Week 5 homework assignment

Anubha Nagar

29 th September, 2021

DSI-EDA

Professor Michael Shepherd

Note: Be warned, it can take some time to choose a good data set! You might find yourself trying one and then realizing you need to switch to another. This is normal. This week's homework is worth 20 points instead of 10 in brightspace.

Part One

1. Choose a data set from the posted google sheet.

link chosen from google sheet:

https://politicscentre.nuffield.ox.ac.uk/whogov-dataset/

Part Two

1. Write 3-5 sentences telling me about this data.

- What was it used for? Why was it collected? How would you describe the data generating process?
- Who created this data, and why?
- What is the unit of analysis (what are the rows/observations)?

Answer:

The data is the largest available dataset on members of government across time(1996- 2016) and 177 countries. This dataset was created to enable researchers to analyze governance data and draw insights. The data generation process in this case made it easier to see years worth of data on one spreadsheet.

Whogov made this data to make it possible to answer many questions on the male female ratios, understanding the trend of number of ministers etc for each country. The rows in the table is per person who is a part of the governance of a country(i.e. minister, leader etc.)

2. Explore the data. What are the dimensions of the data? What types of variables are there?

load dataset

```
library(readr)
WhoGov_within_V1_2 <- read_csv("WhoGov_within_V1.2.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 226711 Columns: 30
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (19): country_isocode, country_name, position, name, title, gender, dead...
## dbl (11): ...1, year, id, birthyear, core, minister, leader, m_finance, m_de...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(WhoGov_within_V1_2)
```

Answer:

```
dim(WhoGov_within_V1_2)
```

```
## [1] 226711     30
```

```
str(WhoGov_within_V1_2)
```

```
## spec_tbl_df [226,711 x 30] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1                : num [1:226711] 1 2 3 4 5 6 7 8 9 10 ...
##  $ year                : num [1:226711] 1966 1966 1966 1966 1966 ...
##  $ country_isocode     : chr [1:226711] "AFG" "AFG" "AFG" "AFG" ...
##  $ country_name        : chr [1:226711] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ id                  : num [1:226711] 1 2 3 4 5 6 7 8 9 10 ...
##  $ position            : chr [1:226711] "King" "Prime Min." "1st Dep. Prime Min." "2nd Dep. Prime Min
##  $ name                : chr [1:226711] "Mohammed Zahir" "Maiwandwal Mohammed Hashim" "Nur Ahmed Etema
##  $ title               : chr [1:226711] NA NA NA NA ...
##  $ gender              : chr [1:226711] "Male" "Male" "Male" "Male" ...
##  $ birthyear           : num [1:226711] NA NA NA NA NA NA NA NA NA NA ...
##  $ deadyear            : chr [1:226711] NA NA NA NA ...
##  $ party               : chr [1:226711] "independent" "pdpa" "independent" "independent" ...
##  $ party_english       : chr [1:226711] "independent" "Progressive Democratic Party of Afghanistan" "
##  $ party_otherlanguage: chr [1:226711] "independent" NA "independent" "independent" ...
##  $ core                : num [1:226711] 1 1 1 1 1 1 1 1 1 1 ...
##  $ minister            : num [1:226711] 0 0 0 0 1 1 1 1 1 1 ...
##  $ leader              : num [1:226711] 0 1 0 0 0 0 0 0 0 0 ...
```

```
##  $ classification     : chr [1:226711] "Member, Royal Family" "Prime Minister" "Deputy Prime Ministe
##  $ portfolio_1        : chr [1:226711] NA NA NA NA ...
##  $ prestige_1         : chr [1:226711] NA NA NA NA ...
##  $ portfolio_2        : chr [1:226711] NA NA NA NA ...
##  $ prestige_2         : chr [1:226711] NA NA NA NA ...
##  $ portfolio_3        : chr [1:226711] NA NA NA NA ...
##  $ prestige_3         : chr [1:226711] NA NA NA NA ...
##  $ portfolio_4        : chr [1:226711] NA NA NA NA ...
##  $ prestige_4         : chr [1:226711] NA NA NA NA ...
##  $ m_finance          : num [1:226711] 0 0 0 0 0 0 0 0 1 0 ...
##  $ m_defense          : num [1:226711] 0 0 0 0 0 0 0 0 0 0 ...
##  $ m_agriculture      : num [1:226711] 0 0 0 0 1 0 0 0 0 0 ...
##  $ m_foreignaffairs   : num [1:226711] 0 0 0 0 0 0 0 0 0 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_double(),
##   ..   year = col_double(),
##   ..   country_isocode = col_character(),
##   ..   country_name = col_character(),
##   ..   id = col_double(),
##   ..   position = col_character(),
##   ..   name = col_character(),
##   ..   title = col_character(),
##   ..   gender = col_character(),
##   ..   birthyear = col_double(),
##   ..   deadyear = col_character(),
##   ..   party = col_character(),
##   ..   party_english = col_character(),
##   ..   party_otherlanguage = col_character(),
##   ..   core = col_double(),
##   ..   minister = col_double(),
##   ..   leader = col_double(),
##   ..   classification = col_character(),
##   ..   portfolio_1 = col_character(),
##   ..   prestige_1 = col_character(),
##   ..   portfolio_2 = col_character(),
##   ..   prestige_2 = col_character(),
##   ..   portfolio_3 = col_character(),
##   ..   prestige_3 = col_character(),
##   ..   portfolio_4 = col_character(),
##   ..   prestige_4 = col_character(),
##   ..   m_finance = col_double(),
##   ..   m_defense = col_double(),
##   ..   m_agriculture = col_double(),
##   ..   m_foreignaffairs = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

The dataset has 226711 rows and 30 columns. Variables consist of both characters and double as shown in the code output.

Part Three

1. Center your analysis on an exploratory (or "motivating") question. Tell me what it is. Use complete sentences.

I chose to do an analysis on the trend of women in governance globally with a special emphasis on Afghanistan. I want to understand what the trend has been globally for women employment in governance and compare those global trends with employment of women in Afghanistan.

2. Explore your question using visualizations.

- In the end, choose two 'final' plots to 'print' and interpret the visualization in 2-5 sentences. Do not show me tons of plots. Only include your final plots in your report. Create a separate file if needed.
- Why did you choose this graphic? What does it show?

```
head(WhoGov_within_V1_2)
```

```
## # A tibble: 6 x 30
##    ...1  year country_isocode country_name    id position   name    title gender
##   <dbl> <dbl> <chr>           <chr>        <dbl> <chr>      <chr>   <chr> <chr>
## 1     1  1966 AFG             Afghanistan      1 King       Mohamm~ <NA>  Male
## 2     2  1966 AFG             Afghanistan      2 Prime Min. Maiwan~ <NA>  Male
## 3     3  1966 AFG             Afghanistan      3 1st Dep. ~ Nur Ah~ <NA>  Male
## 4     4  1966 AFG             Afghanistan      4 2nd Dep. ~ Shaliz~ <NA>  Male
## 5     5  1966 AFG             Afghanistan      5 Min. Of A~ Mohamm~ <NA>  Male
## 6     6  1966 AFG             Afghanistan      6 Min. Of C~ Ali Nur Dr.   Male
## # ... with 21 more variables: birthyear <dbl>, deadyear <chr>, party <chr>,
## #   party_english <chr>, party_otherlanguage <chr>, core <dbl>, minister <dbl>,
## #   leader <dbl>, classification <chr>, portfolio_1 <chr>, prestige_1 <chr>,
## #   portfolio_2 <chr>, prestige_2 <chr>, portfolio_3 <chr>, prestige_3 <chr>,
## #   portfolio_4 <chr>, prestige_4 <chr>, m_finance <dbl>, m_defense <dbl>,
## #   m_agriculture <dbl>, m_foreignaffairs <dbl>
```

Importing libraries:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

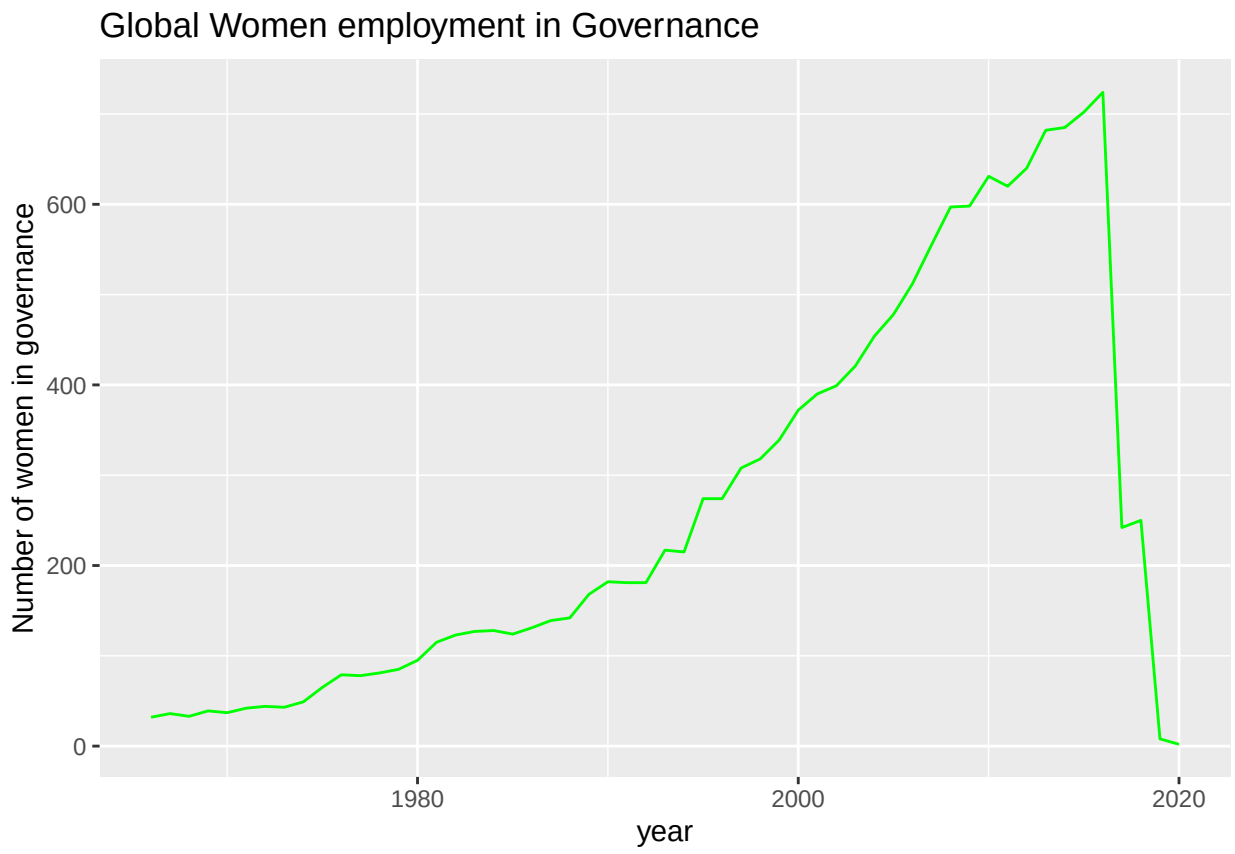
Plot 1:

This plot will help understand the trend of women employment globally in governance

```
women_world<- WhoGov_within_V1_2%>%
  filter(gender == "Female")%>%
  filter(minister == 1)%>%
  group_by(year)%>%
  summarize(n = n())
women_world
```

```
## # A tibble: 55 x 2
##     year     n
##    <dbl> <int>
##  1  1966    32
##  2  1967    36
##  3  1968    33
##  4  1969    39
##  5  1970    37
##  6  1971    42
##  7  1972    44
##  8  1973    43
##  9  1974    49
## 10  1975    65
## # ... with 45 more rows
```

```
ggplot(women_world,aes(x = year, y = n))+ geom_line(color = "green")+ylab("Number of women in governanc
```



Here we can see the employment of women in governance has increased over the years. This steady increase is heartwarming to see that 1965 there were nearly 25 women globally in the government and over 50 years there has been such a huge rise to above 700 women in the government globally. This plot also shows a sudden drop from 2017-2020. This drop is due to insufficient data for the years 2017-2020.

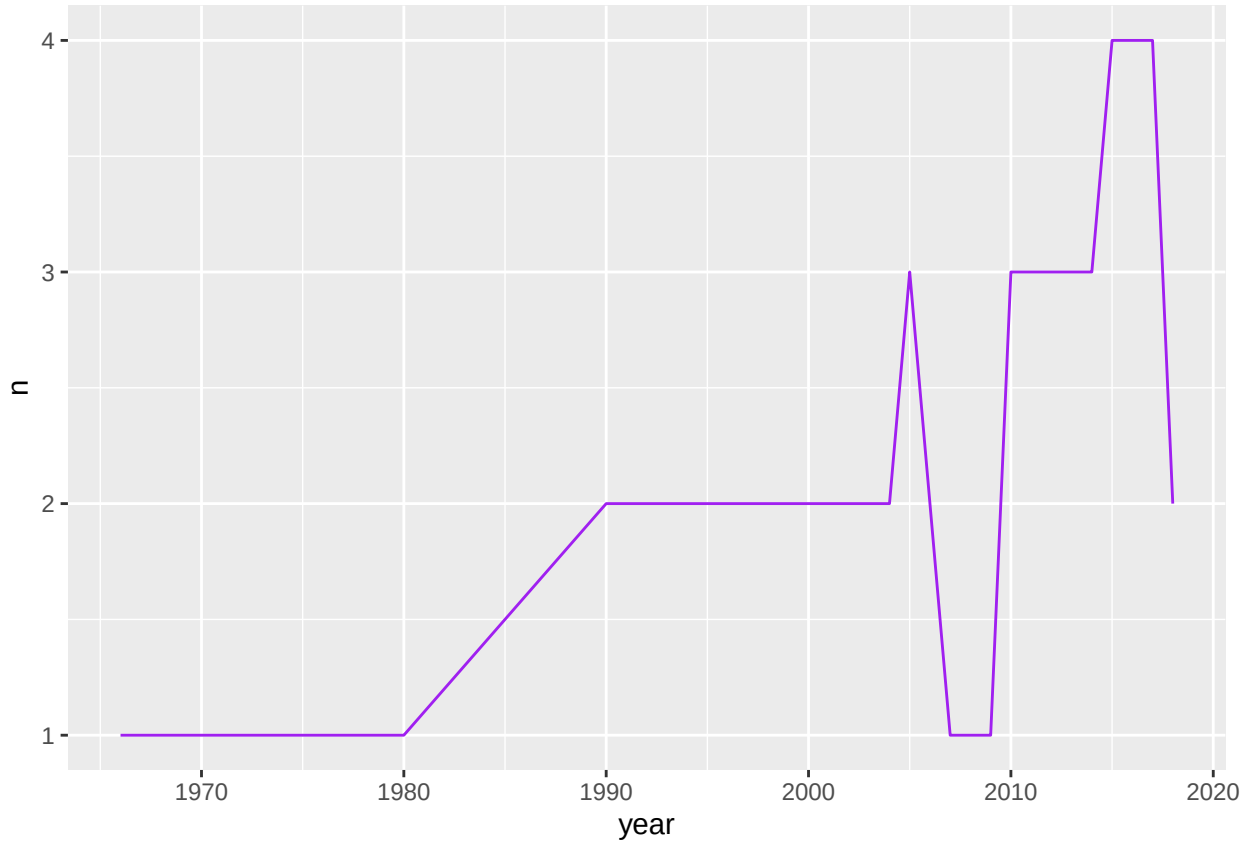
Now let's see the Afghanistan data. Did the number of female ministers increase or decrease over time?

```
b <- WhoGov_within_V1_2%>%
  filter(gender == "Female")%>%
  filter(country_isocode == "AFG")%>%
  filter(minister == 1)%>%
  group_by(year)%>%
  summarize(n = n())
b
```

```
## # A tibble: 25 x 2
##     year     n
##    <dbl> <int>
##  1  1966     1
##  2  1967     1
##  3  1968     1
##  4  1969     1
##  5  1970     1
##  6  1971     1
##  7  1972     1
##  8  1980     1
##  9  1990     2
```

```
## 10  1991     2
## # ... with 15 more rows
```

```
ggplot(b,aes(x = year, y = n))+ geom_line(color = "purple")
```



As we can see that the number of female governance was 0 from 1965- 1980 and it slowly started increasing, but in 2009 again it came down to 1. Even though there was a significant increase in women over the years 1980-2005, the highest number of women at that time went up to only 3, which is very less. In 2015-2016 the number of women working as ministers remained constant. Beyond 2016 there is not enough information to determine a trend. The alarming thing is there is no trend, some years we see a rise but then there are steep falls in value.

These alarming numbers got me wanting to know the percentage of men vs women in the Afghanistan Government!

Percentage of males and females in Afghanistan government:

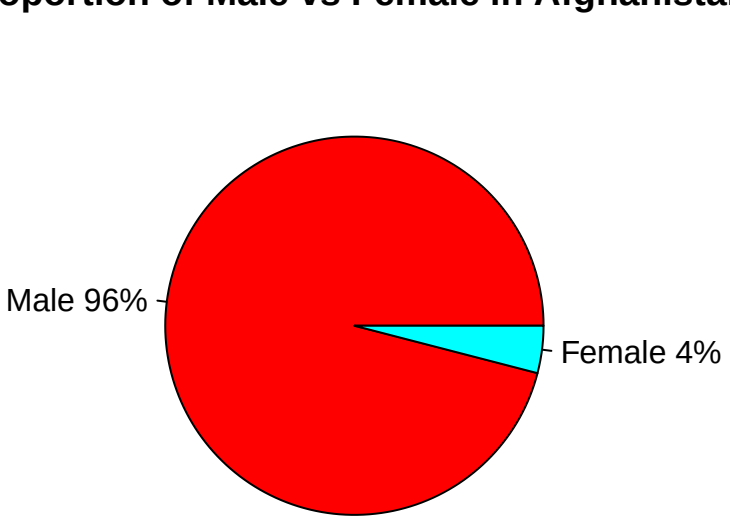
```
male_off <- WhoGov_within_V1_2%>%
  filter(gender == "Male")%>%
  filter(country_isocode == "AFG")%>%
  summarize(n= n())
head(male_off)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  1308
```

```r
female_off <- WhoGov_within_V1_2%>%
  filter(gender == "Female")%>%
  filter(country_isocode == "AFG")%>%
  summarize(n = n())
head(female_off)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    55
```

```r
slices <- c(1308,55)
lbls <- c("Male", "Female")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
   main="Proportion of Male vs Female in Afghanistan")
```



Only 4 % of the government is female, which is such a low number!

As an extension of this analysis, I wanted to check how many women in Afghanistan have taken the role of defense minister,finance minister, foreign affairs and agriculture minister?

```
WhoGov_within_V1_2%>%
  filter(country_isocode == "AFG")%>%
  filter(gender == "Female")%>%
  filter(m_defense == 1)
```

```
## # A tibble: 0 x 30
## # ... with 30 variables: ...1 <dbl>, year <dbl>, country_isocode <chr>,
## #   country_name <chr>, id <dbl>, position <chr>, name <chr>, title <chr>,
## #   gender <chr>, birthyear <dbl>, deadyear <chr>, party <chr>,
## #   party_english <chr>, party_otherlanguage <chr>, core <dbl>, minister <dbl>,
## #   leader <dbl>, classification <chr>, portfolio_1 <chr>, prestige_1 <chr>,
## #   portfolio_2 <chr>, prestige_2 <chr>, portfolio_3 <chr>, prestige_3 <chr>,
## #   portfolio_4 <chr>, prestige_4 <chr>, m_finance <dbl>, m_defense <dbl>, ...
```

```
WhoGov_within_V1_2%>%
  filter(country_isocode == "AFG")%>%
  filter(gender == "Female")%>%
  filter(m_finance == 1)
```

```
## # A tibble: 0 x 30
## # ... with 30 variables: ...1 <dbl>, year <dbl>, country_isocode <chr>,
## #   country_name <chr>, id <dbl>, position <chr>, name <chr>, title <chr>,
## #   gender <chr>, birthyear <dbl>, deadyear <chr>, party <chr>,
## #   party_english <chr>, party_otherlanguage <chr>, core <dbl>, minister <dbl>,
## #   leader <dbl>, classification <chr>, portfolio_1 <chr>, prestige_1 <chr>,
## #   portfolio_2 <chr>, prestige_2 <chr>, portfolio_3 <chr>, prestige_3 <chr>,
## #   portfolio_4 <chr>, prestige_4 <chr>, m_finance <dbl>, m_defense <dbl>, ...
```

```
WhoGov_within_V1_2%>%
  filter(country_isocode == "AFG")%>%
  filter(gender == "Female")%>%
  filter(m_agriculture == 1)
```

```
## # A tibble: 0 x 30
## # ... with 30 variables: ...1 <dbl>, year <dbl>, country_isocode <chr>,
## #   country_name <chr>, id <dbl>, position <chr>, name <chr>, title <chr>,
## #   gender <chr>, birthyear <dbl>, deadyear <chr>, party <chr>,
## #   party_english <chr>, party_otherlanguage <chr>, core <dbl>, minister <dbl>,
## #   leader <dbl>, classification <chr>, portfolio_1 <chr>, prestige_1 <chr>,
## #   portfolio_2 <chr>, prestige_2 <chr>, portfolio_3 <chr>, prestige_3 <chr>,
## #   portfolio_4 <chr>, prestige_4 <chr>, m_finance <dbl>, m_defense <dbl>, ...
```

```
WhoGov_within_V1_2%>%
  filter(country_isocode == "AFG")%>%
  filter(gender == "Female")%>%
  filter(m_foreignaffairs == 1)
```

```
## # A tibble: 0 x 30
## # ... with 30 variables: ...1 <dbl>, year <dbl>, country_isocode <chr>,
## #   country_name <chr>, id <dbl>, position <chr>, name <chr>, title <chr>,
## #   gender <chr>, birthyear <dbl>, deadyear <chr>, party <chr>,
```

```
## #   party_english <chr>, party_otherlanguage <chr>, core <dbl>, minister <dbl>,
## #   leader <dbl>, classification <chr>, portfolio_1 <chr>, prestige_1 <chr>,
## #   portfolio_2 <chr>, prestige_2 <chr>, portfolio_3 <chr>, prestige_3 <chr>,
## #   portfolio_4 <chr>, prestige_4 <chr>, m_finance <dbl>, m_defense <dbl>, ...
```

It is shocking to note that there are no women over all the years(till 2016) to have been a defense, finance, agricultural and foreign affairs minister.

Part Four

1. Answer the following questions:

- how long did it take you to 'choose' this data?

It took me almost a long time to find the correct dataset. I toggled with many datasets to find this dataset. This was one of the hardest part of the assignment.

- how much data wrangling, joining, subsetting, or other forms of data manipulation (such as variable creation, or relabeling) did you have to do? Why did you do it?

I had to do lots of data wrangling to get the data I wanted since everything was so deeply embedded in the data. Also the fact that data was so large, plotting only a specific portion of it required lots of wrangling.

- what did you learn from your data exploration?

I understood many things:

- General Thumb Rules of Analysis in the Real world:
  - The data will be messy and huge and we need to find a way to extract what we want.
  - Understanding the dataset is very important before jumping into solving the question.
  - Drawing inferences from the graphs are as important as drawing the graph itself.
- From the graphs I made I realized:
  - The trend of women in governance has increased globally but in countries such as Afghanistan, the trend is very hard to say (the data may increase gradually but then reduces drastically). Moreover we realize how such less women go into government in Afghanistan. There has never been a finance, defense, agricultural or foreign affairs minister (till 2016), which is super alarming!