

Income Inequality Around the World

Group 12:

Rongmu Sun, Xue Qin,

Kaiyang Li, Tianze Yang



Part I: Introduction



<https://www.pngsucai.com/>

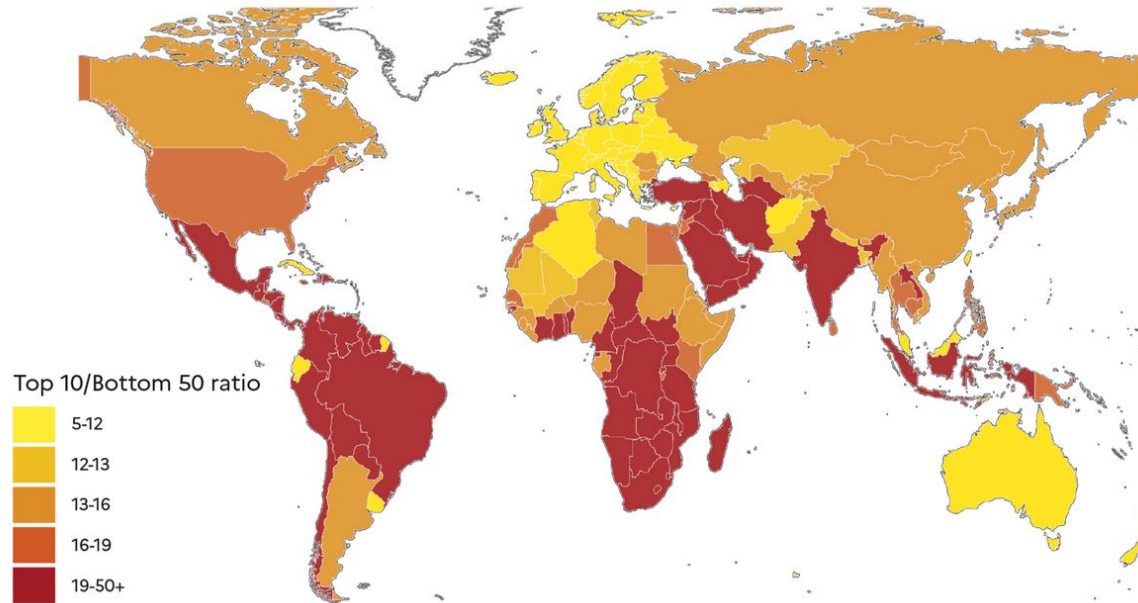
Introduction to Income Inequality

- Inequality is all about the distribution of power and resources, the rights people can exercise, and the opportunities they can access.
- The world is vastly unequal, extreme wealth coexists with extreme poverty
- The poorest 50% of people own just 2% of wealth



Global Trends

- Inequality of income within countries as measured by the ratio of the income of the richest 10% to the poorest 50%, 2021






Data Science Questions


1. What is the global trend about income inequality?
2. What are the factors that may related to the 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 located and income level?
9. Is there a trend that developed country have overall higher average income than developing country?
10. Does the wealthy people from developed region has more income inequality than developing region?

Introduction to Raw Datasets

- Historical income inequality of the world as well as data from USA census of government.
- <https://data.world/m00nlight/incomeinequality>

 **USA-median-mean-income-by-age.xlsx** clean data <https://www.census.gov/data/tables/time-series/demo/inco...> View Download Filter Grid

	 year	 age_group	# number_with_income_male_thousnads	# current_dollars_male_median	# 2017_d
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

- Dataset 1: The mean and median income by age group and gender group

Introduction to Raw Datasets

- Data on the income distribution of different races within the U.S.
- <https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-inequality.html>

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

Introduction to Raw Datasets

- Historical income inequality of the world as well as data from USA census of government.
- <https://data.world/m00nlight/incomeinequality>

 **WIID3.4_19JAN2017New.xlsx**
raw data <https://www.wider.unu.edu/project/wiid-world-income-inequ...>View Download Share Grid

	 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

- Dataset 2: The Income group and income quintile by country and region

Data Clean

- Remove duplicate columns

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

- Add sum of Decile and Quintile

```
{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")])
```

Data Clean

- Assign NA to all blank space

```
{r}
df <- lapply(df, function(x) ifelse(x == "", NA, x))
df <- as.data.frame(df)
```

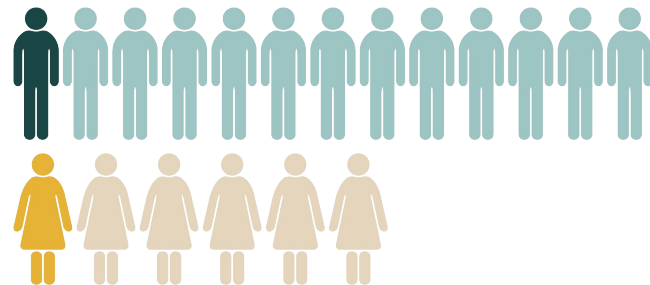
- Assign missing value for all categorical NAs

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

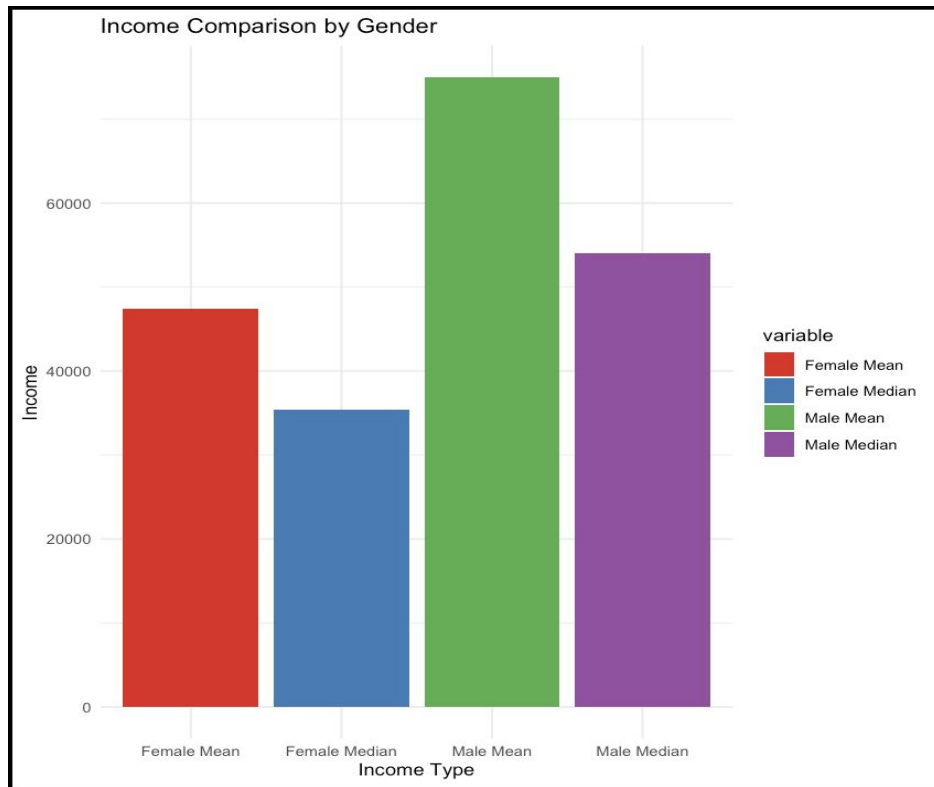
Part II:

Data Analysis : Income Inequality With Gender and Race

EDA on Gender



<https://www.pngsuc.ai.com/>



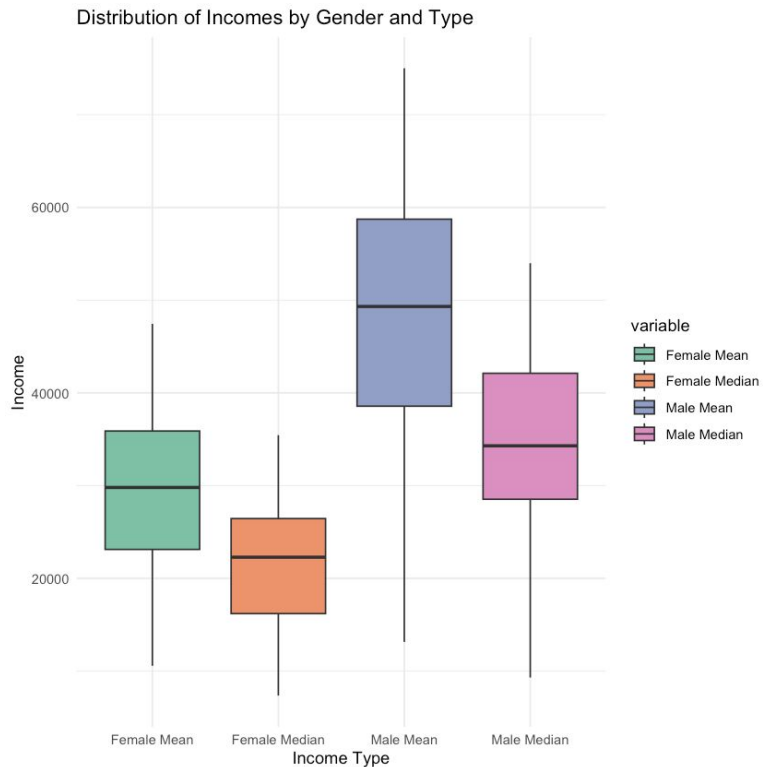
A comparison between the mean and median of the income in male and female.

It can be observed that both the mean and median income has a difference in gender.

EDA on Gender



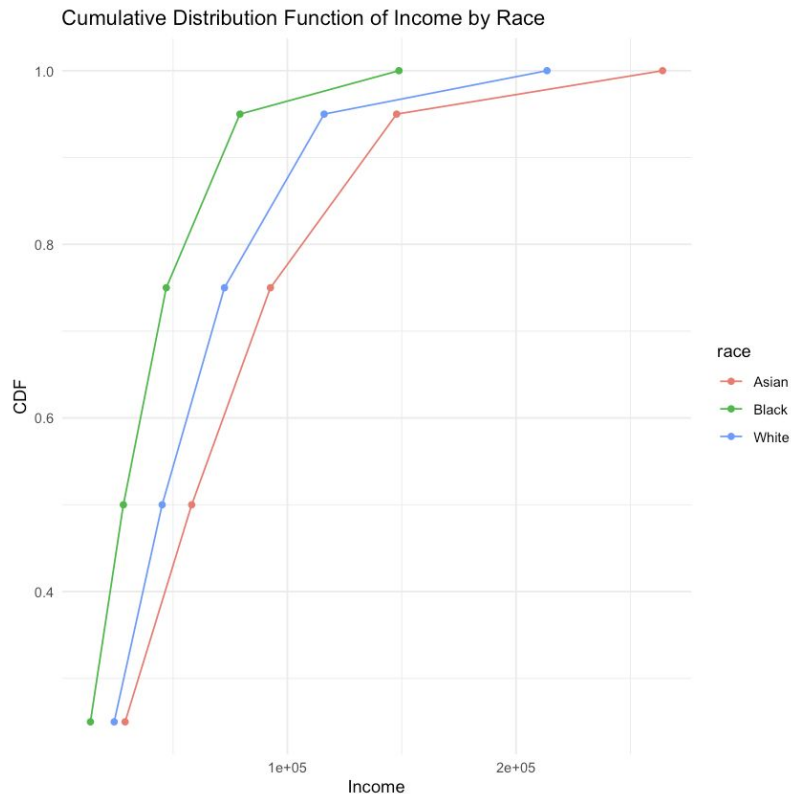
<https://www.pngsucai.com/>



A comparison between the distribution of mean and median income over male and female.

It can be observed that the distribution of income is different between male and female.

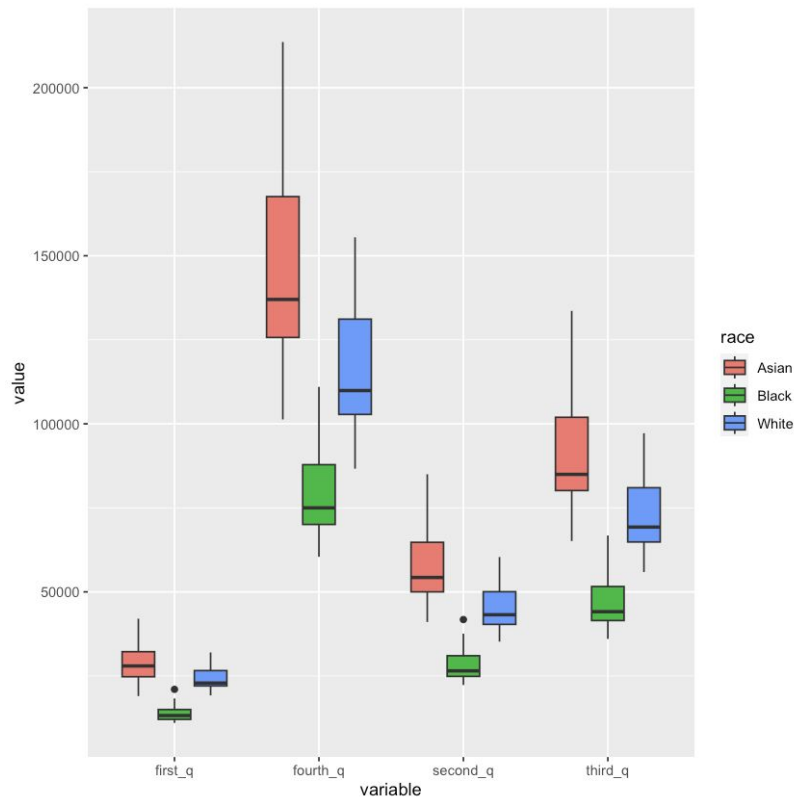
EDA for races issue



A cdf plot of the income for different races in the U.S.

The curves are clearly does not overlapping to each other, indicating not same distribution in income for different races.

EDA for race issue



A comparison between income for every quartile in different races.

It is clear that the difference of income between races is prevail in every quartile instead of for only one or two income group.

Method - Bootstrap

Hypothesis Test:

H0: There is no significant difference between the median of income of male and female in the U.S.

Ha: There is a difference in the median of income of male and female in the U.S.

```
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), repl
}
```

```
diff <- bootstrap_median_m - bootstrap_median_f
```

```
quantile(diff, c(0.025, 0.975))
```

- According to the bootstrap test result, 0 is not in the confidence interval, so the null hypothesis is rejected.
- It can also be seen that within the confidence interval the difference between the two median is quite big.

Method - T-Test

Hypothesis Test:

H0: There is no significant difference between the mean of income of male and female in the U.S.

Ha: There is a difference in the mean of income of male and female in the U.S.

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



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

- According to the t-test on the mean of both male and female income, the p-value is smaller than 0.05 so the null hypothesis is rejected.

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
```

Method - Bootstrap

Hypothesis Test:

H0: There is no significant difference in the second quartile of income for different races in the U.S.

Ha: There is a difference in the second quartile of income for different races in the U.S.

```
# 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

- 0 is not within the 95% confidence interval of the difference between each two races.
- We can reject the null hypothesis and determine there is a difference between the upper limit of second quartile of income in different races.

Method - Permutation test

Hypothesis Test:

H0: There is no significant difference in the lower limit of top 5% of income for different races in the U.S.

Ha: There is a difference in the lower limit in top 5% of income for different races in the U.S.

Permutation test

```
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
```

- We can see the p-value for the permutation test is less than 0.05 (actually 0)
- We can reject the null hypothesis, and determine the lower limits of top 5% income and race are not independent.

Results

1. According to the test above, we can conclude that the difference between male and female's income can be both observed in the mean and median of the value
2. According to the EDA boxplot, we can also observe that male income could be more diverse compared to female's.
3. According to the test and plots, male's income are generally higher than female.
4. The reason behind this would needed more data and social background investigation to determinate.

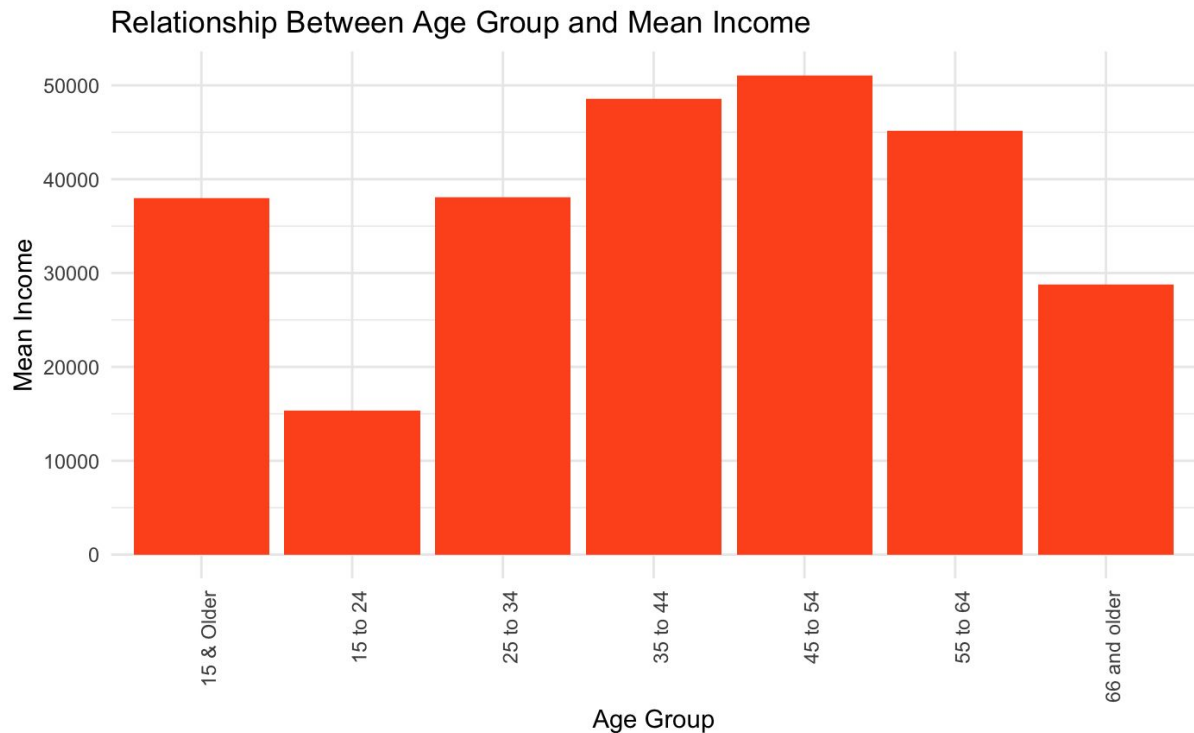
Results

1. According to the test above, we can conclude that the difference between races can be observed in every quartile of income.
2. According to the EDA boxplot, we can also observe that the diversity of income in different races also has difference.
3. The race has lower income in one quartile tends to be low in every quartile, indicating this issue is not for a certain income level
4. The reason behind this would needed more data and social background investigation to determinate.

Part III:

Data Analysis: Income Inequality With Ages and Across Different Income Groups

Income Inequality With Ages - EDA



Income Inequality With Ages - Methods

Hypothesis Test:

H0: There is no difference in mean income across different age groups

Ha: There is a difference in mean income across different age groups.

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 |

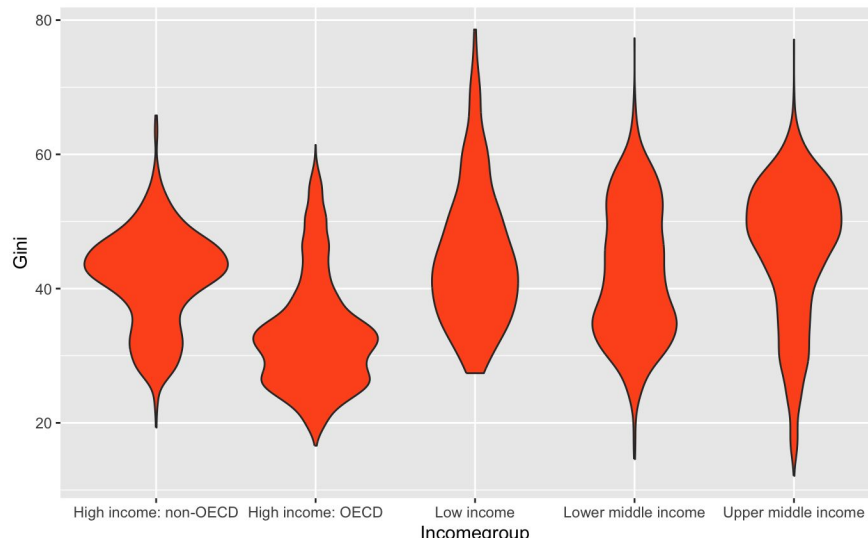
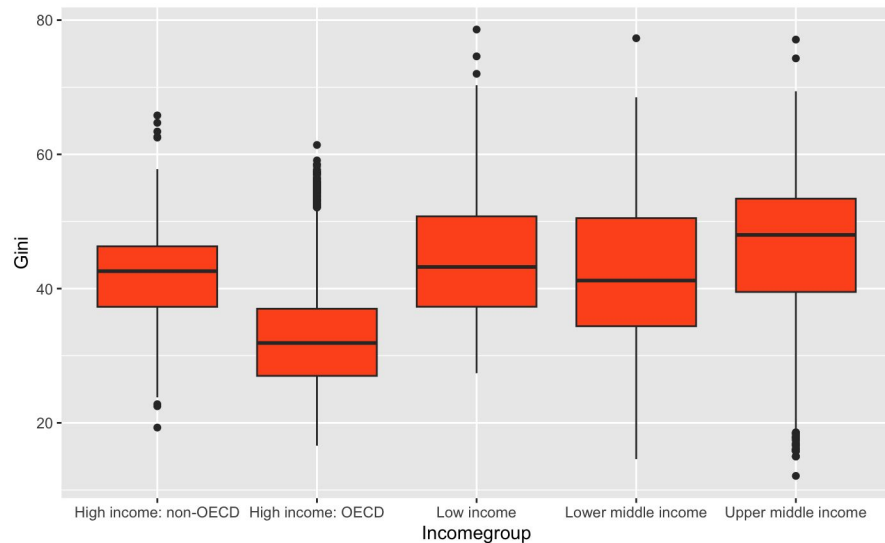
Kruskal-Wallis

```
```{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

Income Inequality With Income Group - EDA



Income Inequality With Income Group

Hypothesis Test:

H0: There is no significant differences in income inequality across different income groups.

Ha: There is a significant differences in income inequality across different income groups.

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

Income Inequality With Income Group

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

We can see that both test get the p-value less than 0.5.

Therefore, we can reject the null hypothesis and conclude that there is significant differences in income inequality across different income group.

Results

- We will reject the null hypothesis for the income inequality for age groups
- We will not reject the null hypothesis for the income inequality across different income groups
- The age groups are intuitively have income inequality for older age would make more money than younger age. However, the test shows that there are no income inequality among them which maybe the reason that a rather small age data or the test methods needs to be improved and modified.
- The income inequality across different income groups still exists even though they have been divided into groups according to their income level.
- Overall, the high-income OECD group exhibiting lower levels of inequality compared to other groups

Part IV:

Data Analysis: Income Inequality With Regions

```

```{r}
summary(asia$Q5)
summary(europe$Q5)
summary(america$Q5)
summary(africa$Q5)
```

```

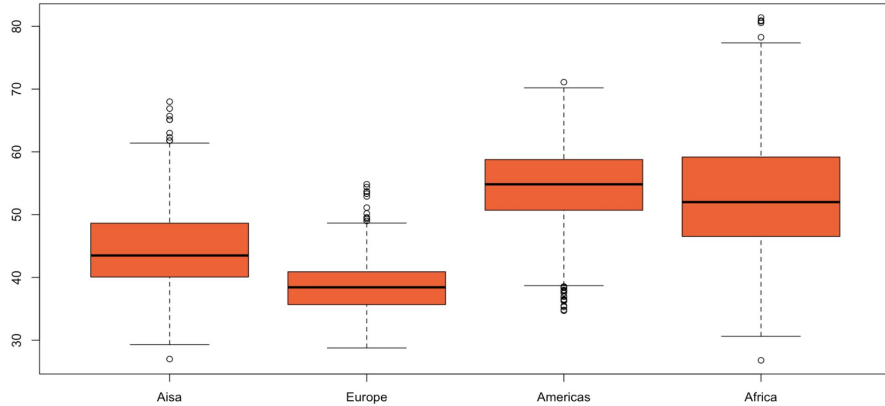
| | | | | | |
|-------|---------|--------|-------|---------|-------|
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 27.00 | 40.06 | 43.50 | 44.64 | 48.65 | 68.00 |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 28.77 | 35.67 | 38.43 | 38.45 | 40.90 | 54.85 |
| Min. | 1st Qu. | Median | 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 |

<https://www.pngsucai.com/>

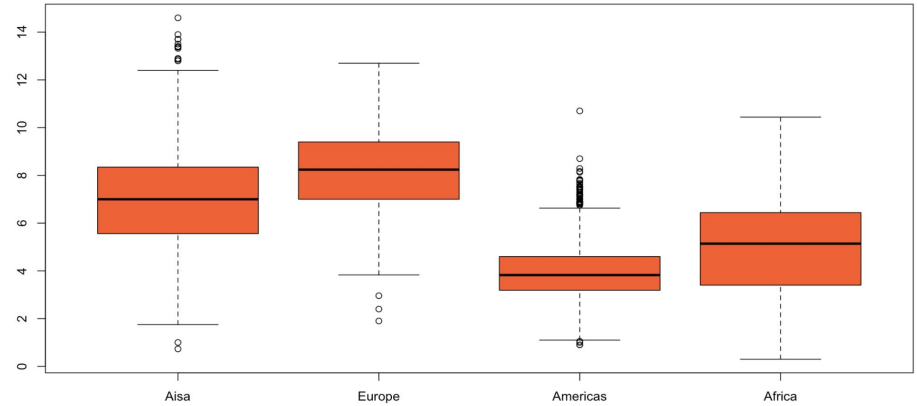


- Q5 means how much income shares the 20% of the population with the top income in the region occupy.
- America has the highest mean, which is 52.54. So, on average, 20% of people with the highest income occupied about 52.54% of total money in America.
- Africa has the maximum number of income shares.

Upper quintile of Incomes



Lower quintile of Incomes



- Wealthy people in America and Africa occupy more income share than in Asia and Europe.
- Africa's range is larger than the other 3 regions, meaning there is more income inequality in Africa.

- Poor people in America and Africa occupy less income share than in Asia and Europe
- Asia's range is larger than the other 3 regions, which means poor people in Asia may occupy more in total income shares.

Method - T-test

Hypothesis Test:

H0: America's upper quintile population of income shares = Asia's upper quintile population of income shares

Ha: America's upper quintile population of income shares > Asia's upper quintile population of income shares

```
# 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)
```
```

Welch Two Sample t-test

```
data: americas_q5 and asia_q5
t = 41.923, df = 1798.1, p-value < 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
```

- P value: $2.2e-16 < 0.05$
- the Welch Two Sample t-test results suggest strong evidence that the mean of "americas_q5" is significantly > the mean of "asia_q5."
- At 95% significance level, reject the null hypothesis

Method - Bootstrap

Hypothesis Test:

H0: America's upper quintile population of income shares = Asia's upper quintile population of income shares

Ha: America's upper quintile population of income shares > Asia's upper quintile population of income shares

```
```{r}
Perform a bootstrap test for ratio of means of Q5
set.seed(1000)
N <- 10000
ratio_mean_boot <- numeric(N)

for (i in 1:N)
{
 americas_sample <- sample(americas_q5, length(americas_q5), replace = TRUE)
 asia_sample <- sample(asia_q5, length(asia_q5), replace = TRUE)
 ratio_mean_boot[i] <- mean(americas_sample) / mean(asia_sample)
}
boot_mean <- mean(ratio_mean_boot)
cat("Bootstrap Mean:", boot_mean, "\n")
one tail
confidence_interval <- 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")
```
```

```
Bootstrap Mean: 1.221895
Bootstrap 95% Confidence Interval for the Mean: 1.211758
Bootstrap variance for the Mean: 3.736707e-05
```

- Variance is a small value like 3.736707e-05 suggests a relatively stable estimate.
- 1 is not in the 95% bootstrap confidence bound
- results suggest strong evidence that the mean of "americas_q5" is significantly > the mean of "asia_q5."
- At 95% significance level, reject the null hypothesis

Results

- T-test: Reject Null
- Bootstrap: Reject Null
- Based on those two methods, we reject the null hypothesis.
- we can conclude that America's upper quintile population of income shares is larger than Asia's upper quintile population of income shares.
- Which means, the wealthy people in America control comparably more money than wealthy people in Asia in their own regions.
- There is more income inequality in America region than Asia region.

Part IV: Conclusion



<https://www.pngsucai.com/>

Answers to Questions

1. What is the global trend about income inequality?

Yes, there's income inequality around the world related all races, population, and groups.

2. What are the factors that may related to the income inequality?

Age, gender, race, region, country, years.....

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, male has 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 the income level.

Answers to Questions

6. Is there a relationship between age and income level?

Intuitively yes, but test result shows that there is no significant difference among them,

7. Are there differences in income inequality across different income groups?

Yes, we still got income inequality even within the same income group.

8. Is there a relationship between the region where the person located and income level?

Yes, people live in more developed areas have higher income.

9. Is there a trend that developed country have overall higher average income than developing country?

Yes, people in developed region, like America has higher income share for wealthy people.

10. Does the wealthy people from developed region has more income inequality than developing region?

Yes, in developed region, wealthy people occupied much higher income share than developing region, so developed countries are more income unequal.

Q & A?

References

- <https://devinit.org/resources/inequality-global-trends/#:~:text=The%20poorest%2050%25%20of%20the,thirds%20of%20global%20income%20inequality.>
- <https://data.world/m00nlight/incomeinequality>
- <https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-inequality.html>