What's the story?

There's diamond in data

Urban Societies & Alcohol

WEEK 4 (May 15, 2016) Assignment 4 –Creating Graphs for your Data (SAS)

DATA: Global Development Indicators from GapMinder

My broader aim at the start of this course was to explore the effects of Urbanization on societies across the globe. How does Urbanization relate to Alcohol Consumption (Male & Female), Income, Suicide Rates, etc. I narrowed down my primary research question to establish a relationship between **Alcohol Consumption** and **Urbanization.** Since I used the GapMinder data set, which had data for a single year, I looked at one year's (2008) data for my chosen variables. My hypothesis was that despite the global financial crisis around 2008, the two variables will be related positively.

I have also included another related variable, **Income per person** or **GDP per Capita** for the year 2008 to see how they related to Urbanization. My hypothesis was that in spite of the economic downturn that year, the relationship would be by and large positive. For this assignment I have included this variable only in the bivariate charts.

My dataset is a modified version of the GapMinder data provided by the course. I collated and used data from The World Bank’s development indicators’ website to make the data more comparable to variables provided by GapMinder. So I had to treat the combined/modified data set as a new one and had to upload and import it to SAS Studio.

———

1. Univariate Graphs for Consumption and Urbanization

——————————————————————————-

Since my data set consists of continuous variables, they needed to be binned. There are two charts for each variable below: one with 10 bins generated by SAS and the second with 4 bins that I created. I have found the graphs generated with bins I have created are easier to interpret, so I have based my comments on them.

The code below is for Urban Rate, but the same code was used for all the variables: I commented out the variables I didn’t need.

```
PROC IMPORT DATAFILE=“/home/snigdhasen0/NewDataSet2008.csv” DBMS=CSV OUT=imported
REPLACE;

run;
Data step1; set imported;
Label urbanrate=“2008 Urban Population (% of total)”
incomeperperson08=“2008 GDP per capita (USD)”
    alcconsumption=“2008 Alcohol Consumption per adult(15+ yrs), liters”;
Keep  urbanrate alcconsumption incomeperperson08 ;
```

```
/*histogram with quantitative variable for urbanization*/
Proc gchart; vbar urbanrate;
run;
```

```
PROC FREQ; tables urbanrate /*alcconsumption; incomeperperson08*/;
```

LINKS TO URBAN RATE UNIVARIATE GRAPHS (Auto bins, My bins)
Urban Rate: Both graphs are similar in shape. The categorical graph is unimodal. The mode, or the bin/value that takes the highest frequency, (74 countries) is between 50 and 75% Urbanized (*My bins*) and 48 countries are 75-100% urbanized; this means that 122 or ~60 % of the countries for which data is available are 50 % or more urbanized . The graph seems skewed to the left as there are more lower frequencies in lower categories.

LINKS TO ALCOHOL CONSUMPTION UNIVARIATE GRAPHS (Auto bins, My bins)
Alcohol Consumption: Again, both graphs are structurally similar and are unimodal, the mode being 4-6 liters. The graph seems skewed to the right as more higher values take lower frequencies. Interestingly, there are more countries (45) that have a high alcohol consumption rate (over 10 liters) than countries (40) that have a low rate (0-2 liters).

---

2. Measures of Center and Spread (Table: Basic Statistical Measures)

[The code below runs the Univariate Procedure for 3 variables: Urban Rate, Alcohol Consumption and Income Per Person. The following output and comments that I have included here are for my two primary variables only, namely Urban Rate and Alcohol Consumption.]

```
PROC IMPORT DATAFILE=“/home/snigdhasen0/NewDataSet2008.csv” DBMS=CSV OUT=imported
REPLACE;

run;
```

```
Data step1; set imported;
Label urbanrate=“2008 Urban Population (% of total)”
incomeperperson08=“2008 GDP per capita (USD)”
      alcconsumption=“2008 Alcohol Consumption per adult(15+ yrs), liters”
    femaleemp08=“2008 Female Employees Age 15+ (% of female population, 15+)”
    totalemp08=“2008 Employees Age 15+ (% of total population, 15+)”;
Keep urbanrate alcconsumption incomeperperson08;
Proc univariate; var urbanrate alcconsumption incomeperperson08;
PROC FREQ; tables urbanrate alcconsumption incomeperperson08 ;
run;
```

The UNIVARIATE Procedure
Variable: urbanrate (2008 Urban Population (% of total))

Urban Rate: The Mean for this variable is 56.7% , while Standard Deviation is 23.8 thus implying a significant spread. The Median Urban Rate is 58%, almost aligned with the average. Note, however, that the Mode is 100%, which is not the Mode we saw in the binned categorical Univariate graph described above. This is because this Proc statement was run with continuous quantitative values; if we look at the table which lists extreme values we find that 5 countries have 100 % urban rate, making it the value with the highest frequency, or the Mode. Binning helps eliminate these peculiarities.

The UNIVARIATE Procedure
Variable: alcconsumption (2008 Alcohol Consumption per adult (15+ yrs), liters)

Alcohol Consumption: The Mean Alcohol Consumption rate is ~7 liters per adult (15+ years), with a Standard Deviation of ~5 l, implying quite a bit of spread. The Median rate is ~6 l of alcohol per adult. As in the case of Urban Rate, since the variable is continuous and not binned, there are 7 modes (of 2 countries each), meaning each of the 7 values takes the highest frequency, which is 2 countries in this case. Again, binning gives a better perspective on the data.

---

3. Urban Rate Vs Alcohol Consumption (ScatterPlot)

To answer my primary research question – Do more urbanized countries have higher alcohol consumption – I used a scatter plot, since my variables are continuous and quantitative.

```
PROC IMPORT DATAFILE=“/home/snigdhasen0/NewDataSet2008.csv” DBMS=CSV OUT=imported
REPLACE;

run;

Data step1; set imported;

Proc sort; by countrynew;

Label urbanrate=“2008 Urban Population (% of total)”

    alcconsumption=“2008 Alcohol Consumption per adult(15+ yrs), liters”;

PROC GPlot; plot alcconsumption*urbanrate;
```

SCATTER PLOT OUTPUT (Link)

As hypothesized, there appears to be a positive relationship between Urbanization and Alcohol Consumption. The plot seems skewed to the left, as the left tail is longer; more countries with lower Urban Rates generally take lower alcohol consumption values.

There is, however, a clutter of highly urbanized countries with low alcohol consumption rates, way below the average of about 6 liters. Looking at the data set, these include wealthy nations like Kuwait (Urban Rate 98.36%, Alcohol Consumption rate 0.1l) , Qatar (95.64%, 1.29 liters) and Saudi Arabia (82.42% 0.34 liters), where religion-based laws and cultural factors explain the low alcohol consumption rate. Singapore (100% , 1.54 liters) is a bit more complicated: while alcohol consumption is not banned here, the laws are prohibitive and alcohol is expensive.

Uganda, on the other hand, is an example of a poorly urbanized country with a high level of alcohol consumption rate: Its Urban Rate (12.98%) is less than its Alcohol Consumption Rate (16.4 liters). No surprise then that the country has an alcohol overconsumption crisis.

Urban Rate& Alcohol Consumption Categorical to Quantitative Bar Chart (link)

For comparison, I binned the Urban Rate and plotted it against the Alcohol Consumption rate. It backs up the scatter plot: left-skewed and linear with slight drop in the most Urbanized group that include a bunch of countries with prohibitive laws against Alcohol Consumption.

```
PROC IMPORT DATAFILE=“/home/snigdhasen0/NewDataSet2008.csv” DBMS=CSV OUT=imported REPLACE;

run;

Data step1; set imported;

Label urbanrate=“2008 Urban Population (% of total)”

    alcconsumption=“2008 Alcohol Consumption per adult(15+yrs),liters”;

Keep urbanrate alcconsumption urbangroup; /*introducing 1 new variable for bin*/

Label urbangroup=“2008 Urban Population (% of total) Grouped”;

/*urban rate data management*/
```

```
if urbanrate LE 25 and urbanrate gt 0 then urbangroup=“0-25  ”;

if urbanrate le 50 and urbanrate GT 25 then urbangroup=“25-50”;

if urbanrate le 75 and urbanrate GT 50 then urbangroup =“50-75”;

if urbanrate GT 75 then urbangroup=“75-100”;

proc freq; tables urbangroup alcconsumption;

PROC Gchart; vbar urbangroup/discrete type=mean sumvar=alcconsumption;
```

Income per Person (GDP per Capita) &Urbanization (link)

My secondary research question was to explore the relationship between Income and Urbanization. For this, I created a Categorical to Quantitative Bar Chart. Clearly Income Per Person and Urbanization are positively related: The higher the income the more urban the society.

Now that we have established a positive relationship between Urbanization and Alcohol Consumption as well as Income and Urbanization, then higher Income per Capita must also mean higher Alcohol Consumption.

For this, I plotted a Categorical to Quantitative Bar Chart to establish the relationship between **Income per Person (GDP per Capita) & Alcohol Consumption (link).**
As expected the relationship is positive and the graph is skewed to the left, as the left tail of smaller values is longer than the right tail of larger values.

To conclude, our hypothesis seems correct: Higher Urbanization and higher Income were generally related positively to higher Alcohol Consumption in 2008.

WEEK 3: Data Management

WEEK 2: First Program in SAS

WEEK 1: Research Question

May 15th, 2016

MORE YOU MIGHT LIKE

Data Management and Visualization

WEEK 3 (May 8, 2016) Assignment 3 – Making Data Management Decisions in SAS

CODE:

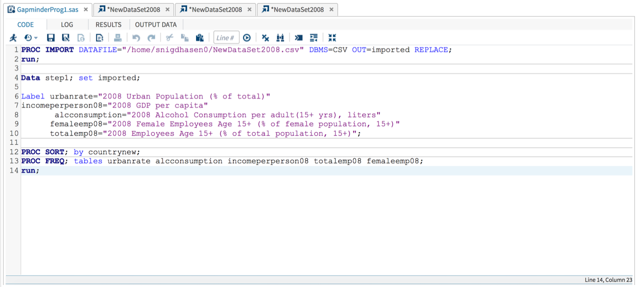
PROC IMPORT DATAFILE="/home/snigdhasen0/NewDataSet2008.csv" DBMS=CSV OUT=imported REPLACE;

run;

Data step1; set imported;

Data Management and Visualization

WEEK 2 (May1, 2016) Assignment 2 – Running your first program in SAS



```
PROC IMPORT DATAFILE="/home/snigdhasen0/NewDataSet2008.csv" DBMS=CSV OUT=imported REPLACE;
run;

Data step1; set imported;

Label urbanrate="2008 Urban Population (% of total)"
incomeperperson08="2008 GDP per capita"
       alcconsumption="2008 Alcohol Consumption per adult(15+ yrs), liters"
      femaleemp08="2008 Female Employees Age 15+ (% of female population, 15+)"
      totalemp08="2008 Employees Age 15+ (% of total population, 15+)";

PROC SORT; by countrynew;
PROC FREQ; tables urbanrate alcconsumption incomeperperson08 totalemp08 femaleemp08;
run;
```

Notes about my data set and the program:

Data Management & Visualization

WEEK 1 (April 24, 2016) Assignment 1 - Developing a Research Question and Creating Your Personal Code Book

DATA source: GapMinder. Having been born and raised in India and now living in the United States, the relative social, cultural, economic and environmental growth and interdependence of countries– and what drives them – have always fascinated me. So I chose to review GapMinder's data provided in this

```
Label urbanrate="2008 Urban Population (% of total)"

incomeperperson08="2008 GDP per capita"

      alcconsumption="2008 Alcohol Consumption per adult(15+ yrs), liters"

    femaleemp08="2008 Female Employees Age 15+ (% of female population, 15+)"

    totalemp08="2008 Employees Age 15+ (% of total population, 15+)";

Keep urbanrate alcconsumption incomeperperson08 totalemp08 urbangroup alcogroup incomegroup empgroup; /*introducing four new variables for bins*/

Label urbangroup="2008 Urban Population (% of total) Grouped"

    alcogroup="2008 Alcohol Consumption per adult(15+ yrs), liters Grouped"

    incomegroup="2008 GDP per capita Grouped"
```

My dataset is a modified version of the GapMinder data provided by the course. I collated and used data from The World Bank's development indicators website to make the data more comparable to the the variables provided in GapMinder. So I had to treat the combined/modified data set as a new one and had to upload and import it to SAS Studio. This threw up a few challenges; I consulted the directions provided by the course, course discussions on this topic as well as online and offline resources on SAS/coding.

RESULTS (Drop Box link to PDF file)

Unlike the discrete variables used as examples in course lectures, the GapMinder data set has continuous variables; it is unclear how or why frequency distribution will help answer the question I am asking about the correlation between two variables. A student on the discussion board and a computer engineer I talked with

course. It covers a year's worth of data for 214 territories.
The data covers 15 indicators, including social, economic, environmental and health indicators.

The Indian economy is morphing ; from a predominantly agricultural economy to a more services-driven, industrial nation. With this transformation comes more education, more female employment, more urbanization, and, significantly, changes in social attitudes.

I often wonder how rapid urbanization will affect the Indian socio-economic fabric, or of any country for that matter. So one of the indicators I chose from the data/codebook is urban rate. Next, I want to examine how other variables relate to rate of urbanization across the globe. So, for example, is there a co-relation between urbanization and alcohol consumption ?

My goal is to investigate the relationship of urbanization with (i) female employment, (ii) alcohol consumption; (iii) alcohol consumption among

```
    empgroup=“2008 Employees Age
15+ (% of total population, 15+)
Grouped”;

/*urban rate data management*/
if urbanrate LE 25 and urbanrate gt 0
then urbangroup=“1.low    ”;
if urbanrate le 50 and urbanrate GT 25
then urbangroup=“2.medium”;
if urbanrate le 75 and urbanrate GT 50
then urbangroup =“3.high”;
if urbanrate GT 75 then
urbangroup=“4.dense ”;

/*alcohol consumption data
management*/
if alcconsumption le 2 and
alcconsumption gt 0 then
alcogroup=“1.light   ”;
if alcconsumption le 6 and
alcconsumption GT 2 then
alcogroup=“2.medium”;
if alcconsumption le 10 and
alcconsumption GT 6 then
alcogroup=“3.high”;
if alcconsumption GT 10 then
alcogroup=“4.heavy”;
```

mentioned grouping and histograms as ways to manage continuous variables; this course has so far not taught us how to do that.

For now, I will look for any pointers the output data may give on my primary research question: how is **Alcohol Consumption** related to **Urbanization** globally in one year (2008)? My hypothesis was the two variables will be related positively.

If we compare the Urban Rate and cumulative percent columns in the first table, we see that 50% or half of the countries included here have urban rates higher than 57%, meaning more than half their populations are urbanized. Six countries (*for names of countries, I referred to my data set*) have 100 percent urban rate: Bermuda, Cayman Island, Hong Kong, Macao China, Monaco and Singapore.

Of these, Alcohol Consumption rates are available only for Singapore, which is 1.54 liters per adult (15 years or older) that year. Data is missing for the other five.

women; (iv) CO2 emissions; (v) and breast cancer, and also, how these variables relate to one another.

But the data provided for this course for breast cancer dates back to 2002; urbanization data is from 2008. Also, alcohol consumption is not divided gender wise. I was not convinced that a credible co-relation could be established with data from years so apart.

I searched the source websites for breast cancer data– the International Agency for Research on Cancer and the World Health Organization – for more recent data: it’s available but in a different format. To collate it into a usable form for this lesson will take more days than this assignment gives me. I intend to pursue that data over a period of time.

So, for this assignment, I have decided to explore the relationship between Urbanization and Alcohol Consumption worldwide for a particular year: **Is alcohol consumption always higher in more Urbanized societies?**

```
/* GDP/income  per person data management*/
if incomeperperson08 le 500 and incomeperperson08 gt 0 then incomegroup=“1.low    ”;
if incomeperperson08 le 5000 and incomeperperson08 gt 500 then incomegroup=“2.medium”;
if incomeperperson08 le 50000 and incomeperperson08 gt 5000 then incomegroup=“3.high”;
if incomeperperson08 gt 50000 then incomegroup=“4.vhigh”;

/*total employment data management*/
if totalemp08 le 55 and totalemp08 gt 0 then empgroup =“1”;
if totalemp08 le 65 and totalemp08 gt 55 then empgroup=“2”;
if totalemp08 le 75 and totalemp08 gt 65 then empgroup =“3”;
if totalemp08 gt 75 then empgroup=“4”;


PROC FREQ; tables urbangroup alcogroup incomegroup empgroup;
run;
```

Notes about my data set, the code and management decisions:

That is less than the median global Alcohol Consumption rate (see second table), which is about 6 liters per adult.

Let’s take a couple more sample points: A country with the median Urban Rate (~50th percentile) and one each below and above the median.

Vanuatu has an Urban Rate of 24.76%, which is in the 10th percentile on the cumulative percent table for Urbanization. It has an Alcohol Consumption rate of 0.96 liters per person, which is in the 14th percentile for Alcohol Consumption (see second table).

Croatia, the country with the median urban rate of 57.28 %, has an Alcohol Consumption rate of 15 liters, which is in the 94th percentile for Alcohol Consumption.

Argentina is 92 % urban, which is in the 91st percentile for Urban Rate. In 2008 Argentinian adults consumed alcohol at the rate of 9.35 liters per adult, placing them in about the 70th percentile for Alcohol Consumption.

I have 2 related topics I want to pursue :

Is Female Employment positively related to Urbanization?

Is Female Employment positively related to Alcohol Consumption?

MY CODE BOOK includes:

Urban rate: 2008 urban population (% of total population)

Alcohol consumption : 2008 average alcohol consumption per adult (age 15+) in liters

Income per person: 2008 Gross Domestic Product (GDP) per capita in constant 2005 U.S. dollars.
[NOTE: *The dataset provided by this course had this variable for the year 2010 in constant 2000 U.S. dollars. The data was sourced from The World Bank’s Work Development Indicators. I will be using the data for 2008, which is readily available from the same source, in constant 2005 U.S. dollars*].

1. My dataset is a modified version of the GapMinder data provided by the course. I collated and used data from The World Bank's development indicators website to make the data more comparable to the variables provided by GapMinder. So I had to treat the combined/modified data set as a new one and had to upload and import it to SAS Studio.

2. Since my data set variables were continuous, I collapsed the data and created 4 "Bins" for each of the 4 variables I chose; the new Bins/variables are: urbangroup alcogroup incomegroup empgroup. I have described the Bins below. I coded out the missing data for each of the Bins.

———

OUTPUT (LINK TO DROPBOX):

My primary research question is: How is **Alcohol Consumption** related to **Urbanization** globally in one year (2008)? My hypothesis is that the two variables are positively related.

That is significantly higher than Vanuatu as expected, but significantly lower than less urbanized Croatia.

Plotting a graph with all variables and/or further analysis will paint a clearer picture of the correlation.

——

Week 1 post. Research Question:

http://seninreverse.tumblr.com/post/143298883105/data-management-visualization

Female Employment Rate: 2008 female employees age 15+ (% of population) ; source The World Bank [*Instead of the 2007 female employees age 15+ (% of population) data provided by GapMinder*]

Employment Rate: 2008 total employees age 125+ (% o f population); Source The World Bank
[Instead of the 2007 total employees age 15+ (% of population) *data provided by GapMinder]*

Literature Review and Hypothesis:

My data sets are for a year that was extraordinary to say the least. The year 2008 saw one of the world's worst financial and economic meltdown: a phenomenal measure of wealth, jobs, houses and lifestyles were wiped out. So the co-relation between the variables I have chosen is one under duress. As a review of a study (referenced below[3]) points out, alcohol consumption may increase in the short term in times of adversity (e.g. financial crisis).

For this I binned each of the 4 variables I chose as described below. I chose the upper and lower limits of the bins in a manner that no bin has too few (<5 %) data points. The interval/progression between bins is constant:

urbangroup (Urban Rate): <=25%= LOW; 25%-50%=MEDIUM; 50%-75%=HIGH; >75%=DENSE

alcogroup (Alcohol Consumption per adult, 15yrs+): <=2 liters =LIGHT; 2l-6l=MEDIUM; 6l-10l=HIGH; >10=HEAVY

incomegroup (GDP per capita in U.S. 2005$): <=$500=LOW; $500-$5,000=MEDIUM; $5000-$50,000=HIGH; >$50,000=V.HIGH

empgroup (Employees, %of total population, 15+): <=55%=1; 55%-65%=2; 65%-75%=3; >75%=4

- In urbangroup, the largest percentage of countries (36.45% or 74 countries) are 50%-75% urbanized.
- Half of the countries in 2008 had a light to medium alcohol

However, since we are comparing the same variables across the globe in a single year, the relationship should still hold. The assumption here is that the downturn had comparable effect on economies worldwide, which may not be the case. Accounting for impact on individual countries is not within the scope of this study.

Studies so far have found a positive co-relation between drinking habits and urbanization in most societies with some exceptions, such as among the population of Greenland Inuit[2], where levels of alcohol consumption couldn’t be related clearly to migration or urbanization.

By and large, more developed nations tend to drink more. Higher incomes and economic development are likely to lead to higher alcohol consumption[1].

A study of a couple dozen countries [5] shows that alcohol consumption is quite strongly associated with urbanization, economic development (GDP), and with the Human Development Index (HDI).

consumption rate of 6 liters or less per adult (15+years). However, a significant number of countries (24%) reported drinking more than 10 liters per adult.

- Citizens of more than half the countries (~56%) earned less than $5,000 per person in 2008.
- Given that empgroup is missing 100 data points, I am considering removing it as a variable in further analysis of my research question.

Previous posts:
Week 2: Running your first program in SAS
Week 1: Research Question

A Swedish study[4] found a similar correlation between urbanization and increased hospital admission rates for alcohol abuse.

My hypothesis is that the positive relationship between alcohol and urbanization is a global phenomenon. There are likely to be some exceptions where social-cultural influences are more deep-rooted, but in most cases, the wealth, easy access to alcohol, relative anonymity and image consciousness associated with urbanization is more likely than not to push up alcohol consumption anywhere. Even in the year that my dataset reflects (2008), I expect to see a strongly positive association in spite of the economic downturn.

References:

1. Summary by David Jernigan on behalf of authors of the book “Alcohol in Developing Societies: A Public Health

Approach”. Alcohol In Developing Societies Summary. World Health Organization

2. Madsen MH, Grønbaek M, Bjerregaard P, Becker U. Urbanization, migration and alcohol use in a population of Greenland Inuit. Source: PubMed

3. M Harvey Brenner, PhD. Trends in Alcohol Consumption and Associated Illnesses. Some Effects of Economic Changes. The American Journal of Public Health December 1975, Vol. 65. No. 12

4. Sundquist, K. and Frank, G. (2004), Urbanization and hospital admission rates for alcohol and drug abuse: a follow-up study of 4.5 million women and men in Sweden. Addiction, 99: 1298–1305. doi: 10.1111/j.1360-0443.2004.00810.x

5. Giora Rahav, Richard Wilsnack, Kim Bloomfield, Gerhard Gmel, Sandra Kuntsche, The Influence of Societal Level Factors on Men’s and Women’s Alcohol Consumption and Alcohol Problems. Alcohol and Alcoholism

Volume 41, Issue suppl 1, 1 October 2006, Oxford University Press Journals,

Show more