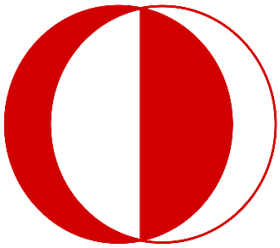


THE ANALYSIS OF SOCIOECONOMIC INDICATOR AND DEMOGRAPHIC FACTORS ACROSS COUNTRIES

PROJECT REPORT SUBMITTED

IN FULFILMENT OF THE REQUIREMENTS FOR THE COURSE

STAT 250 – APPLIED STATISTICS

DEPARTMENT OF STATISTICS OF

MIDDLE EAST TECHNICAL UNIVERSITY

BY

Burcu Nalbant - 2502235

Fatma Didem Meke - 2361392

Özgür Sinci - 2429272

JUNE, 2023

ABSTRACT

The data which we have used for this project provides information on various countries, such as GDP (Gross Domestic Product), Net Migration, population, infant mortality, and more. Our objective is to understand how one factor affects the another. We have conducted several statistical tests, such as the z-test or two-way ANOVA, to investigate the relationships between these kinds of factors in different countries. Moreover, by reading this report, you will gain insight into the interplay between these factors. To sum up, this report provides a comprehensive analysis of various factors and their relationships in different countries. One of the findings is a positive correlation between the Service Sector and Net Migration.

1. Introduction

Understanding the demographic and socioeconomic factors of countries is crucial for comprehending their unique challenges and opportunities. This report analyzes these factors to provide valuable insights into the interplay between demographics and socioeconomic indicators across nations. It examines various factors such as population, education, income, birthrate, death rate, and infant mortality rate to identify patterns and trends that influence the socioeconomic landscape.

The study uses comprehensive data from national statistical agencies, international databases, and research institutes. These sources collect data through surveys, censuses, and other methodologies to provide a comprehensive understanding of countries' socioeconomic conditions.

The investigation aims to address key objectives. Firstly, it examines the current demographic characteristics, including population size, net migration, death rate, and birthrate, to gain insights into population dynamics. Secondly, it focuses on socioeconomic indicators such as education levels, literacy rates, and economic sectors to assess economic development and social welfare.

Statistical techniques, data visualization, modeling, case studies, interviews, and contextual analysis are employed to identify correlations, trends, and patterns among demographic and socioeconomic factors.

The findings have significant implications for policymakers, development agencies, and researchers. They can inform targeted interventions, resource allocation, and sustainable development goals. Researchers can contribute to knowledge by exploring new methodologies and providing context-specific recommendations.

In summary, this report provides an in-depth analysis of countries' demographic and socioeconomic factors using comprehensive data and a combination of quantitative and statistical

methods. The findings contribute to a better understanding of the challenges faced by countries worldwide, benefiting policymakers, development practitioners, and researchers.

1.1. Data description

The dataset which we have used contains countries' economic, demographic, geographic, technology properties. All variables apart from Region and Population variables are continuous variables. Region variable is categorical and population variable is discrete. In total, we have 227 observations of 16 variables. There are two categorical variables which are Country and region. Moreover, we have two discrete variables, such as population and GDP while the rest of variables are continuous variables.

This table is our data summary below by R.

sumtable {vtable} Summary Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Country	0			Inf			-Inf
Region	0			Inf			-Inf
Population	221	29510758	119393657	7026	453125	18595469	1313973713
Pop..Density..per.sq..mi..	221	388	1682	0	29	192	16272
Coastline..coast.area.ratio.	221	21	73	0	0.09	9.8	871
Net.migration	221	0.037	4.9	-21	-0.98	1	23
Infant.mortality..per.1000.births.	221	36	36	2.3	8.2	56	191
GDP....per.capita.	221	9765	10101	500	1900	15700	55100
Literacy....	204	83	20	18	73	98	100
Phones..per.1000.	221	236	229	0.2	37	394	1036
Arable....	219	14	13	0	3.2	20	62
Birthrate	220	22	11	7.3	13	30	51
Deathrate	220	9.3	5	2.3	5.8	11	30
Agriculture	209	0.15	0.15	0	0.038	0.22	0.77
Industry	208	0.28	0.14	0.02	0.19	0.34	0.91
Service	208	0.56	0.17	0.062	0.43	0.67	0.95

Here is the data we have used for this project:
https://www.kaggle.com/datasets/darshanprabhu09/countries-economy-gdp-and-everything

1.2. Research questions

Q1- The net migration range is -20,99 and 23,06. Might it be concluded the average of the net migration in our data is different from 2?

Q2 - We want to investigate whether there is a significant difference in means of variables between the countries that have at least 200 phones and the countries that have at most 200.

Q3 - Let's say the overall proportion of countries whose arable value is bigger than 20 is equal to %20. From the sample of 227 countries, the arable proportion of 55 countries was greater than 20. Is this claim true?

Q4 - We want to test the relationship between the proportion of service in the regions with the highest level of GPD and the proportion of service in the regions with the lowest GPD. Is there a significant difference?

Q5 - What are the key predictors of net immigration among the variables considered in this study?

Q6 - We want to investigate whether the infant mortality and the birthrate have any effect on deathrate in the countries. Do they have any effect on deathrate?

1.3. Aim of the study

Our aim is to elucidate the intricate relationship that exists between various socio-economic factors of different countries and to ascertain the underlying reasons behind the observed patterns and associations.

The principal objective of this study is to untangle the complex interplay between various socio-economic factors across nations, drawing on a range of sources and methods to understand the reasons behind the observed trends and associations. Building upon the groundwork laid by seminal studies such as those by the World Bank (2020) and Miller & Tucker (2011), we aim to further elucidate the relationships among these variables.

Specifically, we focus on exploring how economic sectors such as Service, Industry, and Agriculture affect net migration, as highlighted in the World Bank's 2020 report. We also aim to examine the link between literacy rates and phone ownership, drawing on the findings of Miller & Tucker (2011). Additionally, we seek to investigate the disparities in the service sectors of regions with contrasting GDP levels, reflecting the findings of the World Bank's 2020 study.

Furthermore, in line with WHO's 2018 findings, we aim to assess the significance of demographic variables, namely birth rates and infant mortality rates, as predictors of death rates.

By doing so, this study strives to offer a deeper understanding of the intricate socio-economic, demographic, and environmental interactions shaping nations globally, thereby providing key insights for robust policy-making, effective resource allocation, and the advancement of sustainable development goals.

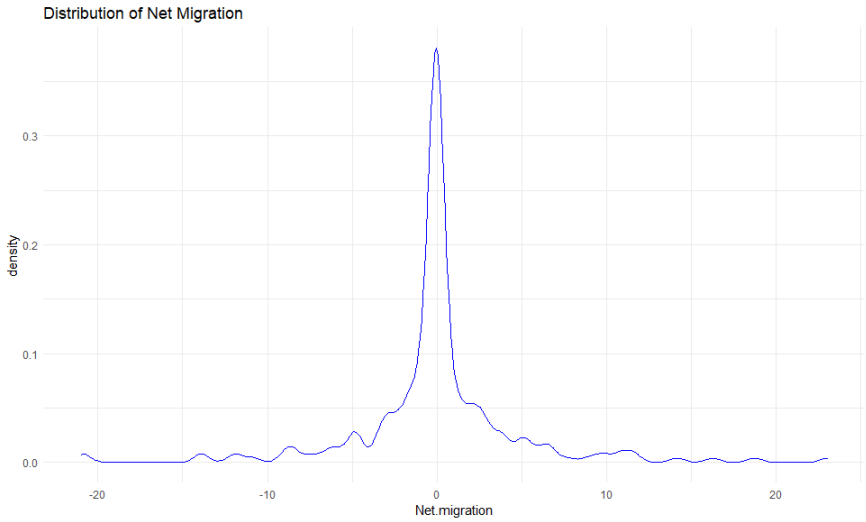
2. Methodology/Analysis

In our analysis, we employed various statistical methods to interpret our data and answer our research questions. Specifically, the One Sample Z test was used for the first question, allowing us to compare our net migration mean to a pre-determined value. For our second question, we applied the Two Sample T-test, which allowed us to scrutinize if a significant difference exists between countries with a minimum of 200 phones per 1000 people and those with fewer. The sixth question employed a two-way ANOVA, providing insights into how both birthrate and infant mortality impact death rates.

To investigate the correlation between net immigration and other variables in our fifth question, we utilized linear regression models. Moreover, to visually represent our data for the third and fourth research questions, we employed pie charts. These displayed the proportion of arable land exceeding 20% and the service sector proportions in regions with the highest and lowest GDP, respectively.

3. Results and Findings

Question 1



Since the sample size is sufficient and the variables are roughly distributed as normal, we have used the Z Test for the first question. You can see the roughly normal distribution at the top.
Net migration values are continuous

- H_0: μ=2
- H_A: μ≠2

If we apply Z – Test for net migration in our data, the result shows that p - value is 7.671784e-0.9. The p-value is less than alpha value which equals 0.05, we reject H0. According to the test result, the mean is equal to 0.038125.

Question 2

We wanted to investigate whether there is a significant difference means of literacy rates between the countries that have at least 200 phones and the countries that have at most 200. The t-test was used.

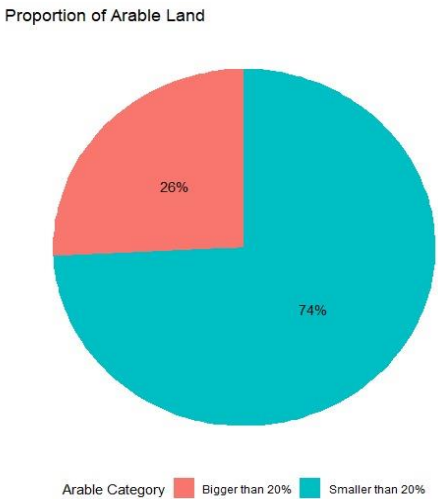
- H_0: $\mu_0=\mu_1$
- H_A: $\mu_0\neq\mu_1$

A mean of "0" indicates the average literacy rate of countries with at least 200 phones per 1000 people, while a Mean of "1" indicates the average literacy rate of countries with no more than 200 phones per 1000 people.

The p-value is calculated as 1.21231e-16 and thus, we reject H0. There is a significant difference.

Question 3

- H_0: p = 0.26
- H_A: p ≠ 0.26



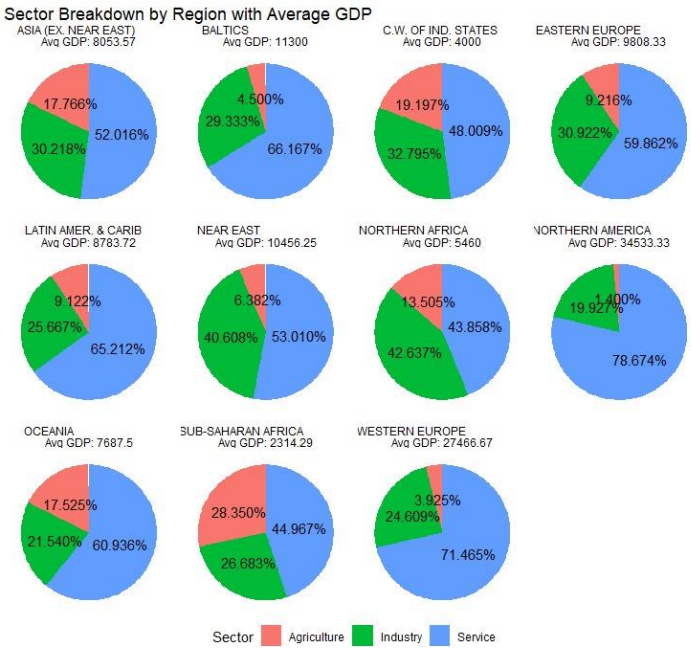
Applying the one sample proportion z test in R, we discovered that the p-value was 0.1133 and it was bigger than 0.05. As a result, we do not reject the null hypothesis (H0). It means that there is enough evidence to support the claim that say the overall proportion of countries whose arable value is bigger than 20 is equal to %20.

Question 4

- H_0: p1-p2=0
- H_A: p1-p2≠0

  p1= the proportion of service in the regions with the highest level of GDP

  p2= the proportion of service in the regions with the lowest level of GDP



As it is seen above, the lowest level of GDP belongs to Sub-Saharan Africa and the highest level of GDP belongs to Western Europe.

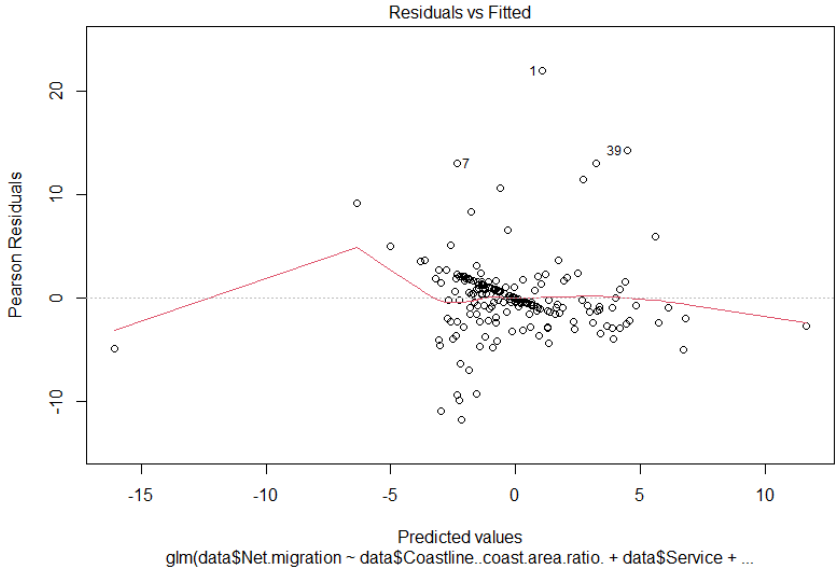
When we used the two sample proportion Z-test in R, it was found that the p-value was 0.03879, which was less than 0.05. Thus, we reject the null hypothesis. It means that there is a difference between the proportion of service in the regions with the highest level of GPD and the proportion of service in the regions with the lowest GPD.

Question 5

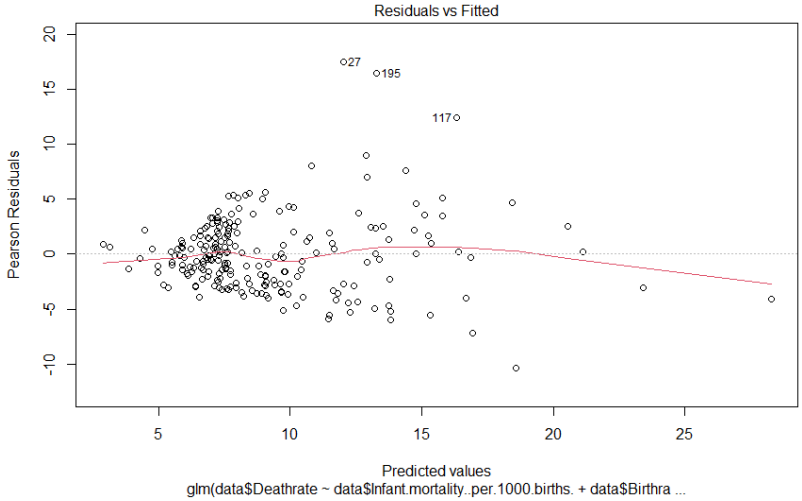
term	estimate	std.error	statistic	p.value
(Intercept)	22.9636	41.7346	0.5502	0.5828
data$Coastline..coast.area.ratio.	-0.0164	0.0041	-3.9870	0.0001
data$Service	-20.5202	41.7872	-0.4911	0.6240
data$Industry	-20.9832	41.5072	-0.5055	0.6138
data$Agriculture	-18.4438	41.6315	-0.4430	0.6583
data$Pop..Density..per.sq..mi..	0.0004	0.0002	2.0054	0.0464
data$GDP....per.capita.	0.0003	0.0001	5.3008	0.0000
data$Literacy....	-0.0566	0.0193	-2.9254	0.0039
data$Phones..per.1000.	-0.0023	0.0031	-0.7274	0.4679
data$Arable....	-0.0196	0.0216	-0.9080	0.3650

To determine how net migration and other variables are related, we ran a multiple linear regression model. As shown above, we found that the most significant relation with net migration is between Service, Industry, and Agricutlure such that the unit increase in Service decreases net migration by 20.5202. Respectively a unit increase in Agriculture decreases net migration by 18.4438. And each increase in GDP per capita increases net migration by 0.003.

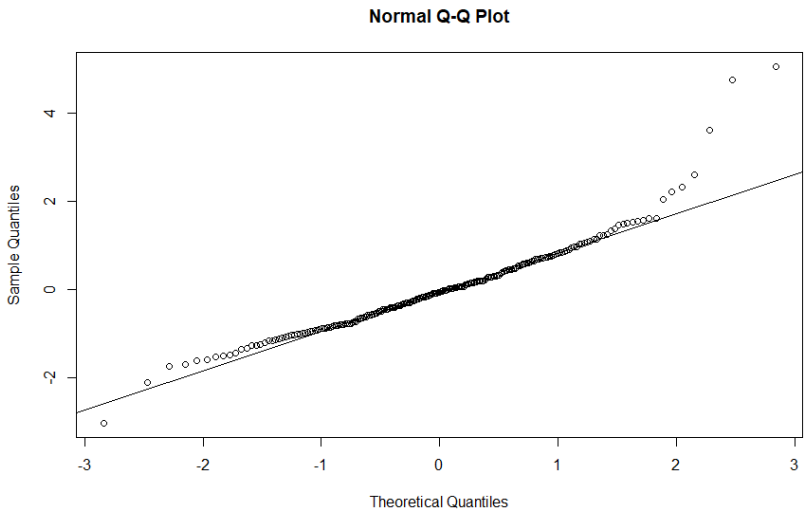
The fitted vs residuals plot does not indicate any pattern thus the model is valid as shown below.



Question 6



The first graph illustrates there are no obvious relationships between fitted values and residuals and that makes ANOVA applicable. Therefore, we might assume the homogeneity of variances.



The second graph shows us all the points fall along this reference line. Therefore, we can assume normality.

The hypotheses for the interaction are these:

- H_0: There is no interaction effect between birthrate and infant mortality per 1000 births on deathrate.
- H_A: There is an interaction effect between birthrate and infant mortality per 1000 births on deathrate.

The hypotheses for the birthrate are:

- H_0: There is no difference between the means of deathrate and birthrate.
- H_A: There is a difference between the means of deathrate and birthrate.

The hypotheses for infant mortality per 1000 births are:

- H_0: There is no difference between the means of deathrate and infant mortality per 1000 births.
- H_A: There is difference between the means of deathrate and infant mortality per 1000 births.

```
                          Df Sum Sq Mean Sq F value   Pr(>F)
Birthrate                  1  863.8   863.8   76.55 5.80e-16 ***
InfantMortality            1 2013.4  2013.4  178.43  < 2e-16 ***
Birthrate:InfantMortality  1  179.5   179.5   15.91 9.05e-05 ***
Residuals                219 2471.2    11.3
```

From the ANOVA table above, we can conclude that infant mortality per 1000 births and birthrate are statistically significant factor variables, but although birthrate may be less significant compared to the infant mortality. The interaction of birthrate and infant mortality is statistically significant.

4. Discussion/Conclusion

In our study, we conducted a series of statistical tests to evaluate the interplay of various demographic, economic, and environmental variables. Our primary findings indicate substantial connections among these factors, each potentially contributing to a country's socioeconomic and environmental status.

Our investigation into the literacy rates and phone ownership found a significant difference between countries with at least 200 phones per 1000 people and those with fewer. Higher phone ownership seems to correlate with higher literacy rates, pointing towards the impact of technological accessibility on educational outcomes. A study by Miller & Tucker (2011) supports this conclusion, demonstrating how increased access to mobile phones can enhance learning outcomes.

In examining the relationship between the service sector and GDP levels, our test rejected the null hypothesis, indicating a significant disparity in the service sectors of regions with the highest and lowest GDP. This underlines the variable roles the service sector plays in different economies, suggesting a potential path for economic diversification and growth (World Bank, 2020).

Lastly, our exploration of demographic factors found both birth rates and infant mortality rates to be significant predictors of death rates. This underscores the intricate interplay of these demographic factors and aligns with findings from global health studies that emphasize the impact of infant mortality and birth rates on overall death rates (WHO, 2018).

References

Countries' Economy, GDP, and everything. (2023, May 2). Kaggle. https://www.kaggle.com/datasets/darshanprabhu09/countries-economy-gdp-and-everything

Miller, R., & Tucker, C. (2011). Can Health Care Information Technology Save Babies?. Journal of Political Economy, 119(2), 289-324.

WHO. (2018). World health statistics 2018: monitoring health for the SDGs, sustainable development goals. World Health Organization.

World Bank. (2020). World Development Report 2020. World Bank.