Agricultural Survey of African Farm Households

Denis Mwaniki

2024-10-10

Contents

Importing Datasets and Preprocesing

Importing Libraries

Reading the data

```
### check working directory
getwd()
```

```
## [1] "D:/Njambanene/Njambanene/R/R/Case Study/FACS"
```

```
setwd("D:/Njambanene/Njambanene/R/R/Case Study/FACS")
### read dataset

dt0 <- readxl::read_xlsx("../FACS/Case_2_Statistics.xlsx")
### preview
glimpse(dt0)
```

```
## Rows: 4,763
## Columns: 14
## $ gender1  <dbl> 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1~
## $ gender2  <dbl> 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 2, 2, 2~
## $ gender3  <dbl> 1, 1, 2, 2, 1, 1, 1, 2, 2, 1, 2, 2, 1, 2, 1, 2, 2, 1, 1, 1, 2~
## $ gender4  <dbl> 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2, 2, 1, 2~
## $ gender5  <dbl> 2, 1, 2, 1, 1, 1, 1, 1, 2, 1, 2, 2, 2, 1, 2, 1, 1, 2, 1, 1, 1~
## $ gender6  <dbl> 1, 2, 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, 2, 2, 2, 1, 2~
## $ gender7  <dbl> 1, 1, 2, 1, 1, 2, 1, 1, 2, 1, 2, 1, NA, NA, 2, 1, NA, 2, 2, 2~
## $ gender8  <dbl> 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 1, 1, NA, NA, 1, 2, NA, 2, 1, 1~
## $ age1     <dbl> 57, 61, 47, 51, 56, 59, 50, 33, 38, 65, 55, 45, 51, 40, 59, 4~
## $ educ1    <dbl> 7, 6, 3, 1, 0, 5, 0, 0, 6, 0, 0, 0, 0, 20, 0, 0, 1, 5, 0, 0, ~
## $ married1 <dbl> 1, 1, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
## $ country  <chr> "burkinafaso", "burkinafaso", "burkinafaso", "burkinafaso", "~
## $ incfarm  <dbl> 375000, 1450000, 750000, 125500, 1240000, 1150000, 856000, 32~
## $ sickdays <dbl> 5, 12, 21, 11, 5, 17, 30, 20, 1, 7, 2, 0, 1, 3, 5, 7, 0, 2, 4~
```

```
head(dt0)
```

```
## # A tibble: 6 x 14
##   gender1 gender2 gender3 gender4 gender5 gender6 gender7 gender8  age1 educ1
##     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1       1       2       1       1       2       1       1       1    57     7
## 2       1       2       1       1       1       2       1       1    61     6
## 3       2       1       2       2       2       1       2       2    47     3
## 4       2       2       2       1       1       1       1       1    51     1
## 5       1       2       1       1       1       2       1       1    56     0
## 6       1       2       1       1       1       1       2       2    59     5
## # i 4 more variables: married1 <dbl>, country <chr>, incfarm <dbl>,
## #   sickdays <dbl>
```

Cleaning

```
#### mapping gender labels

dt1 <-dt0 %>%
  mutate(across(starts_with("gender"),~ recode(as.character(.x)  ,'1' = "Male",'2' = "Female",.missing =

head(dt1)
```

```
## # A tibble: 6 x 14
##   gender1 gender2 gender3 gender4 gender5 gender6 gender7 gender8  age1 educ1
##   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <dbl> <dbl>
## 1 Male    Female  Male    Male    Female  Male    Male    Male       57     7
## 2 Male    Female  Male    Male    Male    Female  Male    Male       61     6
## 3 Female  Male    Female  Female  Female  Male    Female  Female     47     3
## 4 Female  Female  Female  Male    Male    Male    Male    Male       51     1
## 5 Male    Female  Male    Male    Male    Female  Male    Male       56     0
## 6 Male    Female  Male    Male    Male    Male    Female  Female     59     5
## # i 4 more variables: married1 <dbl>, country <chr>, incfarm <dbl>,
## #   sickdays <dbl>
```

```
#### mapping married labels

dt2 <- dt1 %>%
  mutate(married1 = recode(as.character(married1),'1' = "Married",'2' = "Never Married",'3' = "Previous

head(dt2)
```

```
## # A tibble: 6 x 14
##   gender1 gender2 gender3 gender4 gender5 gender6 gender7 gender8  age1 educ1
##   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <dbl> <dbl>
## 1 Male    Female  Male    Male    Female  Male    Male    Male       57     7
## 2 Male    Female  Male    Male    Male    Female  Male    Male       61     6
```

```
## 3 Female  Male    Female  Female  Female  Male    Female  Female     47     3
## 4 Female  Female  Female  Male    Male    Male    Male    Male       51     1
## 5 Male    Female  Male    Male    Male    Female  Male    Male       56     0
## 6 Male    Female  Male    Male    Male    Male    Female  Female     59     5
## # i 4 more variables: married1 <chr>, country <chr>, incfarm <dbl>,
## #   sickdays <dbl>
```

Exploratory Data Analysis

```
#### get size of hh
##### men & women in hh

dt3 <- dt2 %>%
  rowwise() %>%
  mutate(
    num_men = sum(c_across(starts_with("gender")) =="Male",na.rm = T),
    num_women = sum(c_across(starts_with("gender")) =="Female",na.rm = T),
    hh_size = num_men + num_women
  ) %>%
ungroup()

head(dt3)
```

```
## # A tibble: 6 x 17
##   gender1 gender2 gender3 gender4 gender5 gender6 gender7 gender8  age1 educ1
##   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <dbl> <dbl>
## 1 Male    Female  Male    Male    Female  Male    Male    Male       57     7
## 2 Male    Female  Male    Male    Male    Female  Male    Male       61     6
## 3 Female  Male    Female  Female  Female  Male    Female  Female     47     3
## 4 Female  Female  Female  Male    Male    Male    Male    Male       51     1
## 5 Male    Female  Male    Male    Male    Female  Male    Male       56     0
## 6 Male    Female  Male    Male    Male    Male    Female  Female     59     5
## # i 7 more variables: married1 <chr>, country <chr>, incfarm <dbl>,
## #   sickdays <dbl>, num_men <int>, num_women <int>, hh_size <int>
```

```
#### print female in hh

dt4 <- dt3 %>%
  rowwise() %>%
  mutate(
    female_hh = ifelse(num_women>=5,1,0)) %>%
  ungroup()

head(dt4)
```

```
## # A tibble: 6 x 18
##   gender1 gender2 gender3 gender4 gender5 gender6 gender7 gender8  age1 educ1
##   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <dbl> <dbl>
## 1 Male    Female  Male    Male    Female  Male    Male    Male       57     7
## 2 Male    Female  Male    Male    Male    Female  Male    Male       61     6
## 3 Female  Male    Female  Female  Female  Male    Female  Female     47     3
```

```
## 4 Female  Female  Female  Male    Male    Male    Male    Male       51     1
## 5 Male    Female  Male    Male    Male    Female  Male    Male       56     0
## 6 Male    Female  Male    Male    Male    Male    Female  Female     59     5
## # i 8 more variables: married1 <chr>, country <chr>, incfarm <dbl>,
## #   sickdays <dbl>, num_men <int>, num_women <int>, hh_size <int>,
## #   female_hh <dbl>
```

Validation Checks

```
### check NAs
#### total NAs in data
sum(is.na(dt4))
```

```
## [1] 752
```

```
#### atleast a row with NA
sum(rowSums(is.na(dt4))>0)
```

```
## [1] 700
```

```
#### count Nas per col
colSums(is.na(dt4))
```

```
##   gender1   gender2   gender3   gender4   gender5   gender6   gender7   gender8
##         0         0         0         0         0         0         0         0
##      age1     educ1  married1   country   incfarm  sickdays   num_men num_women
##         9       128         6         0        55       554         0         0
##   hh_size female_hh
##         0         0
```

```
#### replace NA in a col w 0
dt4$incfarm[is.na(dt4$incfarm)] <- 0
```

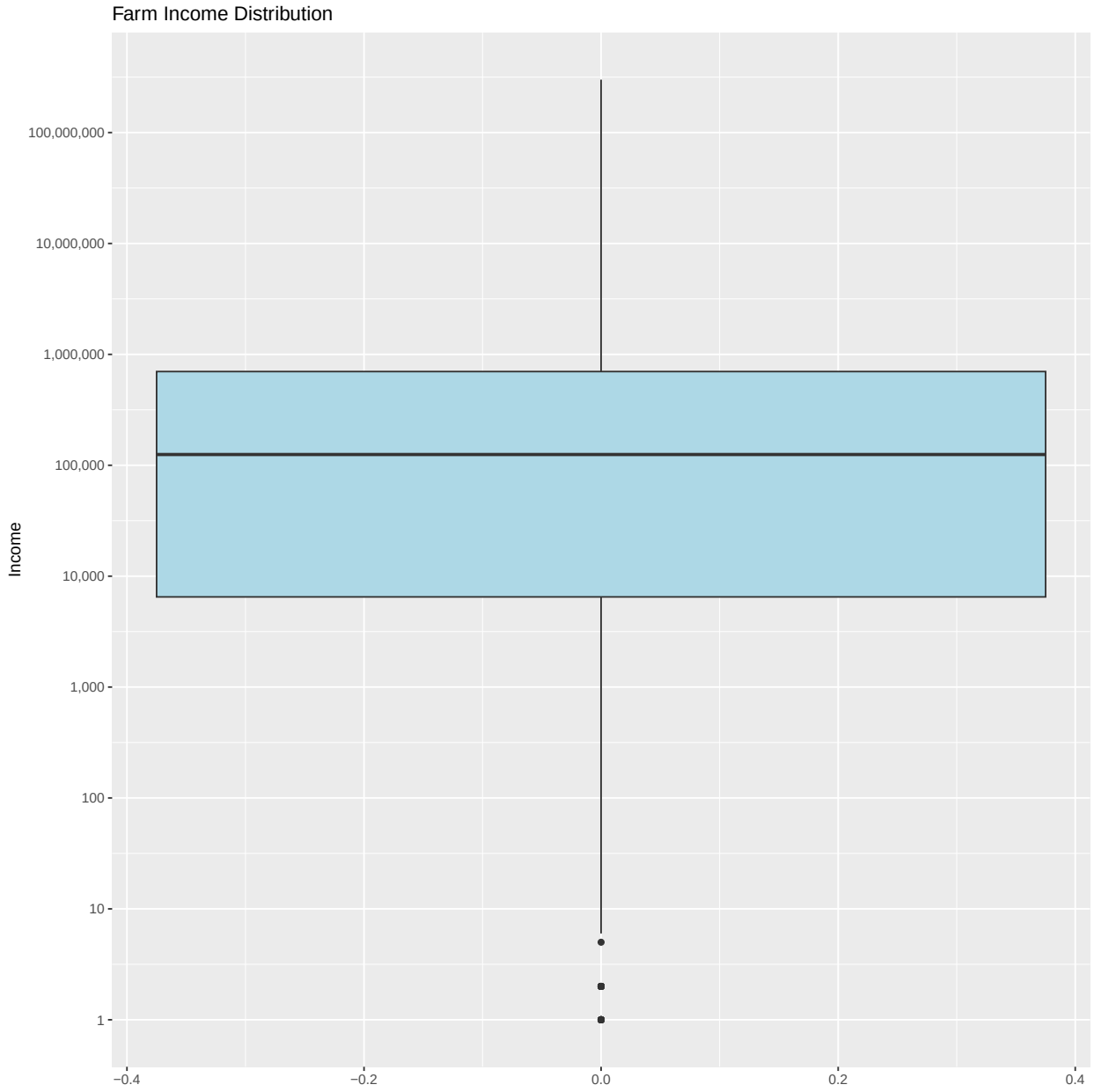
```
### check for outliers

inc_summary <- dt4 %>%
  filter(incfarm > 0 ) %>%
  summarise(
    mean_income = mean(incfarm),
    min_income = min(incfarm),
    q1 = quantile(incfarm,0.25),
    median_income = median(incfarm),
    q3 = quantile(incfarm,0.75),
    max_income = max(incfarm)
  )

print(inc_summary)
```

```
## # A tibble: 1 x 6
##   mean_income min_income    q1 median_income     q3 max_income
##         <dbl>      <dbl> <dbl>         <dbl>  <dbl>      <dbl>
## 1    1275461.          1  6500        125000 700000  300000000
```

Analysis

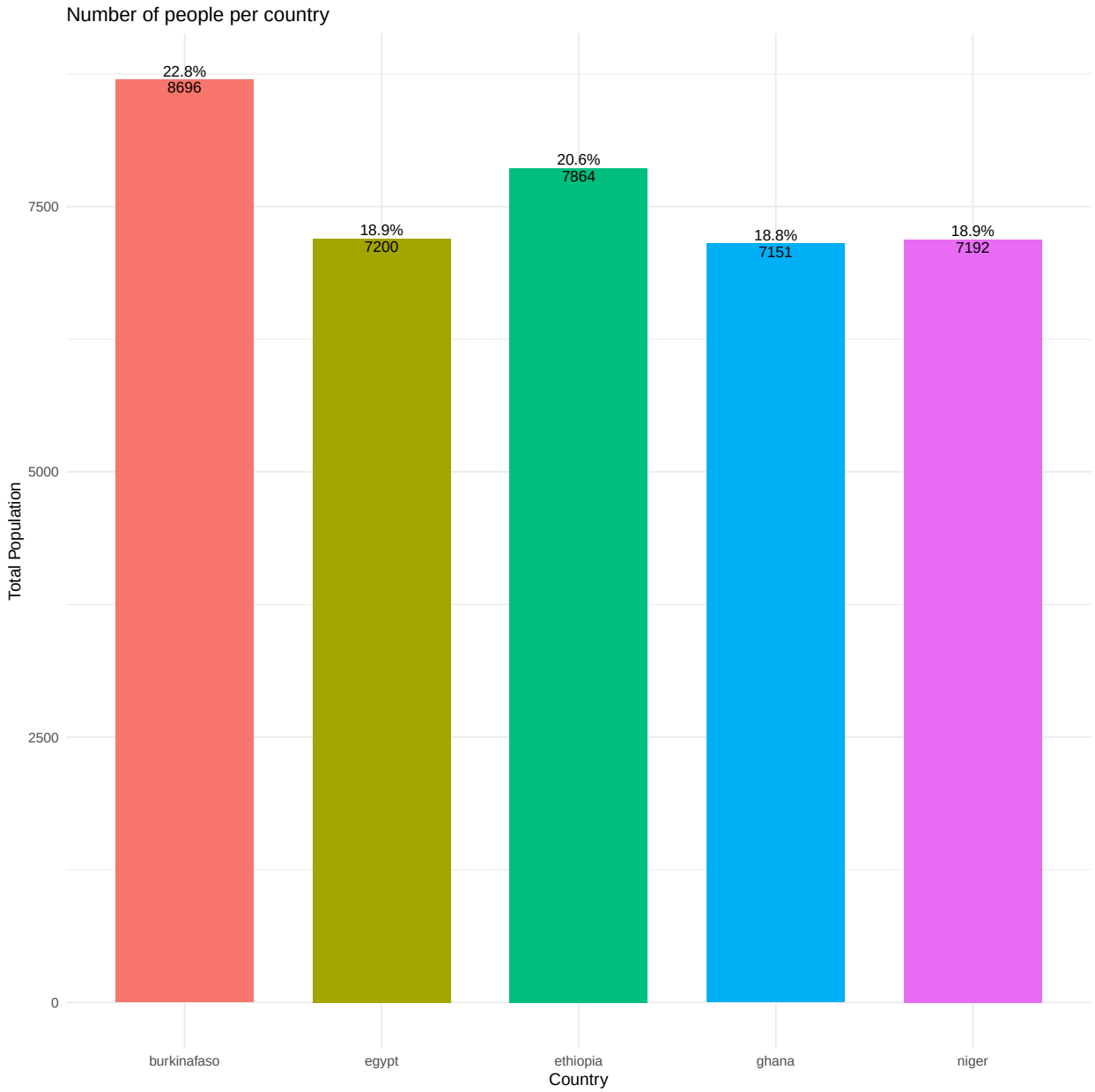
Overall Breakouts

```
#### Plot a boxplot of incfarm

ggplot(dt4, aes(y = incfarm)) +
  geom_boxplot(fill = "lightblue")+
  scale_y_log10(
    labels = scales::comma_format(),
    breaks = scales::log_breaks(n = 10)
  )+
  labs(title = "Farm Income Distribution", y= "Income")
```



```r
#### number of people per country

country_pop <- dt4 %>%
  select(hh_size,country) %>%
  group_by(country) %>%
  summarise(total_pop = sum(hh_size)) %>%
  mutate(percentage = total_pop/sum(total_pop)*100)
```

```r
##### plot 1

ggplot(country_pop,aes(x = country,y= total_pop, fill = country)) +
  geom_bar(stat = "identity",width = 0.7) +
  geom_text(aes(label = paste0(round(percentage,1),"%")),
            vjust = -0.3,size = 3.5) +
  geom_text(aes(label = total_pop),
            vjust = 1.2, size = 3.5) +
  labs(title = "Number of people per country",x = "Country",y= "Total Population") +
  theme_minimal()+
  theme(legend.position = "none")
```



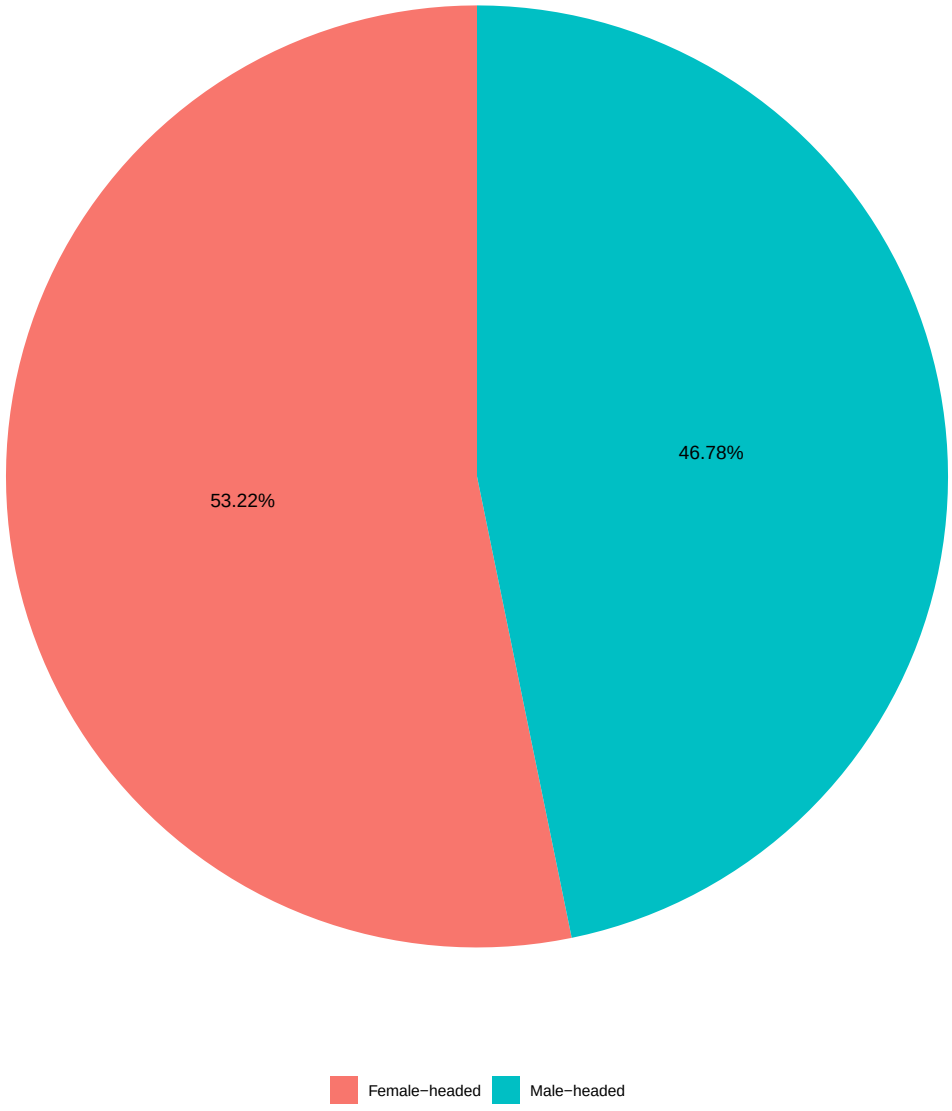

```
#### prop of female hh

proportion_of_fhh <- mean(dt4$female_hh) * 100

print (proportion_of_fhh)
```

```
## [1] 53.22276
```

```
#### gender distr
gender_dt <- data.frame(
  category = c("Female-headed","Male-headed"),
  counts = c(sum(dt4$female_hh),nrow(dt4) - sum(dt4$female_hh))
)
```

```
##### plot 2
ggplot(gender_dt,aes(x = "",y = counts, fill = category)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  geom_text(aes(label = paste0(round(counts/sum(counts) * 100,2),"%")),
            position = position_stack(vjust = 0.5))+
  labs(title = paste0("Proportion of female households: ",
                      round(proportion_of_fhh,2), "%")) +
  theme_void()+
  theme(legend.title = element_blank(),
        legend.position = "bottom")
```

Proportion of female households: 53.22%

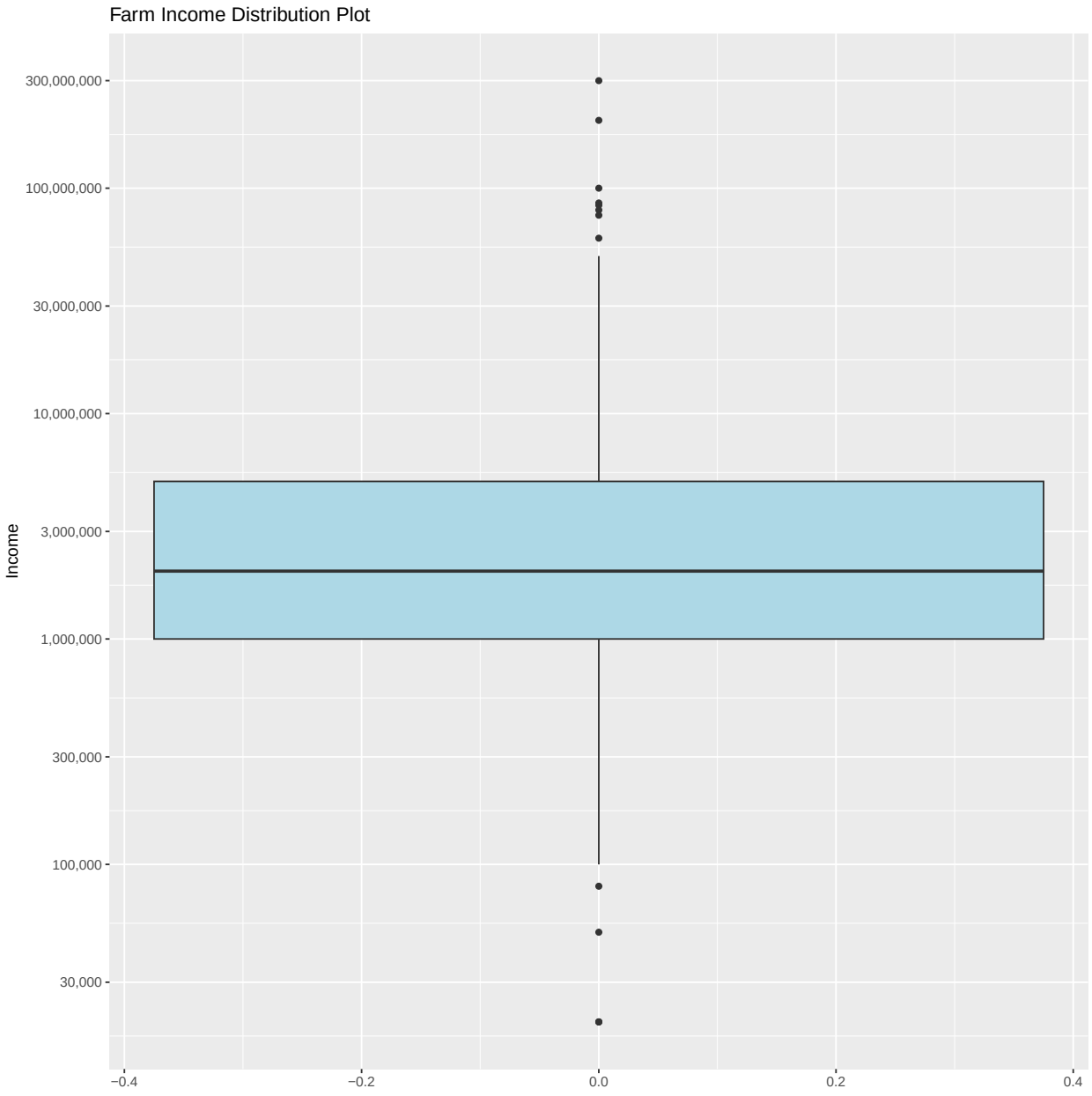


Distribution of farm income in Ghana

```
### filter for ghana
gh_f_inc <- dt4 %>%
  filter(country == "ghana")

### print summary stats
gh_f_inc_summary <- gh_f_inc %>%
  select(country,incfarm) %>%
  summarise(
    mean_income = mean(incfarm),
    median_income = median(incfarm),
    sd_income = sd(incfarm),
    min_income = min(incfarm),
    max_income = max(incfarm)
  )

print(gh_f_inc_summary)
```

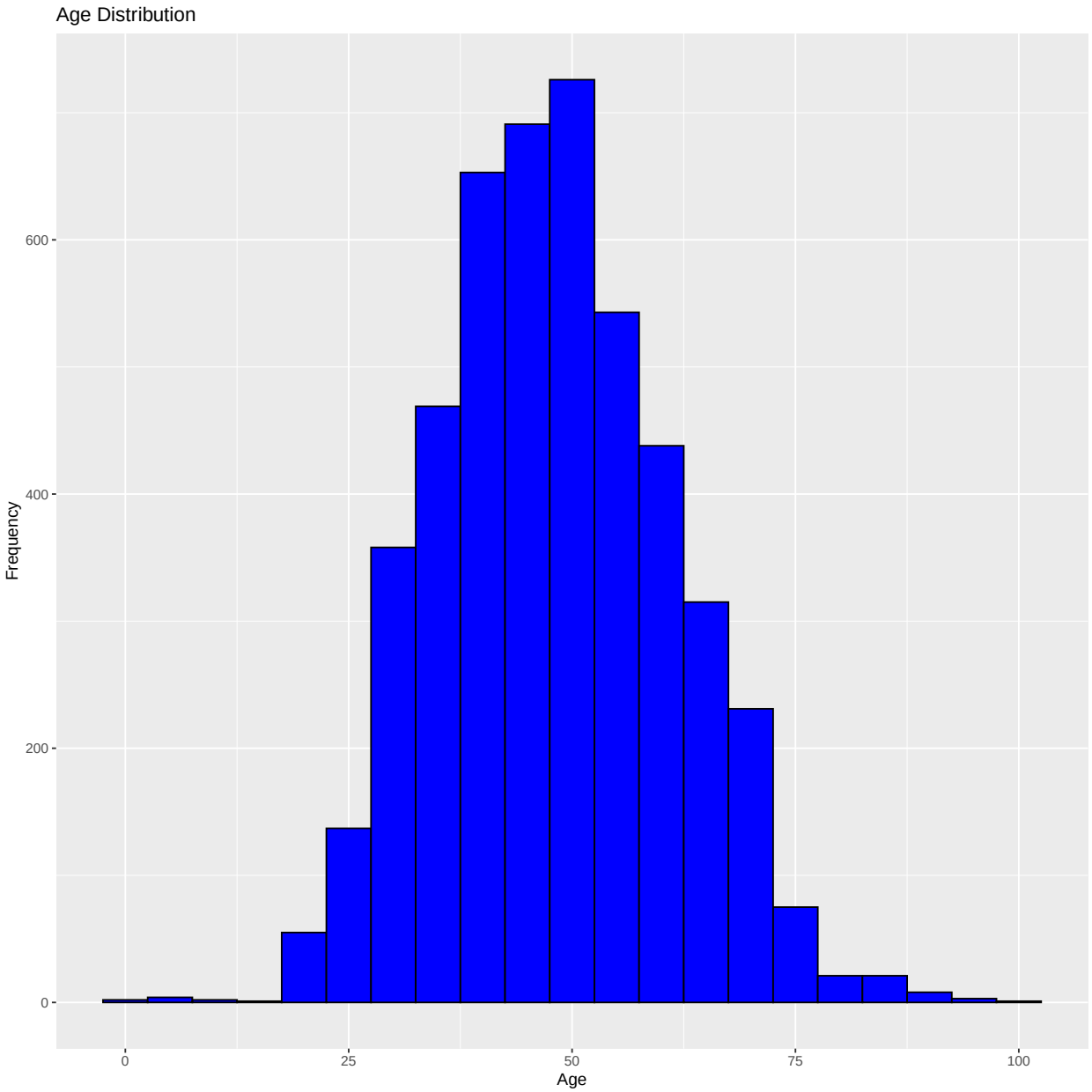
```
## # A tibble: 1 x 5
##   mean_income median_income sd_income min_income max_income
##         <dbl>         <dbl>     <dbl>      <dbl>      <dbl>
## 1    5015704.       2000000 14796194.          0  300000000
```

```
##### plot 3

ggplot(gh_f_inc,aes(y = incfarm))+
  geom_boxplot(fill = "lightblue") +
  scale_y_log10(
    labels = scales::comma_format(),
    breaks = scales::log_breaks(n = 10)
  )+
  labs(title = "Farm Income Distribution Plot",y = "Income")
```



Age Distribution

```
##### plot 4
ggplot(dt4,aes(x = age1)) +
  geom_histogram(binwidth = 5,fill = "blue", color = "black") +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")
```




```
#### gender descriptive stats
gender_analysis <-   gh_f_inc %>%
  group_by(female_hh) %>%
  summarise(
    mean_income = mean(incfarm),
    sd_income = sd(incfarm)
  )

#### gender
t.test(incfarm ~female_hh, data = gh_f_inc)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  incfarm by female_hh
## t = 0.50289, df = 866.65, p-value = 0.6152
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1344390  2270646
## sample estimates:
## mean in group 0 mean in group 1
##         5314095         4850967
```

Overall Conclusion:

- No Significant Difference: The t-test results indicate that there is no statistically significant difference in mean income between group 0 and group 1. The difference observed in the sample means could likely be due to random chance.
- Fail to Reject the Null Hypothesis: The high p-value and the confidence interval including 0 mean that you do not have enough evidence to reject the null hypothesis of no difference in means between the two groups.
- This analysis suggests that, based on the data provided, the incomes in the two groups are not significantly different.

Education Distribution

```
#### education
aov_result <- aov(incfarm ~educ1, data = gh_f_inc)
summary(aov_result)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## educ1         1 1.259e+15 1.259e+15   5.585 0.0183 *
## Residuals   850 1.917e+17 2.255e+14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 42 observations deleted due to missingness
```

Overall Interpretation:

- Statistical Significance: The factor educ1 has a statistically significant effect on the response variable (e.g., income) at the 5% significance level. This means that the difference in the response variable across the levels of educ1 is unlikely to be due to random chance.
- Magnitude of Effect: Although the result is statistically significant, the F value is relatively modest (5.585), suggesting that while educ1 has an effect, it might not explain a large proportion of the variance in the response variable.
- In summary, the ANOVA results indicate that the variable educ1 significantly affects the response variable, suggesting that differences in educ1 levels are associated with differences in the response variable (e.g., income).

Age Distribution

```
#### age
aov_result1 <- aov(incfarm~age1, data = gh_f_inc)
summary(aov_result1)
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## age1          1 2.905e+15 2.905e+15   13.45 0.000259 ***
## Residuals   892 1.926e+17 2.159e+14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overall Interpretation:

- Statistical Significance: The factor age1 has a statistically significant effect on the response variable (e.g., income) at the 5% significance level. This means that the difference in the response variable across the levels of age is unlikely to be due to random chance.
- Magnitude of Effect: Although the result is statistically significant, the F value is fairly high (13.45), suggesting that age1 has an effect, and it may explain a large proportion of the variance in the response variable.
- In summary, the ANOVA results indicate that the variable age1 significantly affects the response variable, suggesting that differences in age1 levels are associated with differences in the response variable (e.g., income).

Sickdays Distribution

```
#### sickdays
aov_result2 <- aov(incfarm~sickdays , data = gh_f_inc)
summary(aov_result2)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## sickdays      1 7.715e+13 7.715e+13   0.309  0.578
## Residuals   684 1.705e+17 2.493e+14
## 208 observations deleted due to missingness
```

Overall Interpretation:

- No Significant Difference: The factor sickdays has no statistically significant effect on the response variable (e.g., income) at the 5% significance level. This means that the difference in the response variable across the levels of sickdays could likely be due to random chance.
- Magnitude of Effect: While the result is not statistically significant, the F value is almost 0 (0.309), suggesting that sickdays has minimal effect, and it cannot explain the variance in the response variable.
- In summary, the ANOVA results indicate that the variable sickdays does not significantly affect the response variable, suggesting that differences in sickdays levels are not associated with differences in the response variable (e.g., income).

```
#### multiple linear regression
lm_result <- lm(incfarm ~ female_hh + educ1 + age1 + sickdays ,data = gh_f_inc)
summary(lm_result)
```

```
##
## Call:
## lm(formula = incfarm ~ female_hh + educ1 + age1 + sickdays, data = gh_f_inc)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -14403613  -4652562  -2290939    685990 288774769
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6089642    2816008  -2.163   0.0309 *
## female_hh     984496    1295918   0.760   0.4477
## educ1         278616     113357   2.458   0.0142 *
## age1          196619      48940   4.018 6.56e-05 ***
## sickdays      -54778      37016  -1.480   0.1394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15770000 on 657 degrees of freedom
##   (232 observations deleted due to missingness)
## Multiple R-squared:  0.02936,    Adjusted R-squared:  0.02345
## F-statistic: 4.969 on 4 and 657 DF,  p-value: 0.0005959
```