Exploring the extent of gender inequality across the world

Aashna Kanuga, Karan Sindwani, Neha Saraf, Raj Biswas

12/9/2018

Introduction

In the past few years, the discussion about gender inequality has been on the rise. Women everywhere are starting to speak up and demand equal rights in all aspects of life. People everywhere - men and women, are involved in making this happen, and a lot of people have been saying that slowly but surely, the situation is improving. But is it truly so? The change that we are seeing, is it consistent throughout the world? Has there been an actual significant improvement over years in multiple verticals of life? Are women really appreciated for the work they do? Do the government and legislation support this fight against gender inequality? These were some of the questions that we were seeking the answers to, and that is why we chose this topic.

A list of team members working on this project, along with their contributions:
1. Aashna Kanuga - Exploratory Data Analysis: Development and Education, and the final report.
2. Karan Sindwani - Exploratory Data Analysis: Government, and the interactive component.
3. Neha Saraf - Exploratory Data Analysis: Education, and the interactive component.
4. Raj Biswas - Exploratory Data Analysis: Employment and Government, and the final report.

Description of Data

Originally, the data we had selected was from the source of a Buzzfeed Article about sexual harassment in the workplace over twenty years. However, after the initial exploration of the data, we found that the data was too small, and a lot of it was overlapping and showing very similar information. We could not generate very useful visualizations from it - that was not already explored in the original article itself - so we decided to look for other datasets. Then we came across a dataset from the Organisation for Economic Co-operation and Development (OECD) about gender inequality across various verticals. It is accessible from the link provided and can be downloaded as distinct or combined CSV files.

We decided to include the following verticals in our analysis:
1. Education: This vertical compares the performance of men and women in terms of education, and also looks at their participation and contribution in the education field. 2. Employment: This vertical looks at how many men and women are in paid work, who works full-time, and how having children and growing older affect women's work patterns and earnings differently to men's.
3. Development: This vertical includes data about the various discriminatory social institutions against women, like violence and harassment, restricted freedom, bias against women and laws across various countries.
4. Governance: This vertical looks at how well women are represented across all aspects of governance.

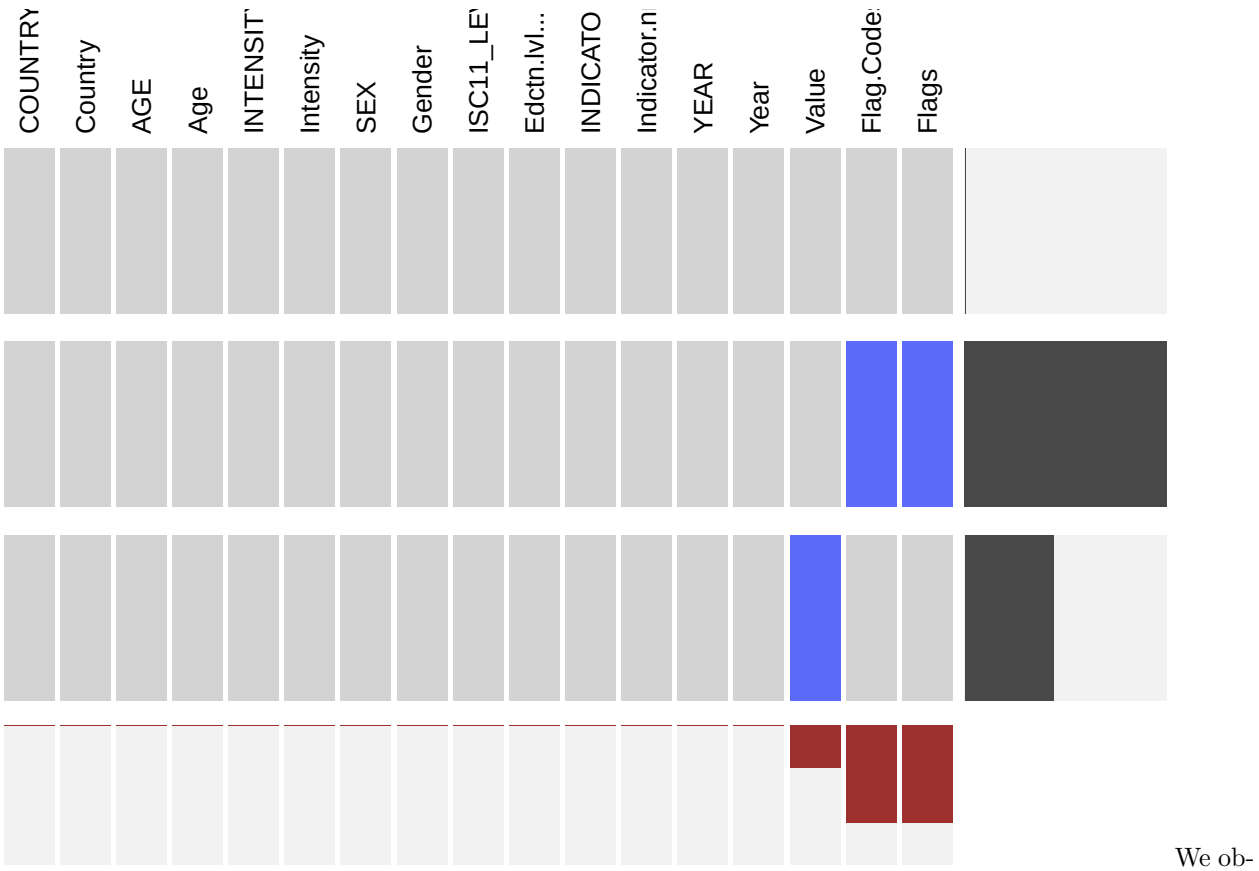
In our analysis, we discover patterns for all variables within each vertical, highlight the gender inequality, and draw conclusions about the gender discrimination for multiple countries across the world.

Analysis of Data Quality

The data itself was not very messy, because it has been used to create some visualization previously. All of the data was in a tidy format, which was very helpful to us. But the few issues that we faced in terms of data quality are:

1. Missing Values: The education sector had a lot of missing values in the files, which we explored using the *visna* function to observe the missing data patterns.
   Here is an example:

```
visna(ed_data_two)
```



We observed that all missing data was contained either in the "Flags" column (which is not to be used in our analysis) and in the "Value" column, which is the primary column we will use in our data analysis. So for all the files in the education sector, we pass the data through a function to eliminate all missing values. Note: We chose not to fill missing values with some data that we compute, say mean or zero value, because that substitution leads to very misleading visualizations, because substituting some fixed value might not represent the actual proportion of the value for that particular gender in some country.

2. Dependency of "Value" column: In each file of every vertical, our dependent column is "Value". But in most of the files, the major issue we faced was that the column was dependent on too many factors, which made it difficult to visualize it, because after a point you cannot simply create one graph for every dependency. So, to overcome this issue, we found the most significant columns on which Value depends (by simply drawing histograms to look at the distribution), and then we grouped our data on the other columns and obtained an average observation, which we could then facet over the significant columns.

3. Very descriptive observations: A lot of our columns had observations that had text that was too long, and unnecessary to show to a user, for which we had to recode some of our columns to include shorter descriptions which are easier to understand.

4. Incomplete data for "Country" and "Time": Although the data did not have too many 'NA's in particular fields, there were entire rows missing which corresponded to data for many tables for different countries for different years. The four verticals that we chose had data for a different number of countries over different timelines, and we chose to work with data that was available, this can be seen in a few of

the heatmaps where some blocks are missing due to the above mentioned reason.

5. Overlap of Data between "Employment" and "Government" verticals: The employment and government verticals had a few tables in common such as wage gap, employment and unemployment rates, etc. We ensured that we picked tables that convey different information and provide different insights into gender inequality in the respective verticals.
6. Duplicate Values: Another issue we faced was that some of the files had many repeating values, because of a glitch in how we could download the data from our source. So we had to make sure that we filter our all repeating values and take only unique ones. For example, in the file "Early_Marriage.csv" we originally had 640 observations (which is not possible because the total countries in our data were lower). The repeating values are due to the same values repeating under specific regions, as well as under the category of "All regions" and also due to multiple income levels category, all of which actually have the same value. So once we subset our data accordingly, we get unique values.
7. Not all countries covered: Our data is also incomplete in the sense that we do not have all values for each and every country in the world, which means our findings cannot be easily generalized for the entire world.

Main Analysis

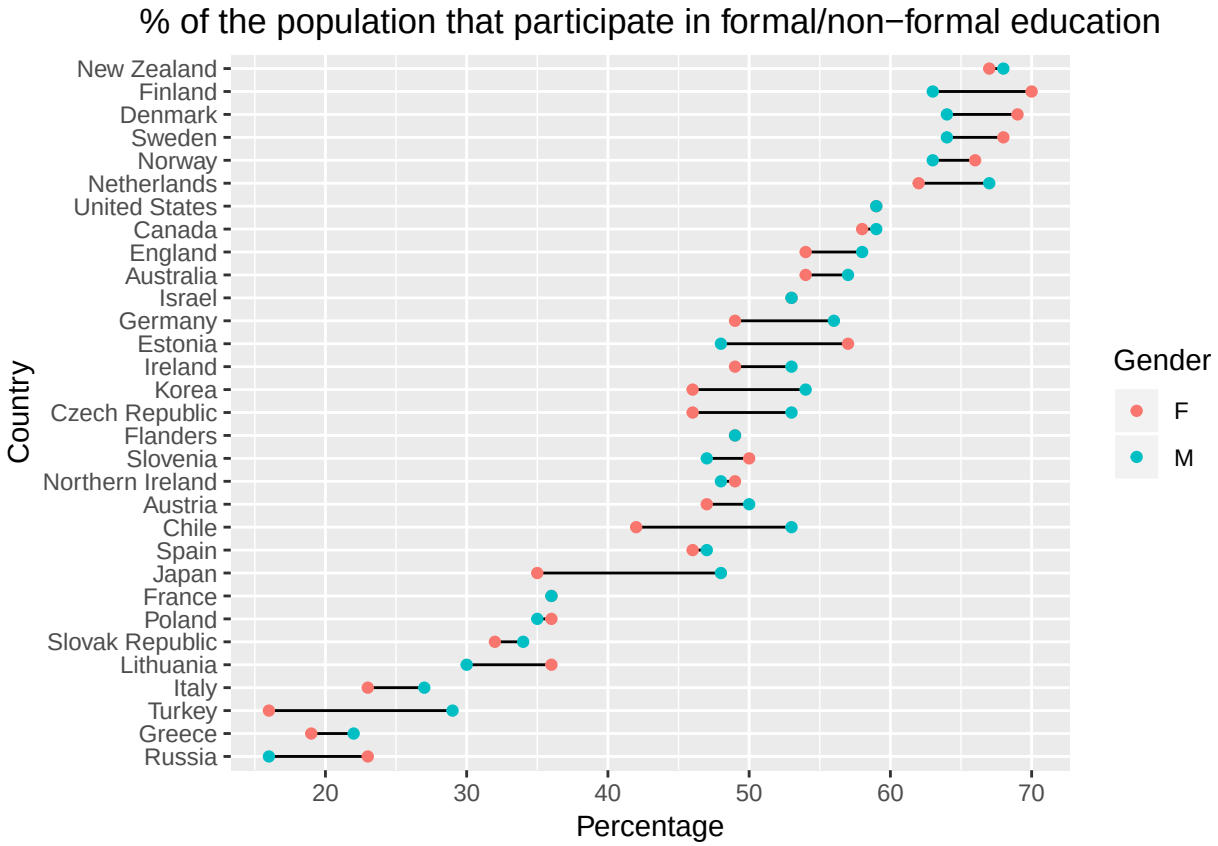
We performed exploratory data analysis on the four mentioned verticals. As a part of the analysis we plotted many graphs and charts for each of the tables according to the type of the data and the number of dimensions to be analyzed. Certain plots convey a lot of information and interesting insights into the data whereas a few of them do not show any patterns. We have included more visualizations that provide insights and a few examples of visualizations that fail to do so.
Link to all the code: https://github.com/kskaran94/EDAV_Project_Fall_2018

The detailed analysis of each of the verticals are as follows -

1. EDUCATION

a) Adult Education and Learning: This indicator presents internationally comparable data on participation in adult learning activities (formal and/or non-formal education). The data for this variable is available for the year 2015. On plotting a Cleveland dot plot for this indicator with the datapoints for women and men in pink and blue respectively, we get the following visualization -

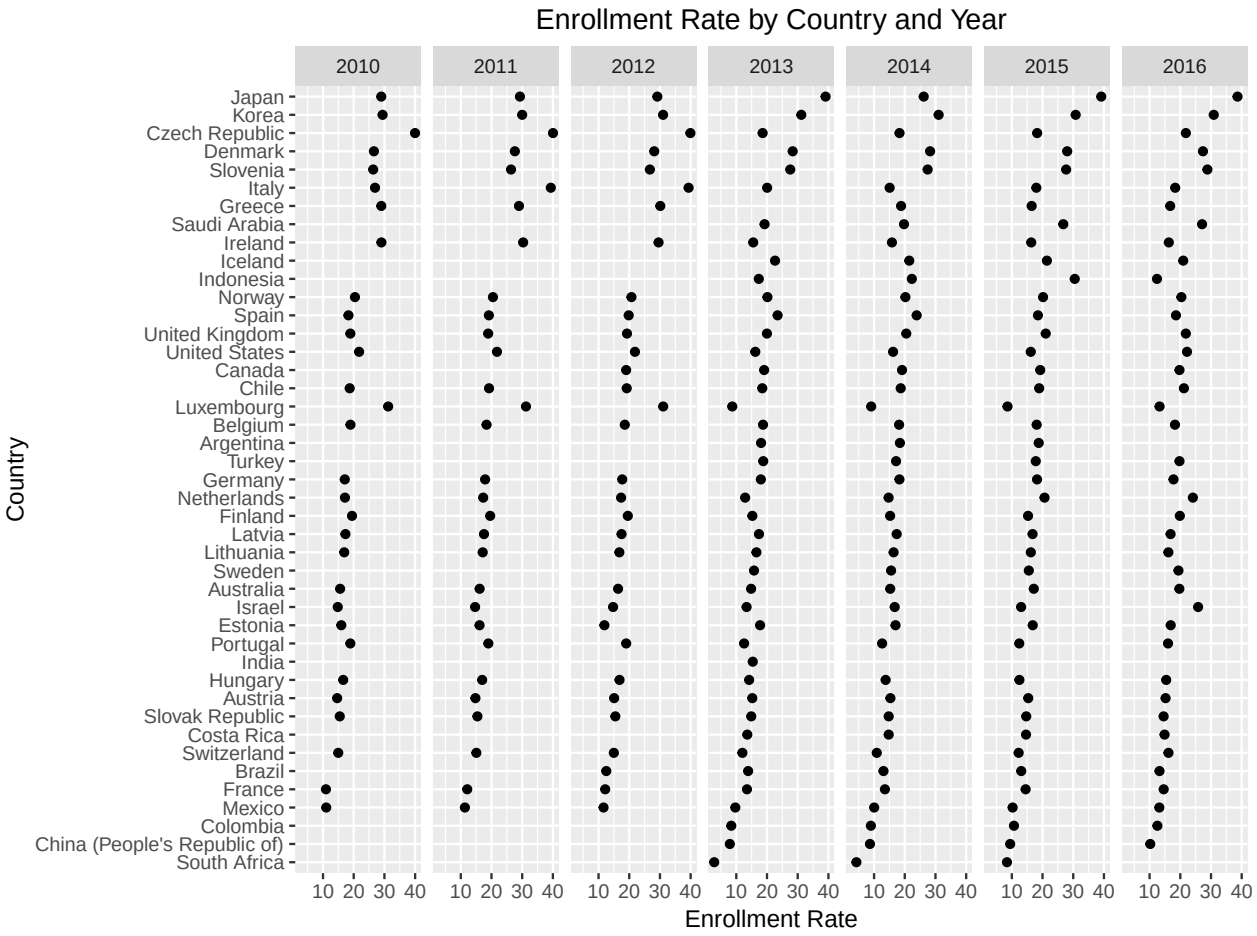
```
ggplot(ed_data_one, aes(y = reorder(Country, Value),x = Value))+
  geom_line(aes(group = Country))+
  geom_point(aes(color = factor(SEX)))+
  labs(x="Percentage", y="Country", color = "Gender")+
  ggtitle("% of the population that participate in formal/non-formal education")+
  theme(plot.title = element_text(hjust = 0.5))
```



We can see that for most countries, more number of men participate in education and learning activities compared to women. Also, in countries with higher population participation in education and learning like Finland, Denmark and Sweden, more number of women participate in education and learning activities compared to men.

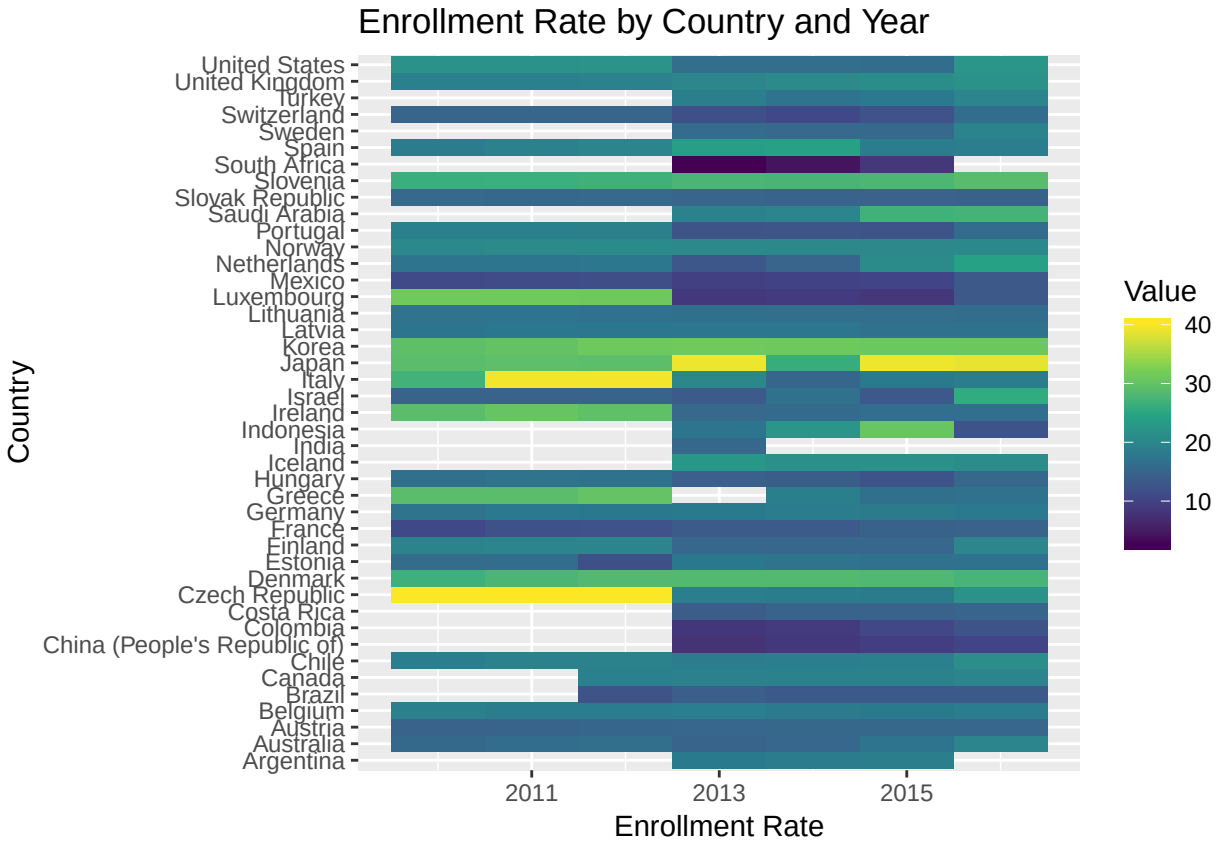
b) Enrollment Rate: Enrolment rate is the percentage of students enrolled in each type of institution over the total of students. The data for this indicator is available from the year 2010 to 2016. On plotting a cleveland plot for the enrollment rate faceted by year and color based on sex, we saw that the enrollment rate for men and women was very similar. Hence we dropped the color based on sex and obtained the following plot -

```
# Cleaning the data by removing NA from Value column
cleaned_data_two <- completeFun(ed_data_two, "Value")
cleaned_data_two <- cleaned_data_two %>%
                    filter(Value > 0)

country_data_year_two <- cleaned_data_two %>%
  group_by(Country, Year) %>%
  summarise(Value = mean(Value))

ggplot(country_data_year_two, aes(y = reorder(Country, Value),x = Value)) +
  geom_point() +
  ggtitle("Enrollment Rate by Country and Year") +
  labs(x = "Enrollment Rate", y = "Country") +
  facet_grid(. ~ Year) +
  theme(plot.title = element_text(hjust = 0.5))
```



The enrollment rate has been more or less the same over the 2010 - 2013 period and then has improved slightly after that. Countries like Japan, Korea and Slovenia have high enrollment rates in general whereas China, South Africa, Colombia and Mexico have very low enrollment rates. We also plotted a heatmap for the same indicator which helps us see the pattern of change of enrollment rates in a better way.

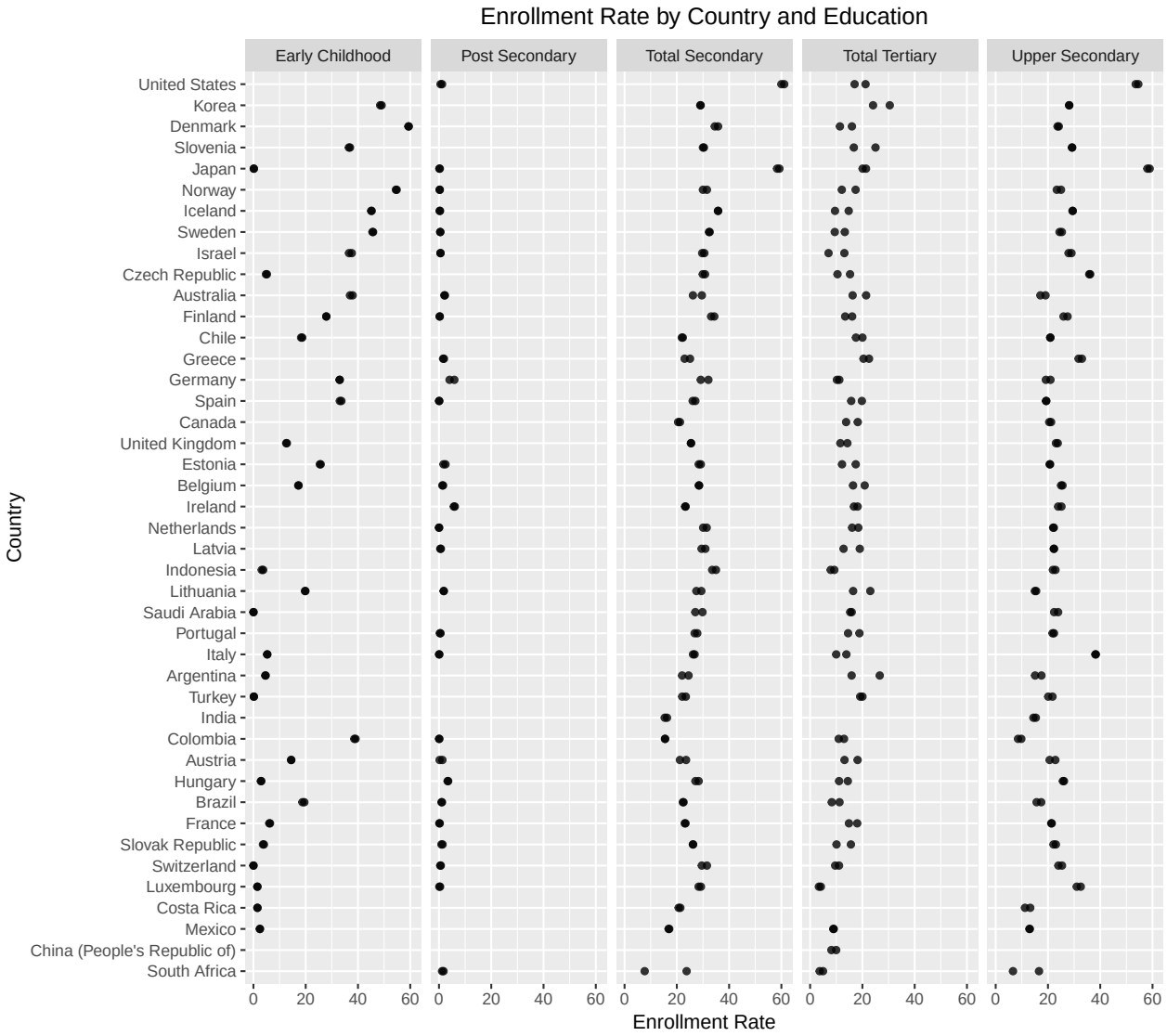
```
ggplot(country_data_year_two, aes(y = Country,x = Year)) +
  geom_tile(aes(fill=Value)) +
  scale_fill_viridis_c(direction = 1) +
  ggtitle("Enrollment Rate by Country and Year") +
  labs(x = "Enrollment Rate", y = "Country")
```



We also plotted the enrollment rates faceted on age groups and obtained the plot below -

```
country_data_education_two <- cleaned_data_two %>%
  group_by(Country, `Education level and programe orientation`, Gender) %>%
  summarise(Value = mean(Value))
```

```
country_data_education_two <- country_data_education_two[country_data_education_two$education_level_sho
ggplot(country_data_education_two, aes(y = reorder(Country, Value), x = Value)) +
  geom_point(alpha = 0.8) +
  ggtitle("Enrollment Rate by Country and Education") +
  labs(x = "Enrollment Rate", y = "Country") +
  facet_grid(. ~ education_level_short) +
  theme(plot.title = element_text(hjust = 0.5))
```

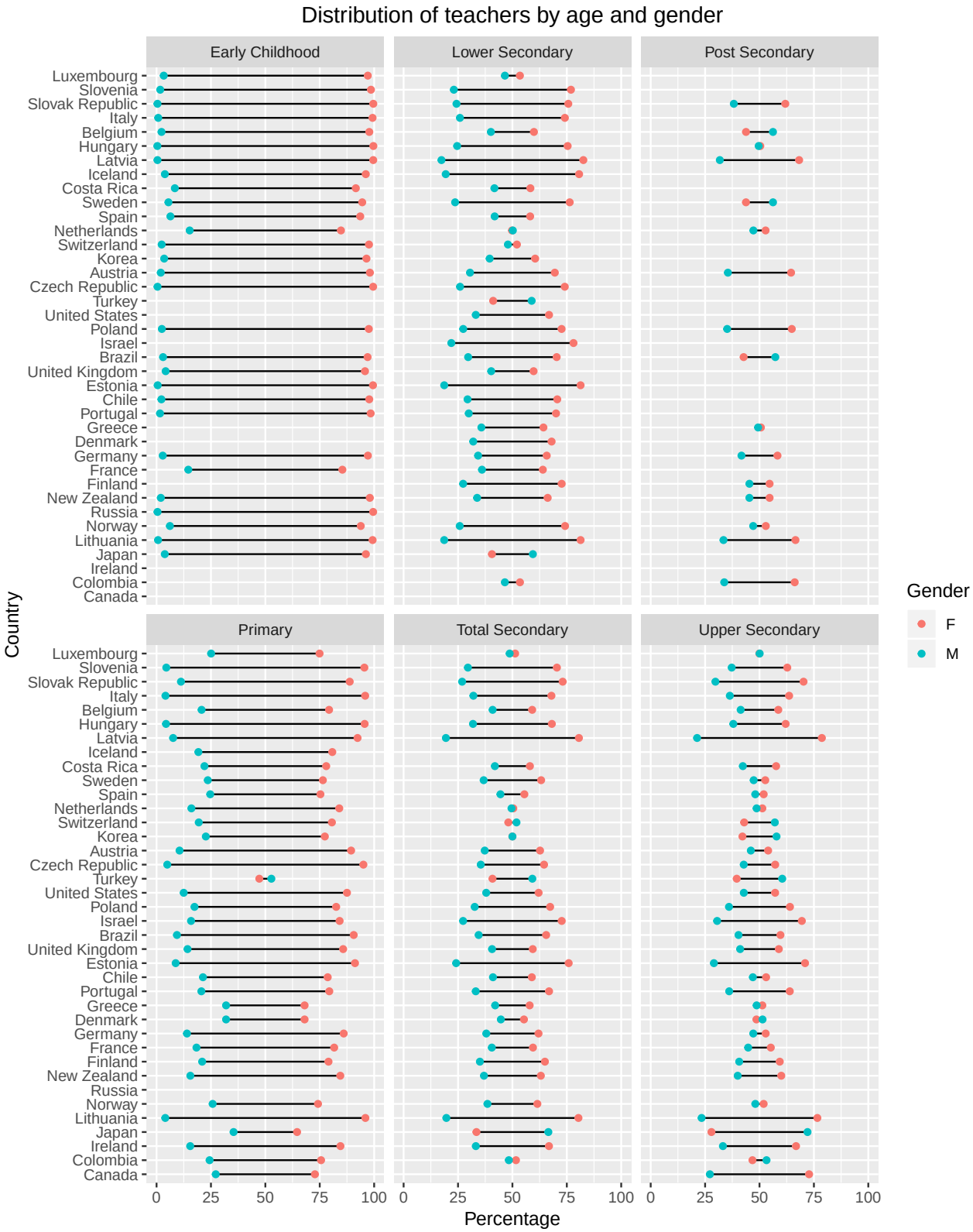


This plot reveals that the enrollment rates are higher for younger age groups and drops as the age increases. We also realized that there was no data for the age group of 3 - 14. The enrollment rates are almost zero for all countries after the age 25.

c) Distribution of Teachers: This indicator shows the distribution of teachers by gender at different levels of education. The data for this indicator is available from the year 2010 to 2016. We picked the average distribution of teachers over these years. We plotted a Cleveland dot-plot and faceted it by the education levels. The datapoints for women and men are pink and blue respectively.

```
country_data_education_three <- cleaned_data_three %>%
  group_by(Country, SEX, `Level of education`) %>%
  summarise(Value = mean(Value))
```

```
ggplot(country_data_education_three, aes(y = reorder(Country, Value),x = Value))+
  geom_line(aes(group = Country))+
  geom_point(aes(color = factor(SEX)))+
  labs(x="Percentage", y="Country", color = "Gender")+
  ggtitle("Distribution of teachers by age and gender")+
  facet_wrap(~education_level_short) +
  theme(plot.title = element_text(hjust = 0.5))
```



As we can see from the plot above, there are more number of women teachers at all levels of education. The difference between the women and men teachers is very high at the early stages of education and reduces as the level of education increases.

d) Distribution of Graduates: Graduates in each educational field as a percentage of the sum of graduates in all fields. This indicator shows the participation of women and men in the educational field. The data for this indicator is available for the year 2016. On creating a Cleveland dot plot for this indicator faceted by different education levels we obtain the following plot -

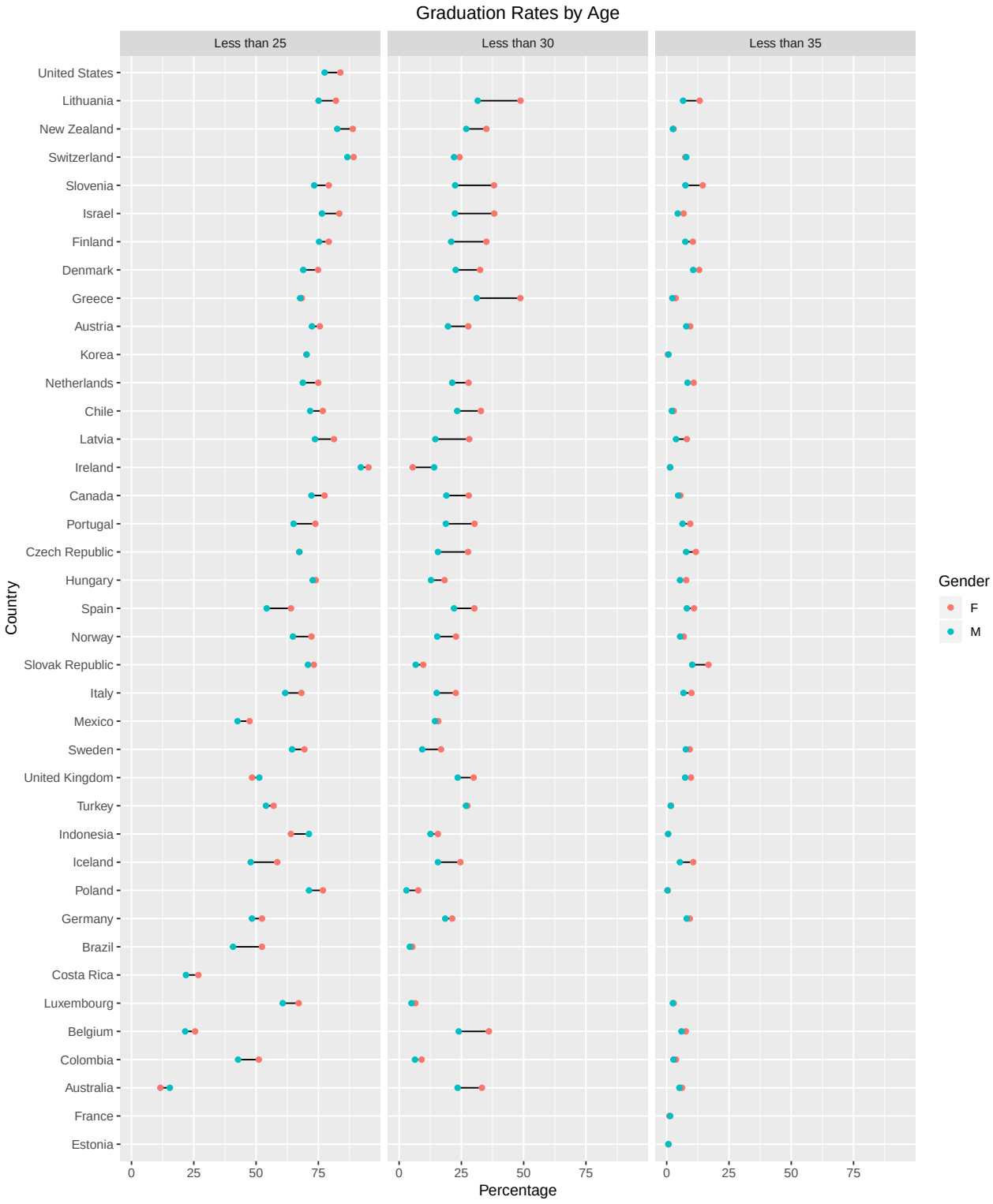
```
country_data_education_four <- cleaned_data_four %>%
  group_by(Country, SEX, Indicator_Short, `Level of education`) %>%
  summarise(Value = mean(Value))
```

```
country_data_education_four <- country_data_education_four[country_data_education_four$Indicator_Short =
ggplot(country_data_education_four, aes(y = reorder(Country, Value),x = Value))+
  geom_line(aes(group = Country))+
  geom_point(aes(color = factor(SEX)))+
  labs(x="Percentage", y="Country", color = "Gender")+
  ggtitle("Distribution of graduates and entrants by Field")+
  facet_wrap(~education_level_short) +
  theme(plot.title = element_text(hjust = 0.5))
```

We did not observe any pattern in difference between men and women graduates in the educational field.

e) Graduation Rate: Graduation rates represent an estimated percentage of an age group expected to graduate a certain level of education at least once in their lifetime. The data for this indicator is available from the year 2010 to 2016. We picked the average distribution of teachers over these years. The datapoints for women and men are pink and blue respectively. We obtained the following plot -

```r
country_data_education_five <- cleaned_data_five %>%
  group_by(Country, SEX, Age) %>%
  summarise(Value = mean(Value))

ggplot(country_data_education_five, aes(y = reorder(Country, Value), x = Value))+
  geom_line(aes(group = Country))+
  geom_point(aes(color = factor(SEX)))+
  labs(x="Percentage", y="Country", color = "Gender")+
  ggtitle("Graduation Rates by Age")+
  facet_wrap(~Age) +
  theme(plot.title = element_text(hjust = 0.5))
```



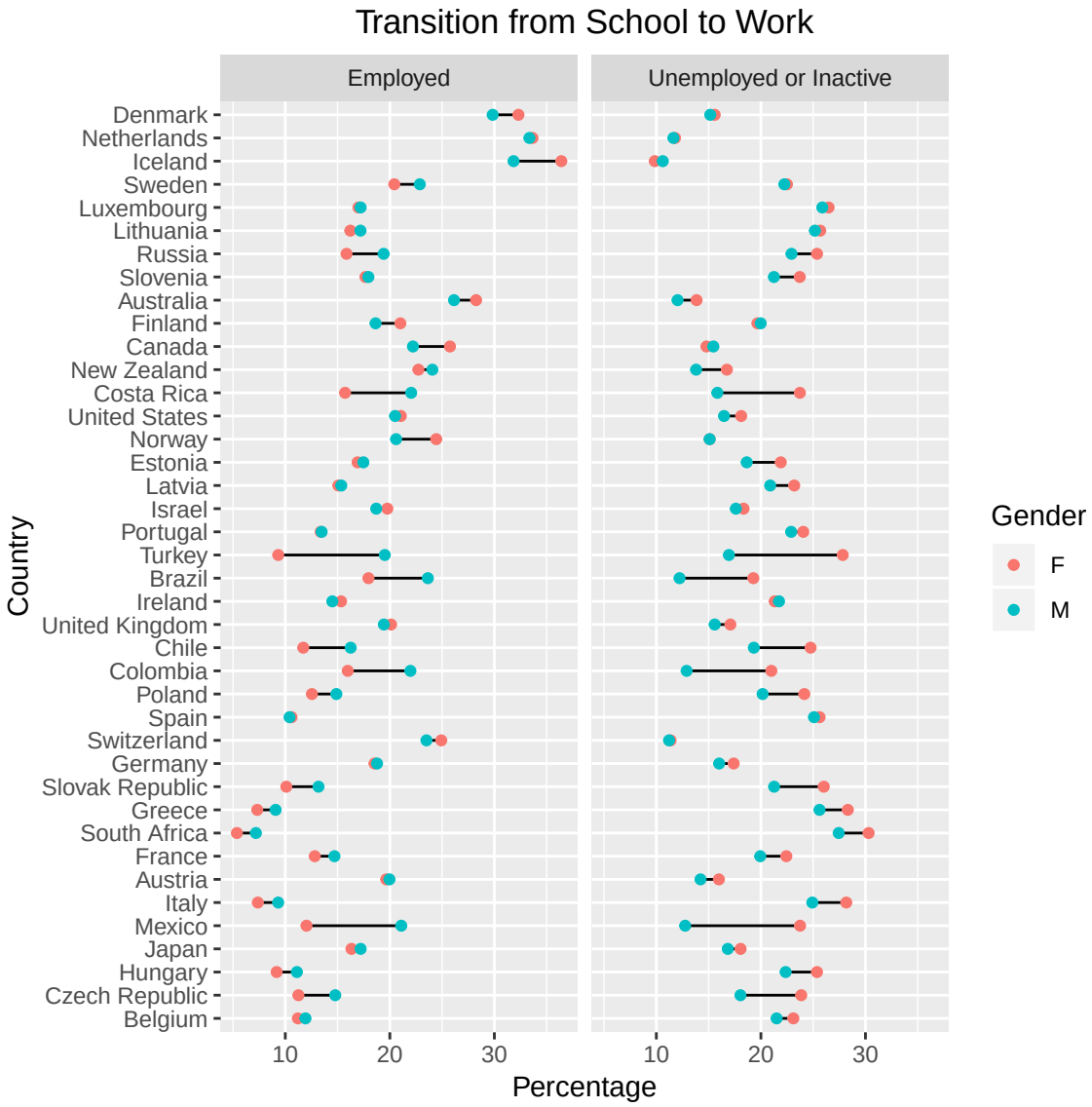
It can be seen from the plot that women have a higher graduation rate compared men and also that the graduation rate drops as the age group increases.

f) Transition from School to Work: This indicator presents internationally comparable data on labour force status and participation in formal education, by educational attainment, age and gender as reported by the labour force survey (LFS) and published in OECD Education at a Glance 2018. For trend data, the

Education at a Glance Database includes data from 1997 to 2017 (or years with available data). We created a plot of the percentage of men and women employed/ unemployed after school. The datapoints for women and men are pink and blue respectively.

```
cleaned_data_six <- cleaned_data_six %>% subset(`Education and labour force status` != "Work-study progr
  group_by(Country, SEX, YEAR, `Education and labour force status`) %>%
  summarise(Value = mean(Value))
cleaned_data_six <- cleaned_data_six[(cleaned_data_six$Country != "European Union 23 members in OECD") &
```

```
cleaned_by_years_six <- cleaned_data_six %>% group_by(Country, status_after_education, SEX) %>% summari

ggplot(subset(cleaned_by_years_six, status_after_education == "Employed" | status_after_education == "U
  geom_line(aes(group = Country))+
  geom_point(aes(color = factor(SEX)))+
  labs(x="Percentage", y="Country", color = "Gender")+
  ggtitle("Transition from School to Work")+
  facet_wrap(~status_after_education) +
  theme(plot.title = element_text(hjust = 0.5))
```

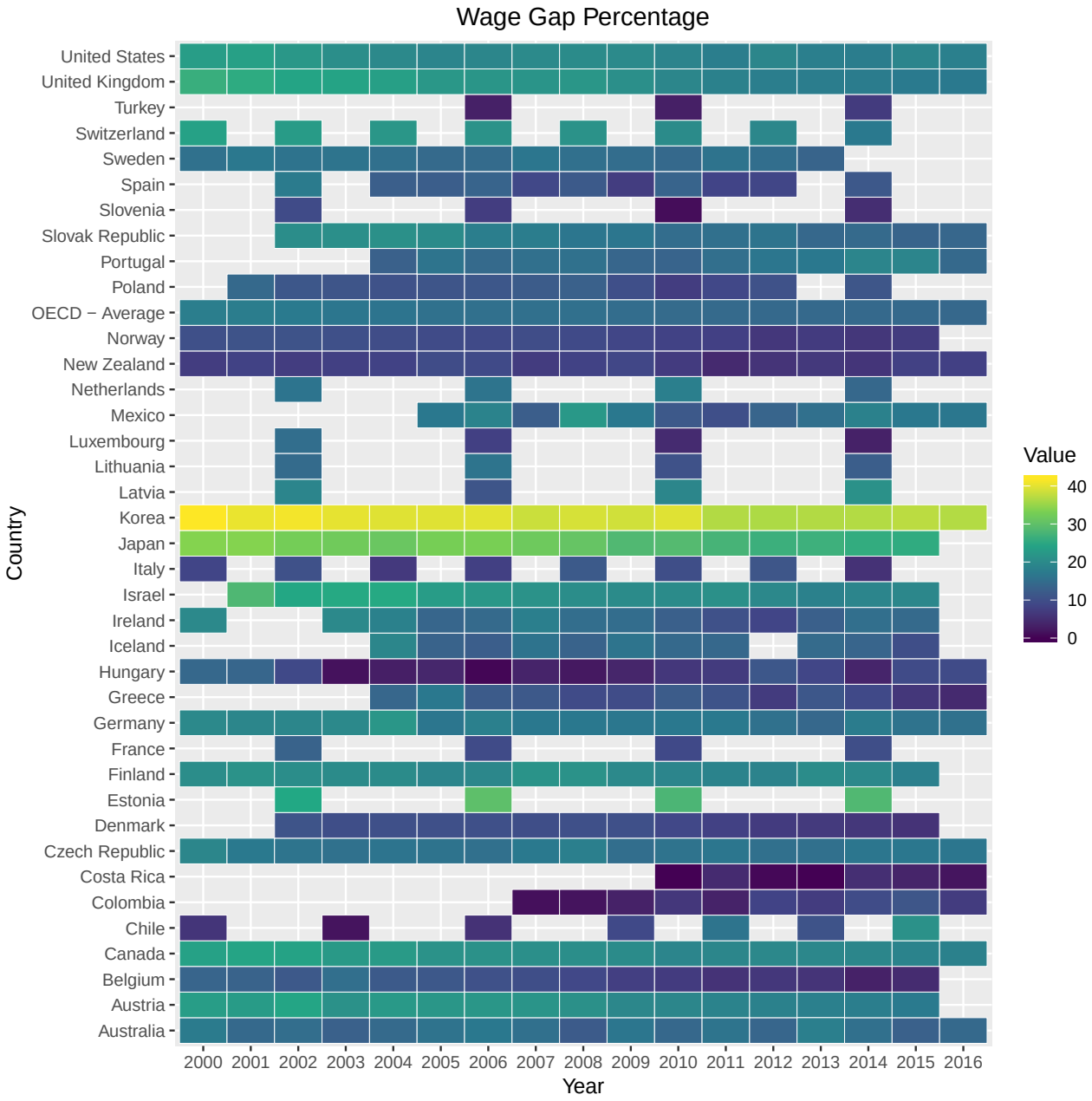


We can see that more men are employed compared women directly after school. It can also be seen that the more women are unemployed or inactive after school.

2. EMPLOYMENT

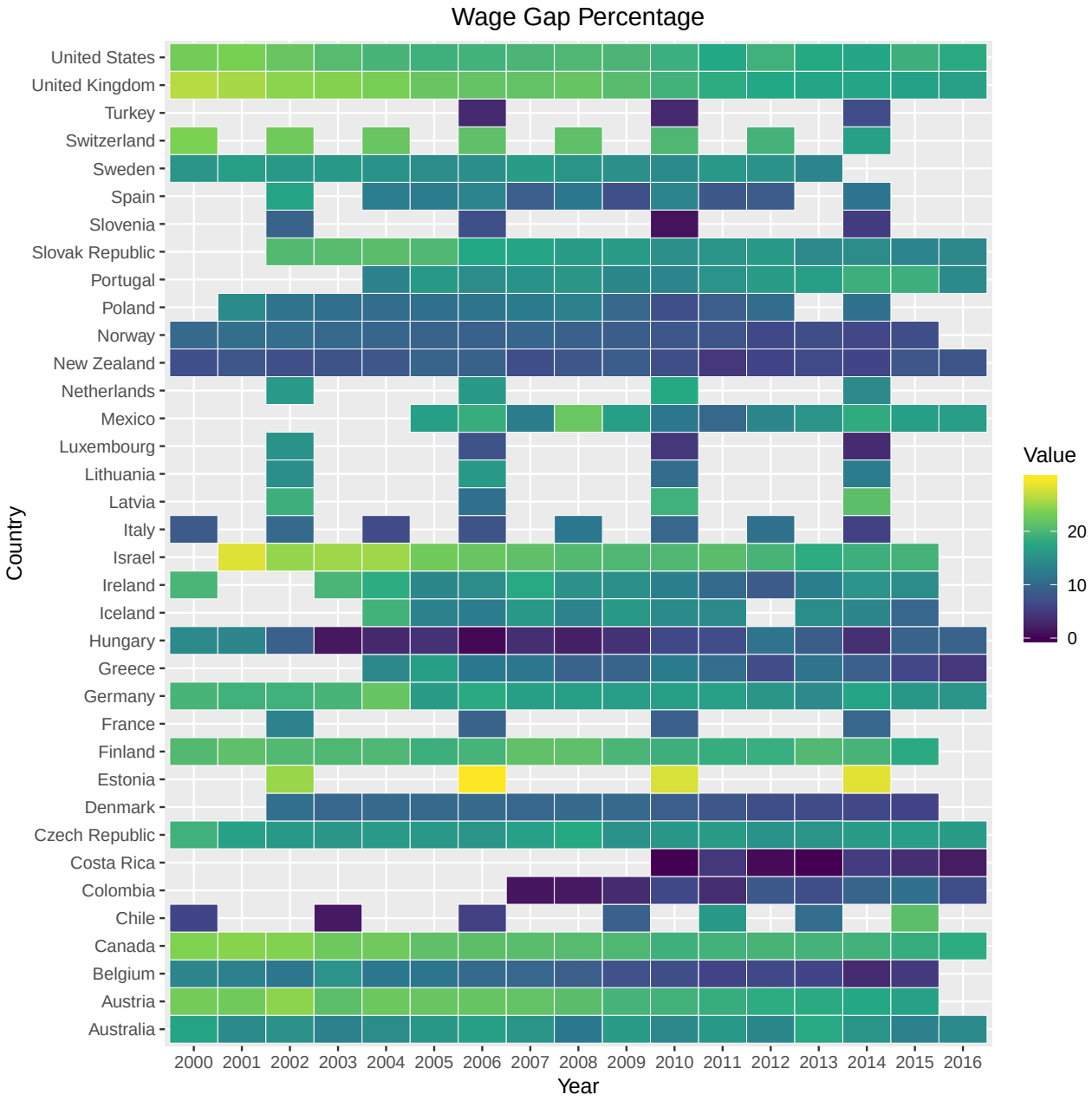
a) Gender Wage Gap: The gender wage gap is the percentage of difference in wages received by Men and Women, ideally there should not be any gender wage gap. The data for this indicator is available from the year 2000 to 2016. On plotting the heatmap for gender wage gap over years for different countries, we get the following visualization.

```
wage_gap$Time <- as.factor(wage_gap$Time)
ggplot(wage_gap, aes(Time, Country)) + geom_tile(aes(fill=Value), color = "white") + scale_fill_viridis
  xlab("Year") + ylab("Country") + ggtitle("Wage Gap Percentage") +
  theme(plot.title = element_text(hjust = 0.5))
```



As we can see in the above heatmap, the gender wage gap is the maximum in Japan and Korea which is surprising as Japan is a developed country and South Korea is widely regarded to be an almost developed country. Since the gender wage gap values for these two countries is so high, it masks the patterns that lie within the data of other countries in the heatmap. Hence, we removed the data from these two countries and obtained the following heatmap.

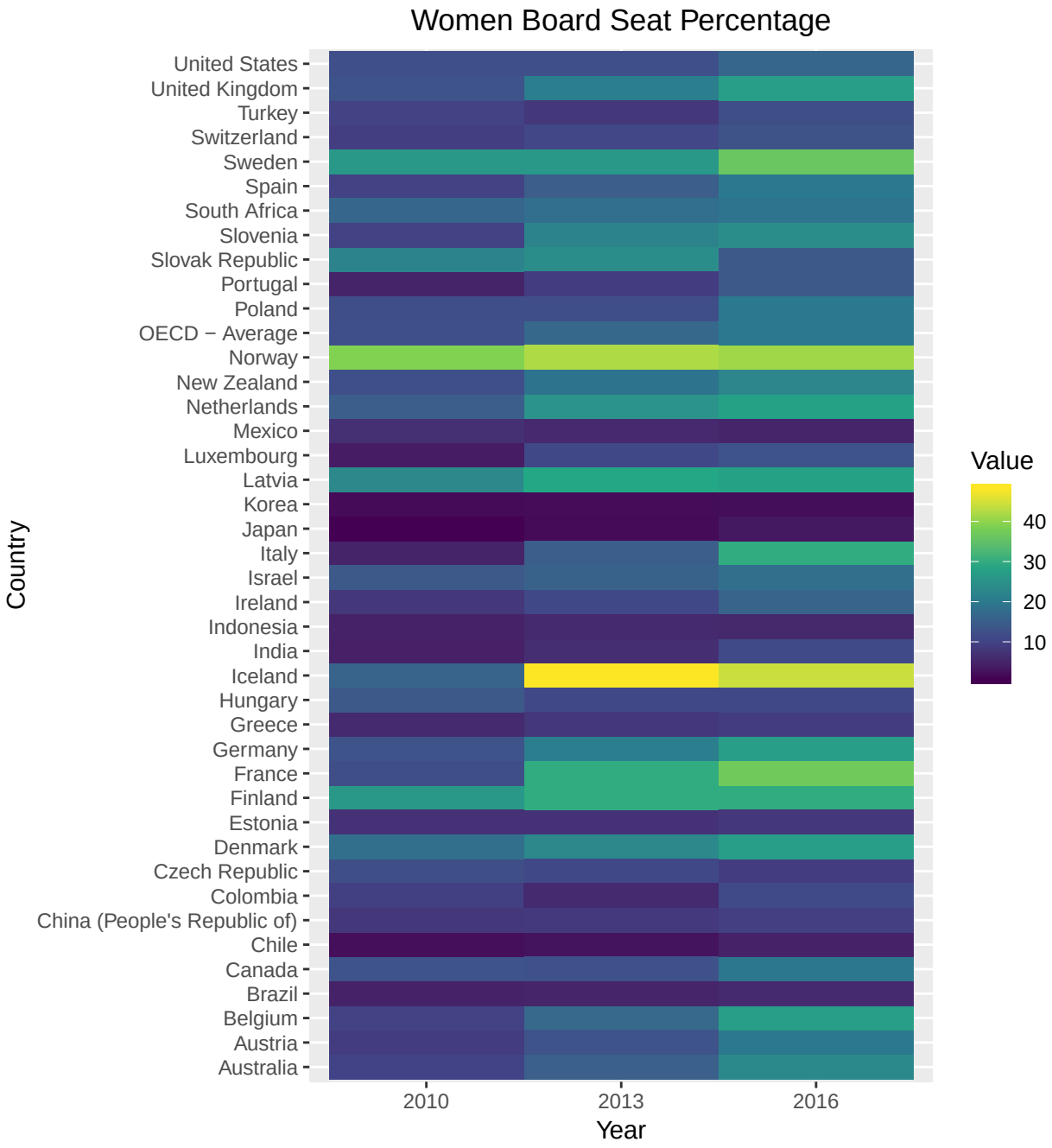
```
wage_gap <- wage_gap[(wage_gap$Country != "Korea") & (wage_gap$Country != "Japan") & (wage_gap$Country
ggplot(wage_gap, aes(Time, Country)) + geom_tile(aes(fill=Value), color = "white") + scale_fill_viridis_
  xlab("Year") + ylab("Country") + ggtitle("Wage Gap Percentage") +
  theme(plot.title = element_text(hjust = 0.5))
```



The above heatmap has more intricate patterns that can be seen now. For most countries, the gender wage gap has reduced with time as the movement from lighter colors to darker colors can be seen from the left to right. The only country where the gender wage gap has increased over the years is Estonia. It is also surprising to see that many under-developed like Costa Rica, Colombia and Hungary have the least gender wage gap.

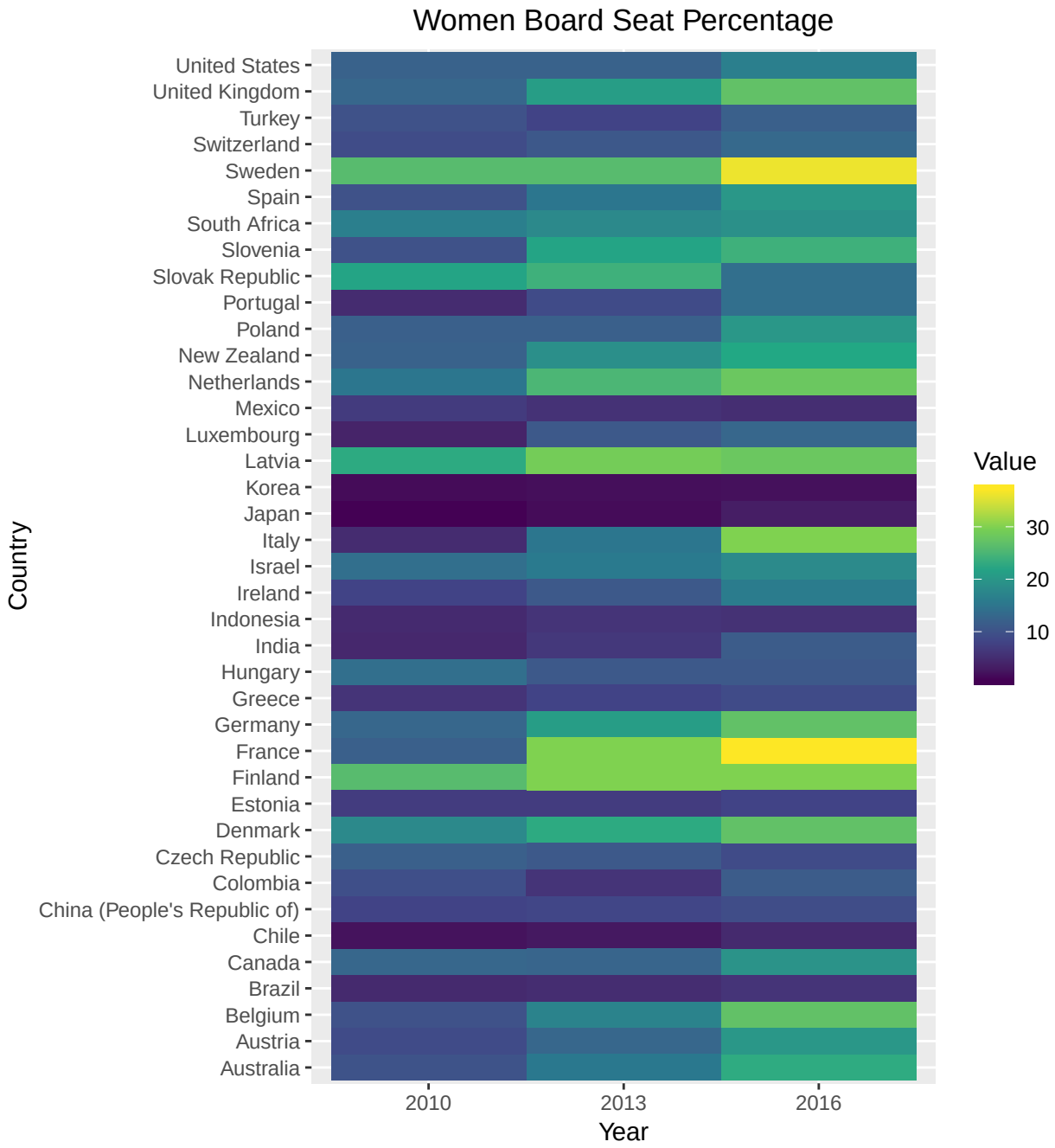
b) Women Board Seat Percentage: This is the percentage of women in the boards of large companies in different countries for the years 2010, 2013 and 2016. The data for this indicator is available for the years 2010, 2013 and 2016. The following visualization was obtained on plotting the heatmap for women board seat percentage -

```
board_seats$Time <- as.factor(board_seats$Time)
ggplot(board_seats, aes(Time, Country)) + geom_tile(aes(fill=Value)) + scale_fill_viridis_c(direction =
  xlab("Year") + ylab("Country") + ggtitle("Women Board Seat Percentage") +
  theme(plot.title = element_text(hjust = 0.5))
```



As we can see above, Norway and Poland have the highest percentage of women in the boards of large companies. Since the women board seat percentage values for these two countries is so high, it masks the patterns that lie within the data of other countries in the heatmap. Hence, we removed the data from these two countries and obtained the following heatmap.

```
board_seats <- board_seats[(board_seats$Country != "Norway") & (board_seats$Country != "Iceland") & (boa
ggplot(board_seats, aes(Time, Country)) + geom_tile(aes(fill=Value)) + scale_fill_viridis_c(direction =
  xlab("Year") + ylab("Country") + ggtitle("Women Board Seat Percentage") +
  theme(plot.title = element_text(hjust = 0.5))
```

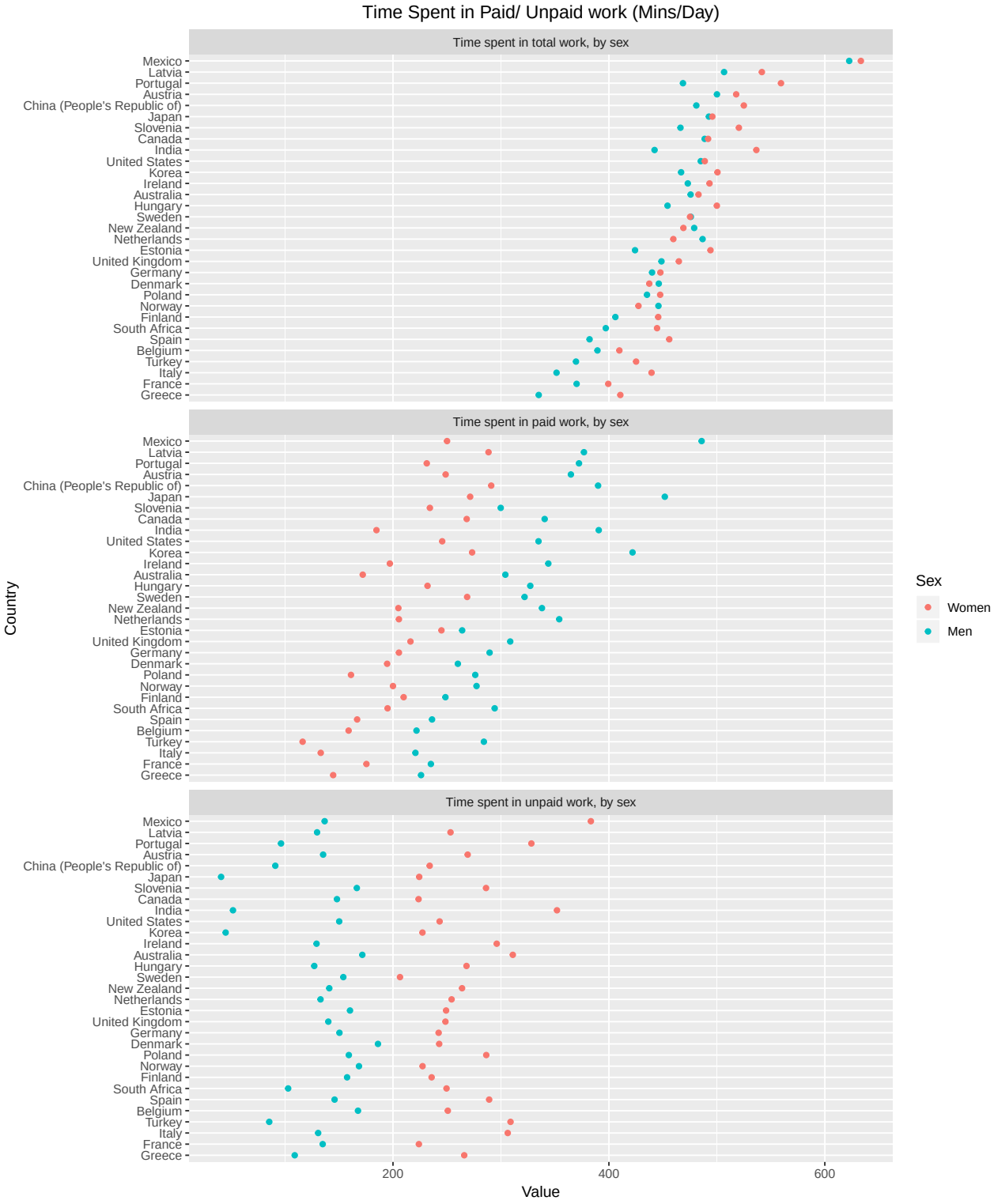


The above heatmap has more intricate patterns that can be seen now. For most countries, the women board percentage values have increased with time as the movement from darker colors to lighter colors can be seen from the left to right. The countries with the lowest values are Japan and Korea followed by Chile, Brazil and Mexico. There has been significant improvement in women board seats in countries like France, Sweden, Italy and Sweden over the recent years.

c) Time Spent in Paid/ Unpaid Work: This variable indicates the time spent by men and women on paid and unpaid work. The data for this indicator is available for the year 2016. We plotted a cleveland dot plot for the time spent and faceted it over total time spent in work, time spent in paid work and time spent in unpaid work. The datapoints for women and men are pink and blue respectively.

```
paid_unpaid$Indicator <- factor(paid_unpaid$Indicator, levels = c("Time spent in total work, by sex", "
paid_unpaid$Sex <- factor(paid_unpaid$Sex, levels = c("Women", "Men"))
```

```
ggplot(paid_unpaid, aes(reorder(Country, Value), Value, color = Sex)) + geom_point() + coord_flip() +
  facet_wrap(~Indicator, ncol  = 1) +
  xlab("Country") + ylab("Value") + ggtitle("Time Spent in Paid/ Unpaid work (Mins/Day)") +
  theme(plot.title = element_text(hjust = 0.5))
```

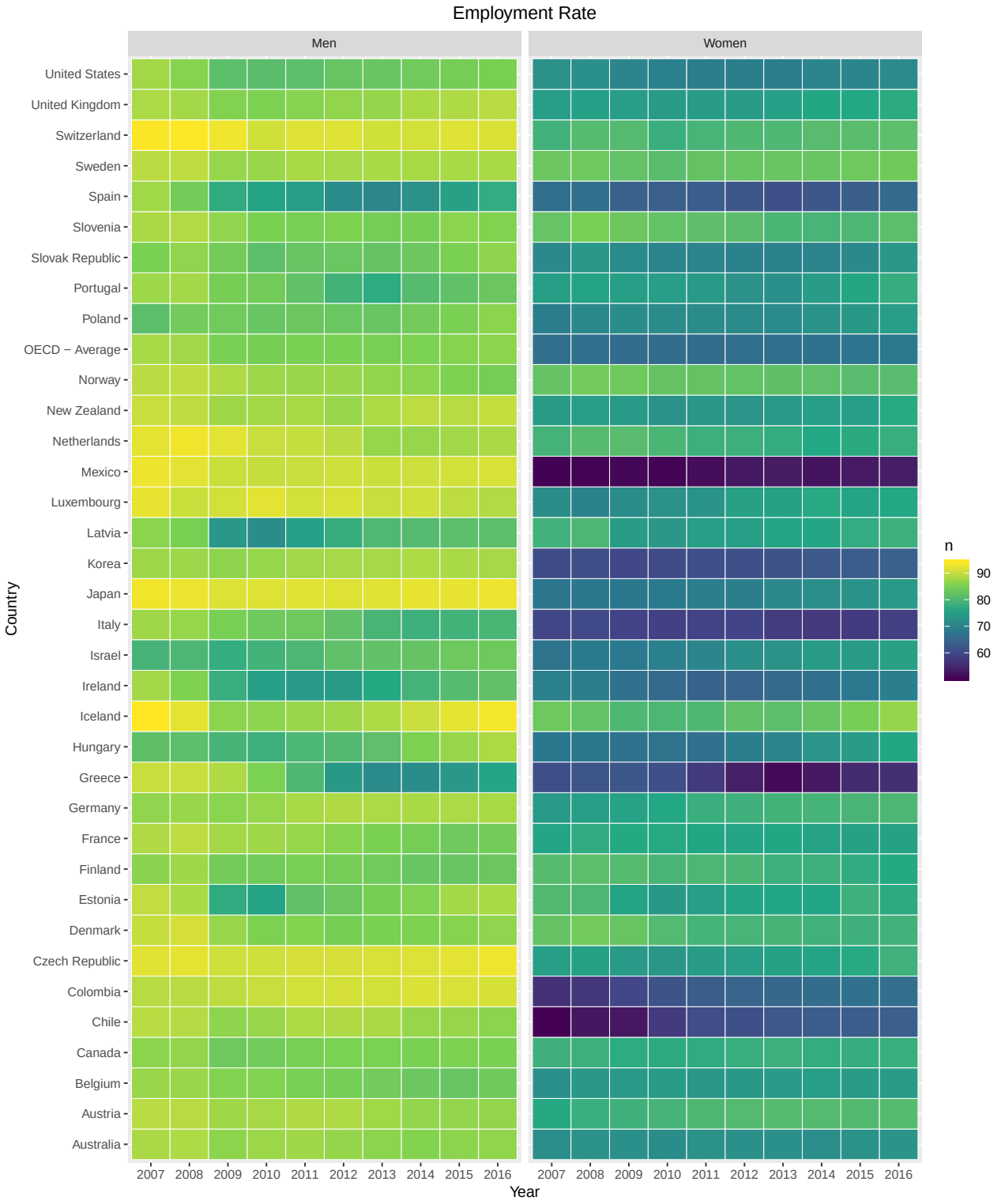


As we can see above, the total time spent by women in work is slightly higher than men. But the surprising

insight that we captured in this plot is that men spend most of that total time in paid work whereas women have to do a much higher amount of unpaid work. It can be seen that women spend the highest amount of time in unpaid work in Mexico, Portugal and India. It is discriminatory that women have to spend more time in unpaid work over men.

d) Employment/ Unemployment Rate: The employment rate is percentage of people who are willing to work and employed. The unemployment rate is the percentage of people who are seeking a job but unemployed. The data for this indicator is available from the year 2007 to 2016. We plotted a heatmap of the employment rate over years, faceted over men and women. Since the employment rate for Turkey was drastically low, it masked the patterns that lie within the data of other countries in the heatmap. Hence, we removed the data from Turkey and obtained the following heatmap.

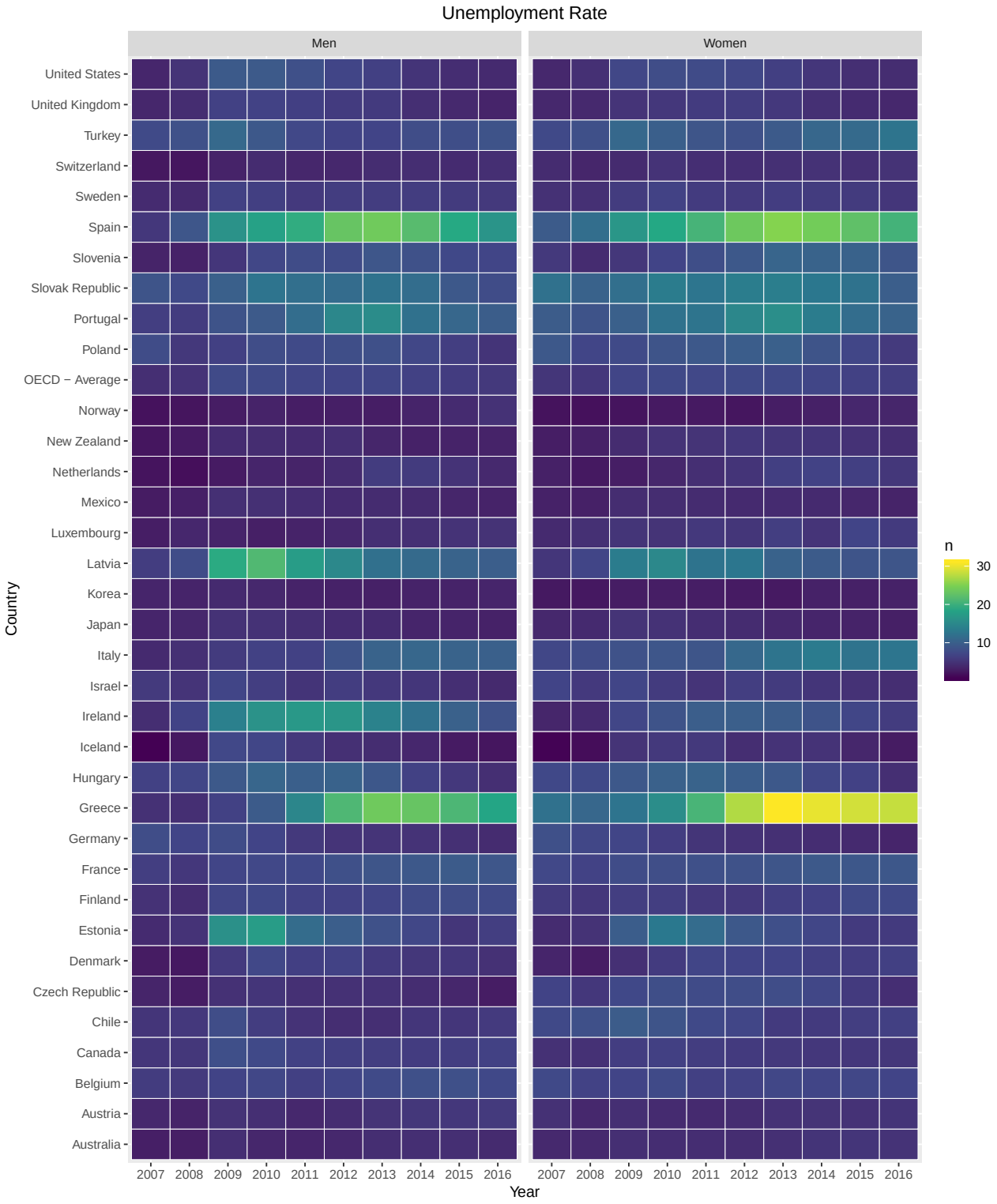
```
employment_unemployment$Time <- as.factor(employment_unemployment$Time)
subset <- employment_unemployment[employment_unemployment$Indicator == "Employment rate, by sex and age
subset <- subset[subset$Country != "Turkey", ]
subset <- subset[subset$Age.Group == "25-54", ]
subset$Year <- substr(subset$Time, 4,7)
subset <- subset %>% group_by(Country, Sex, Year) %>% summarise(n = mean(Value))
ggplot(subset, aes(Year, Country)) + geom_tile(aes(fill=n), color = "white") + scale_fill_viridis_c(dir
  facet_wrap(~Sex, ncol  = 2, scales = "free_x") +
  xlab("Year") + ylab("Country") + ggtitle("Employment Rate") +
  theme(plot.title = element_text(hjust = 0.5))
```



As we can see above, the employment rate of men is much higher compared to women as it can be seen by the difference in the colors of the facets. After Turkey, countries like Mexico, Greece, Colombia and Chile have the least women employment rates. It can also be seen that the employment rates have not changed for most countries over the years and the disparity remains the same which is a cause for concern.

The heatmap for the unemployment rate has been shown below.

```r
subset <- employment_unemployment[employment_unemployment$Indicator == "Unemployment rate, by sex and ag
subset <- subset[subset$Age.Group == "25-54", ]
subset$Year <- substr(subset$Time, 4,7)
subset <- subset %>% group_by(Country, Sex, Year) %>% summarise(n = mean(Value))
ggplot(subset, aes(Year, Country)) + geom_tile(aes(fill=n), color = "white") + scale_fill_viridis_c(dir
  facet_wrap(~Sex, ncol  = 2, scales = "free_x") +
  xlab("Year") + ylab("Country") + ggtitle("Unemployment Rate") +
  theme(plot.title = element_text(hjust = 0.5))
```

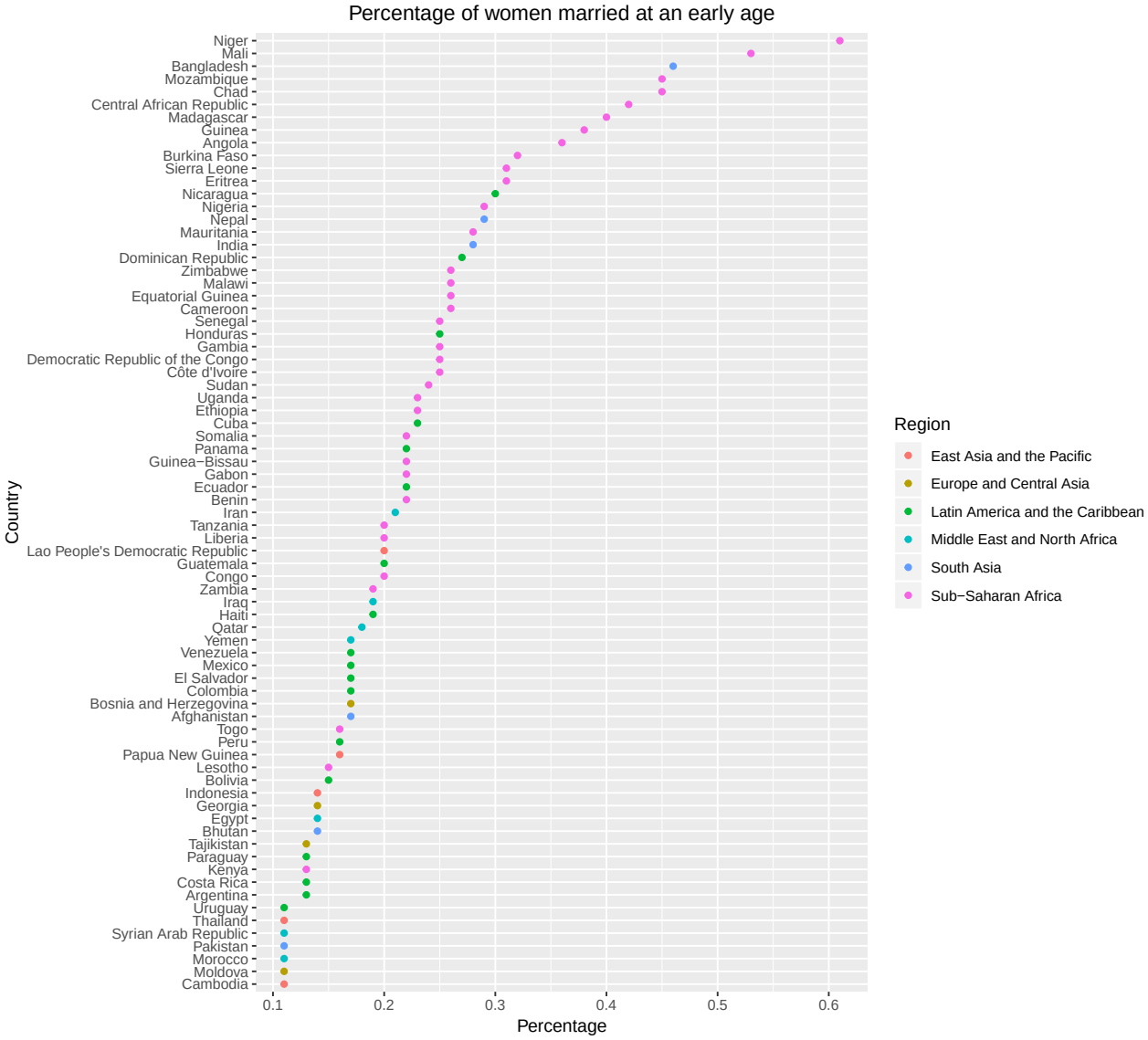


The unemployment rates for men and women are almost the same and there is significant pattern that can be observed in the above heatmap.

3. DEVELOPMENT

The following variables were analyzed (all data is from the year 2014):
a) Early Marriage: This variable depicts the percentage of women who are married before the legal age of marriage in a particular country. We draw a cleaveland dot plot and color different regions to observe what general pattern every region follows. We only include countries for which the value is more than 10%.
Here is the code:

```
#Remove duplicates
early_marriage <- subset(E_M, E_M$Region!="All Regions" & !duplicated(E_M$Country) & E_M$Value>0.1)

ggplot(early_marriage) +
  geom_point(stat = "identity",aes(x=reorder(Country,Value), y=Value, color=Region)) +
  labs(x = "Country", y="Percentage") +
  coord_flip() +
  ggtitle("Percentage of women married at an early age") +
  theme(plot.title = element_text(hjust = 0.5))
```

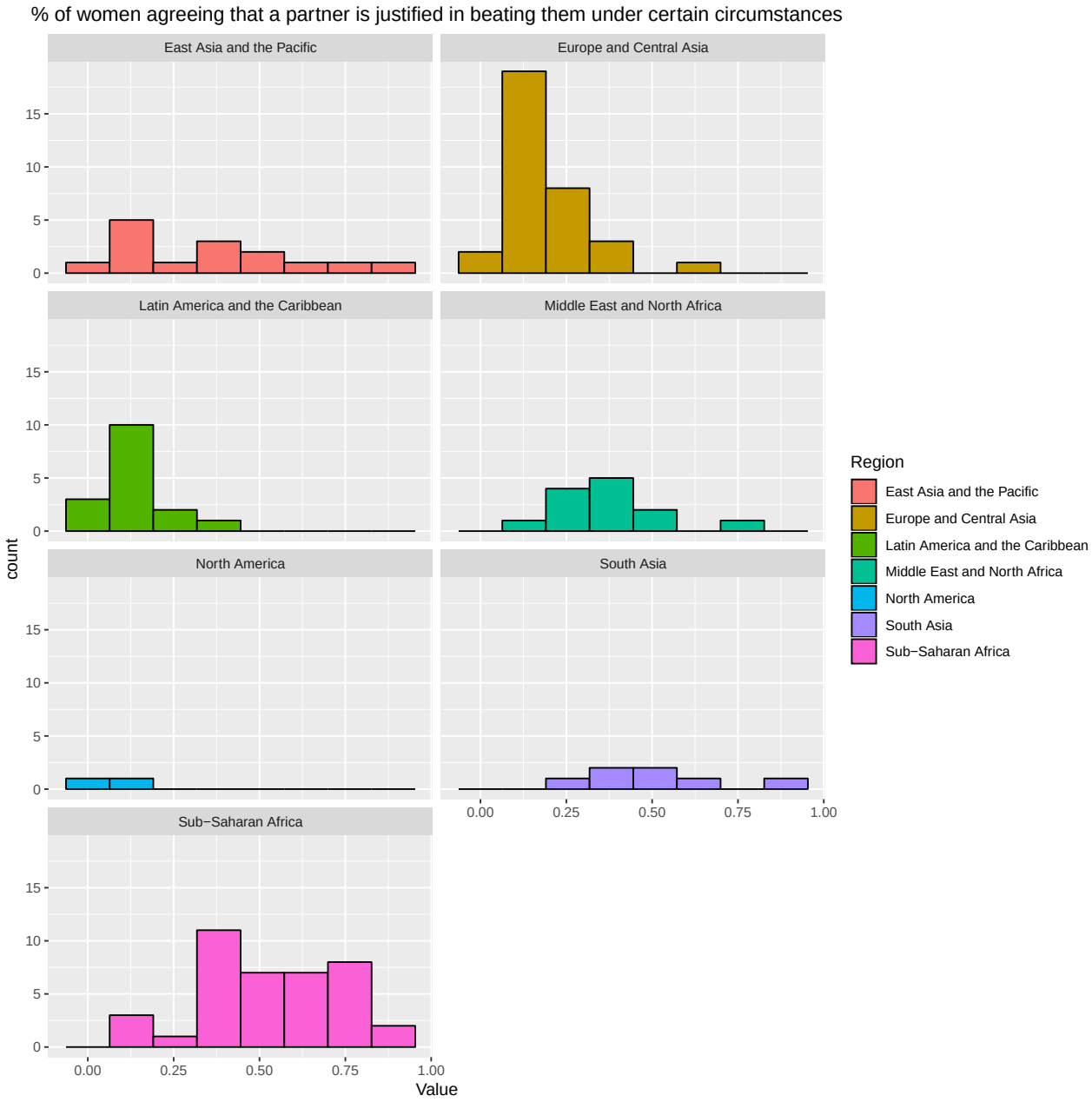


As we can observe, there are a shocking amount of countries in this world where atleast 10% of women are

married off at an early age. Regions like East Asia and the Pacific, Europe and Central Asia and Middle East and North Africa have fewer amount of countries with Value $> 10\%$, and are on the lower end of the graph. Latin America and the Caribbean has a wide range of values, and South Asia and Sub-Saharan Africa have their values in the upper range, with Niger being the maximum value, marrying off 61% of women at an early age.

b) Attitude Towards Violence: This variable depicts perhaps the most saddening result of all - How many women think it is okay for their partner to beat them in certain circumstances? There was simply too much data to plot into one graph, and faceting it by regions and simply displaying the values did not seem to give us a lot of information. Instead, we plot a histogram for value for all regions and observe the distribution:

```
#Remove duplicates
attitudes_towards_violence <- subset(A_T_V, A_T_V$Region!="All Regions" & !duplicated(A_T_V$Country))

ggplot(attitudes_towards_violence, aes(x=Value)) +
  geom_histogram(aes(fill = Region), bins=8, color="black") +
  facet_wrap(~Region, ncol = 2) +
  ggtitle("% of women agreeing that a partner is justified in beating them under certain circumstances"
  theme(plot.title = element_text(hjust = 0.5))
```

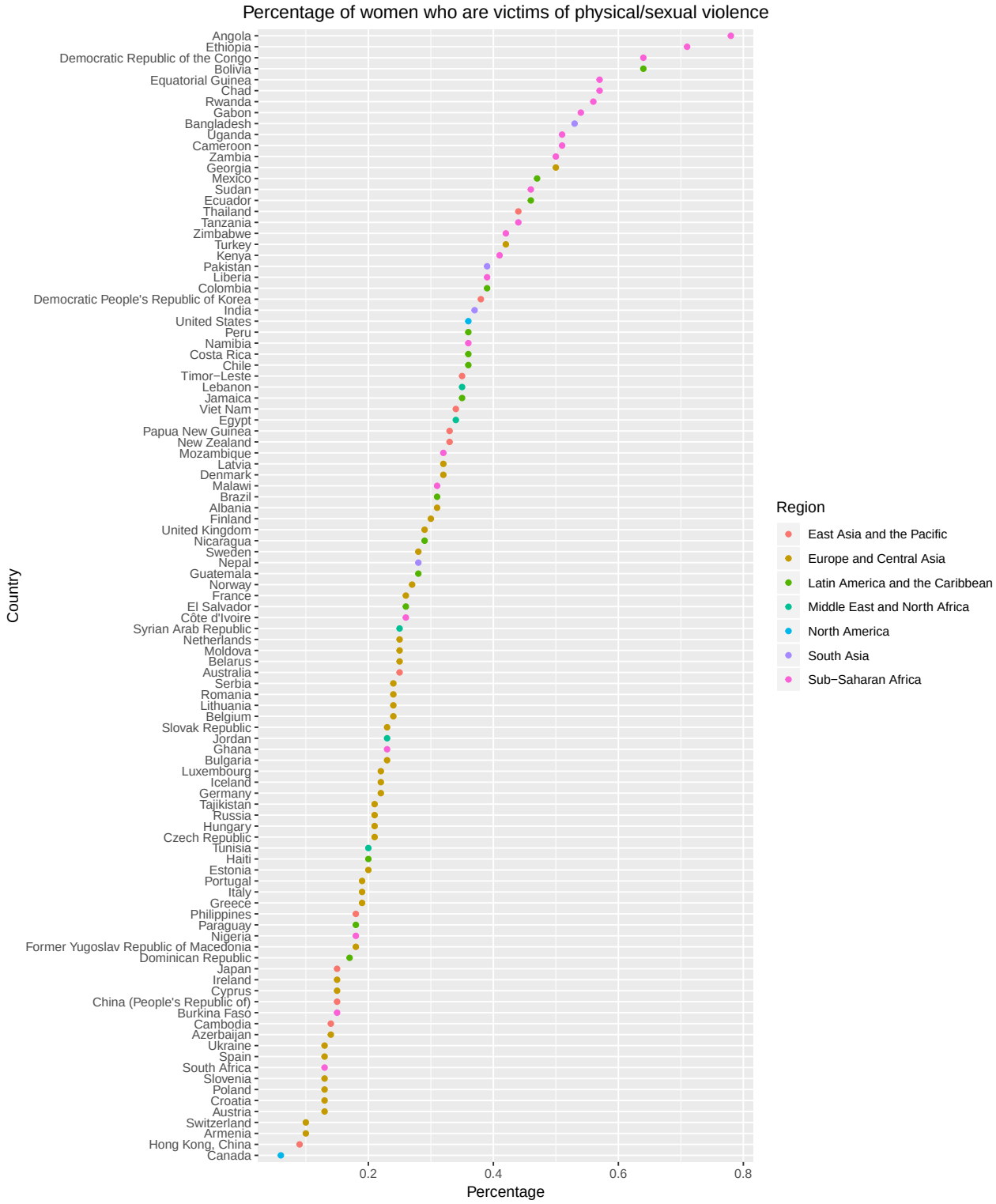


We observe that Europe and Central Asia, and Latin America and the Carribean are skewed towards the right, Sub Saharan Africa is left skewed, The Middle East and North Africa have the highest number of values in the middle, while the rest are uniformly distributed. One thing to note is that even though Europe and Central Asia is left skewed, there are a lot of countries where 6 to 19% of women believe that domestic violence is okay in some cases - not something we would expect from this region.

c) Prevalence Towards Violence: This variable shows how many women in every country are victims of some form of sexual or domestic violence in their life. We draw a Cleaveland dot plot of Countries vs Value as follows:

```
#Remove duplicates
prevelance_of_violence <- subset(P_o_V, P_o_V$Region!="All Regions" & !duplicated(P_o_V$Country))

ggplot(prevelance_of_violence) +
  geom_point(stat = "identity",aes(x=reorder(Country,Value), y=Value, color=Region)) +
```

```
  coord_flip() +
  labs(x = "Country", y="Percentage") +
  ggtitle("Percentage of women who are victims of physical/sexual violence") +
  theme(plot.title = element_text(hjust = 0.5))
```

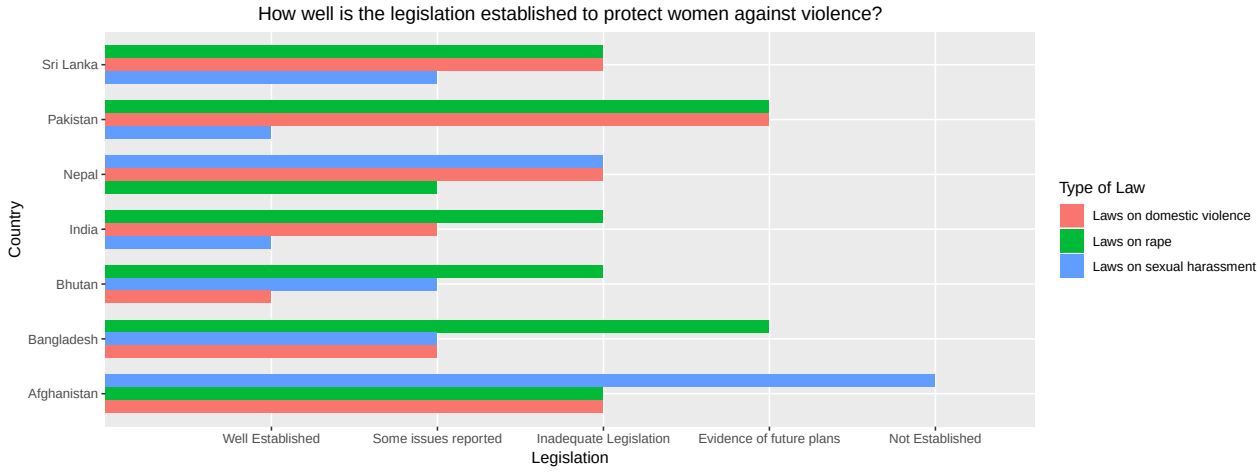


We observe that out of the two countries in North America - Canada and USA, there is a stark difference.

Canada is the lowest of all countries, (about 4%) while USA is quite high at about 37%. Again, the coutnries in the Sub Saharan region have high values and so is South Asia, while Europe and Central Aisa are spread out throught the plot.

d) Laws: This variable describes what the framework of laws is in every country. Three laws are taken into account: Laws against Sexual harassment, Rape and Domestic Violence. We wanted to see is every country has the same extent of legislation for all three, as one would expect. So we plot a dodged bar chart for the countries to see if that is indeed the case. Here we show only the result for the region of South Asia, which had the most interesting results.

```
#Remove duplicates
laws <- subset(L, L$Region!="All regions" & L$INC=="AIC")

#Recode for a better understanding of values
laws$Val_Meaning <- recode(laws$Value, `0.00` = "Well Established", `0.25` = "Some issues reported",
                           `0.50` = "Inadequate Legislation", `0.75` = "Evidence of future plans", `1.0
laws$Val_Meaning <- factor(laws$Val_Meaning, levels = c("Well Established", "Some issues reported", "Ina

ggplot(subset(laws, Region=="South Asia"), aes(x = Country, y=Val_Meaning, fill = Variables)) +
  geom_col(width = 0.7, position = "dodge") +
  labs(x="Country", y="Legislation", fill="Type of Law") +
  coord_flip() +
  ggtitle("How well is the legislation established to protect women against violence?") +
  theme(plot.title = element_text(hjust = 0.5))
```

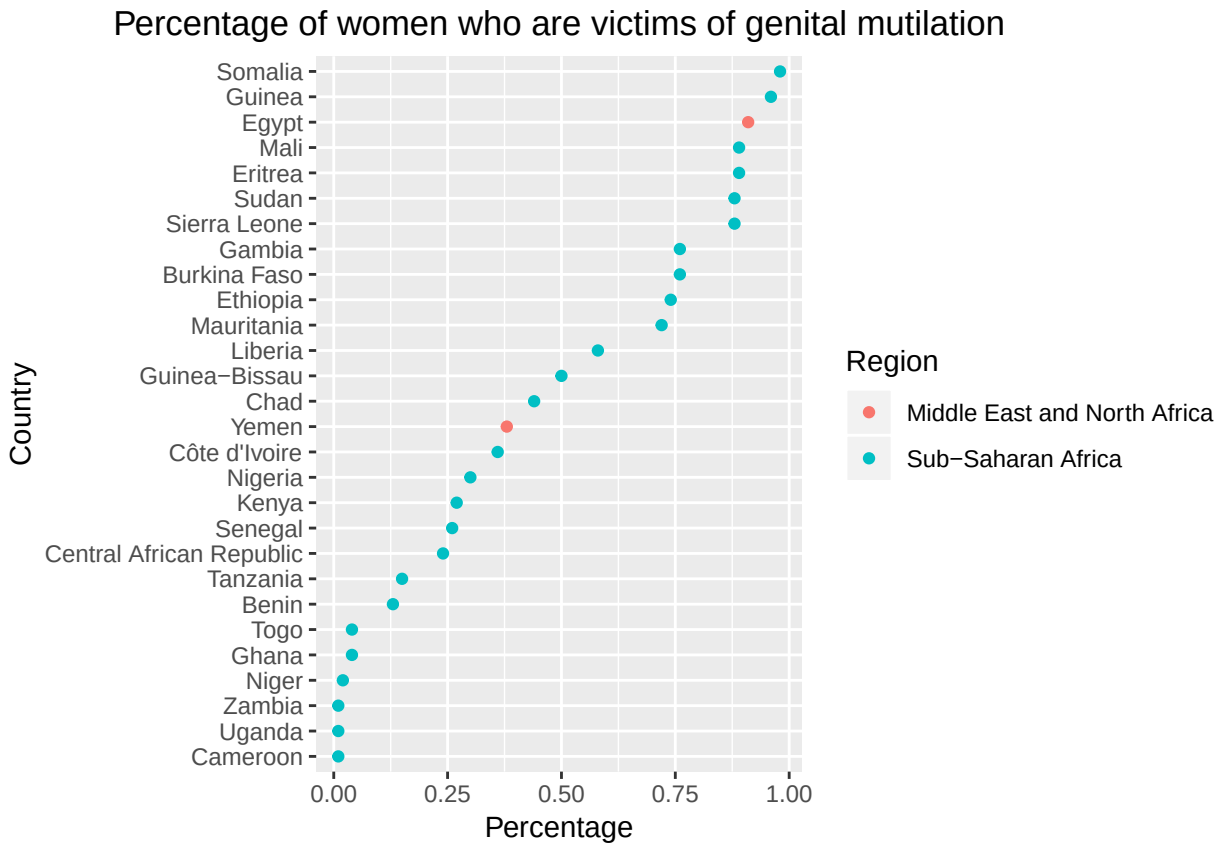


As we can see, the results we see are not what we expected - many countries, who have well established law of one type, do not even have any legislation for others. For example, Pakistan has Well established legislation for sexual harassment, while it has no legislation for either domestic violence or rape (there is evidence of future plans). This means a person going through domestic violence or rape, has no support from the law, while a person going through sexual harassment does. This is clearly a loophole in the legislation for many countries.

e) Female genital mutilation: This variable represents the percentage of women who have suffered from genital mutilation in a country. Here is it's cleaveland dot plot (only showing those countries whose value is > 0):

```
female_genital_mutilation <- subset(F_G_M, F_G_M$Region!="All Regions" & !duplicated(F_G_M$Country) & F

ggplot(female_genital_mutilation) +
  geom_point(stat = "identity",aes(x=reorder(Country,Value), y=Value, color=Region)) +
  coord_flip() +
  labs(x = "Country", y="Percentage") +
```

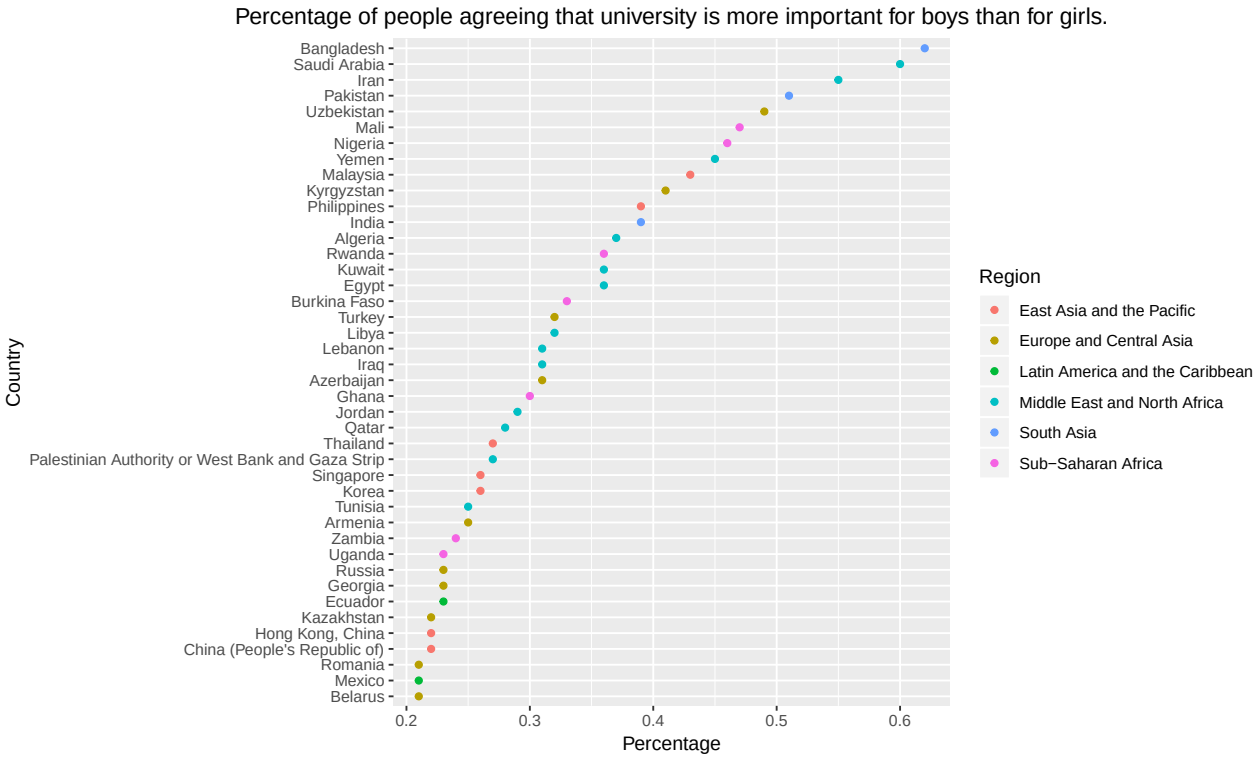
```
  ggtitle("Percentage of women who are victims of genital mutilation") +
  theme(plot.title = element_text(hjust = 0.5))
```



We see that most countries are from Sub-Saharan Africa, and two from Middle East and North Africa. There are so many countries that still follow the practice of genital mutilation, with countries such as Somalia or Guinea having such high percentages that we can say almost every woman has gone through this ordeal. Also observe the sharp jump after Gambia, which is about 75% to Sierra Leone, which is about 87%.

f) Son Education Preference: This variable shows the percentage of people who agree that university is more important for boys than for girls. The scatterplot, showing countries where the percentage is higher than 20% is as follows:

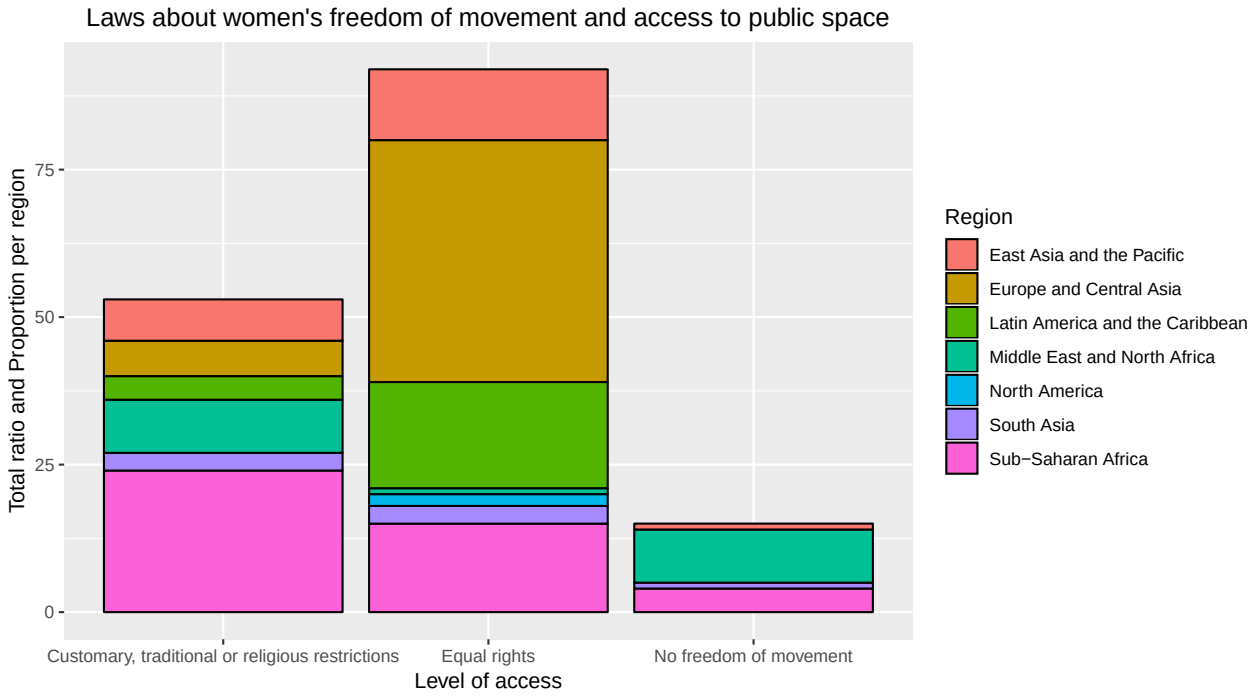
```
son_education_preference <- subset(S_E_P, S_E_P$Region!="All Regions" & !duplicated(S_E_P$Country))

ggplot(subset(son_education_preference, Value>0.2)) +
  geom_point(stat = "identity",aes(x=reorder(Country,Value), y=Value, color=Region)) +
  coord_flip() +
  labs(x = "Country", y="Percentage") +
  ggtitle("Percentage of people agreeing that university is more important for boys than for girls.") +
  theme(plot.title = element_text(hjust = 0.5))
```



We observe that unlike other graphs, the highest values are from South Asia and the Middle East.

g) Access to Public Space: This variable explores the laws about restriction of one of the primary civil liberties of a person. It takes on three values - if a woman's access to public space is not restricted by law, if it is not restricted but some religious or cultural factors influence it, or if there is complete restriction on freedom of movement. We draw a stacked bar chart to explore this:

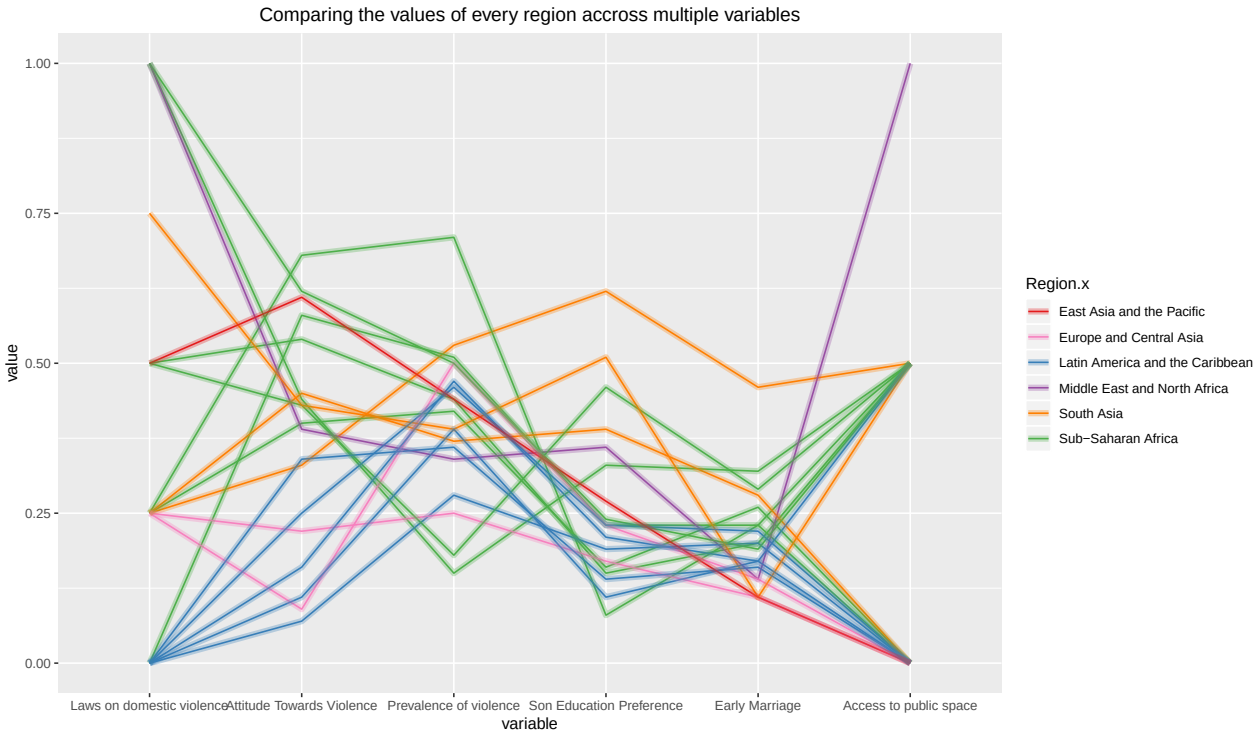
```
access_to_public_space <- subset(A_T_P_S, A_T_P_S$Region!="All Regions" & !duplicated(A_T_P_S$Country))

access_to_public_space$Value_text<-recode(access_to_public_space$Value, `0.0`="Equal rights", `0.5`="Cus

ggplot(access_to_public_space, aes(x=Value_text)) +
  geom_bar(stat = "count", aes(fill = Region),color="black") +
  labs(x = "Level of access", y="Total ratio and Proportion per region") +
  ggtitle("Laws about women's freedom of movement and access to public space") +
  theme(plot.title = element_text(hjust = 0.5))
```



We observe that most countries have equal rights for women in this case, while for some, customs and religions restrict it. And a few countries - majorly from Africa and the Middle east - also have complete restriction of movement. We normally tend to think that such a basic right will be restricted in very few countries, but it is disappointing to see that over 12% countries have complete restrictions, and more than 50% of the countries do not give their women complete freedom.

Finally, we performed some multi-variate analysis for all variables to see if some countries follow a specific pattern. We use a parallel coordinate plot for this:

```
ggparcoord(merged_data, columns=c(8,4,6,10,1,12), groupColumn = "Region.x", scale = "globalminmax", tit
  geom_line(size=2, alpha=0.3) +
  scale_color_manual(values=c(brewer.pal(8, "Set1")[c(1,8,2,4,5,3)])) +
  theme(plot.title = element_text(hjust = 0.5))
```

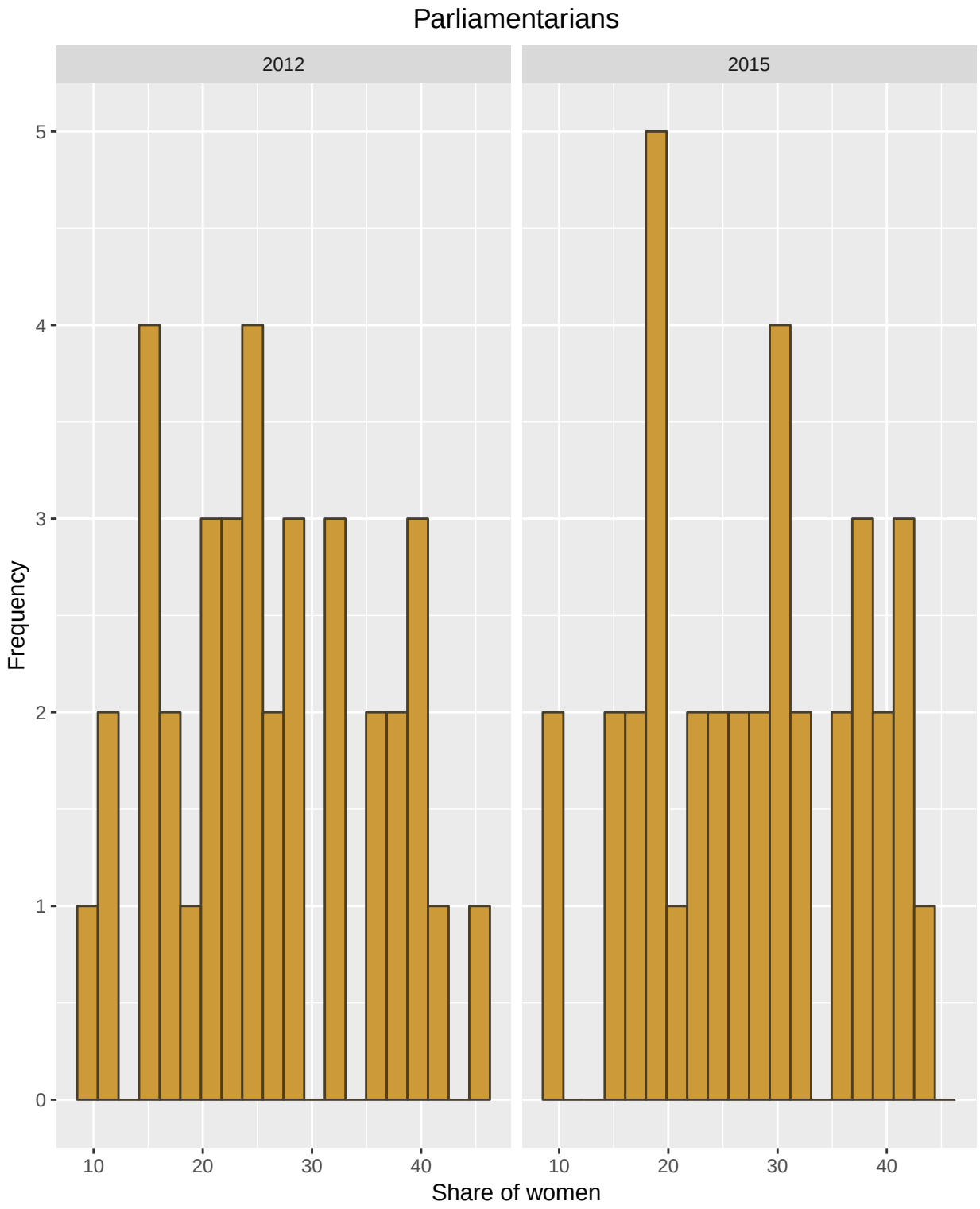


Here, we observe that most regions follow a pattern - Latin America and the Carribean tend to stay on the lower side of the graph, South Asia tends to be on the upper side. East Asia is high initially, but drops to lower values as we move to the right of the graph.

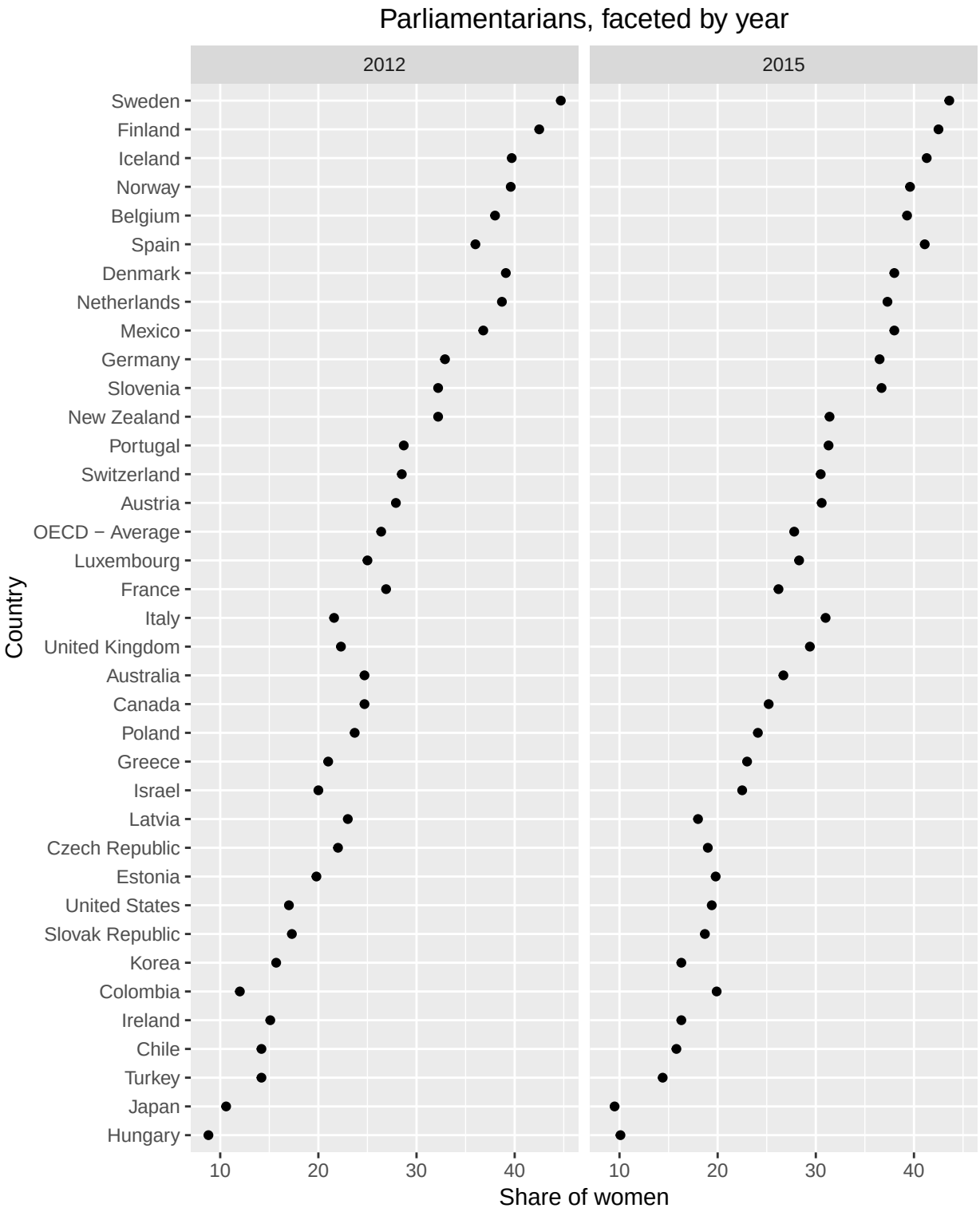
4. GOVERNANCE

a)Share of women Parlimentarians: The variable refers to the percentage of women in parliament for different countries. The data is for years 2012 and 2015.

```
women_parl<- govt_data[govt_data$Indicator == "Share of women parliamentarians",]

ggplot(women_parl, aes(x = Value)) +
  geom_histogram(fill = "#cc9a38", color = "#473e2c", bins=20) +
  ggtitle("Parliamentarians") +
  labs(x = "Share of women", y = "Frequency") +
  facet_grid(. ~ Year) +
  theme(plot.title = element_text(hjust = 0.5))
```



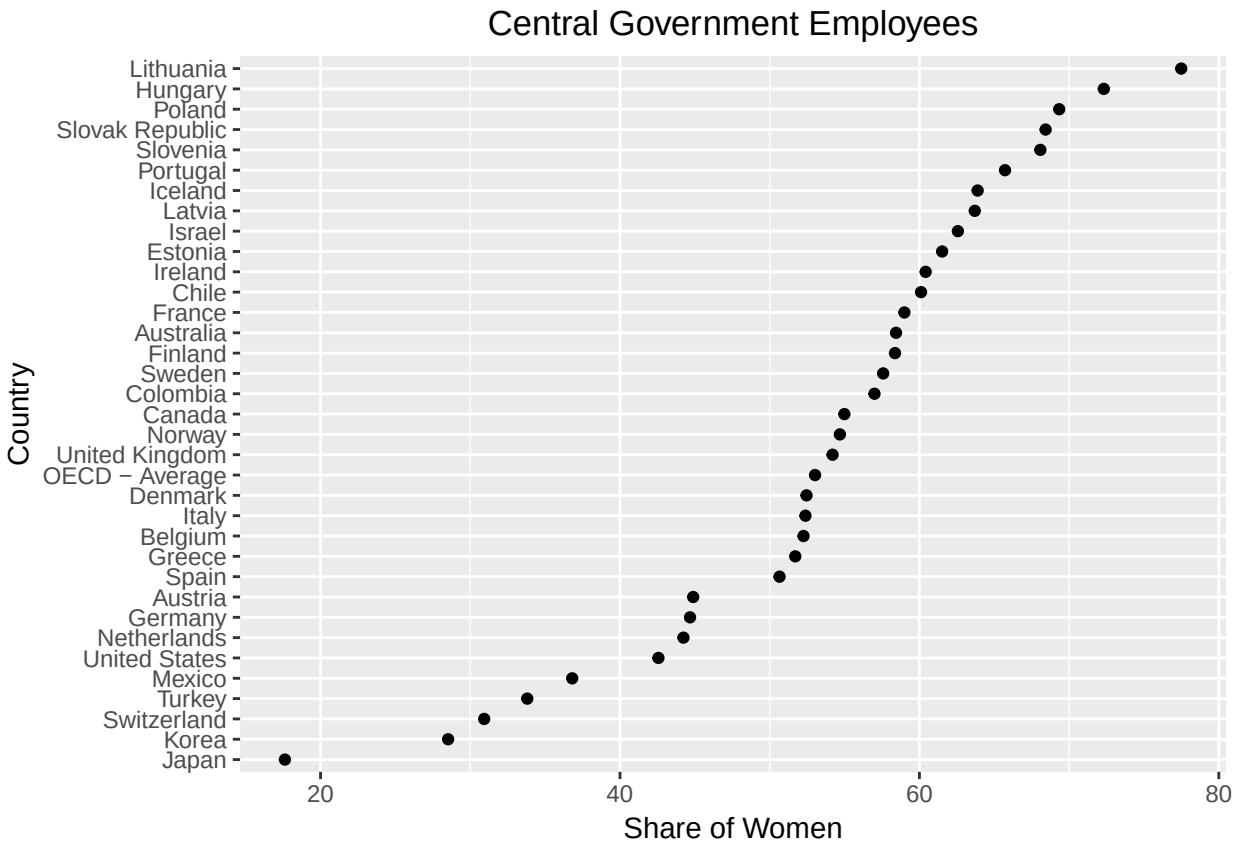

```
ggplot(women_parl, aes(y = reorder(Country, Value), x= Value)) +
  geom_point() +
  ggtitle("Parliamentarians, faceted by year") +
  labs(x = "Share of women", y = "Country") +
  facet_grid(. ~ Year) +
  theme(plot.title = element_text(hjust = 0.5))
```



Histogram shows a distribution of women in parliament for 2012 and 2015. We can see that more higher values in 2015 than 2012.There are very few countries with 50 percent women parliamentarians. The cleveland dot plot shows a distribution of share of women in parliament across countries. We see that higher share is in European countries like Sweden , Finland etc. Ironically, developed countries like Japan and USA still have a male dominant parliament. We see a general trend of increase in women share in parliament from 2012 to 2015 for all countries

b) Share of central government employment: The variable refers to the percentage of women employees in central government. The data is for year 2015.

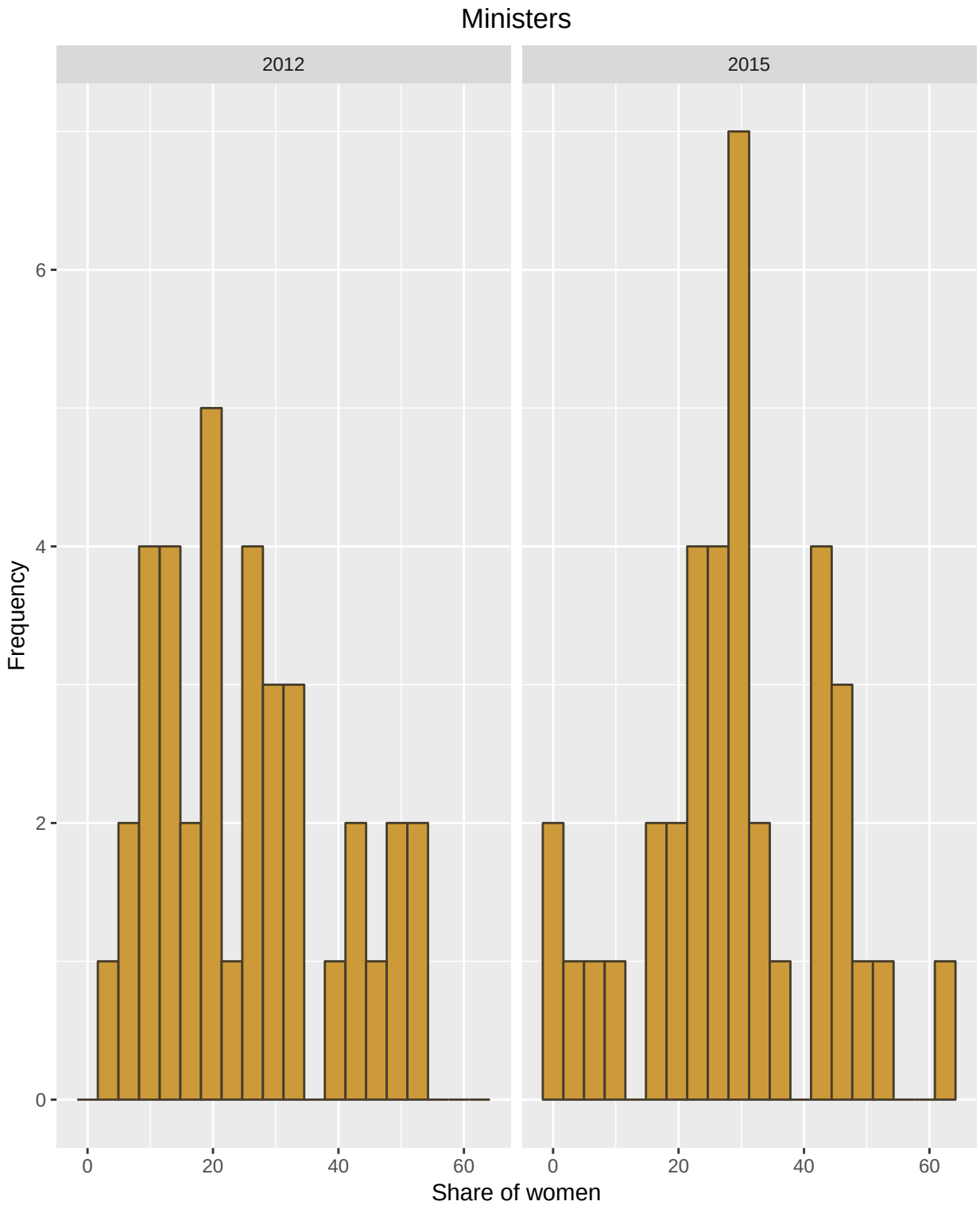
```
central_govt<- govt_data[govt_data$Indicator == "Share of central government employment filled by women

ggplot(central_govt, aes(y = reorder(Country, Value), x= Value)) +
  geom_point() +
  ggtitle("Central Government Employees") +
  labs(x = "Share of Women", y = "Country") +
  theme(plot.title = element_text(hjust = 0.5))
```



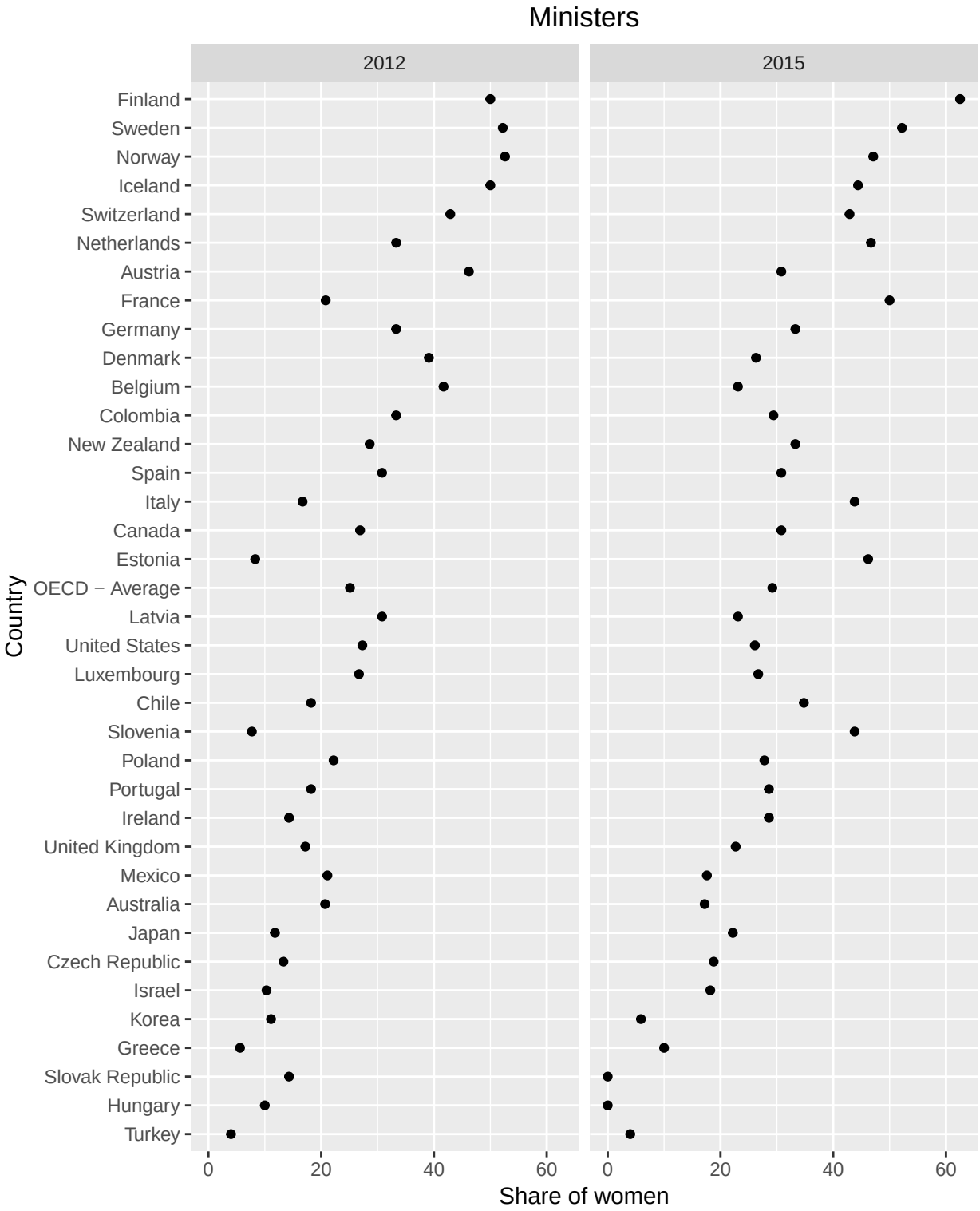
The cleveland dot plot shows a distribution of share of women employees across countries. We see that higher share is in European countries like Poland , Hungary etc. Ironically, developed countries like Japan and USA still have a low female work force in government employees.

c) Share of women ministers: The variable refers to the percentage of women ministers for different countries. The data is for years 2012 and 2015.

```
women_min<- govt_data[govt_data$Indicator == "Share of women ministers",]

ggplot(women_min, aes(x = Value)) +
  geom_histogram(fill = "#cc9a38", color = "#473e2c", bins=20) +
  ggtitle("Ministers") +
  labs(x = "Share of women", y = "Frequency") +
  facet_grid(. ~ Year) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(women_min, aes(y = reorder(Country, Value), x= Value)) +
  geom_point() +
  ggtitle("Ministers") +
  labs(x = "Share of women", y = "Country") +
  facet_grid(. ~ Year) +
  theme(plot.title = element_text(hjust = 0.5))
```

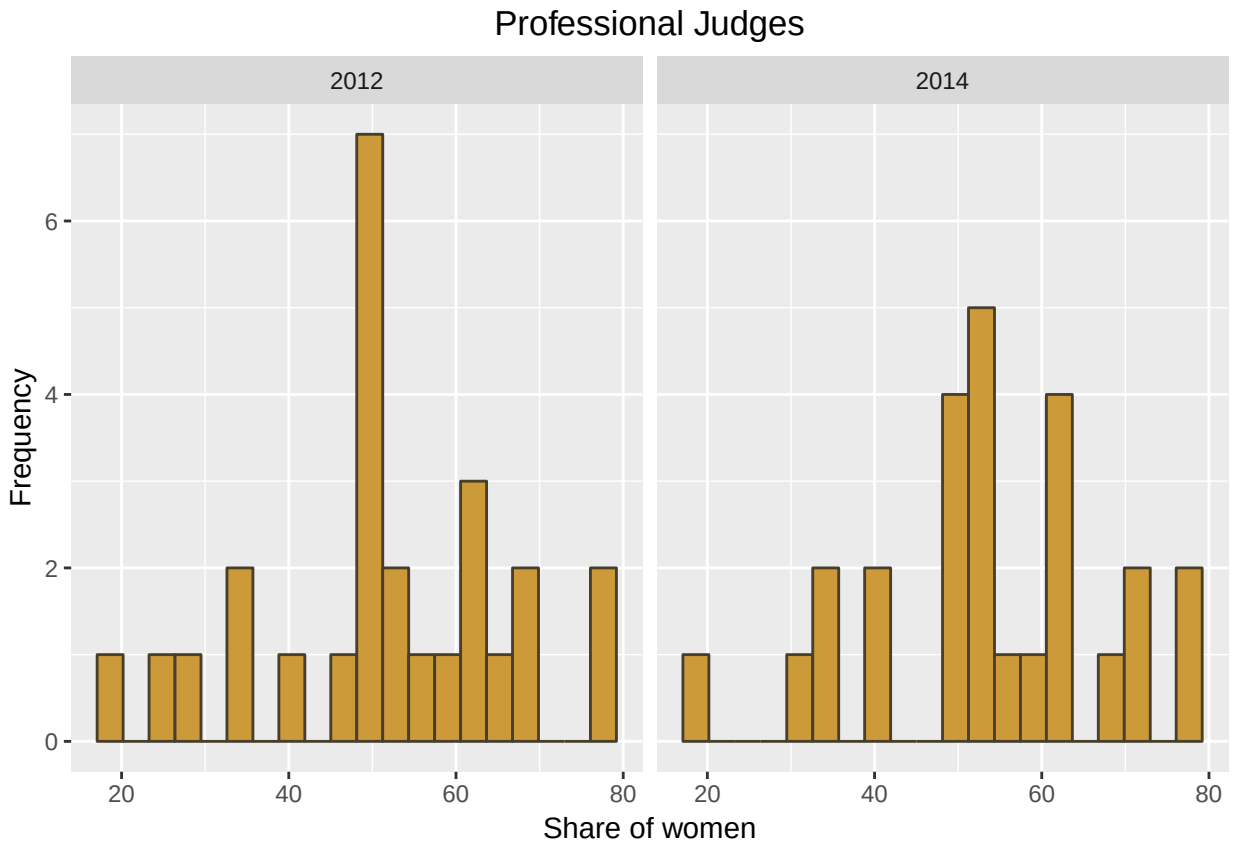


Histogram shows a distribution of women ministers for 2012 and 2015. We can see that more higher values in 2015 than 2012.There are very few countries with 50 percent women ministers. The cleveland dot plot shows a distribution of share of women ministers across countries. We see that higher share is in European countries like Sweden , Finland etc. Ironically, developed countries like Japan and Australia still have male dominant ministeries. We see a general trend of increase in women ministers from 2012 to 2015 for all countries

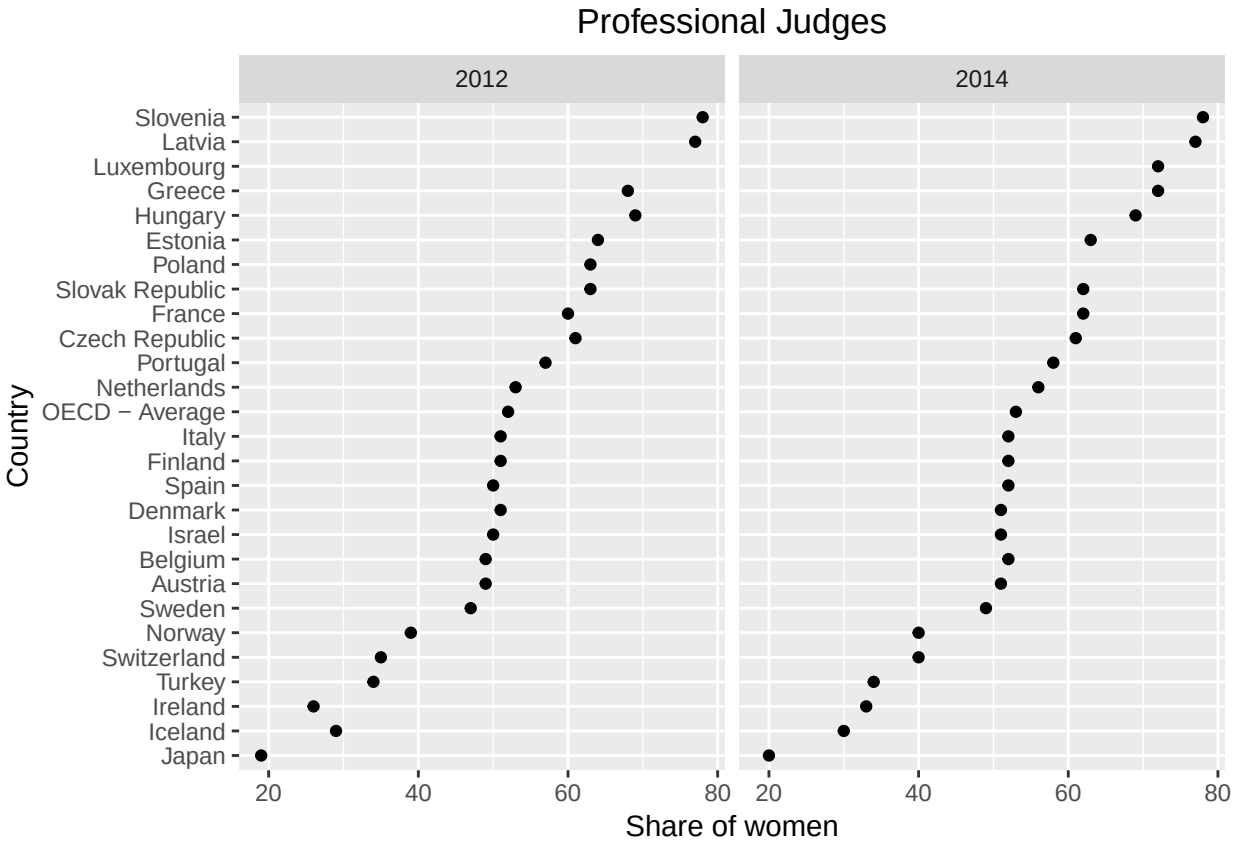
d) Share of professional judges that are women: The variable refers to the share of women judges for

different countries. The data is for years 2012 and 2014.

```
women_judges<- govt_data[govt_data$Indicator == "Share of professional judges that are women",]


ggplot(women_judges, aes(x = Value)) +
  geom_histogram(fill = "#cc9a38", color = "#473e2c", bins=20) +
  ggtitle("Professional Judges") +
  labs(x = "Share of women", y = "Frequency") +
  facet_grid(. ~ Year) +
  theme(plot.title = element_text(hjust = 0.5))
```



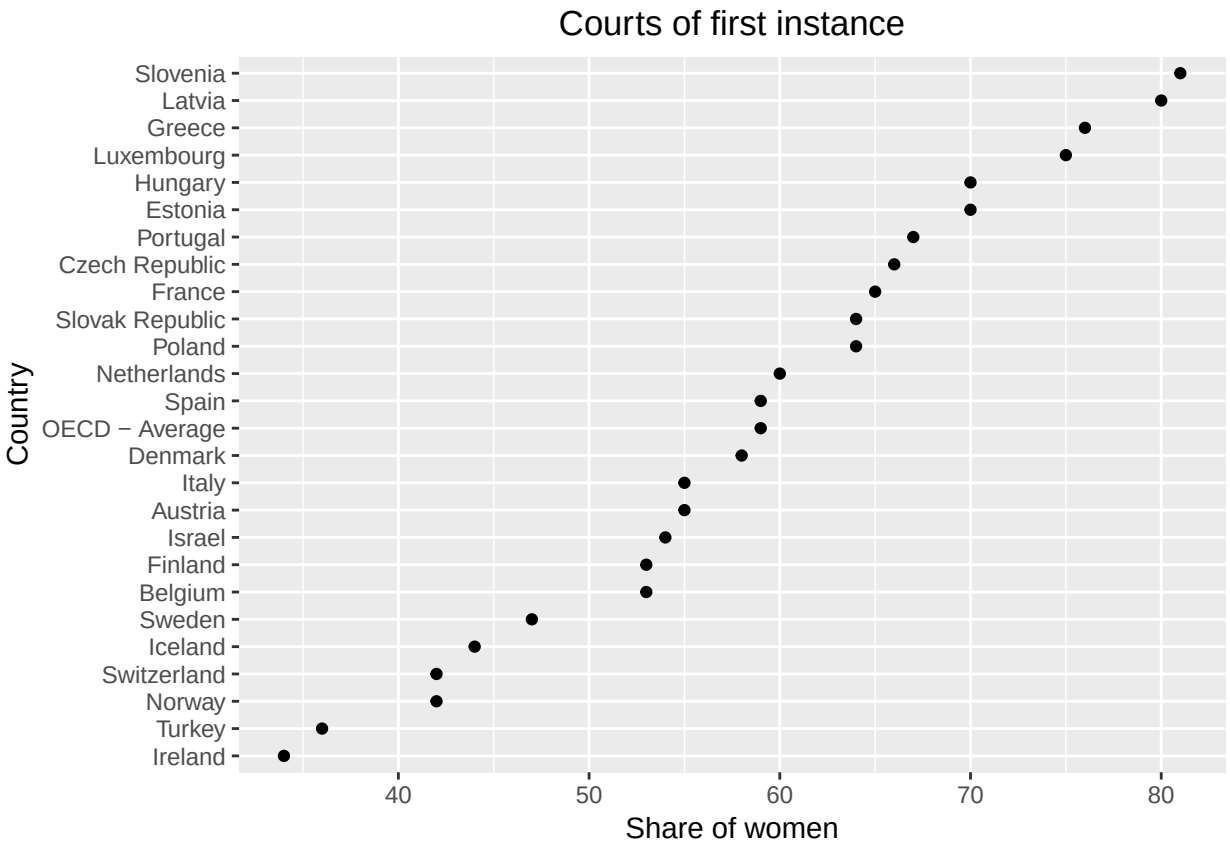

```
ggplot(women_judges, aes(y = reorder(Country, Value), x= Value)) +
  geom_point() +
  ggtitle("Professional Judges") +
  labs(x = "Share of women", y = "Country") +
  facet_grid(. ~ Year) +
  theme(plot.title = element_text(hjust = 0.5))
```



Histogram shows a distribution of women judges for 2012 and 2014. We can see that more higher values in 2014 than 2012.There are very few countries with 50 percent women judges. The cleveland dot plot shows a distribution of share of women judges across countries. We see that higher share is in European countries like Luxemborg , Greece etc. Ironically, developed countries like Japan and Ireland still have a dominance in male judges. We see a general trend of increase in women judges from 2012 to 2014 for all countries

e) Share of women in courts of first instance: The variable refers to the share of women in courts of first instance for different countries. The data is for years 2014.

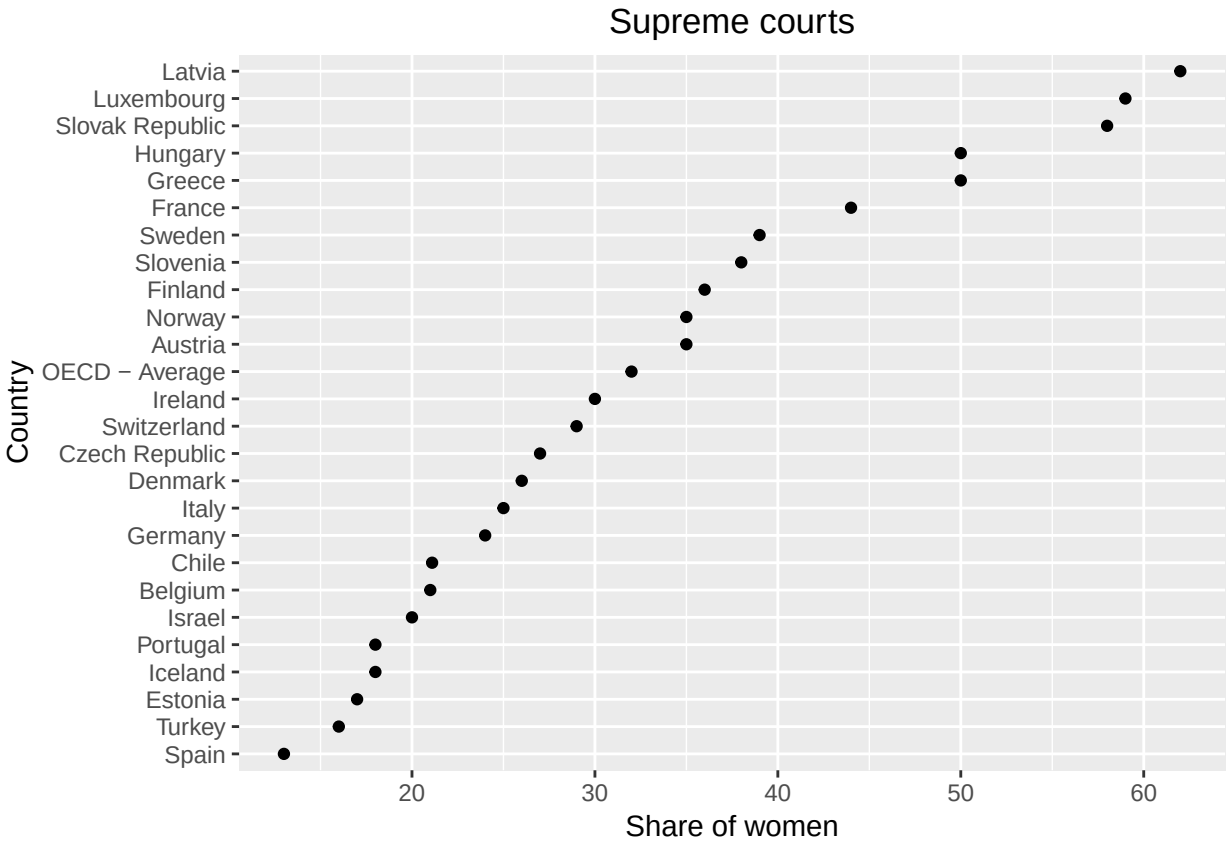
```
women_courts<- govt_data[govt_data$Indicator == "Share of women in courts of first instance",]

ggplot(women_courts, aes(y = reorder(Country, Value), x= Value)) +
  geom_point() +
  ggtitle("Courts of first instance") +
  labs(x = "Share of women", y = "Country") +
  theme(plot.title = element_text(hjust = 0.5))
```



The cleveland dot plot shows a distribution of share of women in courts of first instance across countries. We see that higher share is in European countries like Luxemborg , Greece etc. The data which was available to us is over 40 percent for all countries. Due limited data availability a comparison cannot be drawn or a direct conclusion cannot be made.

f) Share of women in supreme courts: The variable refers to the share of women in supreme courts for different countries. The data is for year 2014.

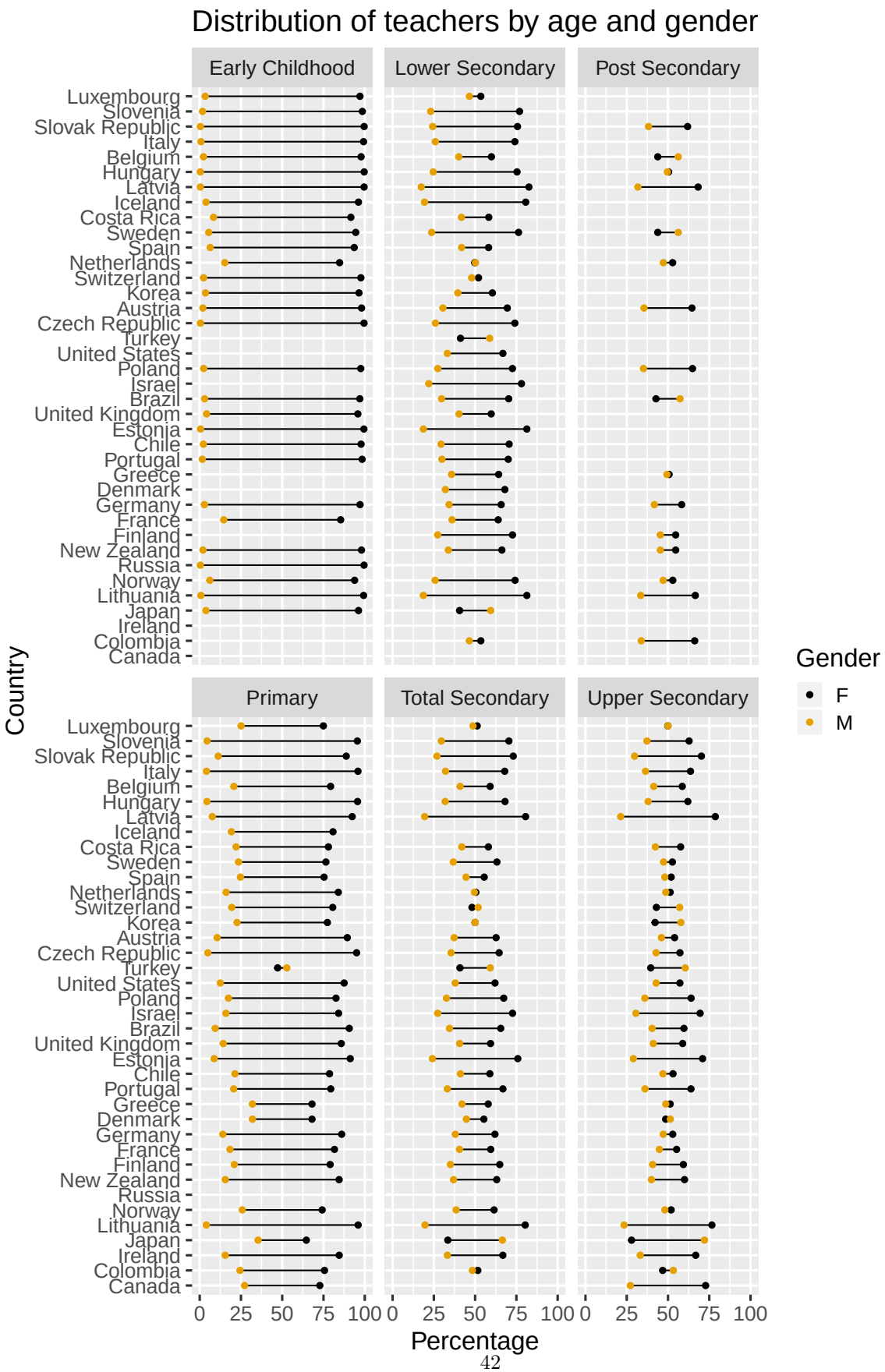
```
women_supreme<- govt_data[govt_data$Indicator == "Share of women in supreme courts",]

ggplot(women_supreme, aes(y = reorder(Country, Value), x= Value)) +
  geom_point() +
  ggtitle("Supreme courts") +
  labs(x = "Share of women", y = "Country") +
  theme(plot.title = element_text(hjust = 0.5))
```



The cleveland dot plot shows a distribution of share of women in supreme courts across countries. We see that higher share is in European countries like France , Greece etc. The data which was available to us is over 40 percent for most of the countries. Due limited data availability a comparison cannot be drawn or a direct conclusion cannot be made.

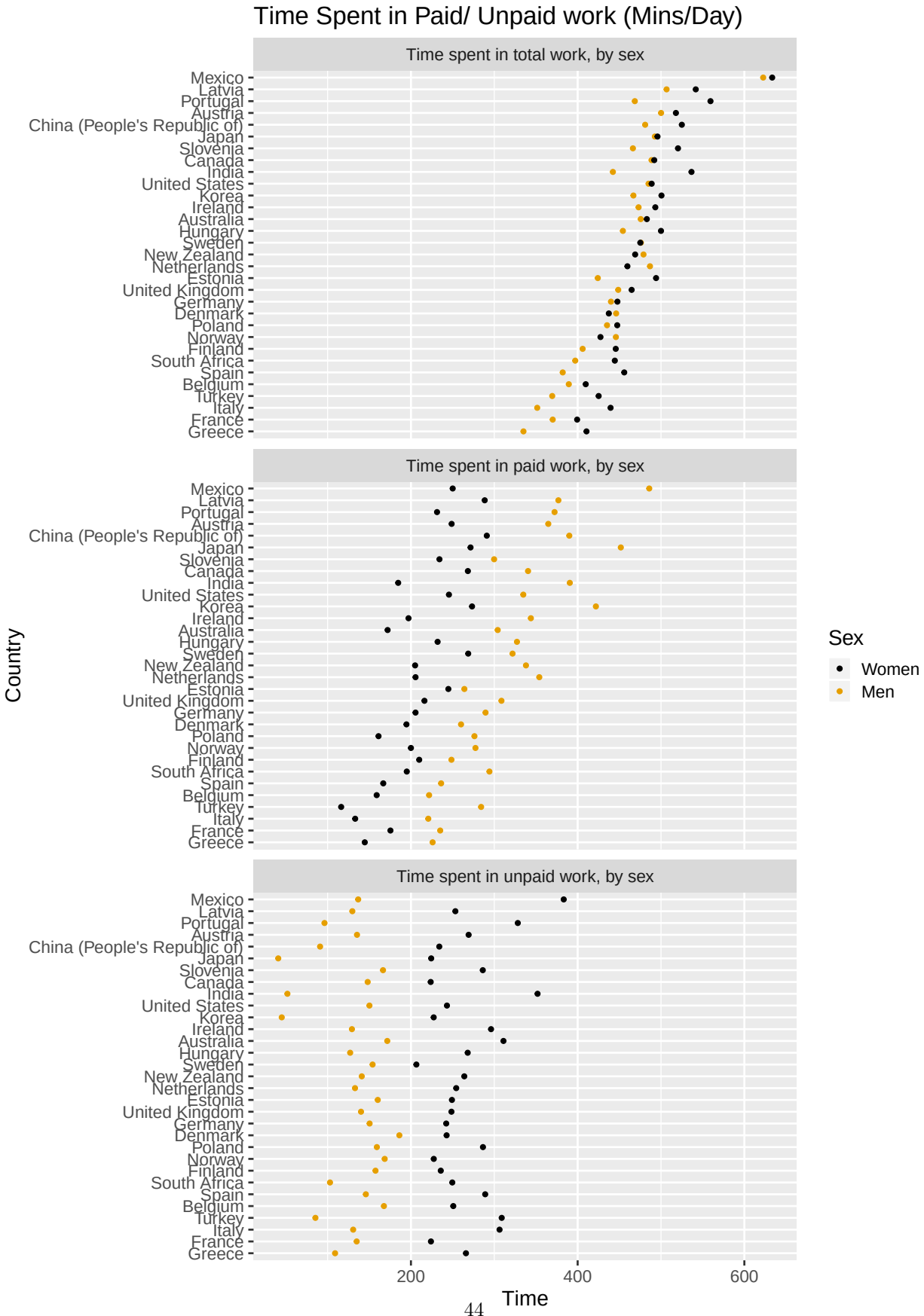
Executive Summary

We set out to explore the scenario of gender inequality in today's times, across multiple verticals of life. After performing a thorough exploratory analysis, we discovered some very interesting insights about the role and current state of women in our society, which we present below:


Distribution of teachers by age and gender
Early Childhood
Lower Secondary
Post Secondary
Primary
Total Secondary
Upper Secondary
Country
Luxembourg
Slovenia
Slovak Republic
Italy
Belgium
Hungary
Latvia
Iceland
Costa Rica
Sweden
Spain
Netherlands
Switzerland
Korea
Austria
Czech Republic
Turkey
United States
Poland
Israel
Brazil
United Kingdom
Estonia
Chile
Portugal
Greece
Denmark
Germany
France
Finland
New Zealand
Russia
Norway
Lithuania
Japan
Ireland
Colombia
Canada
0
25
50
75
100
Percentage
Gender
F
M

The visualization above projects the strong contribution of women in the education domain. The data shows that women surpass men in terms of providing education. Most of the formational education is provided by women while the ratio grows closer as the level of education increases.

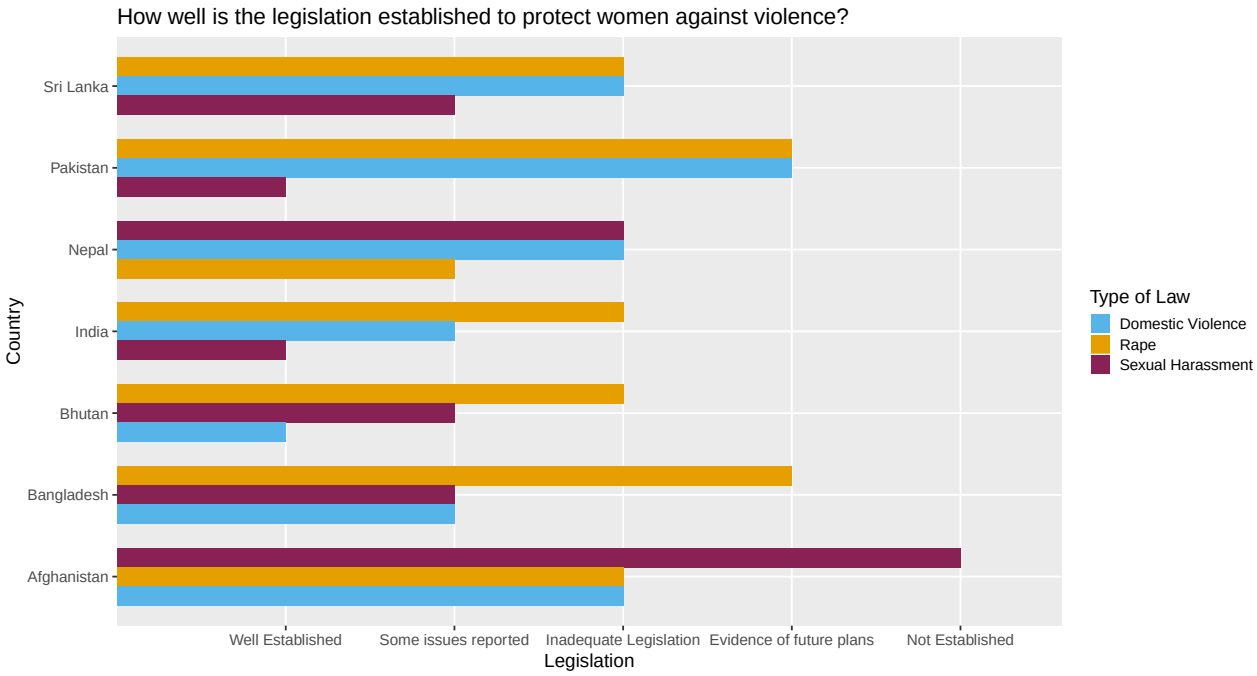
Are women appreciated enough for their contributions?



We can clearly see that women work slightly more than men do every day. But the stark difference we observe

is the amount of time they spend doing paid and unpaid work. Women in general do much more unpaid work than men do, and men spend more time doing work that they get paid for. We can conclude that women, even though they work for a longer time than men do, are not paid or appreciated enough for their work.

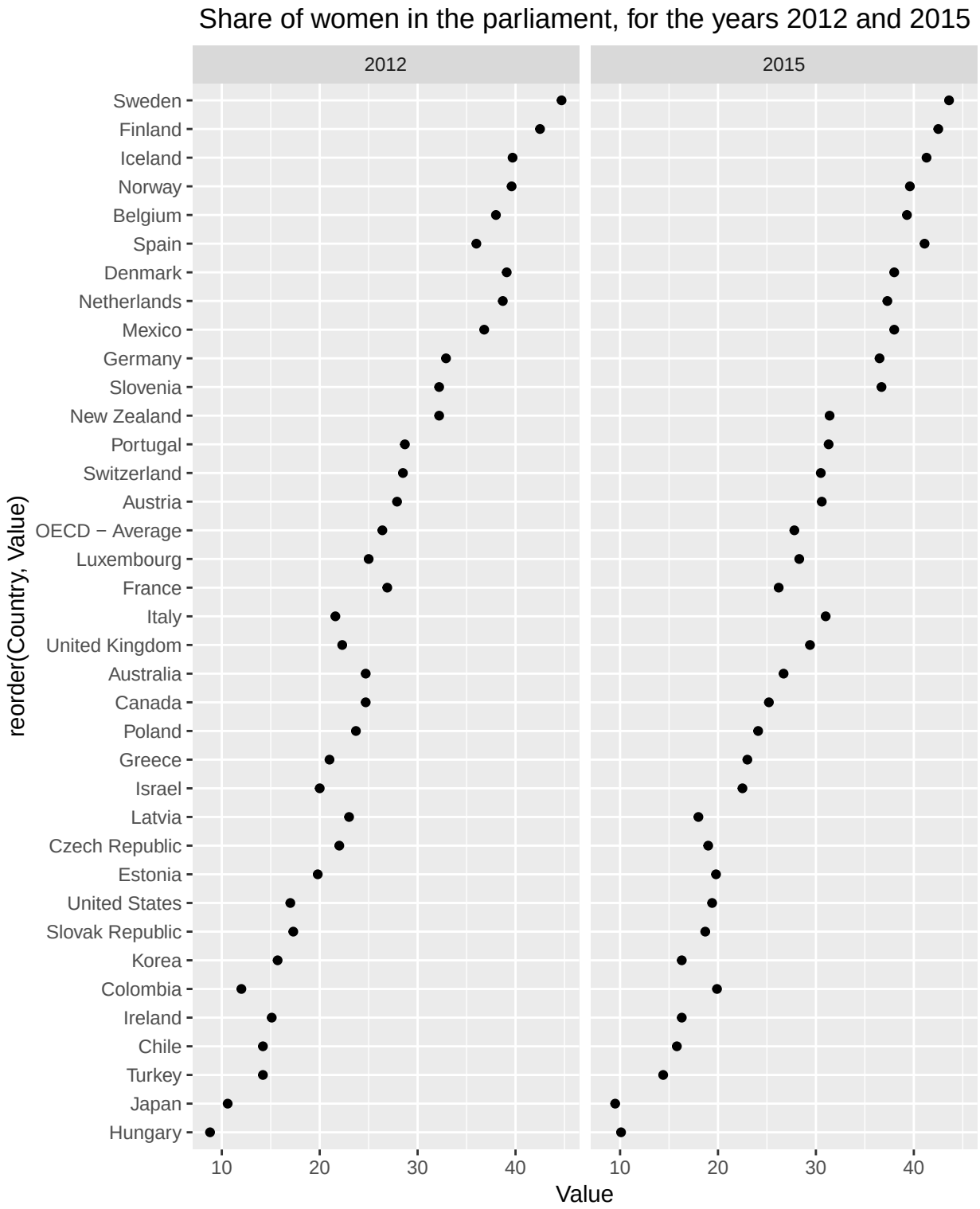
How well does a country's legislation protect women?



One would expect that the legislation for all three types of transgressions against women would be equally established in general. But as we can see from the countries shown above not all of them are given equal importance which makes the environment unsafe for women in some way or the other. All these transgressions are equally serious and should be taken in to consideration.

Are there enough women guiding the legislation to help other women?

```
ggplot(women_parl, aes(y = reorder(Country, Value), x= Value)) +
  geom_point() +
  ggtitle("Share of women in the parliament, for the years 2012 and 2015") +
  facet_grid(. ~ Year) +
  theme(plot.title = element_text(hjust = 0.5))
```



We can observe that some countries have a very good female representation in the parliament, but most of them do not. We can also see from the data that with time, this has barely improved. We believe that the better the representation women have in the parliament, the better they will be able to deal with and improve the condition of gender inequality.

Based on the above observations, it is clear that gender inequality is no where near close to eradication. Women work just as hard, in some cases, even harder, but we do not appreciate them like we should. We

have failed to create a safe space for them in this world, and we should put more efforts into empowering women so that they become equal stakeholders in the society.

Interactive Component

The objective of the interactive component is to see the gender inequality parameters across the world. The verticals under consideration are Employment, Governance, Education and Development. We present a spatial representation of all these parameters on a world map in the form of leaflet plot. There is a menu to select a parameter and within that parameter, there is. a further list of options. Illustratively, one parameter is Governance and within that we can select options such as share of women parliamentarians, share of women ministers etc. A summary of the selected option is displayed at the top of the screen. On clicking a circle on the map, it displays the name of the country and the corresponding value. The Colour Scheme is a convergence Reds which has light color for low values and dark for high values.
Link: https://edav2018.shinyapps.io/gender_inequality/

Conclusion

The data that we had was not available for all countries but a smaller set of OECD countries. Hence, the patterns and outcomes of the analysis may be biased in some way or the other. The data that we had was categorized by geographies, but the same analysis could be performed by categorizing the data based on socio-economic and demographic factors. This kind of an analysis could provide interesting insights as we believe that gender inequality could be dependent on these factors. All our data is at the country-level, and more likely than not, these values may not be consistent across urban, suburban and rural areas, so we could also explore patterns on gender inequality on a more granular level. As a part of our future work, we could perform a better and more accurate analysis as more of this data is made publicly available.

Some of the key learnings while implementing this projects are as follows - a) The transformation of data can bring out very interesting insights that may not be apparent in the data without any transformation.

b) A visualization can communicate a lot information about the data making it very helpful for analysis and summarizing the data.

c) Simple interactive applications like the one developed on Shiny could be really helpful for analysis compared to a static visualization. A lot of information can be depicted in these applications with the option of slicing and filtering data according to the user.

d) Data from different verticals and cross sections could be combined to convey one narrative in a story.