UN SDGs of Asia PROJECT

By Ahmad Ibrahim Waqas

Motivation and Background

-In an era where global interconnectedness has become more evident than ever, the study of regional development patterns takes on a new urgency. The Asian continent, famous for its rapid economic growth, huge population, demographic changes, and unique social challenges, is always a heated discussed topic. Our research is driven by the need to dissect the complex layers of development within Asia and to offer a comprehensive perspective of it.

Data Set Brief Introduction

-Various organization all around measures the development level of a country with different Index, we gonna briefly introduce those three data sets that we included in our project.

- **Sustainable Development Goal Index is** A composite Index set by the United Nations that evaluate the progress of countries towards achieving the 17 development goals including economy, political and environmental perspective and so on. The higher this index is, the better the country is performing.

- ***The Foundation for Peace's Fragile States Index (FSI) is*** an annual ranking that assesses the vulnerability of countries to conflict or collapse, measuring the stresses they face based on social, economic and political perspectives. Basically the higher this index is, the more fragile the country is.

- **Statistical tables from UNICEF's 2023 State of the World's Children report** provide comprehensive data on the well-being of children globally, including series of perspective such as health, education, economic situation and related birth rate and so on

Guiding Question

- Being as the continent with the largest number of developing countries, how exactly is Asia developing? Is it lagging behind or ahead of other places and What should the policy makers know

Introduction

➔ **Research Question 1**

Is it possible to group countries in Asia according to their sustainable development level? Does the cluster result make sense when combined with other data (e.g. FSI Score).

➔ **Research Question 2**

What is the difference between the rate of the share of household income from Asian countries compared to all Western countries' rates?

➔ **Research Question 3**

How does infant mortality rate affect population growth among asian countries? Is there a trend showing they are proportional?

Introduction

This project delves into Asia's developmental landscape and is guided by three research questions, each offering a different perspective through which to understand the continent's trajectory to achieving its Social Development Goals(SDGs) as per data provided by United Nations.

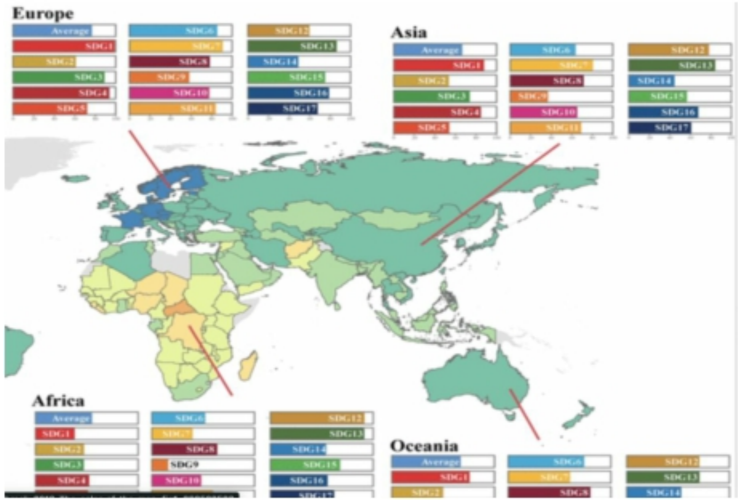
- Firstly, From bustling urban centers to remote rural communities, Asia's diverse tapestry reveals a spectrum of realities shaped by historical legacies, socio-economic structures, and policy frameworks. It will be really meaningful if we delves into the heart of sustainable development in Asia, seeking to classify countries according to their achievements in this area(SDG Score Index), and try to group them so that we can benefit the policy makers by showing them the general pattern.

- Secondly, we examine the contrasting rates of household income share between Asian nations and their Western counterparts. This comparison unveils insights into economic disparities, shedding light on the dynamics of income distribution, economic development, and the pursuit of inclusive growth.

- In our third research question, we navigate the intricacies of population growth and infant mortality rate across Asia. Here, the interplay between newborn healthcare and population dynamics emerges as a pivotal nexus, influencing demographic transitions and well-being.

Research Question 1

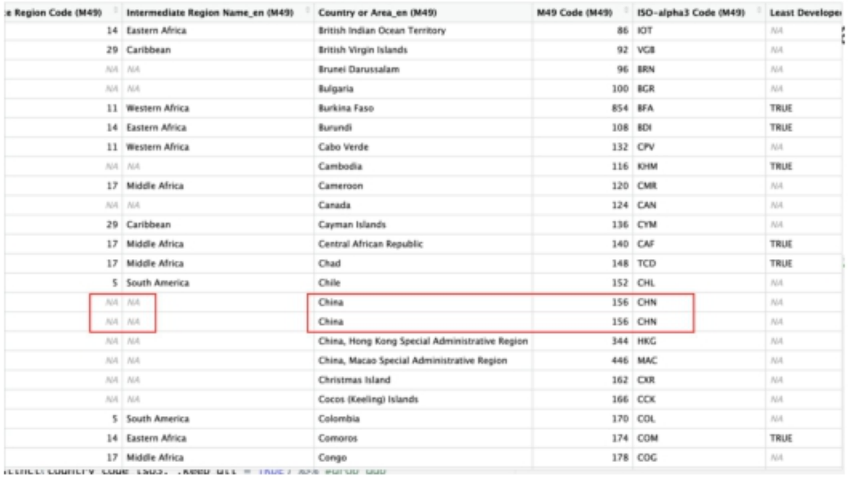
- Is it possible to group countries in Asia according to their sustainable development level? Does the cluster result make sense when combined with other data (e.g. FSI Score)



— Data Cleaning

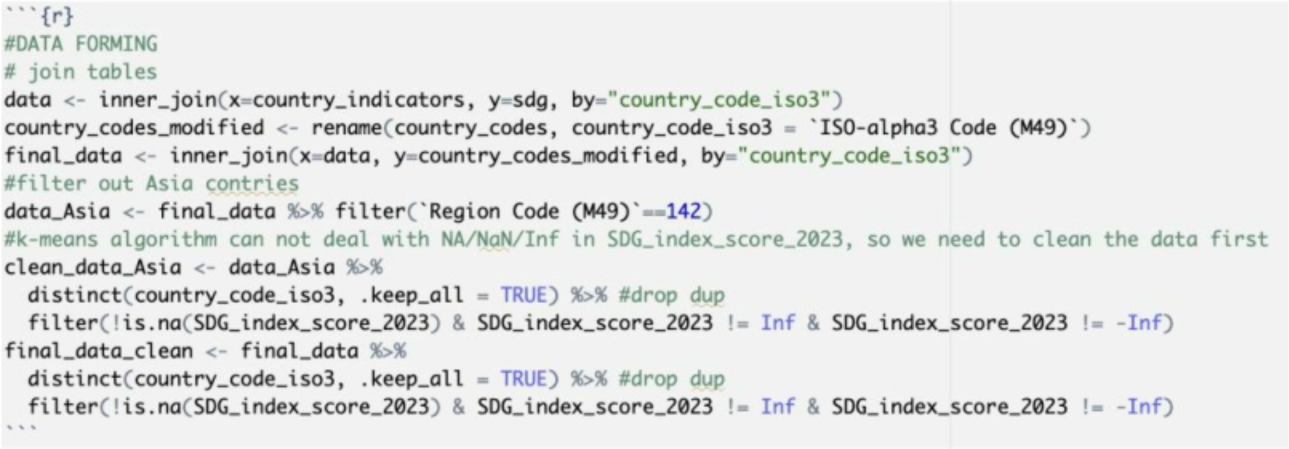
-As we are only studying about Asian countries so we filter out Asian countries first.(By region code here)

-Then we observed some duplicated data in our data set, we decide to drop them for better model accuracy



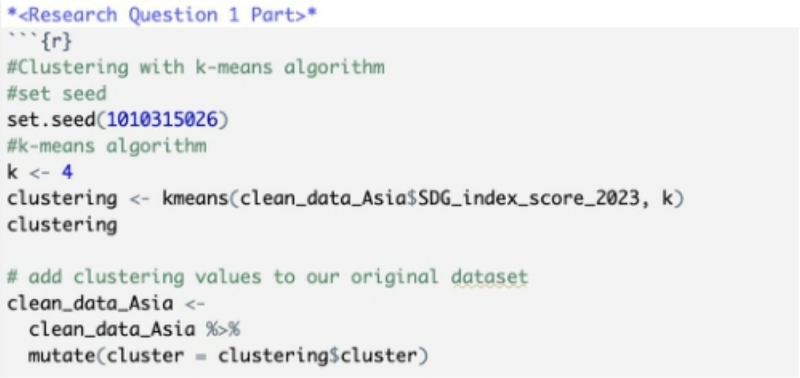
e Region Code (M49)	Intermediate Region Name_en (M49)	Country or Area_en (M49)	M49 Code (M49)	ISO-alpha3 Code (M49)	Least Develope
14	Eastern Africa	British Indian Ocean Territory	86	IOT	NA
29	Caribbean	British Virgin Islands	92	VGB	NA
NA	NA	Brunei Darussalam	96	BRN	NA
NA	NA	Bulgaria	100	BGR	NA
11	Western Africa	Burkina Faso	854	BFA	TRUE
14	Eastern Africa	Burundi	108	BDI	TRUE
11	Western Africa	Cabo Verde	132	CPV	NA
NA	NA	Cambodia	116	KHM	TRUE
17	Middle Africa	Cameroon	120	CMR	NA
NA	NA	Canada	124	CAN	NA
29	Caribbean	Cayman Islands	136	CYM	NA
17	Middle Africa	Central African Republic	140	CAF	TRUE
17	Middle Africa	Chad	148	TCD	TRUE
5	South America	Chile	152	CHL	NA
NA	NA	China	156	CHN	NA
NA	NA	China	156	CHN	NA
NA	NA	China, Hong Kong Special Administrative Region	344	HKG	NA
NA	NA	China, Macao Special Administrative Region	446	MAC	NA
NA	NA	Christmas Island	162	CXR	NA
NA	NA	Cocos (Keeling) Islands	166	CCK	NA
5	South America	Colombia	170	COL	NA
14	Eastern Africa	Comoros	174	COM	TRUE
17	Middle Africa	Congo	178	COG	NA

-As we planned to use k-means algorithm for clustering by SDG Index, and the algorithm cannot handle NA/NaN/Inf values , So we should clean(remove) these observations which contains NA for our input SDG_index_score_2023

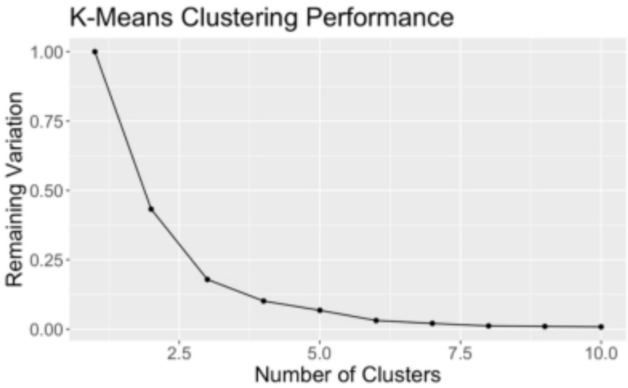


~~~
```{r}
#DATA FORMING
# join tables
data <- inner_join(x=country_indicators, y=sdg, by="country_code_iso3")
country_codes_modified <- rename(country_codes, country_code_iso3 = `ISO-alpha3 Code (M49)`)
final_data <- inner_join(x=data, y=country_codes_modified, by="country_code_iso3")
#filter out Asia contries
data_Asia <- final_data %>% filter(`Region Code (M49)`==142)
#k-means algorithm can not deal with NA/NaN/Inf in SDG_index_score_2023, so we need to clean the data first
clean_data_Asia <- data_Asia %>%
  distinct(country_code_iso3, .keep_all = TRUE) %>% #drop dup
  filter(!is.na(SDG_index_score_2023) & SDG_index_score_2023 != Inf & SDG_index_score_2023 != -Inf)
final_data_clean <- final_data %>%
  distinct(country_code_iso3, .keep_all = TRUE) %>% #drop dup
  filter(!is.na(SDG_index_score_2023) & SDG_index_score_2023 != Inf & SDG_index_score_2023 != -Inf)
```
~~~

– Statistical Analysis and Related Visualization



~~~
*<Research Question 1 Part>*
```{r}
#Clustering with k-means algorithm
#set seed
set.seed(1010315026)
#k-means algorithm
k <- 4
clustering <- kmeans(clean_data_Asia$SDG_index_score_2023, k)
clustering

# add clustering values to our original dataset
clean_data_Asia <-
  clean_data_Asia %>%
  mutate(cluster = clustering$cluster)
~~~



-We use set seed to make sure the clustering is replicable

-Then we perform a k-means algorithm to group countries base on their SDG Index Score

-K-means clustering algorithm helps us assign data to the corresponding clusters based on minimized variance

-Clustering enables decision makers to understand complex dataset's pattern directly and easily, so that they can make efficient decisions that optimize resource allocations.

(e.g. Each groups that shares the similar pattern can share similar policies)

-Then we perform an Elbow Approach to check the optimal numbers of clusters

-The performance curve gets quite flat after 4 clusters, so according to elbow approach, using 3 or 4 clusters would be a reasonable choice

— Statistical Analysis and Related Visualization

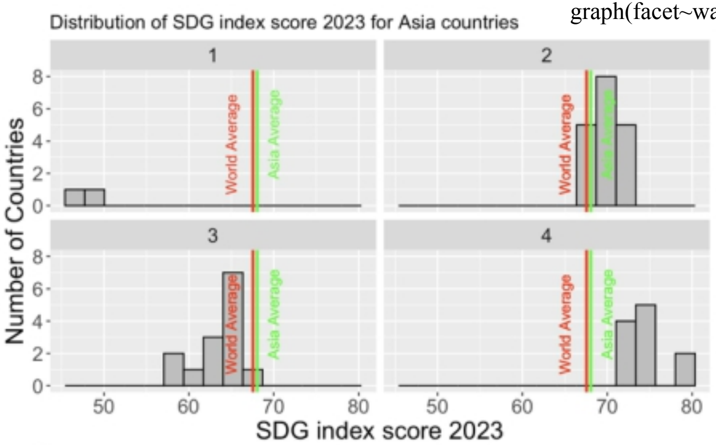
- Why histogram?

Histogram is a great visualization tool for showing the distribution and general pattern.

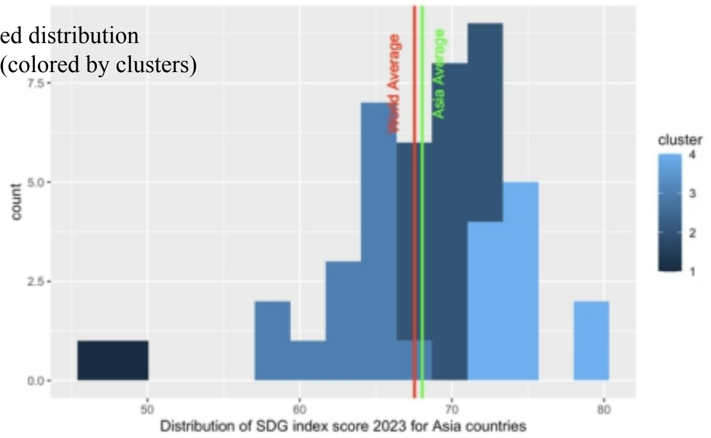
- Why use colored by group and facet~warp?

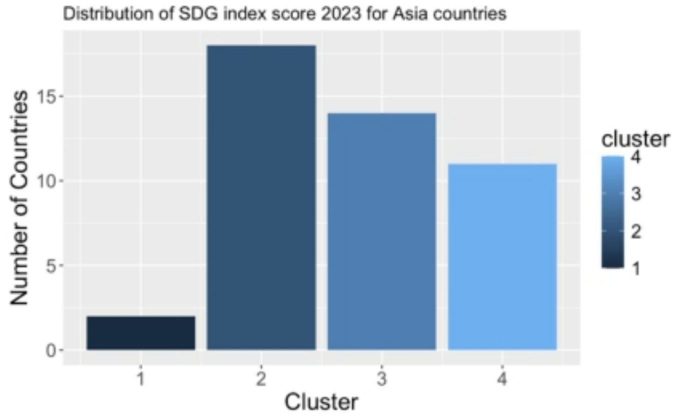
Provide us a more direct and clear view about situation for each cluster and separation between them, which is one of the main purpose of doing clustering compare to the original histogram.

Side by side distribution graph(facet~warp by clusters)



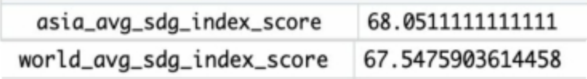
Colored distribution graph(colored by clusters)





Statistical Analysis and Related Visualization

Short Summarization To Previous Clustering



asia_avg_sdg_index_score	68.0511111111111
world_avg_sdg_index_score	67.5475903614458

The distribution of SDG index score is a slightly left skewed normal distribution with one single peak, with most countries' SDG index located near the center, and two cluster with most countries in them next to each other around the center.

Overall the sustainable development level in Asia is preferable, with a mean(68.05) slightly higher than the world average(67.54), and does not have a huge discrepancies because in spite of the first cluster's mean SDG index is significantly lower (47.90) than the average level(68.05), It actually only occupy 4.44% of the Asian countries. The rest 95.56% of Asia countries, although being divided into three different groups, does not have significant difference in mean SDG score(63.62, 69.74 and 74.59), mostly located near the center(68.05).

BUT DOES OUR CLUSTERING MAKE SENSE?

– Statistical Analysis and Related Visualization

Does It make sense for our k-means clustering when we combine our result with other variables?

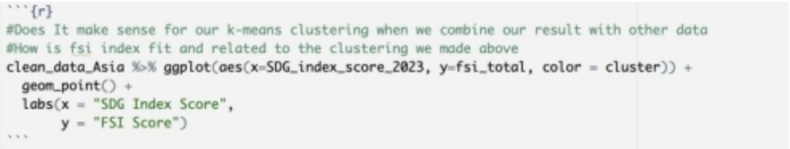
It make sense!

1)With scatter plot colored by clusters, we managed to study how our cluster distributed when two numerical variables are involved.

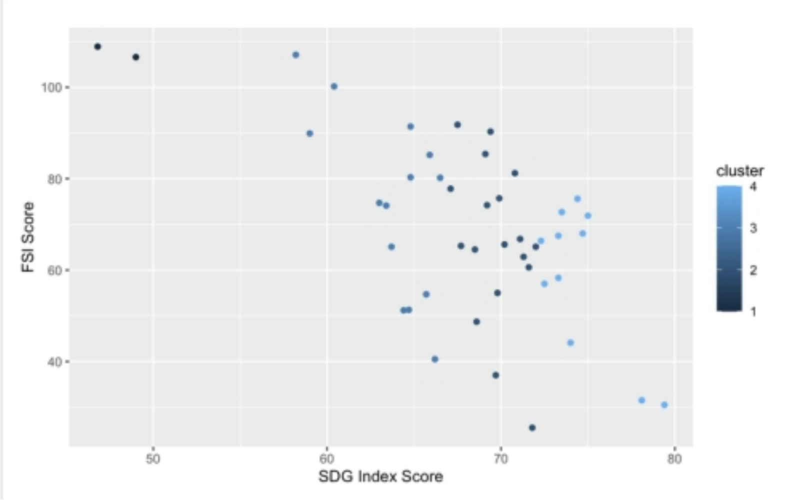
2)The first cluster(Darkest color), the poorly developed group, located at the top left corner, where the FSI Score is highest and SDG Score is lowest

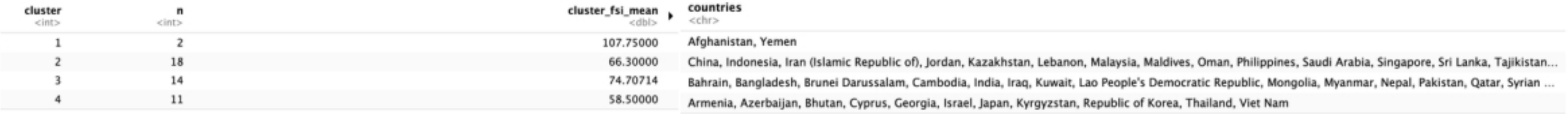
3)The second cluster(second darkest color) has a FSI Score mean of 66.3, which is lower than 74.7 of cluster 3(third darkest color) who has a higher lower mean SDG Score. It is also higher than the FSI Score mean 58.5 of cluster 4, who has the highest average SDG score.

4)FSI Score evaluate a country's stability, the higher the score is, the lower the stability is. More stable the country is, the more likely it is going to have a better development environment and get a higher sustainable development score. All clusters fits this pattern and make sense!



~~~
```{r}
#Does It make sense for our k-means clustering when we combine our result with other data
#How is fsi index fit and related to the clustering we made above
clean_data_Asia %>% ggplot(aes(x=SDG_index_score_2023, y=fsi_total, color = cluster)) +
  geom_point() +
  labs(x = "SDG Index Score",
       y = "FSI Score")
```
~~~





cluster <int>	n <int>	cluster_fsi_mean <dbl>	countries <chr>
1	2	107.75000	Afghanistan, Yemen
2	18	66.30000	China, Indonesia, Iran (Islamic Republic of), Jordan, Kazakhstan, Lebanon, Malaysia, Maldives, Oman, Philippines, Saudi Arabia, Singapore, Sri Lanka, Tajikistan...
3	14	74.70714	Bahrain, Bangladesh, Brunei Darussalam, Cambodia, India, Iraq, Kuwait, Lao People's Democratic Republic, Mongolia, Myanmar, Nepal, Pakistan, Qatar, Syrian ...
4	11	58.50000	Armenia, Azerbaijan, Bhutan, Cyprus, Georgia, Israel, Japan, Kyrgyzstan, Republic of Korea, Thailand, Viet Nam

— Statistical Analysis and Related Visualization

Conclusion and Concerns

So we managed to group Asian countries into 4 different clusters(Poorly, Lower Middle, Upper Middle and Highly) according to their SDG Index Score by k-means algorithm and perform series of advanced histogram to better, clearly explain the complex dataset to the reader, and to show them the pattern efficiently so that they can make resource optimally used.

And with scatter plot, we managed to check whether our clustering result make sense. We combine our clustering based on SDG Score Index with another variable FSI Score index and figure out that overall our cluster for sustainable development level will generally make sense as out clusters follow the fact that the lower FSI is(more stable), the higher SDG is(better development environment). Especially under the situation that our data source are authoritative non profitable organization, meaning that they are generally unbiased and reliable.

However, it's essential to note that while there might be a general trend, the correlation does not necessarily imply causation, and outliers or confundings might exist due to specific national circumstances or other unmeasured factors, our clustering result might still possibly be inaccurate.

Research Question 2

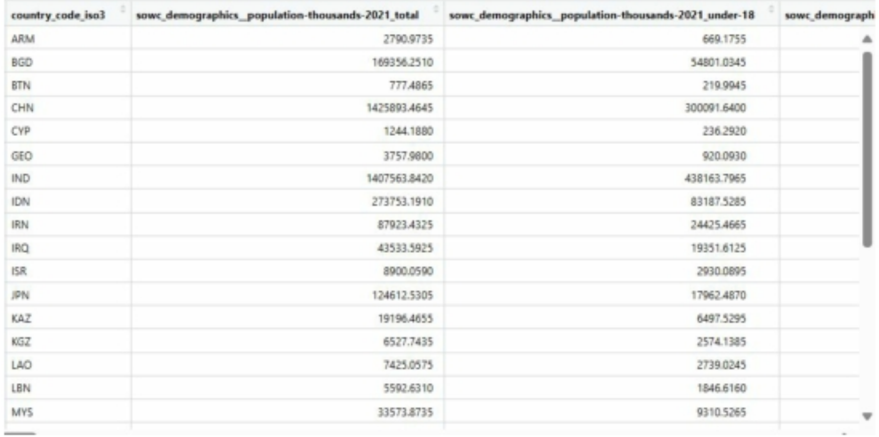
- What is the difference between the rate of the share of household income from Asian countries compared to all Western countries' rates?

– Data Cleaning

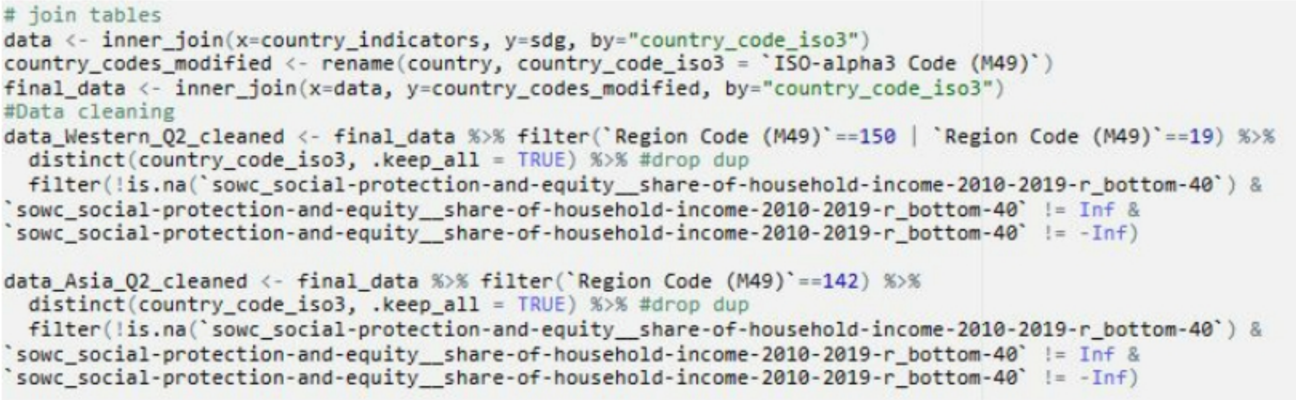
-We have cleaned the original dataset to include only Asian countries there (Have done the same thing for Western Countries)

-Easier to create the two-samples for the following hypothesis test

- Clean for the NA values



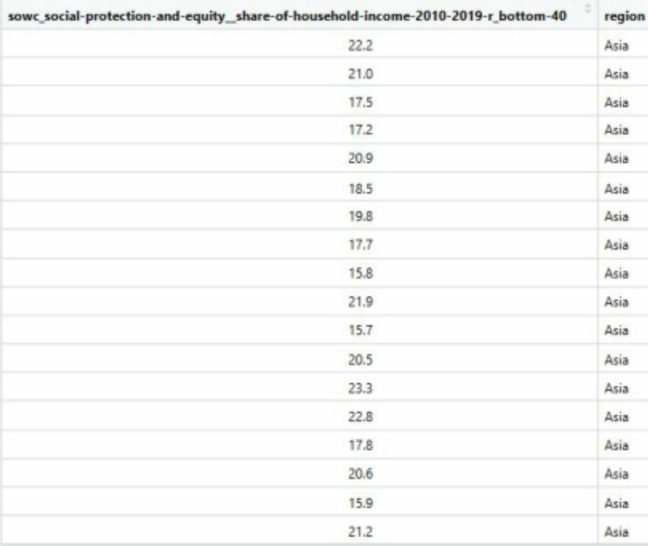
country_code_iso3	sowc_demographics__population-thousands-2021_total	sowc_demographics__population-thousands-2021_under-18	sowc_demograph
ARM	2790.9735	669.1755	
BGD	169356.2510	54801.0345	
BTN	777.4865	219.9945	
CHN	1425893.4645	300091.6400	
CYP	1244.1880	236.2920	
GEO	3757.9800	920.0930	
IND	1407563.8420	438163.7965	
IDN	273753.1910	83187.5285	
IRN	87923.4325	24425.4665	
IRQ	43533.5925	19351.6125	
ISR	8900.0590	2930.0895	
JPN	124612.5305	17962.4870	
KAZ	19196.4655	6497.5295	
KGZ	6527.7435	2574.1385	
LAO	7425.0575	2739.0245	
LBN	5592.6310	1846.6160	
MYS	33573.8735	9310.5265	



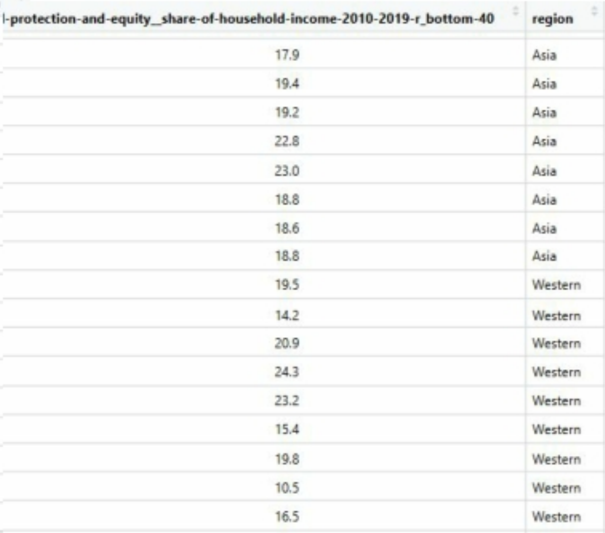
```
# join tables
data <- inner_join(x=country_indicators, y=sdg, by="country_code_iso3")
country_codes_modified <- rename(country, country_code_iso3 = `ISO-alpha3 Code (M49)`)
final_data <- inner_join(x=data, y=country_codes_modified, by="country_code_iso3")
#Data cleaning
data_Western_Q2_cleaned <- final_data %>% filter(`Region Code (M49)`==150 | `Region Code (M49)`==19) %>%
  distinct(country_code_iso3, .keep_all = TRUE) %>% #drop dup
  filter(!is.na(`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40`) &
`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40` != Inf &
`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40` != -Inf)

data_Asia_Q2_cleaned <- final_data %>% filter(`Region Code (M49)`==142) %>%
  distinct(country_code_iso3, .keep_all = TRUE) %>% #drop dup
  filter(!is.na(`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40`) &
`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40` != Inf &
`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40` != -Inf)
```

Data Wrangling

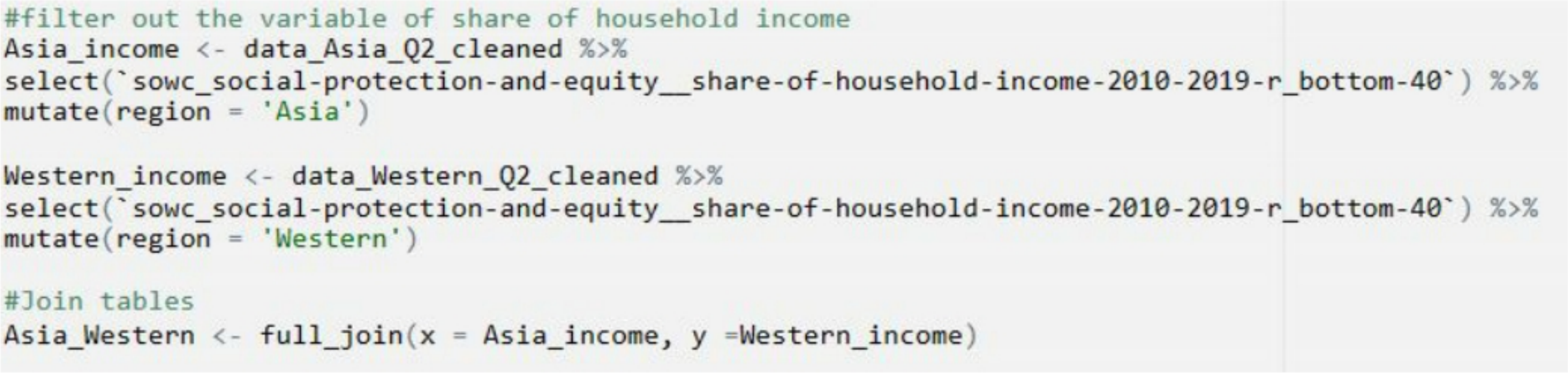
- Filter out the variable of share of household income for both Asian and Western Countries
- Join the two datasets together for a larger sample



sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40	region
22.2	Asia
21.0	Asia
17.5	Asia
17.2	Asia
20.9	Asia
18.5	Asia
19.8	Asia
17.7	Asia
15.8	Asia
21.9	Asia
15.7	Asia
20.5	Asia
23.3	Asia
22.8	Asia
17.8	Asia
20.6	Asia
15.9	Asia
21.2	Asia



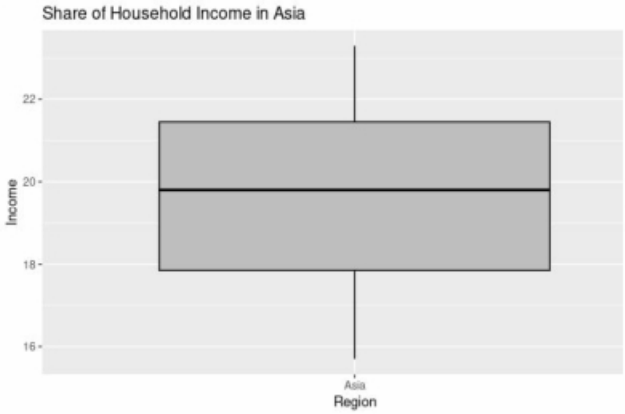
l-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40	region
17.9	Asia
19.4	Asia
19.2	Asia
22.8	Asia
23.0	Asia
18.8	Asia
18.6	Asia
18.8	Asia
19.5	Western
14.2	Western
20.9	Western
24.3	Western
23.2	Western
15.4	Western
19.8	Western
10.5	Western
16.5	Western

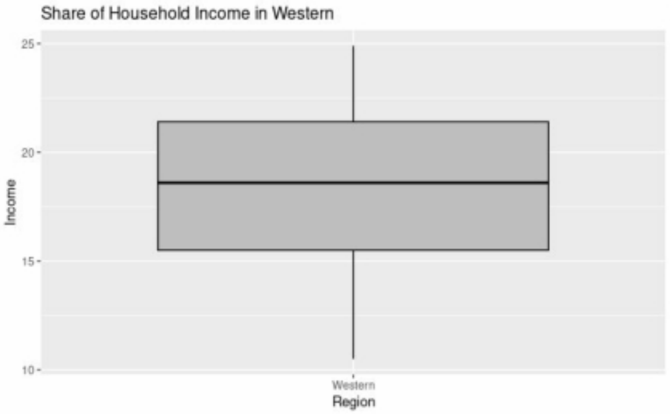


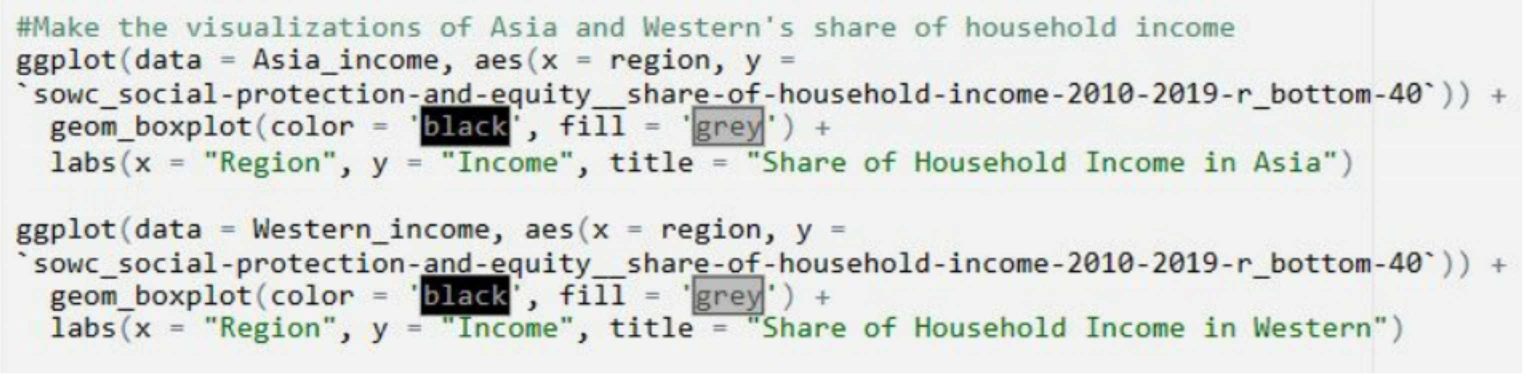
```
#filter out the variable of share of household income
Asia_income <- data_Asia_Q2_cleaned %>%
select(`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40`) %>%
mutate(region = 'Asia')

Western_income <- data_Western_Q2_cleaned %>%
select(`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40`) %>%
mutate(region = 'Western')

#Join tables
Asia_Western <- full_join(x = Asia_income, y =Western_income)
```

– Statistical Analysis and Related Visualization

- Boxplots are useful to show the medians, quantiles between two datasets
- Straightforwardly show the difference


Share of Household Income in Asia

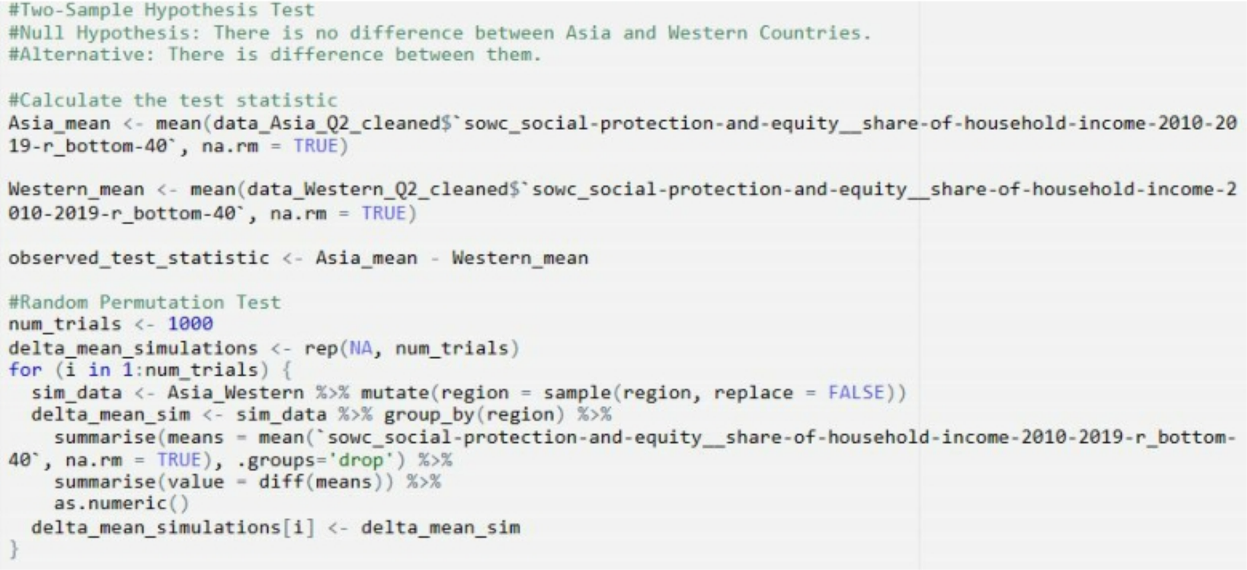

Share of Household Income in Western



```
#Make the visualizations of Asia and Western's share of household income
ggplot(data = Asia_income, aes(x = region, y =
`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40`)) +
  geom_boxplot(color = 'black', fill = 'grey') +
  labs(x = "Region", y = "Income", title = "Share of Household Income in Asia")

ggplot(data = Western_income, aes(x = region, y =
`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-40`)) +
  geom_boxplot(color = 'black', fill = 'grey') +
  labs(x = "Region", y = "Income", title = "Share of Household Income in Western")
```

Two-Sample Hypothesis Test

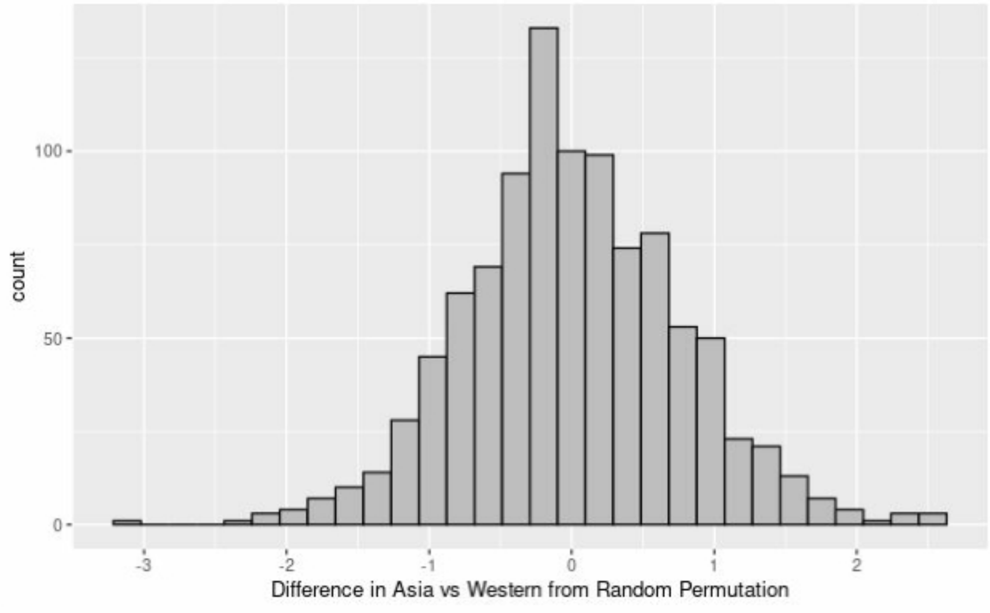
- State the null hypothesis and the alternative hypothesis
- Population: All countries in Asian and Western continents
- Parameter: mean income values between Asia and Western
- Calculate the difference between the mean values of Asian income and Western income
- Perform the permutation test for sample simulation

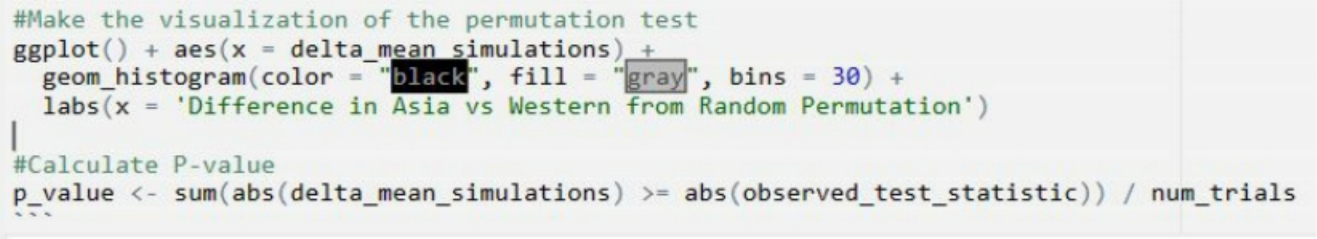


```
#Two-Sample Hypothesis Test
#Null Hypothesis: There is no difference between Asia and Western Countries.
#Alternative: There is difference between them.

#Calculate the test statistic
Asia_mean <- mean(data_Asia_Q2_cleaned$`sowc_social-protection-and-equity__share-of-household-income-2010-20
19-r_bottom-40`, na.rm = TRUE)

Western_mean <- mean(data_Western_Q2_cleaned$`sowc_social-protection-and-equity__share-of-household-income-2
010-2019-r_bottom-40`, na.rm = TRUE)

observed_test_statistic <- Asia_mean - Western_mean

#Random Permutation Test
num_trials <- 1000
delta_mean_simulations <- rep(NA, num_trials)
for (i in 1:num_trials) {
  sim_data <- Asia_Western %>% mutate(region = sample(region, replace = FALSE))
  delta_mean_sim <- sim_data %>% group_by(region) %>%
    summarise(means = mean(`sowc_social-protection-and-equity__share-of-household-income-2010-2019-r_bottom-
40`, na.rm = TRUE), .groups='drop') %>%
    summarise(value = diff(means)) %>%
    as.numeric()
  delta_mean_simulations[i] <- delta_mean_sim
}
```

P-value and distribution

- Calculating p-value using the formula where the delta_mean_simulations represent the simulated test statistic
- Histogram is straightforward to show the difference in Asia vs Western from sampling simulation
  - Unimodal
  - Centered around -0.9
  - Spread from -3 to 2





```
#Make the visualization of the permutation test
ggplot() + aes(x = delta_mean_simulations) +
  geom_histogram(color = "black", fill = "gray", bins = 30) +
  labs(x = 'Difference in Asia vs Western from Random Permutation')

#Calculate P-value
p_value <- sum(abs(delta_mean_simulations) >= abs(observed_test_statistic)) / num_trials
```

Two Sample-Hypothesis Test

Conclusion and Concerns

The p-value calculated here is 0.117, which is higher than the alpha level 0.05. Thus, the null hypothesis would be accepted, which states that there is no difference between the household income in Asia and Western. This conclusion can also be reflected by the boxplots that are made at the beginning, which shows that the mean income between the continents are similar. However, the dataset contains too many non-values and the household income may not represent the whole economic growth of the continents. Hence, there may result in a type II error where it fails to reject the hypothesis. A more specific analysis needs larger samples and more concise values.

Research Question 3

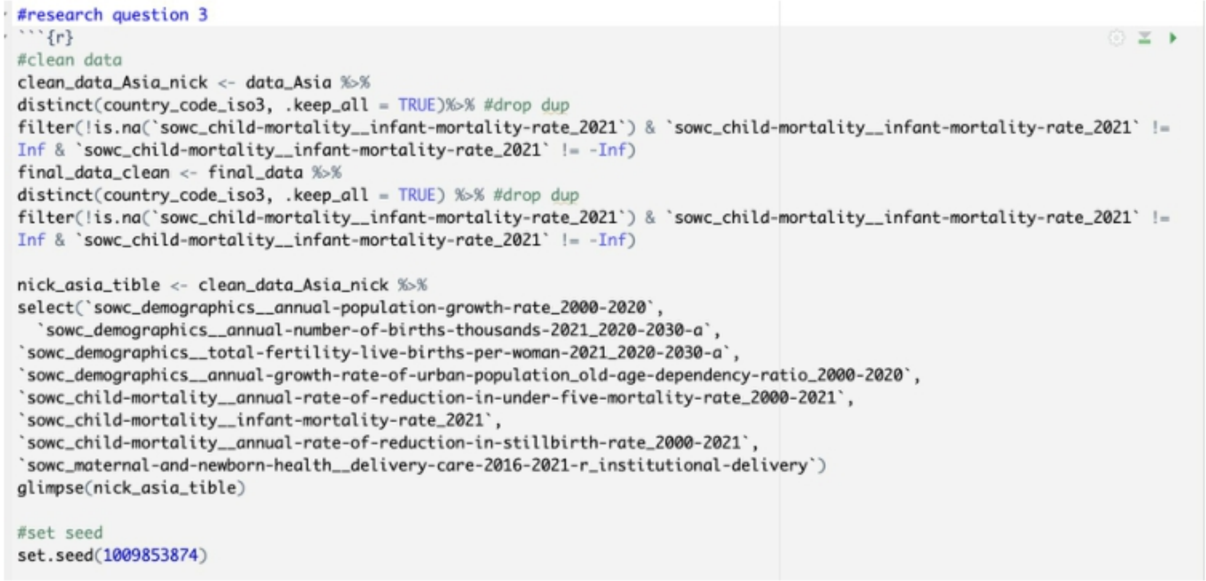
How does infant mortality rate affect population growth among asian countries? Is there a trend showing they are proportional?

Hypothesis

we suspect countries with higher mortality rate also have higher population growth rate, and lower mortality rate also have lower mortality rate. We also suspect there is a certain percentage below such that those countries population growth rate would not be related to the infant mortality rate.

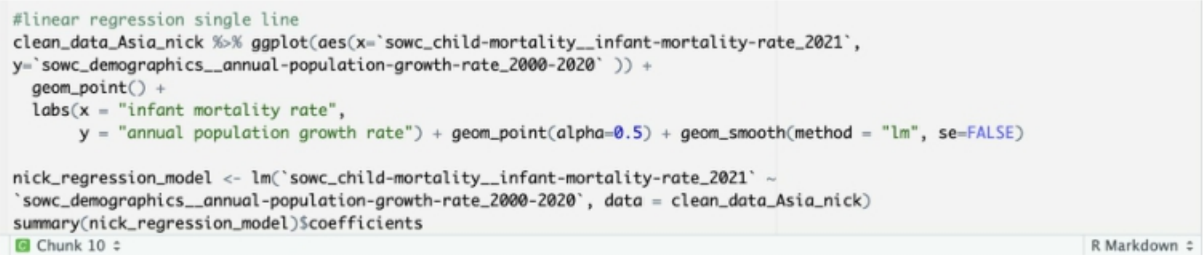
Data cleaning

Same as question 1



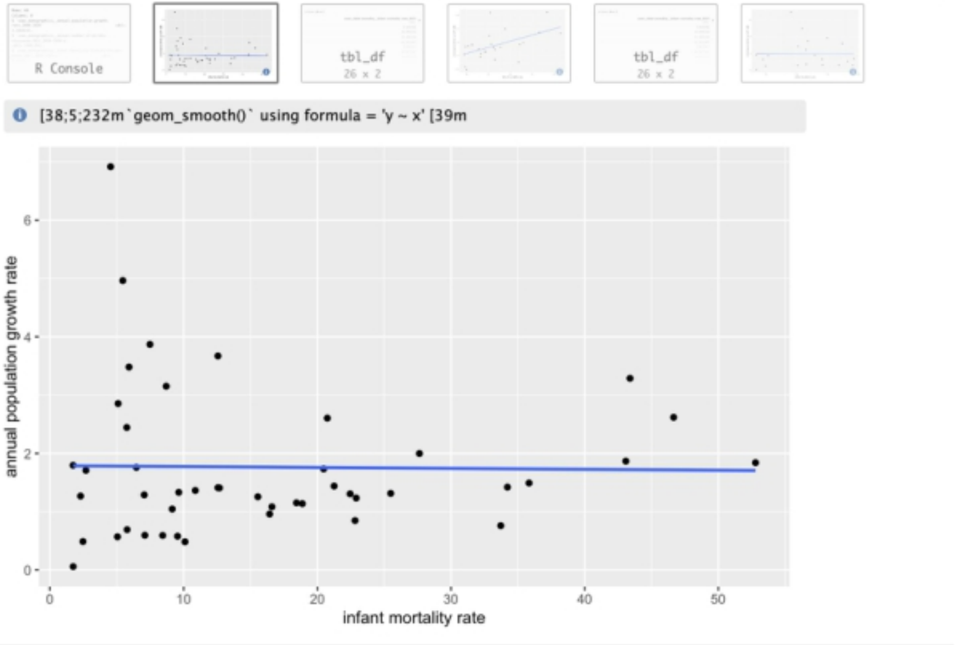
~~~
#research question 3
```{r}
#clean data
clean_data_Asia_nick <- data_Asia %>%
distinct(country_code_iso3, .keep_all = TRUE)%>% #drop dup
filter(!is.na(`sowc_child-mortality__infant-mortality-rate_2021`) & `sowc_child-mortality__infant-mortality-rate_2021` !=
Inf & `sowc_child-mortality__infant-mortality-rate_2021` != -Inf)
final_data_clean <- final_data %>%
distinct(country_code_iso3, .keep_all = TRUE) %>% #drop dup
filter(!is.na(`sowc_child-mortality__infant-mortality-rate_2021`) & `sowc_child-mortality__infant-mortality-rate_2021` !=
Inf & `sowc_child-mortality__infant-mortality-rate_2021` != -Inf)

nick_asia_tible <- clean_data_Asia_nick %>%
select(`sowc_demographics__annual-population-growth-rate_2000-2020`,
  `sowc_demographics__annual-number-of-births-thousands-2021_2020-2030-a`,
`sowc_demographics__total-fertility-live-births-per-woman-2021_2020-2030-a`,
`sowc_demographics__annual-growth-rate-of-urban-population_old-age-dependency-ratio_2000-2020`,
`sowc_child-mortality__annual-rate-of-reduction-in-under-five-mortality-rate_2000-2021`,
`sowc_child-mortality__infant-mortality-rate_2021`,
`sowc_child-mortality__annual-rate-of-reduction-in-stillbirth-rate_2000-2021`,
`sowc_maternal-and-newborn-health__delivery-care-2016-2021-r_institutional-delivery`)
glimpse(nick_asia_tible)

#set seed
set.seed(1009853874)
~~~

Statistical Analysis and Related Visualization



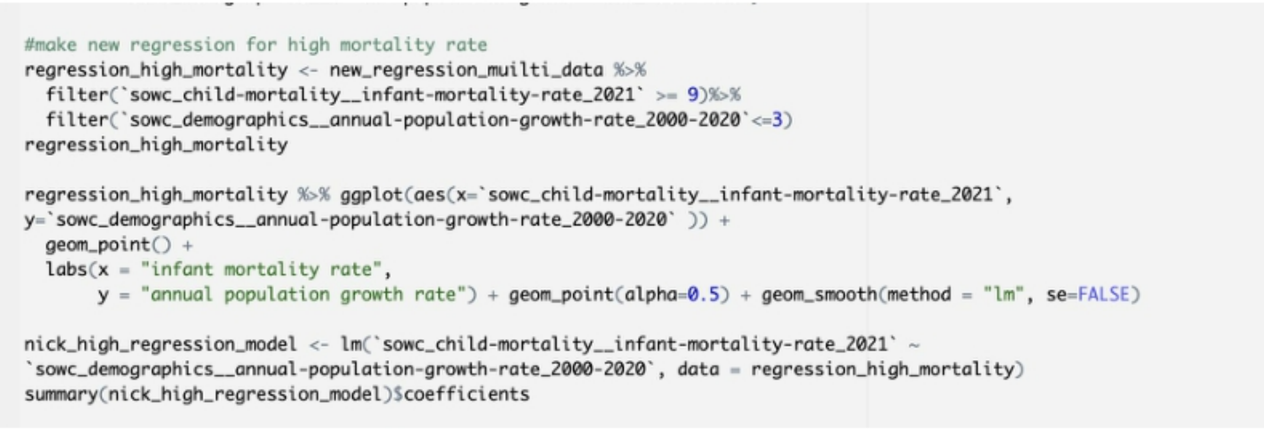
```
#linear regression single line
clean_data_Asia_nick %>% ggplot(aes(x=`sowc_child-mortality__infant-mortality-rate_2021`,
y=`sowc_demographics__annual-population-growth-rate_2000-2020` )) +
  geom_point() +
  labs(x = "infant mortality rate",
       y = "annual population growth rate") + geom_point(alpha=0.5) + geom_smooth(method = "lm", se=FALSE)

nick_regression_model <- lm(`sowc_child-mortality__infant-mortality-rate_2021` ~
`sowc_demographics__annual-population-growth-rate_2000-2020`, data = clean_data_Asia_nick)
summary(nick_regression_model)$coefficients
```

in the scatter graph we could see some extreme values, there are two countries that stand out the most, one of them is having highest population growth rate with low mortality rate, and the other one is quite the opposite, it has the highest infant mortality rate but has lower annual population rate. From what we get from the regression line, we can see the slope of the line is almost zero (-0.161) , which means there is no relation between infant mortality rate and population growth rate in asia. But when we look closely, we can tell that there are two trends that cancel each other out. one is from lower annual population rate and higher population growth rate, other is the opposite. As a result, I will try to make two regression lines.



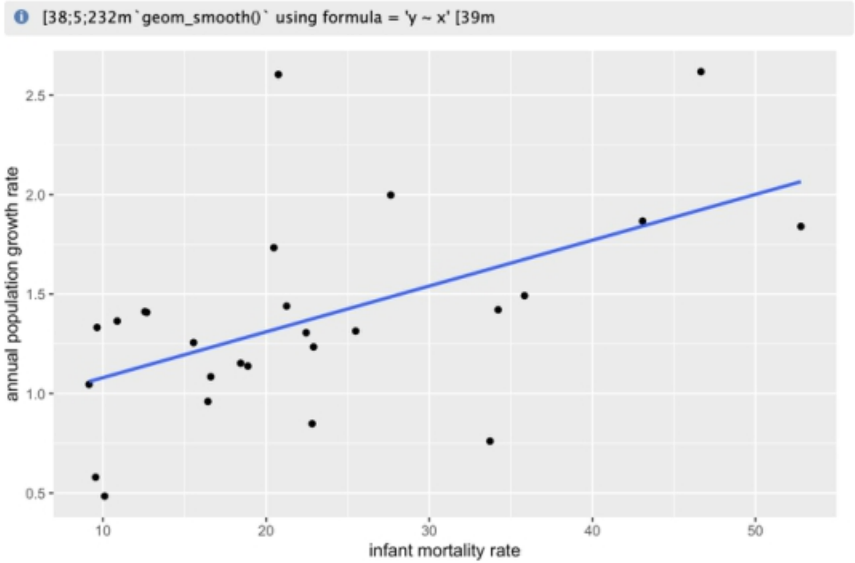
```
[38;5;232m`geom_smooth()` using formula = 'y ~ x' [39m
```

Statistical Analysis and Related Visualization



```
#make new regression for high mortality rate
regression_high_mortality <- new_regression_muilti_data %>%
  filter(`sowc_child-mortality__infant-mortality-rate_2021` >= 9)%>%
  filter(`sowc_demographics__annual-population-growth-rate_2000-2020`<=3)
regression_high_mortality

regression_high_mortality %>% ggplot(aes(x=`sowc_child-mortality__infant-mortality-rate_2021`,
y=`sowc_demographics__annual-population-growth-rate_2000-2020` )) +
  geom_point() +
  labs(x = "infant mortality rate",
       y = "annual population growth rate") + geom_point(alpha=0.5) + geom_smooth(method = "lm", se=FALSE)

nick_high_regression_model <- lm(`sowc_child-mortality__infant-mortality-rate_2021` ~
`sowc_demographics__annual-population-growth-rate_2000-2020`, data = regression_high_mortality)
summary(nick_high_regression_model)$coefficients
```

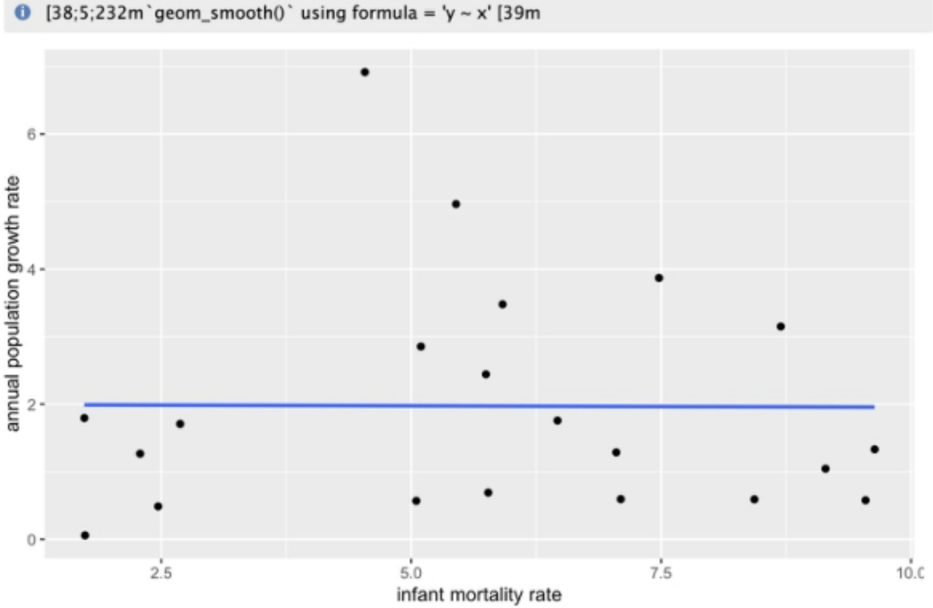
For the higher mortality rate countries, there is a clear upward trend, the slope is about 12.3 , shows how for those countries who have relatively higher mortality rate will likely to have higher annual population growth rate. Which is accurate because we suspect countries with higher mortality rate would also have higher population growth rate.



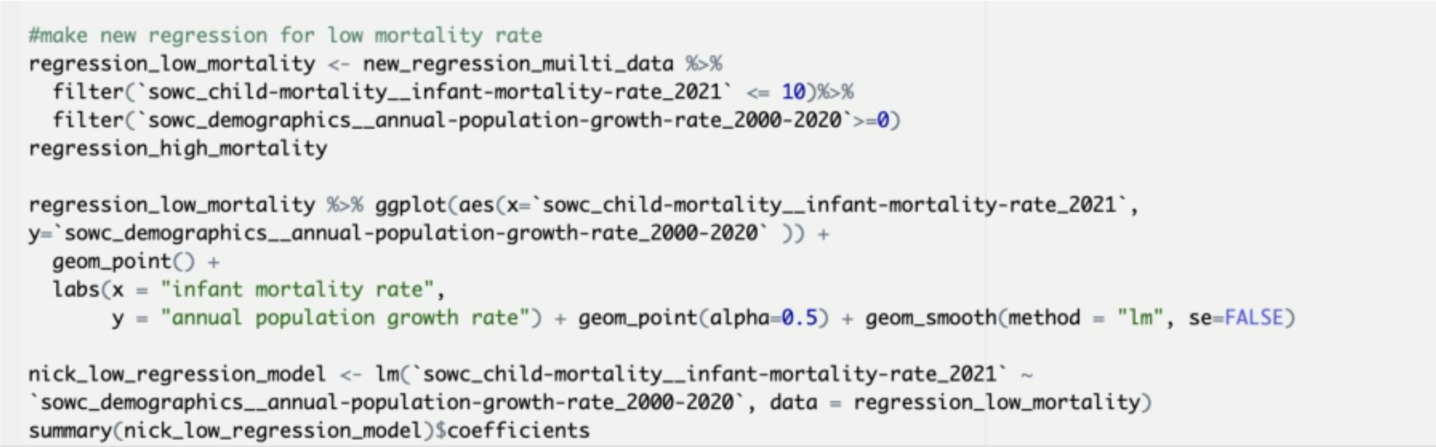
```
ⓘ [38;5;232m`geom_smooth()` using formula = 'y ~ x' [39m
```

Statistical Analysis and Related Visualization

For lower mortality rate countries, the regression line is flat, the slope is -0.00908. Which means that there is no clear trend between infant mortality rate and population growth rate. Which is also intuitive, because we can see under 10 infant mortality rate we can see the regression is flat.



```
[38;5;232m`geom_smooth()` using formula = 'y ~ x' [39m
```



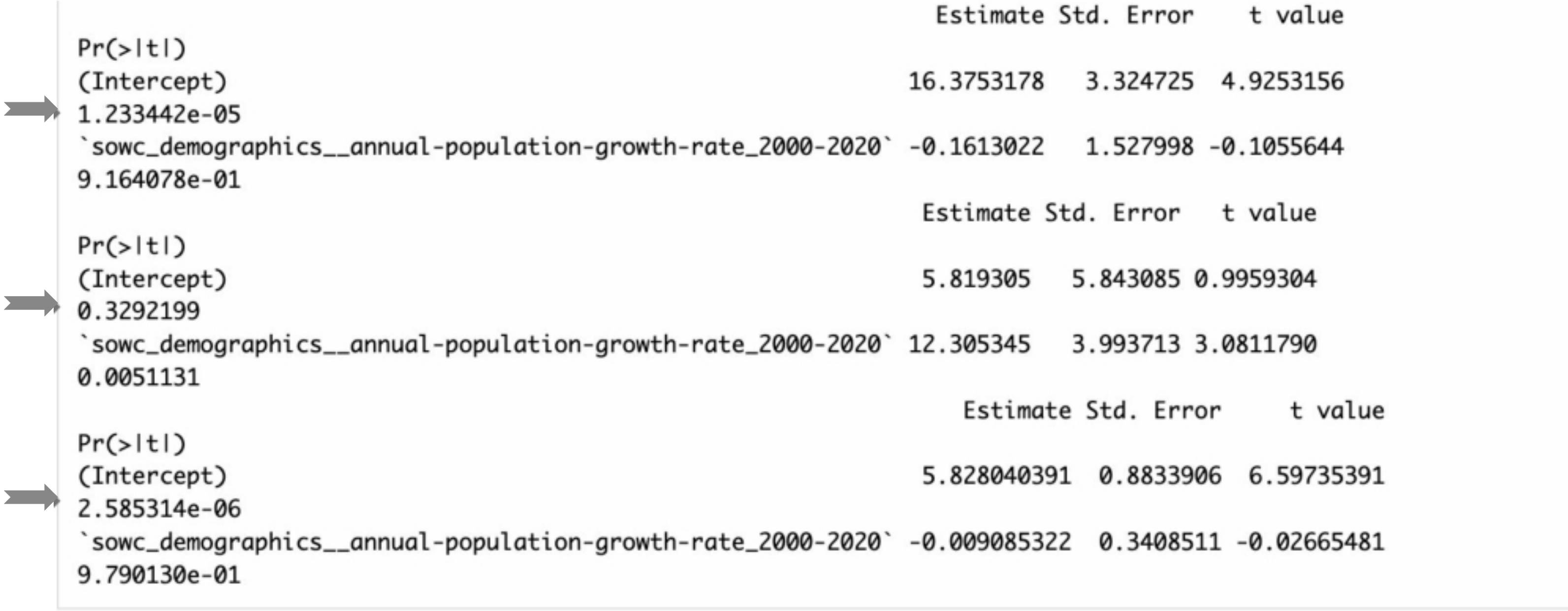
```
#make new regression for low mortality rate
regression_low_mortality <- new_regression_muilti_data %>%
  filter(`sowc_child-mortality__infant-mortality-rate_2021` <= 10)%>%
  filter(`sowc_demographics__annual-population-growth-rate_2000-2020`>=0)
regression_high_mortality

regression_low_mortality %>% ggplot(aes(x=`sowc_child-mortality__infant-mortality-rate_2021`,
y=`sowc_demographics__annual-population-growth-rate_2000-2020` )) +
  geom_point() +
  labs(x = "infant mortality rate",
       y = "annual population growth rate") + geom_point(alpha=0.5) + geom_smooth(method = "lm", se=FALSE)

nick_low_regression_model <- lm(`sowc_child-mortality__infant-mortality-rate_2021` ~
`sowc_demographics__annual-population-growth-rate_2000-2020`, data = regression_low_mortality)
summary(nick_low_regression_model)$coefficients
```

Fitting regression line and interpreting output

Single line

Higher mortality rate

Lower mortality rate



```
                                                              Estimate Std. Error    t value
Pr(>|t|)
(Intercept)                                                 16.3753178   3.324725  4.9253156
1.233442e-05
`sowc_demographics__annual-population-growth-rate_2000-2020` -0.1613022   1.527998 -0.1055644
9.164078e-01
                                                             Estimate Std. Error   t value
Pr(>|t|)
(Intercept)                                                  5.819305   5.843085 0.9959304
0.3292199
`sowc_demographics__annual-population-growth-rate_2000-2020` 12.305345   3.993713 3.0811790
0.0051131
                                                                Estimate Std. Error     t value
Pr(>|t|)
(Intercept)                                                  5.828040391  0.8833906  6.59735391
2.585314e-06
`sowc_demographics__annual-population-growth-rate_2000-2020` -0.009085322  0.3408511 -0.02665481
9.790130e-01
```

Limitations and Concerns

-Firstly, the limitation and concerns we see is that linear regression is sensitive to outliers, which means some of the extreme values would affect our prediction.

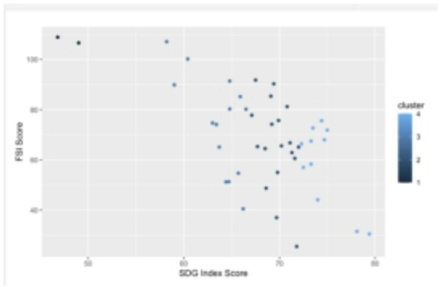
-Second of the limitations is that we only looked at the effect of infant mortality rate, which is only a small part of country population growth. There are other factors that affect the outcome.

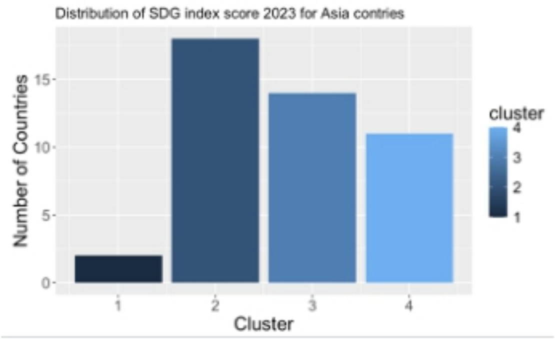
Conclusion - Research Question 1

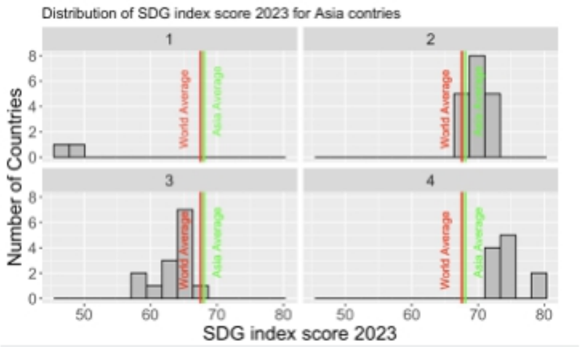
- We examined **disparities in development levels** within **Asian countries** by grouping them into **4 distinct clusters** that represents varying levels of development within Asian countries and combined with the Fragile States Index (FSI) scores to check the validity.

-The first cluster, marked by the darkest color, denotes poorly developed nations, characterized by **high FSI scores** and low **SDG scores** . We can observe a trend where **higher stability correlates with better development and higher SDG scores** . The fourth cluster, having the highest average SDG score, exhibits the lowest FSI score mean.

-This shows the **negative impact** of **instability** on sustainable development efforts.





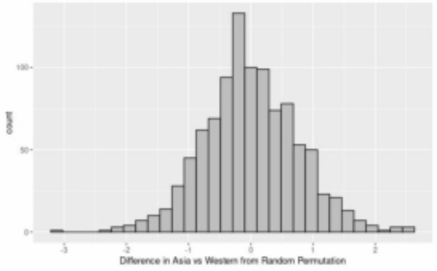


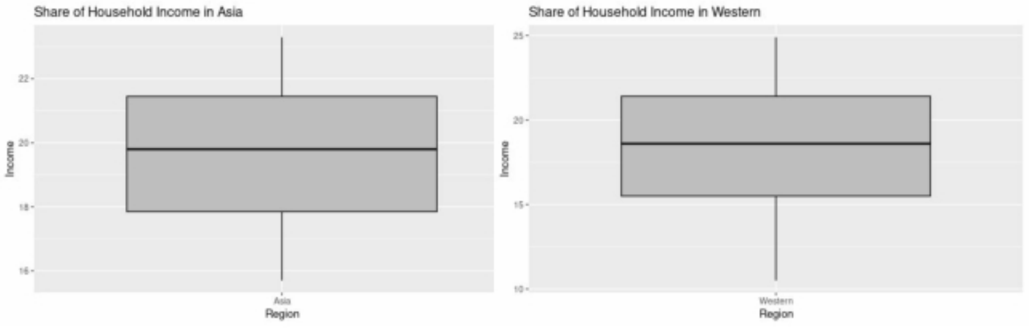
Conclusion - Research Question 2

-We examined the difference in **household income** between **Asian** and **Western** countries to determine **Asia’s economic development**

-With a calculated **p-value of 0.117 exceeding** the **alpha level of 0.05** , we accept the null hypothesis, suggesting there is no significant difference in household incomes between the two regions

-This aligns with observations from the initial boxplot analysis, which showed **similar mean** `s **continents** . However, we acknowledge dataset limitations, such as **non-values** and the **in ability of household income** fully represent **economic landscapes**





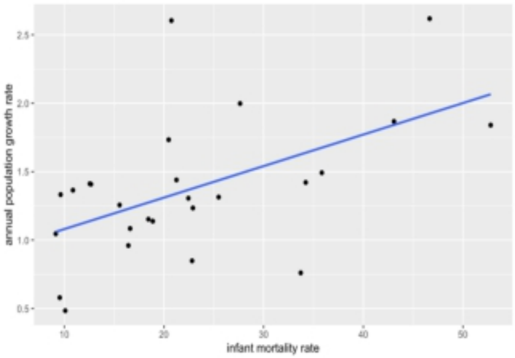
Conclusion - Research Question 3

-We examined the connection between **infant mortality rate** and **population growth** rate in Asian countries

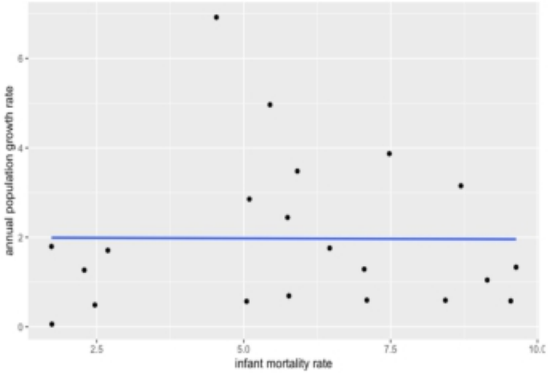
-After closer examination, we made **two linear regression lines** : one indicating an **upward trend for countries with higher mortality rates** , confirming our hypothesis, and the other showing a **flat regression line for countries with lower mortality rates** , indicating no clear relationship

-We have overall determined that the countries have a **healthy annual population rate** despite **different infant mortality rates** indicating **better population growth** of **younger ages**

Higher mortality Rate Countries



Lower mortality Rate Countries



Conclusion

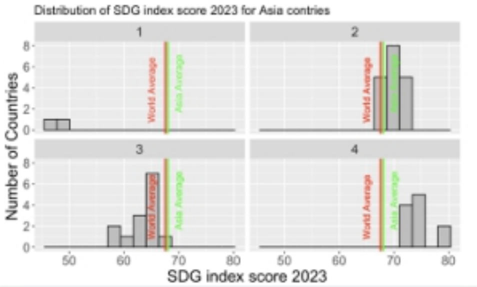
- The conclusions drawn from the analysis of the three research questions collectively provide a comprehensive understanding of Asia's developmental landscape:

1. We have provided the **number of countries** in different **clusters** that signify **instability levels**
2. Showed that there are **no economic disparities between Western and Asian countries**
3. Showed that there is **healthy population growth of the younger generation**

All the research questions enriched our report and made it more meaningful in helping policymakers judge the current situation in Asia through the complex dataset, enabling them to efficiently allocate resources and labour

Although development levels are subjective, we believe that Asia is developing quite well but there are improvements that could be made in terms of development:

1. Finding the countries that are in cluster 1 and 2
2. Finding more factors that cause the countries in these clusters to be far behind (Exploitation, environmental, economic)
3. Making a plan to fix those problems



References and Citations

-"Remove NA Rows in R." *ProgrammingR* , https://www.programmingr.com/examples/remove-na-rows-in-r/ . 2024

-"United Nations Development Programme." *Sustainable Development Goals* , United Nations Development Programme, https://www.undp.org/sustainable-development-goals. , Updated in 2023

-"Fragile States Index." *The Fund for Peace* , https://fragilestatesindex.org/ . Updated in 2023

-"Statistical Tables: The State of the World's Children 2023." *UNICEF Data* , UNICEF, https://data.unicef.org/resources/dataset/the-state-of-the-worlds-children-2023-statistical-tables/ . Updated in 2023