ANALYSIS

# KEY TAKEAWAYS

The aim of this project is to analyze internal migration data in Turkey and investigate the factors such as migration reasons, age, education, and other variables that influence migration. We conducted our research using the internal migration and population datasets available on the Turkish Statistical Institute (TÜİK) website. The conclusions drawn from this project are outlined in the following points:

1. We had assumed that the highest migration due to retirement would be to the Central Anatolia Region or the Aegean Region. To validate this assumption, we created a migration value and cause graph. When looking at this graph, we observed that the highest migration is to the Marmara Region. To obtain more accurate results, we created a migration rate and cause graph. In this graph, we see that the highest migration is to the Black Sea and Aegean regions. So, we realized that part of our assumption is correct. Understanding the importance of creating a Rate graph, we concluded that we should make decisions based on this graph.
2. The age range of the people who migrate the most is seen in the graph as 20-24. The age group of the people who migrated the second most is seen in the graph as 15-19. In the graph of reasons for migration, the most common reason for migration was observed to be education. We can reconcile this incident with this situation.
3. The visual representation of migration numbers across different educational backgrounds highlights a compelling inverse relationship between education and migration. Doctoral and primary school-educated individuals show lower migration rates compared to high school and university-educated individuals. Individuals with higher education levels tend to migrate less. The scatter plot analysis reveals a lack of a straightforward correlation between the average education level and fertility rates across regions in the dataset. While some general trends emerge—like higher education levels in certain regions coinciding with lower fertility rates—outliers such as Central Anatolia and the Mediterranean Region defy these patterns.

# DATA PREPROCESSING

# importing necessary packages  
library(tidyverse)  
library(readxl)  
library(readr)  
library(gridExtra)  
library(dplyr)  
  
population <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx")  
names(population) <- c('IDD','city','regionid','regions','totalpop','male','female','age04','age59','age1014','age1519','age2024','age2529','age3034','age3539','age4044','otherage','unknowns','doctorate','primaryedu','elementaryedu','highschool','literatebutnoschool','notliterate','middleschool','master','university','fertilityrate','electricity','numberofattempts','housingsalesnumbers')  
  
region26 <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx", sheet = "Bolge26")  
names(region26) <- c('region','region2id','workforce15plus','workforce1564','usableincome')  
  
migration <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx")  
population <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx")  
names(population) <- c('IDD','city','regionid','regions','totalpop','male','female','age04','age59','age1014','age1519','age2024','age2529','age3034','age3539','age4044','otherage','unknowns','doctorate','primaryedu','elementaryedu','highschool','literatebutnoschool','notliterate','middleschool','master','university','fertilityrate','electricity','numberofattempts','housingsalesnumbers')  
  
region26 <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx", sheet = "Bolge26")  
names(region26) <- c('region','region2id','workforce15plus','workforce1564','usableincome')  
  
migration <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx", sheet = "Goc Bilgileri")  
  
population <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx")  
names(population) <- c('IDD','city','regionid','regions','totalpop','male','female','age04','age59','age1014','age1519','age2024','age2529','age3034','age3539','age4044','otherage','unknowns','doctorate','primaryedu','elementaryedu','highschool','literatebutnoschool','notliterate','middleschool','master','university','fertilityrate','electricity','numberofattempts','housingsalesnumbers')  
  
region26 <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx", sheet = "Bolge26")  
names(region26) <- c('region','region2id','workforce15plus','workforce1564','usableincome')  
  
migration <- read\_excel("C:\\Users\\melek\\desktop\\DataVizards\_FinalDataFrame.xlsx", sheet = "Goc Bilgileri")  
  
names(migration) <- c('IDD','male2','female2','turnbackfamily','unknowns2','betterlifecond','others','education','retirement','buyhome','familymig','finjob','maritalstatuschange','health','appointment','age04\_2','age59\_2','age1014\_2','age1519\_2','age2024\_2','age2529\_2','age3034\_2','age3539\_2','age4044\_2','university2','highschool2','middleschool2', 'elementaryschool2')  
  
  
# tidy migration dataset  
migration$city <- population$city   
migration$regions <- population$regions  
  
tidy\_data\_gender <- migration |> pivot\_longer(c(male2,female2),names\_to = "gender2",values\_to = "gender\_value2")  
  
  
tidy\_data\_causes <- tidy\_data\_gender |> pivot\_longer(c(turnbackfamily,unknowns2,betterlifecond,others,education,  
 retirement,buyhome,familymig,finjob,maritalstatuschange,health,appointment),names\_to = "migrationcauses",values\_to = "migrationcauses\_value")  
  
tidy\_data\_age <- tidy\_data\_causes |> pivot\_longer(c(age04\_2,age59\_2,age1014\_2,age1519\_2,  
 age2024\_2,age2529\_2,age3034\_2,age3539\_2,age4044\_2),names\_to = "agerange2",values\_to = "agerange\_value2")  
  
last\_migration\_tidy\_data <- tidy\_data\_age |> pivot\_longer(c(university2,highschool2,middleschool2,elementaryschool2),names\_to = "education2",values\_to = "education\_value2")  
head(last\_migration\_tidy\_data)

# A tibble: 6 × 11  
 IDD city regions gender2 gender\_value2 migrationcauses  
 <dbl> <chr> <chr> <chr> <dbl> <chr>   
1 1 Adana Akdeniz Bölgesi male2 25200 turnbackfamily   
2 1 Adana Akdeniz Bölgesi male2 25200 turnbackfamily   
3 1 Adana Akdeniz Bölgesi male2 25200 turnbackfamily   
4 1 Adana Akdeniz Bölgesi male2 25200 turnbackfamily   
5 1 Adana Akdeniz Bölgesi male2 25200 turnbackfamily   
6 1 Adana Akdeniz Bölgesi male2 25200 turnbackfamily   
# ℹ 5 more variables: migrationcauses\_value <dbl>, agerange2 <chr>,  
# agerange\_value2 <dbl>, education2 <chr>, education\_value2 <dbl>

# tidy region26 dataset  
tidy\_region26 <- region26 |> pivot\_longer(c(workforce15plus,workforce1564),names\_to = "workforce",values\_to = "workforce\_values")  
head(tidy\_region26)

# A tibble: 6 × 5  
 region region2id usableincome workforce workforce\_values  
 <chr> <chr> <dbl> <chr> <dbl>  
1 Adana, Mersin TR62 0.382 workforce1… 1579  
2 Adana, Mersin TR62 0.382 workforce1… 1528  
3 Ağrı, Kars, Iğdır, Ardahan TRA2 0.381 workforce1… 383  
4 Ağrı, Kars, Iğdır, Ardahan TRA2 0.381 workforce1… 369  
5 Ankara TR51 0.353 workforce1… 2341  
6 Ankara TR51 0.353 workforce1… 2308

# tidy population dataset  
tidy\_data\_gender2 <- population |> pivot\_longer(c(male,female),names\_to = "gender",values\_to = "gender\_value")   
  
tidy\_data\_age <- tidy\_data\_gender2 |> pivot\_longer(c(age04,age59,age1014,age1519,  
 age2024,age2529,age3034,age3539,age4044,otherage),names\_to = "agerange",values\_to = "agerange\_value")  
  
tidy\_literate <- tidy\_data\_age |> pivot\_longer(c(unknowns,literatebutnoschool,notliterate),names\_to = "literate",values\_to = "literate\_values")  
  
last\_tidy\_data\_education2 <- tidy\_literate|> pivot\_longer(c(university,highschool,middleschool,elementaryedu,doctorate,primaryedu,master),names\_to = "education",values\_to = "education\_value")  
  
last\_population\_tidy\_data <- last\_tidy\_data\_education2 |> pivot\_longer(c(fertilityrate,electricity,numberofattempts,housingsalesnumbers),names\_to = "othervariables",values\_to = "othervariables\_value")  
head(last\_population\_tidy\_data)

# A tibble: 6 × 15  
 IDD city regionid regions totalpop gender gender\_value agerange  
 <dbl> <chr> <chr> <chr> <dbl> <chr> <dbl> <chr>   
1 1 Adana A Akdeniz Bölgesi 2263373 male 1130862 age04   
2 1 Adana A Akdeniz Bölgesi 2263373 male 1130862 age04   
3 1 Adana A Akdeniz Bölgesi 2263373 male 1130862 age04   
4 1 Adana A Akdeniz Bölgesi 2263373 male 1130862 age04   
5 1 Adana A Akdeniz Bölgesi 2263373 male 1130862 age04   
6 1 Adana A Akdeniz Bölgesi 2263373 male 1130862 age04   
# ℹ 7 more variables: agerange\_value <dbl>, literate <chr>,  
# literate\_values <dbl>, education <chr>, education\_value <dbl>,  
# othervariables <chr>, othervariables\_value <dbl>

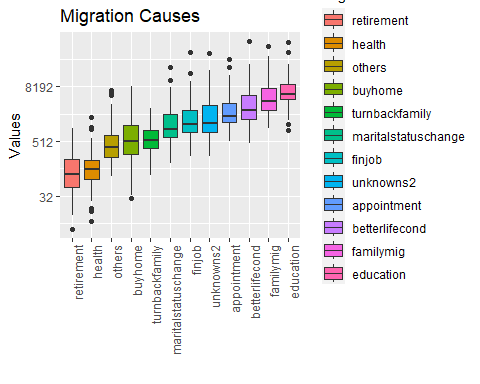
population$`region2id`<- ""  
for (i in 1:nrow(population)) {  
 city <- population$city[i]  
   
 for (j in 1:length(region26$region)) {  
 if (grepl(city, region26$region[j])) {  
 population$`region2id`[i] <- paste(region26$`region2id`[j])  
 break   
 }  
 }  
}

Now that all data preparations are complete, we can begin the process of visualization :)

# DATA VISUALIZATION

## Migration Causes VS. Migration Values

library(tidyverse)  
library(readxl)  
library(readr)  
  
# Migration Causes Box Plot  
last\_migration\_tidy\_data |> mutate(migrationcauses = reorder(migrationcauses, migrationcauses\_value, FUN = median)) |>  
 ggplot(aes(migrationcauses,migrationcauses\_value,fill = migrationcauses)) +  
 geom\_boxplot() + ggtitle("Migration Causes") + xlab("Migration Causes") +  
 ylab("Values") + theme(axis.text.x = element\_text(angle= 90, hjust = 1)) + xlab("") +  
 scale\_y\_continuous(trans = "log2")

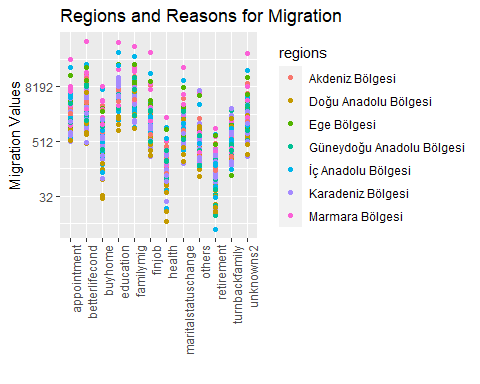


To have an overview of data on migration causes we plotted a box plot to see what reasons cause migration the most. Box plot shows that education is the most common reason for migration followed by migration due to a family member and better life conditions.

When plotting the box plot we scaled y values to log2 scale to see the distribution better and we ordered the box plots to make it easier to understand which are the most and the least causes of migration.

## Regions and Reasons for Migration

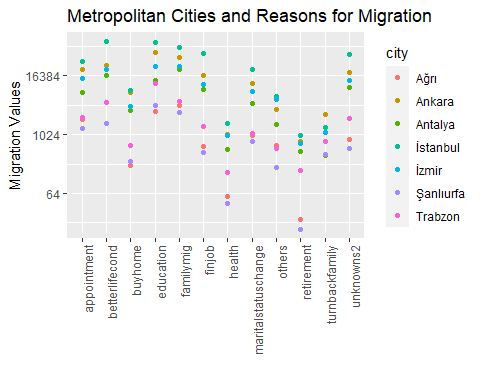
region\_plot <- last\_migration\_tidy\_data |> select(migrationcauses,migrationcauses\_value,regions) |>  
 ggplot(aes(migrationcauses,migrationcauses\_value)) + geom\_point(aes(color = regions)) +  
 ggtitle("Regions and Reasons for Migration") + xlab("Migration Causes") +  
 ylab("Migration Values") + theme(axis.text.x = element\_text(angle= 90, hjust = 1)) + xlab("") + scale\_y\_continuous(trans = "log2")   
print(region\_plot)



When we analyze migration reasons on a regional basis, we observe that under the category of ‘different reasons,’ the highest migration has occurred to the Marmara region. Following the Marmara region, the second-highest migration took place to the Central Anatolia region. This drew our attention to a few points. We had expected that the highest migration due to retirement would be to the Central Anatolian Region or Aegean Region, but it turned out that the majority of migration happened to the Marmara region.

## Metropolitan Cities and Reasons for Migration

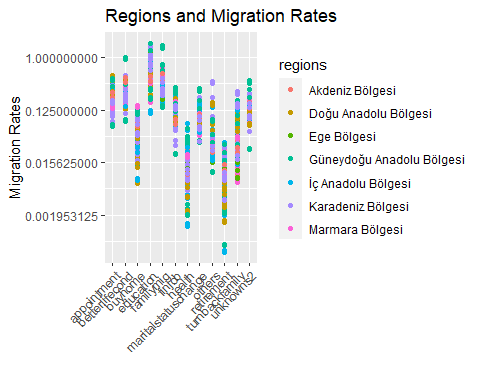
city\_plot <- last\_migration\_tidy\_data |> filter(city %in% c("Ankara","İstanbul","İzmir","Antalya","Trabzon","Ağrı","Şanlıurfa")) |> select(migrationcauses,migrationcauses\_value,city) |>  
 ggplot(aes(migrationcauses,migrationcauses\_value)) + geom\_point(aes(color = city)) +  
 ggtitle("Metropolitan Cities and Reasons for Migration") + xlab("Migration Causes") +  
 ylab("Migration Values") + theme(axis.text.x = element\_text(angle= 90, hjust = 1)) + xlab("") + scale\_y\_continuous(trans = "log2")   
print(city\_plot)



While creating this graph, we included the most populous cities of each region because plotting for all 81 provinces would be quite challenging. As seen in this graph, the highest migration, due to different reasons, has been to Istanbul, followed by Ankara. What surprised us again is that the highest migration due to retirement is to İstanbul and Ankara.”

## Regions and Migration Rates

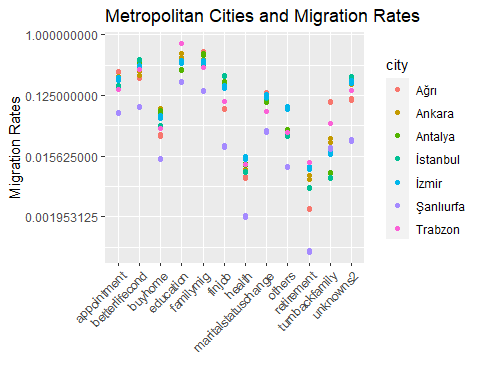
migration\_gender2 <- last\_migration\_tidy\_data |> mutate(rate = (last\_migration\_tidy\_data$migrationcauses\_value) / (last\_migration\_tidy\_data$gender\_value2))  
  
region\_rate\_plot <- migration\_gender2 |> ggplot(aes(migrationcauses,rate)) + geom\_point(aes(color = regions)) +  
 ggtitle("Regions and Migration Rates") + xlab("Migration Causes") +  
 ylab("Migration Rates") + theme(axis.text.x = element\_text(angle= 45, hjust = 1)) + xlab("") + scale\_y\_continuous(trans = "log2")  
 print(region\_rate\_plot)



In the above plots , plotted values were migration values with respect to migration reason and we got some results that we were not expecting. To understand the situation deeply we plotted migration rates rather than migration values and result was more of what we were expected to see. For example migration due to retirement was actually occurring to Central Anatolian Region and Black Sea Region.

## Metropolitan Cities and Migration Rates

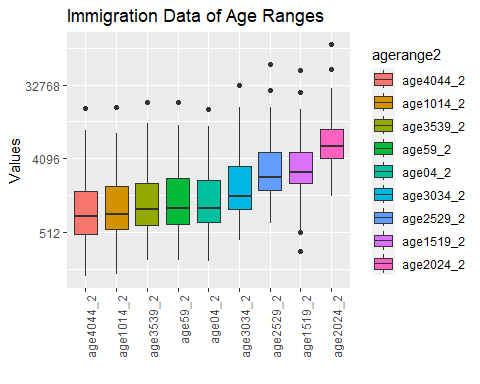
migration\_gender2 <- last\_migration\_tidy\_data |> mutate(rate = (last\_migration\_tidy\_data$migrationcauses\_value) / (last\_migration\_tidy\_data$gender\_value2))  
city\_rate\_plot <- migration\_gender2 |> filter(city %in% c("Ankara","İstanbul","İzmir","Antalya","Trabzon","Ağrı","Şanlıurfa")) |> ggplot(aes(migrationcauses,rate)) + geom\_point(aes(color = city)) +  
 ggtitle("Metropolitan Cities and Migration Rates") + xlab("Migration Causes") +  
 ylab("Migration Rates") + theme(axis.text.x = element\_text(angle= 45, hjust = 1)) + xlab("") + scale\_y\_continuous(trans = "log2")   
 print(city\_rate\_plot)



In the same fashion, the city that people migrate the most due to retirement is Trabzon followed by İzmir and Ankara.

## Immigration Data of Age Ranges

last\_migration\_tidy\_data |> mutate(agerange2 = reorder(agerange2, agerange\_value2, FUN = median)) |>  
 ggplot(aes(agerange2,agerange\_value2,fill = agerange2)) +  
 geom\_boxplot() + ggtitle("Immigration Data of Age Ranges") + xlab("Age Ranges") +  
 ylab("Values") + theme(axis.text.x = element\_text(angle= 90, hjust = 1)) + xlab("") +  
 scale\_y\_continuous(trans = "log2")



When we look at the graph, we see that the age range of the people who migrate the most is 20-24. The 15-19 age range ranks second. The age range of people who migrate the least is observed to be 40-44. When we look at the Immigration Data of Age Ranges, we see that the most immigration age range is 20-24. We can say that education was the highest value among the reasons for immigration and this is consistent with this. Likewise, the second place is 15-19. It can be interpreted that high school students also migrate for education. “finjob” ranks in the middle list of reasons for migration. If we consider 25-29 as the age for finding a job, we can comment that people aged 25-29 do not migrate only for the purpose of finding a job. For example, they may have migrated due to better living conditions. When we look at the graph, we see that the migration age of 40-44 migrates the least compared to other ages. In our opinion, we can conclude from the graph that people who are used to a certain order no longer want change. According to the chart, 10-14 years old is the second to last migration age range. “familymig” ranked second among the reasons for migration. We can say that there is a contradiction here. We would have expected a higher number of people migrating for health reasons. Health is very important for everyone, but unfortunately health services are not developed everywhere. Health problems are a factor that can be seen in all age groups, so we would expect the values to be high.

## Correlations Between Migration Reasons and Age Ranges.

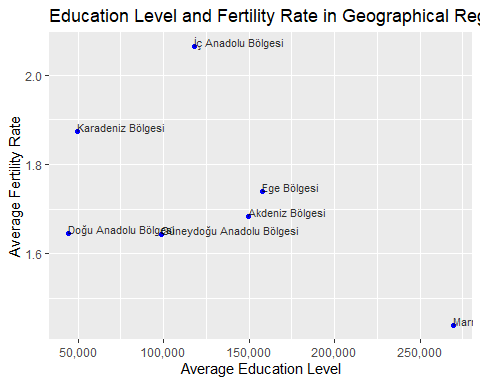
cor(last\_migration\_tidy\_data$migrationcauses\_value, last\_migration\_tidy\_data$agerange\_value2)

[1] 0.4977804

This value indicates a moderate positive relationship between migration reasons and age ranges.

## Education Level and Fertility Rate in Geographical Regions

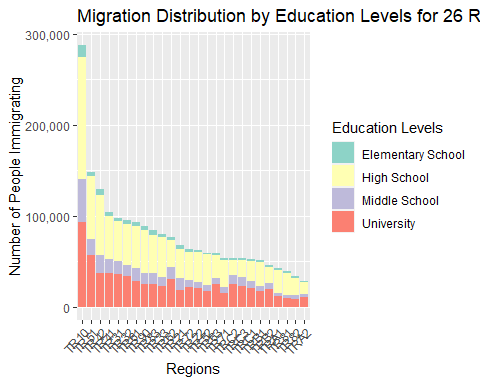
library(tidyverse)  
# Extracting necessary columns from the 'last\_tidy\_data\_education2' dataset  
data <- last\_tidy\_data\_education2 %>%  
 select(regions, education, education\_value, fertilityrate)   
# Calculating average education and fertility rate per region  
education\_fertility <- data %>%  
 group\_by(regions) %>%  
 summarise(  
 avg\_education = mean(education\_value), # Calculating the mean education value for each region  
 avg\_fertility = mean(fertilityrate) # Calculating the mean fertility rate for each region  
 )  
# Creating a scatter plot to visualize the relationship between education level and fertility rate across regions  
ggplot(education\_fertility, aes(x = avg\_education, y = avg\_fertility, label = regions)) +  
 geom\_point(color = "blue") + # Adding points to represent each region  
 geom\_text(hjust = 0, vjust = 0, size = 3, color = "black", alpha = 0.8) + # Adding text labels for regions  
 labs(  
 title = "Education Level and Fertility Rate in Geographical Regions", # Title of the plot  
 x = "Average Education Level", # X-axis label  
 y = "Average Fertility Rate" # Y-axis label  
 ) +  
 scale\_x\_continuous(labels = scales::comma, breaks = seq(0, 300000, by = 50000)) # Formatting X-axis labels



The scatter plot does not show a clear correlation between the average education level and fertility rate across regions. The points representing different regions are scattered without showing a clear trend of increase or decrease. The region with the highest average education level is Marmara, with about 250,000 units of education value. The region with the lowest average education level is Eastern Anatolia Region, with about 50,000 units of education value. The region with the highest average fertility rate is Central Anatolia Region, with above 2.0 children per woman. The region with the lowest average fertility rate is the Marmara Region, with below 1.5 children per woman. There are some outliers in the data, such as Central Anatolia Region, which has a relatively high average education level but also a high average fertility rate, and Mediterranean Region, which has a relatively low average education level but also a low average fertility rate. These outliers may indicate that there are other factors influencing the fertility rate besides the education level, such as culture, religion, income, etc. The surprising point in this data is that I expected the Eastern Anatolia Region to have the highest fertility rate. The low population registration rates in that region may be one of the reasons affecting the data output. According to TUIK data, the population registration rate of the Eastern Anatolia Region is 97.92%, which is 98.82% below the Turkey average. This may mean that some of the children born in this region are not registered.

## Migration Distribution by Education Levels for 26 Regions

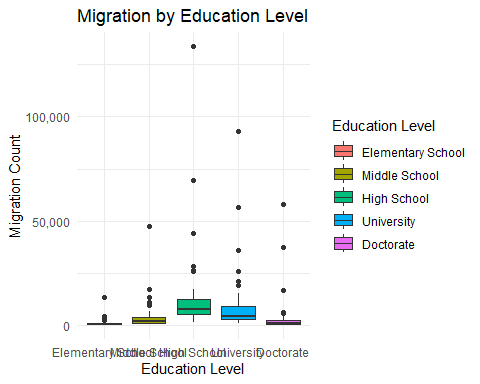
library(tidyverse)  
# Merging population and migration datasets based on the 'city' column  
merged\_data <- population %>%  
 inner\_join(migration, by = "city") %>%  
 # Selecting specific columns from the merged dataset  
 select(city, region2id, university2, highschool2, middleschool2, elementaryschool2)  
  
# Transforming the merged data into a tidy format using pivot\_longer  
merged\_data\_tidy <- merged\_data %>%  
 pivot\_longer(  
 cols = c(university2, highschool2, middleschool2, elementaryschool2),  
 names\_to = "education\_level",  
 values\_to = "migration\_count"  
 ) %>%  
 # Grouping by region and education level to calculate total migration counts  
 group\_by(region2id, education\_level) %>%  
 summarise(total\_migration = sum(migration\_count, na.rm = TRUE)) %>%  
 ungroup() %>%  
 # Renaming education levels for better readability  
 mutate(  
 education\_level = case\_when(  
 education\_level == "university2" ~ "University",  
 education\_level == "middleschool2" ~ "Middle School",  
 education\_level == "highschool2" ~ "High School",  
 education\_level == "elementaryschool2" ~ "Elementary School",  
 TRUE ~ as.character(education\_level)  
 )  
 )   
  
# Sorting regions based on total migration counts  
sorted\_regions <- merged\_data\_tidy %>%  
 group\_by(region2id) %>%  
 summarise(total\_migration = sum(total\_migration, na.rm = TRUE)) %>%  
 arrange(desc(total\_migration)) %>%   
 pull(region2id)  
  
# Reordering factor levels of 'region2id' based on sorted regions  
merged\_data\_tidy$region2id <- factor(merged\_data\_tidy$region2id, levels = sorted\_regions)  
  
# Creating a stacked bar plot to visualize migration distribution by education levels across regions  
ggplot(merged\_data\_tidy, aes(x = region2id, y = total\_migration, fill = education\_level)) +  
 geom\_bar(stat = "identity", position = "stack") + # Stacked bar plot  
 labs(  
 title = "Migration Distribution by Education Levels for 26 Regions", # Title of the plot  
 x = "Regions", # X-axis label  
 y = "Number of People Immigrating", # Y-axis label  
 fill = "Education Levels" # Legend label  
 ) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + # Rotating x-axis labels for better readability  
 scale\_fill\_brewer(palette = "Set3") + # Setting color palette for education levels  
 scale\_y\_continuous(labels = scales::comma) # Formatting Y-axis labels using commas for better readability



It is seen that immigrants with high education levels mostly migrate to regions numbered TR10, TR51 and TR42. These regions cover the northwestern, western and southwestern regions of Turkey. It may indicate that these regions are more attractive in terms of education and job opportunities. It is seen that immigrants with low education level mostly migrate to TR81, TR82 and TRA2 regions. These regions cover the northeastern, eastern and southeastern regions of Turkey. It may indicate that these regions are poorer in terms of education and job opportunities. It may indicate that other Regions are average in terms of education and job opportunities.

## Migration by Education Level

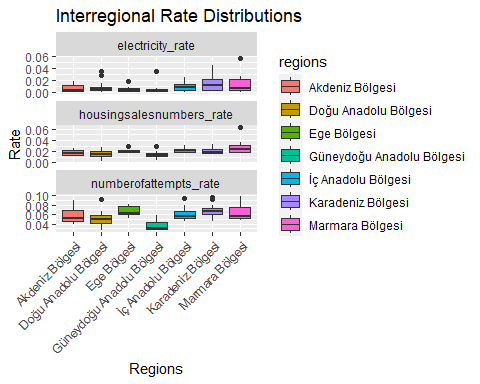
library(tidyverse)  
  
# Merging two datasets 'population' and 'migration' based on the common column 'city'  
merged\_data <- inner\_join(population, migration, by = "city") %>%  
 # Selecting specific columns from the merged dataset  
 select(city, region2id, university2, doctorate, highschool2, elementaryschool2, middleschool2, finjob)  
  
# Reshaping the data using pivot\_longer to gather migration counts for different education levels  
education\_migration <- merged\_data %>%  
 pivot\_longer(  
 cols = c(university2, doctorate, highschool2, elementaryschool2, middleschool2),  
 names\_to = "education\_level",  
 values\_to = "migration\_count"  
 ) %>%  
 # Renaming and categorizing education levels for better visualization  
 mutate(  
 education\_level = case\_when(  
 education\_level == "university2" ~ "University",  
 education\_level == "doctorate" ~ "Doctorate",  
 education\_level == "highschool2" ~ "High School",  
 education\_level == "elementaryschool2" ~ "Elementary School",  
 education\_level == "middleschool2" ~ "Middle School",  
 TRUE ~ as.character(education\_level)  
 )  
 )  
  
# Setting the order of the education levels for better visualization  
education\_migration$education\_level <- factor(education\_migration$education\_level,   
 levels = c("Elementary School", "Middle School", "High School", "University", "Doctorate"))  
  
# Creating a boxplot to visualize the migration count across different education levels  
ggplot(education\_migration, aes(x = education\_level, y = migration\_count, fill = education\_level)) +  
 geom\_boxplot() + # Adding boxplot geometry  
 labs(  
 title = "Migration by Education Level", # Title of the plot  
 x = "Education Level", # X-axis label  
 y = "Migration Count", # Y-axis label  
 fill = "Education Level" # Legend label  
 ) +  
 theme\_minimal() + # Applying a minimal theme to the plot  
 scale\_y\_continuous(labels = scales::comma) # Formatting Y-axis labels with commas for thousands



On the x axis, there are four education levels: doctorate, primary school, high school and university. On the Y axis, the number of people migrating is shown with values ranging from 0 to 100,000. In the graph, the number of migrations of individuals at doctoral or primary school level is much less than that of individuals at high school and university levels. In the graph, the highest number of migrations is seen among individuals at university level. This is followed by individuals at the high school level. It is seen that individuals with a higher level of education migrate less than individuals with a lower level of education. This may indicate that individuals with higher levels of education are more dependent on working conditions or find fewer job opportunities. It is seen that individuals with low education levels migrate more than individuals with high education levels. This may indicate that individuals with lower levels of education strive more to improve their living conditions or are under more pressure.

## Inter-regional Rate Distributions

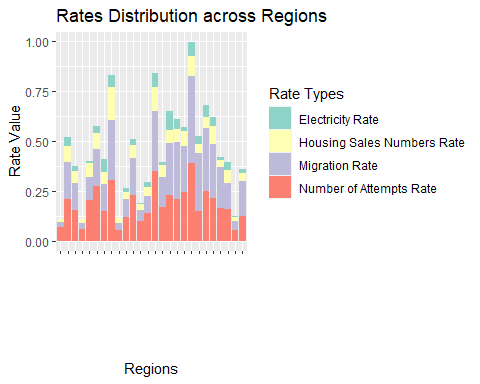
migration\_rate <- (migration$male2+migration$female2)/population$totalpop  
numberofattempts\_rate <- population$numberofattempts/population$totalpop  
housingsalesnumbers\_rate <- population$housingsalesnumbers/population$totalpop  
electricity\_rate <- population$electricity/population$totalpop  
educated\_rate <- (population$highschool + population$university + population$master + population$doctorate)/population$totalpop  
  
merged\_data <- cbind(population, numberofattempts\_rate, housingsalesnumbers\_rate, electricity\_rate)  
  
merged\_data$regions <- factor(merged\_data$regions)   
  
  
  
  
merged\_data\_long <- merged\_data %>%  
 pivot\_longer(cols = c(numberofattempts\_rate, housingsalesnumbers\_rate, electricity\_rate),  
 names\_to = "variable", values\_to = "rate")  
  
  
ggplot(merged\_data\_long, aes(x = regions, y = rate, fill = regions)) +  
 geom\_boxplot() +  
 labs(x = "Regions", y = "Rate", title = "Interregional Rate Distributions") +  
 facet\_wrap(~ variable, scales = "free\_y", ncol = 1) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



ratios <- data.frame(  
 electricity\_rate ,   
 housingsalesnumbers\_rate,  
 numberofattempts\_rate,  
 migration\_rate,  
 educated\_rate  
)  
  
correlation\_matrix <- cor(ratios)  
  
print(correlation\_matrix)

electricity\_rate housingsalesnumbers\_rate  
electricity\_rate 1.0000000 0.2673481  
housingsalesnumbers\_rate 0.2673481 1.0000000  
numberofattempts\_rate 0.4605411 0.4822953  
migration\_rate 0.6837392 0.2184274  
educated\_rate 0.6228220 0.2711951  
 numberofattempts\_rate migration\_rate educated\_rate  
electricity\_rate 0.4605411 0.6837392 0.6228220  
housingsalesnumbers\_rate 0.4822953 0.2184274 0.2711951  
numberofattempts\_rate 1.0000000 0.5059387 0.6996717  
migration\_rate 0.5059387 1.0000000 0.8014049  
educated\_rate 0.6996717 0.8014049 1.0000000

population$`Migration Rate` <- (migration$male2 + migration$female2) / population$totalpop  
population$`Number of Attempts Rate` <- population$numberofattempts / population$totalpop  
population$`Housing Sales Numbers Rate` <- population$housingsalesnumbers / population$totalpop  
population$`Electricity Rate` <- population$electricity / population$totalpop  
  
  
rates\_data <- population %>%  
 select(city, region2id, `Migration Rate`, `Number of Attempts Rate`, `Housing Sales Numbers Rate`, `Electricity Rate`)  
  
rates\_data\_tidy <- rates\_data %>%  
 pivot\_longer(  
 cols = c(`Migration Rate`, `Number of Attempts Rate`, `Housing Sales Numbers Rate`, `Electricity Rate`),  
 names\_to = "rate\_type",  
 values\_to = "rate\_value"  
 )  
  
ggplot(rates\_data\_tidy, aes(x = region2id, y = rate\_value, fill = rate\_type)) +  
 geom\_bar(stat = "identity", position = "stack") +  
 labs(  
 title = "Rates Distribution across Regions",  
 x = "Regions",  
 y = "Rate Value",  
 fill = "Rate Types"  
 ) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 50)) +  
 scale\_fill\_brewer(palette = "Set3") +  
 scale\_y\_continuous(labels = scales::comma)



The high education level of the Marmara Region may cause electricity consumption to be high in this region. However, other regions appear to have lower-than-expected electricity consumption. While high usage is expected due to illegal electricity use, especially in Eastern Anatolia, the data do not support this prediction.

Since the Aegean Region is a summer resort, it was expected that house sales would be high. The fact that Southeastern Anatolia has the lowest housing sales rates is an expected result due to the development level of the region.

Number of Initiatives and Touristic Regions: The fact that the Black Sea is a region with an unexpectedly high number of initiatives may be due to the fact that it is a touristic region. This supports our prediction, as the Aegean Region ranks second as a touristic region. It is relatively expected that Southeastern Anatolia has the lowest number of startups. The fact that the Black Sea is a region with an unexpectedly high number of initiatives may be due to the fact that it is a touristic region. This supports our prediction, as the Aegean Region ranks second as a touristic region. It is relatively expected that Southeastern Anatolia has the lowest number of startups.

When looking at correlation analyses, although the correlation between electricity consumption and migration rates has the highest value, this relationship cannot be said to be linear. The observed high value of the correlation between education level and electricity consumption (close to 1) supports that high education level may be related to electricity consumption. We obtained results that support our opinion stated in first sentence.

In the analyzes made on a regional basis, no significant relationship is seen. It is observed that each region is associated with different data sets at different levels compared to the other.

## Workforce and Usable Income

ggplot(region26, aes(x = workforce15plus, y = usableincome, label = region2id)) +  
geom\_point(color = "blue", size = 3) +   
geom\_text(vjust = -0.5, hjust = -0.5) +  
labs(title = "Workforce ve Usable Income", x = "Workforce", y = "Usable Income")

