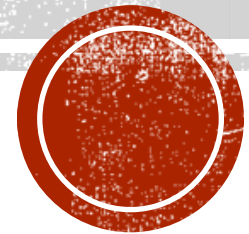


DATA PREPROCESSING

For panel data



0. DATA SOURCE

- <http://databank.worldbank.org/>
- 1. GDP(constant 2010 US\$)New.csv
- 2. co2 emissionsNew.csv
- 3. commercial bank branchesNew.csv
- 4. foreign direct investment net inflowsNew.csv
- 5. gini indexNew.csv
- 6. government expenditure on educationNew.csv
- 7. internetUsersNew.csv
- 8. labor forceNew.csv
- 9. life expectancyNew.csv
- 10. mobile cellular subscriptionsNew.csv
- 11. railways goods transportedNew.csv
- 12. school life expectancyNew.csv
- 13. technical articlesNew.csv
- 14. time required to start a businessNew.csv
- 15. unemployment rateNew.csv



1. DATA CLEANING



1.1 CHOOSE YEAR BETWEEN 2004 – 2014

E4																
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Series Name	Series Code	Country	Country Code	1990 [YR199	1991 [YR199	1992 [YR199	1993 [YR199	1994 [YR199	1995 [YR199	1996 [YR199	1997 [YR199	1998 [YR199	1999 [YR199	2000 [YR199	2001 [YR199
2	GINI index (V SI.POV.GINI		Afghanistan	AFG
3	GINI index (V SI.POV.GINI		Albania	ALB	27.01
4	GINI index (V SI.POV.GINI		Algeria	DZA
5	GINI index (V SI.POV.GINI		Angola	AGO
6	GINI index (V SI.POV.GINI		Argentina	ARG	..	46.76	45.47	44.86	45.92	48.9	49.52	49.11	50.73	49.79
7	GINI index (V SI.POV.GINI		Armenia	ARM	44.42	36.22
8	GINI index (V SI.POV.GINI		Australia	AUS	33.72
9	GINI index (V SI.POV.GINI		Austria	AUT
10	GINI index (V SI.POV.GINI		Azerbaijan	AZE	34.65
11	GINI index (V SI.POV.GINI		Bangladesh	BGD	..	27.57	32.94
12	GINI index (V SI.POV.GINI		Belarus	BLR	21.6	..	28.76	32.25	32
13	GINI index (V SI.POV.GINI		Belgium	BEL
14	GINI index (V SI.POV.GINI		Belize	BLZ	60.25	60.91	57.55	56.59	60.43	54.91	53.26
15	GINI index (V SI.POV.GINI		Benin	BEN
16	GINI index (V SI.POV.GINI		Bhutan	BTN
17	GINI index (V SI.POV.GINI		Bolivia	BOL	42.04	..	49.11	58.16	..	58.1
18	GINI index (V SI.POV.GINI		Bosnia and H	BIH
19	GINI index (V SI.POV.GINI		Botswana	BWA	60.79



1.2 DROP COUNTRIES THAT MISS DATA OVER 3 YEARS

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Series Name	Series Code	Country	Country Code	1990 [YR199	1991 [YR199	1992 [YR199	1993 [YR199	1994 [YR199	1995 [YR199	1996 [YR199	1997 [YR199	1998 [YR199	1999 [YR199	2000 [YR199
2	GINI index (V SI.POV.GINI		Afghanistan	AFG
3	GINI index (V SI.POV.GINI		Albania	ALB	27.01
4	GINI index (V SI.POV.GINI		Algeria	DZA
5	GINI index (V SI.POV.GINI		Angola	AGO
6	GINI index (V SI.POV.GINI		Argentina	ARG	..	46.76	45.47	44.86	45.92	48.9	49.52	49.11	50.73	49.79	..
7	GINI index (V SI.POV.GINI		Armenia	ARM	44.42	36.22	..
8	GINI index (V SI.POV.GINI		Australia	AUS	33.72
9	GINI index (V SI.POV.GINI		Austria	AUT
10	GINI index (V SI.POV.GINI		Azerbaijan	AZE	34.65
11	GINI index (V SI.POV.GINI		Bangladesh	BGD	..	27.57	32.94
12	GINI index (V SI.POV.GINI		Belarus	BLR	21.6	..	28.76	32.25	32	..
13	GINI index (V SI.POV.GINI		Belgium	BEL
14	GINI index (V SI.POV.GINI		Belize	BLZ	60.25	60.91	57.55	56.59	60.43	54.91	53.26	..
15	GINI index (V SI.POV.GINI		Benin	BEN
16	GINI index (V SI.POV.GINI		Bhutan	BTN
17	GINI index (V SI.POV.GINI		Bolivia	BOL	42.04	..	49.11	58.16	..	58.1	..
18	GINI index (V SI.POV.GINI		Bosnia and H	BIH
19	GINI index (V SI.POV.GINI		Botswana	BWA	60.79

One of 15 tables, Gini index



1.3 FILL MISSING DATA WITH YEAR AVG

```
26 from sklearn.preprocessing import Imputer
27 • imputer = Imputer(missing_values = 'NaN', strategy = 'mean', axis = 0)
28   imputer = imputer.fit(E[:, 0:columnsE])
29   E[:, 0:columnsE] = imputer.transform(E[:, 0:columnsE])
30   expect.to_csv("commercial bank branchesNew.csv")
```



1.4 ARRANGE COMBINATION TO PICK DATASETS

intersection	list	27	['Brazil', 'Slovenia', 'Estonia', 'Iran, Isl...
--------------	------	----	---

```
Console 1/A
233
In [111]: for m in fileName:
...:     fileM = pd.read_csv(m)
...:     print(str(fileM.shape[0]) + "      "+m)
241     GDP(constant 2010 US$)New.csv
233     co2 emissionsNew.csv
201     commercial bank branchesNew.csv
2563     final dataset.csv
215     foreign direct investment net inflowsNew.csv
90     gini indexNew.csv
104     government expenditure on educationNew.csv
242     internetUsersNew.csv
231     labor forceNew.csv
240     life expectancyNew.csv
228     mobile cellular subscriptionsNew.csv
96     railways goods transportedNew.csv
121     school life expectancyNew.csv
239     technical articlesNew.csv
200     time required to start a businessNew.csv
199     unemployment rateNew.csv
```



209

7

192

8

185

9

172

10

159

11

142

12

89

13

56

14

1.5 PICK DATASETS

1. CO2 emission kg/\$, constant 2010
2. Foreign direct investment net inflows, \$, constant 2010
3. Commercial bank branches, per 100,000 people
4. Time required to start a business, days
5. labor force, people
6. Internet users, per 1,000 people
7. Life expectancy, years
8. Mobile cellular subscriptions, per 100 people
9. Unemployment rate
10. Technical articles



2 PUT ALL DATA IN ONE TABLE

A	B	C	D	E	F	G	H	I	J	K	L
	Country Name	Year	GDP(constant 2010 US\$)	co2 emissions	commercial bank branches	foreign direct investment in	internetUsers	labor force	life expectancy	mobile cellular	technology
0	Afghanistan	2004	8781610175	0.108152489	0.383742856	186900000	0.10580903	6478539	57.89121951	2.49805547	
1	Afghanistan	2005	9762978844	0.135968132	0.609975301	271000000	1.224148084	6795212	58.51558537	4.82686537	
2	Afghanistan	2006	10305228125	0.16012746	0.964169781	238000000	2.107123645	6982165	59.12595122	9.83316402	
3	Afghanistan	2007	11721187594	0.1939684	1.287792643	188690000	1.9	7139660	59.70885366	17.7162433	
4	Afghanistan	2008	12144482858	0.346334138	1.531646932	46033740	1.84	7292595	60.25673171	29.2203738	
5	Afghanistan	2009	14697331941	0.46057897	2.300990369	197512727.5	3.55	7474251	60.76460976	37.894937	
6	Afghanistan	2010	15936800636	0.531062425	2.463293515	54200551	4	7707349	61.23546341	45.7781747	
7	Afghanistan	2011	16911126453	0.723810211	2.290036916	57620844	5	8050184	61.6742439	60.32632	
8	Afghanistan	2012	19352203806	0.555766729	2.211204172	47226787.51	5.454545455	8458402	62.09297561	65.4521935	
9	Afghanistan	2013	19731337261	0.507546796	2.340217032	37638586	5.9	8916157	62.50160976	70.6613589	
10	Afghanistan	2014	19990317161	0.490698818	2.397348601	43510167.3	7	9397624	62.90268293	74.8828424	
11	Albania	2004	8766856445	0.475165987	9.670207822	341285112.5	2.420387798	1302765	75.15709756	39.1639567	
12	Albania	2005	9268392518	0.458949056	11.66806544	262479012.6	6.043890864	1284116	75.35702439	47.8780275	
13	Albania	2006	9771760096	0.398906744	13.76968007	325138316.8	9.609991316	1270016	75.5612439	60.0673424	
14	Albania	2007	10348293942	0.379517341	18.48652355	652275603.7	15.03611541	1254859	75.79492683	73.3503842	
15	Albania	2008	11127520474	0.393145176	23.28914326	1240972849	23.86	1239840	76.07265854	58.9123515	
16	Albania	2009	11500292411	0.38072058	23.72983753	1343091150	41.2	1225724	76.39487805	78.1845877	
17	Albania	2010	11926953259	0.385548421	23.81053625	1089416366	45	1216574	76.74790244	85.468247	
18	Albania	2011	1233100568	0.428428100	23.84044138	1040425306	40	1200573	77.10846341	88.2015346	

