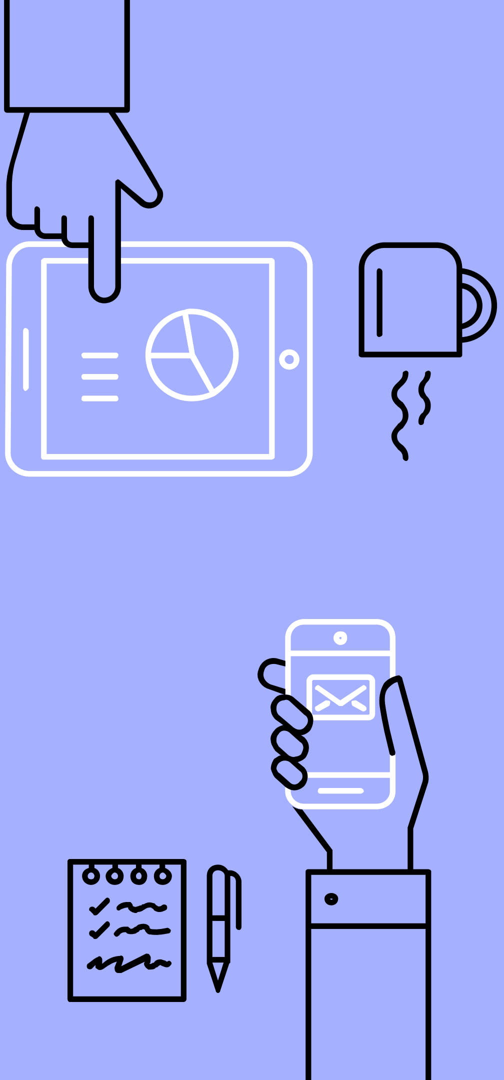
Visualizations





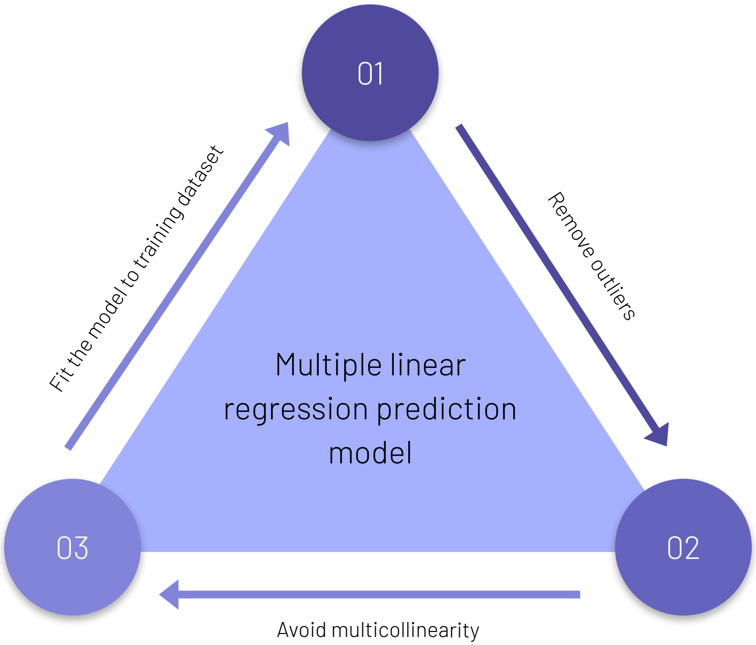




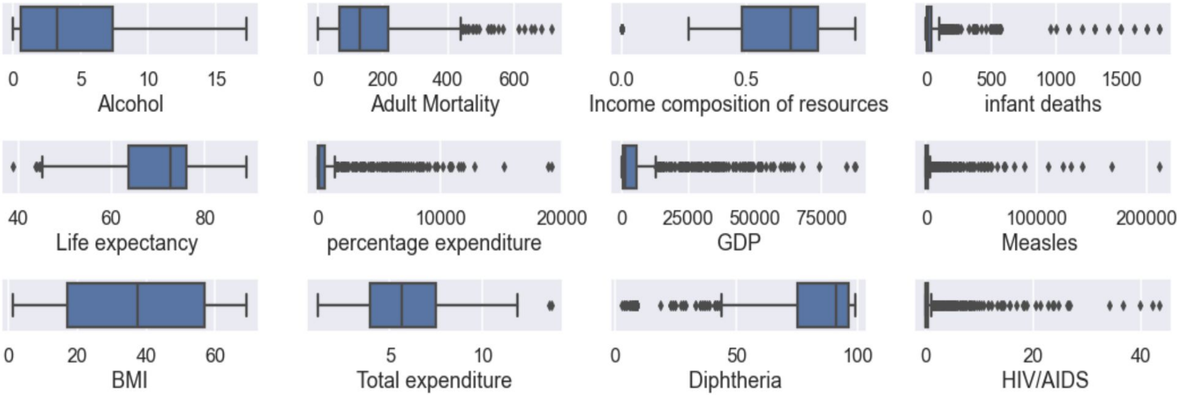


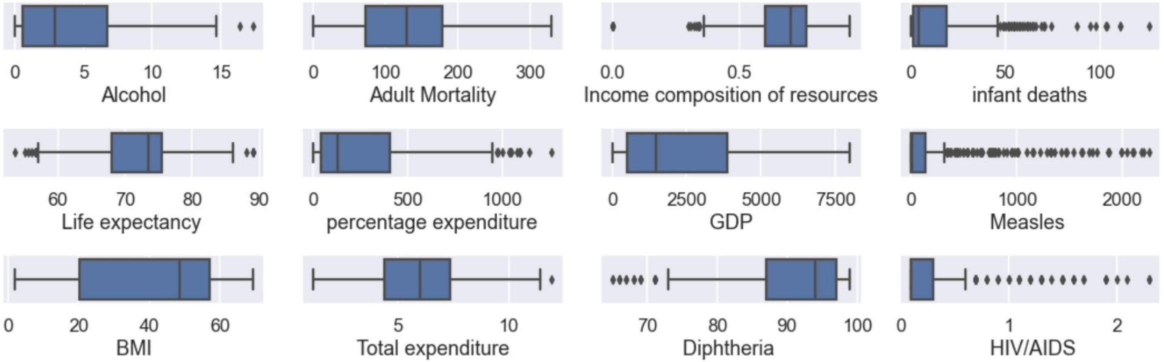
Multiple Linear Regression Prediction model

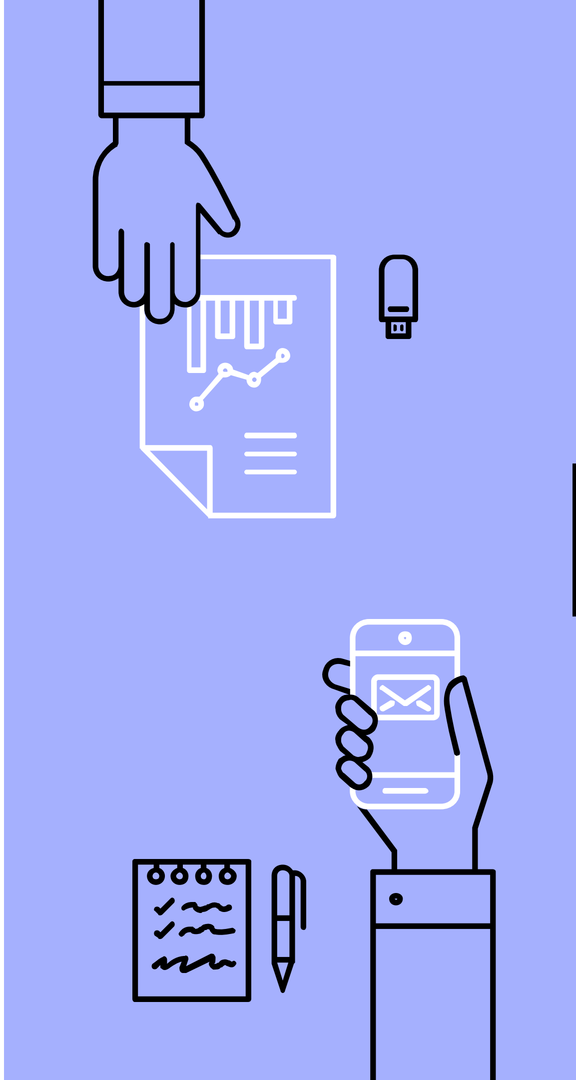
Steps


01
Remove outliers
02
Avoid multicollinearity
03
Fit the model to training dataset
Multiple linear regression prediction model

1


0 5 10 15
Alcohol
0 200 400 600
Adult Mortality
0.0 0.5
Income composition of resources
0 500 1000 1500
infant deaths
40 60 80
Life expectancy
0 10000 20000
percentage expenditure
0 25000 50000 75000
GDP
0 100000 200000
Measles
0 20 40 60
BMI
5 10
Total expenditure
0 50 100
Diphtheria
0 20 40
HIV/AIDS


0 5 10 15
Alcohol
0 100 200 300
Adult Mortality
0.0 0.5
Income composition of resources
0 50 100
infant deaths
60 70 80 90
Life expectancy
0 500 1000
percentage expenditure
0 2500 5000 7500
GDP
0 1000 2000
Measles
0 20 40 60
BMI
5 10
Total expenditure
70 80 90 100
Diphtheria
0 1 2
HIV/AIDS



2

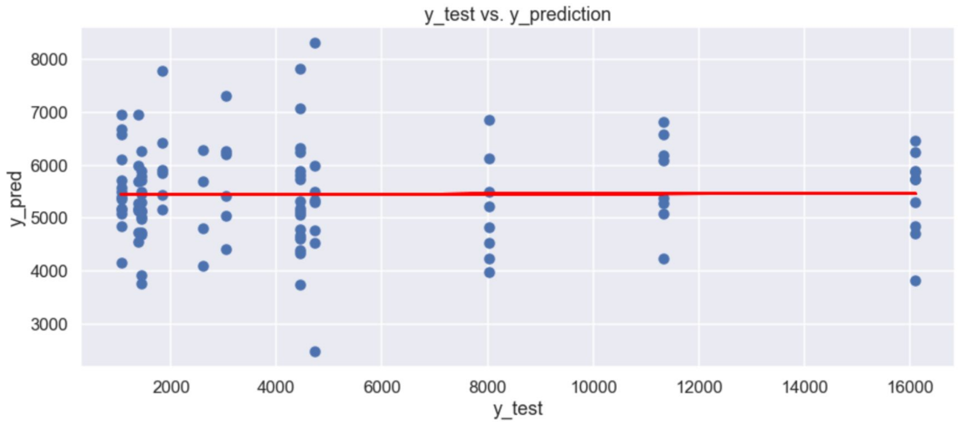
```
In [85]: preprocessing(X)
```

	VIF	Features
0	37.188662	Life expectancy
1	3.031046	Alcohol
2	45.396187	Income composition of resources
3	4.387596	Adult Mortality
4	7.611101	GDP
5	1.836985	infant deaths
6	6.765698	percentage expenditure
7	1.432079	Measles
8	8.099446	BMI
9	5.792687	Total expenditure
10	15.188202	Diphtheria
11	1.776908	HIV/AIDS

```
In [86]: # drop those columns to avoid multicollinearity
X.drop(['Life expectancy ', 'Income composition of resources', 'GDP', 'percentage expenditure',' BMI ', 'Total expen
preprocessing(X)
```

	VIF	Features
0	1.568358	Alcohol
1	2.849879	Adult Mortality
2	1.797224	infant deaths
3	1.331247	Measles
4	1.505919	HIV/AIDS

3



In [90]:
```
model_1 = sms.OLS(y_train, x_train).fit()
model_1.summary()
```

Out[90]:

OLS Regression Results

Dep. Variable:	Count	**R-squared (uncentered):**	0.465
Model:	OLS	**Adj. R-squared (uncentered):**	0.456
Method:	Least Squares	**F-statistic:**	53.01
Date:	Thu, 09 Dec 2021	**Prob (F-statistic):**	1.74e-39
Time:	02:16:39	**Log-Likelihood:**	-3096.3
No. Observations:	310	**AIC:**	6203.
Df Residuals:	305	**BIC:**	6221.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]

Let’s improve the model!

OUR PROCESS of improving

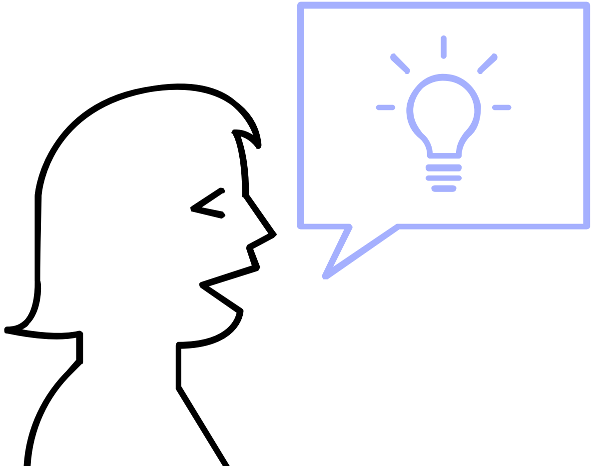
Increase independent variables

Change dependent variable to Gaussian distribution

Run the model again!!

0.931

This is the final R-squared



X2

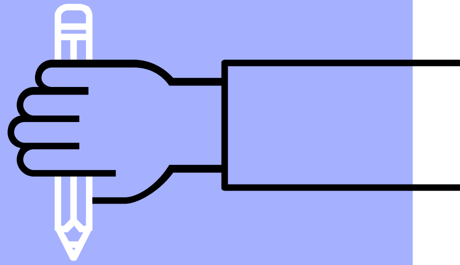
We doubled the model performance

Over 100 lines

A lot of cells

100%

Our total effort



Thanks for your time