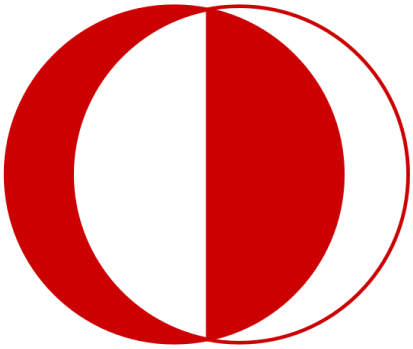


Middle East Technical University

Department of Statistics

STAT 250

APPLIED STATISTICS

2023 WORLD COUNTRIES INDICATORS:

Economic, Social and Demographic Analysis

June 2025

Submitted to Prof. Dr. Burçak Başbuğ Erkan

Aykut Ünyazıcı 2354280

Rabia Demircan 2623221

Timuçin Eke 2549244

Table of Contents

1.INTRODUCTION

1.1 Abstract

This study was created by bringing together multiple datasets. These analyses include economic, social, health, education, environmental, and demographic indicators for more than 190 countries for the year 2023. The dataset contains 35 variables such as life expectancy, CO2 emissions, education rates, tax revenue, health expenditures, and agricultural lands. The study examines differences, similarities, comparisons, and possible relationships between countries using statistical analysis.

The main purpose of the study is to determine the relationships between countries and continents. Especially, in this study, our aim is to determine the differences and similarities between Turkey's and other countries' data. Data tidying was made by R and visualizations, and some essential tests were performed by R and Tableau.

1.2 Data Description

The data set has 195 countries and 35 variables. Some of the variables of the countries are empty. Because of that, some of the countries cannot compare specific topics. Also, some of the numeric variables' have mistakes such as some of them written as integers and others written as float numbers. This mistake was fixed, and numerical variables were written in float format.

The descriptions of the variables were taken from Kaggle.

Country	Name of country	Qualitative
Density	Population density is measured in persons per square kilometer	Quantitative
Abbreviation	Abbreviation or code representing the country	Quantitative
Agricultural Land	Percentage of land area used for agricultural purposes	Quantitative

Land Area	Total land area of the country in square kilometers	Quantitative
Armed Forces Size	Size of the armed forces in the country	Quantitative
Birth Rate	Number of births per 1,000 population per year	Quantitative
Calling Code	International calling code for the country	Categorical
Capital/Major city	Name of the capital city of the country	Categorical
Co2-Emission	Carbon dioxide emissions in tons	Quantitative
CPI	Consumer Price Index, a measure of inflation	Quantitative
CPI Change	Percentage change in the CPI compared to the previous year	Quantitative
Currency-Code	Currency code used in the country	Qualitative
Fertility Rate	Average number of children born to a woman during her lifetime	Quantitative
Forested Area	Percentage of land area covered by forest	Quantitative
Gasoline Price	Price of gasoline per liter in local currency	Quantitative

GDP	Gross Domestic Product, is the total value of goods and services produced in the country	Quantitative
Gross primary education enrollment	Gross enrollment ratio for primary education	Quantitative
Gros tertiary education enrollment	Gross enrollment ratio for tertiary education	Quantitative
Infant mortality	Number of deaths per 1000 live births before reaching one year of age	Quantitative
Largest city	Name of the country's largest city	Qualitative
Life expectancy	The average number of years a newborn is expected to live	Quantitative
Maternal mortality ration	Number of maternal deaths per 100,000 live births.	Quantitative
Minimum wage	Minimum wage level in local currency.	Quantitative
Official language	Official language spoken in the country	Qualitative
Out-of-pocket health expenditure	Percentage of total health expenditure paid out of pocket by individuals.	Quantitative
Physicians per thousand	Number of physicians per thousand people.	Quantitative

Population	Total population of the country	Quantitative
Population: Labor force participation	Percentage of the population that is part of the labor face.	Quantitative
Tax revenue	Tax revenue as a percentage of GDP	Quantitative
Total tax rate	Overall tax burden as a percentage of commercial profits.	Quantitative
Unemployment rate	Percentage of the labor force that is unemployed.	Quantitative
Urban population	Percentage of the population living in urban areas.	Quantitative
Latitude	Latitude coordinate of the country's location.	Quantitative
Longitude	Longitude coordinate of the country's location.	Quantitative

2. DATA ANALYSIS

2.1 Research Question 1

Are the average agricultural lands in European countries' countries equal to the EU target of 37%?

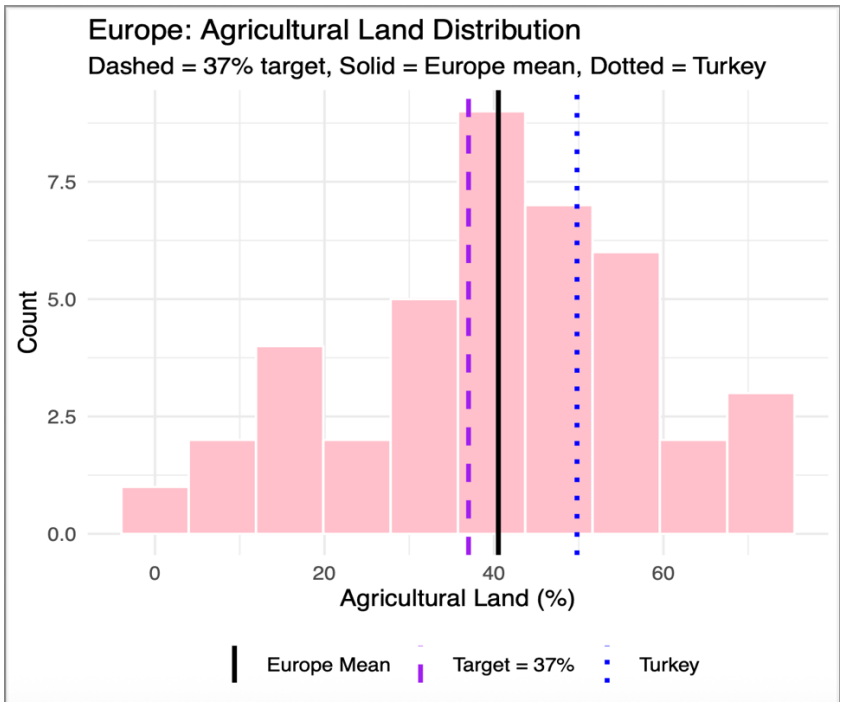


Figure 1.1

In the histogram, 44 Europe countries' agricultural land percentages are shown. The purple dashed line represents the EU target as 37%, Europe's mean which is 38.93 % land shown with the black solid line in the histogram. Most of the countries' agricultural land percentages are around 34% and 40%. Europe means fall right off the target so it's higher than 37%.

Also, there is a dotted line in the figure which represents Turkey's agricultural land proportion. We know that it is 53%. To sum up, Turkey's variable is higher than Europe, so Turkey is more advanced than most of the EU countries in protecting its agriculture.

For the test of the question, the first Shapiro-Wilk test was applied since it is normal and there is no outlier, tested with a box plot, all assumptions are satisfied: one-sample t-test can apply. The result of the test shows us There is not enough evidence to reject the claim that: "The average of agricultural lands in European countries equal to the EU target of 37%".

2.2 Research Question 2

Is the average gasoline price different between European and North American countries?

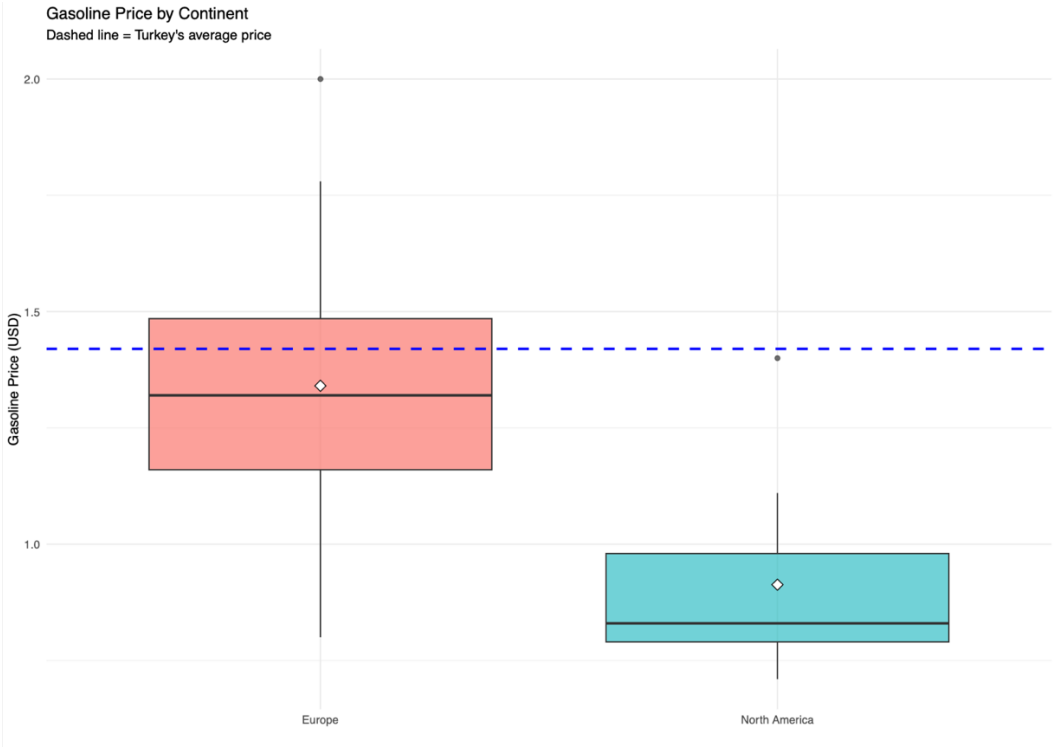


Figure 2.1

There are two boxplots in the figure, the left box represents Europe and the right one as North America. The lines in the middle of the boxes show the median and the upper and lower lines show the 3rd and 1st quartiles. The white rhombs in the middle of the boxes show the average gasoline price of each group. (Europe mean≈ 1.34USD, North America≈ 0.91USD)

European median and mean are above North America so we can see that European average gasoline price is higher than North America.

Also, the blue dotted line represents Turkey's gasoline price as 1.42 USD. So, Turkey's price is relatively high compared to North America and Europe.

The test step of the question applies the normality test and variance test. Since all assumptions are satisfied two-sample t-test for equal variances applied. At the level of %5 significance, there is enough evidence to the support claim that: "The average gasoline prices of the European and North American countries are different".

2.3 Research Question 3

Does having a higher CPI also mean a higher CPI increase?

CPI (Consumer Price Index) is an indicator to track inflation. The baseline of 100 was set for CPI values in 1984. So, a country with 130 CPI means its inflation is %30 higher than in 1984. CPI change represents the difference between last year and this year. Is there a relation between CPI and its increase rate?

This question requires a regression model to detect if two variables change together. After checking the assumption, we can use a simple linear regression model.

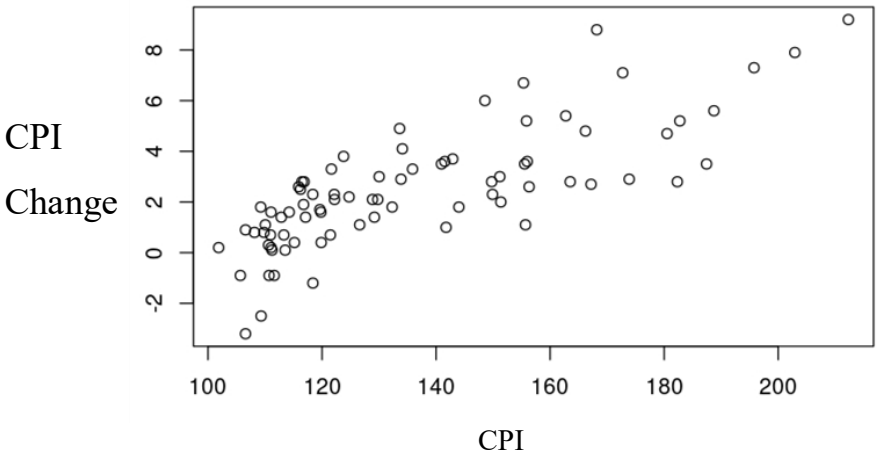


Figure 3.1

Figure 3.1 Shows a possible linear corr between variables. Assumptions must be cl before using simple linear regression.

Linearity: Two variables have a linear relation.

Normality of residuals: Residual errors must be normally distributed.

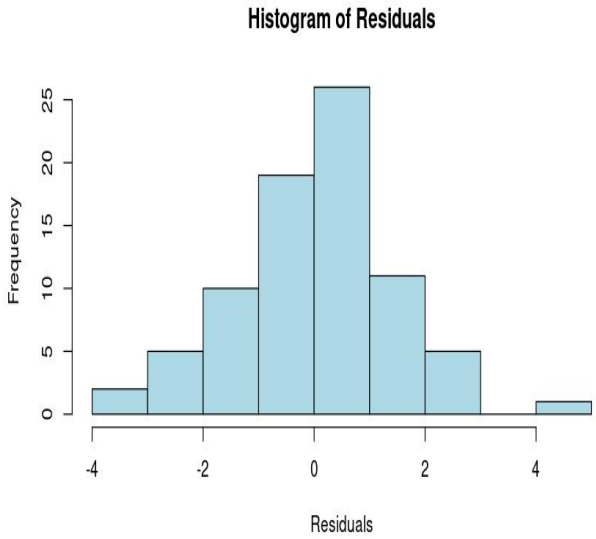


Figure 3.2

Result of Shapiro-Wilk test is 0.9317. Based on test results and Figure 3.2 we can assume normality.

Homoscedasticity: Variances of residuals must be consistent.

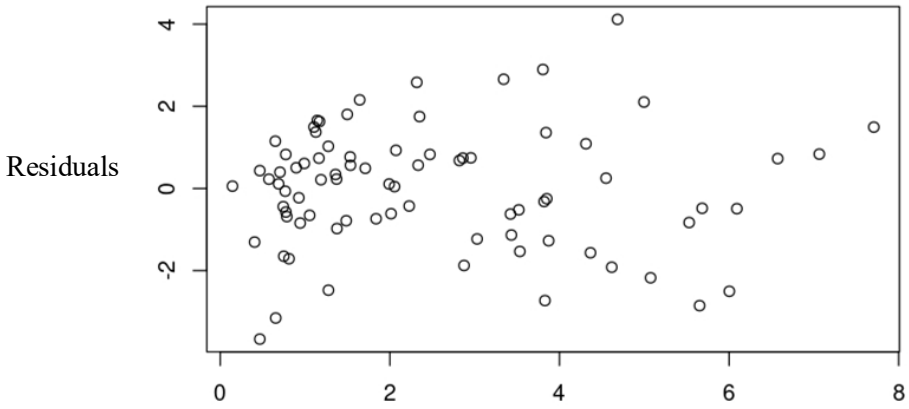


Figure 3.3

Figure 3.3 shows no clear correlation between residuals and fitted values. With R-Square value of -0.01299 variances of error terms are consistent.

Multicollinearity: With 0.605 r-squared value there is no multicollinearity.

Since all assumptions are met, we can use a simple linear regression model.

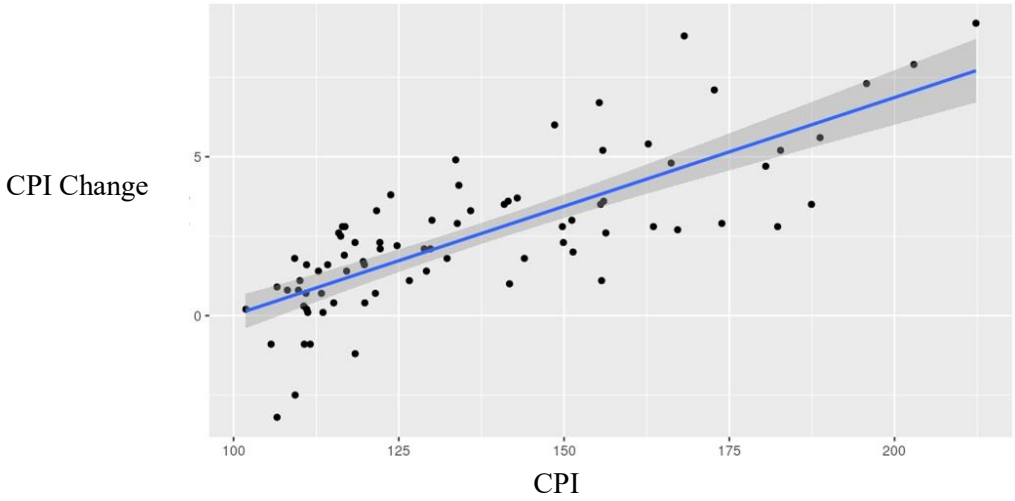


Figure 3.4

$\hat{y} = -6.83231 + 0.06848x1$

To test the significance of coefficients:

$H_0: \beta_1 = 0$

$H_1: \beta \neq 0$

The P-value for the significance of the coefficient is smaller than 2e-16. We can reject the null hypothesis that "Higher CPI values also mean higher CPI increase rate".

2.4 Research Question 4

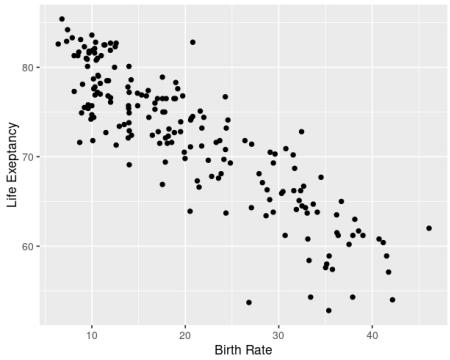
How can we estimate Turkey's life expectancy based on other variables?

After analyzing the data three variables stand out for estimation. These variables have clear linear strong relations with life expectancy. These variables are birth rate, infant mortality, and fertility rate.

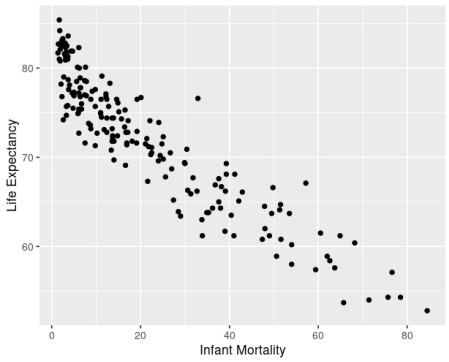
Although we have three variables to create a multiple linear regression model. First, we should check if this model is proper by checking assumptions.

Linearity: Dependent variables and independent variables must have a linear relationship.

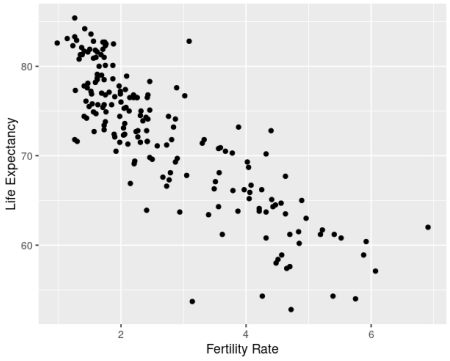
Life Exp. Vs Birth Rate



Life Exp. Vs Infant Mortality



Life Exp. Vs Fertility Rate



Three independent variables have a linear relation with the dependent variable.

No Multicollinearity: Two independent variables must not have a high correlation.

After calculating r-square values we see that the Fertility Rate and Birth Rate have 0.9622 r-square values. That indicates multicollinearity. So, we must remove one of these variables. We decided to keep the Birth Rate since it is more suitable for linear regression.

Multivariate Normality: Random errors must be normally distributed. To check this, we can look at the histogram of random errors.

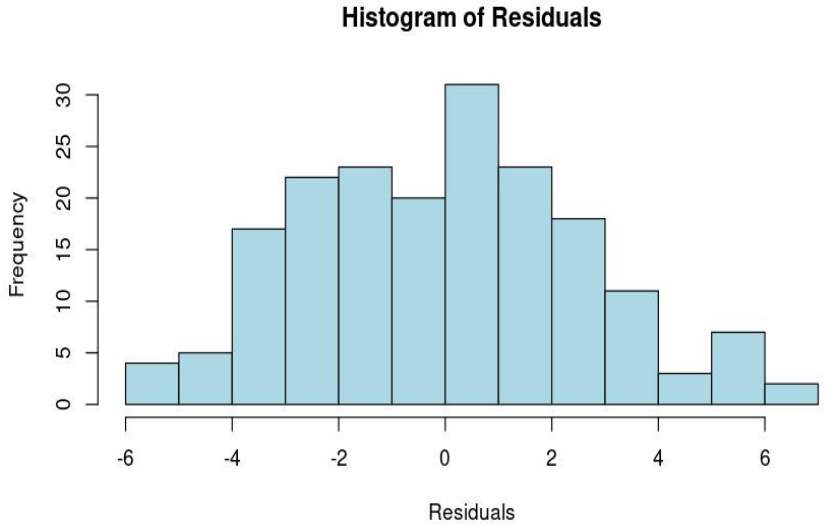


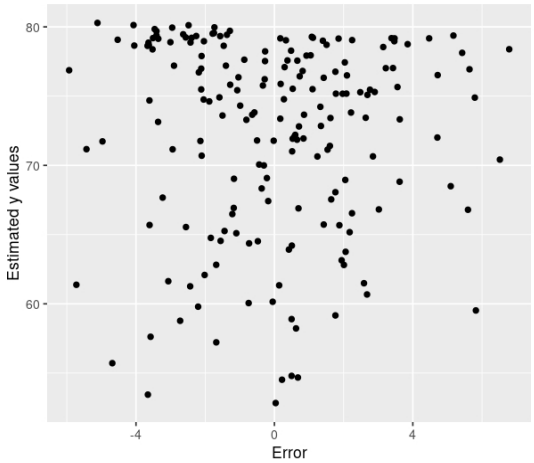
Figure 4.4

Both visual methods and Shapiro-Wilk test, with 0.2832 p-value, indicate normality for random errors.

Variance Inflation Factor: With VIF equal to 7.692308 it is within limits.

Sample Size and Variable Types: All variables are continuous, and their level is ratio. Also, the sample size is large enough.

Homoscedasticity: Variances of error terms should be consistent among independent variables. A scatter plot between residuals and predicted values is a good way to test it.



This plot shows no clear relation. This indicates error terms are consistent among independent variables.

Figure 4.5

Since all requirements are met, we can use multiple linear regression. The formula for the fitted model is:

$$\hat{y} = 82.22 - 0.2205x_1 - 0.2555x_2$$

Testing the significance of the model:

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \beta_1 \neq 0, \beta_2 \neq 0$$

With a p-value smaller than 2e-16, we can reject null hypothesis and apply the fitted model to Turkey. Turkey's birth rate is 16.03 and infant mortality is 17 so life exp is 76.36033.

2.5 Research Question 5

Do countries from different continents have different forested areas on average?

To test this, we randomly selected ten countries from Africa, Europe, and Asia. First using a box plot we will examine data visually.

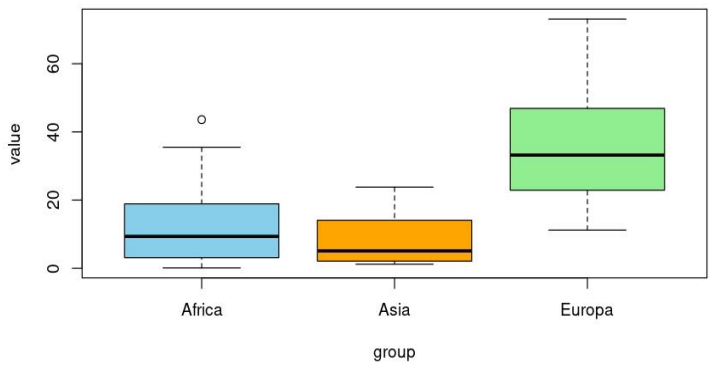


Figure 5.1

Figure 5.1 shows that countries from Asia and Africa have similar average forested area but, European countries seem to have significantly higher average than both Asian and African countries. To test this hypothesis One-Way ANOVA would be logical. But before apply test first we must check assumptions for One-Way ANOVA.

1-Indipendence: Samples are randomly selected using R and they are independent.

2-Normality of residuals: Visual inspection and the Shapiro-Wilk test both show normality.

3-Variances must be homogenous: After using Barlett's test p-value is 0.052. So, we can assume that variances are homogenous among variables.

Since all assumptions are met, we can use One-Way ANOVA.

Do all continents have an equal average forested area?

```
            Df Sum Sq Mean Sq F value Pr(>F)
group        2   4012  2005.8   9.622  7e-04 ***
Residuals   27   5628   208.5
```

ANOVA table gives a p-value of 7e-4 so at a 5% significance level, we can reject the null hypothesis and say at least one continent has a different average forested area. Based on Figure 5.1 we can assume Europe has a higher average than Africa and Asia. But to be sure we can test if African and Asian countries have the same average. If we cannot reject null hypothesis, we can say different mean belongs to Europe.

Shapiro-Wilk test shows that all averages are approx. Normally distributed.

Do Asian and African countries have the same average forested area?

The P-value for the two mean test is 0.5080. At 0.05 significant level, we cannot reject the null hypothesis. This test doesn't tell us that these two mean are equal. However, since the p-value is very high we can assume different mean cannot be one of these continents.

So, the conclusion is while average forested areas of Asian and African countries are not statistically significant Europe's average is significantly higher.

2.6 Research Question 6

Does the number of doctors in a country make a meaningful difference in how many infants survive their first year? Specifically, are countries with more physicians per 1,000 people less likely to have high infant mortality rates than those with fewer?

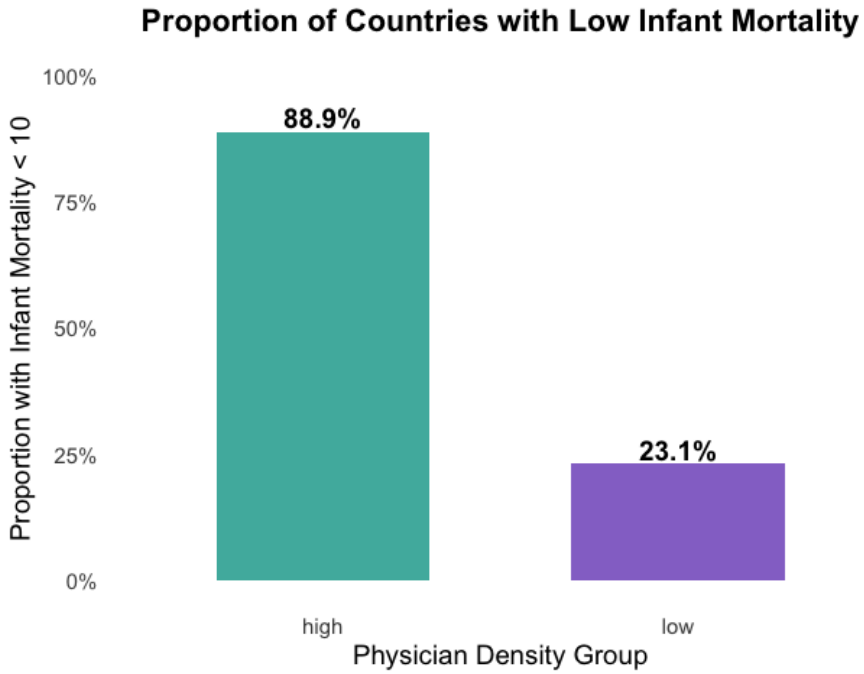


Figure 6.1

To examine how access to doctors might influence child survival rates, we looked at whether the share of countries with low infant mortality meaning fewer than 10 deaths for every 1,000 live births varies depending on how many physicians they have per 1,000 people.

We sorted the countries into two categories based on how many doctors they have per 1,000 people: those with high physician density (3 or more doctors) and those with low physician density (fewer than 3 doctors).

The outcome variable was binary: Low infant mortality: 1 if infant mortality < 10, otherwise 0

Is the proportion of countries with low infant mortality is the same in both groups?

A contingency table was created to compare the proportion of low infant mortality across both physician-density groups. Since some cell counts were small, we applied both the **t**wo-proportion z-test and Fisher's Exact Test to evaluate the hypothesis.

Before running the two-proportion z-test, we made sure the conditions for using it were met. Each country was counted as a separate case, so the data points were independent. We also used a simple yes-or-no setup for whether a country had low infant mortality—marked as 1 if the rate was under 10, and 0 if it wasn't. Both groups had enough countries to make the test valid, with at least five countries showing each outcome. Even though the countries weren't picked randomly, that's common in observational research and doesn't make the test invalid. So overall, the assumptions held up well enough to go ahead with the z-test.

Because the p-value was significantly lower than 0.05 in both tests, we rejected the null hypothesis. This provides strong statistical evidence that the proportion of countries with low infant mortality does differ based on physician density. In this case, countries with more physicians per 1,000 people are much more likely to have low infant mortality rates.

The percentage of countries displaying low infant mortality is significantly higher in the high physician density group (88.9%) than in the low group (23.1%).

3. OUTCOME & CONCLUSION

Countries' indicators are provided by our analysis so that all factors impacting them can be investigated. We conducted several analyses to find the predictors using the dataset, which includes data from multiple countries on several continents.

The key findings are as follows:

The average agricultural land area of European countries is equal to the target set by the EU; the target is 37% and the average of Europe is 40%, which shows that European countries have achieved its target in this regard. Also, European countries' average forested area is higher than the African and Asian continents. There is no meaningful difference between Africa and Asia.

In addition, Turkey is much higher than the average and it shows us that Turkey is more successful and at a better level in protecting agricultural areas, but it is not as successful as in the forested area comparison.

In the other one, gasoline price data were used. The average gasoline price is higher in European countries than in North America. Also, Turkey's value is included in the analysis. Turkey's value is higher than both Europe and North America. Also in the data set, we can recognize it Turkey is one of the countries with the highest gasoline prices.

Also, Turkey's life expectancy is based on 3 variables; these are birth rate, infant mortality, and fertility rate. These relationships are strong negative relationships. As birth rate, infant mortality, and fertility rate increase, the value of life expectancy is decreasing.

Our last finding is, that there is a negative relationship between number of the doctors in the country and infant mortality. It's seen in the analysis as the number of physicians increasing infant mortality which means the number of babies who can live until one year old is decreasing.

In conclusion, our analysis shows that a combination of demographic, economic, educational, and social indicators conduce to national-level success. However, because of the size of the dataset, it is not possible to compare all variables and see the relationship between them. So, dataset might require deeper analysis rather than direct analysis.

4. SUGGESTION FOR FUTURE WORK

As mentioned before, because of the size of the dataset; it is not possible to compare all variables and see the relationship between them. For this reason, there are several directions for future work.

Time Series Analysis: The variables are collected in 2023, if there is another year variable, it will allow a better examination of relationships and trends over time. (After the analysis relationship can be trend, seasonal, or no relationship)

Psychological and Cultural Variable: Indicators such as national happiness, and behaviors towards education or technology. There are lots of variables in the data set but most of them are about environmental and economic. Adding these kinds of variables will provide a richer dimension to the analysis.

In the last question of the research, we want to search for other factors for instance healthcare system efficiency and, the status of economic or regional effects that might also contribute. After adding these factors, the answer to the last question will be more meaningful.

5.REFERENCE

Elgiriyewithana, N. (2023). Global Country Information 2023. https://doi.org/10.5281/zenodo.8165229