The Relationship Between Educational Level and The Number of First Cousin Marriages

Welcome to my project page.

Keep an eye on this space to stay updated with my project activities.

1. Project Overview and Scope

In this project, I am investigating the relationship between the level of education and the change in first cousin marriages over time in 81 provinces of Turkey. Specifically, I am testing the hypothesis that provinces with higher secondary and higher education completion rates exhibit lower cousin marriage rates.

2. Data

2.1 Data Source

- **Cousin Marriage Data**

  Exported from the Turkish Statistical Institute database, this table contains, for each province and year, the total number of marriages, the number of marriages between first cousins, and the proportion of cousin marriages (%). Coverage: 2010-2024.

- **Education Level Data**:

  Also from TUIK, this table provides the count of individuals by education level (illiterate, literate without diploma, primary school, secondary school, high school & equivalents, universities, master, doctorate) and sex, for each province and year. Coverage: 2010-2023.

2.2 General Information About Data

- **Cousin Marriage Dataset**

  - Columns per year: total marriages, cousin marriages, and cousin-marriage proportion (%)

  - Rows: 81 provinces/14 years

- **Education Dataset**

  - Columns: year, province code & name, total population by education level and sex

- Education levels: illiterate, literate without diploma, primary, lower secondary, high school, universities, master, doctorate,unknown

- Rows: 81 provinces /8 education categories with 2 sexes / 15 years

2.3 Reason of Choice

chose these data sets because it is necessary to examine the level of education from both a cultural and social (inbreeding) point of view. Demonstrating an inverse relationship between education levels and cousin-marriage rates may support targeted interventions in areas with low education levels. Moreover, provincial granularity offers insights into regional policy needs.

2.4 Preprocessing

- **Loading & Header Processing**

  1. Read both Excel files using `readxl` to preserve raw header rows.

```r
library(readxl)
akraba_raw <- read_excel("akraba_evliligi.xlsx",
                         col_names = FALSE)
```

```
New names:
* `` -> `...1`
* `` -> `...2`
* `` -> `...3`
* `` -> `...4`
* `` -> `...5`
* `` -> `...6`
* `` -> `...7`
* `` -> `...8`
* `` -> `...9`
* `` -> `...10`
* `` -> `...11`
* `` -> `...12`
* `` -> `...13`
* `` -> `...14`
* `` -> `...15`
* `` -> `...16`
* `` -> `...17`
* `` -> `...18`
* `` -> `...19`
* `` -> `...20`
* `` -> `...21`
* `` -> `...22`
* `` -> `...23`
* `` -> `...24`
* `` -> `...25`
* `` -> `...26`
* `` -> `...27`
* `` -> `...28`
* `` -> `...29`
* `` -> `...30`
```

```
* `` -> `...31`
* `` -> `...32`
* `` -> `...33`
* `` -> `...34`
* `` -> `...35`
* `` -> `...36`
* `` -> `...37`
* `` -> `...38`
* `` -> `...39`
* `` -> `...40`
* `` -> `...41`
* `` -> `...42`
* `` -> `...43`
* `` -> `...44`
* `` -> `...45`
* `` -> `...46`
* `` -> `...47`
* `` -> `...48`
* `` -> `...49`
* `` -> `...50`
* `` -> `...51`
* `` -> `...52`
* `` -> `...53`
* `` -> `...54`
* `` -> `...55`
* `` -> `...56`
* `` -> `...57`
* `` -> `...58`
* `` -> `...59`
* `` -> `...60`
```

```r
egitim_raw <- read_excel("egitim_durumu.xlsx",
                         col_names = FALSE)
```

```
New names:
* `` -> `...1`
* `` -> `...2`
* `` -> `...3`
* `` -> `...4`
* `` -> `...5`
* `` -> `...6`
* `` -> `...7`
* `` -> `...8`
* `` -> `...9`
* `` -> `...10`
* `` -> `...11`
* `` -> `...12`
* `` -> `...13`
* `` -> `...14`
* `` -> `...15`
* `` -> `...16`
* `` -> `...17`
* `` -> `...18`
* `` -> `...19`
```

```
* `` -> `...20`
* `` -> `...21`
* `` -> `...22`
* `` -> `...23`
* `` -> `...24`
* `` -> `...25`
* `` -> `...26`
* `` -> `...27`
* `` -> `...28`
* `` -> `...29`
* `` -> `...30`
* `` -> `...31`
* `` -> `...32`
* `` -> `...33`
* `` -> `...34`
* `` -> `...35`
* `` -> `...36`
* `` -> `...37`
* `` -> `...38`
* `` -> `...39`
* `` -> `...40`
* `` -> `...41`
* `` -> `...42`
* `` -> `...43`
* `` -> `...44`
* `` -> `...45`
* `` -> `...46`
* `` -> `...47`
* `` -> `...48`
* `` -> `...49`
```

When the data obtained from TUIK are examined without manipulation, it is seen that its structure is not suitable for reading in R.

I have brought two data into tidy form by making Reshaping to make it tidy

- **Education Data Tidy**
  Clean headers and reshape education data into (`Province, Year, Education_level_sex,`)
  Cousin Marriage Data Tidy Similarly clean headers and reshape cousin marriage data into (`Province, Year, Number of marriages, Number of marriages between first cousins, Proportion of marriages between first cousins (%)`)

```
tidy_education <- read_excel("tidy_education.xlsx")

tidy_marriages <- read_excel("tidy_marriages.xlsx")
```

Data recorded as R.Data and column names are shown.

```
save(tidy_marriages, tidy_education, file = "data.RData")

# tidy_marriages
print(names(tidy_marriages))
```

```
[1] "Province"
[2] "Year"
[3] "Number_of_marriages"
[4] "Number_of_marriages_between_first_cousins"
[5] "Proportion"
```

```
# tidy_education
print(names(tidy_education))
```

```
 [1] "Province"                        "Year"
 [3] "Illiterate_Total"                "Illiterate_Male"
 [5] "Illiterate_Female"               "Literate_without_diploma_Total"
 [7] "Literate_without_diploma_Male"   "Literate_without_diploma_Female"
 [9] "Primary1_school_Total"           "Primary1_school_Male"
[11] "Primary1_school_Female"          "Primary2_school_Total"
[13] "Primary2_school_Male"            "Primary2_school_Female"
[15] "Lower_secondary_school_Total"    "Lower_secondary_school_Male"
[17] "Lower_secondary_school_Female"   "High_school_Total"
[19] "High_school_Male"                "High_school_Female"
[21] "Universities_Total"              "Universities_Male"
[23] "Universities_Female"             "Master_Total"
[25] "Master_Male"                     "Master_Female"
[27] "Doctorate_Total"                 "Doctorate_Male"
[29] "Doctorate_Female"                "Unknown_Total"
[31] "Unknown_Male"                    "Unknown_Female"
```

3. Analysis

3.1 Exploratory Data Analysis

```
marriage_years  <- sort(unique(tidy_marriages$Year))
education_years <- sort(unique(tidy_education$Year))

cat("Marriages data covers years:", marriage_years, "\n")
```

```
Marriages data covers years: 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
```

```
cat("Education data covers years:", education_years, "\n")
```

```
Education data covers years: 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
```

```
# --- Provinces ---
provinces <- sort(unique(tidy_marriages$Province))
cat("Number of provinces in marriages data:", length(provinces), "\n")
```

```
Number of provinces in marriages data: 82
```

```
cat("Sample provinces:", paste(head(provinces, 10), collapse = ", "), "...\n\n")
```

Sample provinces: Adana, Adıyaman, Afyonkarahisar, Ağrı, Aksaray, Amasya, Ankara, Antalya, Ard

```
# --- Education Levels  ---
edu_cols    <- names(tidy_education)[-(1:2)]
edu_levels <- unique(sub("_(Total|Male|Female)$", "", edu_cols))
cat("Education levels available:", paste(edu_levels, collapse = ", "), "\n")
```

Education levels available: Illiterate, Literate_without_diploma, Primary1_school, Primary2_sc

3.2 Trend Analysis

Tidy data has been combined according to city and year to use the data more effectively.

```
combined_data <- merge(tidy_education, tidy_marriages, by = c("Province", "Year"))
```

How the educational levels of people in Turkey change by year is shown with line graphs.

```
library(ggplot2)
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

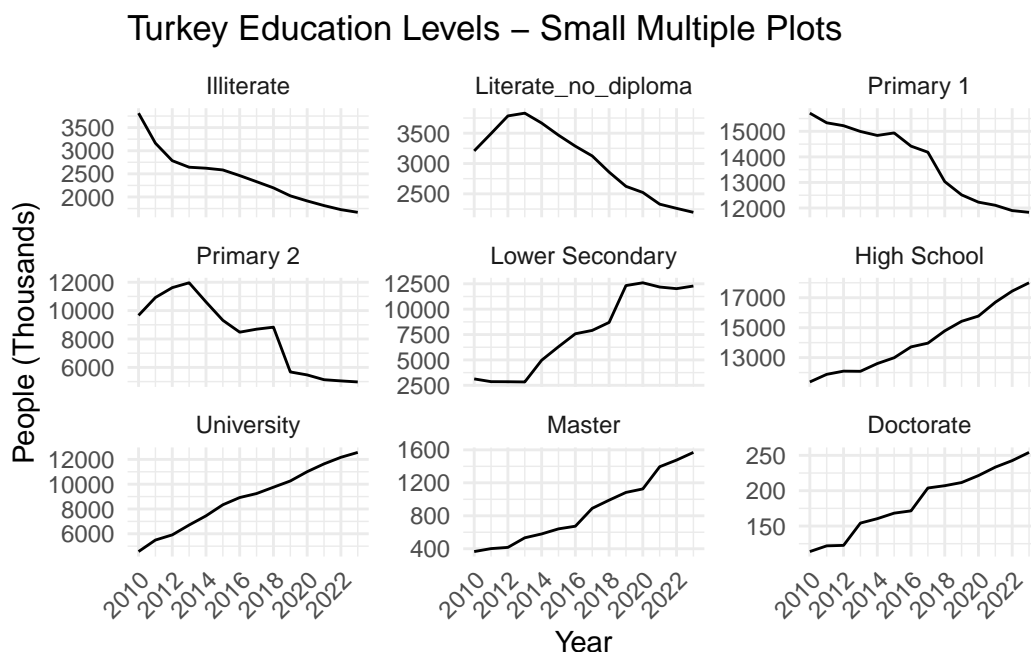   intersect, setdiff, setequal, union

```
library(tidyr)
library(stringi)

combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  select(Year, Illiterate_Total, Literate_without_diploma_Total,
         Primary1_school_Total, Primary2_school_Total,
         Lower_secondary_school_Total, High_school_Total,
         Universities_Total, Master_Total, Doctorate_Total) |>
  mutate(Year = as.numeric(Year)) |>
  pivot_longer(
    cols = -Year,
    names_to = "Education_Level",
    values_to = "People"
```

```
  ) |>
  filter(!is.na(People)) |>
  mutate(People = People / 1000) |>
  mutate(Education_Level = recode(Education_Level,
                          "Illiterate_Total" = "Illiterate",
                          "Literate_without_diploma_Total" = "Literate_no_diploma",
                          "Primary1_school_Total" = "Primary 1",
                          "Primary2_school_Total" = "Primary 2",
                          "Lower_secondary_school_Total" = "Lower Secondary",
                          "High_school_Total" = "High School",
                          "Universities_Total" = "University",
                          "Master_Total" = "Master",
                          "Doctorate_Total" = "Doctorate")) |>
  mutate(Education_Level = factor(Education_Level, levels = c(
    "Illiterate", "Literate_no_diploma", "Primary 1", "Primary 2",
    "Lower Secondary", "High School", "University", "Master", "Doctorate"
  ))) |>
  ggplot(aes(x = Year, y = People)) +
  geom_line() +
  facet_wrap(~ Education_Level, scales = "free_y", nrow = 3) +
  scale_x_continuous(breaks = seq(2010, 2024, by = 2)) +
  labs(title = "Turkey Education Levels – Small Multiple Plots",
       x = "Year", y = "People (Thousands)") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```


Turkey Education Levels – Small Multiple Plots

In the same way, it was shown how the percentage of first cousin marriages in Turkey varies by year.

```
library(ggplot2)
library(dplyr)
library(stringi)
```

```
combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  mutate(Year = as.numeric(Year)) |>
  ggplot(aes(x = Year, y = Proportion)) +
  geom_point(color = "steelblue") +
  geom_text(aes(label = paste0(round(Proportion, 2), "%")),hjust=-0.3, vjust =-0.3, size = 3)
  scale_x_continuous(breaks = seq(2010, 2024, by = 2)) +
  labs(title = "Turkey - First Cousin Marriage Percentage Over Years",
       x = "Year", y = "Percentage") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```
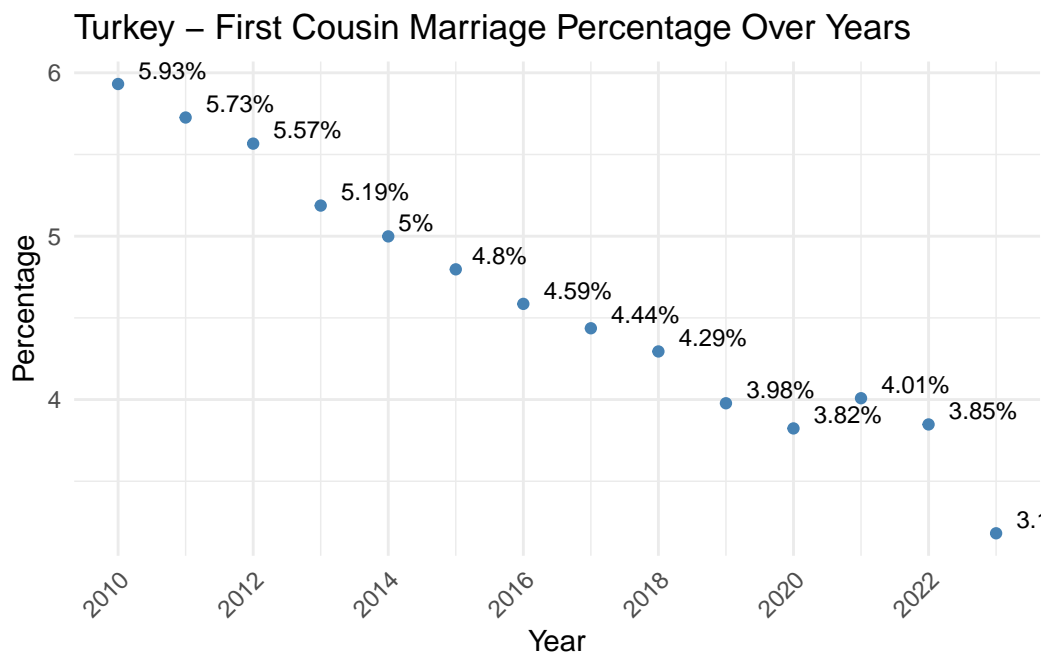

The increase after **2020** is thought to be because people could not socialize after the global pandemic and spent time with their closest acquaintances.

3.3 Model Fitting

To observe the compatibility of the two data models, first of all, the data according to the increase in the level of education for Turkey.

3.3.1 Illiterate Population & Cousin Marriage Rate

The data suggests a possible positive association between the proportion of illiterate individuals and cousin marriage rates in Turkey. As literacy improves over time, a downward shift in consanguineous marriage patterns can be observed.

```r
edu_var <- "Illiterate_Total"

plot_data <- combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  mutate(Year = as.numeric(Year),
         Education_k = .data[[edu_var]] / 1000,
         Proportion_percent = Proportion) |>
  select(Year, Education_k, Proportion_percent)


min_y <- min(plot_data$Education_k)
max_y <- max(plot_data$Education_k)

plot_data <- plot_data |>
  mutate(Proportion_scaled = ((Proportion_percent - min(Proportion_percent)) /
                              (max(Proportion_percent) - min(Proportion_percent))) *
                             (max_y - min_y) + min_y)


ggplot(plot_data, aes(x = Year)) +
  geom_line(aes(y = Education_k, color = "Illiterate (1000s)"), size = 1.2) +
  geom_point(aes(y = Proportion_scaled, color = "Cousin Marriage (%)"), size = 2.5) +
  geom_text(aes(y = Proportion_scaled, label = paste0(round(Proportion_percent, 2), "%")),
            vjust = -0.8, size = 3) +
  scale_x_continuous(breaks = seq(2010, 2024, 2)) +
  scale_y_continuous(
    name = "Illiterate Individuals (1000s)",
    sec.axis = sec_axis(
      trans = ~ (.-min_y) * (max(plot_data$Proportion_percent) - min(plot_data$Proportion_perc
                  (max_y - min_y) + min(plot_data$Proportion_percent),
      name = "Cousin Marriage (%)"
    )
  ) +
  labs(
    title = "Illiterate Population & Cousin Marriage Rate (Turkey)",
    x = "Year", color = ""
  ) +
  theme_minimal() +
  theme(legend.position = "top", axis.text.x = element_text(angle = 45, hjust = 1))
```
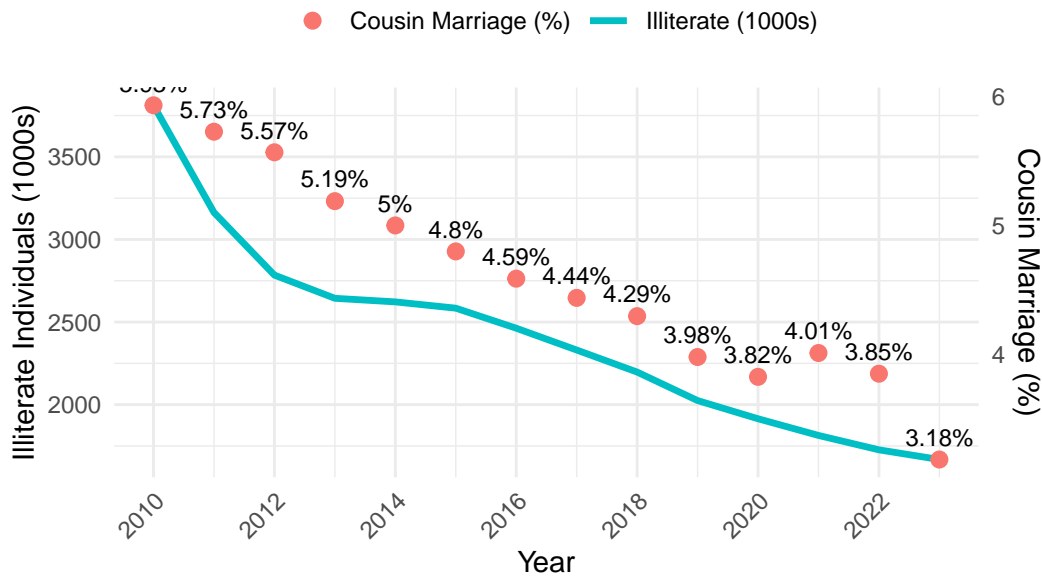
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.


Warning: The `trans` argument of `sec_axis()` is deprecated as of ggplot2 3.5.0.
i Please use the `transform` argument instead.
```

Illiterate Population & Cousin Marriage Rate (Turkey)

3.3.2 Primary School Graduates & Cousin Marriage (%)

The expansion of primary education in Turkey has played a key role in shaping social behaviors. This chart shows that increases in primary school graduation are associated with reductions in cousin marriage rates over time.

```r
plot_data <- combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  mutate(
    Year = as.numeric(Year),
    Education_k = (Primary1_school_Total + Primary2_school_Total) / 1000,
    Proportion_percent = Proportion
  ) |>
  select(Year, Education_k, Proportion_percent)

min_y <- min(plot_data$Education_k)
max_y <- max(plot_data$Education_k)

plot_data <- plot_data |>
  mutate(Proportion_scaled = ((Proportion_percent - min(Proportion_percent)) /
                             (max(Proportion_percent) - min(Proportion_percent))) *
                             (max_y - min_y) + min_y)


ggplot(plot_data, aes(x = Year)) +
  geom_line(aes(y = Education_k, color = "Primary School (1000s)"), size = 1.2) +
  geom_point(aes(y = Proportion_scaled, color = "Cousin Marriage (%)"), size = 2.5) +
  geom_text(aes(y = Proportion_scaled, label = paste0(round(Proportion_percent, 2), "%")),
            vjust = -0.8, size = 3) +
  scale_x_continuous(breaks = seq(2010, 2024, 2)) +
  scale_y_continuous(
```

```
    name = "Primary School Graduates (1000s)",
    sec.axis = sec_axis(
      trans = ~ (.-min_y) * (max(plot_data$Proportion_percent) - min(plot_data$Proportion_perc
                     (max_y - min_y) + min(plot_data$Proportion_percent),
      name = "Cousin Marriage (%)"
    )
  ) +
  labs(
    title = "Primary School Graduates & Cousin Marriage Rate (Turkey)",
    x = "Year", color = ""
  ) +
  theme_minimal() +
  theme(legend.position = "top", axis.text.x = element_text(angle = 45, hjust = 1))
```


Primary School Graduates & Cousin Marriage Rate (Turkey)
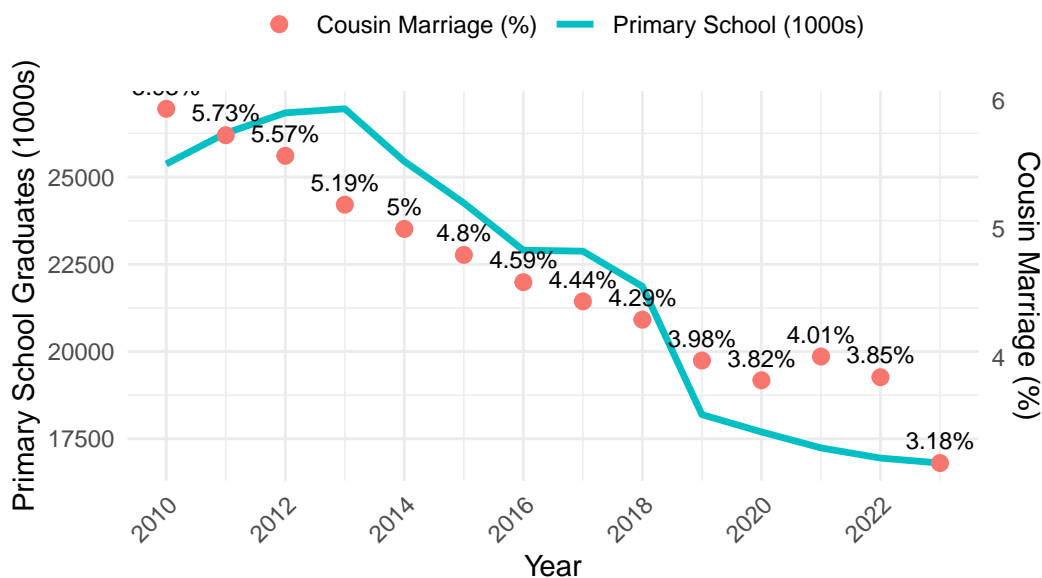
3.3.3 Lower Secondary School Graduates & Cousin Marriage (%)

In Turkey, the transition to the 4+4+4 education model in the 2012-2013 academic year explains the acceleration increase in the graph.

```
edu_var <- "Lower_secondary_school_Total"

plot_data <- combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  mutate(
    Year = as.numeric(Year),
    Education_k = .data[[edu_var]] / 1000,
    Proportion_percent = Proportion
  ) |>
  select(Year, Education_k, Proportion_percent)
```

```
min_y <- min(plot_data$Education_k)
max_y <- max(plot_data$Education_k)

plot_data <- plot_data |>
  mutate(Proportion_scaled = ((Proportion_percent - min(Proportion_percent)) /
                              (max(Proportion_percent) - min(Proportion_percent))) *
                             (max_y - min_y) + min_y)

ggplot(plot_data, aes(x = Year)) +
  geom_line(aes(y = Education_k, color = "Lower Secondary School (1000s)"), size = 1.2) +
  geom_point(aes(y = Proportion_scaled, color = "Cousin Marriage (%)"), size = 2.5) +
  geom_text(aes(y = Proportion_scaled, label = paste0(round(Proportion_percent, 2), "%")),
            vjust = -0.8, size = 3) +
  scale_x_continuous(breaks = seq(2010, 2024, 2)) +
  scale_y_continuous(
    name = "Lower Secondary School Graduates (1000s)",
    sec.axis = sec_axis(
      trans = ~ (.-min_y) * (max(plot_data$Proportion_percent) - min(plot_data$Proportion_perc
                   (max_y - min_y) + min(plot_data$Proportion_percent),
      name = "Cousin Marriage (%)"
    )
  ) +
  labs(
    title = "Lower Secondary School Graduates & Cousin Marriage Rate (Turkey)",
    x = "Year", color = ""
  ) +
  theme_minimal() +
  theme(legend.position = "top", axis.text.x = element_text(angle = 45, hjust = 1))
```
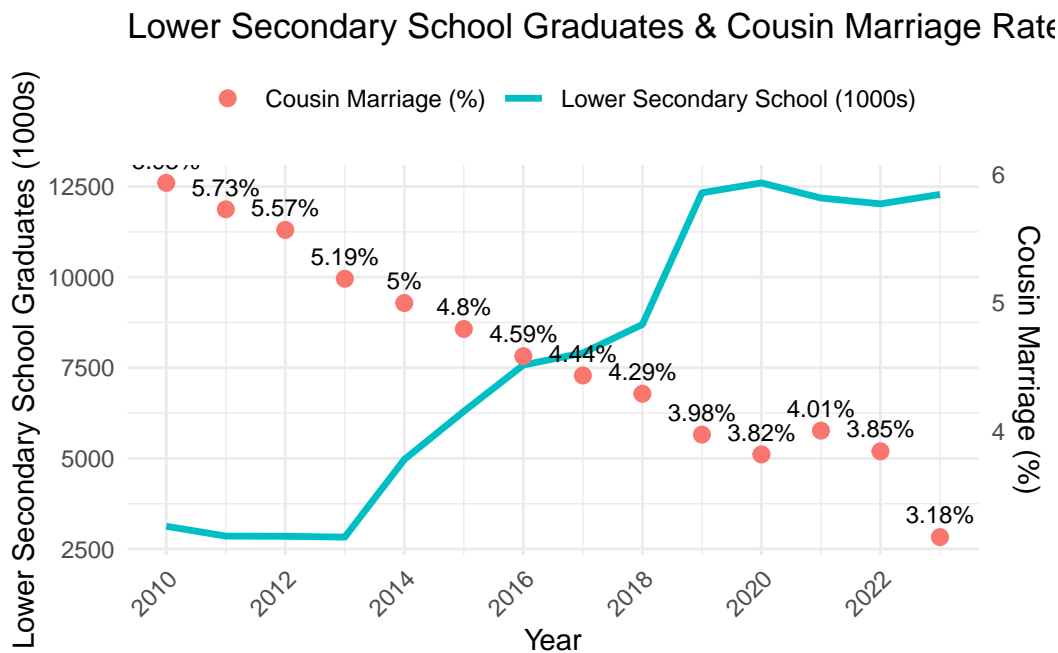

Lower Secondary School Graduates & Cousin Marriage Rate (Turkey)

3.3.4 High School Graduates& Cousin Marriage (%)

In Turkey, as the number of high school graduates has increased steadily over the years, the rate of cousin marriages has shown a noticeable decline. This trend implies a possible inverse relationship

between secondary education completion and consanguineous marriage practices.

```
edu_var <- "High_school_Total"

plot_data <- combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  mutate(
    Year = as.numeric(Year),
    Education_k = .data[[edu_var]] / 1000,
    Proportion_percent = Proportion
  ) |>
  select(Year, Education_k, Proportion_percent)

min_y <- min(plot_data$Education_k)
max_y <- max(plot_data$Education_k)

plot_data <- plot_data |>
  mutate(Proportion_scaled = ((Proportion_percent - min(Proportion_percent)) /
                              (max(Proportion_percent) - min(Proportion_percent))) *
                             (max_y - min_y) + min_y)

ggplot(plot_data, aes(x = Year)) +
  geom_line(aes(y = Education_k, color = "High School (1000s)"), size = 1.2) +
  geom_point(aes(y = Proportion_scaled, color = "Cousin Marriage (%)"), size = 2.5) +
  geom_text(aes(y = Proportion_scaled, label = paste0(round(Proportion_percent, 2), "%")),
            vjust = -0.8, size = 3) +
  scale_x_continuous(breaks = seq(2010, 2024, 2)) +
  scale_y_continuous(
    name = "High School Graduates (1000s)",
    sec.axis = sec_axis(
      trans = ~ (.-min_y) * (max(plot_data$Proportion_percent) - min(plot_data$Proportion_perc
                  (max_y - min_y) + min(plot_data$Proportion_percent),
      name = "Cousin Marriage (%)"
    )
  ) +
  labs(
    title = "High School Graduates & Cousin Marriage Rate (Turkey)",
    x = "Year", color = ""
  ) +
  theme_minimal() +
  theme(legend.position = "top", axis.text.x = element_text(angle = 45, hjust = 1))
```
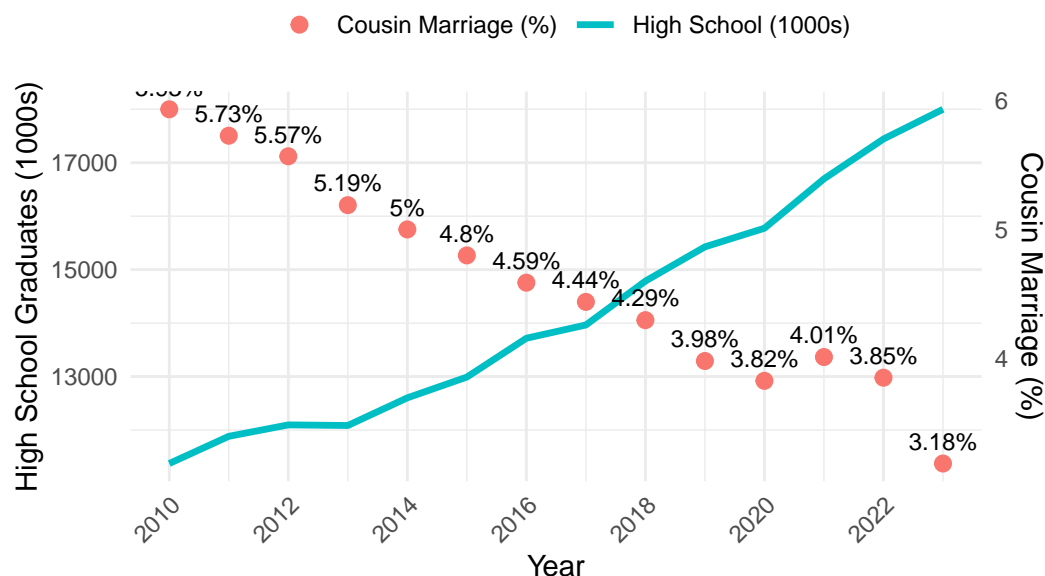
High School Graduates & Cousin Marriage Rate (Turkey)

3.5.5 Universities Graduates & Cousin Marriage (%)

Over the years, as the number of university graduates in Turkey increased, a steady decline in the rate of cousin marriages can be observed. This trend suggests a potential negative correlation between higher education attainment and consanguineous marriage preferences.

```r
edu_var <- "Universities_Total"

plot_data <- combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  mutate(
    Year = as.numeric(Year),
    Education_k = .data[[edu_var]] / 1000,
    Proportion_percent = Proportion
  ) |>
  select(Year, Education_k, Proportion_percent)

min_y <- min(plot_data$Education_k)
max_y <- max(plot_data$Education_k)

plot_data <- plot_data |>
  mutate(Proportion_scaled = ((Proportion_percent - min(Proportion_percent)) /
                              (max(Proportion_percent) - min(Proportion_percent))) *
                             (max_y - min_y) + min_y)

ggplot(plot_data, aes(x = Year)) +
  geom_line(aes(y = Education_k, color = "University (1000s)"), size = 1.2) +
  geom_point(aes(y = Proportion_scaled, color = "Cousin Marriage (%)"), size = 2.5) +
  geom_text(aes(y = Proportion_scaled, label = paste0(round(Proportion_percent, 2), "%")),
            vjust = -0.8, size = 3) +
  scale_x_continuous(breaks = seq(2010, 2024, 2)) +
```

```
scale_y_continuous(
  name = "University Graduates (1000s)",
  sec.axis = sec_axis(
    trans = ~ (.-min_y) * (max(plot_data$Proportion_percent) - min(plot_data$Proportion_perc
                   (max_y - min_y) + min(plot_data$Proportion_percent),
    name = "Cousin Marriage (%)"
  )
) +
labs(
  title = "University Graduates & Cousin Marriage Rate (Turkey)",
  x = "Year", color = ""
) +
theme_minimal() +
theme(legend.position = "top", axis.text.x = element_text(angle = 45, hjust = 1))
```

University Graduates & Cousin Marriage Rate (Turkey)
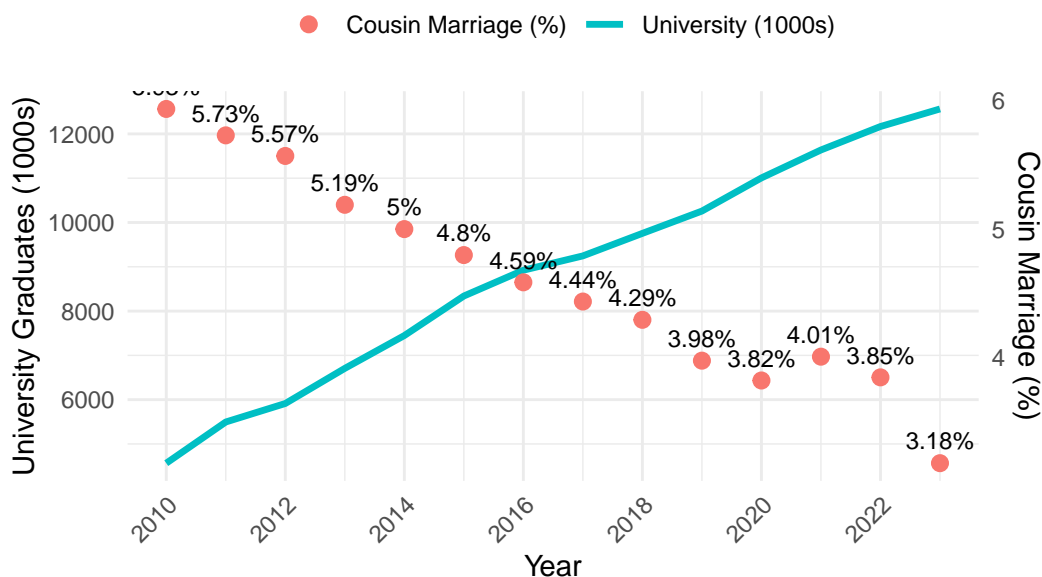
3.5.5 Masters Degree Graduates & Cousin Marriage (%)

A rising trend in master???s degree graduations appears to align with a gradual decrease in cousin marriage rates. This relationship may indicate that individuals with postgraduate education are less likely to engage in consanguineous marriages.

```
edu_var <- "Master_Total"

plot_data <- combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  mutate(
    Year = as.numeric(Year),
    Education_k = .data[[edu_var]] / 1000,
    Proportion_percent = Proportion
  ) |>
```

```
    select(Year, Education_k, Proportion_percent)

min_y <- min(plot_data$Education_k)
max_y <- max(plot_data$Education_k)

plot_data <- plot_data |>
  mutate(Proportion_scaled = ((Proportion_percent - min(Proportion_percent)) /
                                (max(Proportion_percent) - min(Proportion_percent))) *
                               (max_y - min_y) + min_y)

ggplot(plot_data, aes(x = Year)) +
  geom_line(aes(y = Education_k, color = "Master's Degree (1000s)"), size = 1.2) +
  geom_point(aes(y = Proportion_scaled, color = "Cousin Marriage (%)"), size = 2.5) +
  geom_text(aes(y = Proportion_scaled, label = paste0(round(Proportion_percent, 2), "%")),
            vjust = -0.8, size = 3) +
  scale_x_continuous(breaks = seq(2010, 2024, 2)) +
  scale_y_continuous(
    name = "Master's Degree Graduates (1000s)",
    sec.axis = sec_axis(
      trans = ~ (.-min_y) * (max(plot_data$Proportion_percent) - min(plot_data$Proportion_perc
                    (max_y - min_y) + min(plot_data$Proportion_percent),
      name = "Cousin Marriage (%)"
    )
  ) +
  labs(
    title = "Master's Degree Graduates & Cousin Marriage Rate (Turkey)",
    x = "Year", color = ""
  ) +
  theme_minimal() +
  theme(legend.position = "top", axis.text.x = element_text(angle = 45, hjust = 1))
```
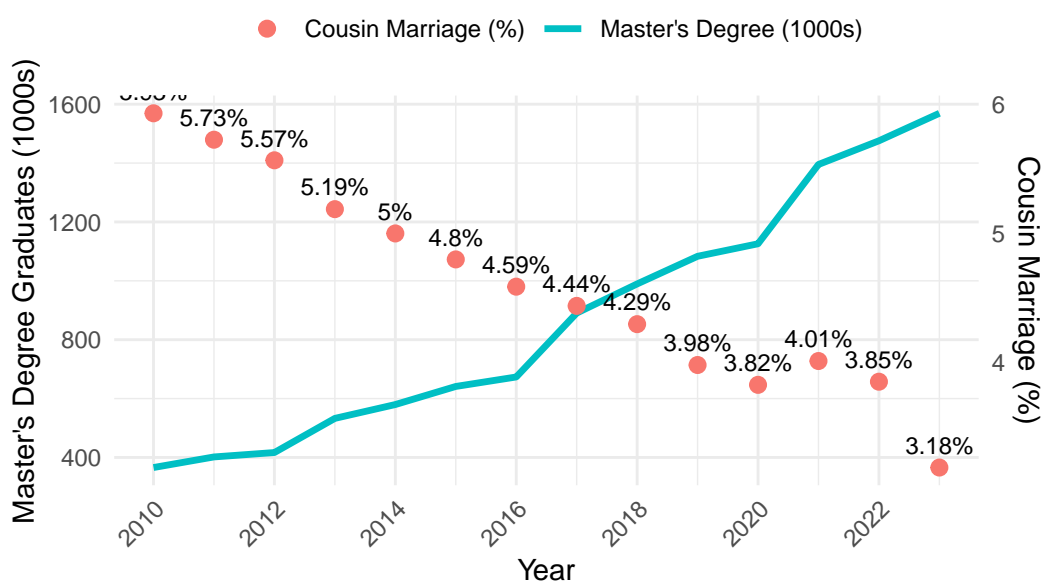

Master's Degree Graduates & Cousin Marriage Rate (Turkey)

3.5.6 Doctorate Graduates & Cousin Marriage (%)

As the number of individuals attaining doctoral degrees increases, cousin marriage rates show a significant downward trend. This supports the hypothesis that advanced education is inversely associated with traditional marriage practices such as consanguinity.

```
edu_var <- "Doctorate_Total"

plot_data <- combined_data |>
  mutate(Province = stri_trans_general(Province, "Latin-ASCII"),
         Province = tolower(Province)) |>
  filter(Province == "turkiye") |>
  mutate(
    Year = as.numeric(Year),
    Education_k = .data[[edu_var]] / 1000,
    Proportion_percent = Proportion
  ) |>
  select(Year, Education_k, Proportion_percent)

min_y <- min(plot_data$Education_k)
max_y <- max(plot_data$Education_k)

plot_data <- plot_data |>
  mutate(Proportion_scaled = ((Proportion_percent - min(Proportion_percent)) /
                               (max(Proportion_percent) - min(Proportion_percent))) *
                              (max_y - min_y) + min_y)

ggplot(plot_data, aes(x = Year)) +
  geom_line(aes(y = Education_k, color = "Doctorate (1000s)"), size = 1.2) +
  geom_point(aes(y = Proportion_scaled, color = "Cousin Marriage (%)"), size = 2.5) +
  geom_text(aes(y = Proportion_scaled, label = paste0(round(Proportion_percent, 2), "%")),
            vjust = -0.8, size = 3) +
  scale_x_continuous(breaks = seq(2010, 2024, 2)) +
  scale_y_continuous(
    name = "Doctorate Graduates (1000s)",
    sec.axis = sec_axis(
      trans = ~ (.-min_y) * (max(plot_data$Proportion_percent) - min(plot_data$Proportion_perc
                     (max_y - min_y) + min(plot_data$Proportion_percent),
      name = "Cousin Marriage (%)"
    )
  ) +
  labs(
    title = "Doctorate Graduates & Cousin Marriage Rate (Turkey)",
    x = "Year", color = ""
  ) +
  theme_minimal() +
  theme(legend.position = "top", axis.text.x = element_text(angle = 45, hjust = 1))
```
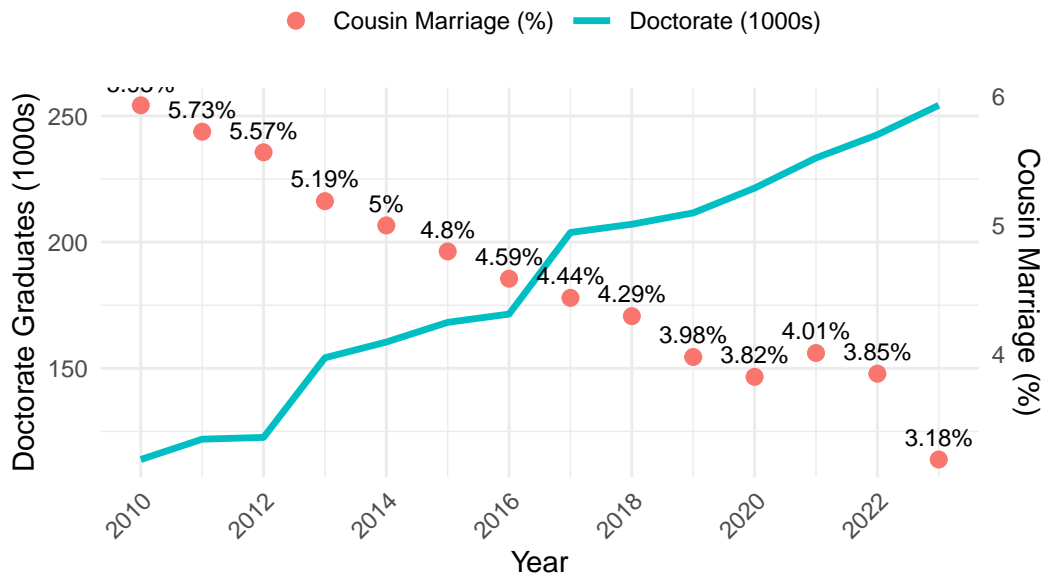
Doctorate Graduates & Cousin Marriage Rate (Turkey)

3.5.7 Heat Maps of Education Levels in Turkey

The animated maps below illustrate the spatial distribution of high school and higher education graduates across Turkish provinces over time. Color gradients represent the number of graduates (in thousands), while the labels highlight the top five and bottom five provinces each year.

3.5.7.1 High School Heat Map

```r
library(readxl)
library(dplyr)
library(stringi)
library(ggplot2)
library(sf)
```

```
Linking to GEOS 3.13.1, GDAL 3.10.2, PROJ 9.5.1; sf_use_s2() is TRUE
```

```r
library(gganimate)
```

```r
turkiye <- st_read("turkey-geojson.json") |>
  mutate(il_adi = tolower(name),
         il_adi = stri_trans_general(il_adi, "Latin-ASCII")) |>
  mutate(il_adi = case_when(
    il_adi == "afyon" ~ "afyonkarahisar",
    il_adi == "eskisehir" ~ "eskisehir",
    il_adi == "mugla" ~ "mugla",
    il_adi == "sanliurfa" ~ "sanliurfa",
    il_adi == "kirikkale" ~ "kirikkale",
    TRUE ~ il_adi
  ))
```

```
Reading layer `turkey-geojson' from data source
  `C:\Users\yagmu\Desktop\emu660-spring2025-yyaslan\turkey-geojson.json'
  using driver `GeoJSON'
Simple feature collection with 81 features and 1 field
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 25.66514 ymin: 35.81543 xmax: 44.83384 ymax: 42.10541
Geodetic CRS:  WGS 84
```

```r
edu <- read_excel("tidy_education.xlsx") |>
  mutate(
    Province = stri_trans_general(Province, "Latin-ASCII"),
    Province = tolower(trimws(Province))
  ) |>
  filter(Province != "turkiye") |>
  select(Year, il_adi = Province, mezun = High_school_Total) |>
  mutate(mezun = mezun / 1000)

harita_df <- expand.grid(il_adi = unique(turkiye$il_adi), Year = unique(edu$Year)) |>
  left_join(edu, by = c("il_adi", "Year")) |>
  left_join(turkiye, by = "il_adi") |>
  st_as_sf()

etiketler <- harita_df |>
  group_by(Year) |>
  mutate(
    sira = rank(mezun, ties.method = "first"),
    etiket = case_when(
      sira <= 5 | sira >= (n() - 4) ~ paste0(round(mezun), "k"),
      TRUE ~ NA_character_
    ),
    x = st_coordinates(st_centroid(geometry))[,1],
    y = st_coordinates(st_centroid(geometry))[,2]
  ) |>
  ungroup() |>
  mutate(
    x = ifelse(il_adi == "istanbul", x + 0.7, x),
    y = ifelse(il_adi == "istanbul", y + 0.3, y)
  )


p_high <- ggplot(etiketler) +
  geom_sf(aes(fill = mezun, geometry = geometry), color = "white", linewidth = 0.2) +
  geom_text(data = subset(etiketler, !is.na(etiket)),
            aes(x = x, y = y, label = etiket),
            size = 3, color = "black") +
  scale_fill_viridis_c(option = "plasma", direction = -1, na.value = "grey90") +
  labs(
    title = "{closest_state} High School Graduates (in Thousand)",
    fill = "Thousand People"
  ) +
  theme_minimal() +
  transition_states(Year, transition_length = 2, state_length = 1) +
```
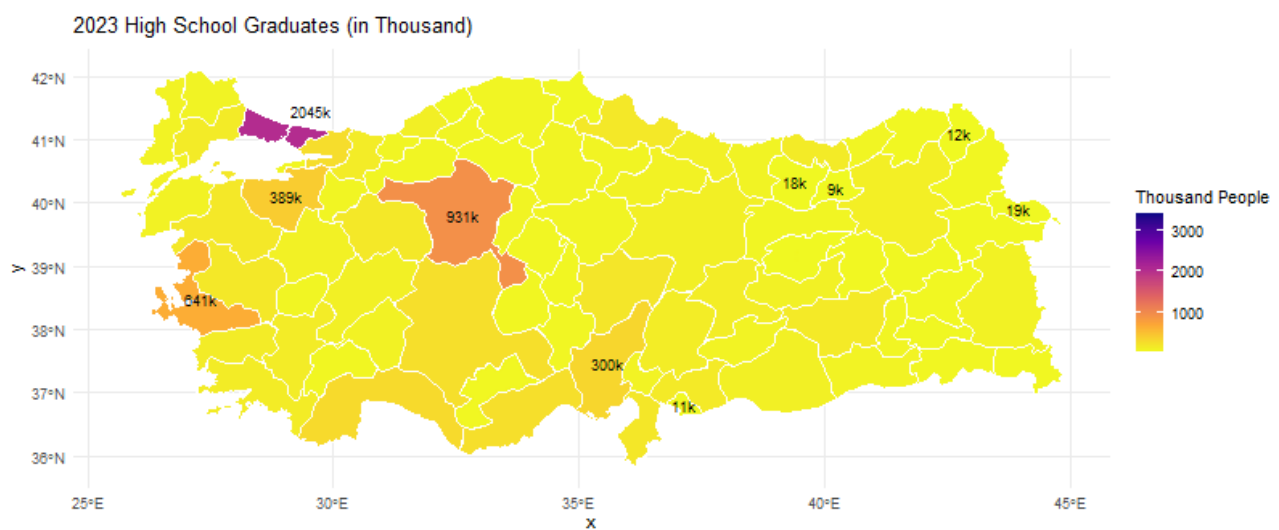
```
    ease_aes("cubic-in-out")


anim_save("lise.gif", animate(p_high, width = 800, height = 600, fps = 2, duration = 10))
library(magick)
image_read("lise.gif") %>%
  .[length(.)] %>%  # son kare
  image_convert(format = "png", type = "truecolor") %>%
  image_write("lise.png")
```


3.5.7.2 University Heat Map

```
edu_uni <- read_excel("tidy_education.xlsx") |>
  mutate(
    Province = stri_trans_general(Province, "Latin-ASCII"),
    Province = tolower(trimws(Province))
  ) |>
  filter(Province != "turkiye") |>
  select(Year, il_adi = Province, mezun = Universities_Total) |>
  mutate(mezun = mezun / 1000)


harita_uni <- expand.grid(il_adi = unique(turkiye$il_adi), Year = unique(edu_uni$Year)) |>
  left_join(edu_uni, by = c("il_adi", "Year")) |>
  left_join(turkiye, by = "il_adi") |>
```
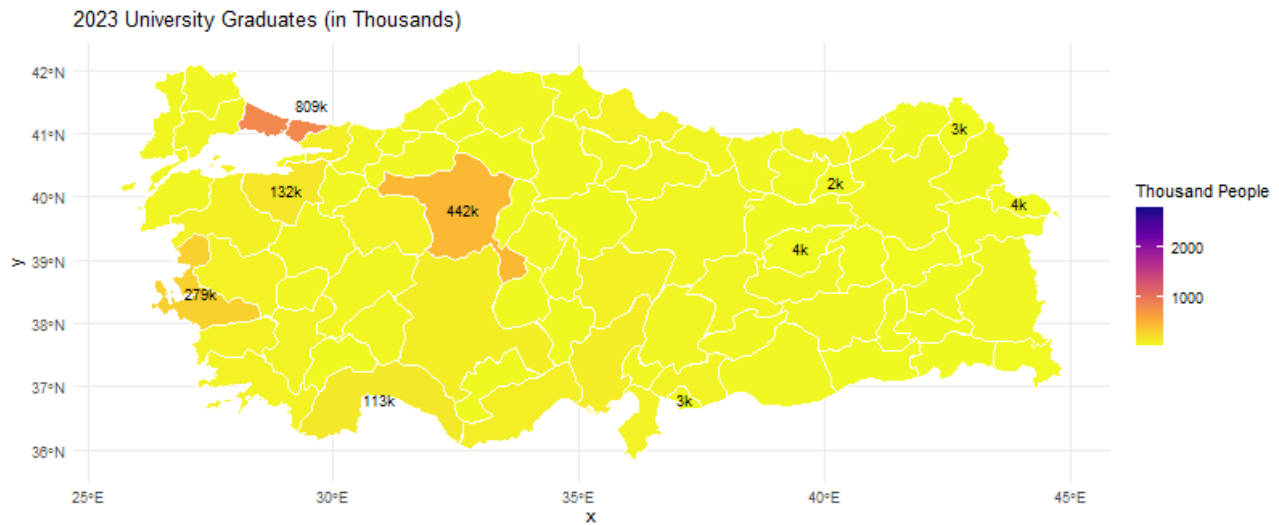
```r
  st_as_sf()


etiketler_uni <- harita_uni |>
  group_by(Year) |>
  mutate(
    sira = rank(mezun, ties.method = "first"),
    etiket = case_when(
      sira <= 5 | sira >= (n() - 4) ~ paste0(round(mezun), "k"),
      TRUE ~ NA_character_
    ),
    x = st_coordinates(st_centroid(geometry))[,1],
    y = st_coordinates(st_centroid(geometry))[,2],
    x = ifelse(il_adi == "istanbul", x + 0.7, x),
    y = ifelse(il_adi == "istanbul", y + 0.3, y)
  ) |>
  ungroup()


p_uni <- ggplot(etiketler_uni) +
  geom_sf(aes(fill = mezun, geometry = geometry), color = "white", linewidth = 0.2) +
  geom_text(data = subset(etiketler_uni, !is.na(etiket)),
            aes(x = x, y = y, label = etiket),
            size = 3, color = "black") +
  scale_fill_viridis_c(option = "plasma", direction = -1, na.value = "grey90") +
  labs(
    title = "{closest_state} University Graduates (in Thousands)",
    fill = "Thousand People"
  ) +
  theme_minimal() +
  transition_states(Year, transition_length = 2, state_length = 1) +
  ease_aes("cubic-in-out")


anim_save("uni.gif", animate(p_uni, width = 800, height = 600, fps = 2, duration = 10))
library(magick)
image_read("uni.gif") %>%
  .[length(.)] %>%
  image_convert(format = "png", type = "truecolor") %>%
  image_write("uni.png")
```

2023 University Graduates (in Thousands)

3.5.7.3 Master's Heat Map

```
edu_master <- read_excel("tidy_education.xlsx") |>
  mutate(
    Province = stri_trans_general(Province, "Latin-ASCII"),
    Province = tolower(trimws(Province))
  ) |>
  filter(Province != "turkiye") |>
  select(Year, il_adi = Province, mezun = Master_Total) |>
  mutate(mezun = mezun / 1000)

harita_master <- expand.grid(il_adi = unique(turkiye$il_adi), Year = unique(edu_master$Year)) |>
  left_join(edu_master, by = c("il_adi", "Year")) |>
  left_join(turkiye, by = "il_adi") |>
  st_as_sf()

etiketler_master <- harita_master |>
  group_by(Year) |>
  mutate(
    sira = rank(mezun, ties.method = "first"),
      etiket = case_when(
  sira <= 5 | sira >= (n() - 4) ~ ifelse(mezun < 1,
                                      as.character(round(mezun * 1000)),
                                      paste0(round(mezun), "k")),
  TRUE ~ NA_character_
```

```r
  ),
    x = st_coordinates(st_centroid(geometry))[,1],
    y = st_coordinates(st_centroid(geometry))[,2],
    x = ifelse(il_adi == "istanbul", x + 0.7, x),
    y = ifelse(il_adi == "istanbul", y + 0.3, y)
  ) |>
  ungroup()

p_master <- ggplot(etiketler_master) +
  geom_sf(aes(fill = mezun, geometry = geometry), color = "white", linewidth = 0.2) +
  geom_text(data = subset(etiketler_master, !is.na(etiket)),
            aes(x = x, y = y, label = etiket),
            size = 3, color = "black") +
  scale_fill_viridis_c(option = "plasma", direction = -1, na.value = "grey90") +
  labs(
    title = "{closest_state} Masters Graduates (in Thousands)",
    fill = "Thousand People"
  ) +
  theme_minimal() +
  transition_states(Year, transition_length = 2, state_length = 1) +
  ease_aes("cubic-in-out")
```

```r
anim_save("master.gif", animate(p_master, width = 800, height = 600, fps = 2, duration = 10))
library(magick)
image_read("master.gif") %>%
  .[length(.)] %>%
  image_convert(format = "png", type = "truecolor") %>%
  image_write("master.png")
```

2023 Masters Graduates (in Thousands)
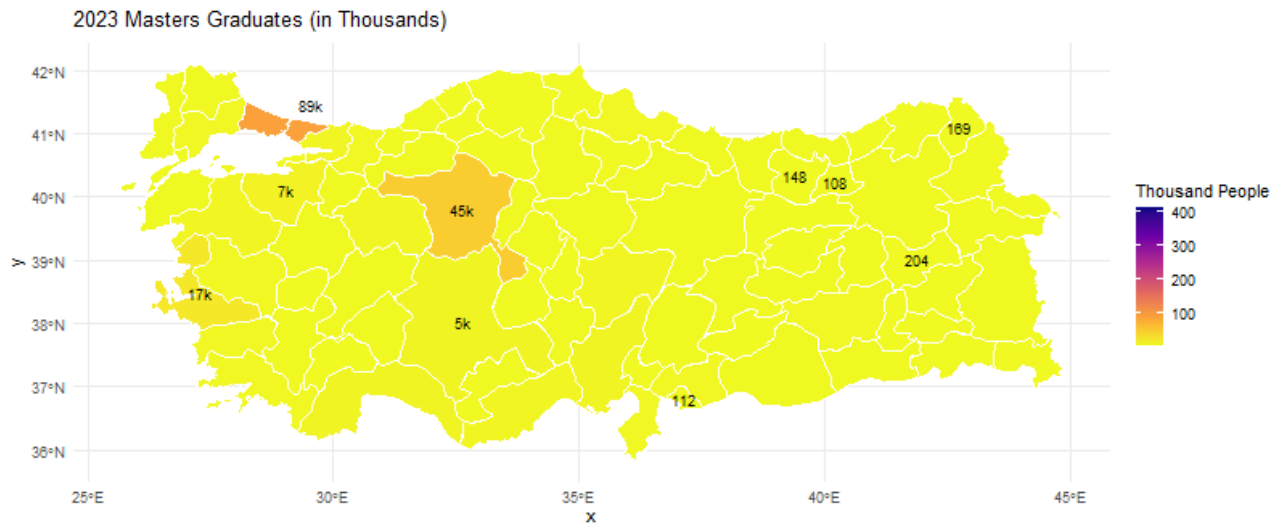
3.5.7.4 Doctorate Heat Map

```
edu_phd <- read_excel("tidy_education.xlsx") |>
  mutate(
    Province = stri_trans_general(Province, "Latin-ASCII"),
    Province = tolower(trimws(Province))
  ) |>
  filter(Province != "turkiye") |>
  select(Year, il_adi = Province, mezun = Doctorate_Total) |>
  mutate(mezun = mezun / 1000)

harita_phd <- expand.grid(il_adi = unique(turkiye$il_adi), Year = unique(edu_phd$Year)) |>
  left_join(edu_phd, by = c("il_adi", "Year")) |>
  left_join(turkiye, by = "il_adi") |>
  st_as_sf()

etiketler_phd <- harita_phd |>
  group_by(Year) |>
  mutate(
    sira = rank(mezun, ties.method = "first"),
    etiket = case_when(
  sira <= 5 | sira >= (n() - 4) ~ ifelse(mezun < 1,
                                    as.character(round(mezun * 1000)),
                                    paste0(round(mezun), "k")),
  TRUE ~ NA_character_
```

```r
  ),
    x = st_coordinates(st_centroid(geometry))[,1],
    y = st_coordinates(st_centroid(geometry))[,2],
    x = ifelse(il_adi == "istanbul", x + 0.7, x),
    y = ifelse(il_adi == "istanbul", y + 0.3, y)
  ) |>
  ungroup()

p_phd <- ggplot(etiketler_phd) +
  geom_sf(aes(fill = mezun, geometry = geometry), color = "white", linewidth = 0.2) +
  geom_text(data = subset(etiketler_phd, !is.na(etiket)),
            aes(x = x, y = y, label = etiket),
            size = 3, color = "black") +
  scale_fill_viridis_c(option = "plasma", direction = -1, na.value = "grey90") +
  labs(
    title = "{closest_state} Doctorate Graduates (in Thousands)",
    fill = "Thousand People"
  ) +
  theme_minimal() +
  transition_states(Year, transition_length = 2, state_length = 1) +
  ease_aes("cubic-in-out")
```

```r
anim_save("phd.gif", animate(p_phd, width = 800, height = 600, fps = 2, duration = 10))
library(magick)
image_read("phd.gif") %>%
  .[length(.)] %>%
  image_convert(format = "png", type = "truecolor") %>%
  image_write("phd.png")
```

2023 Doctorate Graduates (in Thousands)
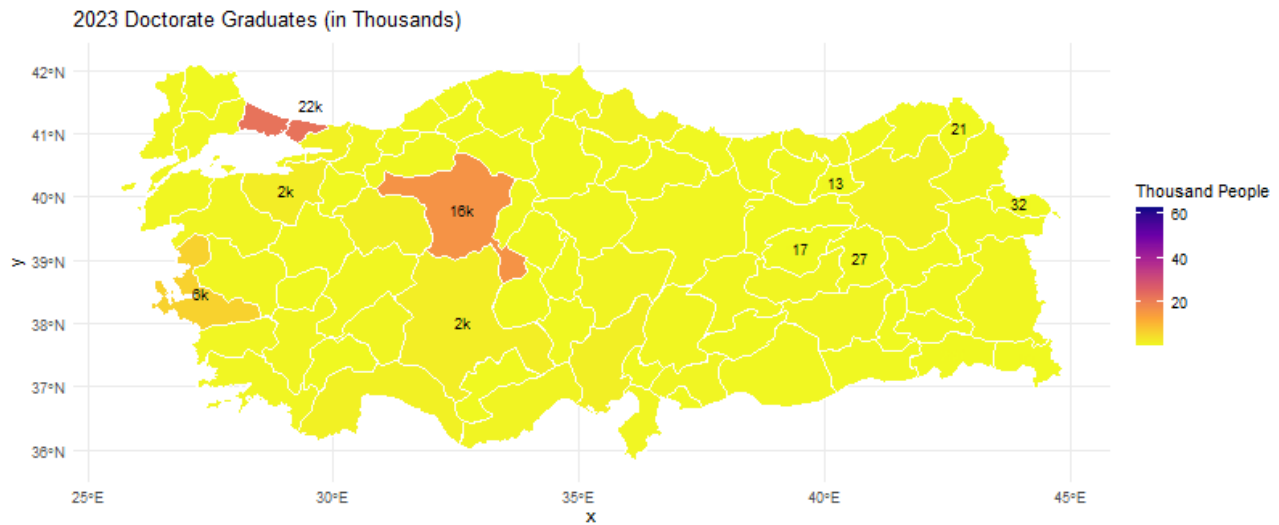
3.5.8 Heat Maps of Cousin Marriage in Turkey

This animated map shows how cousin marriage rates change across provinces in Turkey over the years. Darker colors mean higher rates. We can see that eastern and southeastern provinces usually have the highest rates, while western provinces have much lower ones.

```r
harita_evlilik <- expand.grid(il_adi = unique(turkiye$il_adi), Year = unique(combined_data$Yea
  left_join(combined_data |>
            mutate(il_adi = tolower(stri_trans_general(Province, "Latin-ASCII"))),
          by = c("il_adi", "Year")) |>
  left_join(turkiye, by = "il_adi") |>
  st_as_sf()

etiketler_evlilik <- harita_evlilik |>
  group_by(Year) |>
  mutate(
    sira = rank(Proportion, ties.method = "first"),
   etiket = case_when(
  sira <= 5 | sira >= (n() - 4) ~ paste0(round(Proportion, 1), "%"),
  TRUE ~ NA_character_
),

    x = st_coordinates(st_centroid(geometry))[,1],
    y = st_coordinates(st_centroid(geometry))[,2],
    x = ifelse(il_adi == "istanbul", x + 1.5, x),
```

```r
    y = ifelse(il_adi == "istanbul", y + 1.0, y)
  ) |>
  ungroup()

p_evlilik <- ggplot(etiketler_evlilik) +
  geom_sf(aes(fill = Proportion, geometry = geometry), color = "white", linewidth = 0.2) +
  geom_text(data = subset(etiketler_evlilik, !is.na(etiket)),
            aes(x = x, y = y, label = etiket),
            size = 3, color = "black") +
  scale_fill_distiller(
  palette = "YlOrRd",
  direction = 1,
  na.value = "grey90",
  name = "Proportion (%)",
  labels = waiver()
)+
  labs(
    title = "{closest_state} Consanguineous Marriage Rate by Province",
    fill = "Proportion"
  ) +
  theme_minimal() +
  transition_states(Year, transition_length = 2, state_length = 1) +
  ease_aes("cubic-in-out")
```

```r
anim_save("consanguinity_rate.gif", animate(p_evlilik, width = 800, height = 600, fps = 2, dur
library(magick)
image_read("consanguinity_rate.gif") %>%
  .[length(.)] %>%
  image_convert(format = "png", type = "truecolor") %>%
  image_write("consanguinity_rate.png")
```

2023 Consanguineous Marriage Rate by Province
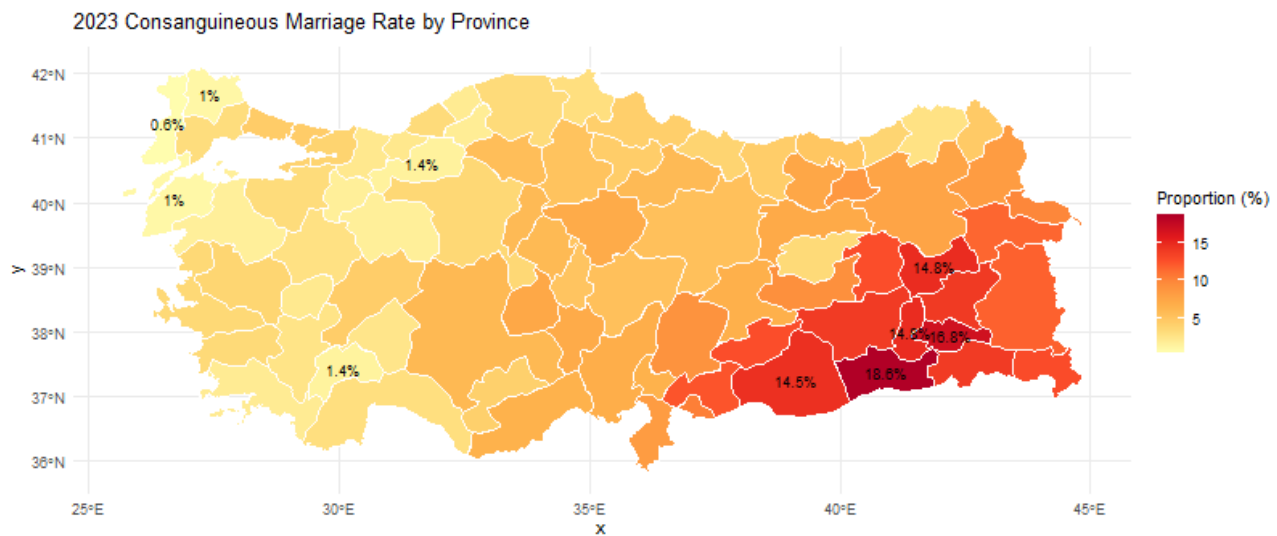
3.4 Regression Analysis

In this analysis, we wanted to see if education levels have an effect on cousin marriage rates in Turkey. For this, we built a regression model where the cousin marriage rate is the outcome, and the shares of high school, university, master???s, and PhD graduates are the predictors.

The results show that provinces with a higher percentage of only high school graduates tend to have higher cousin marriage rates. On the other hand, places with more universities, master???s, and especially PhD graduates tend to have lower rates.

This supports the idea that as education level increases, cousin marriage becomes less common. While the model doesn???t explain everything (since cultural factors also matter), it still gives a clear overall pattern.

```
library(dplyr)
combined_data <- combined_data |>
  mutate(
    total_edu = High_school_Total + Universities_Total + Master_Total + Doctorate_Total,
    hs_ratio = High_school_Total / total_edu,
    uni_ratio = Universities_Total / total_edu,
    master_ratio = Master_Total / total_edu,
    phd_ratio = Doctorate_Total / total_edu
  )

model <- lm(Proportion ~ hs_ratio + uni_ratio + master_ratio + phd_ratio, data = combined_data
summary(model)
```

28

```
Call:
lm(formula = Proportion ~ hs_ratio + uni_ratio + master_ratio +
    phd_ratio, data = combined_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4538 -2.4797 -0.8696  1.5759 12.2465

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -543.47      54.89  -9.901   <2e-16 ***
hs_ratio       557.66      55.02  10.136   <2e-16 ***
uni_ratio      539.25      55.12   9.784   <2e-16 ***
master_ratio   563.43      63.95   8.811   <2e-16 ***
phd_ratio          NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.441 on 1144 degrees of freedom
Multiple R-squared:  0.2071,    Adjusted R-squared:  0.2051
F-statistic: 99.62 on 3 and 1144 DF,  p-value: < 2.2e-16
```

In this regression model, we examined how the relative share of graduates at different education levels affects cousin marriage rates. The dependent variable is the cousin marriage proportion, and the predictors are the ratios of high school, university, master???s, and PhD graduates among all graduates.

The results show that high school, university, and masters graduate shares are all positively associated with cousin marriage rate. Surprisingly, the coefficient signs are all positive, but the interpretation depends on the structure of the ratios.

The coefficient for **phd_ratio is marked as NA** due to perfect multicollinearity ??? meaning its value is automatically determined by the others and cannot be estimated separately. This happens because the ratios sum to 1, so one of them must be dropped from the model.

The model explains around 20.5% of the variation in cousin marriage rates (Adjusted R?? = 0.205), which is reasonable given that cultural and regional factors are not included. **The model is statistically significant overall (p < 0.001).**

4. Results and Key Takeaways

This study examines the relationship between education levels and first cousin marriage rates based on provinces in Turkey, both by visualizing the relationship on a map and by performing regression analysis.

A normalized data set was created by combining education data and inbreeding data, thus enabling comparisons to be made both over time and between provinces.

The established regression model shows that the rates of first cousin marriages are lower in provinces with a high level of education. This suggests that there may be an inverse relationship between educational level and first cousin marriages.

A predictive analysis was performed with a linear regression model using normalized training ratios. The model showed significant results and explained about 20% of the change in the first cousin

marriages rate. He can talk about a socio-cultural structure to explain the rest of the part. But the scope of this research includes educational levels.

These findings show that increasing access to higher education, especially in regions where first-cousin marriage is common, may indirectly reduce these rates.