MiR Training Session

Getting to know ggplot2

Shelmith Nyagathiri Kariuki

July 03, 2020

Contents

Dataset: Financial Inclusion in Africa

0.1 Set up

```
knitr::opts_chunk$set(echo = TRUE,
                      message=FALSE,warning=FALSE,
                      fig.width = 8, fig.height = 6)
```

0.2 Install and load the packages required

```
###  create a vector of packages to be installed
pkgs <- c("tidyverse","DT","lubridate","patchwork")

###  Check if there are packages you want to load, that are not already installed.
```

```
miss_pkgs <- pkgs[!pkgs %in% installed.packages()[,1]]

###  Installing the missing packages
if(length(miss_pkgs)>0){
  install.packages(miss_pkgs)
}

###  Loading all the packages
invisible(lapply(pkgs,library,character.only=TRUE))

###  Remove the objects that are no longer required
rm(miss_pkgs)
rm(pkgs)
```

```
# Set the theme
# mir_theme <- theme(plot.title = element_text(size = 12, #18
#                                              #family = "Source Sans Pro Semibold",
#                                              face = "italic", hjust = 0.5),
#                axis.line = element_line(color = "black", size = 1),
#                axis.title = element_text(size = 16),
#                axis.text = element_text(size = 14),
#                panel.background = element_rect(fill = NA),
#                plot.caption = element_text(size = 14),
#                legend.title = element_blank(),
#                legend.position = "bottom")

mir_theme <- theme(plot.title = element_text(size = 12, #18
                                             #family = "Source Sans Pro Semibold",
                                             face = "italic", hjust = 0.5),
               axis.line = element_line(color = "black", size = 1),
               axis.title = element_text(size = 10),
               axis.text = element_text(size = 9),
               panel.background = element_rect(fill = NA),
               plot.caption = element_text(size = 8),
               legend.title = element_blank(),
               legend.position = "bottom")
```

1. Read in the dataset

```
df <- read_csv("Train_v2.csv")
```

2. Convert character variables to factor variables

```
## 2.1. bank_account
df <- df %>%
  mutate(bank_account = fct_relevel(bank_account, "No", "Yes"))

## 2.2. location_type
df <- df %>%
  mutate(location_type = fct_relevel(location_type, "Rural" , "Urban"))

## 2.3. cellphone_access
df <- df %>%
```

```
  mutate(cellphone_access = fct_relevel(cellphone_access,  "No", "Yes"))

## 2.4. gender_of_respondent
df <- df %>%
  mutate(gender_of_respondent = fct_relevel(gender_of_respondent,  "Female" ,"Male"))


## 2.5. relationship_with_head
df <- df %>%
  mutate(relationship_with_head = fct_relevel(relationship_with_head,
        "Child","Spouse", "Parent", "Head of Household",
        "Other relative",   "Other non-relatives" ))

## 2.6. marital_status
df <- df %>%
  mutate(marital_status = fct_relevel(marital_status,
        "Single/Never Married","Divorced/Seperated","Widowed",
        "Married/Living together","Dont know"))

## 2.7. education_level
df <- df %>%
  mutate(education_level = fct_relevel(education_level,
        "No formal education", "Primary education", "Secondary education",
        "Tertiary education", "Vocational/Specialised training","Other/Dont know/RTA"))

## 2.8. job_type
```

3. Generate 2 new variables i.e monthly income and date they registered for xyz

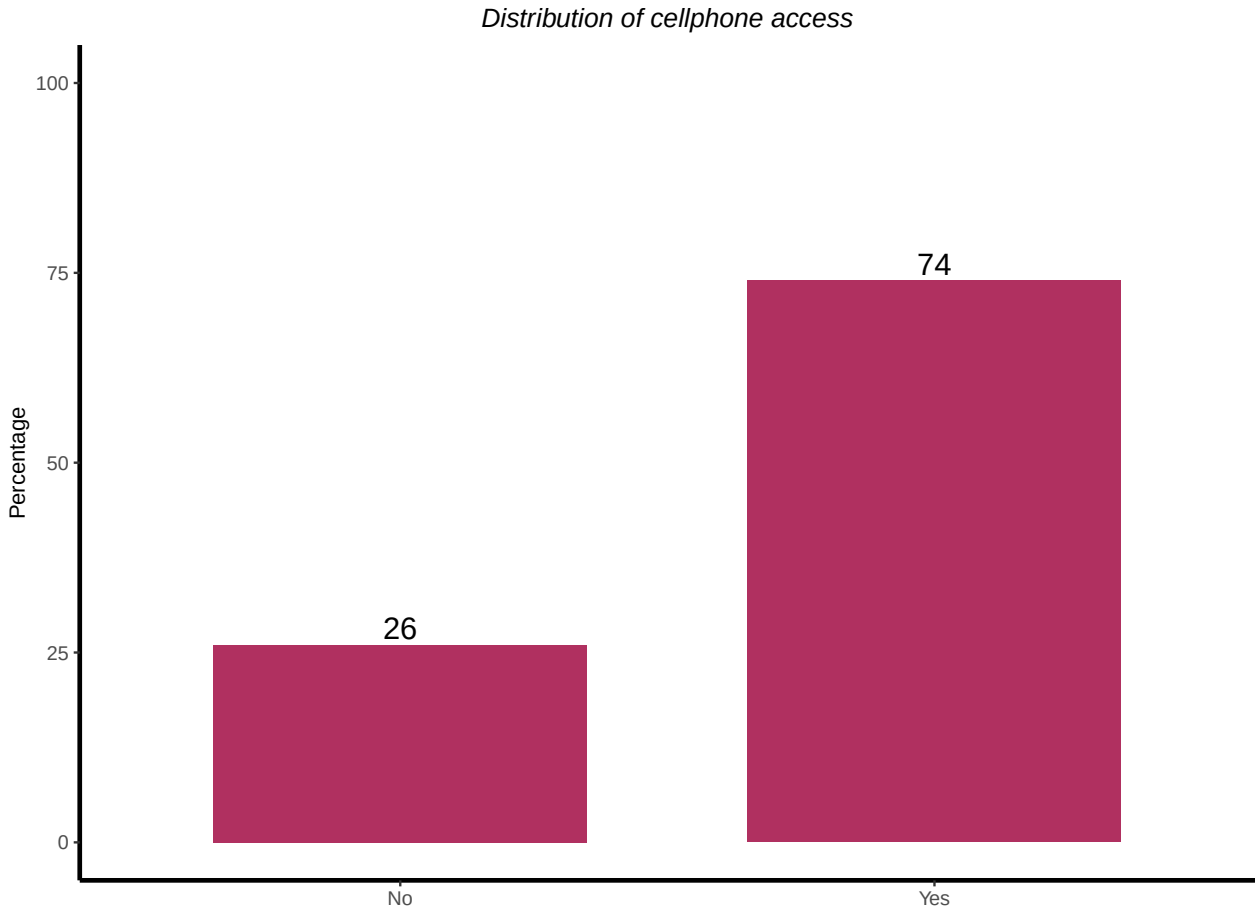
```
 ## 3.1 monthly income

set.seed(2020)
income_values <- sample(c(5000 : 150000), nrow(df))

df <- df %>%
  mutate(income = income_values,
         income = ifelse(job_type == "No Income", NA, income))

## 3.2 date of registering for xyz
date_vec <- sample(seq(as.Date('2014/01/01'), as.Date('2014/12/31'), by="day"), nrow(df), replace = T)

df <- df %>%
  mutate(date = date_vec)

## Generate the month
df <- df %>%
  mutate(month = month(date, abbr = T, label = T))
```

4. Main graphs

4.1 Bar graph

4.1.1 Single categorical variable Distribution of cellphone access

```
## First generate a table
tab1 <- df %>%
  group_by(cellphone_access) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = round(n/sum(n) *100, 0))

## Plot a graph
plot1 <- ggplot(data = tab1, aes(x = cellphone_access, y = perc))+
  geom_bar(stat = "identity", fill = "maroon", width = 0.7)+
  geom_text(aes(label = perc), size = 5, hjust = 0.5, vjust = -0.25)+
  mir_theme +
  labs(title = "Distribution of cellphone access",
       x = "",y = "Percentage")+
       #caption = "Twitter:@Shel_Kariuki")+
  ylim(c(0,100))
plot1
```

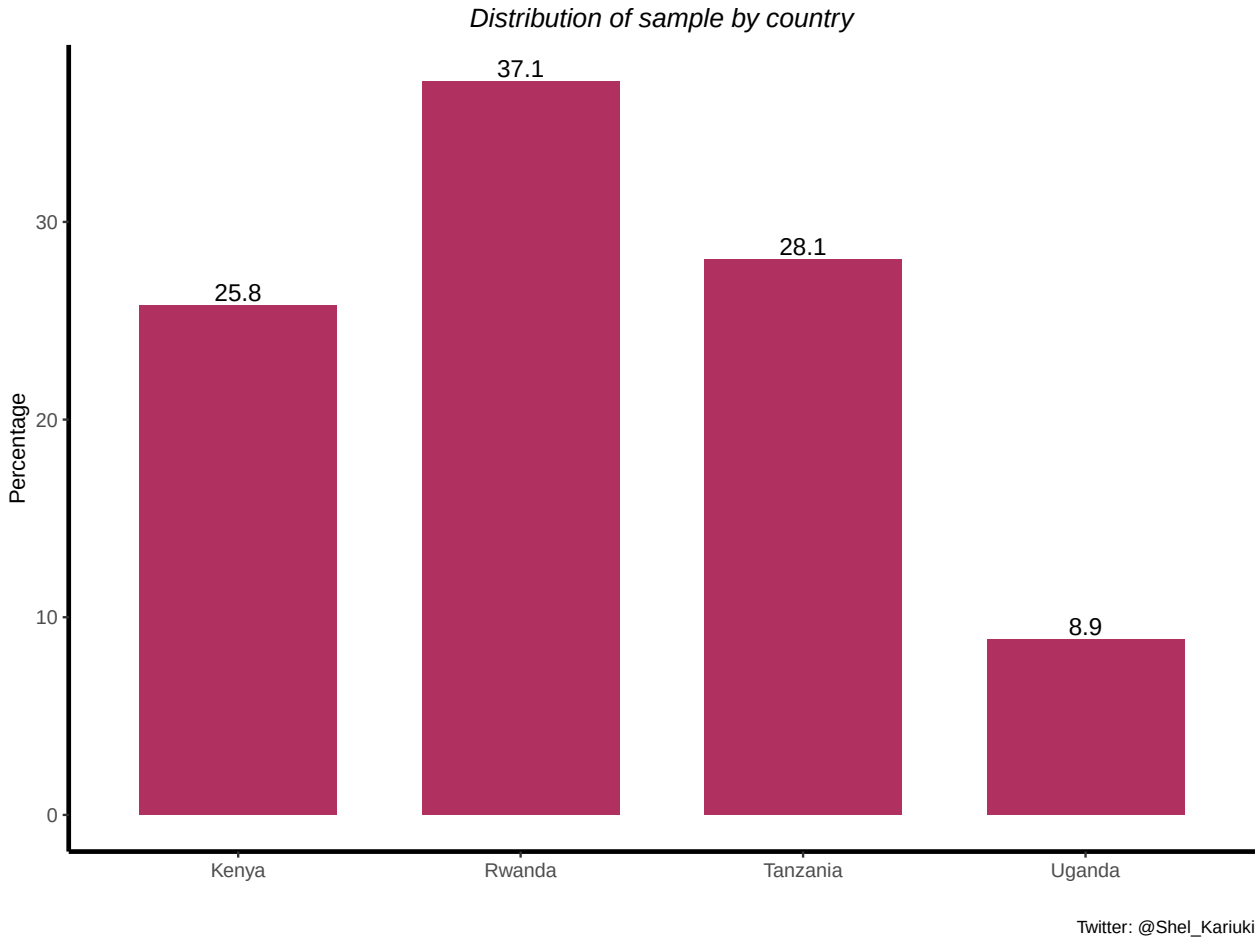



```
#rstudio_blue <- "#4AA4DE"
```

Distribution of sample by country

```
## table

tab2 <- df %>%
  group_by(country) %>%
```

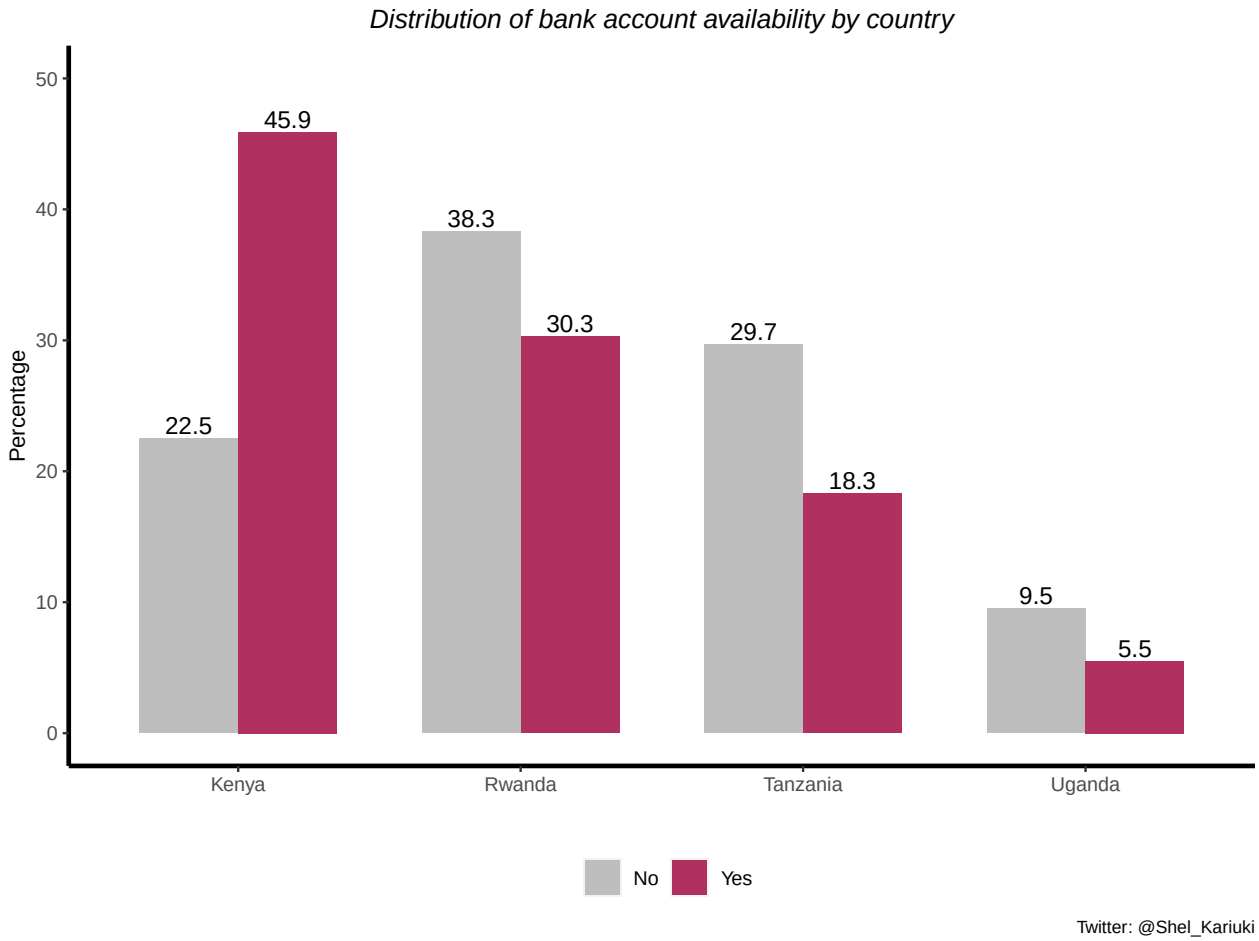
```
  count() %>%
  ungroup() %>%
  mutate(perc = round(n / sum(n) *100,1))

## graph

plot2 <- ggplot(data = tab2, aes(x = country, y=perc)) +
           geom_bar(stat = "identity", fill = "maroon", width = 0.7) +
          geom_text(aes(label = perc), size = 4 , hjust = 0.5 , vjust = -0.25)+
          mir_theme +
        labs(title = "Distribution of sample by country",
             x = "", y = "Percentage",
             caption = "Twitter: @Shel_Kariuki")
plot2
```



4.1.2 Two categorical variables (main variable and grouping variable) Distribution of bank account availability by country

```
## table

tab3 <- df %>%
  group_by(country, bank_account) %>%
  count() %>%
  ungroup() %>%
```

```
  group_by(bank_account) %>%
  mutate(perc = round(n / sum(n) *100,1))

## graph

plot3 <- ggplot(data = tab3, aes(x = country, y=perc, fill = bank_account)) +
           geom_bar(stat = "identity",  width = 0.7, position = "dodge") +
          geom_text(aes(label = perc), size = 4 , hjust = 0.5 , vjust = -0.25,
                    position = position_dodge(width = 0.7))+
          mir_theme +
        scale_fill_manual(values = c("grey", "maroon"))+
        labs(title = "Distribution of bank account availability by country",
             x = "", y = "Percentage",
             caption = "Twitter: @Shel_Kariuki")+
   ylim(c(0, 50))
plot3
```



4.2 Line graph

4.2.1 Single categorical variable Distribution of registrations by month

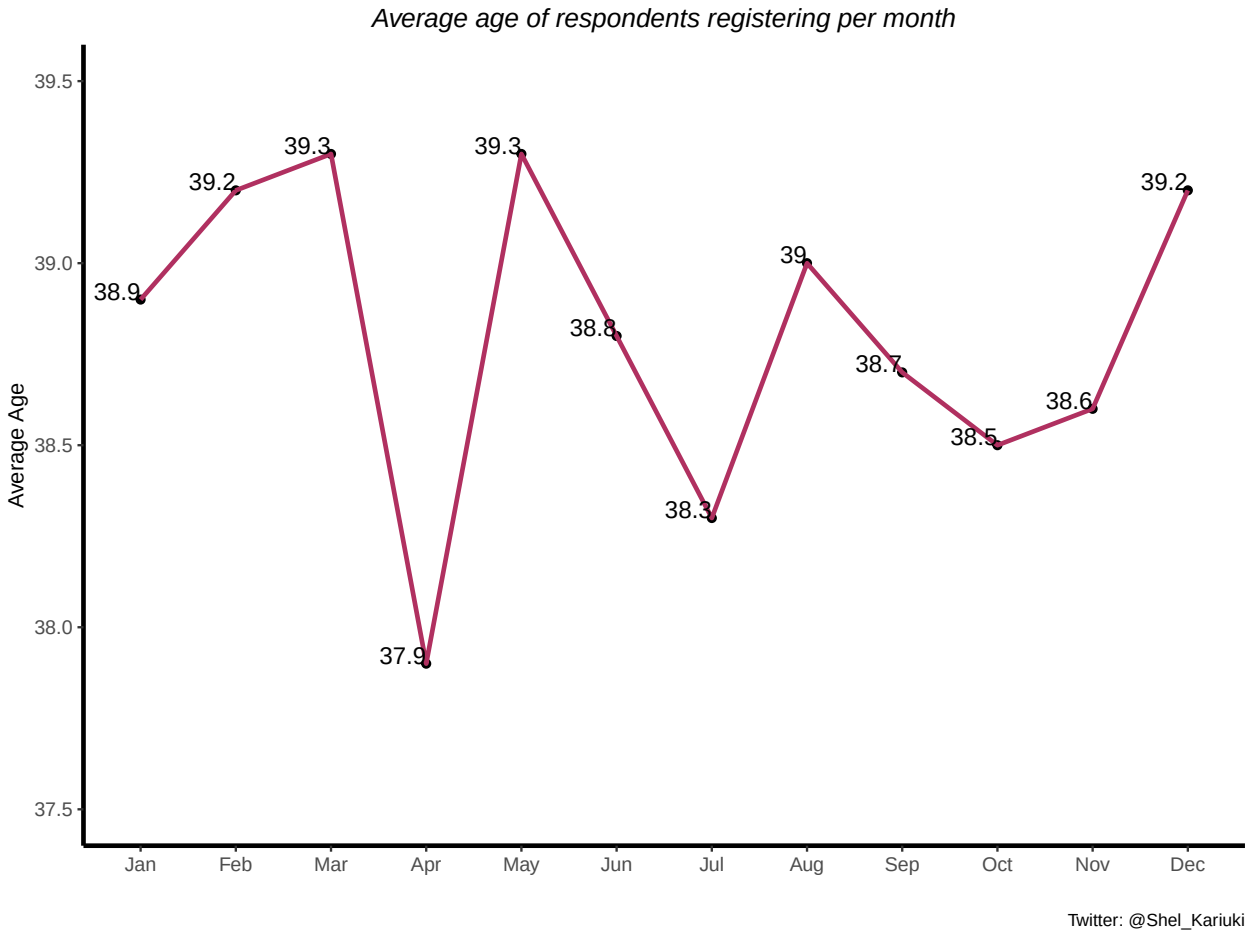
```
## table
tab4 <- df %>%
  group_by(month) %>%
```

```
  count() %>%
  ungroup() %>%
  mutate(perc = round(n/sum(n)*100,1))

## graph
plot4 <- ggplot(data = tab4, aes(x = month, y = perc, group = 1)) +
          geom_point()+
          geom_line(stat = "identity", color = "maroon", size = 1)+
          geom_text(aes(label = perc), size = 5, hjust = 0.5, vjust = -0.25)+
          mir_theme +
          labs(title = "Distribution of registrations by month",
               x = "", y = "Percentage",
               caption = "Twitter: @Shel_Kariuki")+
  ylim(c(7.5,9))

plot4
```

Average age of respondents registering per month

```
 ## table
tab5 <- df %>%
  group_by(month) %>%
  summarise(avg_age = round(mean(age_of_respondent, na.rm = T),1))

## plot
plot5 <- ggplot(data = tab5, aes(x = month, y = avg_age, group = 1))+
          geom_point()+
          geom_line(stat = "identity", size = 1, color = "maroon", linetype = "solid")+
          geom_text(aes(label = avg_age), vjust = 0, hjust = 1)+
          mir_theme +
```

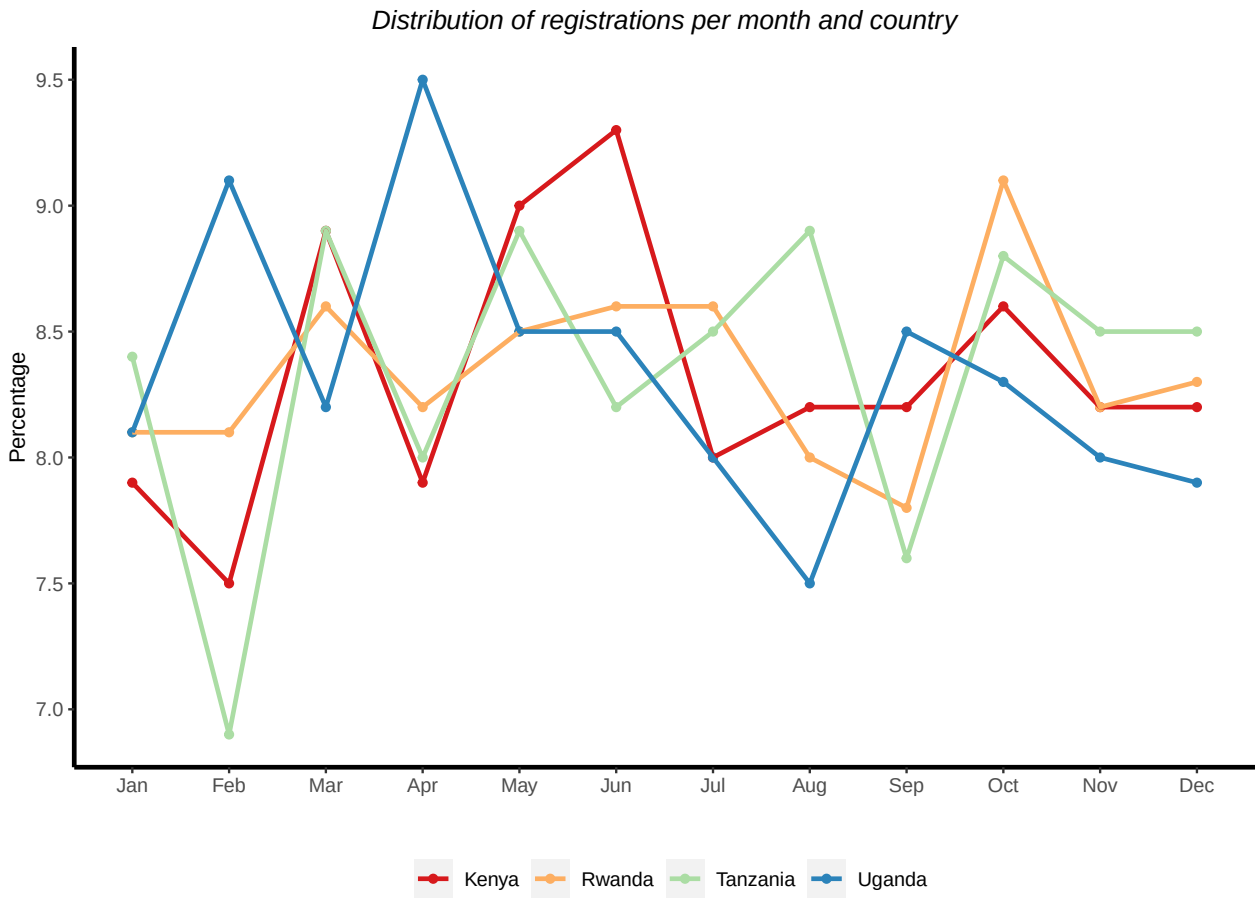
```
          labs(title = "Average age of respondents registering per month",
               x = "", y = "Average Age",
               caption = "Twitter: @Shel_Kariuki")+
  ylim(c(37.5,39.5))
plot5
```



4.2.2 Two categorical variables (main variable and grouping variable) Distribution of registrations per month and country

```
## table
tab6 <- df %>%
  group_by(month, country) %>%
  count() %>%
  ungroup() %>%
  group_by(country) %>%
  mutate(perc = round(n/sum(n)*100,1))

## plot
plot6 <- ggplot(data = tab6, aes(x = month, y = perc, group = country, color = country))+
          geom_point()+
          geom_line(stat = "identity", size = 1, linetype = "solid")+
          #geom_text(aes(label = perc), vjust = 0, hjust = 1)+
          mir_theme +
  scale_color_brewer(palette = "Spectral")+
```

```
          labs(title = "Distribution of registrations per month and country",
               x = "", y = "Percentage",
               caption = "Twitter: @Shel_Kariuki")

plot6
```



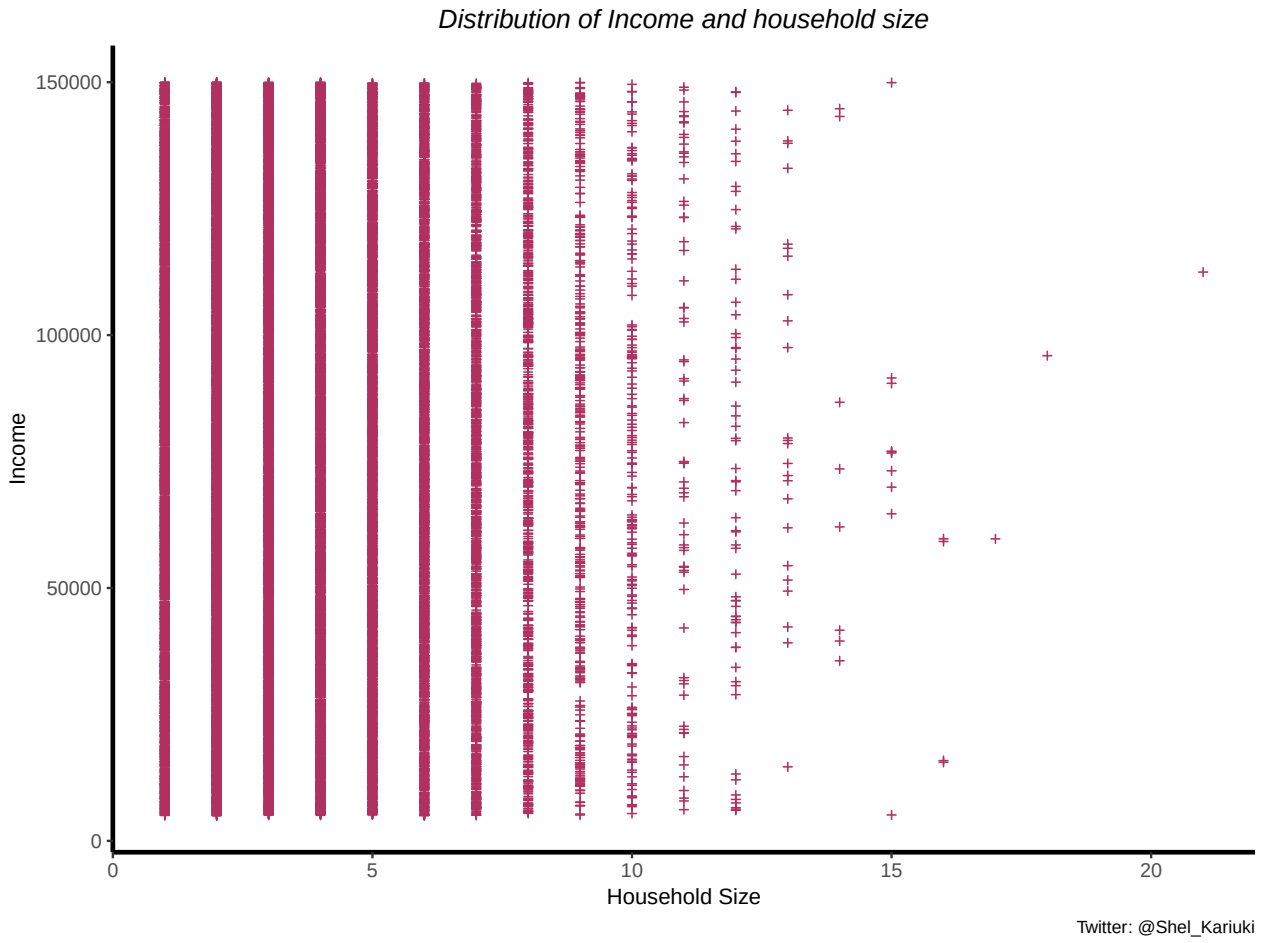

```
## table
# tab6 <- df %>%
#   group_by(month, country) %>%
#   count() %>%
#   ungroup() %>%
#   group_by(country) %>%
#   mutate(perc = round(n/sum(n)*100,1))
#
# ## graph
# plot6 <- ggplot(data = tab6, aes(x = month, y = perc, group = country, color = country))+
#           geom_point()+
#           geom_line(stat = "identity", size = 1)+
#           #geom_text(aes(label = perc), hjust = 1.5, vjust = 0)+
#           mir_theme+
#           scale_color_brewer(palette = "Spectral")+
#           labs(title = "Registrations by month",
#                x = "", y = "Percentage",
#                caption = "Twitter: @Shel_Kariuki")
```

```
# plot6
```

4.3 Scatter plot

4.3.1 2 continuous variables Distribution of Income and household size

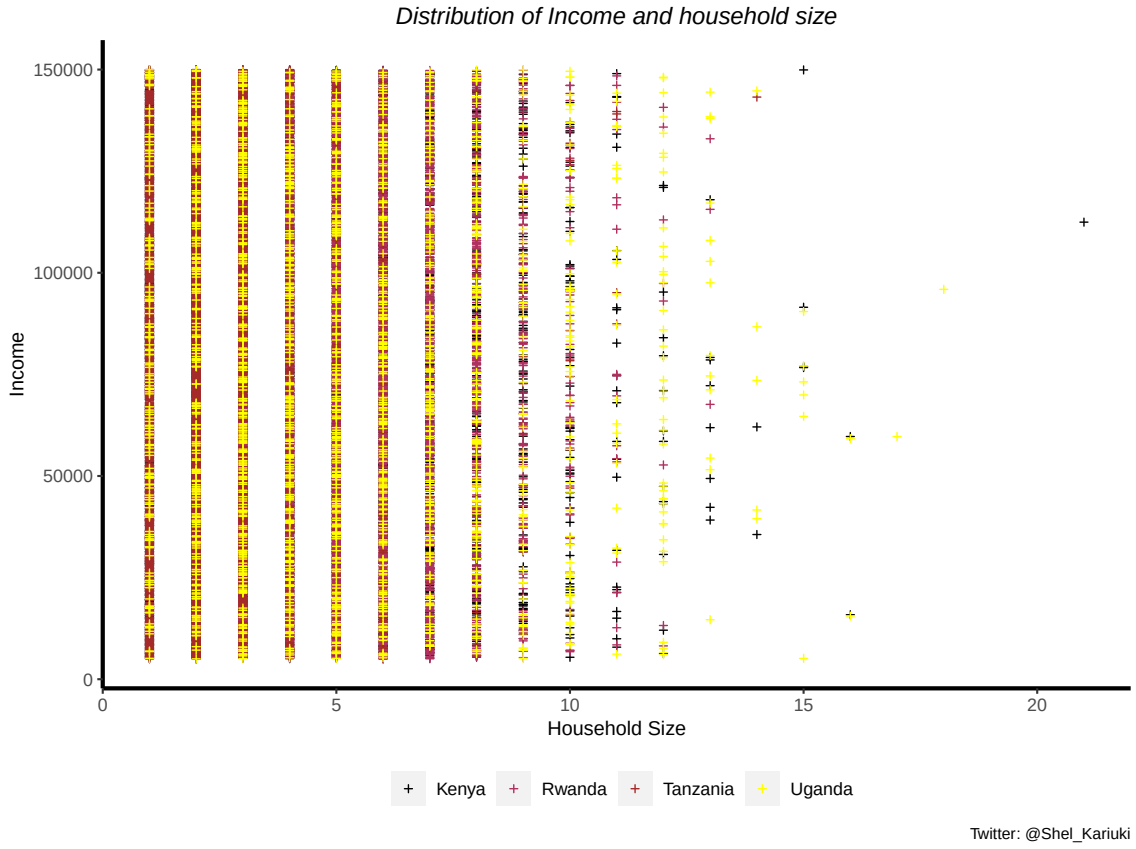
```
plot7 <- ggplot(data = df, aes(x = household_size,  y = income))+
          geom_point(size = 1, color = "maroon", shape = 3)+
          mir_theme +
  labs(title = "Distribution of Income and household size",
       x = "Household Size", y = "Income",
       caption = "Twitter: @Shel_Kariuki")
plot7
```




```
plot8 <- ggplot(data = df, aes(x = household_size,  y = income, color = country))+
          geom_point(size = 1, shape = 3)+
          mir_theme +
          scale_color_manual(values = c("black","maroon","brown","yellow"))+
          labs(title = "Distribution of Income and household size",
            x = "Household Size", y = "Income",
          caption = "Twitter: @Shel_Kariuki")
```

```
plot8
```

4.3.2 2 continuous variables and a grouping variable

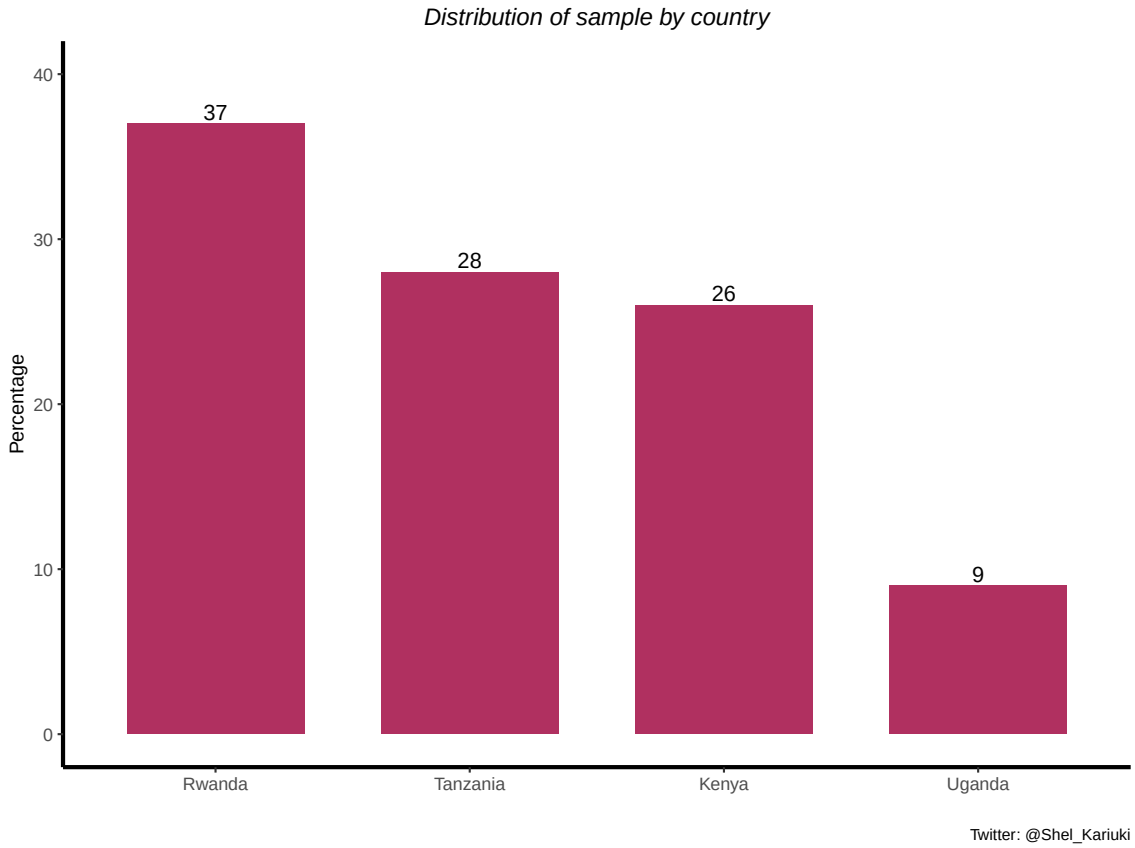


5. Other neat tricks

```
## table
tab2 <- df %>%
  group_by(country) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = round( n/ sum(n)*100,0))
```

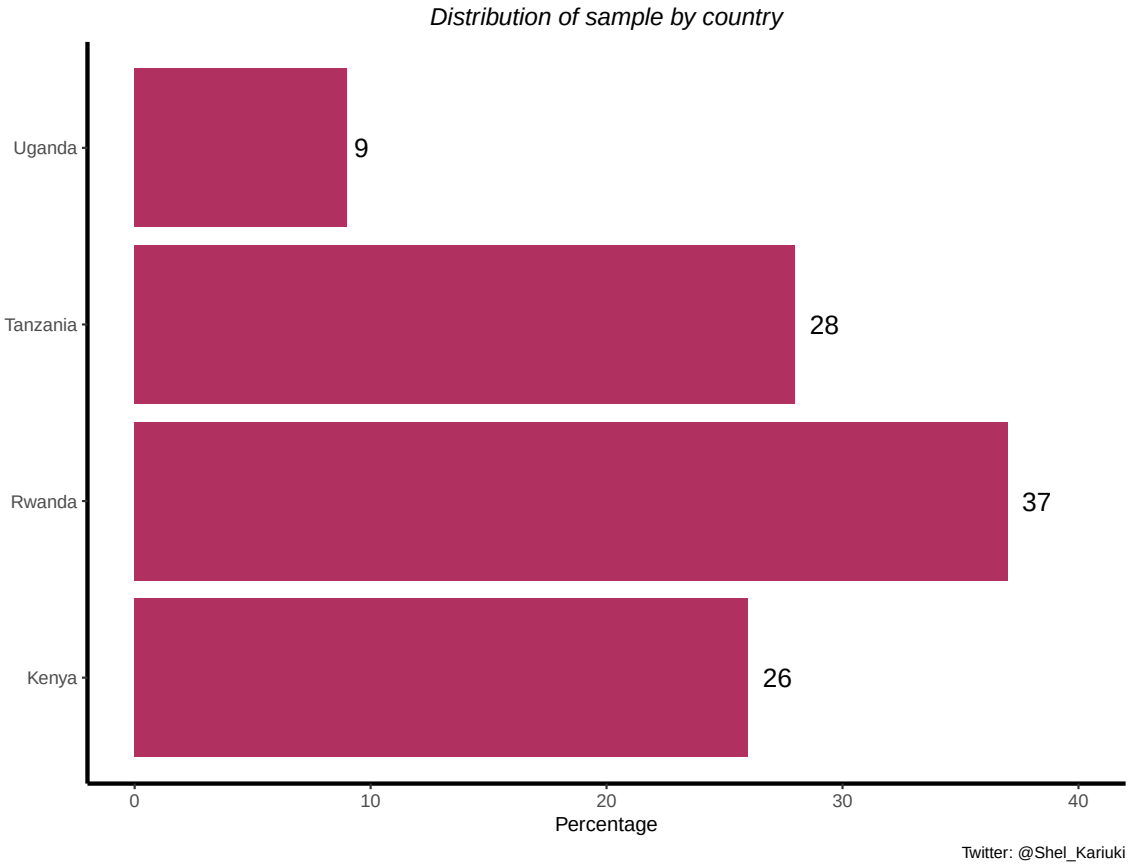
```
## graph

plot2b <- ggplot(data = tab2, aes(x = reorder(country,-perc), y=perc)) +
            geom_bar(stat = "identity", fill = "maroon", width = 0.7) +
          geom_text(aes(label = perc), size = 4 , hjust = 0.5 , vjust = -0.25)+
          mir_theme +
        labs(title = "Distribution of sample by country",
             x = "", y = "Percentage",
             caption = "Twitter: @Shel_Kariuki")+
        ylim(c(0,40))
plot2b
```

5.1 Reordering bar graphs by ascending order of the y axis




```
 ## graph
plot2c <- ggplot(data = tab2, aes(x = country, y = perc ))+
          geom_bar(stat = "identity", fill = "maroon")+
          geom_text(aes(label = perc), hjust = -0.5, vjust = 0.5, size = 4.5)+
          mir_theme+
          labs(title = "Distribution of sample by country",
               x = "", y = "Percentage",
               caption = "Twitter: @Shel_Kariuki")+
         coord_flip()+
         ylim(c(0,40))
plot2c
```

5.2 Flipping graphs

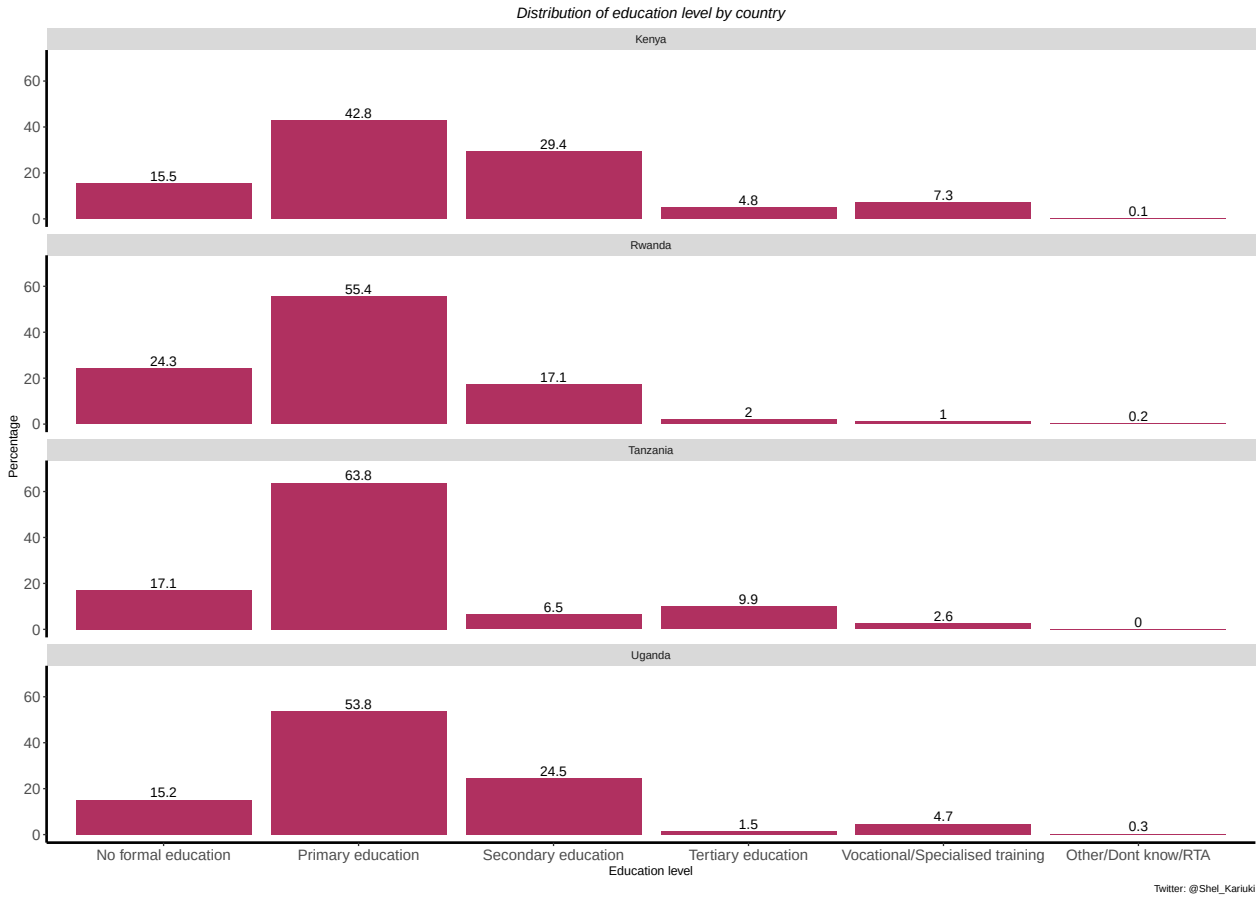


5.3 Faceting plots

5.3.1 Facet wrap: faceting a plot by one variable

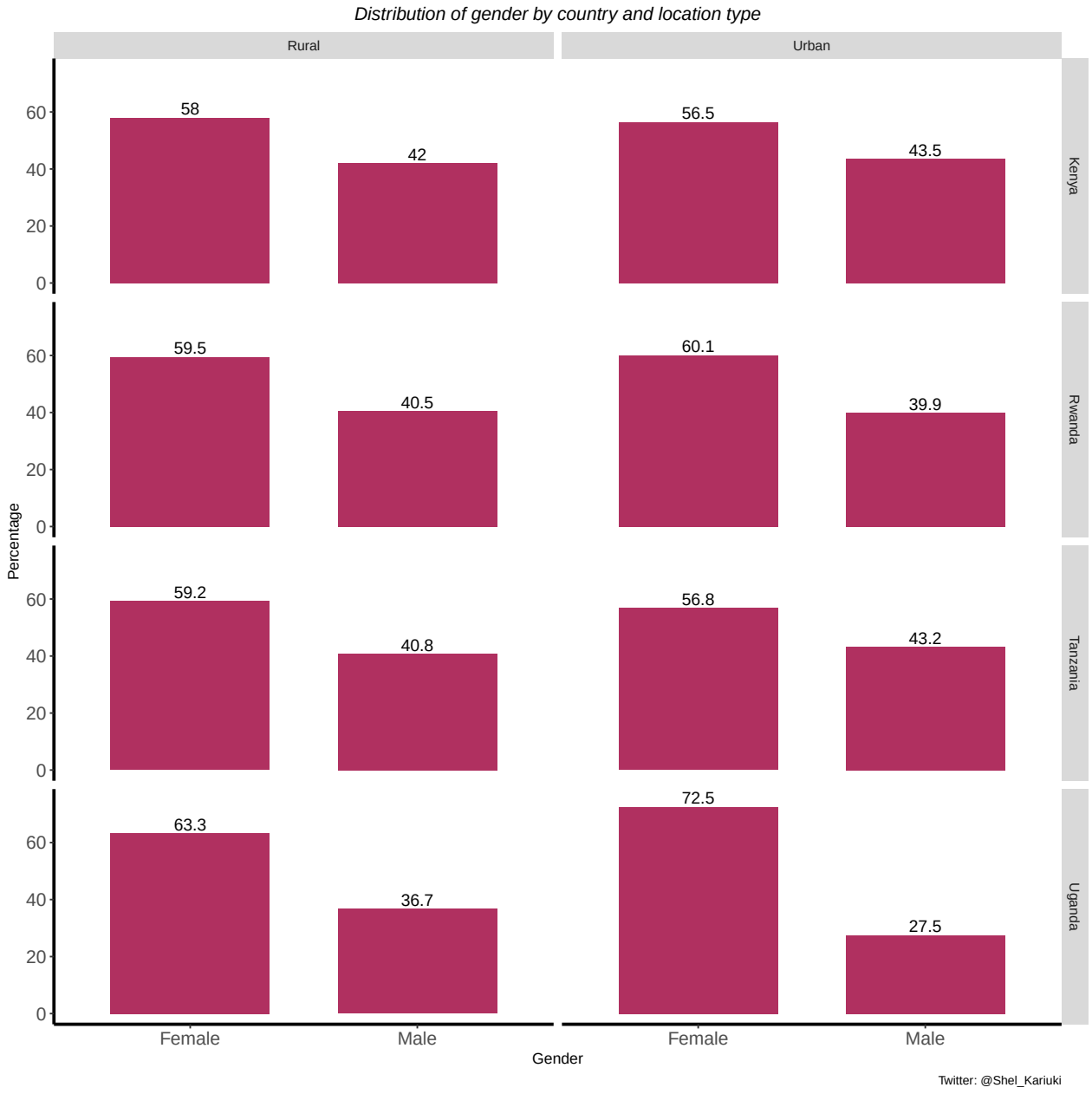
Distribution of education level by country

```
## table
tab9 <- df %>%
  group_by(education_level, country) %>%
  count() %>%
  ungroup() %>%
  group_by(country) %>%
  mutate(perc = round(n/sum(n) *100,1))

##plot
plot9 <- ggplot(data = tab9, aes(x = education_level, y = perc))+
          geom_bar(stat = "identity", fill = "maroon")+
          geom_text(aes(label = perc), vjust = -0.25, hjust = 0.5)+
          mir_theme+
          theme(axis.text = element_text(size = 12))+
          labs(title = "Distribution of education level by country",
               x = "Education level", y = "Percentage",
               caption = "Twitter: @Shel_Kariuki")+
          facet_wrap(~country, ncol = 1)+
          ylim(c(0,70))
plot9
```



5.3.2 Facet grid: Faceting a plot by two variables Distribution of gender by country and location type

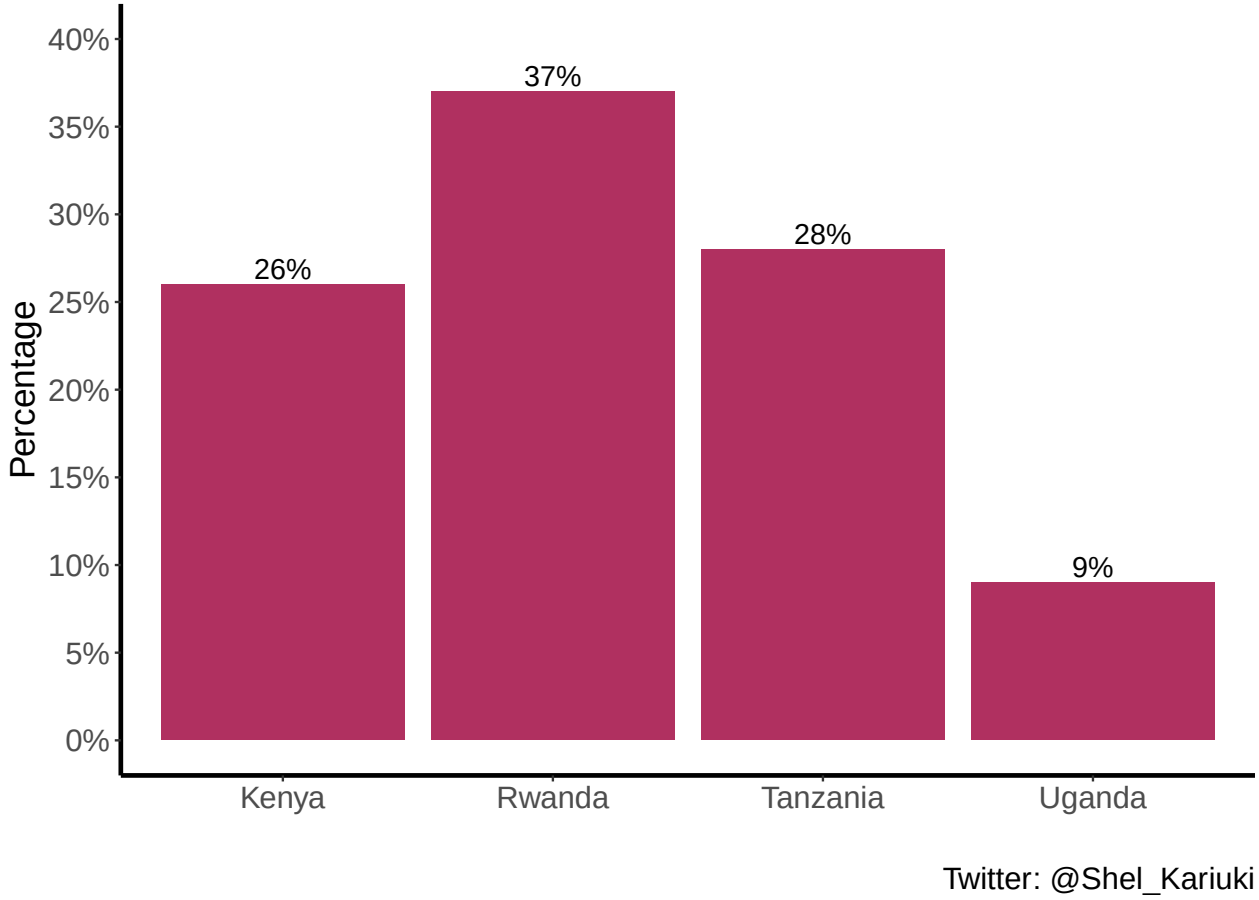
```
## table
tab10 <- df %>%
  group_by(location_type, country, gender_of_respondent) %>%
  count() %>%
  ungroup() %>%
  group_by(country,location_type) %>%
  mutate(perc = round(n/sum(n) *100,1))

##plot
plot10 <- ggplot(data = tab10, aes(x = gender_of_respondent, y = perc))+
          geom_bar(stat = "identity", fill = "maroon", width = 0.7)+
          geom_text(aes(label = perc), vjust = -0.25, hjust = 0.5)+
          mir_theme+
          theme(axis.text = element_text(size = 12))+
          labs(title = "Distribution of gender by country and location type",
               x = "Gender", y = "Percentage",
               caption = "Twitter: @Shel_Kariuki")+
          facet_grid(country ~ location_type)+
          ylim(c(0, 75))
plot10
```




```
## graph
plot2d <- ggplot(data = tab2, aes(x = country, y = perc ))+
          geom_bar(stat = "identity", fill = "maroon")+
          geom_text(aes(label = paste0(perc,"%")), hjust = 0.5, vjust = -0.25, size = 4.5)+
          theme(plot.title = element_text(size = 12, #family = "Source Sans Pro Semibold",
                                          face = "italic", hjust = 0.5),
                panel.background = element_rect(fill = NA),
                axis.line = element_line(size = 1, colour = "black"),
                axis.text = element_text(size = 14),
                axis.title = element_text(size = 16),
                plot.caption = element_text(size = 14))+
          ylim(c(0,40))+
```

```
          scale_y_continuous(limits = c(0, 40), breaks = seq(0, 40, by = 5),
           labels = function(x) paste0(x, "%"))+
          labs(title = "Distribution of sample by country",
               x = "", y = "Percentage",
               caption = "Twitter: @Shel_Kariuki")
plot2d
```

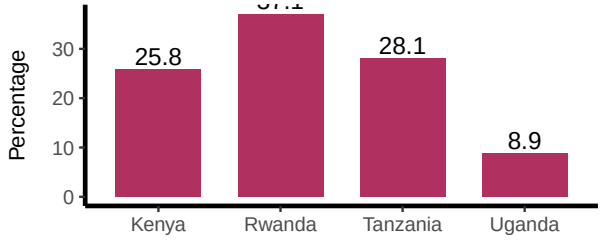
5.4 Adding % marks on the y axis and the text labels, and altering the breaks of the y axis




```
# p1 <-  plot2 / plot3
# p1
end <- (plot2 + plot3) / (plot5 + plot6)
end + plot_annotation(tag_levels = "I")
```

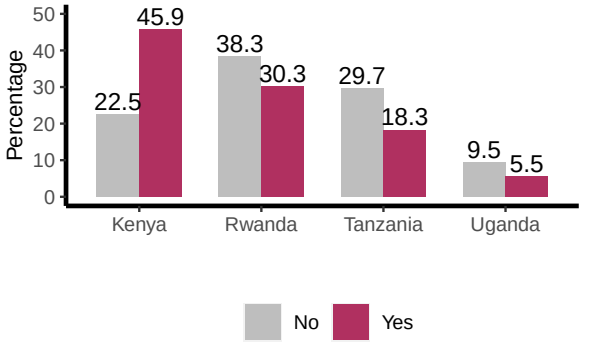
5.5 Using patch work: a package for combining multiple plots

I



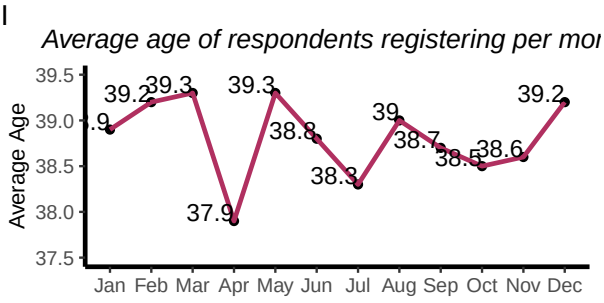
Twitter: @Shel_Kariuki

II



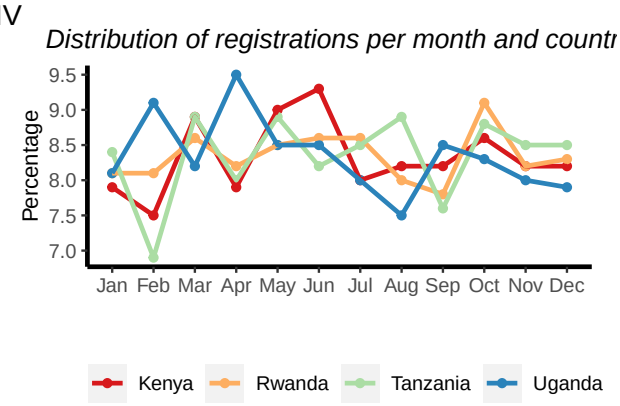
Twitter: @Shel_Kariuki

III



Twitter: @Shel_Kariuki

IV



Twitter: @Shel_Kariuki