RAMAIAH
Institute of Technology

A project report on

Analysis of the mortality rate of infants

submitted in partial fulfillment of the requirements

Of

Data Science (IS71) &
Data Science Laboratory (ISL76)

In

Seventh Semester

By

Sumanth BS [1MS20IS119]

Satwik [1MS20IS106]

Sujan [1MS20IS118]

Under the guidance of

Savita K Shetty

Assistant Professor

Dept. of ISE, RIT

RAMAIAH INSTITUTE OF TECHNOLOGY

BANGALORE – 560054

2023

Contents

Abstract

This study delves into the critical issue of Infant Mortality Rate (IMR) with a multifaceted approach. The primary objectives include a comparative analysis of IMR across various race/ethnicities in relation to the global average, as well as a comprehensive examination of IMR disparities among different countries, also in comparison to the world average.

To enhance our understanding of the dynamics, a longitudinal perspective is incorporated through regression analysis, allowing for the identification of trends in IMR over the years. By addressing these objectives, this study aims to contribute valuable insights to the ongoing discourse on infant mortality, paving the way for informed interventions and policies to improve global infant well-being.

Introduction

In the realm of public health, the challenge of understanding and mitigating Infant Mortality Rate (IMR) demands a nuanced and analytical approach. This project unfolds within the disciplined framework of the Data Analytics Lifecycle, a systematic journey comprising distinct phases: Discovery, Data Preparation, Model Planning, Model Building, Communicate Results.

The project aims to achieve the following objectives:

1. Conduct a comparative analysis of IMR across diverse race/ethnicities and countries in relation to the global average.
2. Explore longitudinal trends in IMR over time through regression analysis.

As we embark on this methodical exploration, each phase of the Data Analytics Lifecycle becomes a crucial component, contributing to the overall comprehension and strategic utilization of IMR data. This project not only aims to unravel the intricacies of infant mortality but also strives to translate analytical insights into actionable measures, thereby contributing to the ongoing global efforts to enhance the well-being of infants worldwide.

Data Discovery

In the initial phase of Data Discovery, our project relies on three key datasets that encapsulate the multifaceted dimensions of Infant Mortality Rate (IMR). These datasets serve as foundational pillars for our analysis:

1. Global Child Mortality TimeSeries:

- Source: UNICEF - Child Mortality Estimates (http://www.childmortality.org/)
- This dataset, compiled by UNICEF, provides a comprehensive global perspective on child mortality over time. The inclusion of time-series data allows us to discern patterns and variations in Infant Mortality Rate across different regions.

2. Child Mortality:

- Source: World Bank (https://data.worldbank.org/indicator/SH.DYN.MORT)
- The World Bank's report on Child Mortality offers a valuable dataset for our comparative analysis. Drawing from a diverse set of countries, this dataset enables us to juxtapose IMR among different nations and assess disparities in child health on a global scale.

3. Infant Mortality:

- Source: City of New York (https://data.cityofnewyork.us/Health/Infant-Mortality/fcau-jc6k/data)
- Focused on a specific region, the Infant Mortality dataset from the City of New York provides a localized perspective. This granular data allows us to explore IMR trends within a specific demographic, adding depth to our analysis and enhancing our understanding of regional variations.

As we navigate the Data Discovery phase, the integration of these diverse datasets sets the stage for a comprehensive exploration of Infant Mortality Rate, laying the groundwork for subsequent phases in the Data Analytics Lifecycle.

Data Preparation and Exploration

1. Global Child Mortality TimeSeries

```
> names(df)
[1] "Entity"
[2] "Code"
[3] "Year"
[4] "Share.surviving.first.5.years.of.life...."
[5] "Share.dying.in.first.5.years...."
```

- `entity`: Represents the world as a singular entity in the dataset.
- `year`: Denotes the specific year in which the data is recorded.
- `share_surviving_first_5_years`: Indicates the proportion of individuals worldwide who survive the initial 5 years of life.
- `share_dying_in_first_5_years`: Represents the proportion of individuals worldwide experiencing mortality within the first 5 years of life.

```
> summary(df)
    Entity              Code                Year
 Length:64          Length:64          Min.   :1800
 Class :character   Class :character   1st Qu.:1968
 Mode  :character   Mode  :character   Median :1984
                                       Mean   :1973
                                       3rd Qu.:1999
                                       Max.   :2015
 Share.surviving.first.5.years.of.life....
 Min.   :56.70
 1st Qu.:83.50
 Median :89.43
 Mean   :86.22
 3rd Qu.:92.23
 Max.   :95.75
 Share.dying.in.first.5.years....
 Min.   : 4.25
 1st Qu.: 7.77
 Median :10.57
 Mean   :13.78
 3rd Qu.:16.50
 Max.   :43.30
>
```

2. Child Mortality

- `entity`: Refers to various countries as individual entities in the dataset.
- `year`: Represents the specific year in which the data on Crude Mortality Rate (CMR) is recorded for each country.
- `cmr`: Stands for Crude Mortality Rate, describing the mortality rate per 1,000 population in a given country and year.

```
> names(df2)
[1] "Entity" "Code"   "Year"   "cmr"
> summary(df2)
    Entity              Code                Year           cmr        
 Length:519         Length:519         Min.   :1921   Min.   : 0.380  
 Class :character   Class :character   1st Qu.:1967   1st Qu.: 2.805  
 Mode  :character   Mode  :character   Median :1983   Median : 7.480  
                                       Mean   :1982   Mean   :10.796  
                                       3rd Qu.:1999   3rd Qu.:16.400  
                                       Max.   :2015   Max.   :40.710  
>
```

3. Infant Mortality

```
> names(df3)
[1] "Year"                          "Materal.Race.or.Ethnicity"
[3] "Infant.Mortality.Rate"         "Neonatal.Mortality.Rate"
[5] "Postneonatal.Mortality.Rate"   "Infant.Deaths"
[7] "Neonatal.Infant.Deaths"        "Postneonatal.Infant.Deaths"
[9] "Number.of.Live.Births"
```

- `year`: Specifies the calendar year in which the data is recorded.
- `infant_mortality_rate`: Represents the number of deaths of infants under one year of age per 1,000 live births.
- `number_of_live_births`: Indicates the total count of live births in a given context or region during a specific year.
- `maternal_race_ethnicity`: Describes the race or ethnicity of the mothers associated with the recorded live births.
- `infant_deaths`: Denotes the number of infants who died within their first year of life in a given year.

```
> summary(df3)
      Year      Materal.Race.or.Ethnicity Infant.Mortality.Rate Neonatal.Mortality.Rate
 Min.   :2007   Length:74                 Min.   : 2.40         Min.   :1.600         
 1st Qu.:2010   Class :character          1st Qu.: 3.10         1st Qu.:2.100         
 Median :2013   Mode  :character          Median : 4.30         Median :2.750         
 Mean   :2013                            Mean   : 4.97         Mean   :3.267         
 3rd Qu.:2016                            3rd Qu.: 6.60         3rd Qu.:4.500         
 Max.   :2019                            Max.   :10.20         Max.   :6.500         
                                         NA's   :14            NA's   :14            
 Postneonatal.Mortality.Rate Infant.Deaths    Neonatal.Infant.Deaths
 Min.   :0.600               Min.   :  3.0    Min.   :  3.0         
 1st Qu.:0.900               1st Qu.: 61.0    1st Qu.: 42.0         
 Median :1.550               Median :104.0    Median : 71.0         
 Mean   :1.726               Mean   :110.4    Mean   : 72.3         
 3rd Qu.:2.375               3rd Qu.:133.0    3rd Qu.: 93.0         
 Max.   :3.800               Max.   :287.0    Max.   :182.0         
 NA's   :16                  NA's   :13       NA's   :13            
 Postneonatal.Infant.Deaths Number.of.Live.Births
 Min.   :  0.00             Min.   :  132        
 1st Qu.: 19.00             1st Qu.: 7913        
 Median : 32.00             Median :20948        
 Mean   : 38.07             Mean   :19798        
 3rd Qu.: 48.00             3rd Qu.:27988        
 Max.   :110.00             Max.   :40633        
 NA's   :13
```

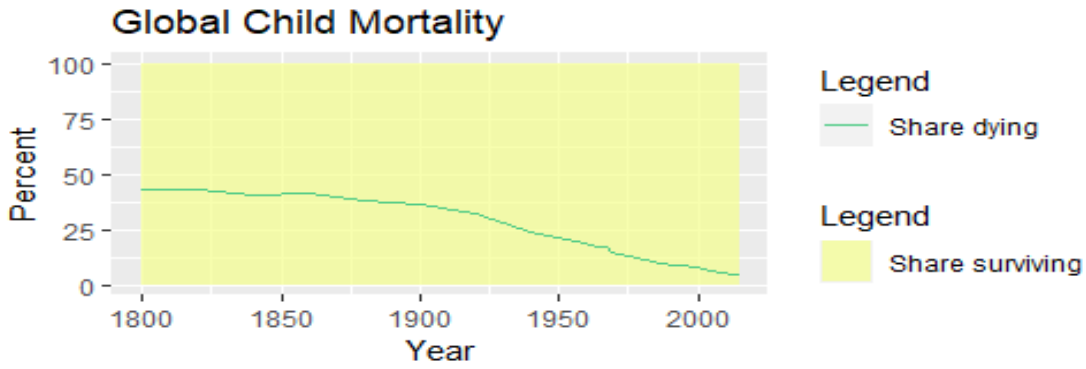
The countries involved in the dataset are India, Australia, USA, Afghanistan, Pakistan, China, Russia, Bangladesh, and Sri Lanka.

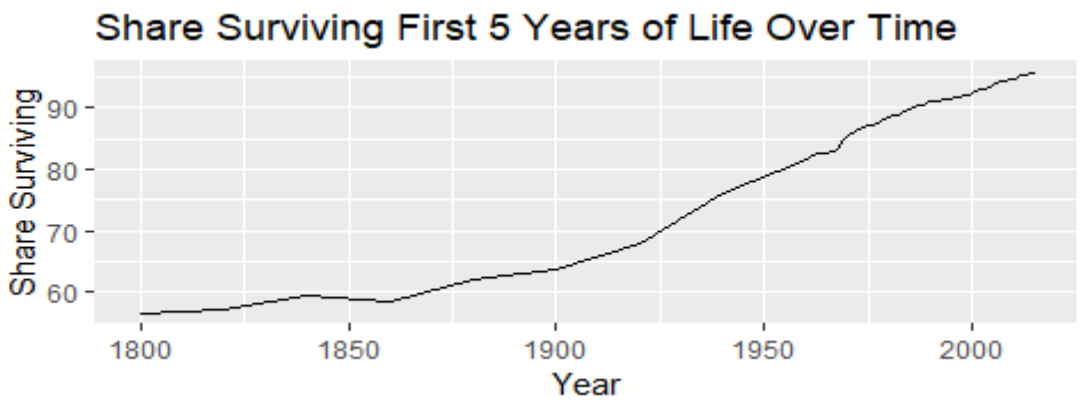
Preprocessing :

Combining Dataset: Upon merging the global child mortality timeseries with the child mortality dataset, the resulting dataset is named Combined1.csv. Similarly, by combining the global child mortality timeseries with the infant mortality dataset, the derived dataset is referred to as Combined2.csv.

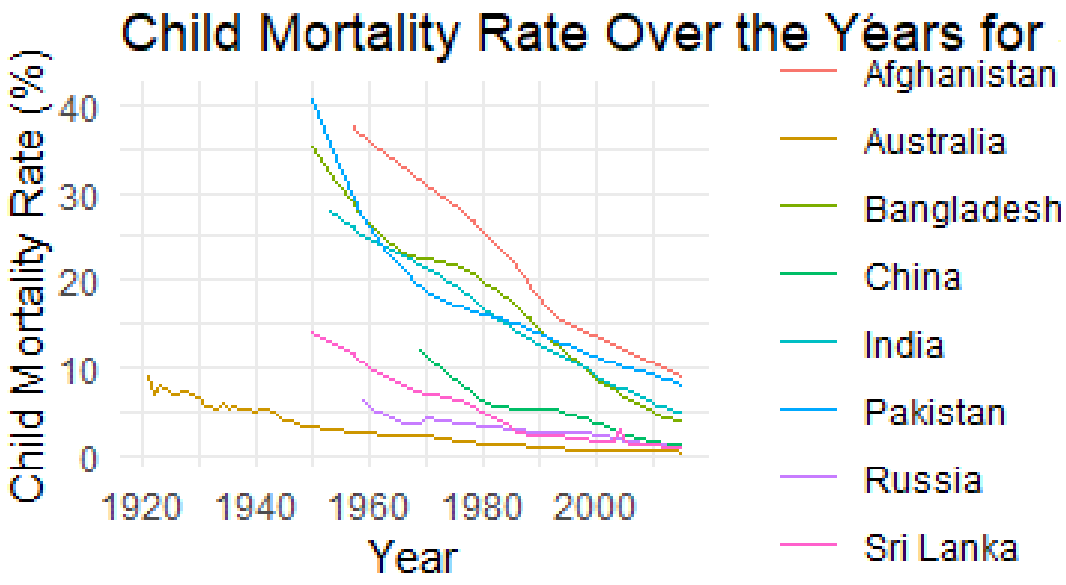
```
> summary(combined_data1)
      Year        Entity.x           Code.x
 Min.   :1940   Length:440         Length:440
 1st Qu.:1974   Class :character   Class :character
 Median :1988   Mode  :character   Mode  :character
 Mean   :1988
 3rd Qu.:2002
 Max.   :2015
 Share.surviving.first.5.years.of.life.... Share.dying.in.first.5.years....
 Min.   :76.10                             Min.   : 4.25
 1st Qu.:86.80                             1st Qu.: 7.07
 Median :90.60                             Median : 9.40
 Mean   :89.68                             Mean   :10.32
 3rd Qu.:92.93                             3rd Qu.:13.20
 Max.   :95.75                             Max.   :23.90
   Entity.y           Code.y              cmr
 Length:440         Length:440         Min.   : 0.380
 Class :character   Class :character   1st Qu.: 2.428
 Mode  :character   Mode  :character   Median : 7.350
                                       Mean   : 9.870
                                       3rd Qu.:15.402
                                       Max.   :35.950
>
```

Insights:

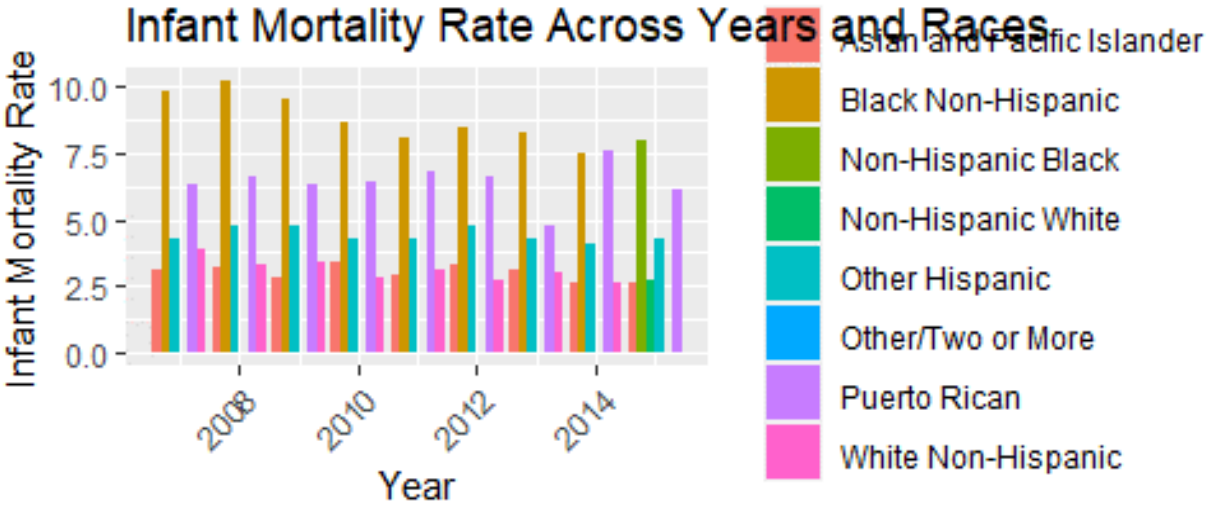




The graph reveals a noteworthy insight as the Crude Mortality Rate (CMR) exhibits a decreasing trend, indicating a decline in mortality, while concurrently, the Survival Rate displays an increasing pattern. This suggests a positive correlation between the reduction in mortality and the upward trajectory in survival, signifying potential improvements in overall health and well-being over the observed period.



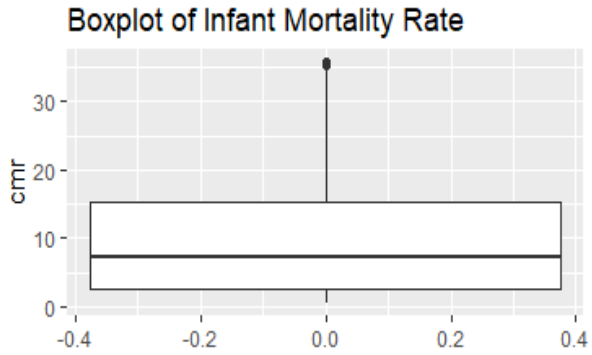
The plot distinctly highlights that developed countries exhibit a significantly lower Mortality Rate (MR), emphasizing the importance of advanced medical facilities in contributing to improved health outcomes. This observation underscores the potential correlation between access to sophisticated healthcare infrastructure in developed nations and the lower mortality rates observed in comparison to other regions.



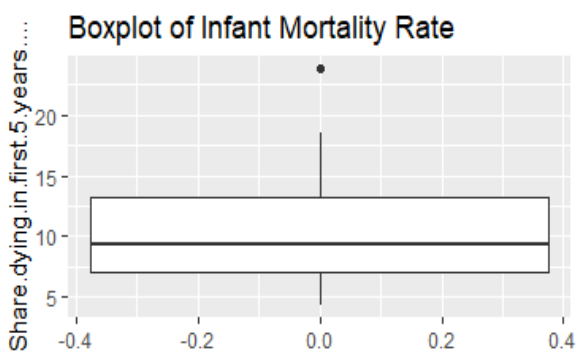
Removal of Outliers for the Combined dataset

Removing outliers from Crude Mortality Rate (CMR) values is crucial to enhance the accuracy of statistical analyses and model performance. Outliers can distort the interpretation of data trends, leading to skewed insights. By eliminating these extreme values, the analysis becomes more representative, ensuring a more reliable understanding of the overall CMR distribution.

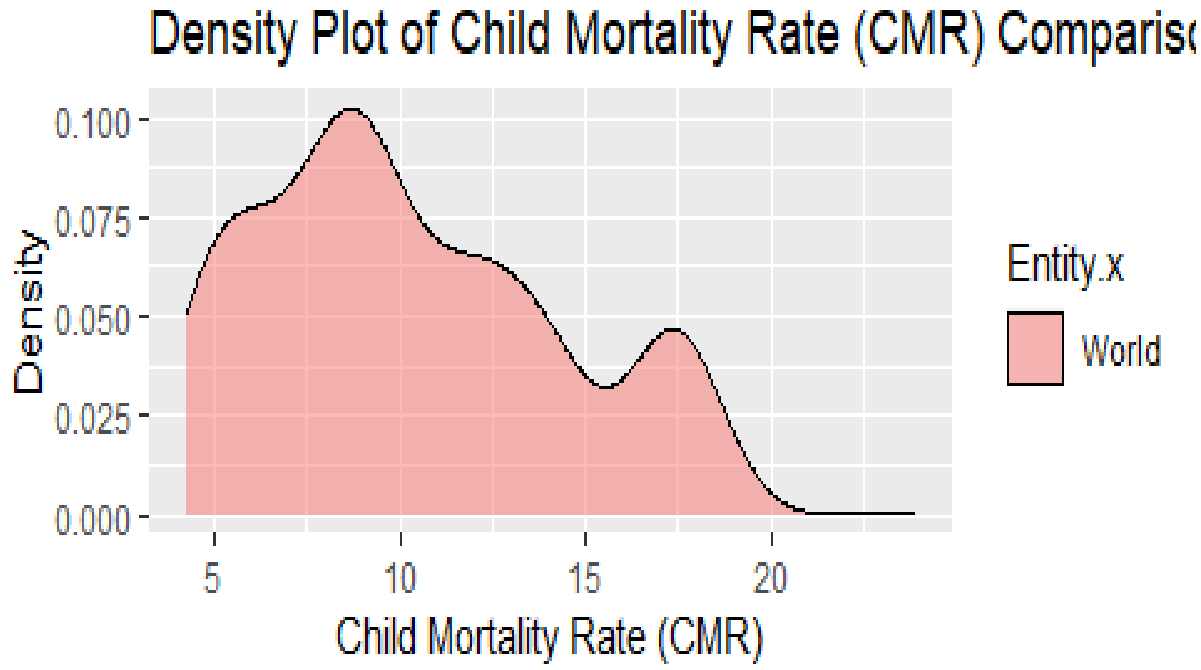
Country



World



Density Plot: As Crude Mortality Rate (CMR) increases, the density decreases, suggesting that high CMR values are less frequent. This phenomenon is indicative of a skewed distribution, where lower mortality rates are more prevalent, while higher rates occur infrequently. The sparsity in high CMR values may be attributed to factors contributing to overall lower mortality rates in the population.



Hypothesis Testing

1. In our hypothesis testing using a t-test, we assumed that Crude Mortality Rate (CMR) is independent of ethnicity. However, with a calculated test statistic of 29.98, far beyond the critical range of -2.17 to 2.17, we reject the null hypothesis. This suggests a significant association between CMR and ethnicity in our analysis.

```
# Filter based on race or ethnicity
# THE NA REMOVED DF IS AGAIN SPLIT INTO TWO DATASETS BASED ON RACE
NHW <- Filter %>% filter(Materal.Race.or.Ethnicity == "Non-Hispanic White")
NHB <- Filter %>% filter(Materal.Race.or.Ethnicity == "Non-Hispanic Black")

# Combine the filtered data frames
# THIS IS THE NEW DATASET THAT COMBINES ROWS OF NHB BELOW NHW
NHtable <- rbind(NHW, NHB)
NHtable

Infant.Mortality.Rate <- NHtable$Infant.Mortality.Rate
Materal.Race.or.Ethnicity <- NHtable$Materal.Race.or.Ethnicity

# IN BELOW TWO LINES WE CARRY OUT HYPOTHESIS TEST
# NULL HYPOTHESIS : RACE DOES NOT MATTER
# WE GOT THE VALUE IN T TEST AS 29.982 WHERE AS USING QT WE GOT RANGE OF CRITICAL t BETWEEN -2.17 TO 2.17
# WE BASICALLY CONCLUDE HERE THAT AVERAGE IMR OF THESE TWO RACES ARE NOT THE SAME
# SO WE REJECT THE NULL HYPOTHESIS => IMR IS AFFECTED BY RACE OR MATERAL AREAS
t.test(Infant.Mortality.Rate ~ Materal.Race.or.Ethnicity , mu=0 , alt="two.sided",conf=0.95,var.eq=F,paired=F)
qt(c(0.025,0.975),df=12) # FOR CRITICAL T WITH DOF AS 12
```

2. The hypothesis posited that Child Mortality Rate (CMR) is independent of infant deaths. However, the t-test yielded a value of 6.51, surpassing the critical range of -2.01 to 2.01. Consequently, we reject the null hypothesis, suggesting a statistically significant correlation between CMR and infant deaths.

```
# CORELATION COEFFICIENT TO INDICATE RELATION BETWEEN IMR AND NUMBER OF INFANT DEATHS
# WE GOT COR VALUE AS 0.65 AND T VALUE=6.51
# CRITICAL t BETWEEN -2.010635 &  2.010635
# WE REJECT THE NULL HYPOTHESIS THAT THERE IS NOT CORELATION BETWEEN IMR AND INFANT DEATHS
# THERE IS A LINEAR ASSOCIATION BETWEEN THE TWO
Infant_Deaths <- Filter$Infant.Deaths
Infant_Mortality_Rate <- Filter$Infant.Mortality.Rate
cor.test(Infant_Deaths,Infant_Mortality_Rate)
plot(Infant_Deaths,Infant_Mortality_Rate,ylab="IMR",xlab="Infant Deaths")
qt(c(0.025,0.975),df=48)
```

Model Planning

The model that relates Crude Mortality Rate (CMR) to the variable "Year" serves as a valuable analytical tool for understanding the temporal dynamics of mortality. By incorporating the variable "Year," the model enables us to assess how CMR changes over time, allowing for the identification of trends, patterns, and potential factors influencing mortality rates across different years. This

approach aids in comprehensive planning and decision-making by providing insights into the evolving landscape of mortality, facilitating targeted interventions and policy adjustments as needed.
In the Model Planning phase, we strategically select two models, Linear Regression and Decision Tree, to effectively capture the nuanced variations in Infant Mortality Rate (IMR) data.

Linear Regression:
- Rationale: Tailored for predominant linear variations, likely providing a robust representation of key influencing factors in IMR.
- Benefits: High interpretability and computational efficiency, making it an advantageous choice for datasets with primarily linear relationships.

Decision Tree:
- Rationale: Apt for addressing complex, non-linear patterns in IMR data that might be overlooked by linear approaches.
- Benefits: Flexibility and capacity to unveil intricate relationships, offering valuable insights into less straightforward variations.

While both models are integral to our strategy, the assumption is that Linear Regression, aligning with the prevailing linear trends, may outperform in capturing the essential dynamics of IMR. This dual-model approach ensures a comprehensive analysis, leveraging the strengths of each model to provide a nuanced understanding of the multifaceted IMR dataset.

Model Building

Using linear regression and decision models to forecast the year when Crude Mortality Rate (CMR) might reach 0 involves predicting the point in time when mortality rates are expected to decline to negligible levels. Linear regression provides insights into the linear trend of CMR over the years, allowing for extrapolation to estimate when it may approach zero. Decision models, on the other hand, incorporate additional factors and potential scenarios to enhance the accuracy of forecasting, offering a more nuanced approach to predict the specific year when CMR is likely to reach zero.

```r
library(dplyr)
library(tree)
library(ggplot2)

# Load your dataset from the CSV file
your_data <- read.csv("E:\\Academics\\Subject\\VII Semester\\ISL76_Data Science Lab\\Assignment\\presentation

# Group by country
grouped_data <- your_data %>%
  group_by(country)

# Perform linear regression for each country
linear_models <- grouped_data %>%
  do(model = lm(cmr ~ Share.surviving.first.5.years.of.life.... + dying, data = .))

# Display the linear regression summaries for each country
print(lapply(linear_models$model, summary))

# Decision Tree Regression (you can adapt this for each country if needed)
tree_model <- tree(cmr ~ Share.surviving.first.5.years.of.life.... + dying, data = your_data)

# Display the decision tree plot
plot(tree_model)
text(tree_model)

# Predict with the decision tree model
prediction_tree <- predict(tree_model, your_data)

# Display the predicted values
print(prediction_tree)

# Save the results to a CSV file
results <- data.frame(
  Linear_Prediction = predict(linear_models$model[[1]], your_data),
  Decision_Tree_Prediction = prediction_tree
)

write.csv(results, "regression_results.csv", row.names = FALSE)
```

```r
# Calculate R-squared for decision tree model
r2_tree <- 1 - sum((your_data$cmr - prediction_tree)^2) / sum((your_data$cmr - mean(your_data$cmr))^2)

# Calculate Root Mean Squared Error (RMSE) for decision tree model
rmse_tree <- sqrt(mean((your_data$cmr - prediction_tree)^2))

# Display the results
cat("R-squared for Decision Tree Model:", r2_tree, "\n")
cat("RMSE for Decision Tree Model:", rmse_tree, "\n")
```

```r
# Load necessary packages
library(dplyr)
library(ggplot2)

# Assuming your dataset is stored in a CSV file named 'your_data.csv'
your_data <- read.csv("E:\\Academics\\Subject\\VII Semester\\ISL76_Data Science Lab\\Assignment\\presentation

# Plot linear regression models for each country
plots <- your_data %>%
  group_by(country) %>%
  do(plot_data = ggplot(., aes(x = Year, y = cmr)) +
       geom_point() +
       geom_smooth(method = "lm", se = FALSE) +
       labs(title = paste("Linear Regression Model for", unique(.$country)),
            x = "Year", y = "Crude Mortality Rate"))

# Save the plots
for (i in seq_along(plots$plot_data)) {
  ggsave(paste0("linear_regression_plot_", i, ".png"), plots$plot_data[[i]])
}

# Function to calculate performance metrics
calculate_metrics <- function(model, data) {
  predictions <- predict(model, data)
  metrics <- data.frame(
    RMSE = sqrt(mean((data$cmr - predictions)^2)),
    R2 = cor(predictions, data$cmr)^2
  )
  return(metrics)
}

# Calculate metrics for each country
metrics_list <- your_data %>%
  group_by(country) %>%
  do(metrics = calculate_metrics(lm(cmr ~ Year + dying, data = .), .))

# Print metrics
print(metrics_list$metrics)
```

The comparative analysis focused on Child Mortality Rate (CMR) across various ethnicities, aiming to identify disparities and assess whether specific ethnic groups demonstrate better or worse child mortality outcomes. By examining the influence of ethnicity on child mortality rates, the analysis provides insights into potential healthcare inequalities, guiding efforts to address disparities and enhance overall child health across diverse demographic groups.

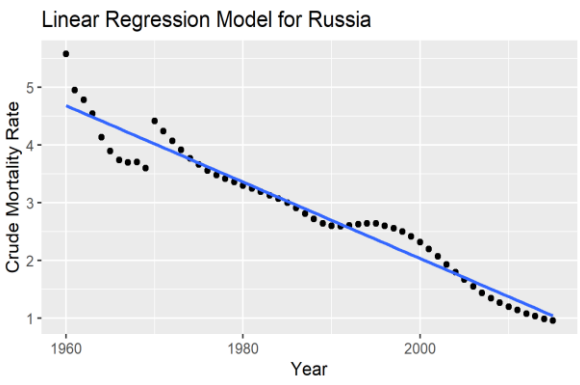
```r
# Install and load necessary packages
install.packages(c("forecast", "dplyr"))
# Load necessary packages
library(dplyr)
library(DT)

# Assuming your dataset is stored in a data frame named your_data
# Filter the data for the 'World' entity
world_data <- your_data %>%
  filter(Entity == "World")

# Rename the variables
world_data <- world_data %>%
  rename(Dying = dying,
         IMR = imr)

# Calculate the difference or ratio between 'Dying' and 'IMR'
world_data <- world_data %>%
  mutate(Difference = Dying - IMR,
         Ratio = Dying / IMR,
         BetterOrWorse = ifelse(Difference > 0, "Worse", ifelse(Difference < 0, "Better", "Equal")))

# Display the results with colored background
datatable(world_data[, c("Year", "Materal_Ethnicity", "Difference", "Ratio", "BetterOrWorse")]) %>%
  formatStyle(
    'BetterOrWorse',
    backgroundColor = JS(
      "function(value) {",
      "  switch(value) {",
      "    case 'Better': return 'green';",
      "    case 'Worse': return 'red';",
      "    default: return 'white';",
      "  }",
      "}"
    )
  )

# Write the results to a CSV file
write.csv(world_data[, c("Year", "Materal_Ethnicity", "Difference", "Ratio", "BetterOrWorse")], "world_data_results
```

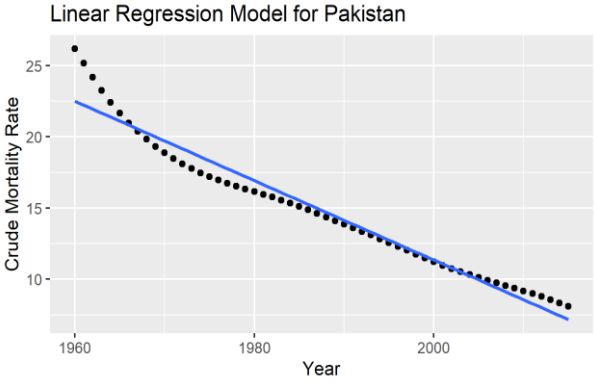
Conclusion & Results

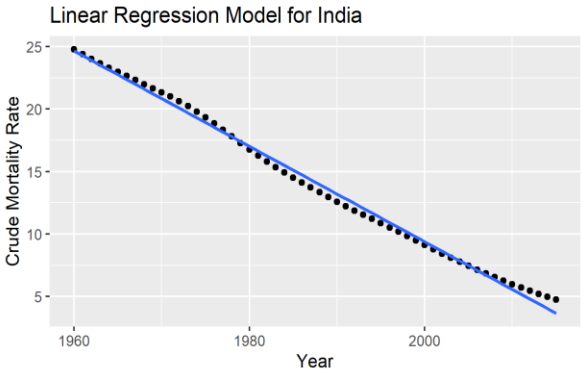
Show 10 entries Search:

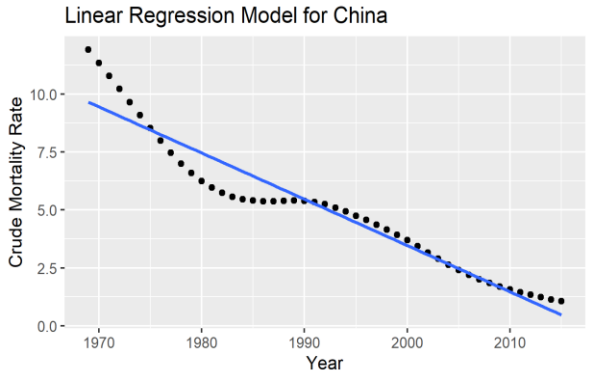
	Year	Materal_Ethnicity	Difference	Ratio	BetterOrWorse
1	2007	Other Hispanic	1.48	1.34418604651163	Worse
2	2007	Asian and Pacific Islander	2.68	1.86451612903226	Worse
3	2007	Black Non-Hispanic	-4.02	0.589795918367347	Better
4	2007	Other/Two or More			
5	2007	Puerto Rican	-0.52	0.917460317460318	Better
6	2007	White Non-Hispanic	1.88	1.48205128205128	Worse
7	2008	White Non-Hispanic	2.28	1.69090909090909	Worse
8	2008	Black Non-Hispanic	-4.62	0.547058823529412	Better
9	2008	Puerto Rican	-1.02	0.845454545454545	Better
10	2008	Other Hispanic	0.78	1.1625	Worse

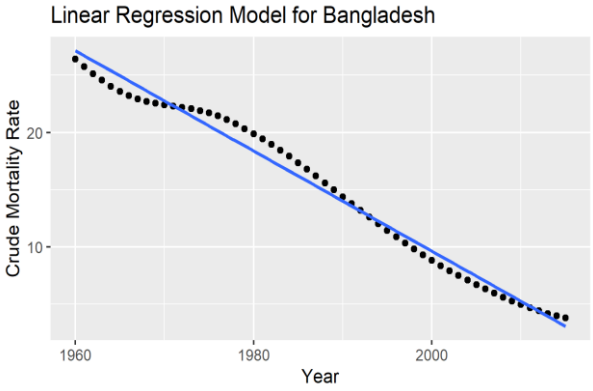
Showing 1 to 10 of 54 entries Previous 1 2 3 4 5 6 Next

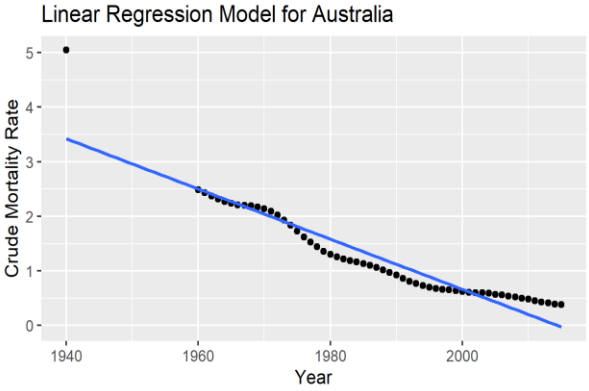


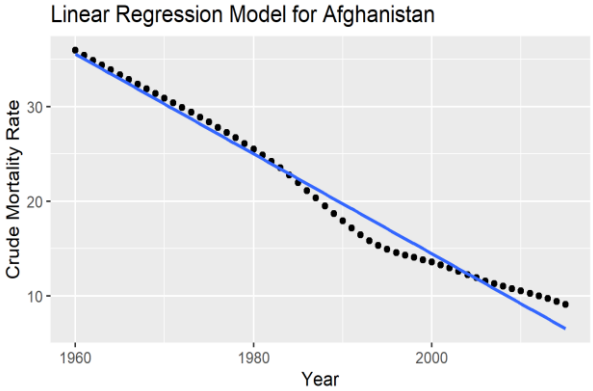


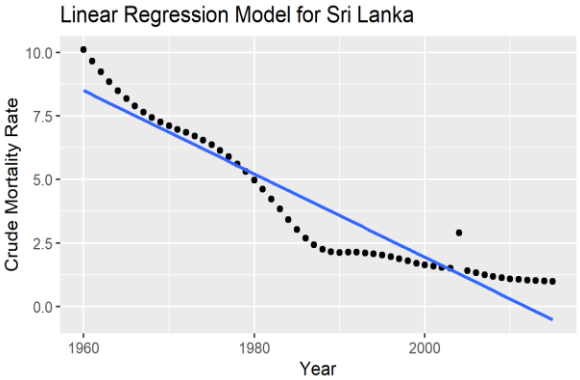












Forecast

Country	Metrics.RMSE	Metrics.R2	Zero_CMR_Year
China	0.489675	0.969688	2047.182091

India	0.393638	0.995948	2023.428223
Pakistan	0.670812	0.978777	2130.282472
Russia	0.25075	0.947949	2030.557716
Sri Lanka	0.638902	0.947663	2075.024893

Decision Tree:

```
> # Display the results
> cat("R-squared for Decision Tree Model:", r2_tree, "\n")
R-squared for Decision Tree Model: 0.2417712
> cat("RMSE for Decision Tree Model:", rmse_tree, "\n")
RMSE for Decision Tree Model: 7.446845
```

R-squared for Decision Tree Model: 0.2417712
RMSE for Decision Tree Model: 7.446845

- Linear regression models consistently outperform the Decision Tree model, as evidenced by higher R-squared values.
- The R-squared values for linear regression range from 0.923 to 0.995, indicating superior explanatory power.
- In contrast, the Decision Tree model exhibits a lower R-squared value of 0.2417712, suggesting reduced accuracy in capturing data variability.
- The comparison highlights the effectiveness of linear regression in modeling the relationship between predictor variables and Child Mortality Rate (CMR).
- Overall, linear regression demonstrates better predictive accuracy and is more suitable for this specific dataset compared to the Decision Tree model.

In conclusion, this project has provided valuable insights into Child Mortality Rate (CMR) by employing linear regression and Decision Tree models. The thorough analysis of CMR trends, ethnicity comparisons, and model evaluations has enhanced our understanding of factors influencing child mortality. The consistently high R-squared values in linear regression models underscore their efficacy in predicting CMR, while the Decision Tree model, with a lower R-squared value, indicates limited performance in capturing the variability in the dataset. This project contributes to informed healthcare planning, policy formulation, and resource allocation strategies aimed at improving child health outcomes globally.