DATA 606 Project Proposal

Priya Shaji

3/30/2019

### Data Preparation

# load data  
adolescent\_fertility\_rates <- read.csv("https://raw.githubusercontent.com/PriyaShaji/Adolescent-Pregnancy-Final-Projects/master/Adolecent\_Fertility\_Rates.csv", header=TRUE, check.names = FALSE)  
gendered\_financial\_indicators <- read.csv("https://raw.githubusercontent.com/PriyaShaji/Adolescent-Pregnancy-Final-Projects/master/Gendered\_Financial\_Indicators.csv", header=TRUE, check.names = FALSE)  
gendered\_world\_indicators <- read.csv("https://raw.githubusercontent.com/PriyaShaji/Adolescent-Pregnancy-Final-Projects/master/Gender\_World%20\_Indicators.csv", header=TRUE, check.names = FALSE)

### Research question

**You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.**

By doing this porject , I want to research about the following questions:

1. Which countries have the most significant increasing/decreasing adolescent fertility rates.
2. Determine potential reasons for what in particular makes countries with decreasing or consistent fertility rates different from countries with higher or increasing fertility rates?

### Cases

**What are the cases, and how many are there?**

Each country forms it’s own case and demonstrates rate of adolescent fertility for women aged 15 to 19 years old. There are 261 countries, therefore there are 261 cases.

### Data collection

**Describe the method of data collection.**

The World Bank has an up-to-date (as of 2015) data-set with adolescent fertility rates for 261 countries ranging 45 years. It also has financial indicators for each country broken down by gender.

### Type of study

**What type of study is this (observational/experiment)?**

This is an observational study looking at data from 1960 to 2015 for the most populous countries in the world.

### Data Source

**If you collected the data, state self-collected. If not, provide a citation/link.**

Fertility Data is found here: <https://data.worldbank.org/indicator/SP.ADO.TFRT>

World Development Indicators are found here: <http://wdi.worldbank.org/table/WV.5> <https://data.worldbank.org/topic/gender>

### Dependent Variable

**What is the response variable? Is it quantitative or qualitative?**

The response variable is quantitative value (numerical value) demonstrating a weighted average of births per 1,000 women ages 15-19. This is used to determine fertility rates for adolecent girls.

### Independent Variable

**You should have two independent variables, one quantitative and one qualitative.**

The explanatory variable is world development indicators and are also numerical (some are percentage, one is age, another is binary). Sample variables include: “Life Expectancy”, “% with Account at a Financial Institution”, “% Women in Parliaments”, and “Nondiscrimination clause mentions gender in the constitution”

### Relevant summary statistics

**Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.**

library(tidyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

tidy\_adolescent\_fertility\_rates <- gather(adolescent\_fertility\_rates, "year", "n", 5:60)   
colnames(tidy\_adolescent\_fertility\_rates)[colnames(tidy\_adolescent\_fertility\_rates) == "Country Name"] <- "Country"  
tidy\_adolescent\_fertility\_rates <- select(tidy\_adolescent\_fertility\_rates, one\_of("Country", "year", "n"))  
head(tidy\_adolescent\_fertility\_rates)

## Country year n  
## 1 Aruba 1960 106.2062  
## 2 Afghanistan 1960 145.3210  
## 3 Angola 1960 234.6840  
## 4 Albania 1960 54.4408  
## 5 Andorra 1960 NA  
## 6 Arab World 1960 133.5946

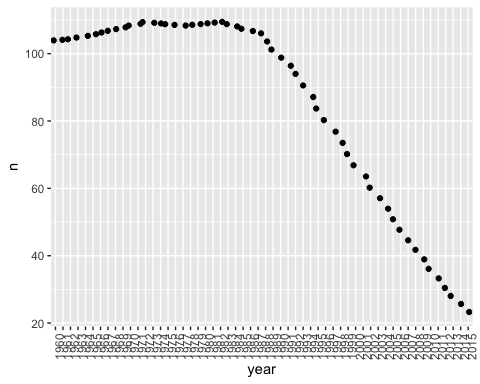
summary for adolescent fertility rate tidy dataset

summary(tidy\_adolescent\_fertility\_rates)

## Country year n   
## Afghanistan : 56 Length:14784 Min. : 0.5222   
## Albania : 56 Class :character 1st Qu.: 34.0233   
## Algeria : 56 Mode :character Median : 66.7703   
## American Samoa: 56 Mean : 77.2695   
## Andorra : 56 3rd Qu.:114.1903   
## Angola : 56 Max. :235.3200   
## (Other) :14448 NA's :1344

India’s Fertility Rates

library(ggplot2)  
  
India\_Fertility <- filter(tidy\_adolescent\_fertility\_rates, Country=="India")  
  
India <- ggplot(India\_Fertility, aes(year, n))  
India + geom\_jitter() + theme(axis.text.x = element\_text(angle = 90, hjust = 1))



World Indicators Exploratory Variables

tidy\_gendered\_world\_indicators <- gather(gendered\_world\_indicators, "year", "n", 5:60)   
colnames(tidy\_gendered\_world\_indicators)[colnames(tidy\_gendered\_world\_indicators) == "Country Name"] <- "Country"  
colnames(tidy\_gendered\_world\_indicators)[colnames(tidy\_gendered\_world\_indicators) == "Indicator Name"] <- "Indicator"  
tidy\_gendered\_world\_indicators <- select(tidy\_gendered\_world\_indicators, one\_of("Country", "Indicator","year", "n"))  
head(tidy\_gendered\_world\_indicators)

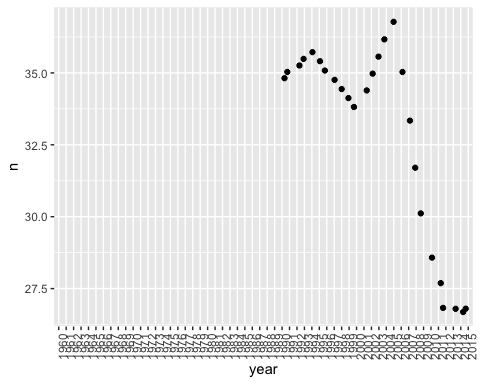
## Country  
## 1 Aruba  
## 2 Aruba  
## 3 Aruba  
## 4 Aruba  
## 5 Aruba  
## 6 Aruba  
## Indicator  
## 1 Mobile account, female (% age 15+)  
## 2 Mobile account, male (% age 15+)  
## 3 Account at a financial institution, female (% age 15+)  
## 4 Account at a financial institution, male (% age 15+)  
## 5 Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)  
## 6 Female headed households (% of households with a female head)  
## year n  
## 1 1960 NA  
## 2 1960 NA  
## 3 1960 NA  
## 4 1960 NA  
## 5 1960 NA  
## 6 1960 NA

India’s Female Labor force Participlation Rate

Sample Exploratory Variable we can use for analysis

India\_Labor\_Force <- filter(tidy\_gendered\_world\_indicators, (Country=="India") & (Indicator == "Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate)"))  
  
India\_Labor\_Force <- ggplot(India\_Labor\_Force, aes(year, n))  
India\_Labor\_Force + geom\_jitter() + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

## Warning: Removed 30 rows containing missing values (geom\_point).



We can make a lot of linear regression comparions between various variables, comparing it to adolescent fertility, and use statistical inference techniques to determine best variables to use.

We can generalize the top 10 countries which have the highest fertility rates and reserach resons for the same using statistical analysis.