Probability and Statistics

MT2005-Probability and Statistics

**Course Instructor**

Dr. Harris Khurram

**Project**

Overseas Pakistani Population and Remittance

**Authors**

Shahmeer Hussain (21F-9432)

Noman Ijaz(21F-9513)

**Date of submission**

15 May, 2023

Spring 2023



Department of Software Engineering

FAST–National University of Computer & Emerging Sciences Chiniot – Faisalabad Campus

**Overseas Pakistani Population and Remittance Dataset**

**Problem Statement:**

The project aims to analyze and visualize data related to overseas Pakistanis, focusing on factors such as remittance, population by country, population by continent, and year. The goal is to gain insights into the trends and patterns associated with the overseas Pakistani community and their contributions to the country's economy. By exploring the data and creating informative visualizations, the project aims to answer questions such as:

How has the remittance from overseas Pakistanis evolved over the years? Are there any noticeable trends or fluctuations?

What are the countries with the highest overseas Pakistani population, and how does it vary across different regions and continents?

Is there any correlation between the year and the remittance amount?

Can a regression model be built to predict future remittance based on the year?

How does the distribution of remittance values look like? Are there any significant clusters or outliers?

**Objectives:**

1. Collect and import the overseas Pakistani data from a CSV file into the Shiny application.

2. Display the imported data in a tabular format for easy reference and exploration.

3. Generate histograms to visualize the distribution of remittance, population by country, population by continent, and year.

4. Create pie charts to represent the proportion of remittance, population by country, population by continent, and year.

5. Construct bar plots to compare and visualize the remittance, population by country, population by continent, and year.

6. Compute summary statistics for the dataset, including measures of central tendency and variability.

7. Implement a regression model to analyze the relationship between remittance and year, providing insights into the potential future remittance trends.

8. Develop a distribution plot to visualize the density and shape of the remittance values.

9. Customize the appearance of the Shiny application, including the arrangement and colors of the tabs and the body.

10. Ensure the application is user-friendly, allowing users to upload their own CSV file for analysis and visualization.

11. Provide clear and informative labels, titles, and axis descriptions for all plots and visualizations.

12. Document the findings and insights obtained from the data analysis and visualizations in a concise and understandable manner.

13. Test the Shiny application for functionality and address any potential errors or bugs.

14. Present the results and conclusions of the project, highlighting the key insights and implications.

By achieving these objectives, the project aims to facilitate the exploration, analysis, and visualization of overseas Pakistani data, enabling a better understanding of their contributions and patterns over time.

**Data Discerption:**

<https://www.kaggle.com/datasets/tayyarhussain/overseas-pakistani-population-and-remittance>

This dataset provides a comprehensive overview of the overseas Pakistani population across different continents and countries, along with their remittance statistics. The dataset includes information about the country, article, continent, and year, as well as the number of overseas Pakistani residents and their remittance amounts in US dollars. The dataset also contains additional information on the overseas Pakistani population by continent and their overall population.

Country: The name of the country where the Pakistani expatriates are residing.

Article: Any relevant article or information associated with the country, such as its political or economic situation.

Continent: The name of the continent where the country is located.

Overseas Pakistani population (By Country): The estimated number of Pakistani expatriates residing in the country.

Extra Information: Any additional information that is relevant to the Pakistani expatriate population in the country.

Overseas Pakistani Population by Continent: The estimated number of Pakistani expatriates residing in the continent.

Population (by Continent): The estimated population of the continent.

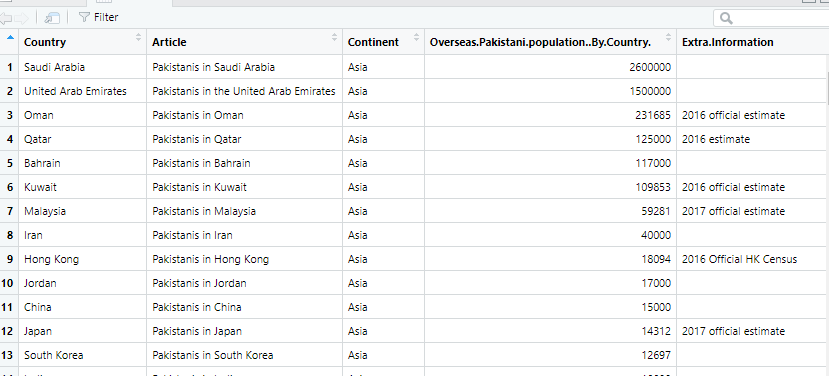
Year: The year to which the data corresponds Remittance received.

Remittance ($ billion): The amount of remittance in US dollars sent by the Pakistani expatriate population from the country to Pakistan during the corresponding year.

This dataset is an essential resource for researchers, policymakers, and anyone interested in studying or analyzing the trends and patterns of the Pakistani diaspora around the world. The data can help researchers identify areas of high remittance inflows and study their economic impact on the country of origin. It can also be used by policymakers to formulate policies that cater to the needs and challenges of Pakistani expatriates in different countries. Finally, this dataset can also be used to understand the social and cultural dynamics of the Pakistani community in different parts of the world.

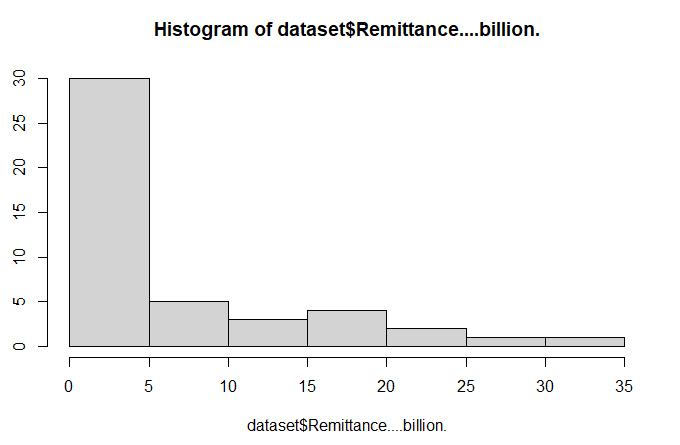
**Outputs:**

**Dataset:**

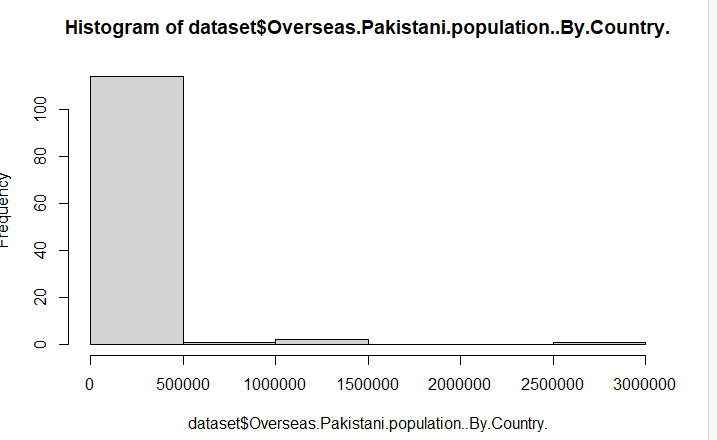


**Histograms:**

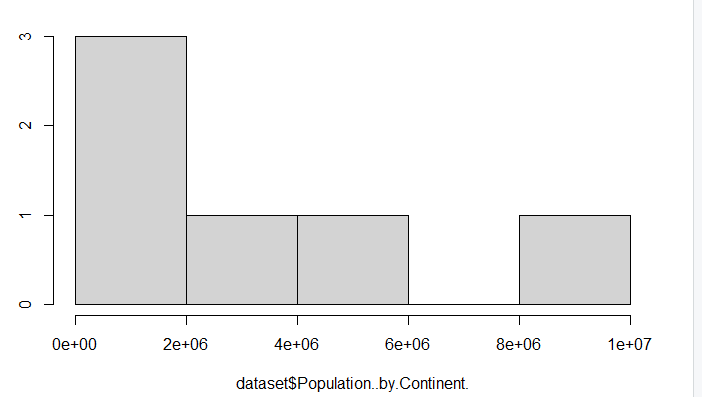
**Remittance:**



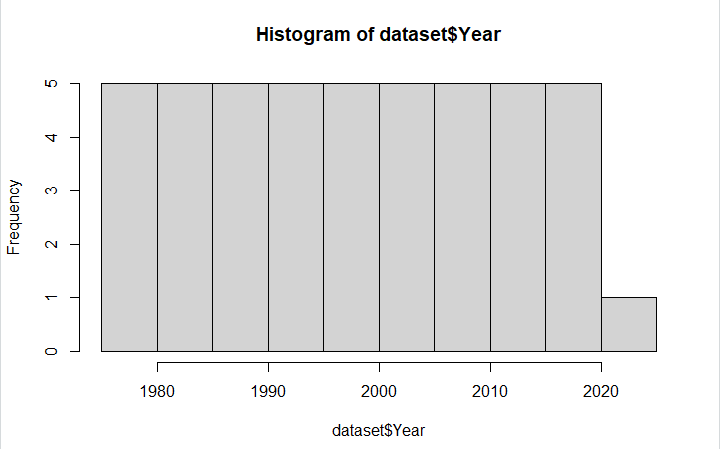
**Overseas Population by country:**



**Population By Continent:**

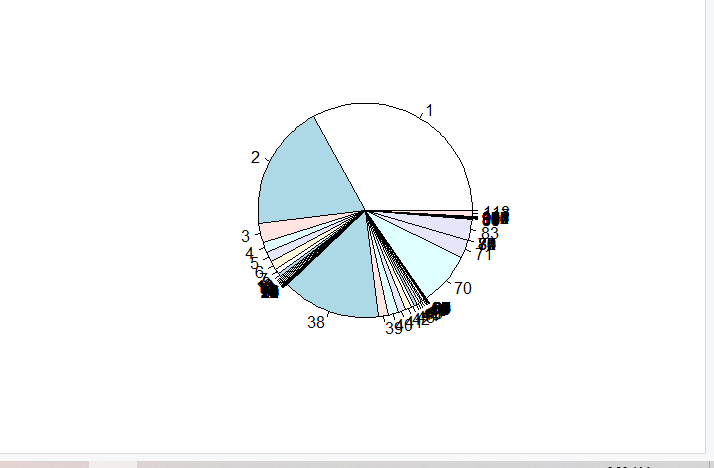


**Year:**



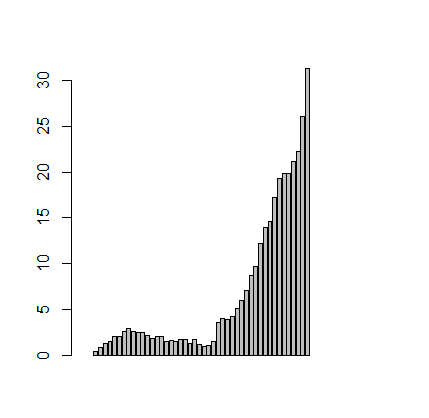
**Pie Charts :**

**Overseas Pakistani Population By country:**

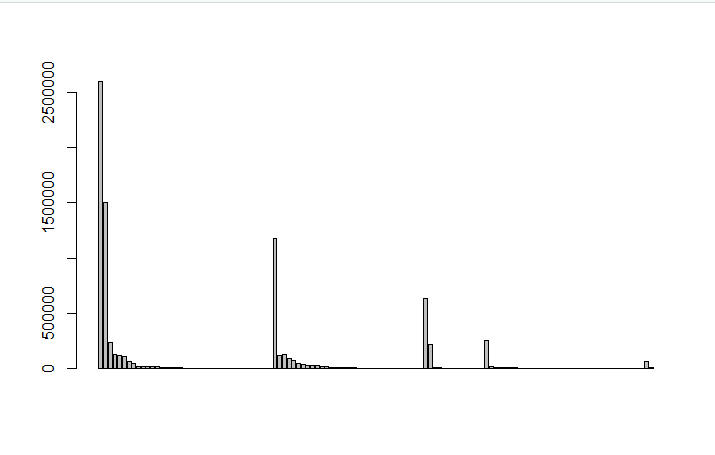


**Bar Plots :**

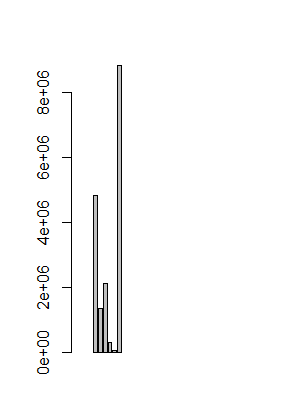
**Remittance:**



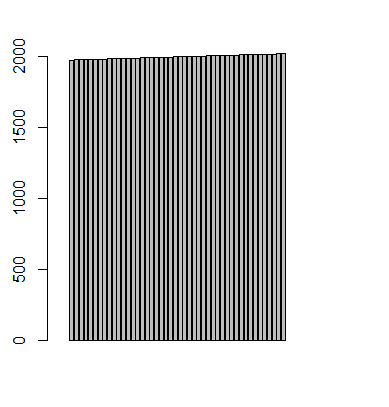
**Overseas Pakistani Population By country:**



**Overseas Pakistani Population By continent:**



**Year:**



**Mean:**

**Remittance:**



**Overseas Pakistani Population By country:**



**Overseas Pakistani Population By continent:**



**Year:**



**Median:**

**Remittance:**



**Overseas Pakistani Population By country:**



**Overseas Pakistani Population By continent:**



**Year:**



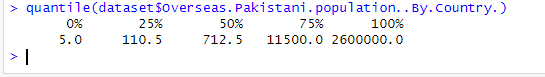
**Variance:**

**Overseas Pakistani Population By country:**

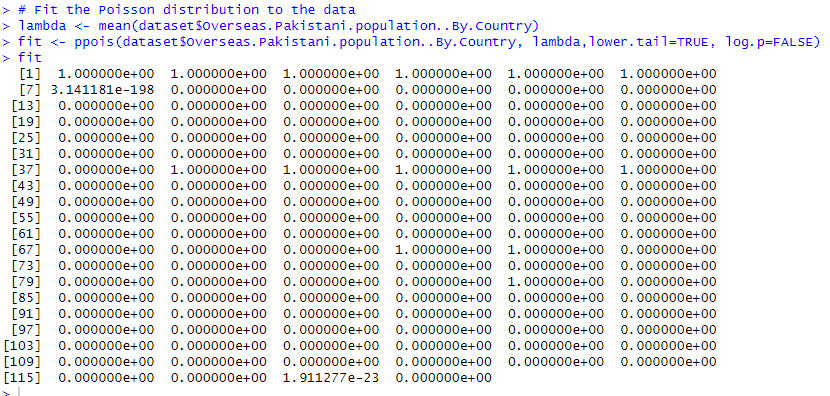


**Quantile:**

**Overseas Pakistani Population By country:**

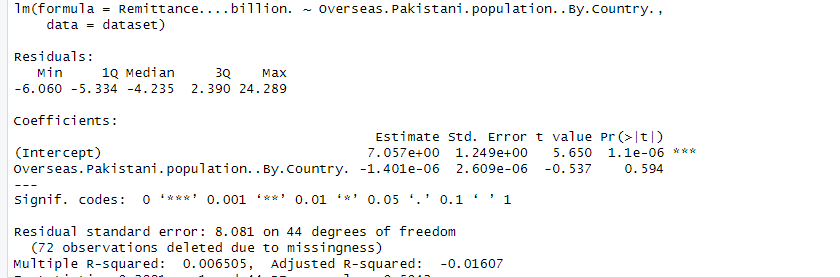


**Poisson Distribution:**



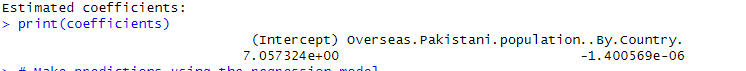
**Linear Regression Model:**

I have chosen the linear Regression model

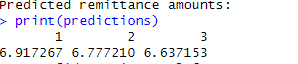


**Predictions:**

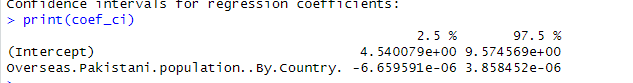
**Estimated:**



**Predicted:**



**Confidence intervals:**



**R code:**

library(ggplot2)

library(dplyr)

library(tidyr)

library(readr)

# Replace "path/to/your/dataset.csv" with the actual path to your dataset file

dataset <- read.csv("C:/Users/shahmeer/Desktop/Overseas Pakistani Population and Remittance By Years.csv")

View(dataset)

hist(dataset$Remittance....billion.)

hist(dataset$Overseas.Pakistani.population..By.Country.)

hist(dataset$Population..by.Continent.)

hist(dataset$Year)

pie(dataset$Remittance....billion.)

pie(dataset$Overseas.Pakistani.population..By.Country.)

pie(dataset$Population..by.Continent.)

pie(dataset$Year)

barplot(dataset$Remittance....billion.)

barplot(dataset$Overseas.Pakistani.population..By.Country.)

barplot(dataset$Population..by.Continent.)

barplot(dataset$Year)

mean(dataset$Remittance....billion.)

mean(dataset$Overseas.Pakistani.population..By.Country.)

mean(dataset$Population..by.Continent.)

mean(dataset$Year)

median(dataset$Remittance....billion.)

median(dataset$Overseas.Pakistani.population..By.Country.)

median(dataset$Population..by.Continent.)

median(dataset$Year)

var(dataset$Remittance....billion.)

var(dataset$Overseas.Pakistani.population..By.Country.)

var(dataset$Population..by.Continent.)

var(dataset$Year)

quantile(dataset$Remittance....billion.)

quantile(dataset$Overseas.Pakistani.population..By.Country.)

quantile(dataset$Population..by.Continent.)

quantile(dataset$Year)

data <- read.csv("C:/Users/shahmeer/Desktop/Overseas Pakistani Population and Remittance By Years.csv")

# Perform exploratory data analysis and determine the probability distribution

# Analyze the variables in the dataset and check their distributions

summary(data)

# Assuming 'dataset' is the variable containing your dataset

# Assuming 'Overseas.Pakistani.population..By.Country' is the column of interest

# Fit the Poisson distribution to the data

lambda <- mean(dataset$Overseas.Pakistani.population..By.Country)

fit <- ppois(dataset$Overseas.Pakistani.population..By.Country, lambda,lower.tail=TRUE, log.p=FALSE)

fit

# Access the estimated parameters

lambda\_hat <- fit$estimate

# Print the estimated parameter

cat("Estimated lambda (Poisson):", lambda\_hat, "\n")

# Assuming 'dataset' is the variable containing your dataset

# Assuming 'Remittance....billion.' is the dependent variable

# Assuming 'Overseas.Pakistani.population..By.Country.' is the independent variable

# Fit the linear regression model

model <- lm(Remittance....billion. ~ Overseas.Pakistani.population..By.Country., data = dataset)

# Print the summary of the regression model

summary(model)

# Access the estimated coefficients

coefficients <- coef(model)

# Print the estimated coefficients

cat("Estimated coefficients:\n")

print(coefficients)

# Make predictions using the regression model

new\_data <- data.frame(Overseas.Pakistani.population..By.Country. = c(100000, 200000, 300000))

predictions <- predict(model, newdata = new\_data)

# Print the predicted values

cat("Predicted remittance amounts:\n")

print(predictions)

# Assuming 'dataset' is the variable containing your dataset

# Assuming 'Remittance....billion.' is the dependent variable

# Assuming 'Overseas.Pakistani.population..By.Country.' is the independent variable

# Confidence interval for mean

mean <- confint(t.test(dataset$Remittance....billion.))

cat("Confidence interval for mean:\n")

print(mean)

# Confidence interval for median

median <- quantile(dataset$Remittance....billion., probs = c(0.025, 0.975))

cat("Confidence interval for median:\n")

print(median)

# Confidence interval for regression coefficients

coef\_ci <- confint(model)

cat("Confidence intervals for regression coefficients:\n")

print(coef\_ci)

**R Shinny Code:**

library(shiny)

library(ggplot2)

library(dplyr)

library(tidyr)

library(readr)

# UI

ui <- fluidPage(

titlePanel("Overseas Pakistani Data Analysis"),

sidebarLayout(

sidebarPanel(

fileInput("file", "Choose CSV File",

accept = c(".csv")

),

br(),

actionButton("analyzeBtn", "Analyze Data")

),

mainPanel(

tabsetPanel(

id = "tabset", # Set an ID for tabsetPanel

tabPanel("Data", dataTableOutput("data\_table")),

tabPanel("Histograms",

plotOutput("hist\_remittance"),

plotOutput("hist\_population\_country"),

plotOutput("hist\_population\_continent"),

plotOutput("hist\_year")

),

tabPanel("Pie Charts",

plotOutput("pie\_remittance"),

plotOutput("pie\_population\_country"),

plotOutput("pie\_population\_continent"),

plotOutput("pie\_year")

),

tabPanel("Bar Plots",

plotOutput("bar\_remittance"),

plotOutput("bar\_population\_country"),

plotOutput("bar\_population\_continent"),

plotOutput("bar\_year")

),

tabPanel("Summary Statistics",

verbatimTextOutput("summary\_stats")

),

tabPanel("Regression Model and Distribution",

plotOutput("regression\_plot"),

plotOutput("distribution\_plot")

)

)

)

)

)

# Server

server <- function(input, output) {

dataset <- reactive({

req(input$file)

read.csv(input$file$datapath)

})

output$data\_table <- renderDataTable({

dataset()

})

output$hist\_remittance <- renderPlot({

ggplot(dataset(), aes(x = Remittance....billion.)) +

geom\_histogram(fill = "lightblue", bins = 20) +

labs(x = "Remittance (billion)", y = "Frequency", title = "Histogram of Remittance")

})

output$hist\_population\_country <- renderPlot({

ggplot(dataset(), aes(x = Overseas.Pakistani.population..By.Country.)) +

geom\_histogram(fill = "lightblue", bins = 20) +

labs(x = "Overseas Pakistani Population (by Country)", y = "Frequency", title = "Histogram of Overseas Pakistani Population (by Country)")

})

output$hist\_population\_continent <- renderPlot({

ggplot(dataset(), aes(x = Population..by.Continent.)) +

geom\_histogram(fill = "lightblue", bins = 20) +

labs(x = "Population (by Continent)", y = "Frequency", title = "Histogram of Population (by Continent)")

})

output$hist\_year <- renderPlot({

ggplot(dataset(), aes(x = Year)) +

geom\_histogram(fill = "lightblue", bins = 20) +

labs(x= "Year", y = "Frequency", title = "Histogram of Year")

})

output$pie\_remittance <- renderPlot({

ggplot(dataset(), aes(x = "", y = Remittance....billion.)) +

geom\_bar(stat = "identity", fill = rainbow(nrow(dataset()))) +

coord\_polar(theta = "y") +

labs(title = "Pie Chart of Remittance")

})

output$pie\_population\_country <- renderPlot({

ggplot(dataset(), aes(x = "", y = Overseas.Pakistani.population..By.Country.)) +

geom\_bar(stat = "identity",fill = rainbow(nrow(dataset()))) +

coord\_polar(theta ="y") +

labs(title = "Pie Chart of Overseas Pakistani Population (by Country)")

})

output$pie\_population\_continent <- renderPlot({

ggplot(dataset(), aes(x = "", y = Population..by.Continent.)) +

geom\_bar(stat = "identity", fill = rainbow(nrow(dataset()))) +

coord\_polar(theta = "y") +

labs(title = "Pie Chart of Population (by Continent)")

})

output$pie\_year <- renderPlot({

ggplot(dataset(), aes(x = "", y = Year)) +

geom\_bar(stat = "identity", fill = rainbow(nrow(dataset()))) +

coord\_polar(theta = "y") +

labs(title = "Pie Chart of Year")

})

output$bar\_remittance <- renderPlot({

ggplot(dataset(), aes(x = "", y = Remittance....billion.)) +

geom\_bar(stat = "identity", fill = "lightblue") +

labs(title = "Bar Plot of Remittance")

})

output$bar\_population\_country <- renderPlot({

ggplot(dataset(), aes(x = "", y = Overseas.Pakistani.population..By.Country.)) +

geom\_bar(stat = "identity", fill = "lightblue") +

labs(title = "Bar Plot of Overseas Pakistani Population (by Country)")

})

output$bar\_population\_continent <- renderPlot({

ggplot(dataset(), aes(x = "", y = Population..by.Continent.)) +

geom\_bar(stat = "identity", fill = "lightblue") +

labs(title = "Bar Plot of Population (by Continent)")

})

output$bar\_year <- renderPlot({

ggplot(dataset(), aes(x = "", y = Year)) +

geom\_bar(stat = "identity", fill = "lightblue") +

labs(title = "Bar Plot of Year")

})

output$summary\_stats <- renderPrint({

stats <- summary(dataset())

stats$stats <- NULL # Remove the "stats" column from the summary table

stats

})

output$regression\_plot <- renderPlot({

# Perform regression analysis and create a plot

# Replace the code below with your own regression analysis code

lm\_model <- lm(Remittance....billion. ~ Year, data = dataset())

plot(dataset()$Year, dataset()$Remittance....billion.,

xlab = "Year", ylab = "Remittance (billion)",

main = "Regression Analysis: Remittance over Time")

abline(lm\_model, col = "red")

})

output$distribution\_plot <- renderPlot({

# Create a distribution plot

# Replace the code below with your own distribution plot code

ggplot(dataset(), aes(x = Remittance....billion.)) +

geom\_density(fill = "lightblue") +

labs(x = "Remittance (billion)", y = "Density",

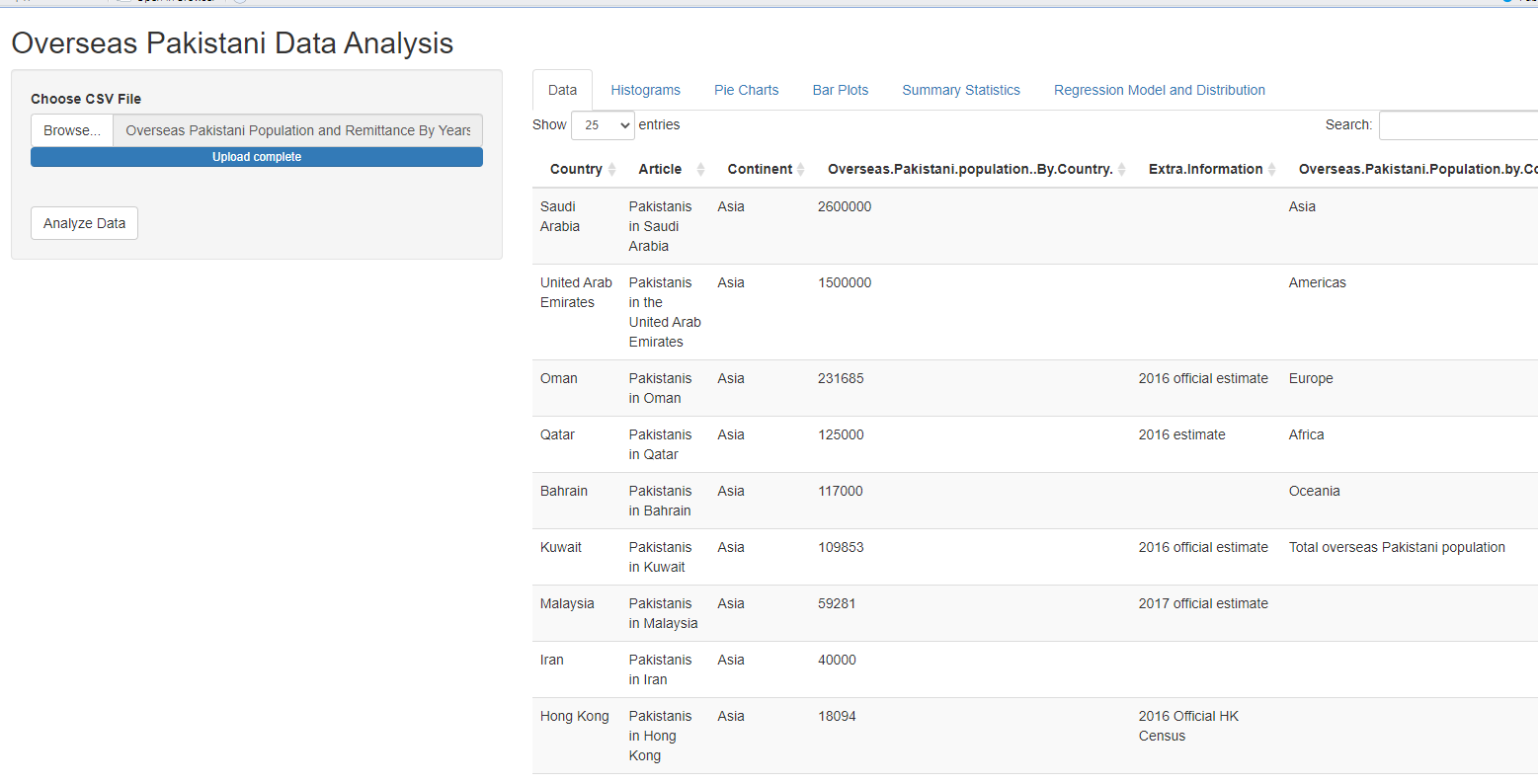
title = "Distribution of Remittance")

})

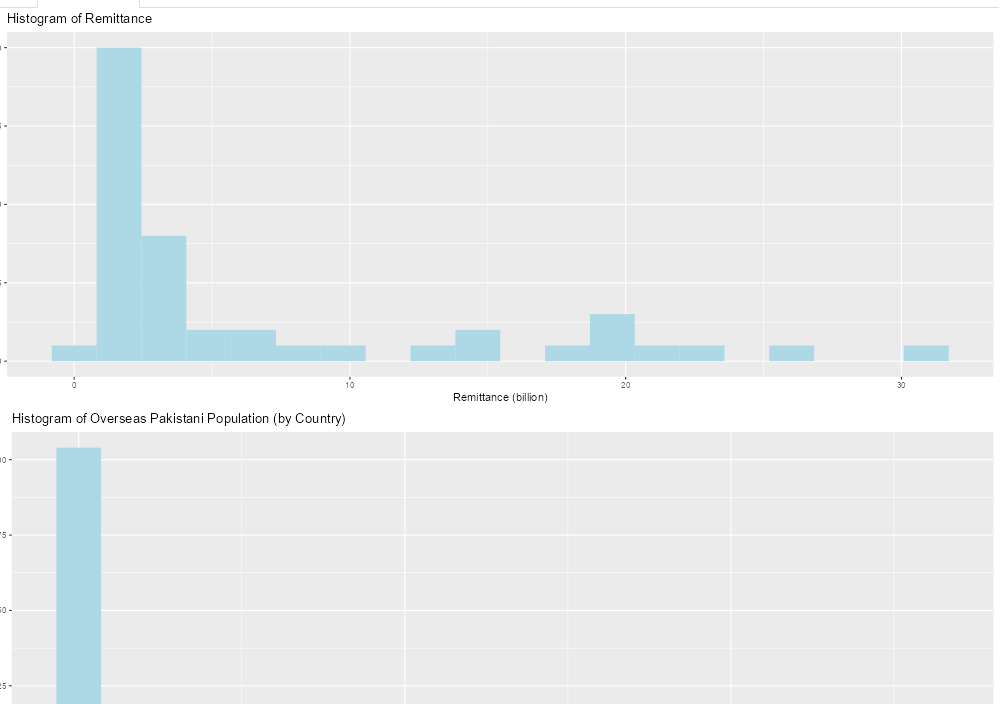
}

shinyApp(ui, server)

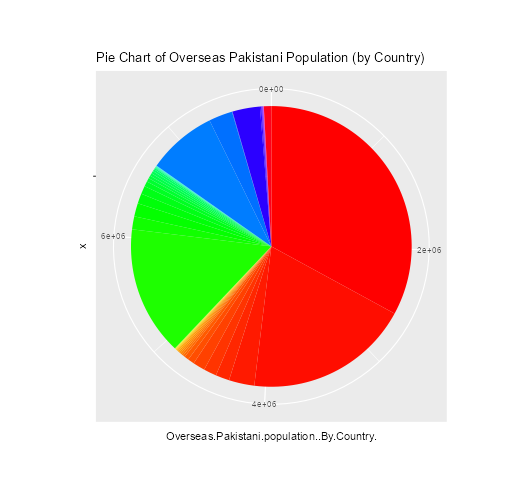
**Outputs:**



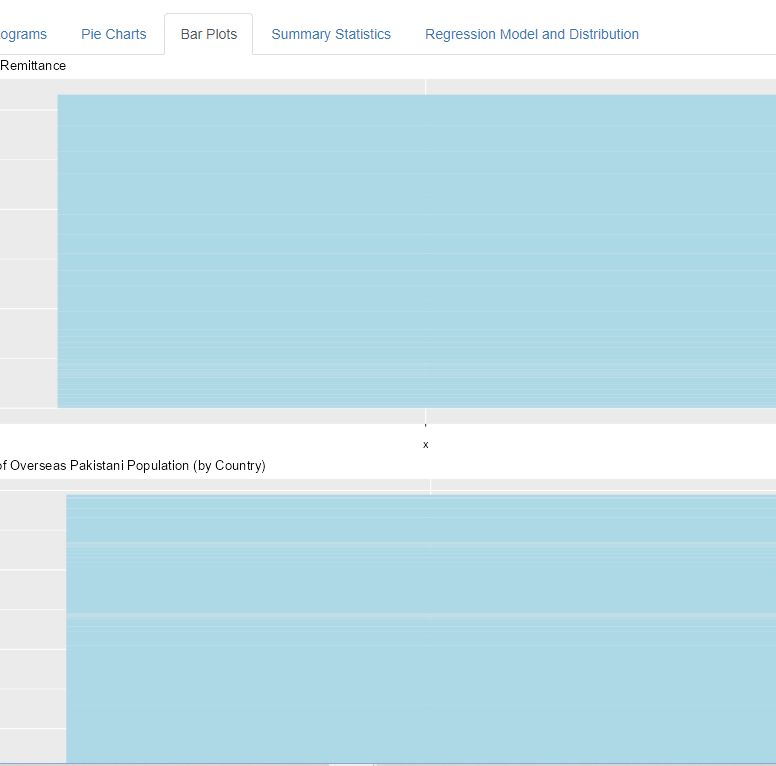
**Histograms:**



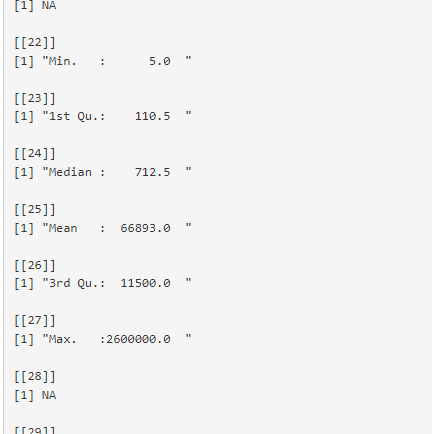
**Pie Chart:**



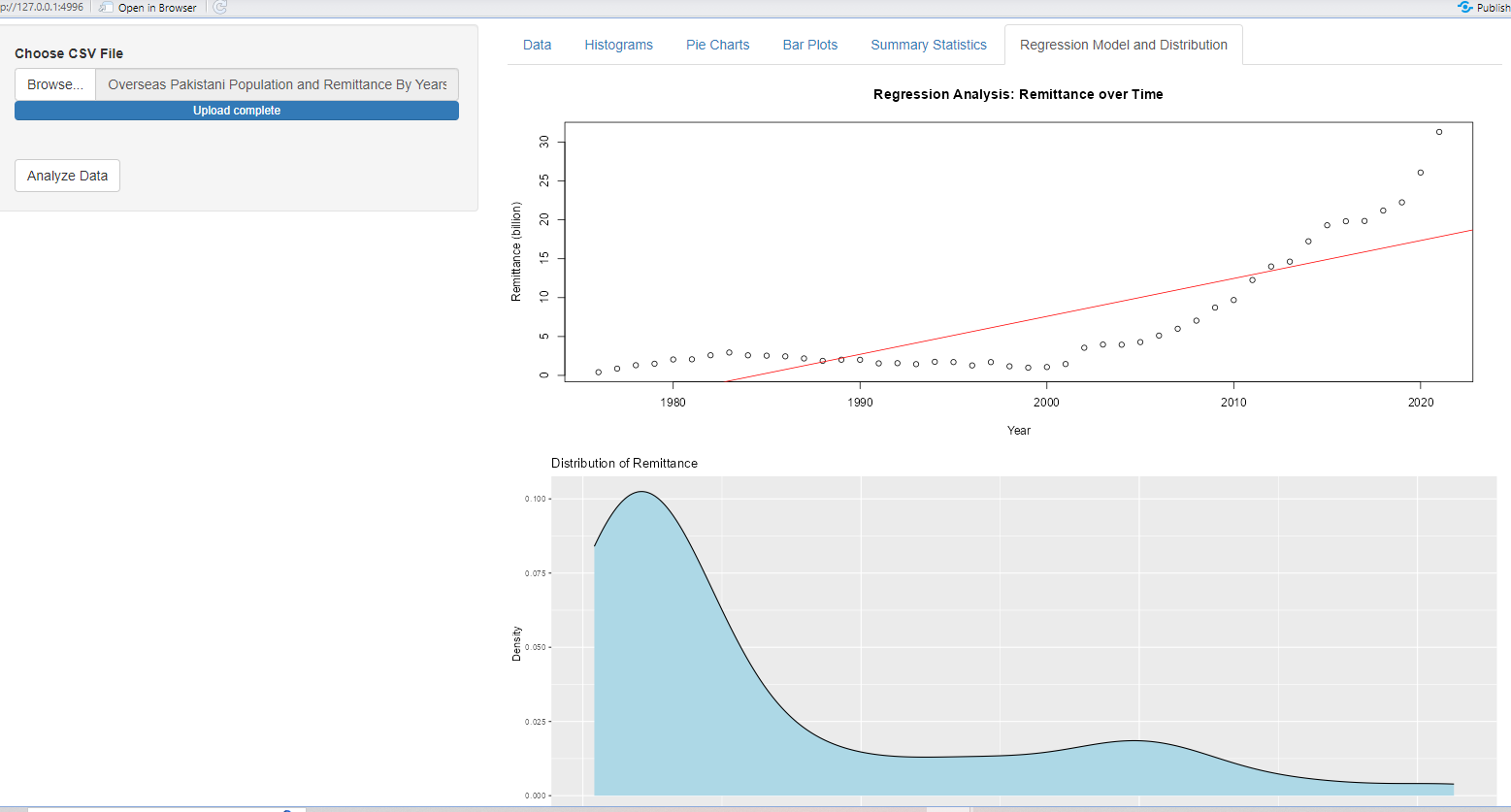
**Bar Plots:**



**Summary Statistics:**



**Regression Models and Distribution:**



**Conclusion:**

In conclusion, this project analyzed data on the overseas Pakistani population and their remittance contributions. The dataset revealed insights into the distribution of overseas Pakistanis by country and continent, as well as the range of remittance amounts. These findings highlight the significant role of overseas Pakistanis in supporting the country's economy. The analysis provides valuable information for policymakers, researchers, and stakeholders, aiding in the identification of regions with higher concentrations of overseas Pakistanis and informing policy decisions related to economic development and welfare programs. Further data cleaning and imputation may be needed due to missing values in some variables. Overall, this project enhances understanding of the demographic and economic aspects of overseas Pakistanis and their impact on the homeland. In conclusion, this project analyzed data on the overseas Pakistani population and their remittance contributions. The dataset revealed insights into the distribution of overseas Pakistanis by country and continent, as well as the range of remittance amounts. These findings highlight the significant role of overseas Pakistanis in supporting the country's economy. The analysis provides valuable information for policymakers, researchers, and stakeholders, aiding in the identification of regions with higher concentrations of overseas Pakistanis and informing policy decisions related to economic development and welfare programs. Further data cleaning and imputation may be needed due to missing values in some variables. Overall, this project enhances understanding of the demographic and economic aspects of overseas Pakistanis and their impact on the homeland.