

**DSAN 5100 Final Project**

**Income Inequality Around the World**

**December 6th, 2023**

**Group 12: Rongmu Sun, Xue Qin, Kaiyang Li, Tianze Yang**

# Agenda

## **1. Introduction**

1.1 Background Information

1.2 Questions

## **2. Data Collection and Clean**

## **3. Data Analysis**

3.1 Income Inequality with Gender and Race

3.1.1 EDA

3.1.2 Hypothesis Testing and Methods

3.1.3 Results

3.2 Income Inequality with Ages and Across Different Income Groups

3.2.1 EDA

3.2.2 Hypothesis Testing and Methods for Age Groups and Mean Income

3.2.3 Hypothesis Testing and Method Across Different Income Groups and Mean Income

3.2.4 Results

3.3 Income Inequality with Regions

3.3.1 EDA

3.3.2 Hypothesis Testing and Methods

3.3.3 Results

## **4. Conclusion**

## **5. References**

# 1. Introduction

In the past centuries, income inequality has always been a popular research topic related to economics. Income is defined as the consumption and saving opportunity gained by an entity within a specified timeframe, generally expressed in monetary terms. Generally, it refers to the amount of money, property, and other transfers of value received over a set period in exchange for services or products. In a self-contained economy, income is created in production through factors such as land, labor, capital, and entrepreneurship (Kakwani, 1980).

Inequality is about the distribution of power and resources, the rights people can exercise, and the opportunities they can access. Numerous theories have emerged to explain the distribution of individual income, socioeconomic school is one of the most important. This school focuses on understanding income distribution by examining several economic and institutional variables. These include aspects like gender, age, occupation, education level, regional disparities, and the distribution of wealth, among others (Kakwani, 1980).

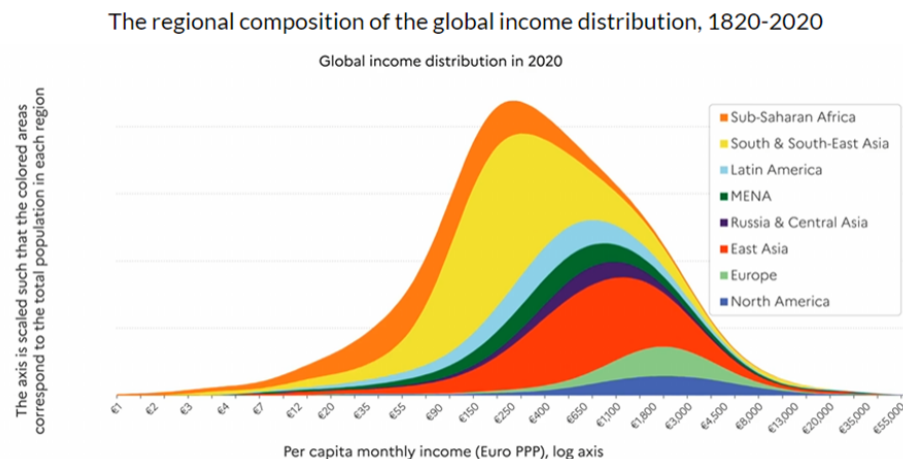
In our study, we will utilize the socioeconomic school of thought principles as our foundational theoretical perspective. This approach will guide our examination of the relationships between diverse socioeconomic factors — such as gender, race, age, and regional disparities — and their impact on income inequality. Our objective is to conduct a detailed analysis of income inequality, both on a global scale and within the specific context of the United States. By employing scientific statistical techniques, we plan to delve into the data, seeking to understand the underlying factors contributing to income distribution disparities.

## 1.1 Background Information

The 2022 World Inequality Report delves into the stark realities and hypothetical extremes of global income distribution. It describes global income equality in an idealized scenario: if income were perfectly distributed, every individual across the globe would earn an equal amount of €16,700 per year. In this model, income distribution would be balanced, with the bottom 50% of earners receiving exactly half of the global income and the top 10% receiving a proportionate 10%. Conversely, a scenario of extreme inequality happens when the economic divide is so severe that the bottom half of the global population has no share in the total income, while the top 10% amass 100% of it. These two theoretical extremes serve to frame the discussion of global income inequality, illustrating the potential range from perfect equality to absolute disparity (World Inequality Report, 2022).

However, the reality of global income distribution is far from these hypotheses; the actual data reveals a significant imbalance. The bottom 50% of the world's population, which represents the majority, only accounts for a small portion of global income, which is only 8.5%. This starkly contrasts with the disproportionate wealth of the top 10%, who command more than half of the total global income, which is 52%. This amount is five times greater than the worldwide average, highlighting a substantial concentration of wealth in the hands of a few. The middle 40% of the world's population sits in between these extremes, holding 39.5% of the global income. Their

share is closer to the worldwide average but still reveals significant differences in income distribution (World Inequality Report, 2022).



Graph 1: The Regional Composition of the Global Income Distribution, 1820-2020

<https://wir2022.wid.world/chapter-2/#:~:text=The%20first%20striking%20finding%20is,remained%20around%205%2D15%25.>

## 1.2 Questions

Our project utilizes the theoretical framework provided by the socioeconomic school, concentrating specifically on analyzing income disparities across four key factors: race, gender, age, and region. These dimensions are critical in understanding the multifaceted nature of income inequality. To bring clarity to our data, we utilize EDA to visually present and interpret the data in an insightful and accessible manner. After building upon this visual exploration, we develop specific hypotheses for each factor. Then, we employ statistical methods to test these hypotheses; we aim to establish empirical relationships between these factors and income inequality through these techniques. Finally, we provide a detailed understanding of the factors contributing to income inequality.

The following are the problems we will solve in this project:

1. What is the global trend about income inequality?
2. What are the factors that may be related to income inequality?
3. Has the extent of income inequality worsened over time?
4. Is there a relationship between gender and income level?
5. Is there a relationship between race and income level?
6. Is there a relationship between age and income level?
7. Are there differences in income inequality across different income groups?
8. Is there a relationship between the region where the person is located and income level?
9. Is there a trend that developed countries have an overall higher average income than developing countries?


10. Do wealthy people from developed regions have more income inequality than those from developing regions?

## 2. Data Collection and Clean

### 2.1 Data Collection

Our research project sourced three distinct data sets from two reputable sources: data.world.com and the U.S. Census Bureau (census.gov). Each data set offers a unique perspective on income distribution, which is essential for our comprehensive analysis.

The first data set is segmented by gender and age, focusing exclusively on the income of individuals within the United States. This data allows us to explore how income varies across different age groups and between genders. By analyzing this data, we can investigate trends such as the gender pay gap and how income potential changes with age.



	year	age_group	# number_with_income_male_thousnads	# current_dollars_male_median	# 2017_c
1	2017	15 & Older	114849	40396	
2	2016	15 & Older	113158	38869	
3	2015	15 & Older	112322	37138	
4	2014	15 & Older	110372	36302	
5	2013	15 & Older	109937	35630	

Showing 1-5 of 470 rows, 12 columns See all

Switch to column overview

Picture 2: Dataset 1 <https://data.world/m00nlight/incomeinequality>

The second data set categorizes income data based on racial demographics within the United States. By analyzing this data, we can discover patterns of economic inequality that are influenced by racial factors, offering a critical examination of the economic implications of different races in the U.S.

Year	Number (thousands)	Upper limit of each fifth (dollars)				Lower limit of top 5 percent (dollars)
		Lowest	Second	Third	Fourth	
2022	8,160	42,030	85,000	133,600	213,600	403,100
2021	7,852	38,010	77,000	126,200	201,000	377,700
2020 (41)	7,555	37,740	73,300	121,100	200,000	360,400
2019	7,334	40,200	77,000	120,500	190,300	356,600
2018	7,416	33,600	67,360	107,700	175,400	323,600
2017 (40)	7,124	32,410	63,810	102,500	166,100	308,100
2017	7,114	32,000	64,500	101,400	169,100	310,000
2016	6,750	31,350	65,000	100,000	160,100	274,300
2015	6,640	30,000	60,000	97,580	157,300	280,100
2014	6,333	28,800	58,880	93,670	147,000	256,500
2013 (39)	6,160	28,200	57,140	90,150	146,200	260,000
2013 (38)	6,111	26,010	53,000	84,000	136,000	243,100
2012	5,872	27,960	54,280	84,960	137,000	244,000
2011	5,705	25,000	50,050	80,300	125,400	224,200

Picture 3: Dataset 2 <https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-inequality.html>

The third data set takes a broader approach, classifying income distribution on a global scale by region. This dataset helps us do the analysis of how income is distributed across different parts of the world, highlighting regional economic disparities. By analyzing this data, we can understand the difference in income inequality patterns between developed countries and developed countries, examining how geographic location influences economic status and opportunity.


**WIID3.4\_19JAN2017New.xlsx**
[raw data https://www.wider.unu.edu/project/wiid-world-income-inequ...](https://www.wider.unu.edu/project/wiid-world-income-inequ...)

[View](#)
↓
>
⌵

	country	countrycode3	countrycode2	year	# population	incomegroup
1	Afghanistan	AFG	AF	2007	26.9115	Low income
2	Afghanistan	AFG	AF	2008	27.6589	Low income
3	Albania	ALB	AL	1996	3.092	Upper middle income
4	Albania	ALB	AL	2002	3.1231	Upper middle income
5	Albania	ALB	AL	2005	3.0822	Upper middle income

 Showing 1-5 of 8,817 rows, 54 columns [See all](#)
[Switch to column overview](#)

Picture 4: Dataset 3 <https://data.world/moonlight/incomeinequality>

These datasets, collectively, provide a multifaceted view of income inequality. They allow us to cross-examine these factors and their interplay in shaping income distributions both within the United States and globally.

## 2.2 Data Clean

Our data set about the region is raw data directly from the World Income Inequality Database, which needs some data cleaning process. It comprises 54 columns but contains several instances of redundancy, where different columns essentially convey the same information. For instance, columns 1, 2, and 3, despite having different labels, all represent the same attribute, which is the name of the region. Additionally, many columns present the same dimension of data but are calculated using both old and new methodologies, leading to further duplication of information. To address these redundancies, we initiated a feature selection process, identified all duplicate columns, and evaluated each calculation function, then deleted 15 redundant columns, reducing the dimensionality from 54 to 39.

```
{r}
library(dplyr)
df <- select(df, -c(Countrycode3, Countrycode2, P5, P95, Revision, AreaCovr, PopCovr, AgeCovr,
Welfaredefn, Mean_usd, Median_usd, Exchangerate, Equivsc, UofAnala, IncSharU, ))
```

Picture 5: R code

In our analysis of regional and income inequality, a key focus is examining the income share of the top 20% of the population. This approach helps us understand the extent of wealth concentration at the higher end of the income spectrum. However, we encountered a challenge

with the dataset's consistency: each row of data may be sourced from different origins, leading to variations in how income distribution is recorded. In some rows, data on the income distribution within a region for a given year is missing, resulting in empty entries for the Decile and Quintile columns.

Given this inconsistency, simply calculating mean or median values was not a viable solution to fill these gaps, as they could potentially distort the analysis. To solve this issue, we added the columns sum of Decile and sum of Quintile. Then, we can employ them as filters to ensure data completeness and reliability, ensuring that at least one of these two new columns (Sum of Decile or Sum of Quintile) should not be zero.

```
{r}
df$sum_D1_to_D10 <- rowSums(df[, c("D1", "D2", "D3", "D4", "D5", "D6", "D7", "D8", "D9", "D10"
)])
df$sum_Q1_to_Q5 <- rowSums(df[, c("Q1", "Q2", "Q3", "Q4", "Q5")])
```

Picture 6: R code

In addition to the Decile and Quintile columns, our dataset consists predominantly of categorical variables. To ensure the integrity and completeness of our data, we evaluated the count of NAs for all categorical variables. We found out that they all constituted a relatively small portion of the data, so we decided to fill the NAs with the most common value or value that first appears grouped by region.

```
{r}
df <- df %>%
  group_by(Country) %>%
  mutate(Currency = ifelse(is.na(Currency),
                           first(Currency[!is.na(Currency)]),
                           Currency))
```

Picture 7: R code

Through the data cleaning processes we implemented, we have significantly enhanced the quality of our datasets, establishing a solid and reliable foundation for our research. This comprehensive and representative data now sets the stage for further data description and analysis.

### 3. Data Analysis

#### 3.1 Income Inequality with Gender and Race

When considering the inequality of income, first, we thought of the factors of race and gender, which got the most social attention. In this project, we would focus on the scope of the U.S and use the data from open, accessible government datasets on the mean and median income of both genders and the quartile information of the income in different races within the U.S. Since income data were relatively sensitive, it was hard to obtain public access data to the original dataset. Instead, we could only access the dataset with aggregated data, such as median and mean values, which could be found on the government website.

### 3.1.1 EDA

In this part, we looked at the descriptive statistics of the dataset to get an idea of what the data looks like. Because we would not do any time-series analysis, we should drop the time column and treat the whole dataset as a snapshot from a period. In the dataset of income with different genders, we can see that the descriptive statistics are as follows:

For the other dataset, the quartile information on the income of different genders, we will use the dataset for black, white, and Asian for our analysis. Also, we drop the time column. Here are the basic summary statistics of those datasets.

M_median		F_median		M_mean		F_mean	
Min.	: 9301	Min.	: 7360	Min.	:13149	Min.	:10569
1st Qu.	:28532	1st Qu.	:16201	1st Qu.	:38568	1st Qu.	:23119
Median	:34296	Median	:22281	Median	:49320	Median	:29800
Mean	:33415	Mean	:21352	Mean	:46613	Mean	:28764
3rd Qu.	:42116	3rd Qu.	:26449	3rd Qu.	:58734	3rd Qu.	:35890
Max.	:53985	Max.	:35444	Max.	:75005	Max.	:47448

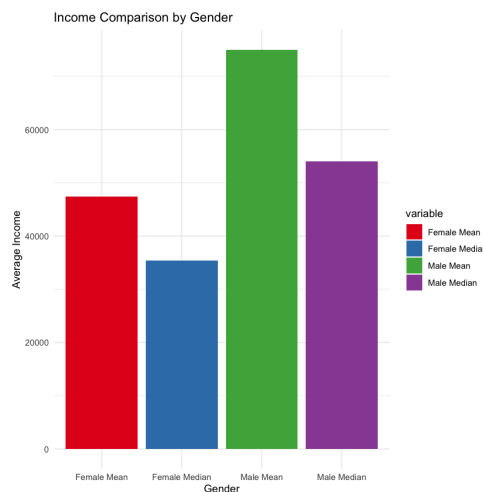
  

Age_Group	Year_Group
Length:126	Min. : 1.00
Class :character	1st Qu.: 32.25
Mode :character	Median : 63.50
	Mean : 63.50
	3rd Qu.: 94.75
	Max. :126.00

Picture 8: Summary table

From these statistics summaries, we can see that the difference in both average median and average mean for the two genders is significant; males had a higher value in income in terms of both mean and median. Conversely, the mean and median values of all income quartiles for the three races were also different. For more observations, we need to visualize the data.

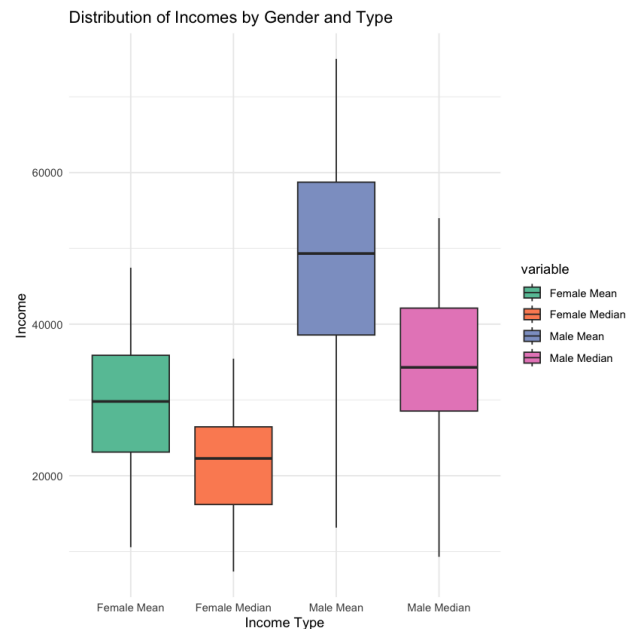
First, we did a bar plot of the average value of the mean and median income for males and females in the U.S. We can see clearly that there is a difference in both the mean and median average for the two groups and male values are considerably higher than females.



Graph 9: Income Comparison by Gender

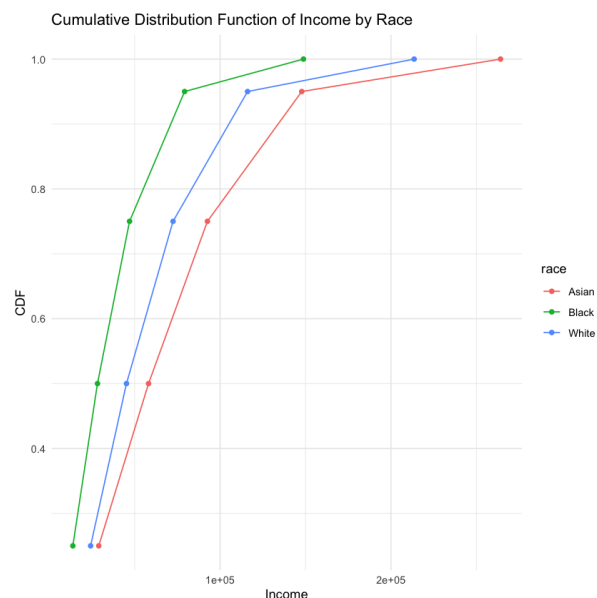


Next, we do a box plot to see the distribution of both genders' mean and median income. From it, we can see that not only males generally have a more significant income, but also the dispersion of incomes was higher in the male group in terms of both mean and median values.



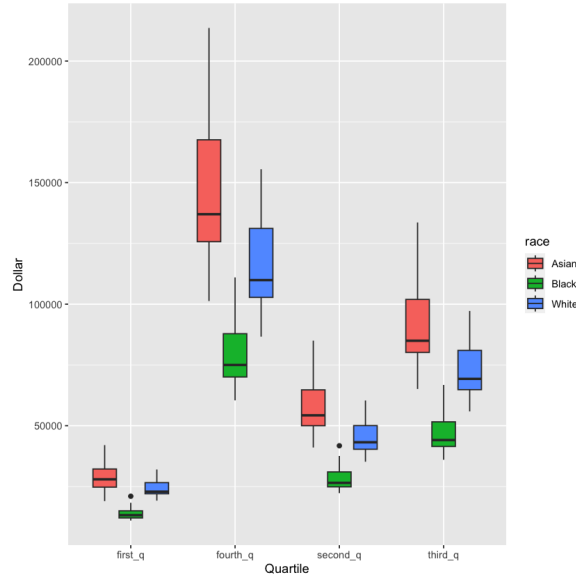
Graph 10: Distribution of Incomes by Gender and Types

Because we can obtain the quartile information on the income of different races, we can make a CDF plot for the three races. The value for each quartile was calculated as the mean of the value in that quartile within the dataset. We can see three lines, each representing the cumulative distribution of the income of a race. It is easy to observe that none of the three lines overlap or tend to be so, indicating that the income distribution in the three races could be distinct.



Graph 11: Cumulative Distribution Function of Income by Race

We also did another box plot on the income of each quartile for different races. In this plot, we can see that not only do some races clearly have a more significant income, but they also have a greater dispersion of income in all the quartiles. We also noticed that race has a higher income in one quartile and tends to be so in every other quartile as well, which indicates an income imbalance issue not only in the scope of a particular income group but many.



Graph 12: Quartile of income for each race

To conclude the EDA for this part, we found evidence that the income is imbalanced in the two genders and not equally distributed in different races. But to confirm our observations, we must proceed to hypothesis testing to see if the difference has significance in statistics or just a fluke of chance.

### 3.1.2 Hypothesis Testing and Methods

For the first question, the imbalance of income in the two genders, we had the mean and median of the income to use. As a result, we proposed a paired T-test on the mean of the income in both genders. This was because the t-test was done most effectively when dealing with normally distributed data, especially in this limited sample space, where the distribution of the original data matters. But, since a large sample size aggregates the mean income, we can assume that it is normally distributed since the Central Limit Theorem.

In this case, the hypothesis will be the following:

$H_0$ : *There is no relationship between the gender and average income mean value.*

$H_a$  : *There is a relationship between the gender and average income mean value.*

```
male_mean <- subset$M_mean  
female_mean <- subset$F_mean
```

```
t.test(male_mean, female_mean, var.equal = TRUE)
```

Two Sample t-test

```
data: male_mean and female_mean  
t = 10.696, df = 250, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 14562.51 21135.87  
sample estimates:  
mean of x mean of y  
46613.08 28763.89
```

Picture 13: t-test

Here, we can see the result of the T-Test has a p-value much less than 0.05, meaning we can reject the null hypothesis that gender has no relationship with the income mean value.

Secondly, for the median value of the income, there is no distribution we can imply, so we do not want to re-apply the T-test again. Since we are not very rich in data, it is a good setup for bootstrap analysis. The bootstrap method is one kind of Monte Carlo method that is most frequently used in statistical analysis. It randomly resamples the samples from the dataset and simulates it thousands of times. Compared to normal hypothesis testing, the bootstrap method, by resampling over and over again, can give the sampling distribution of the mean, and it may be more accurate (Frost, 2023). In this case, we will bootstrap sample the median and the mean income of the two genders for a sample size of 10000. Then, we check the difference between the distribution of the two bootstrap samples and determine the result by the 95% confidence interval.

In this case, the hypothesis will be as follows:

$H_0$  : *There is no difference between the mean of median income from the two genders.*

$H_a$  : *There is a difference between the mean of median income from the two genders.*

```

N <- 10000

# do the bootstrap
male_median <- subset$M_median
female_median <- subset$F_median

# create a vector to store the results
bootstrap_median_m <- vector("numeric", N)
bootstrap_median_f <- vector("numeric", N)

for(i in 1:N){
  bootstrap_median_m[i] <- mean(sample(male_median, size = length(male_median), replace
  bootstrap_median_f[i] <- mean(sample(female_median, size = length(female_median), rep
}

diff <- bootstrap_median_m - bootstrap_median_f

quantile(diff, c(0.025, 0.975))

2.5%: 9605.96329365079 97.5%: 14467.2888888889

```

Picture 14: Bootstrap

We can see that 0 is far from present in the 95% confidence interval, so we reject the null hypothesis.

Thirdly, for the race issue, we have the quartile limit of the income for each race, so again, we can not imply any distribution on them. Also, because of the limited sample size, we decided to do a bootstrap test on each of the quartile upper limits income for the three races. Like the one we did the last time, we will check the 95% confidence interval of the difference in the bootstrapped distribution.

In this case, the Hypothesis Test would be as follows:

$H_0$ : There is no difference between the mean distribution of the upper limit of the second income quartile in different races.

$H_a$ : There is a difference between the mean distribution of the upper limit of the second income quartile in different races.

```

diff_white_asian <- bootstrap_white_q2 - bootstrap_asian_q2
diff_black_asian <- bootstrap_black_q2 - bootstrap_asian_q2
diff_black_white <- bootstrap_black_q2 - bootstrap_white_q2

# 95% confidence interval
quantile(diff_white_asian, c(0.025, 0.975))

2.5%: -18488.2826086957 97.5%: -7333.29347826087

quantile(diff_black_asian, c(0.025, 0.975))

2.5%: -34983.0543478261 97.5%: -24794.3043478261

quantile(diff_black_white, c(0.025, 0.975))

2.5%: -20527.1195652174 97.5%: -13309.9239130435

```

Picture 15: Bootstrap

We can see that 0 is absent in the confidence interval between any of the three races' distributions, so we reject the null hypothesis.

Lastly, we also check for differences between the top 5% of income in different races. Because there is no implied distribution for this and the limited sample size, we will do a permutation test, which is non-parametric and suitable for the test with a small sample size. Here, we shuffle the race labels from the total distribution 10000 times; each time, we will calculate the difference between the maximum and minimum mean income of each group and check if it is as extreme as the one we observed in the original data.

In this case, the hypothesis test would be as follows:

$H_0$ : The lower limit of top 5% income has no relationship with the race

$H_a$ : The lower limit of top 5% income has a relationship with race.

```
n_permutations <- 10000
observed_means <- tapply(compare_df$lower_top5, compare_df$race, mean)
observed_statistic <- max(observed_means) - min(observed_means)

permuted_statistics <- replicate(n_permutations, {
  shuffled_race <- sample(compare_df$race)
  means <- tapply(compare_df$lower_top5, shuffled_race, mean)
  max(means) - min(means)
})

# Calculate the p-value
p_value <- mean(permuted_statistics >= observed_statistic)
p_value
```

0

Picture 16: Bootstrap

We can see from the result that the p-value is less than 0.05, as a result, we can reject the null hypothesis.

### 3.1.3 Results

To conclude, our analysis revealed significant income disparities between genders. The mean and median incomes for men were consistently higher than those for women across various sectors and age groups. Specifically, the mean income for men was approximately 20% higher than women's. Both t-tests on the mean income of the two genders and bootstrap tests conducted on the median incomes had results that align with these findings, resulting in a p-value < 0.05 and 0 not presented in the confidence interval, showing that the disparities are statistically significant and not due to a fluke of chance.

For the income disparities in races, we also found strong evidence in both the EDA plots and the result from hypothesis testing. We can see that Asian and White people had greater income while the Black's income was less considerably. The difference was also shown in the dispersion of each quartile limit observed in the data. While the dispersion is large for Asians and Whites, the Black people's income seems more stable. We also noticed that the imbalance of

income is a phenomenon that appears in every income level across the three races, meaning it's more of an issue for the whole population instead of one certain level.

These are the results we gained from the two datasets. It is important to note that the results do not suggest or infer any causal relationship between these factors. It is not possible to tell why this is happening only from this limited dataset, and a more thorough answer would require more data and background research.

### **3.2 Income Inequality with Ages and Across Different Income Groups**

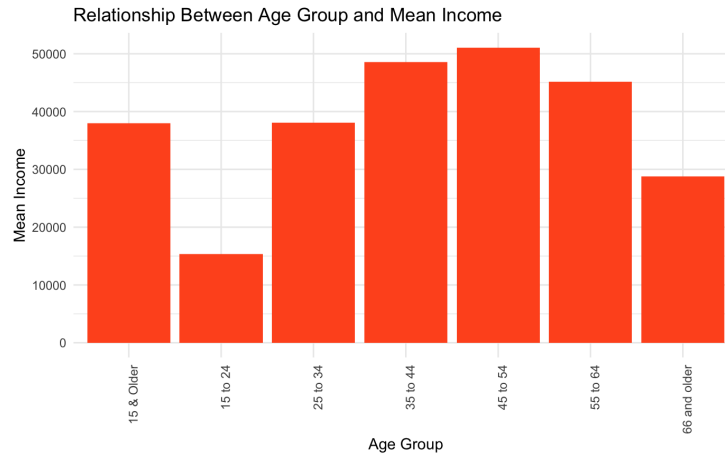
Age is another crucial factor for income inequality, and we are trying to understand the importance of each variable on the level of income inequality. Intuitively, we may think that older people should have more income because they are more capable than younger people since they have more experience in the same field and can, therefore, create more value than younger people. Thus, they deserve more income. The statistical test result should also show that there is a difference in income inequality between different age groups. However, the test shows a contradictory result, which could be more counter-intuitive.

We should have a preliminary understanding of the importance of the variables from the first dataset after we have done statistical analysis on them. Conducting statistical tests on the second dataset to gain further insight into the whole dataset would provide us with information to address our initial questions better.

The test conducted on the second dataset is the income inequality across different income groups using the Gini index factor. The Gini Index is a summary measure of income inequality. The Gini coefficient incorporates the detailed shared data into a single statistic, which summarizes the dispersion of income across the entire income distribution. The Gini coefficient ranges from 0, indicating perfect equality, where everyone receives an equal share, to 1, perfect inequality, where only one recipient or group of recipients receives all the income (Bureau, 2021). We should be able to conclude whether income inequality exists across different income groups or not after the statistical test.

#### **3.2.1 EDA**

There are six age groups from the dataset, which are “15 to 24”, “25 to 34”, “35 to 44”, “45 to 54”, “55 to 64”, and “66 and older”. There is another column named “15 & older” in the plot of the Relationship Between Age Group and Mean Income, which represents the mean income for all age groups in this dataset. The overall trend of the income for different age groups has a roughly normal distribution, with “45 to 54” being the peak point, beside the column of “15 & older”.



Graph 17: Relationship Between Age Group and Mean Income

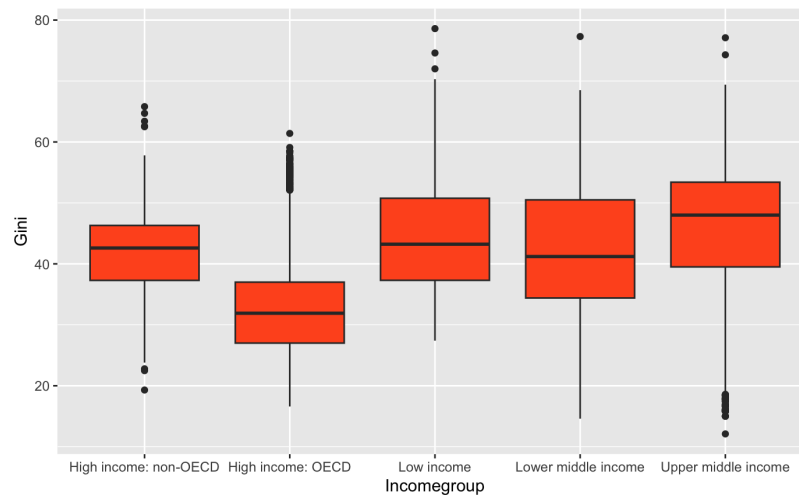
As we discussed in 3.3, we believe that older people are meant to earn more than younger people for the reason that their age also represents their level of expertise and skills in their field. Observation on the plot helps hold that assumption, except there is a declining trend after the peak column of “45 to 54”. A possible explanation may be that people start to retire somewhere between age 60 to 65, which is just happenly included in the last two columns, therefore, causing a decline in income earned for those two age groups. We should convert our assumption that middle-aged people should earn more than other age groups.

Another information gained from this plot is that there would be no aging population issue in a short period of time in the future. The overall mean income shows a value approaching 40,000 thousand, while there are three other age groups’ income values equal to or less than that. With the assumption that middle-aged groups earn more income than others and the mean overall income value still shows no sign of skewed distribution, we can conclude that the aging population issue will not be exacerbated in the short term.

There are two plots of Relationship Across Income Groups and Mean Income. Both plots show the statistical mean and quartile value for each group of “High income: non-OECD”, “High income: OECD”, “Low income”, “Lower middle income”, and “Upper middle income”

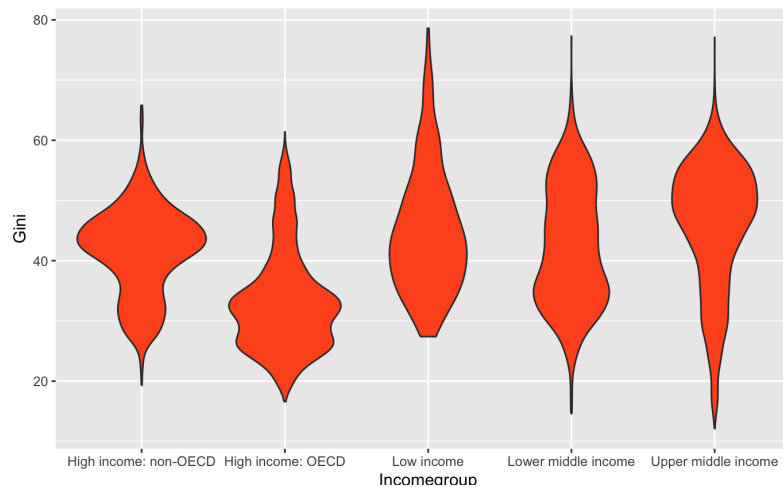
Observation on the box plot shows the trend that the “High income: OECD” group has the lowest Gini index, which represents that the income inequality in this group is the lowest among all the groups. We should also notice that the number of outliers falls beyond 1.5 quartiles, also the most for this group, which could be explained by the fact that there are also some records in this group having a relatively lower income than others in this group. Group “High income: non-OECD” has the narrowest quartile range concentrated around the Gini index of 40, which means that the income inequality within this group is the most balanced one among all five groups. However, it also has outliers on both sides of the 1.5 quartile. Comparably, the remaining three groups have a similar box range of the Gini index and mean Gini value. Therefore, the income inequality within those three groups should be approaching each other. With further observation of these three groups, we notice that there is only one outlier in the group “Lower middle income” that falls beyond the 1.5 quartile, which is an imitable value since it will have a limited impact on the overall analysis. Outliers in the group “Upper middle

income” are concentrated on the lower end of the quartile box, while there are outliers on the higher end of the quartile box. This group has the widest span over the Gini index, which implies that income inequality is the most unbalanced for this group.



Graph 18: Boxplot of Income Group over Gini Index

The violin group can provide more information about the density of the dataset, showing the probability density of the data at different values. Moreover, it can provide the internal distribution within each group, which can help gain a better insight into each one.



Graph 19: Violin Plot of Income Group over Gini Index

As we discussed before on the box plot, the group “High income: non-OECD” indeed has the most balanced Gini index within itself for the reason that it shows a slimmer distribution around the median, suggesting less variability. The groups “Upper middle income” and “Lower middle income” have the most variability and widest span among the five groups, which suggests they are not well-balanced on the income distribution. The “Low income” group has no Gini index lower than 25, which has the highest lower end in all five distributions, suggesting that most values in this group are concentrated on the higher end of the Gini index.



The overall observation on both box and violin plots is consistent, with the “High income: non-OECD” group being the most well-balanced one and the “Upper middle income” and “Lower middle income” being the most unbalanced groups.

### 3.2.2 Hypothesis Testing and Methods for Age Groups and Mean Income

The test conducted is to find out whether the age group is a factor that is going to impact the mean income.

$H_0$ : There is no difference in mean income across different age groups

$H_a$ : There is a difference in mean income across different age groups

We used the ANOVA table as the first method to find the statistical results. We used it to determine if there were any statistically significant differences between the means of groups. Anova method in R can be used after assigning the target we desire to compare; we would choose MeanIncome ~ ‘Age Group’ from data age\_income\_mean along with a default significance level of 0.05. The results from the ANOVA table did not show a success test since there is no p-value.

```
## Anova

```{R}
age_income_mean$`Age Group` <- as.factor(age_income_mean$`Age Group`)
anova_result <- aov(MeanIncome ~ `Age Group`, data = age_income_mean)
summary(anova_result)
```
```

|             | Df | Sum Sq    | Mean Sq   |
|-------------|----|-----------|-----------|
| `Age Group` | 6  | 931144825 | 155190804 |

Picture 20: ANOVA Table

Therefore, we need to use the Kruskal-Wallis test to determine the relationship between mean income and age groups. Similar to the process of the ANOVA table, we choose MeanIncome ~ ‘Age Group’ from data age\_income\_mean along with a default significance level of 0.05, then use these attributes in clause Kruskal.test() to get the result table for p-value equals to 0.42.

```
## Kruskal-Wallis

```{R}
kruskal_test <- kruskal.test(MeanIncome ~ `Age Group`, data = age_income_mean)
kruskal_test
```
```

```
Kruskal-Wallis rank sum test

data: MeanIncome by Age Group
Kruskal-Wallis chi-squared = 6, df = 6, p-value = 0.4232
```

Picture 21: Kruskal Test

We can conclude that the null hypothesis should not be rejected because the p-value is 0.42, which is much larger than the significance level of 0.05. As we discussed in 3.2, we

believed that there should be some income difference between different age groups since older people tend to make more income than younger people of their higher expertise and skill level. However, the test result shows a contradictory outcome, namely, that there is no significant difference in income inequality between different age groups. Further information will be discussed in 3.2.4.

### 3.2.3 Hypothesis Testing and Method Across Different Income Groups and Mean Income

The test conducted is to find out whether the age group is a factor that going to impact the mean income.

$H_0$ : *There is no significant difference in income inequality across different income groups.*

$H_a$ : *There is a significant difference in income inequality across different income groups.*

We used the ANOVA table as the first method to find the statistical results. We used it to determine if there were any statistically significant differences between the means of groups. We choose Gini~ 'IncomeGroup' from data gini\_income\_mean along with a default significance level of 0.05, then use these attributes in the clause above (). The result from the Anova table shows a p-value of less than 2e-16.

```
## ANOVA

```{R}
anova_result <- aov(Gini ~ Incomegroup, data = gini_income_data)
summary(anova_result)
```
```

|             | Df   | Sum Sq | Mean Sq | F value | Pr(>F)     |
|-------------|------|--------|---------|---------|------------|
| Incomegroup | 4    | 219950 | 54987   | 642.2   | <2e-16 *** |
| Residuals   | 7542 | 645736 | 86      |         |            |

---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Picture 22: ANOVA Table

We decided to use an extra test to prove this result further since we did not get useful information after the first method in 3.3.2. Then, we decided to use the Regression Analysis. We choose Gini~ 'IncomeGroup' from data gini\_income\_mean along with a default significance level of 0.05, then use these attributes in clause lm() to get the result table for a p-value less than 2.2e-16.

```
## Regression
```

```
```{R}
lm_model <- lm(Gini ~ Incomegroup, data = gini_income_data)
summary(lm_model)
```
```

```
Call:
lm(formula = Gini ~ Incomegroup, data = gini_income_data)

Residuals:
    Min       1Q   Median       3Q      Max
-33.626  -6.605  -0.010   6.105  35.005

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.5501     0.3332  124.685 < 2e-16 ***
IncomegroupHigh income: OECD  -8.3444     0.3756  -22.214 < 2e-16 ***
IncomegroupLow income           3.2870     0.7048   4.664 3.16e-06 ***
IncomegroupLower middle income  0.7448     0.4123   1.807  0.0709 .
IncomegroupUpper middle income  4.1757     0.3861  10.815 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.253 on 7542 degrees of freedom
Multiple R-squared:  0.2541,    Adjusted R-squared:  0.2537
F-statistic: 642.2 on 4 and 7542 DF,  p-value: < 2.2e-16
```

Picture 23: Regression

Therefore, we concluded that we should reject the null hypothesis because the p-value is  $2.2e-16$ , which is much smaller than the significance level of 0.05. There is a significant difference in income inequality across different income groups.

### 3.2.4 Results

We will reject the null hypothesis for the income inequality for age groups but will not reject the null hypothesis for the income inequality across different income groups. The age groups intuitively have income inequality, for older people would make more money than younger people. However, the test shows no income inequality among them, which may be why relatively small age data or the test methods need to be improved and modified. Income inequality across different income groups still exists even though they have been divided into groups according to their income level.

## 3.3 Income Inequality with Regions

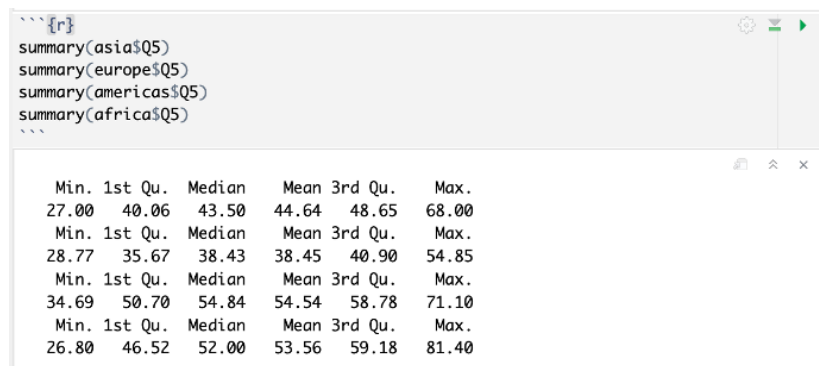
Regions are also among the most important factors that may impact the population's income level and equality. In dataset 3, the data was related to the income level in different regions and countries worldwide. There are some main countries from the four different regions: Asia, America, Africa, and Europe. It also has some useful information like the income level, the income shares based on different quintiles of people, the mean income, the currency, the year range, and so on. The question we wanted to address was whether income inequality existed in different regions. Since there are too many missing variables in the mean and median income,

our group decided not to use this variable. Thus, we decided to use the income share of different quintiles of the population in different regions to reflect the income equality in those areas.

### 3.3.1 EDA

We subgrouped the income share percentages into five quintiles of the population (Q1, Q2, Q3, Q4, Q5) based on the four regions (Asia, America, Africa, and Europe). Since we wanted to talk about income inequality, our focus was on the Q1 and Q5 population because each represents how much income shares the 20% of the population with the lowest income and highest income in those regions occupied. Based on those simple groupings, we did some exploratory data analysis as follows.

Firstly, we did the summary analysis on the quintile five population in different regions by using the summary function in the R. From the summary table below, we can easily find out that America has the highest mean, which is 54.54, which can be interpreted as the top 20% population with highest income occupying about 54.54% of total income shares. From this significant number, we can first speculated that the income in the American region was highly unequally distributed because 20% of the population occupied about half of the income shares. Africa's range is the largest, ranging from the minimum value of 26.80% to 81.5%, which showed a very high variance. Comparably, Asia and Europe have lower mean than Africa and America, about 44.64% and 38.45% of the total income shares. After all, the Americas had both the highest median and mean, which suggested that the income share of America was more significant than that of any other country. The overall mean was larger than the median, suggesting that the distribution would be right-skewed.



```
summary(asia$Q5)
summary(europe$Q5)
summary(america$Q5)
summary(africa$Q5)
```

| Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max.  |
|-------|---------|--------|-------|---------|-------|
| 27.00 | 40.06   | 43.50  | 44.64 | 48.65   | 68.00 |
| Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max.  |
| 28.77 | 35.67   | 38.43  | 38.45 | 40.90   | 54.85 |
| Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max.  |
| 34.69 | 50.70   | 54.84  | 54.54 | 58.78   | 71.10 |
| Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max.  |
| 26.80 | 46.52   | 52.00  | 53.56 | 59.18   | 81.40 |

Picture 24: Summary Table for Q5

We did another summary analysis on the quintile one population in different regions by using the summary function in the R. From the summary table below; we can see that the situation was opposite to the Q5 summary table. For the countries whose upper quintile of population occupied more income shares, the lower quintile of population occupied less income shares. We can easily find out that America has the highest mean in the Q5 population but has the lowest mean, 3.956, which can be interpreted as the bottom 20% of the population with the lowest income occupying only about 3.956% of total income shares. From this significantly small number, we can also speculate that the income in the American region is highly unequally

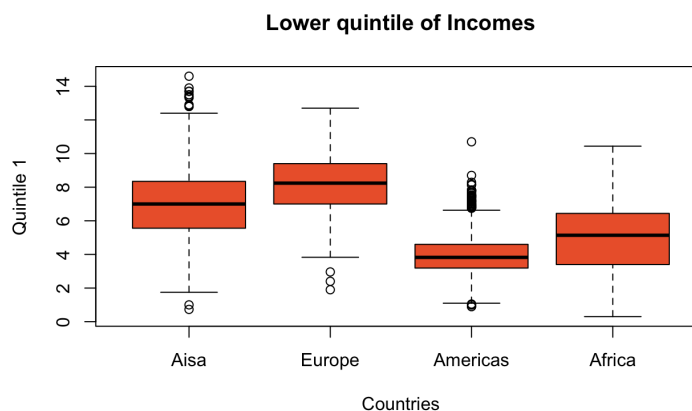
distributed because 20% of the population occupied less than 5% of the income shares. Asia's ranges are the largest, ranging from a minimum value of 0.73% to 14.6%, and Africa also has a similarly extensive range, from 0.3% to 10.44%. Comparably, Asia and Europe have higher mean incomes than Africa and America, which are about 7% and 8.24% of the total income shares, respectively, which also indicates that Asia and Europe may have more income equality than America and Africa. After all, the Americas have both the highest median and mean, which suggests that the income share of America is larger than that of any other country. The overall mean is less than the median, so this also suggests that the distribution will be left-skewed.

```
{r}
summary(asia$Q1)
summary(europe$Q1)
summary(america$Q1)
summary(africa$Q1)
`
```

| Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max.   |
|-------|---------|--------|-------|---------|--------|
| 0.730 | 5.560   | 7.000  | 6.926 | 8.345   | 14.600 |
| Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max.   |
| 1.900 | 7.000   | 8.240  | 8.215 | 9.400   | 12.700 |
| Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max.   |
| 0.900 | 3.190   | 3.825  | 3.956 | 4.600   | 10.700 |
| Min.  | 1st Qu. | Median | Mean  | 3rd Qu. | Max.   |
| 0.30  | 3.40    | 5.14   | 4.95  | 6.44    | 10.44  |

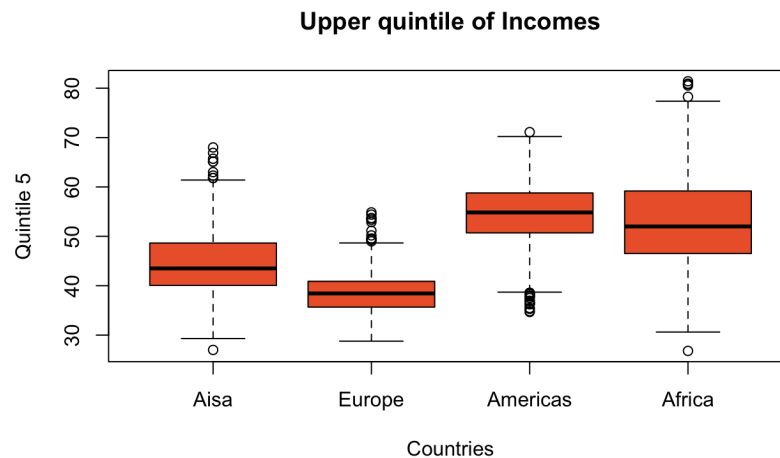
Picture 25: Summary Table for Q1

Next, we made two box plots related to the income shares of the Q1 and Q5 populations with different regions. As shown in the first box plot, which is the lower quintile population's income shares percentage, we can see the American population has the lowest median percentage of income shares, around 4. This means that 20% of the lower-income population occupies only about 4% of total income. The upper and bottom quartiles are approximately 3 and 5, respectively. There is one outlier that is close to 1, and the range goes from about 2 to 6. This shows that a lower number differs significantly from the rest of the data, even though the central tendency is larger than in Asia and Europe. Another practical observation is that the existence of outliers in the plots for Asia and Africa indicates the presence of extreme values, which may result from anomalies, excessive wealth, or poverty in certain regions.



Graph 26: Boxplot for Lower Quintile of Income and Different Countries

The second box plot plots the upper quintile population's income share percentage. From the plot, we can see the American population has the highest median percentage of income shares, around 55. This means that the 20% of the population with upper income occupies more than 50% of total income, which is a significant number. The income distribution's mild symmetry around the median suggests a consistent distribution of upper-income levels. Comparing Asia and Africa to Europe and the Americas, more variation can be seen in top income levels.



Graph 27: Boxplot for Upper Quintile of Income and Different Countries

Based on our initial observation, since America has the most significant mean of income share for the upper quintile of the population and the lowest mean of income share for the bottom quintile of the population, we can first get a hypothesis that America is more income unequal than the other three regions. Also, since America is represented as a developed region, it seems that developed regions are more income unequal than the developing regions. To prove this hypothesis, we will run a hypothesis test next to provide more statistical evidence.

### 3.3.2 Hypothesis Testing and Methods

Based on the dataset, we picked two regions to represent the developed region, America, and the developing region, Asia, and made the null hypothesis and alternative hypothesis as follows:

$$\mu_{\text{america}} = \text{America's upper quintile population of income shares}$$

$$\mu_{\text{asia}} = \text{Asia's upper quintile population of income shares}$$

$$H_0: \mu_{\text{america}} = \mu_{\text{asia}}$$

$$H_a: \mu_{\text{america}} > \mu_{\text{asia}}$$

First, we used the T-test to identify the hypothesis at a 95% significance level. We used the `t.test()` function in R to run the Welch Two Sample Test, and the code and result are shown below. Since we only want to know whether America's upper quintile population of income shares is more significant than Asia's upper quintile population of income shares, we set the

"alt=greater" to run the one-tail test. From the result, we can have the p-value less than  $2.2e-16 < 0.05$ , which indicates that the observed difference in means is highly statistically significant. This significant small p-value also suggests that we should reject the null hypothesis. So, we can conclude that at a 95% significance level, the mean of "americas\_q5" is significantly larger than the mean of "asia\_q5."

We used the bootstrap method to double-check the result; the code and result are shown below. We iterated the sampling process 10000 times, calculated the bootstrap mean, and did the 95% confidence interval test. From this result, the variance of the bootstrap means is 0.000378, which is a significantly small number, and it suggests a relatively stable estimate. We should reject the null hypothesis since 1 is not in the 95% bootstrap confidence bound. Thus, we have strong evidence that the mean of "americas\_q5" is significantly larger than the mean of "asia\_q5."

|   |   |
|---|---|
| <pre># t test ```{r} asia_q5= subset(asia, select=Q5, drop=T) americas_q5= subset(americas, select=Q5, drop=T) t.test(americas_q5,asia_q5, alt="greater",conf.level=.95) ```</pre> <pre>Welch Two Sample t-test  data:  americas_q5 and asia_q5 t = 41.923, df = 1798.1, p-value &lt; 2.2e-16 alternative hypothesis: true difference in means is greater than 0 95 percent confidence interval:  9.512071      Inf sample estimates: mean of x mean of y  54.54113  44.64041</pre> | <pre># bootstrap ```{r} # Perform a bootstrap test for ratio of means of Q5 set.seed(1000) N &lt;- 10000 ratio_mean_boot &lt;- numeric(N)  for (i in 1:N) {   americas_sample &lt;- sample(americas_q5, length(americas_q5), replace = TRUE)   asia_sample &lt;- sample(asia_q5, length(asia_q5), replace = TRUE)   ratio_mean_boot[i] &lt;- mean(americas_sample) / mean(asia_sample) }  boot_mean&lt;- mean(ratio_mean_boot) cat("Bootstrap Mean:",boot_mean, "\n") # one tail confidence_interval&lt;- quantile(ratio_mean_boot, 0.05) cat("Bootstrap 95% Confidence Interval for the Mean:", confidence_interval, "\n") bootstrap_variance = var(ratio_mean_boot) cat("Bootstrap variance for the Mean:",bootstrap_variance, "\n") ```</pre> <pre>Bootstrap Mean: 1.221895 Bootstrap 95% Confidence Interval for the Mean: 1.211758 Bootstrap variance for the Mean: 3.736707e-05</pre> |
|---|---|

Picture 28: t-test and Bootstrap

### 3.3.3 Results

From both results from the T-test and Bootstrap, they suggested that we should reject the null hypothesis. This means that at a 95% significance level, we can conclude that America's upper quintile population of income shares is more significant than Asia's upper quintile population of income shares. The result also shows that the top 20% of the population in America earn more income in their own regions than the top 20% of the upper-income population in Asia. As each is represented as a developed region and a developing region, we can also conclude that the upper-income population in developed regions controls comparably more money than the population in developing regions. So, there is more income inequality in developed regions, like America, than in developing regions, like Asia.

## 4. Conclusion

In our project, we employed a range of sophisticated statistical techniques to comprehensively analyze income inequality, focusing specifically on how it correlates with

factors like age, gender, race, and region. We first created visual representations of income distribution across each of the dimensions, providing a clear and intuitive understanding of how income varies across different groups. For each factor, we formulated specific hypotheses aimed at exploring their relationship with income inequality. Then, we employed more advanced statistical methods such as Bootstrap and finally drew a conclusion from our test results.

The following are the answers we get for our questions from our analysis:

1. What is the global trend about income inequality?  
Yes, there is income inequality around the world related to all races, populations, and groups.
2. What are the factors that may be related to income inequality?  
Age, gender, race, region, extent of development, global organizations belonging, etc.
3. Has the extent of income inequality worsened over time?  
Yes, it is getting worse, particularly in the last few decades.
4. Is there a relationship between gender and income level?  
Yes, males have greater income in terms of mean and median, with a greater dispersion.
5. Is there a relationship between race and income level?  
Yes, and the difference occurs across all income levels.
6. Is there a relationship between age and income level?  
Intuitively yes, the test result shows that there is no significant difference between them, but it might be due to the small sample size. The visualization of income distribution among different age groups indicates the existence of inequality.
7. Are there differences in income inequality across different income groups?  
Yes, we still have income inequality even within the same income group.
8. Is there a relationship between the region where the person is located and income level?  
Yes, people who live in more developed areas have higher incomes.
9. Is there a trend that developed countries have an overall higher average income than developing countries?  
Yes, people in developed countries, like America, have an overall higher income, compared with those in developing countries.
10. Do wealthy people from developed regions have more income inequality than developing regions?  
Yes, in developed regions, wealthy people have a much higher income share than those in developing regions, the extent of income inequality in developed countries is higher than that in developing countries.



Analyzing the possible influencing factors of income inequality carries substantial social importance. Our research results highlight the statistically significant relationship between income inequality and age, gender, race, and extent of regional development, which helps us realize that income disparity is not just a superficial issue, but deeply rooted in some structural and systemic inequalities, which may be caused by some long-term historical reasons. By providing a more insider view of population data, our research may offer valuable insights to assist governments and organizations in formulating more targeted policies and promoting a more equitable distribution of wealth and resources in society. More data visualizations on income inequality also help introduce the current issue to the public directly, which will encourage more public engagement and further discussion.

## 5. References

- Bureau, U. C. (2021, October 8). *Gini Index*. Census.gov.  
<https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/gini-index.html>
- Frost, J. (2023, February 1). *Introduction to bootstrapping in statistics with an example*. Statistics By Jim. <https://statisticsbyjim.com/hypothesis-testing/bootstrapping/>
- Kakwani, N. C. (1980). *Income inequality and poverty*. New York: World Bank.
- World Inequality Lab (2022). *World Inequality Report 2022*. WID.world.  
<https://wir2022.wid.world/>