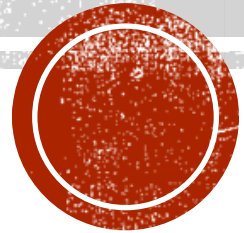
DATA PREPROCESSING

For panel data



0. DATA SOURCE

11 years *
15 regressors *
220 (average) countries ≈ 36, 300

- http://databank.worldbank.org/

1. GDP(constant 2010 US$).csv
2. co2 emissions.csv
3. commercial bank branches.csv
4. foreign direct investment net inflows.csv
5. gini index.csv
6. government expenditure on education.csv
7. internetUsers.csv
8. labor force.csv
9. life expectancy.csv
10. mobile cellular subscriptions.csv
11. railways goods transported.csv
12. school life expectancy.csv
13. technical articles.csv
14. time required to start a business.csv
15. unemployment rate.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	
1	Series Name	Series Code	Country	Country Cod	1990 [YR199	1991 [YR199	1992 [YR199	1993 [YR199	1994 [YR199	1995 [YR199	1996 [YR199	1997 [YR199	1998 [YR199	1999 [YR199	20(
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

E4 fx ..

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Series Name	Series Code	Country	Country Cod	1990 [YR199	1991 [YR199	1992 [YR199	1993 [YR199	1994 [YR199	1995 [YR199	1996 [YR199	1997 [YR199	1998 [YR199	1999 [YR199	20(
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

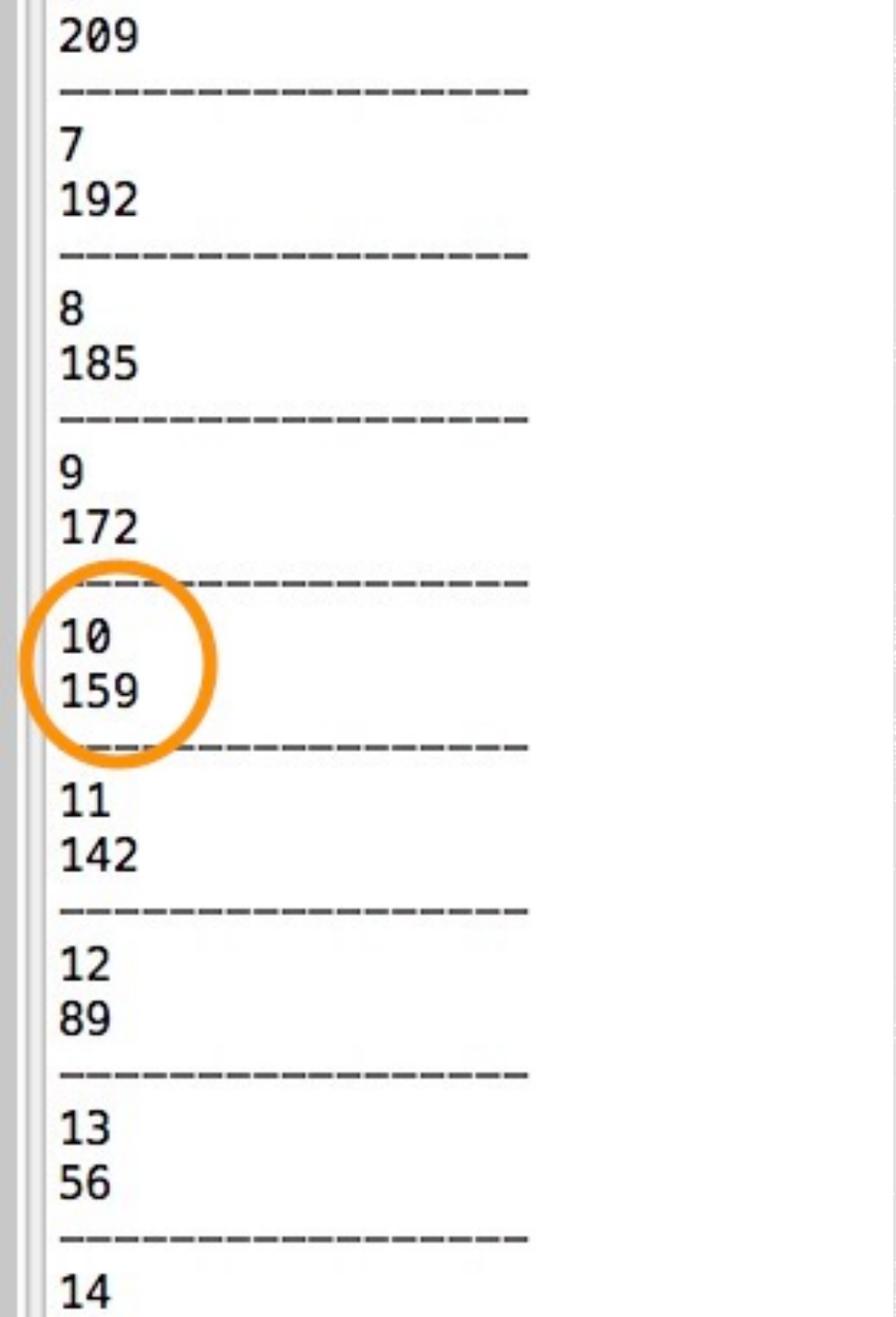
```
26    from sklearn.preprocessing import Imputer
27    imputer = Imputer(missing_values = 'NaN', strategy = 'mean', axis = 0)
28    imputer = imputer.fit(E[:, 0:columsE])
29    E[:, 0:columsE] = imputer.transform(E[:, 0:columsE])
30    expect.to_csv("commercial bank branchesNew.csv")
```

1.4 ARRANGE COMBINATION TO PICK DATASETS

intersection	list	27	['Brazil', 'Slovenia', 'Estonia', 'Iran, Isl…

Console 1/A

```
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
```



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 bossiness, 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 DATA INTEGRATION

	Country Nam	Year	GDP(constant 2010 US$)	co2 emissions	commercial bank branches	foreign direct investment n	internetUsers	labor force	life expectancy	mobile cellul	tec
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	12231090568	0.428428109	23.84044128	1049425306	49	1209673	77.10846341	98.2915346	

3 DATA REDUCTION

The coefficient is statistically significant p-value < 0.1

```
Fixed-effects (within) regression               Number of obs      =       1,562
Group variable: year                            Number of groups   =          11

R-sq:                                           Obs per group:
     within  = 0.9868                                         min =         142
     between = 0.9582                                         avg =       142.0
     overall = 0.9860                                         max =         142

                                                F(10,1541)         =     11500.74
corr(u_i, Xb)  = -0.0519                        Prob > F           =      0.0000
```

gdpconstant2010us	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
co2emissionskgper2010usofgdp	-1.82e+11	6.08e+10	-2.98	0.003	-3.01e+11	-6.23e+10
commercialbankbranchesper100000a	-9.92e+07	2.10e+09	-0.05	0.962	-4.22e+09	4.02e+09
foreigndirectinvestmentnetinflow	2.379906	.2372237	10.03	0.000	1.914591	2.845222
internetusersper100people	-3.68e+08	1.81e+09	-0.20	0.839	-3.92e+09	3.18e+09
laborforce	247.9013	88.8373	2.79	0.005	73.6465	422.156
lifeexpectancy	4.77e+09	4.95e+09	0.96	0.335	-4.93e+09	1.45e+10
mobilecellularsubscriptionsper10	7.14e+07	1.13e+09	0.06	0.949	-2.14e+09	2.28e+09
technicalarticles	3.21e+07	306882.8	104.76	0.000	3.15e+07	3.28e+07
timerequiredtostartabusiness	2.83e+08	4.55e+08	0.62	0.534	-6.09e+08	1.17e+09
unemploymentrate	-7.78e+08	5.17e+09	-0.15	0.880	-1.09e+10	9.36e+09
_cons	-1.40e+11	3.00e+11	-0.47	0.640	-7.29e+11	4.48e+11
sigma_u	2.416e+11					
sigma_e	9.792e+11					
rho	.05739609	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(10, 1541) = 5.15                     Prob > F = 0.0000
```

4 DATA STANDARDIZATION

co2emissio~p	comm~100000a	foreigndir~w	internetus~e	laborforce	lifeexpect~y	mobilecel~10	technicala~s	timerequir~s	unemployme~e	z2co2	z2ta
.1081525	.3837429	1.869e+08	.105809	6478539	57.89122	2.498055	6.8	9	8.5	-.9132215	-.3054654
.1359681	.6099753	2.710e+08	1.224148	6795212	58.51559	4.826865	8.4	9	8.5	-.8499383	-.3054587
.1601275	.9641698	2.380e+08	2.107124	6982165	59.12595	9.833164	10.2	9	8.8	-.7949736	-.3054512
.1939684	1.287793	1.887e+08	1.9	7139660	59.70885	17.71624	13	9	8.4	-.7179823	-.3054395
.3463341	1.531647	46033740	1.84	7292595	60.25673	29.22037	12.7	9	8.9	-.3713361	-.3054408
.460579	2.30099	1.975e+08	3.55	7474251	60.76461	37.89494	21.5	7	8.1	-.1114185	-.3054041
.5310624	2.463294	54200551	4	7707349	61.23546	45.77818	29.3	7	8.7	.0489379	-.3053716
.7238102	2.290037	57620844	5	8050184	61.67424	60.32632	43.2	7	8.9	.487457	-.3053137
.5557667	2.211204	47226788	5.454545	8458402	62.09298	65.45219	35	7	8.5	.1051424	-.3053479
.5075468	2.340217	37638586	5.9	8916157	62.50161	70.66136	26.5	5	9.2	-.0045624	-.3053833
.4906988	2.397349	43510167	7	9397624	62.90268	74.88284	26.5	7	9.1	-.0428931	-.3053833
.475166	9.670208	3.413e+08	2.420388	1302765	75.1571	39.16396	21.9	39	12.6	-.0782317	-.3054025
.4589491	11.66807	2.625e+08	6.043891	1284116	75.35703	47.87803	24.5	40	12.5	-.1151267	-.3053916
.3989067	13.76968	3.251e+08	9.609991	1270016	75.56124	60.06734	31.8	38	12.4	-.2517286	-.3053612
.3795173	18.48652	6.523e+08	15.03612	1254859	75.79493	73.35039	40.5	36	13.5	-.2958412	-.3053249
.3931452	23.28914	1.241e+09	23.86	1239840	76.07266	58.91235	51.5	9	13	-.2648366	-.3052791