Measurement Final

Jinane AMAL

I. Some useful infos :

- The Countries are identified in 2 different waves "country" and "country_mod"

As I will focus on studying Spain on this easy share version of the data , and will use 2 other countries to compare them to Spain especially to answer question 2 :

It will be of great use to highlight what's the "country" and "country_mod" identifiers for these 3 countries for further code lines .

"country", "country_mod", "language" for Spain are in this order : 15 - 724 - 15 "country", "country_mod" , "language" for Austria are in this order : 11 - 40 - 11 "country", "country_mod" , "language" for Bulgaria are in this order : 51 - 100 - 51

II. Loading the libraries

```
library(here)
library(dplyr)
library(vroom)
library(tidyr)
library(ggplot2)
library(knitr)
library(rstatix)
library(ineq)
theme_set(theme_bw())
here::i_am("measurements.Rproj")
```

III Loading the data & Data cleaning process :

Filter to keep only wave 7 to keep a base for the questions in which we will use for other countries

```
wave7data <-load("easySHARE_rel8_0_0.rda")
```

```
# Filter data to include only wave 7
wave7data <- easySHARE_rel8_0_0|>
  filter(wave == 7)
# Now 'wave7data' contains only the data from wave 7
```

Erase column 7

```
wave7data <- wave7data |>
  filter(wave == 7) |>
  select(-wave)
```

Drop wavepart column = 7:

Since it is the same for everyone.

```
wave7data <- wave7data |>
  select(-wavepart)
```

Drop langage column :

Since we've noticed that for the chosen countries the "country" code is the same as the "langage" code

```
wave7data <- wave7data |>
  select(-language)
```

Drop int_year as all the individuals have been interviewed the same year 2017

```
wave7data <- wave7data |>
  select(-int_year)
```

Drop country mod

```
wave7data <- wave7data |>
  select(-country_mod)
```

1| Spain statistics' description :

We now have 4704 observations for 102 variables after dropping certain repetitive variables :

```
wave7dataSP <- wave7data |>
  filter(country == 15)
```

2| First descriptive statistic variable is the Household count :

When we have the same 6 digits in the middle, we know that we're observing a same household. For this reason I generated this code to count how many people are part of a same household so we don't have doubles to get more reliable results.

```
household_countSP <- wave7dataSP|>
  mutate(household_id = substr(mergeid, 1, 10))|>
  group_by(household_id) |>
  summarise(count = n())
```

3| Second descriptive statistic variable is the number respondents move out of the sample :

Identifying the number of respondents who may have moved out across waves identified by "=="B"" at last char with the help of the variable hhid 's column, thanks to the following code which counts how many of them are present in wave 7

```
count_B <- sum(substr(wave7dataSP$hhid, nchar(wave7dataSP$hhid), nchar(wave7dataSP$hhid))
```

4| Focus on the couples of the sample

I decided to count how many in the coupleid column have the same id to clearly see how many couples there are . Normally for each couple id count i should get a 2 Surprisingly after further observation of this code's output, I get sometimes a one instead of 2 in the coupleid column because the partner's supposed same id doesn't appear. We can assume that these individuals are indeed a couple but that their partner chose not to respond to the wave 7 I still decided to count these individuals whose partners don't appear in the couple id column, we thus have 1834 couples

```
nbocouplesw7sp <- wave7dataSP|>
  group_by(substr(coupleid, 1, 15))|>
  summarise(count = n())
```

The population of interest: wave 7; country focus spain topics of interest:1) Demographics 2) Household composition 3) Social support & network 4) Childhood conditions 5) Health and behavior 6) Functional limitation indices 7) Work & money units –>indiv housecold and marital/couple sample size –> tot of above

In compliance with the easyshare guide 8 descriptive

Descriptive Statistics:

Focus on the age of the sample:

```
wave7datainf30sp <- wave7dataSP |>
  filter(age < 30)
```

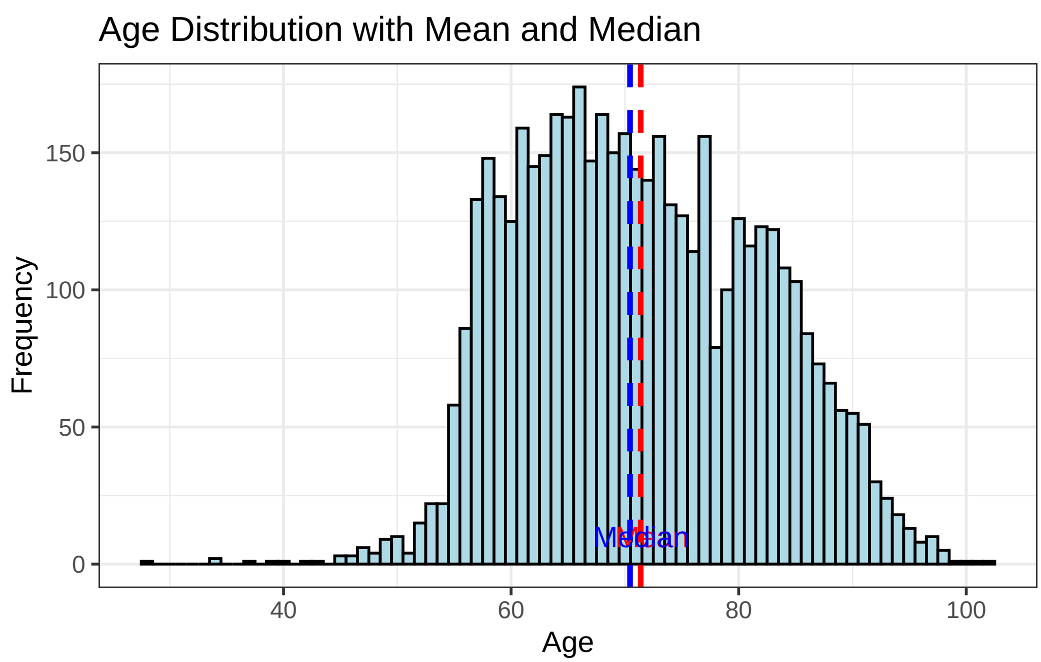
```
wave7datainf44sp <- wave7dataSP|>
  filter(age < 44)
```

Based on these filters , we notice that most of the individuals interviewed are seniors which are overly represented in the sample which we will prove in the following age graph . This is an issue because they are not representative of the whole population in reality and normally we should be studying a more spain representative sample knowing that the median age in Spain is 44 and out of the 4708 observations only 8 are either 44 years or younger .

Age distribution for Spain (Actual median age in Spain = 44 years old)

```
# Calculate mean and median
mean_age <- mean(wave7dataSP$age)
median_age <- median(wave7dataSP$age)

# Create a histogram of age distribution with mean and median markers
ggplot(wave7dataSP, aes(x = age)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  geom_vline(xintercept = mean_age, color = "red", linetype = "dashed", size = 1) +
  geom_vline(xintercept = median_age, color = "blue", linetype = "dashed", size = 1) +
  labs(title = "Age Distribution with Mean and Median",
       x = "Age",
       y = "Frequency") +
  annotate("text", x = mean_age + 1, y = 10, label = "Mean", color = "red", size = 4) +
  annotate("text", x = median_age + 1, y = 10, label = "Median", color = "blue", size = 4)
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Which gender is slighly predominant: Females respondents

```
proportion_femalesp <- mean(wave7dataSP$female)

print(paste("Proportion of females in spain:", proportion_femalesp))
```

```
[1] "Proportion of females in spain: 0.559736394557823"
```

Focus on the income of the sample this time :

```
wave7dataSP <- wave7dataSP|>
  filter(thinc_m != -10)
```

Statistics Table linking age and income:

```
age_summary <- summary(wave7dataSP$age)
income_summary <- summary(wave7dataSP$thinc_m)

summary_table <- data.frame(
  Variable = c("Age", "Income"),
  Min = c(min(age_summary), min(income_summary)),
  Max = c(max(age_summary), max(income_summary)),
  Mean = c(mean(age_summary), mean(income_summary)),
  Median = c(median(age_summary), median(income_summary)),
  SD = c(sd(age_summary), sd(income_summary))
)

print(summary_table)
```

```
  Variable  Min      Max        Mean      Median          SD
1      Age 46.6     98.0    73.74661    74.38984     16.91341
2   Income  0.0 516154.9 97300.39889 17107.71463 205353.07312
```

Focus on employment:

As we can expect, the vast majority of the sample is already retired as seen through the code identifier (1) in the table below which has a proportion of 53.6% followed by home makers (5) with 27.8% ratio then the employed or self employed with 6.9% : We should keep in mind that there is a difference between these categories . Indeed , by definition:A homemaker is someone, typically a spouse or parent, who manages household tasks and cares for the family without being formally employed outside the home. Self-employed are individuals who work for themselves rather than for an employer. They manage their own business or freelance work, often providing services or selling products to clients . They are responsible for their own income, taxes, and business operations. Unemployed individuals are those who are willing and able to work but are currently without a job. They may be actively seeking employment opportunities but have not yet found suitable work.

Note also that we need to get rid of -15 and -12 and 97 codes which identify missing infos, people who didn't or weren't fond of responding to these questions in this partiular wave at least .

```
category_proportions <- wave7dataSP |>
  group_by(ep005_) |>
  summarise(count = n()) |>
  mutate(proportion = count / sum(count))
print(category_proportions)
```

```
# A tibble: 8 x 3
  ep005_ count proportion
   <int> <int>      <dbl>
1    -15    24    0.0188
2    -12     2    0.00156
3      1   686    0.536
4      2    88    0.0688
5      3    19    0.0148
6      4    60    0.0469
7      5   356    0.278
8     97    45    0.0352
```

Focus on the number of years spent on education "eduyears" variable here:

```
edu7dataSP <- wave7dataSP %>%
  filter(eduyears_mod >= 0)

mean_eduyearsSP <- mean(edu7dataSP$eduyears_mod)
print(paste("Mean of eduyears_mod after filtering:", mean_eduyearsSP))
```

```
[1] "Mean of eduyears_mod after filtering: 7.25204731574158"
```

Focus on physical strenght "maxgrip" :

```
maxgrip7dataSP <- wave7dataSP |>
  filter(maxgrip >= 0)

mean_max_grip <- mean(maxgrip7dataSP$maxgrip)

print(paste("Mean of max_grip after filtering:", mean_max_grip))
```

```
[1] "Mean of max_grip after filtering: 26.2420494699647"
```

First Country comparison Focus on Austria :

```
wave7dataAT <- wave7data |>
  filter(country == 11)
```

The income of the sample for Austria: (AT)

```
wave7dataAT <- wave7dataAT|>
  filter(thinc_m != -10)
```

```
income_summaryAT <- summary(wave7dataAT$thinc_m)
print(income_summaryAT)
```

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
   733.5  16691.5  23550.3  28882.6  34228.4 447464.9 
```

Second Country comparison Focus on Denmark :

```
wave7dataDK <- wave7data |>
  filter(country == 18)
```

The income of the sample for Denmark: (DK)

```
wave7dataDK <- wave7dataDK|>
  filter(thinc_m != -10)
```

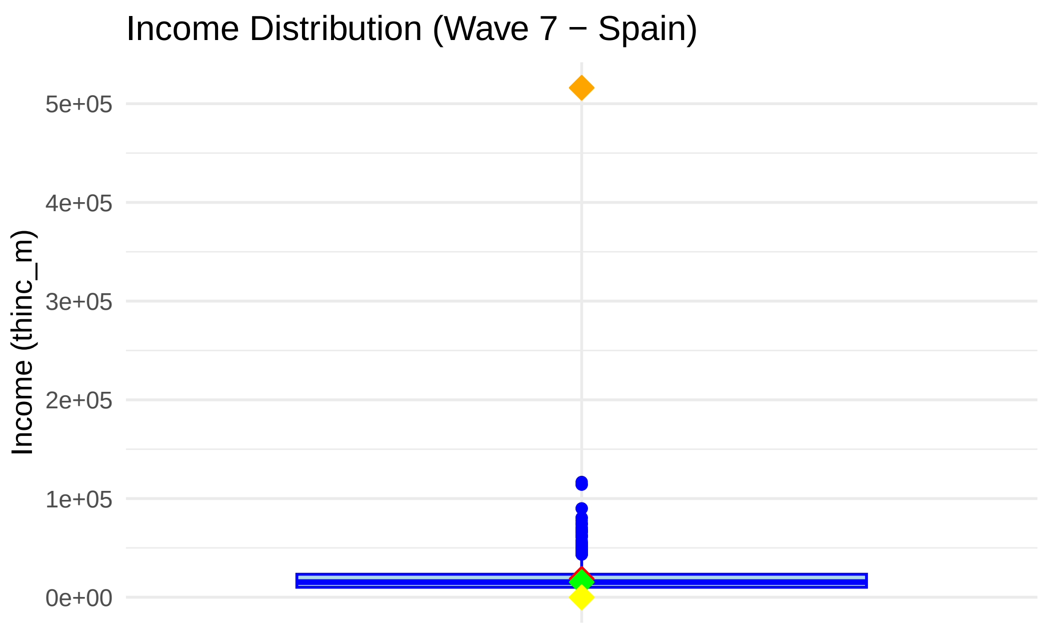
```
income_summaryDK <- summary(wave7dataDK$thinc_m)
print(income_summaryDK)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   16455   25563   29914   41147  117995
```

Going back to our reference country Spain:

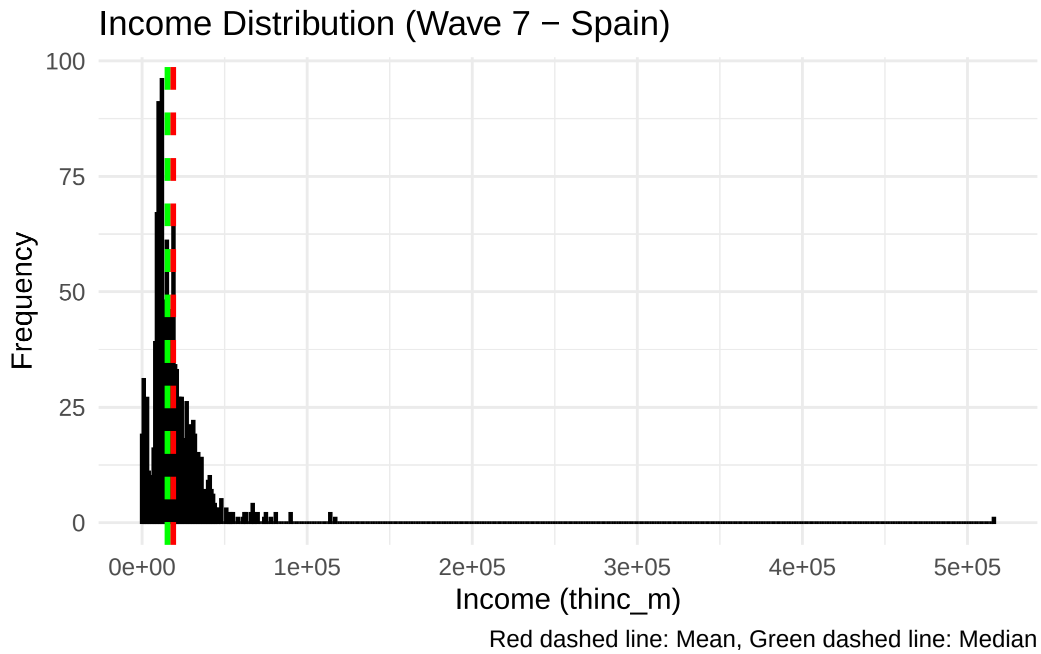
Graphical representation: average houshold income for spain Option 1

```
ggplot(wave7dataSP, aes(x = "", y = thinc_m)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "red", color = "re
  stat_summary(fun = median, geom = "point", shape = 23, size = 3, fill = "green", color =
  stat_summary(fun = min, geom = "point", shape = 23, size = 3, fill = "yellow", color = "
  stat_summary(fun = max, geom = "point", shape = 23, size = 3, fill = "orange", color = "
  labs(title = "Income Distribution (Wave 7 - Spain)",
       x = NULL,
       y = "Income (thinc_m)",
       fill = "Statistic") +
  theme_minimal()
```



Graphical representation: average houshold income for spain Option 2

```
ggplot(wave7dataSP, aes(x = thinc_m)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(thinc_m)), color = "red", linetype = "dashed", size = 1
  geom_vline(aes(xintercept = median(thinc_m)), color = "green", linetype = "dashed", size
  labs(title = "Income Distribution (Wave 7 - Spain)",
       x = "Income (thinc_m)",
       y = "Frequency",
       caption = "Red dashed line: Mean, Green dashed line: Median") +
  theme_minimal()
```



Getting rid of extreme values wave7sp keeping the 0

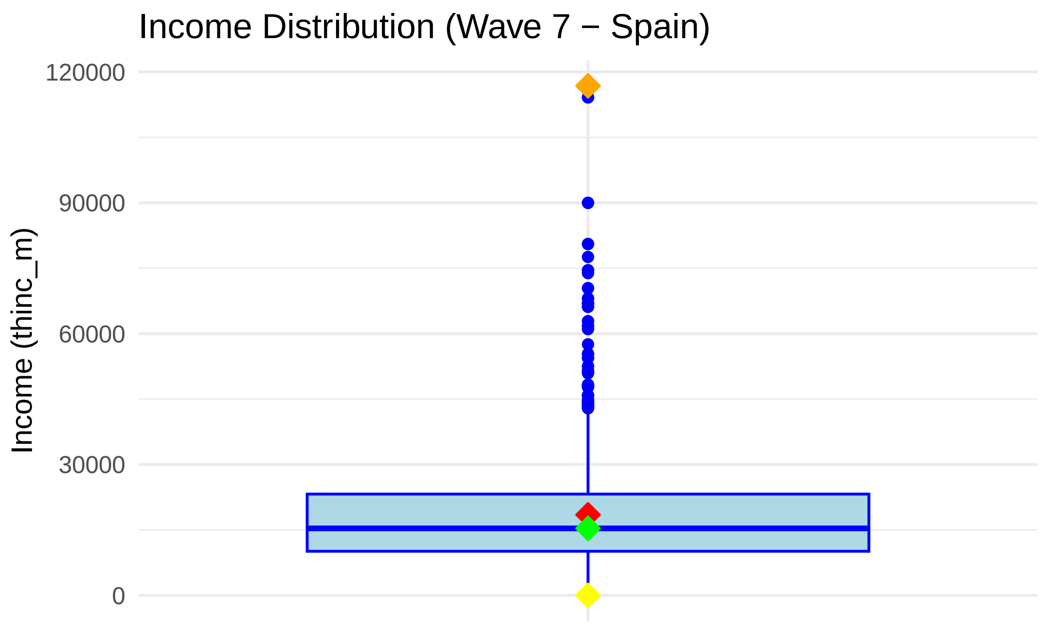
```
filtered_wave7dataSP <- subset(wave7dataSP, thinc_m >= 0 & thinc_m <= 120000)

income_summarySP <- summary(filtered_wave7dataSP$thinc_m)

print(income_summarySP)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   10105   15352   18475   23216  116804
```

Graphical representation: average houshold income for spain after getting rid of outliers option 1

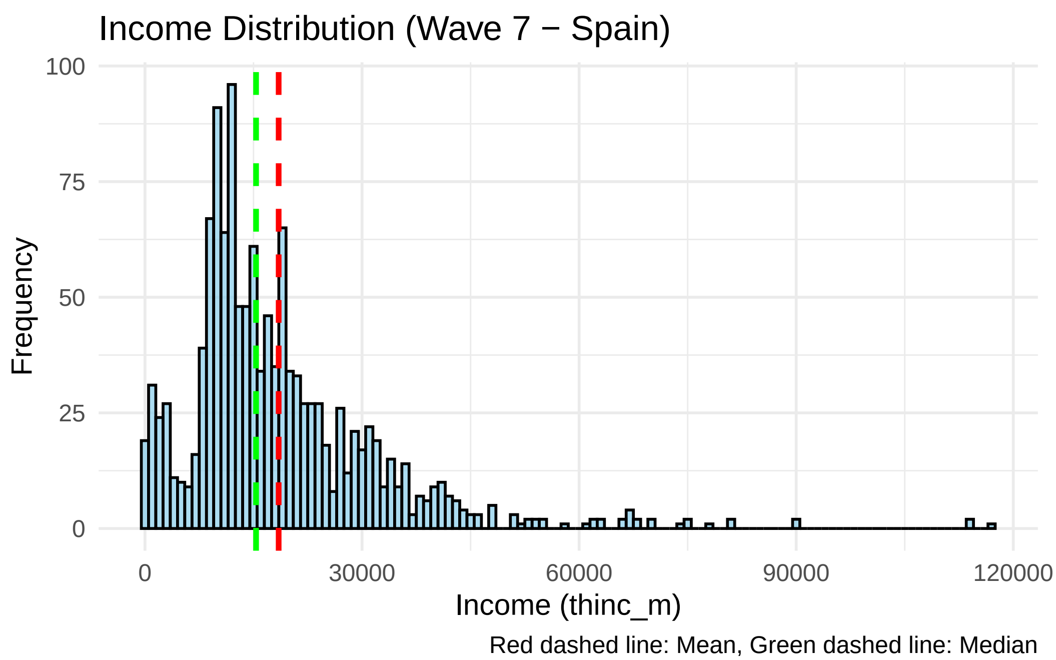
```
ggplot(filtered_wave7dataSP, aes(x = "", y = thinc_m)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "red", color = "re
  stat_summary(fun = median, geom = "point", shape = 23, size = 3, fill = "green", color =
  stat_summary(fun = min, geom = "point", shape = 23, size = 3, fill = "yellow", color = "
```

```r
  stat_summary(fun = max, geom = "point", shape = 23, size = 3, fill = "orange", color = "
  labs(title = "Income Distribution (Wave 7 - Spain)",
       x = NULL,
       y = "Income (thinc_m)",
       fill = "Statistic") +
  theme_minimal()
```



Graphical representation: average houshold income for spain after getting rid of outliers option 2

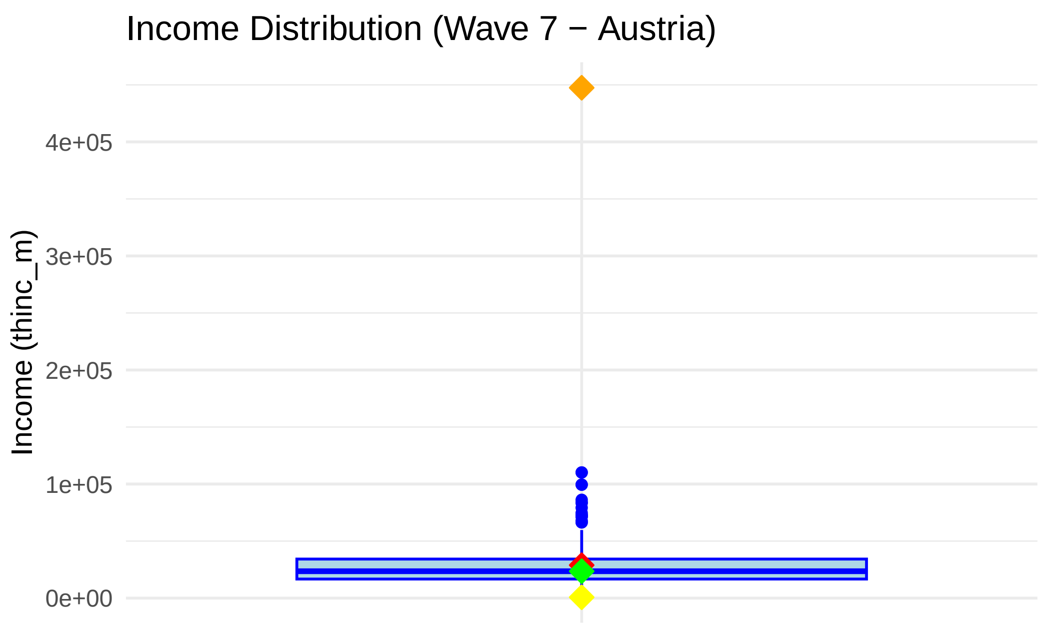
```r
ggplot(filtered_wave7dataSP, aes(x = thinc_m)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(thinc_m)), color = "red", linetype = "dashed", size = 1
  geom_vline(aes(xintercept = median(thinc_m)), color = "green", linetype = "dashed", size
  labs(title = "Income Distribution (Wave 7 - Spain)",
       x = "Income (thinc_m)",
       y = "Frequency",
       caption = "Red dashed line: Mean, Green dashed line: Median") +
  theme_minimal()
```



Our first comparison country Austria:

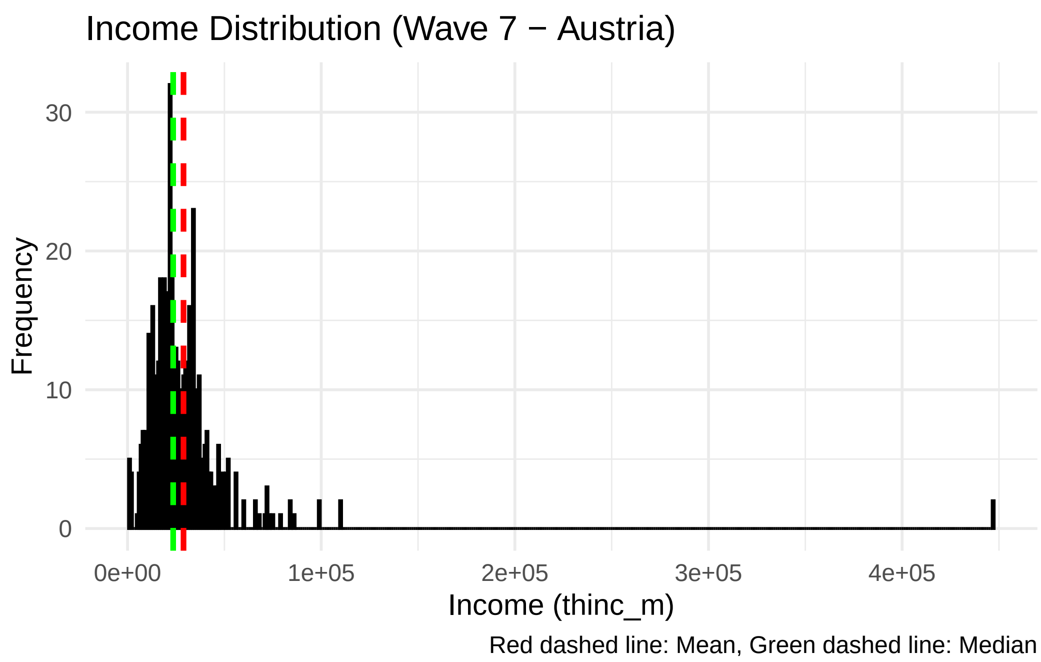
Graphical representation: average income for Austria option 1

```
ggplot(wave7dataAT, aes(x = "", y = thinc_m)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "red", color = "re
  stat_summary(fun = median, geom = "point", shape = 23, size = 3, fill = "green", color =
  stat_summary(fun = min, geom = "point", shape = 23, size = 3, fill = "yellow", color = "
  stat_summary(fun = max, geom = "point", shape = 23, size = 3, fill = "orange", color = "
  labs(title = "Income Distribution (Wave 7 - Austria)",
       x = NULL,
       y = "Income (thinc_m)",
       fill = "Statistic") +
  theme_minimal()
```



Graphical representation: average houshold income for Austria Option 2

```
ggplot(wave7dataAT, aes(x = thinc_m)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(thinc_m)), color = "red", linetype = "dashed", size = 1
  geom_vline(aes(xintercept = median(thinc_m)), color = "green", linetype = "dashed", size
  labs(title = "Income Distribution (Wave 7 - Austria)",
       x = "Income (thinc_m)",
       y = "Frequency",
       caption = "Red dashed line: Mean, Green dashed line: Median") +
  theme_minimal()
```



Getting rid of extreme values wave7AT

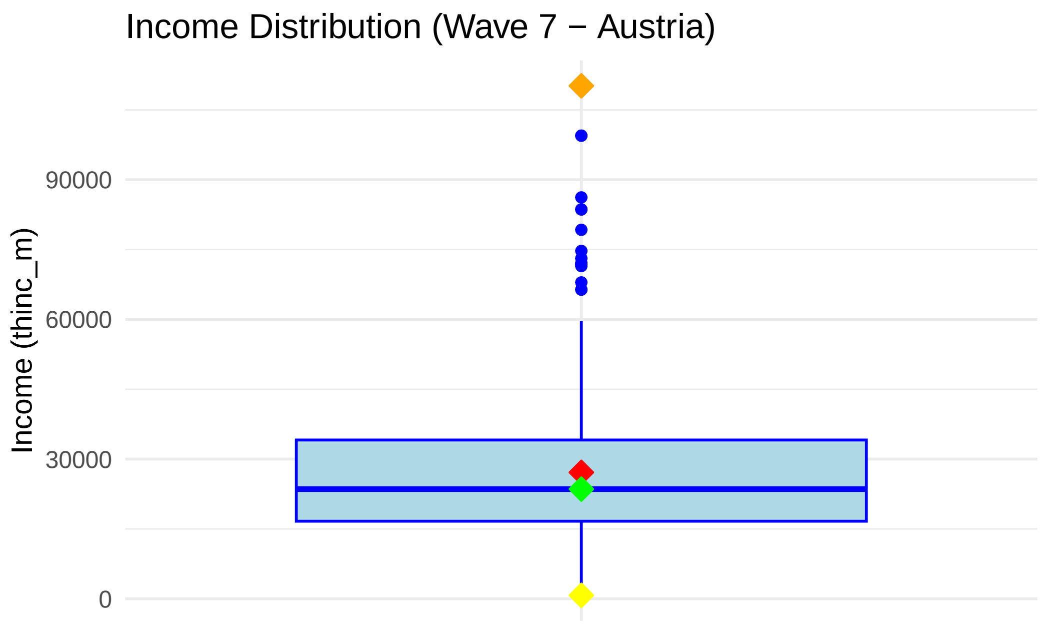
```
filtered_wave7dataAT <- subset(wave7dataAT, thinc_m >= 0 & thinc_m <= 120000)

income_summaryAT <- summary(filtered_wave7dataAT$thinc_m)

print(income_summaryAT)
```

```
  Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 733.5  16649.2  23550.3  27142.1  34093.0 110153.4
```

Graphical representation: average houshold income for Austria after getting rid of outliers option 1

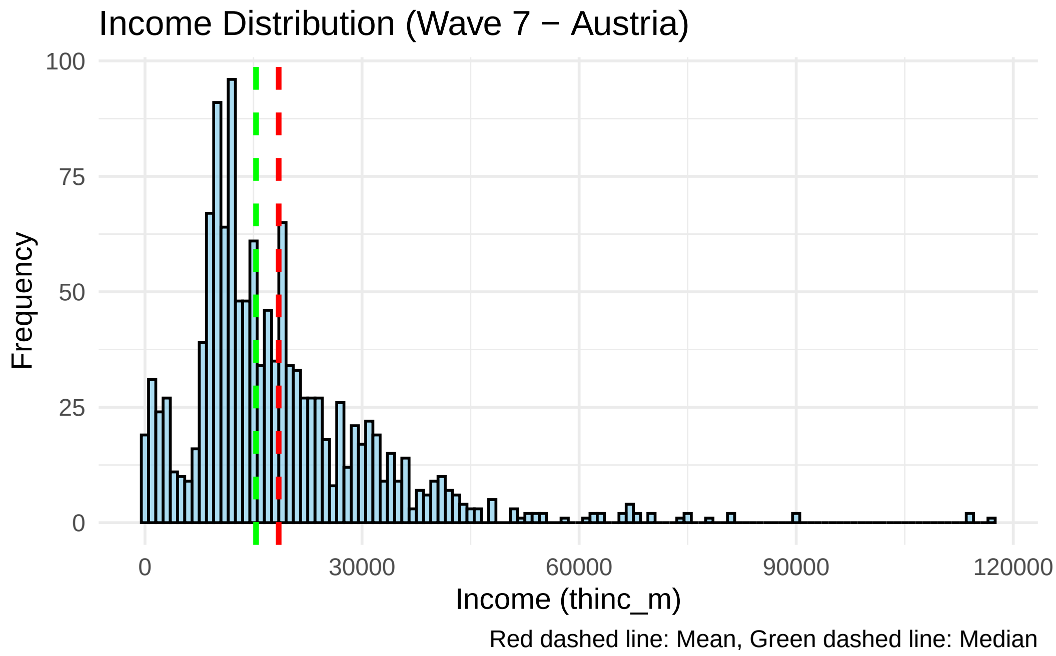
```
ggplot(filtered_wave7dataAT, aes(x = "", y = thinc_m)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "red", color = "re
  stat_summary(fun = median, geom = "point", shape = 23, size = 3, fill = "green", color =
  stat_summary(fun = min, geom = "point", shape = 23, size = 3, fill = "yellow", color = "
```

```
  stat_summary(fun = max, geom = "point", shape = 23, size = 3, fill = "orange", color = "
  labs(title = "Income Distribution (Wave 7 - Austria)",
       x = NULL,
       y = "Income (thinc_m)",
       fill = "Statistic") +
  theme_minimal()
```



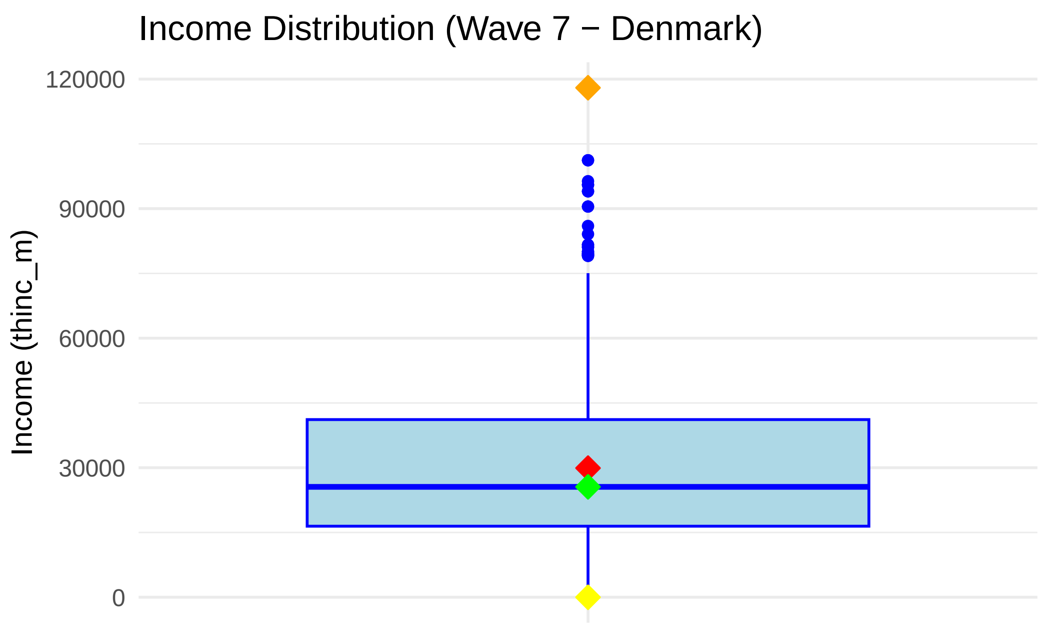
Graphical representation: average houshold income for Austria after getting rid of outliers option 2

```
ggplot(filtered_wave7dataSP, aes(x = thinc_m)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(thinc_m)), color = "red", linetype = "dashed", size = 1
  geom_vline(aes(xintercept = median(thinc_m)), color = "green", linetype = "dashed", size
  labs(title = "Income Distribution (Wave 7 - Austria)",
       x = "Income (thinc_m)",
       y = "Frequency",
       caption = "Red dashed line: Mean, Green dashed line: Median") +
  theme_minimal()
```



Graphical representation: average houshold income for Denmark Option 1

```
ggplot(wave7dataDK, aes(x = "", y = thinc_m)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "red", color = "re
  stat_summary(fun = median, geom = "point", shape = 23, size = 3, fill = "green", color =
  stat_summary(fun = min, geom = "point", shape = 23, size = 3, fill = "yellow", color = "
  stat_summary(fun = max, geom = "point", shape = 23, size = 3, fill = "orange", color = "
  labs(title = "Income Distribution (Wave 7 - Denmark)",
       x = NULL,
       y = "Income (thinc_m)",
       fill = "Statistic") +
  theme_minimal()
```



Getting rid of extreme values wave7AT

```
filtered_wave7dataDK <- subset(wave7dataDK, thinc_m >= 0 & thinc_m <= 120000)

income_summaryDK <- summary(filtered_wave7dataDK$thinc_m)

print(income_summaryDK)
```

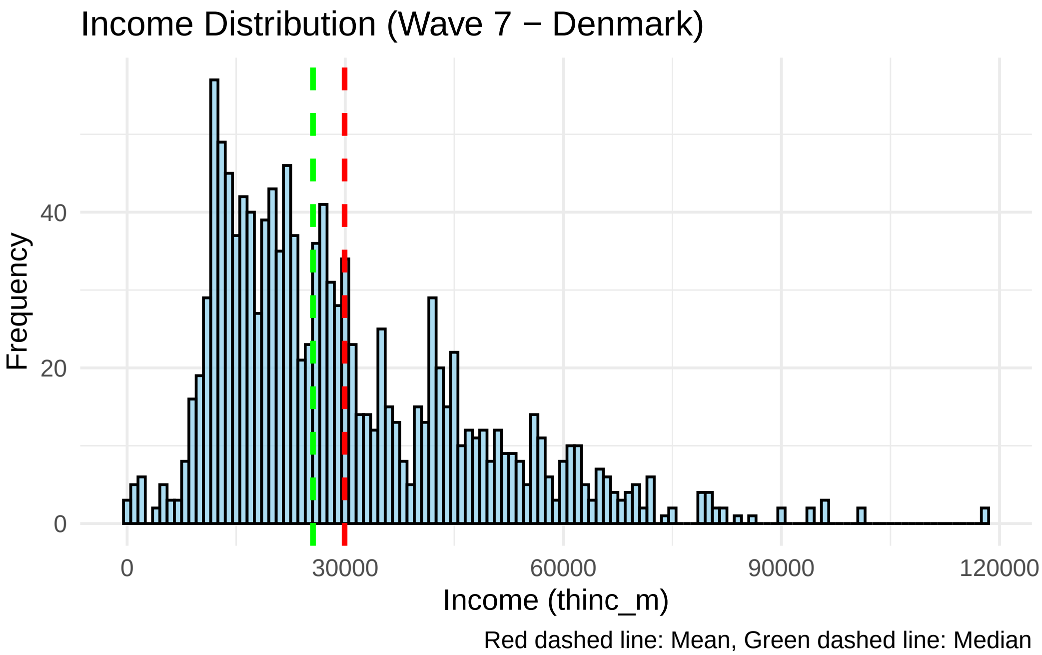
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0   16455   25563   29914   41147  117995
```

Our second comparison country Denmark :

Graphical representation: average houshold income for Denmark Option 2

```
ggplot(filtered_wave7dataDK, aes(x = thinc_m)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(thinc_m)), color = "red", linetype = "dashed", size = 1
  geom_vline(aes(xintercept = median(thinc_m)), color = "green", linetype = "dashed", size
```

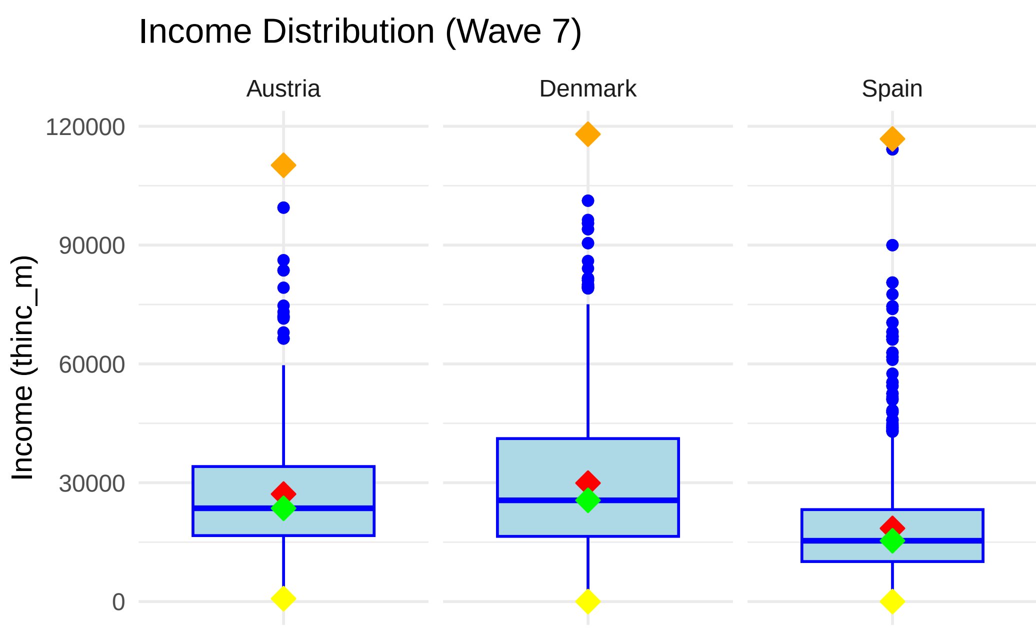
```
  labs(title = "Income Distribution (Wave 7 - Denmark)",
       x = "Income (thinc_m)",
       y = "Frequency",
       caption = "Red dashed line: Mean, Green dashed line: Median") +
  theme_minimal()
```



Side by side boxplots

```
combined_data <- rbind(
  transform(filtered_wave7dataDK, Country = "Denmark"),
  transform(filtered_wave7dataSP, Country = "Spain"),
  transform(filtered_wave7dataAT, Country = "Austria")
)
ggplot(combined_data, aes(x = "", y = thinc_m)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "red", color = "re
  stat_summary(fun = median, geom = "point", shape = 23, size = 3, fill = "green", color =
  stat_summary(fun = min, geom = "point", shape = 23, size = 3, fill = "yellow", color = "
  stat_summary(fun = max, geom = "point", shape = 23, size = 3, fill = "orange", color = "
  facet_wrap(~Country) +
```

```
  labs(title = "Income Distribution (Wave 7)",
       x = NULL,
       y = "Income (thinc_m)",
       fill = "Statistic") +
  theme_minimal()
```



Significance test (for the comparisons)

ANOVA : Testing income difference's significance :

```
# Combining the three datasets
combined_data <- rbind(
  mutate(filtered_wave7dataDK, Country = "Denmark"),
  mutate(filtered_wave7dataSP, Country = "Spain"),
  mutate(filtered_wave7dataAT, Country = "Austria")
)

# Perform one-way ANOVA
anova_result <- aov(thinc_m ~ Country, data = combined_data)
```

```r
# Check ANOVA summary
summary(anova_result)
```

```
               Df    Sum Sq   Mean Sq F value Pr(>F)
Country         2 8.734e+10 4.367e+10   173.6 <2e-16 ***
Residuals    3041 7.649e+11 2.515e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Perform post-hoc tests (e.g., Tukey's HSD) for pairwise comparisons if ANOVA is signific
if (summary(anova_result)[[1]][[5]][[1]] < 0.05) { # Check if p-value is less than 0.05
  posthoc_result <- TukeyHSD(anova_result)
  print(posthoc_result)
} else {
  print("ANOVA result is not significant. Post-hoc tests are not conducted.")
}
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = thinc_m ~ Country, data = combined_data)

$Country
                      diff         lwr       upr     p adj
Denmark-Austria   2771.924    783.8428  4760.005 0.0031273
Spain-Austria    -8667.435 -10656.5752 -6678.295 0.0000000
Spain-Denmark   -11439.359 -12908.5337 -9970.185 0.0000000
```

#The different measures of income inequality :

```r
# Extract thinc_m variable from filtered_wave7dataDK
thinc_m_spain <- na.omit(filtered_wave7dataDK$thinc_m)
# Compute Gini coefficient for Spain
gini_spain <- ineq::Gini(thinc_m_spain)
# Compute Palma ratio for Spain
palma_spain <- sum(thinc_m_spain[order(thinc_m_spain, decreasing = TRUE)][1:round(0.1*leng
                sum(thinc_m_spain[order(thinc_m_spain)][1:round(0.4*length(thinc_m_spain))
# Compute P90/P10 ratio for Spain
p90p10_spain <- quantile(thinc_m_spain, probs = 0.9, na.rm = TRUE) / quantile(thinc_m_spai
# Print results
```

```r
print(paste("Gini coefficient for Spain:", round(gini_spain, 3)))
```

```
[1] "Gini coefficient for Spain: 0.321"
```

```r
print(paste("Palma ratio for Spain:", round(palma_spain, 3)))
```

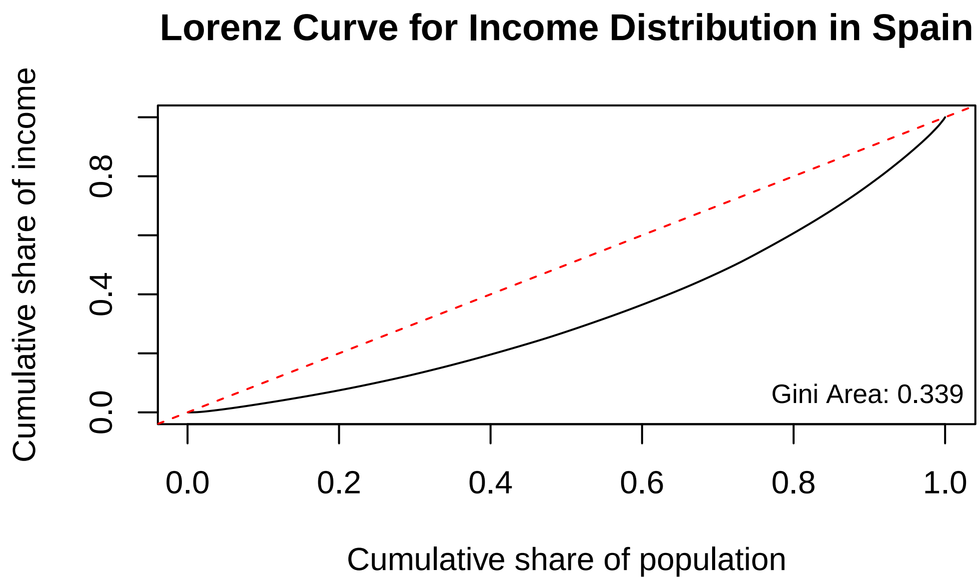
```
[1] "Palma ratio for Spain: 1.16"
```

```r
print(paste("P90/P10 ratio for Spain:", round(p90p10_spain, 3)))
```

```
[1] "P90/P10 ratio for Spain: 4.625"
```

LORENZ CURVE

```r
# Extract thinc_m variable from filtered_wave7dataDK
thinc_m_spain <- na.omit(filtered_wave7dataDK$thinc_m)

# Sort income data
thinc_m_sorted <- sort(thinc_m_spain)

# Calculate cumulative share of income
cumulative_income_share <- cumsum(thinc_m_sorted) / sum(thinc_m_sorted)

# Calculate cumulative share of population
cumulative_population_share <- seq_along(thinc_m_sorted) / length(thinc_m_sorted)

# Plot Lorenz curve with 45 degrees equality line and surface area calculation
plot(cumulative_population_share, cumulative_income_share, type = "l",
     xlab = "Cumulative share of population", ylab = "Cumulative share of income",
     main = "Lorenz Curve for Income Distribution in Spain")
# Add equality line
abline(0, 1, col = "red", lty = 2)  # 45-degree equality line

# Calculate area between the Lorenz curve and the equality line (Gini coefficient)
gini_area <- sum(diff(cumulative_population_share) * (cumulative_income_share[-1] + cumula
legend("bottomright", legend = paste("Gini Area:", round(gini_area, 3)), bty = "n", cex =
```



The three different measures of poverty and their interpretations

POVERTY headcount ratio

```
poverty_line <- quantile(thinc_m_spain, probs = 0.5, na.rm = TRUE) * 0.6

poverty_headcount <- sum(thinc_m_spain < poverty_line) / length(thinc_m_spain)

print(paste("Poverty Headcount Ratio for Spain:", round(poverty_headcount * 100, 2), "%"))
```

```
[1] "Poverty Headcount Ratio for Spain: 21.65 %"
```

Poverty gap index

```
poverty_gap <- mean(pmax(0, poverty_line - thinc_m_spain)) / poverty_line

print(paste("Poverty Gap Index for Spain:", round(poverty_gap, 4)))
```

```
[1] "Poverty Gap Index for Spain: 0.0549"
```

Squared poverty gap index

We chose the subjective variable poverty 'co007_':

```
squared_poverty_gap <- mean(pmax(0, poverty_line - thinc_m_spain)^2) / poverty_line^2

print(paste("Squared Poverty Gap Index for Spain:", round(squared_poverty_gap, 4)))
```

```
[1] "Squared Poverty Gap Index for Spain: 0.0235"
```

Measures of inequality based on health outcomes at the individual level :

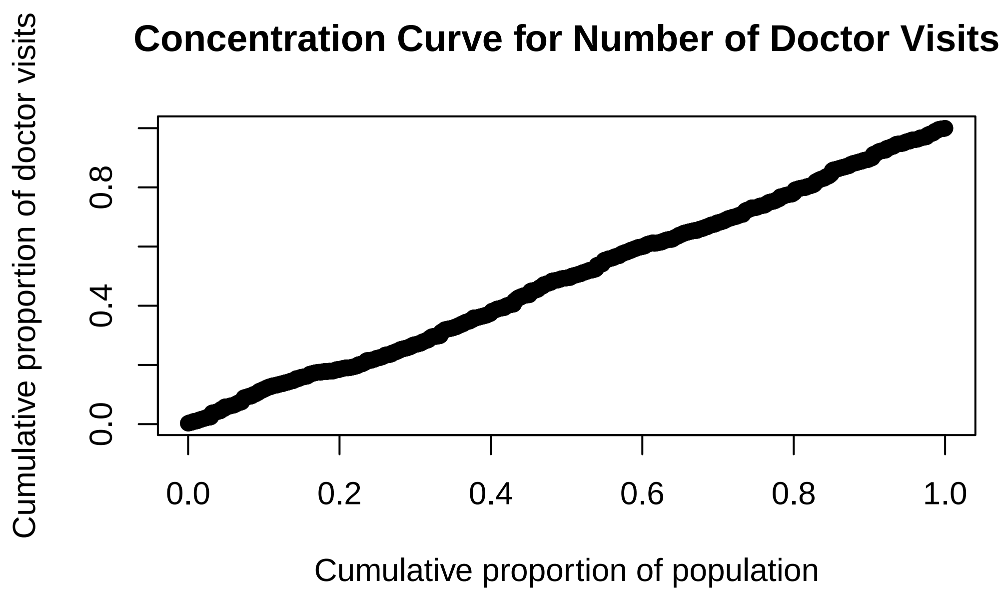
Doctor visits annually 'hc002_mod':

```
# Subseting the data to include only the relevant variables and remove negative values or

docvisitsdata <- wave7dataSP |>
  filter(thinc_m > 0, hc002_mod >= 0)

income_rank <- rank(docvisitsdata$thinc_m) / length(docvisitsdata$thinc_m)

mean_doctor_visits <- mean(docvisitsdata$hc002_mod)

covariance <- cov(docvisitsdata$hc002_mod, income_rank)

concentration_index <- 2 * covariance / mean_doctor_visits

print(paste("Concentration Index for number of doctor visits (excluding negative values):"
```

```
[1] "Concentration Index for number of doctor visits (excluding negative values): -0.025"
```

CI curve :

```
# Sorting the data by the fractional rank of hc002_mod (number of doctor visits)
doc_visits_data <- wave7dataSP[order(income_rank), ]

# Cumulative proportion of doctor visits
cumulative_doc_visits <- cumsum(doc_visits_data$hc002_mod) / sum(doc_visits_data$hc002_mod

# Cumulative proportion of the population
cumulative_population <- seq(0, 1, length.out = nrow(doc_visits_data))

# The concentration curve plot
plot(cumulative_population, cumulative_doc_visits, type = "b",
     xlab = "Cumulative proportion of population",
     ylab = "Cumulative proportion of doctor visits",
     main = "Concentration Curve for Number of Doctor Visits")
```



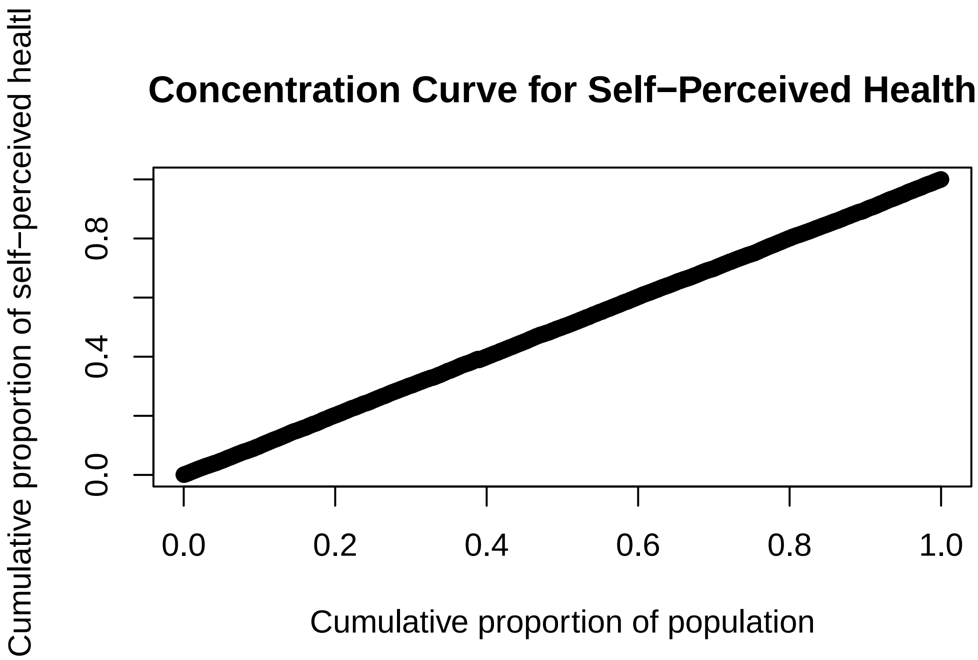
CI Value of the Self-perceived health variable study:

```
# Subseting the data to include only the relevant variables and remove negative values or
perceivedhealtdata <- wave7dataSP |>
  filter(thinc_m > 0, sphus >= 1)

income_rank <- rank(perceivedhealtdata$thinc_m) / length(perceivedhealtdata$thinc_m)

mean_health_status <- mean(perceivedhealtdata$sphus)

covariance <- cov(perceivedhealtdata$sphus, income_rank)

concentration_index <- 2 * covariance / mean_health_status

print(paste("Concentration Index for self-perceived health status:", round(concentration_i
```

```
[1] "Concentration Index for self-perceived health status: -0.031"
```

CI Curve of the Self-perceived health variable study

```
# Sorting the data by the fractional rank of sphus (self-perceived health)
perceived_health_data <- wave7dataSP[order(income_rank), ]

# Computing cumulative proportion of self-perceived health
cumulative_perceived_health <- cumsum(perceived_health_data$sphus) / sum(perceived_health_

# Computing cumulative proportion of the population
cumulative_population <- seq(0, 1, length.out = nrow(perceived_health_data))

# Ploting the concentration curve
plot(cumulative_population, cumulative_perceived_health, type = "b",
     xlab = "Cumulative proportion of population",
     ylab = "Cumulative proportion of self-perceived health",
     main = "Concentration Curve for Self-Perceived Health")
```



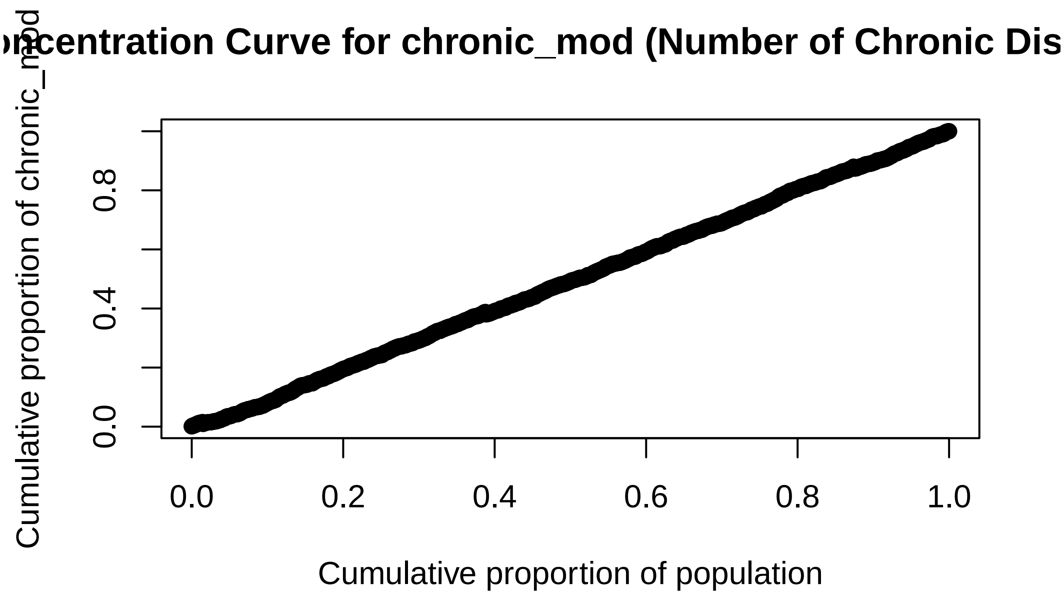
CI value of the number of chronic diseases variable study:

```
chrondisdata <- wave7dataSP %>%
  filter(thinc_m > 0, chronic_mod >= 0)  # Remove negative values for thinc_m and chronic_

income_rank <- rank(chrondisdata$thinc_m) / length(chrondisdata$thinc_m)

mean_chronic_diseases <- mean(chrondisdata$chronic_mod)

covariance <- cov(chrondisdata$chronic_mod, income_rank)

concentration_index <- 2 * covariance / mean_chronic_diseases

print(paste("Concentration Index for number of chronic diseases:", round(concentration_ind
```

```
[1] "Concentration Index for number of chronic diseases: -0.029"
```

CI curve of the number of chronic diseases variable study :

```
# Sorting the data by the fractional rank of chronic_mod
strdata <- wave7dataSP[order(income_rank), ]

# Computing the mean number of chronic diseases
mean_chronic_mod <- mean(strdata$chronic_mod)

# Calculating the fractional rank of chronic_mod
rank_chronic_mod <- rank(strdata$chronic_mod) / length(strdata$chronic_mod)

# Computing the covariance between chronic_mod and rank
covariance <- cov(strdata$chronic_mod, rank_chronic_mod)

# Computing the concentration index
concentration_index <- 2 * covariance / mean_chronic_mod

# Computing cumulative proportion of chronic_mod
cumulative_chronic_mod <- cumsum(strdata$chronic_mod) / sum(strdata$chronic_mod)

# Computing cumulative proportion of the population
cumulative_population <- seq(0, 1, length.out = nrow(strdata))

# Ploting the concentration curve
plot(cumulative_population, cumulative_chronic_mod, type = "b",
     xlab = "Cumulative proportion of population",
     ylab = "Cumulative proportion of chronic_mod",
     main = "Concentration Curve for chronic_mod (Number of Chronic Diseases)")
```



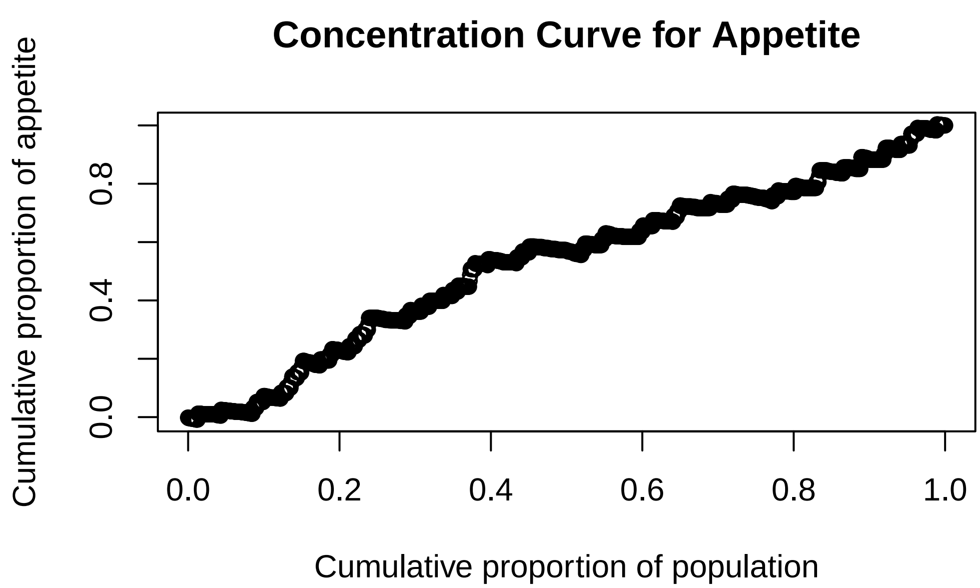
CI of the appetite variable study

```r
# Subseting the data to include only the relevant variables and remove negative values
appetitdata <- wave7dataSP |>
  filter(euro8 >= 0)  # Remove negative values for euro8

# Ranking individuals by socioeconomic status (thinc_m) and normalize ranks
income_rank <- rank(appetitdata$thinc_m) / length(appetitdata$thinc_m)

# Computing the mean appetite (0 or 1)
mean_appetite <- mean(appetitdata$euro8)

# Calculating the covariance between appetite and income ranks
covariance <- cov(appetitdata$euro8, income_rank)

# Computing the Concentration Index
concentration_index <- 2 * covariance / mean_appetite

# The Concentration Index
print(paste("Concentration Index for appetite:", round(concentration_index, 3)))
```

```
[1] "Concentration Index for appetite: -0.22"
```

CI of the appetite variable study

```
# Sorting the data by the fractional rank of euro8 (appetite)
appetite_data <- wave7dataSP[order(income_rank), ]

# Computing cumulative proportion of appetite
cumulative_appetite <- cumsum(appetite_data$euro8) / sum(appetite_data$euro8)

# Computing cumulative proportion of the population
cumulative_population <- seq(0, 1, length.out = nrow(appetite_data))

# Ploting the concentration curve
plot(cumulative_population, cumulative_appetite, type = "b",
     xlab = "Cumulative proportion of population",
     ylab = "Cumulative proportion of appetite",
     main = "Concentration Curve for Appetite")
```



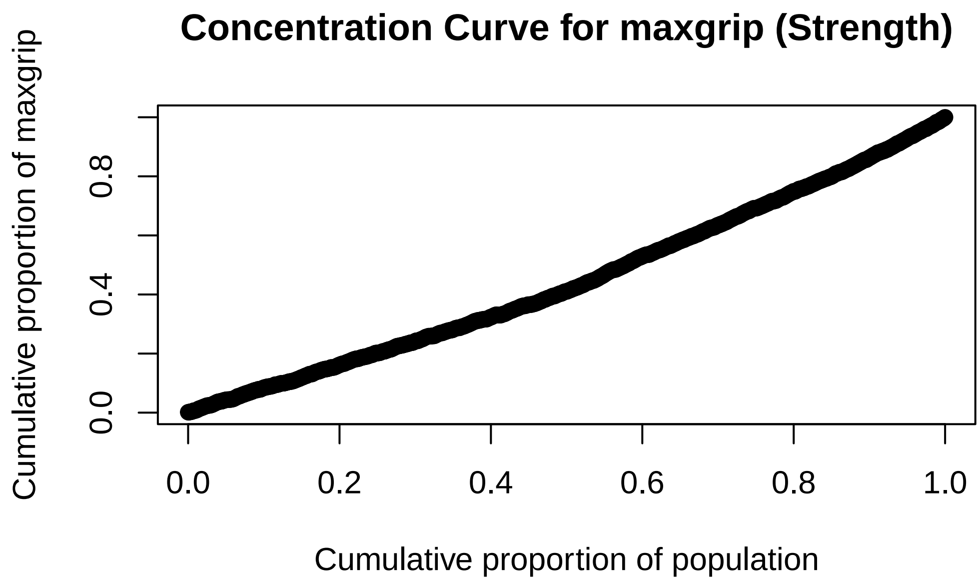
CI of the strenght variable study 'maxgrip'

```
# Subseting the data to include only the relevant variables and remove negative values
strenghtdata <- wave7dataSP %>%
  filter(thinc_m >= 0)  # Remove negative values for thinc_m

# Ranking individuals by socioeconomic status (thinc_m) and normalize ranks
income_rank <- rank(strenghtdata$thinc_m) / length(strenghtdata$thinc_m)

# Computing the mean grip strength
mean_grip <- mean(strenghtdata$maxgrip)

# Calculating the covariance between grip strength and income ranks
covariance <- cov(strenghtdata$maxgrip, income_rank)

# Computing the Concentration Index
concentration_index <- 2 * covariance / mean_grip

# The Concentration Index
print(paste("Concentration Index for grip strength (maxgrip):", round(concentration_index,
```

```
[1] "Concentration Index for grip strength (maxgrip): 0.104"
```

CI curve of the strenght variable study 'maxgrip'

```
# SortING the data by the fractional rank of maxgrip
strdata <- wave7dataSP[order(income_rank), ]

# Computing cumulative proportion of maxgrip
cumulative_maxgrip <- cumsum(strdata$maxgrip) / sum(strdata$maxgrip)

# Computing cumulative proportion of the population
cumulative_population <- seq(0, 1, length.out = nrow(strdata))

# Ploting the concentration curve
plot(cumulative_population, cumulative_maxgrip, type = "b",
     xlab = "Cumulative proportion of population",
     ylab = "Cumulative proportion of maxgrip",
     main = "Concentration Curve for maxgrip (Strength)")
```



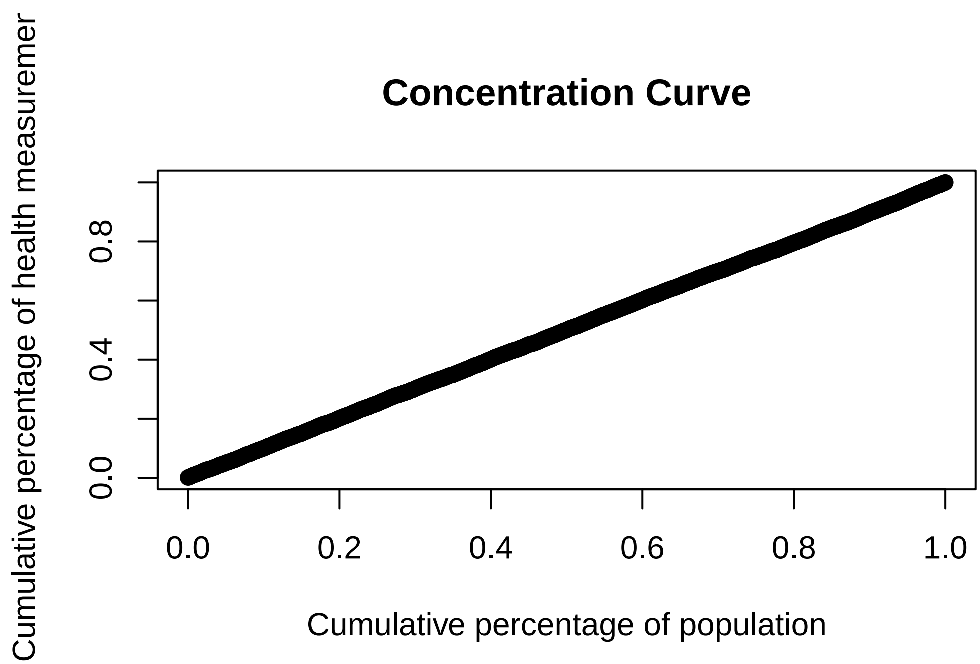
CI of the hospital stays number variable study

```
# Subseting the data to include only the relevant variables and remove negative values
hospital_data <- wave7dataSP %>%
  filter(hc012_ >= 0)  # Remove negative values for hc012_

# Ranking individuals by socioeconomic status (thinc_m) and normalize ranks
income_rank <- rank(hospital_data$thinc_m) / length(hospital_data$thinc_m)

# Computing the mean hospital stay (hc012_)
mean_hospital_stay <- mean(hospital_data$hc012_)

# Calculating the covariance between hospital stay and income ranks
covariance <- cov(hospital_data$hc012_, income_rank)

# Computing the Concentration Index
concentration_index <- 2 * covariance / mean_hospital_stay

# The Concentration Index
print(paste("Concentration Index for hospital stay (hc012_):", round(concentration_index,
```

```
[1] "Concentration Index for hospital stay (hc012_): 0.001"
```

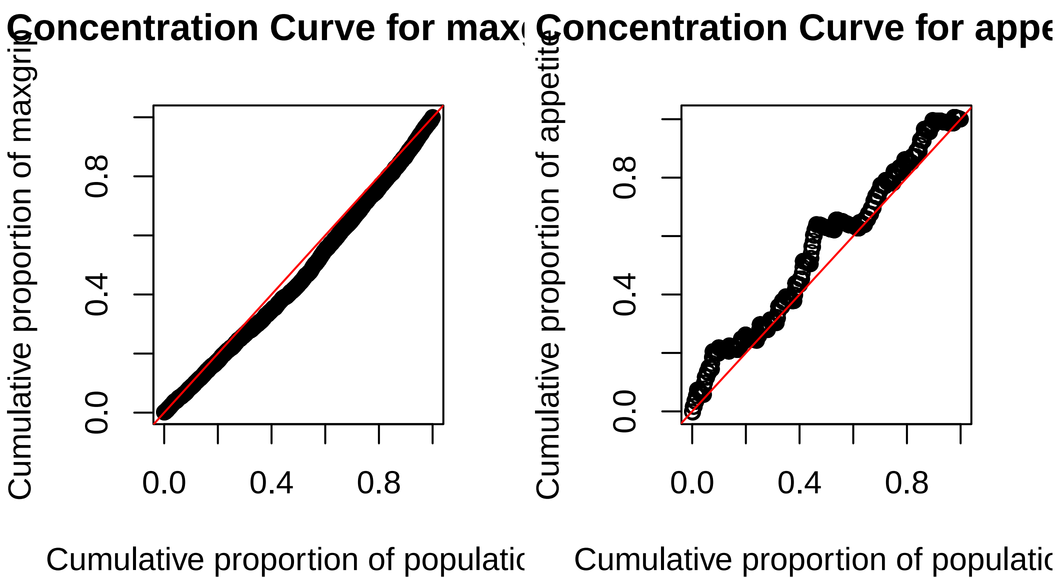
CI curve of the hospital stays number variable study

```r
# Sorting the data by income rank
sorted_data <- hospital_data[order(income_rank), ]

# Computing cumulative percentage of the population
cumulative_percentage <- seq(0, 1, length.out = nrow(sorted_data))

# Computing cumulative percentage of the health measurement variable (mean hospital stay)
cumulative_health_measurement <- cumsum(sorted_data$hc012_) / sum(sorted_data$hc012_)

plot(cumulative_percentage, cumulative_health_measurement, type = "b",
     xlab = "Cumulative percentage of population",
     ylab = "Cumulative percentage of health measurement",
     main = "Concentration Curve")
```



Visualise plots:

```
par(mfrow = c(1, 2))
# Sort the data by the fractional rank of maxgrip
strdata <- wave7dataSP[order(income_rank), ]

# Compute cumulative proportion of maxgrip
cumulative_maxgrip <- cumsum(strdata$maxgrip) / sum(strdata$maxgrip)

# Compute cumulative proportion of the population
cumulative_population <- seq(0, 1, length.out = nrow(strdata))

# Calculate the concentration index for maxgrip
concentration_index_maxgrip <- 2 * cov(strdata$maxgrip, income_rank) / mean(strdata$maxgri

# Plot the concentration curve for maxgrip
plot(cumulative_population, cumulative_maxgrip, type = "b",
     xlab = "Cumulative proportion of population",
     ylab = "Cumulative proportion of maxgrip",
     main = "Concentration Curve for maxgrip")
abline(0, 1, col = "red")  # Add the 45-degree line

# Now repeat the process for the appetite variable
# Sort the data by the fractional rank of appetite
appetite_data <- wave7dataSP[order(income_rank), ]

# Compute cumulative proportion of appetite
cumulative_appetite <- cumsum(appetite_data$euro8) / sum(appetite_data$euro8)

# Calculate the concentration index for appetite
concentration_index_appetite <- 2 * cov(appetite_data$euro8, income_rank) / mean(appetite_

# Plot the concentration curve for appetite
plot(cumulative_population, cumulative_appetite, type = "b",
     xlab = "Cumulative proportion of population",
     ylab = "Cumulative proportion of appetite",
     main = "Concentration Curve for appetite")
abline(0, 1, col = "red")  # Add the 45-degree line
```



What can explain health inequalities across individuals ? Propose an econometric analysis and discuss your results:

What variables do we think are statistically significant and why are they not indeed the case:

Linear Regression with some dummies:

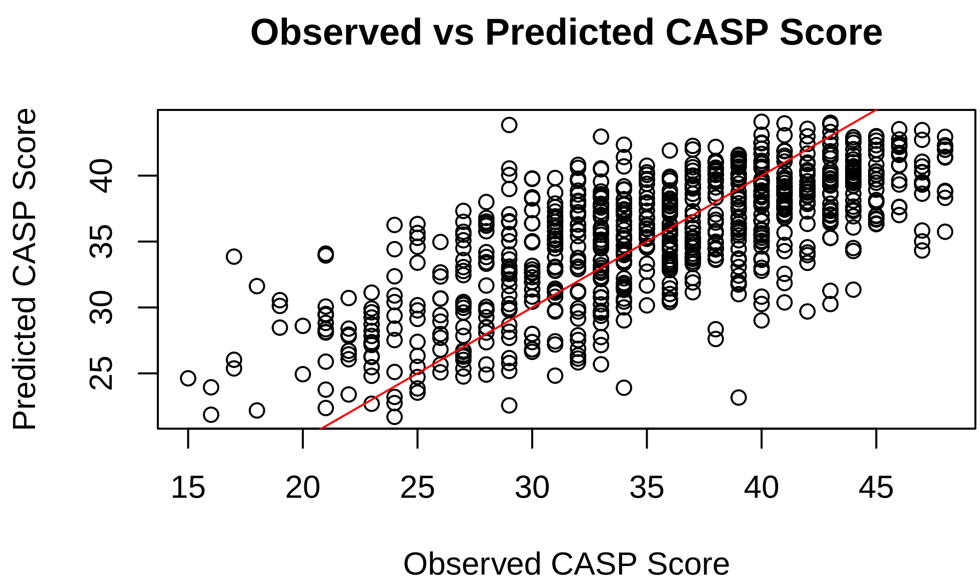
```
# Filter the dataset to exclude negative values and missing values
filtdata <- wave7dataSP %>%
  filter(casp >= 0, mobilityind >= 0, lgmuscle >= 0,
         numeracy_2 >= 0, smoking >= 0, maxgrip >= 0, euro2 >= 0,
         euro4 >= 0, euro5 >= 0, euro6 >= 0, euro10 >= 0, chronic_mod >= 0, sphus>= 0, thi

# Perform the regression analysis
regression_model <- lm(casp ~    mobilityind + lgmuscle + numeracy_2 + smoking + maxgrip +
                      euro10 + chronic_mod + sphus + thinc_m , data = filtdata)

# View summary of the regression model
summary(regression_model)
```

```
Call:
lm(formula = casp ~ mobilityind + lgmuscle + numeracy_2 + smoking +
    maxgrip + euro2 + euro5 + euro6 + euro10 + chronic_mod +
    sphus + thinc_m, data = filtdata)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8557  -3.1503   0.2228   3.3525  15.8342

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.038e+01  1.217e+00  33.192  < 2e-16 ***
mobilityind -5.211e-01  2.367e-01  -2.202  0.02796 *
lgmuscle    -2.238e-01  2.129e-01  -1.051  0.29358
numeracy_2   3.234e-01  1.035e-01   3.124  0.00185 **
smoking      2.452e-01  1.449e-01   1.692  0.09103 .
maxgrip      3.938e-02  2.088e-02   1.886  0.05972 .
euro2       -3.298e+00  4.206e-01  -7.840 1.45e-14 ***
euro5       -9.811e-01  3.758e-01  -2.611  0.00920 **
euro6       -2.797e+00  4.919e-01  -5.686 1.82e-08 ***
euro10      -2.084e+00  4.210e-01  -4.950 9.08e-07 ***
chronic_mod -8.800e-02  1.515e-01  -0.581  0.56154
sphus       -1.585e+00  2.179e-01  -7.277 8.19e-13 ***
thinc_m      4.500e-05  1.496e-05   3.008  0.00271 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.883 on 793 degrees of freedom
Multiple R-squared:  0.4961,    Adjusted R-squared:  0.4885
F-statistic: 65.07 on 12 and 793 DF,  p-value: < 2.2e-16
```

"Observed vs Predicted CASP Score" Final Graph:

```r
# Filter out negative values for all relevant variables
filtdata <- subset(wave7dataSP,
                   mobilityind >= 0 &
                   lgmuscle >= 0 &
                   numeracy_2 >= 0 &
                   smoking >= 0 &
```

```r
                   maxgrip >= 0 &
                   euro2 >= 0 &
                   euro5 >= 0 &
                   euro6 >= 0 &
                   euro10 >= 0 &
                   chronic_mod >= 0 &
                   sphus >= 0 &
                   thinc_m >= 0 &
                   casp >= 0)

# Fit the regression model
model <- lm(casp ~ mobilityind + lgmuscle + numeracy_2 + smoking +
             maxgrip + euro2 + euro5 + euro6 + euro10 + chronic_mod +
             sphus + thinc_m, data = filtdata)

# Plot the regression line
plot(filtdata$casp,
     predict(model),
     xlab = "Observed CASP Score",
     ylab = "Predicted CASP Score",
     main = "Observed vs Predicted CASP Score")

# Add a reference line with a slope of 1 (for perfect prediction)
abline(a = 0, b = 1, col = "red")
```


Observed vs Predicted CASP Score
Predicted CASP Score
25
30
35
40
15
20
25
30
35
40
45
Observed CASP Score