Data Management and Visualization

Developing a Research Question and Creating Your Personal Code Book

STEP 1: Choose a data set that you would like to work with.

I am choosing GapMinder dataset.

STEP 2. Identify a specific topic of interest

I am exploring is there a relationship on Polity scores with life expectancy.

STEP 3. Prepare a codebook of your own (i.e., print individual pages or copy screen and paste into a new document) from the larger codebook that includes the questions/items/variables that measure your selected topics.)

In [4]:

```
df = pd.read_csv("gapminder.csv")
df.columns
```

Out[4]:

```
Index(['country', 'incomeperperson', 'alcconsumption', 'armedforcesrate', 'breastcancerper100th', 'co2emission
s', 'femaleemployrate', 'hivrate', 'internetuserate', 'lifeexpectancy', 'oilperperson', 'polityscore', 'relectr
icperperson', 'suicideper100th', 'employrate', 'urbanrate'], dtype='object')
```

In [5]:

```
df
```

Out[5]:

	country	incomeperperson	alcconsumption	armedforcesrate	breastcancerper100th	co2emissions	femaleemployrate	hivrate	int
0	Afghanistan		.03	.5696534	26.8	75944000	25.6000003814697		3.654
1	Albania	1914.99655094922	7.29	1.0247361	57.4	223747333.333333	42.0999984741211		44.98
2	Algeria	2231.99333515006	.69	2.306817	23.5	2932108666.66667	31.7000007629394	.1	12.50
3	Andorra	21943.3398976022	10.17						
4	Angola	1381.00426770244	5.57	1.4613288	23.1	248358000	69.4000015258789	2	9.999
...	...	...	...	...	...	...	...	...	
208	Vietnam	722.807558834445	3.91	1.0853671	16.2	1425435000	67.5999984741211	.4	27.85
209	West Bank and Gaza			5.9360854		14241333.3333333	11.3000001907349		36.42
210	Yemen, Rep.	610.3573673206	.2	2.3162346	35.1	234864666.666667	20.2999992370605		12.34
211	Zambia	432.226336974583	3.56	.3413352	13	132025666.666667	53.5	13.5	10.1
212	Zimbabwe	320.771889948584	4.96	1.0327854	19	590219666.666666	58.0999984741211	14.3	11.50

213 rows × 16 columns

Data Dictionary

Field	Description
country	Unique Identifier
incomeperperson	2010 Gross Domestic Product per capita in constant 2000 US$
alcconsumption	2008 alcohol consumption per adult (age 15+), litres
armedforcesrate	Armed forces personnel (% of total labor force)
breastcancerper100th	2002 breast cancer new cases per 100,000 female
co2emissions	2006 cumulative CO2 emission (metric tons)
femaleemployrate	2007 female employees age 15+ (% of population)
hivrate	2009 estimated HIV Prevalence % - (Ages 15-49)
internetuserate	2010 Internet users (per 100 people)
lifeexpectancy	2011 life expectancy at birth (years)
oilperperson	2010 oil Consumption per capita (tonnes per year and person)
polityscore	2009 Democracy score (Polity)
relectricperperson	2008 residential electricity consumption, per person (kWh)
suicideper100th	2005 Suicide, age adjusted, per 100 000
employrate	2007 total employees age 15+ (% of population)
urbanrate	2008 urban population (% of total)

STEP 4. Identify a second topic that you would like to explore in terms of its association with your original topic

The second one is has employment rate influence urban rates.

STEP 5. Add questions/items/variables documenting this second topic to your personal codebook

STEP 6. Perform a literature review to see what research has been previously done on this topic.

Ref 1: Health advocacy with Gapminder animated statistics

Ref 2: Formalizing students’ informal statistical reasoning on real data: Using Gapminder to follow the cycle of inquiry and visual analyses

Ref 3: USE OF TED.COM and GAPMINDER.ORG IN TEACHING APPLICATIONS OF MATHEMATICS AND STATISTICS

STEP 7. Based on your literature review, develop a hypothesis about what you believe the association might be between these topics. Be sure to integrate the specific variables you selected into the hypothesis.

Hypothesis suggested: Has suicide rate influenced by HIV rate on victims?

In [ ]: