Data607-MajorAssignment-Project2-Data Transformation

Vinayak Kamath

3/7/2020

Data Transformation

Below three of the "wide" datasets identified in the Week 6 Discussion items have been used for this exercise.

1. Set 1 - Bank stocks from 2007 (*Discussion Thread by Jeff Shamp*)

2. Set 2 - UNICEF dataset on Under 5 Mortality (*Discussion Thread by Samuel Bellows*)

3. Set 3 - Hospital Consumer Assessment of Healthcare Providers and Systems response data by state (*Discussion Thread by Thomas Hill*)

We will practice tidying and transformations on these data sets.and We will performing few analysis points as discussed/requested in the discussion threads.

Bank Stocks

Reading the CSV file from GIT repository and loading into dataframe:

```
theUrl <- "https://raw.githubusercontent.com/kamathvk1982/Data607-MajorAssignment-Project2/master/banks
banks.full.df <- read.csv(file = theUrl, header = T , sep = ',', na.strings=c("NA","NaN", "") )

# Creating new data frame with reduced columns for current analysis:
banks.df <-  banks.full.df %>%
  select(c(date=Bank.Ticker, date.for.split=Bank.Ticker, bac.close=BAC.3, bac.volume=BAC.4, c.close=C.3
  filter(date!= 'Stock Info' ,  date!= 'Date'  ) %>%
  separate(date.for.split,  c("date.year", "date.month")) %>%
  unite("date.year.month", date.year:date.month, sep='-')

kable(head(banks.df))
```

date	date.year.month	bac.close	bac.volume	c.close	c.volume	jpm.close	jpm.volume	gs.close	gs.volume
2006-01-03	2006-01	47.08	16296700	492.9	1537660	40.19	12839400	128.87	6188700
2006-01-04	2006-01	46.58	17757900	483.8	1871020	39.62	13491800	127.09	4862000
2006-01-05	2006-01	46.64	14970900	486.2	1143160	39.74	8109400	127.04	3717600
2006-01-06	2006-01	46.57	12599800	486.2	1370250	40.02	7966900	128.84	4319600
2006-01-09	2006-01	46.6	15620000	483.9	1680740	40.67	16575200	130.39	4723500
2006-01-10	2006-01	46.21	15634800	485.4	1365960	40.73	16614900	132.03	5539800

```
# Next, we will tidy the data by reshaping the data layput in the table by using tidyr->gather function
banks.tidy.df <- gather(banks.df, key = "key", value = "value", bac.close , bac.volume , c.close , c.vol

banks.tidy.df$value  <- as.numeric(as.character(banks.tidy.df$value))

kable(head(banks.tidy.df))
```

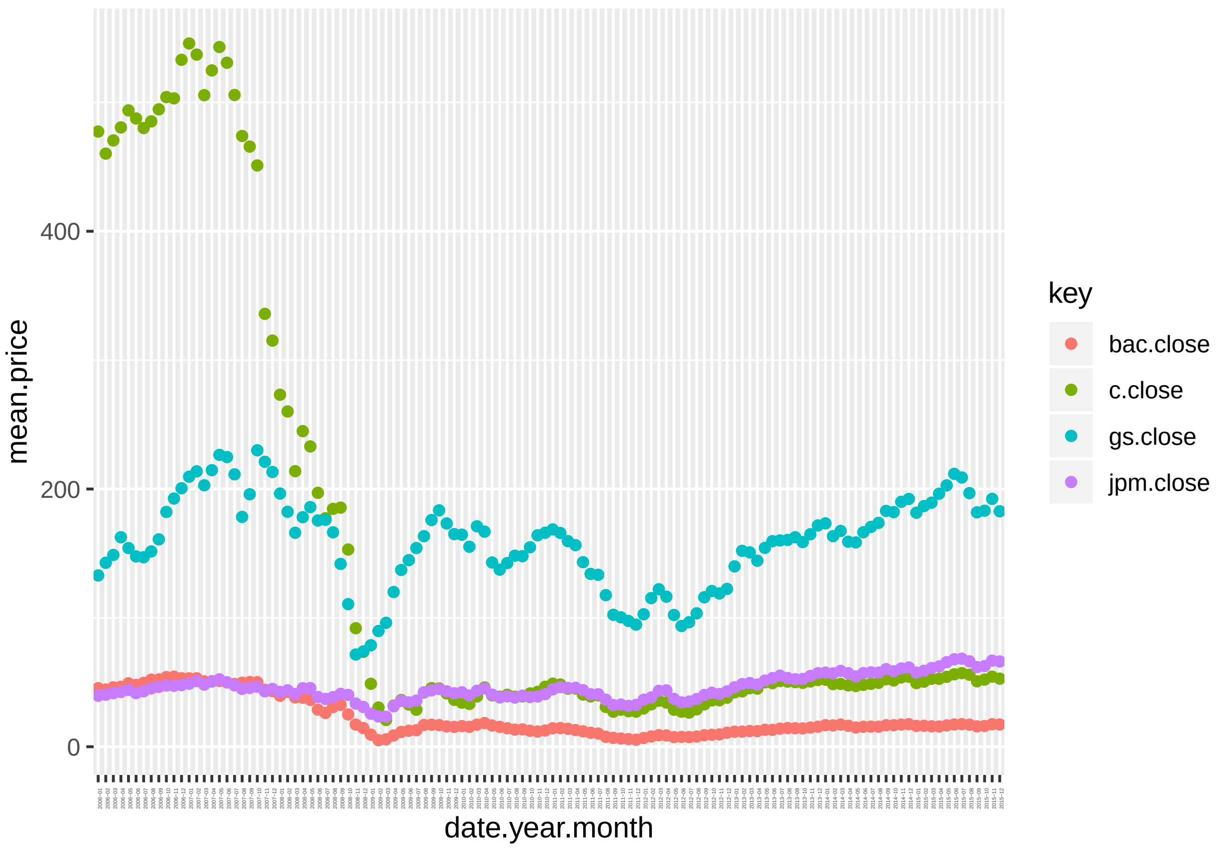
date	date.year.month	key	value
2006-01-03	2006-01	bac.close	47.08
2006-01-04	2006-01	bac.close	46.58
2006-01-05	2006-01	bac.close	46.64
2006-01-06	2006-01	bac.close	46.57
2006-01-09	2006-01	bac.close	46.60
2006-01-10	2006-01	bac.close	46.21

Analysis 1 Piloting the closing balance and volume traded of each bank using point chart:

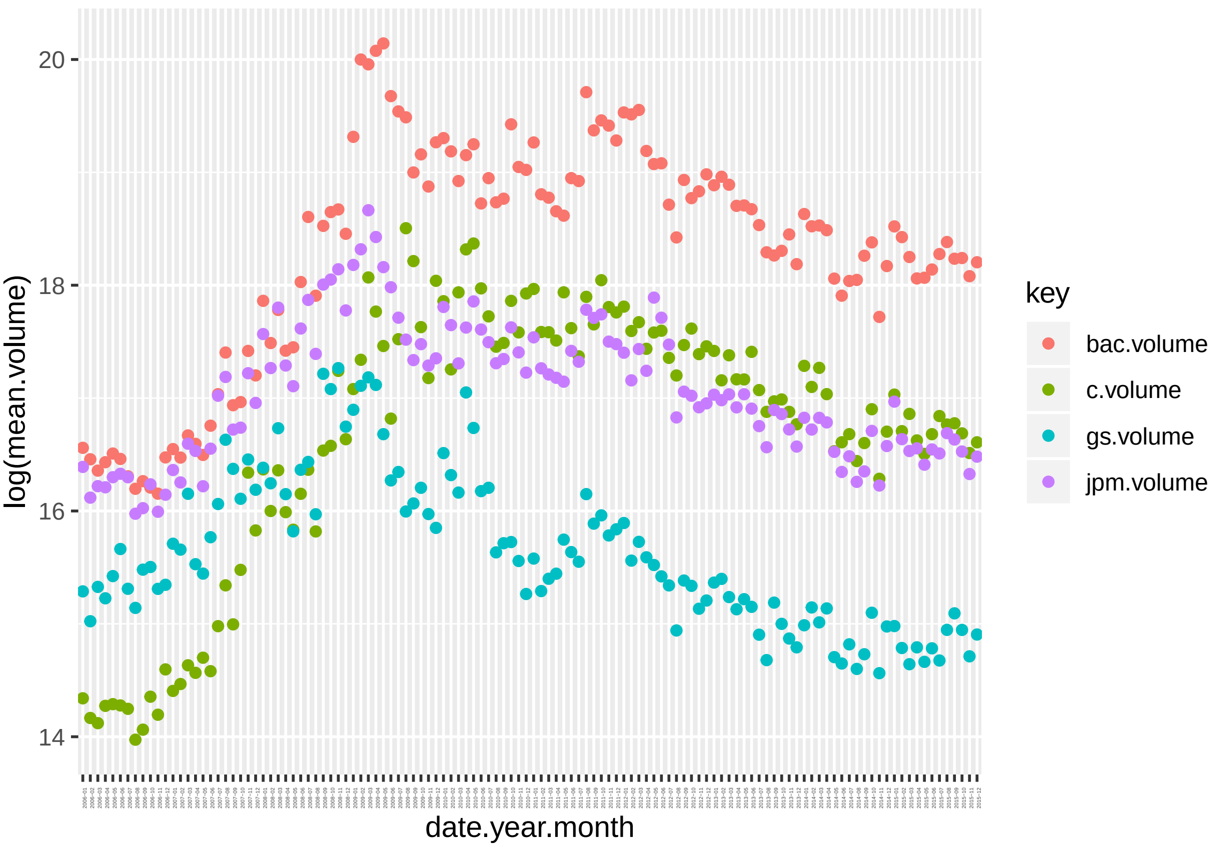
```
plot.price <- banks.tidy.df %>%
  filter(grepl('close', key) ) %>%
   group_by(key,date.year.month) %>%
     summarize(mean.price= mean(as.double(value)))  %>%
       ggplot(aes(x=date.year.month, y=mean.price, colour=key)) +
  theme(axis.text.x = element_text(angle = 90, size = 2)) +
  geom_point()

plot.volume <- banks.tidy.df %>%
  filter(grepl('volume', key) ) %>%
   group_by(key,date.year.month) %>%
     summarize(mean.volume= mean(as.double(value)))  %>%
       ggplot(aes(x=date.year.month, y=log(mean.volume), colour=key)) +
  theme(axis.text.x = element_text(angle = 90, size = 2)) +
  geom_point()

plot.price
```


mean.price
400
200
0
date.year.month
key
bac.close
c.close
gs.close
jpm.close
2006–01
2006–02
2006–03
2006–04
2006–05
2006–06
2006–07
2006–08
2006–09
2006–10
2006–11
2006–12
2007–01
2007–02
2007–03
2007–04
2007–05
2007–06
2007–07
2007–08
2007–09
2007–10
2007–11
2007–12
2008–01
2008–02
2008–03
2008–04
2008–05
2008–06
2008–07
2008–08
2008–09
2008–10
2008–11
2008–12
2009–01
2009–02
2009–03
2009–04
2009–05
2009–06
2009–07
2009–08
2009–09
2009–10
2009–11
2009–12
2010–01
2010–02
2010–03
2010–04
2010–05
2010–06
2010–07
2010–08
2010–09
2010–10
2010–11
2010–12
2011–01
2011–02
2011–03
2011–04
2011–05
2011–06
2011–07
2011–08
2011–09
2011–10
2011–11
2011–12
2012–01
2012–02
2012–03
2012–04
2012–05
2012–06
2012–07
2012–08
2012–09
2012–10
2012–11
2012–12
2013–01
2013–02
2013–03
2013–04
2013–05
2013–06
2013–07
2013–08
2013–09
2013–10
2013–11
2013–12
2014–01
2014–02
2014–03
2014–04
2014–05
2014–06
2014–07
2014–08
2014–09
2014–10
2014–11
2014–12
2015–01
2015–02
2015–03
2015–04
2015–05
2015–06
2015–07
2015–08
2015–09
2015–10
2015–11
2015–12

```
plot.volume
```



Based on the above point chart we can see how the stock prices for banks have been impacted during recession.

Analysis 2 Comparing for Citi and JP Morgan; Getting the mean of the prices for the year 2008; the peak of the recession:

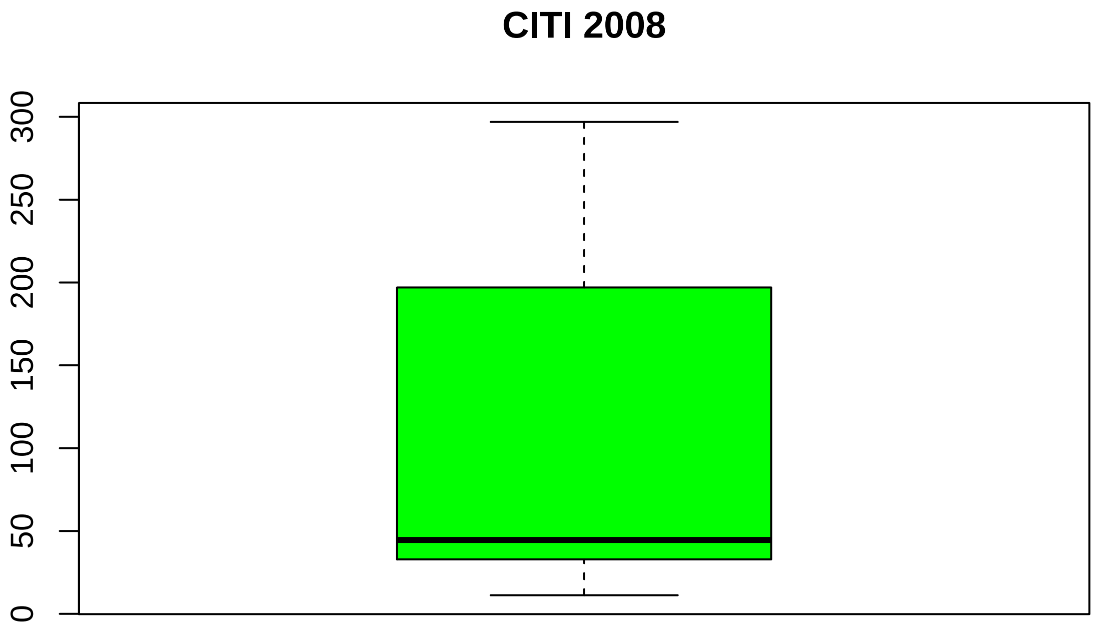
```
data.citi.2008 <- banks.tidy.df %>%
  filter(grepl('c.close', key) , grepl('2008', date.year.month)) %>%
   separate(date.year.month,  c("date.year", "date.month"))

data.jpm.2008 <- banks.tidy.df %>%
  filter(grepl('jpm.close', key) , grepl('2008', date.year.month)) %>%
   separate(date.year.month,  c("date.year", "date.month"))

#52 week data for Citi for 2008
summary(as.double(data.citi.2008$value))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.25   32.91   44.59  110.62  196.97  296.90
```

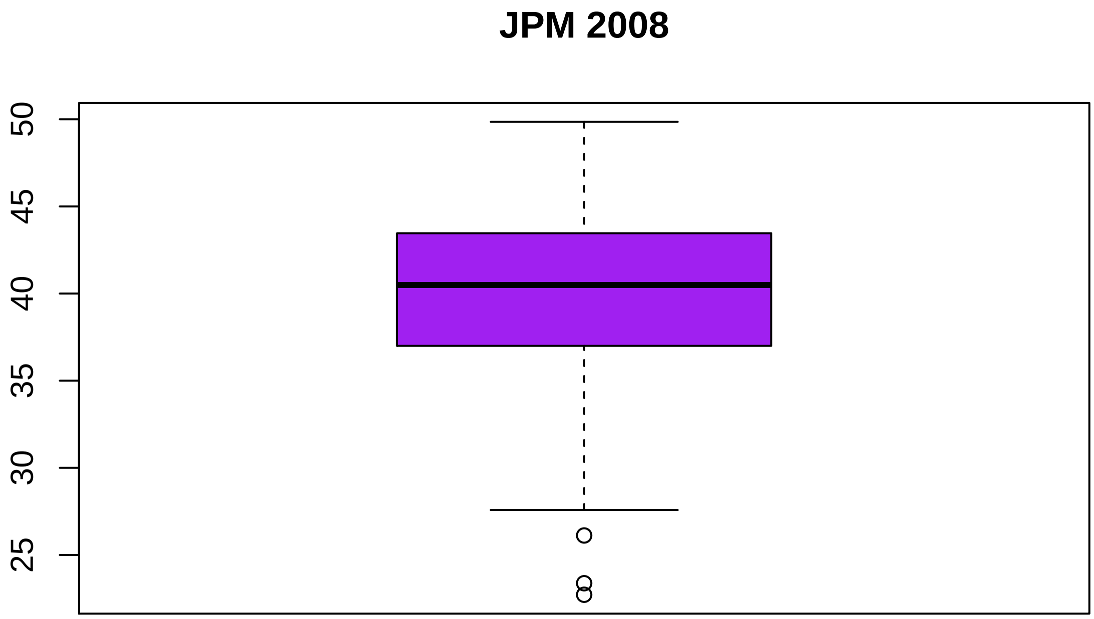
```
boxplot(as.double(data.citi.2008$value), main="CITI 2008", col = "green")
```



```
#52 week data for JPM for 2008
summary(as.double(data.jpm.2008$value))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.72   37.00   40.49   39.83   43.46   49.85
```

```
boxplot(as.double(data.jpm.2008$value), main="JPM 2008", col = "purple")
```



Based on the above box plot shapes we can say that Citi had a much bigger or higher price dip then JPM.

UNICEF

Reading the CSV file from GIT repository and loading into dataframe:

```
theUrl <- "https://raw.githubusercontent.com/kamathvk1982/Data607-MajorAssignment-Project2/master/unice
unicef.full.df <- read.csv(file = theUrl, header = T , sep = ',', na.strings=c("NA","NaN", "") )
dim((unicef.full.df))
```

```
## [1] 196  67
```

Data Transformation and Tidy using dplyr and tidyr:

```
# Next, we will tidy the data by reshaping the data layput in the table by using tidyr->gather function
unicef.tidy.df <- gather(unicef.full.df, key = "Year", value = "Value", -CountryName)
unicef.tidy.df$Value  <- as.numeric(as.character(unicef.tidy.df$Value))
unicef.tidy.df$CountryName  <- str_trim(as.character(unicef.tidy.df$CountryName))

# Use the tidyr->drop_na function to drop the row on column Status having NA value:
unicef.tidy.df <- drop_na(unicef.tidy.df, Value)
```

```
# Use the sub  function to drop the 'U5MR.' from new column Year:
unicef.tidy.df$Year  <- sub('U5MR.','',unicef.tidy.df$Year)

kable(head(unicef.tidy.df))
```

	CountryName	Year	Value
9	Australia	1950	31.6
31	Canada	1950	48.7
48	Benin	1950	348.2
49	Denmark	1950	34.1
51	Dominican Republic	1950	156.0
58	Fiji	1950	135.7

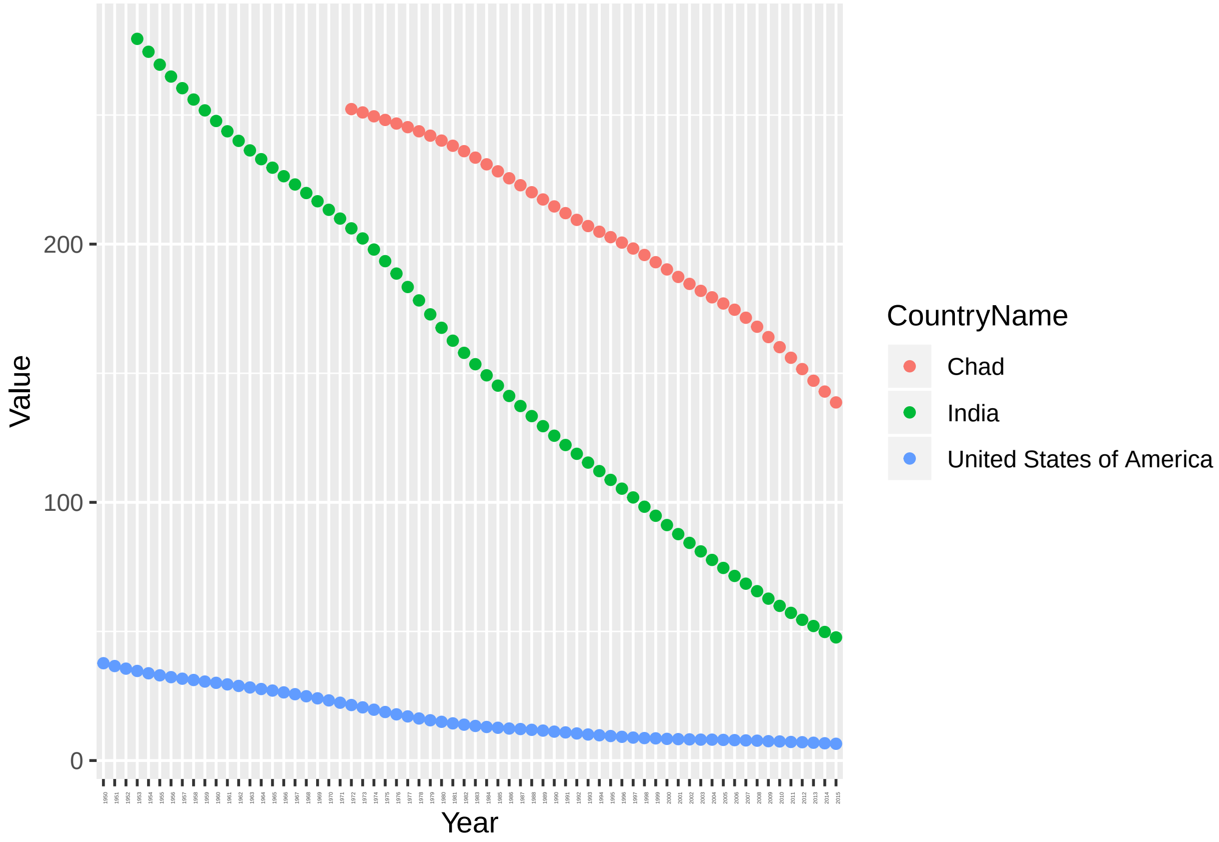
Analysis 1 Comparing a Developed Nation **United States of America** , a Developing nation **India** and a Under Developed Nation **Chad**:

```
unicef.set1.df <-  unicef.tidy.df %>%
  filter(grepl('United States of America|India|Chad' , CountryName  )) %>%
  arrange(Year, CountryName)

kable(head(unicef.set1.df))
```

CountryName	Year	Value
United States of America	1950	37.7
United States of America	1951	36.6
United States of America	1952	35.6
India	1953	279.5
United States of America	1953	34.7
India	1954	274.5

```
unicef.set1.df %>%
       ggplot(aes(x=Year, y=Value, colour=CountryName)) +
  theme(axis.text.x = element_text(angle = 90, size = 2)) +
  geom_point()
```

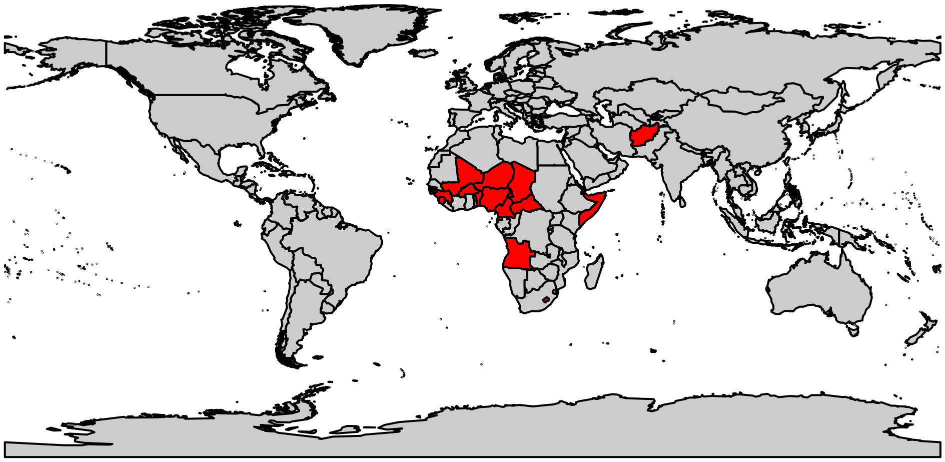


Based on the above Chart, we can see that the under 5 mortality is coming down for all three countries; but comparatively the counts are still high for Under Developed Countries and for Developing Countries .

Analysis 2 List of countries where the under 5 mortality as of 2015 is still greater then 85:

```r
# Filtering for 2015 and greater then 85:
unicef.set2.df <-  unicef.tidy.df %>%
  filter( Year=='2015' ,  Value > 85 ) %>%
  arrange( desc(Value))


# Ploting on world map:
data(wrld_simpl)
myCountries = wrld_simpl@data$NAME %in% names(table(unicef.set2.df$CountryName))
plot(wrld_simpl, col = c(gray(.80), "red")[myCountries+1])
```



Based on the above World Map plotting we can see that most of these countries are in Continent Africa.

Hospital Consumer Assessment

Reading the CSV file from GIT repository and loading into dataframe:

```
theUrl <- "https://raw.githubusercontent.com/kamathvk1982/Data607-MajorAssignment-Project2/master/HCAHPS
hcahps.full.df <- read.csv(file = theUrl, header = T , sep = ',', na.strings=c("NA","NaN", "Not Availabl

hcahps.full.df$HCAHPS.Answer.Percent  <- as.numeric(as.character(hcahps.full.df$HCAHPS.Answer.Percent))

kable(head(hcahps.full.df))
```

State	HCAHPS.Question	HCAHP
AK	Patients who reported that their nurses “Always” communicated well	H_COM
AK	Patients who reported that their nurses “Sometimes” or “Never” communicated well	H_COM
AK	Patients who reported that their nurses “Usually” communicated well	H_COM
AK	Patients who reported that their nurses “Always” treated them with courtesy and respect	H_NUR
AK	Patients who reported that their nurses “Sometimes” or “Never” treated them with courtesy and respect	H_NUR
AK	Patients who reported that their nurses “Usually” treated them with courtesy and respect	H_NUR

Analysis 1 Measure RESPECT (treated patients with courtesy and respect) for Nurses and Doctors for NJ and near by States:

```
# we will create the required dataset using select, filter and separate function for data transformatio
respect.df <-  hcahps.full.df %>%
  select(c(State, Measure.ID = HCAHPS.Measure.ID,Answer.Percent=HCAHPS.Answer.Percent )) %>%
  filter(grepl('CT|NY|PA|NJ' , State    ) ,  grepl('RESPECT' , Measure.ID   ) ) %>%
  separate(Measure.ID,  c("Type", "Response"), sep = '_RESPECT_')


# Next, we will tidy the data by reshaping the data layput in the table by using tidyr->spread function
respect.tidy.df <- spread(respect.df, key = Response, value = Answer.Percent  )
colnames(respect.tidy.df) <- c("State", "Type", "Always", "Sometimes.or.Never", "Usually")

kable(respect.tidy.df)
```

State	Type	Always	Sometimes.or.Never	Usually
CT	H_DOCTOR	84	3	13
CT	H_NURSE	85	4	11
NJ	H_DOCTOR	83	4	13
NJ	H_NURSE	84	4	12
NY	H_DOCTOR	84	4	12
NY	H_NURSE	84	4	12
PA	H_DOCTOR	87	3	10
PA	H_NURSE	88	2	10

```
respect.tidy.df %>% group_by(State, Type) %>% summarise (Positive.Ind = sum(as.integer(Always)+as.integ
```

```
## # A tibble: 8 x 3
## # Groups:   State [4]
##   State Type     Positive.Ind
##   <fct> <chr>           <int>
## 1 CT    H_DOCTOR           97
## 2 CT    H_NURSE            96
## 3 NJ    H_DOCTOR           96
## 4 NJ    H_NURSE            96
## 5 NY    H_DOCTOR           96
## 6 NY    H_NURSE            96
## 7 PA    H_DOCTOR           97
## 8 PA    H_NURSE            98
```

If we treat "Usually" and "Always" as a positive indicator then we can say that in State of CT the Doctors did better and in the State of PA the Nurses did better; for NJ and NY state the Doctors and Nurses were voted equally.

Analysis 2 Measure RATING (a rating of 9 or 10 on a scale from 0 (lowest) to 10 (highest)) for HOSPITALS for NJ and near by States:

```
# we will create the required dataset using select, filter and separate function for data transformatio
rating.df <-  hcahps.full.df %>%
  select(c(State, Measure.ID = HCAHPS.Measure.ID,Answer.Percent=HCAHPS.Answer.Percent )) %>%
  filter(grepl('CT|NY|PA|NJ' , State    ) ,  grepl('RATING' , Measure.ID    ) ) %>%
  separate(Measure.ID,  c("Type", "Response"), sep = '_RATING_')

# Next, we will tidy the data by reshaping the data layput in the table by using tidyr->spread function
rating.tidy.df <- spread(rating.df, key = Response, value = Answer.Percent  )
colnames(rating.tidy.df) <- c("State", "Type", "low.6.or.Lower", "medium.7.or.8", "high.9.or.10")

kable(rating.tidy.df)
```

State	Type	low.6.or.Lower	medium.7.or.8	high.9.or.10
CT	H_HSP	10	20	70
NJ	H_HSP	11	23	66
NY	H_HSP	10	24	66
PA	H_HSP	8	20	72

```
rating.tidy.df %>% group_by(State, Type) %>% summarise (Positive.Ind = sum(as.integer(medium.7.or.8)+as
```

```
## # A tibble: 4 x 3
## # Groups:   State [4]
##   State Type  Positive.Ind
##   <fct> <chr>        <int>
## 1 CT    H_HSP           90
## 2 NJ    H_HSP           89
## 3 NY    H_HSP           90
## 4 PA    H_HSP           92
```

If we treat "rating of 7 or 8 [medium]" and "rating of 9 or 10 [high]" as a positive indicator then we can say that in State of PA the Hospitals did better then other states.

Analysis 3 Measure Doctors and Nurses across applicable questions in all States:

```
# we will create the required dataset using select, filter and separate function for data transformatio
# BY looking at the data values we can say that in column Measure.ID any pattern of DOCTOR or COMP_2 ca
# and any pattern of NURSE or COMP_1 can be treated as NURSE data:

compare.doctor.df <-  hcahps.full.df %>%
  select(c(State, Measure.ID = HCAHPS.Measure.ID,Measure.ID2 = HCAHPS.Measure.ID, Answer.Percent=HCAHPS
  filter(   grepl('DOCTOR|COMP_2' , Measure.ID  ), grepl('_A_P|_U_P' , Measure.ID2  ) )

compare.nurse.df <-  hcahps.full.df %>%
  select(c(State, Measure.ID = HCAHPS.Measure.ID,Measure.ID2 = HCAHPS.Measure.ID, Answer.Percent=HCAHPS
  filter(   grepl('NURSE|COMP_1' , Measure.ID   ), grepl('_A_P|_U_P' , Measure.ID2  ) )
```

Next, we will calculate the sum of the scores:

```
## [1] "Total of Positive Response Score for Doctors is 19842"
```

```
## [1] "Total of Positive Response Score for Nurses is 19934"
```

If we treat “Usually [_U_P]” and “Always [_A_P]” as a positive indicator then we can say that the Nurses had a better result then Doctors across all states