Exploring the World of Data: A Journey with R,

Shainyravishika

5 min read · May 30, 2024

Hey There! In this article, I'll walk you through my experience exploring and making predictions on an interesting dataset using R. By the end, you'll have a deeper understanding of how data analysis and predictions.

Let's begin!!

Tourism is a multifaceted industry that significantly impact economies,cultures,and environments worldwide. In this article, we delve into the complex dynamics of tourism using “**Tourism Has Dataset**” . By analyzing global tourism data serves multiple purposes, including understanding travel trends, identifying popular destinations, evaluating economic impacts, and informing policy decisions. Additionally, analyzing this data can aid in sustainable tourism development.

Exploring Data

First of all, I identified the sources from which I could gather relevant dataset for my analysis. Then I collected the dataset from **“https://www.kaggle.com/datasets”** and imported dataset as a **.csv** file formats into R.

Import Dataset

```r
data=read.csv("C:/Users/shain/Downloads/international-tourist new.csv")
print(data)
```

country <chr>	Year <int>	International.tourist.arrivals.by.region <dbl>
Africa	1995	12832774
Americas	1995	101567080
East Asia and the Pacific	1995	114378800
Europe	1995	299340380
Middle East	1995	10119565
Not classified	1995	10150115
South Asia	1995	4779760
Africa	1996	14155691
Americas	1996	109251144
East Asia and the Pacific	1996	125587384

1-10 of 189 rows | 1-3 of 4 columns Previous 1 2 3 4 5 6 ... 19 Next

Once loaded, I used functions like `head()` to take a look at the first few rows of the dataset and get an idea of its structure.

Display Data Rows

```r
head(data,5)
```

	country <chr>	Year <int>	International.tourist.arrivals.by.region <dbl>
1	Africa	1995	12832774
2	Americas	1995	101567080
3	East Asia and the Pacific	1995	114378800
4	Europe	1995	299340380
5	Middle East	1995	10119565

5 rows | 1-4 of 4 columns

After that to understand the structure of the dataset, I used functions like `summary()`

Summary of Dataset

```{r}
summary(data)
```

```
   country              Year      International.tourist.arrivals.by.region
 Length:189         Min.   :1995   Min.   :  4779760
 Class :character   1st Qu.:2001   1st Qu.: 19199696
 Mode  :character   Median :2008   Median : 37785040
                    Mean   :2008   Mean   :153332124
                    3rd Qu.:2015   3rd Qu.:203407280
                    Max.   :2021   Max.   :878253500
  percentage
 Length:189
 Class :character
 Mode  :character
```

Then clean the dataset remove any errors or missing values that could be affect the analysis

Identify Missing Values

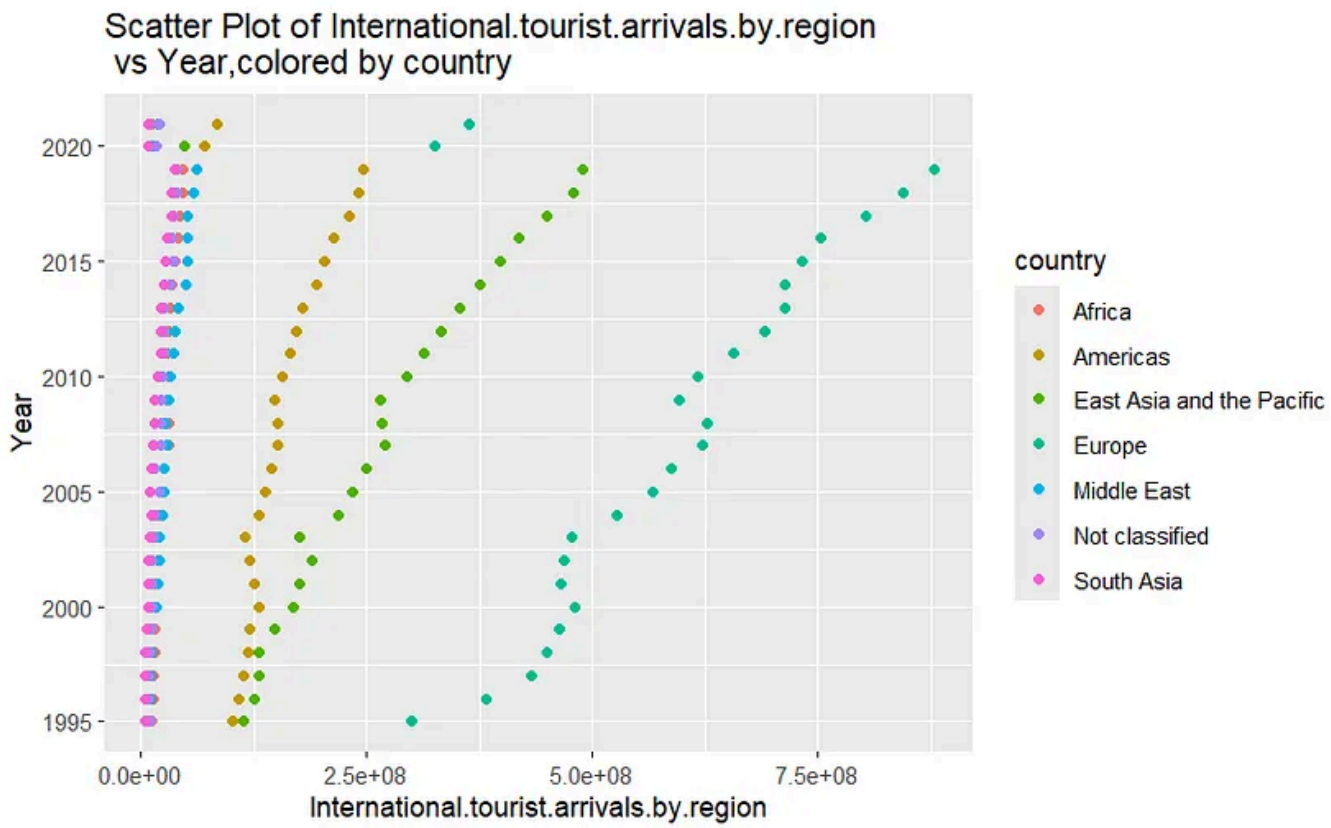
```{r}
is.na(data)
```

```
      country  Year International.tourist.arrivals.by.region percentage
 [1,]   FALSE FALSE                                    FALSE      FALSE
 [2,]   FALSE FALSE                                    FALSE      FALSE
 [3,]   FALSE FALSE                                    FALSE      FALSE
 [4,]   FALSE FALSE                                    FALSE      FALSE
 [5,]   FALSE FALSE                                    FALSE      FALSE
 [6,]   FALSE FALSE                                    FALSE      FALSE
 [7,]   FALSE FALSE                                    FALSE      FALSE
 [8,]   FALSE FALSE                                    FALSE      FALSE
 [9,]   FALSE FALSE                                    FALSE      FALSE
[10,]   FALSE FALSE                                    FALSE      FALSE
[11,]   FALSE FALSE                                    FALSE      FALSE
[12,]   FALSE FALSE                                    FALSE      FALSE
[13,]   FALSE FALSE                                    FALSE      FALSE
[14,]   FALSE FALSE                                    FALSE      FALSE
[15,]   FALSE FALSE                                    FALSE      FALSE
[16,]   FALSE FALSE                                    FALSE      FALSE
[17,]   FALSE FALSE                                    FALSE      FALSE
[18,]   FALSE FALSE                                    FALSE      FALSE
[19,]   FALSE FALSE                                    FALSE      FALSE
[20,]   FALSE FALSE                                    FALSE      FALSE
[21,]   FALSE FALSE                                    FALSE      FALSE
[22,]   FALSE FALSE                                    FALSE      FALSE
[23,]   FALSE FALSE                                    FALSE      FALSE
[24,]   FALSE FALSE                                    FALSE      FALSE
[25,]   FALSE FALSE                                    FALSE      FALSE
[26,]   FALSE FALSE                                    FALSE      FALSE
[27,]   FALSE FALSE                                    FALSE      FALSE
[28,]   FALSE FALSE                                    FALSE      FALSE
[29,]   FALSE FALSE                                    FALSE      FALSE
[30,]   FALSE FALSE                                    FALSE      FALSE
[31,]   FALSE FALSE                                    FALSE      FALSE
```

To further understand the dataset, I created visualizations using packages like `ggplot2` one of the most popular packages for data visualization in R:

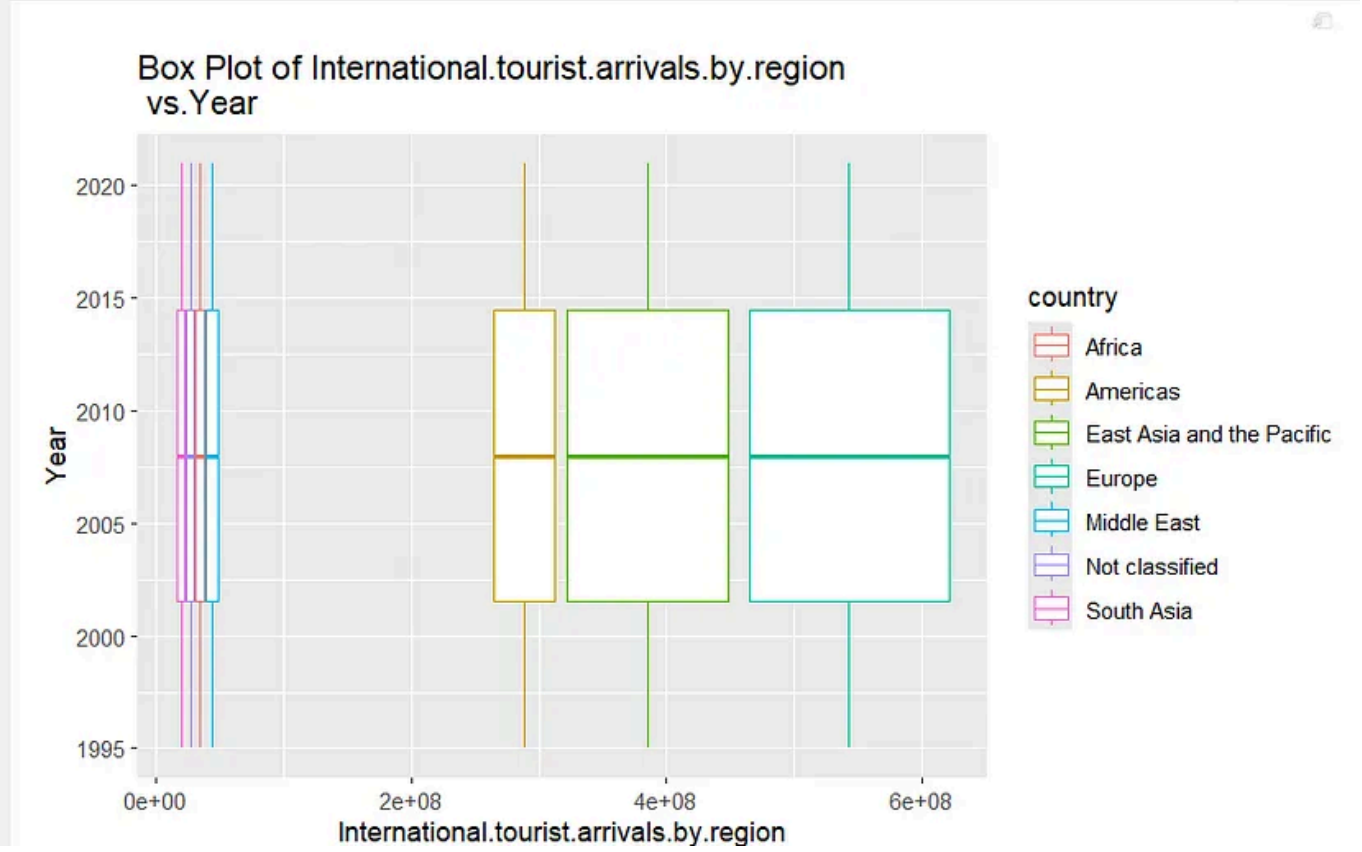
Data Visualization

```{r}
library(ggplot2)
ggplot(data,aes(x=International.tourist.arrivals.by.region
,y=Year,color=country))+
  geom_point()+
ggtitle("Scatter Plot of International.tourist.arrivals.by.region
 vs Year,colored by country")
```



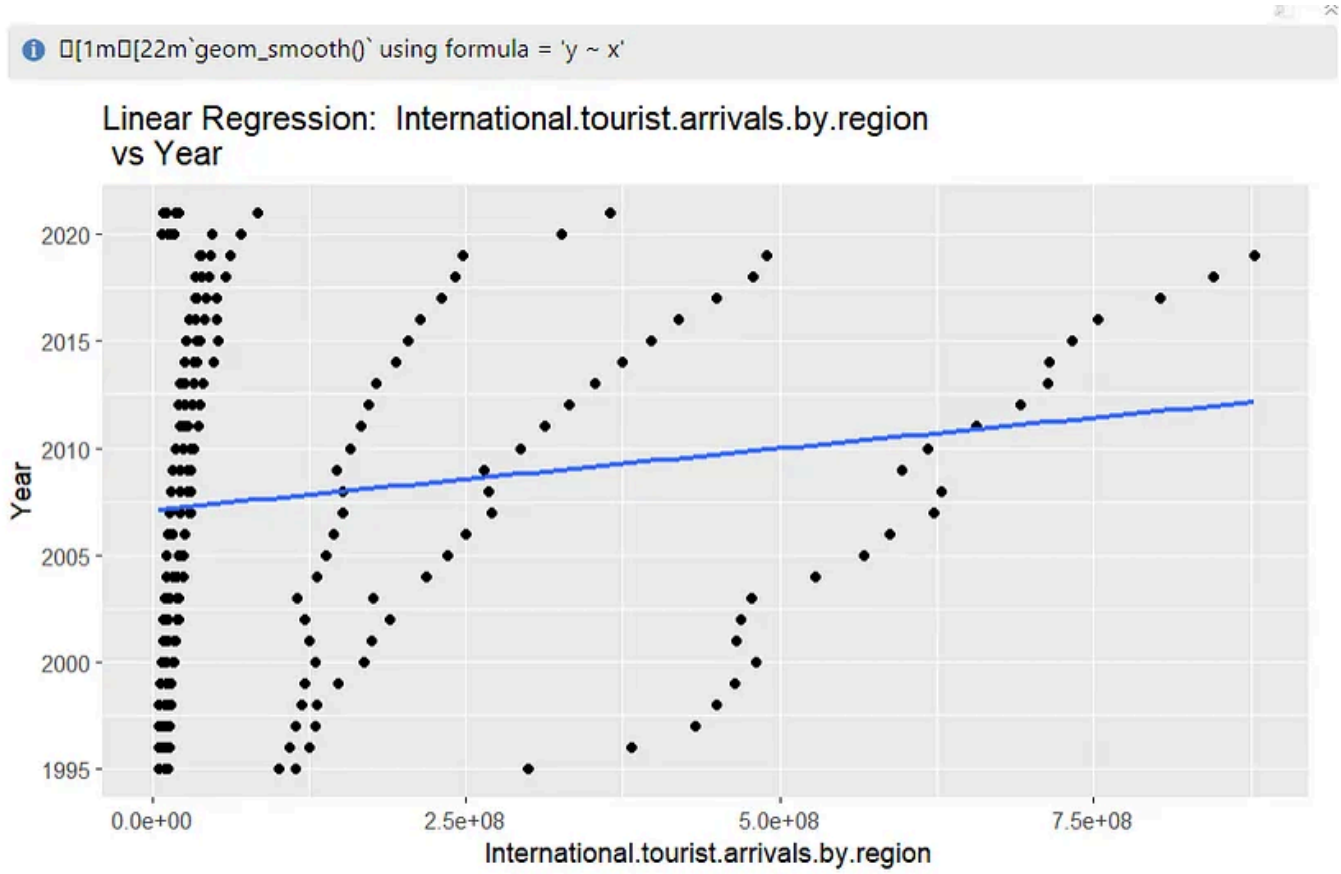
Box Plot

```{r}
ggplot(data,aes(x=International.tourist.arrivals.by.region
,y=Year,color=country))+
  geom_boxplot() +
  ggtitle("Box Plot of International.tourist.arrivals.by.region
 vs.Year")
```



Liner Regression

```{r}
ggplot(data, aes(x = International.tourist.arrivals.by.region
, y = Year)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = " International.tourist.arrivals.by.region
", y = "Year") +
  ggtitle("Linear Regression:  International.tourist.arrivals.by.region
 vs Year ")
```



Interesting things

While exploring the datasets, I discovered ,the number of tourists visiting the international region varied significantly throughout the year. For example, there was a notable increase and decrease in tourist arrivals in Europe counties and South Asia countries respectively.....

Working with R Shiny

R Shiny is important for several reasons, particularly in the realm of data analysis, visualization, and communication. Let's see how I used R Shiny web application to make interactive graphs and charts

```r
#
# This is a Shiny web application. You can run the application by clicking
# the 'Run App' button above.
#
# Find out more about building applications with Shiny here:
#
#    https://shiny.posit.co/
#

library(shiny)
library(datasets)

# Define UI for application that draws a histogram
ui <- fluidPage(

    # Application title
    titlePanel("Toursim Data Explore"),

    # Sidebar with a slider input for number of bins
    sidebarLayout(
        sidebarPanel(
            selectInput("Country",
                        "select country:",
                        choices = unique(data$country)),
             actionButton("plot_hist","Plot Histogram"),
             actionButton("plot_scatter","Plot Scatterplot")
        ),

        # Show a plot of the country
        mainPanel(
           plotOutput("Plot")
        )
    )
)
```

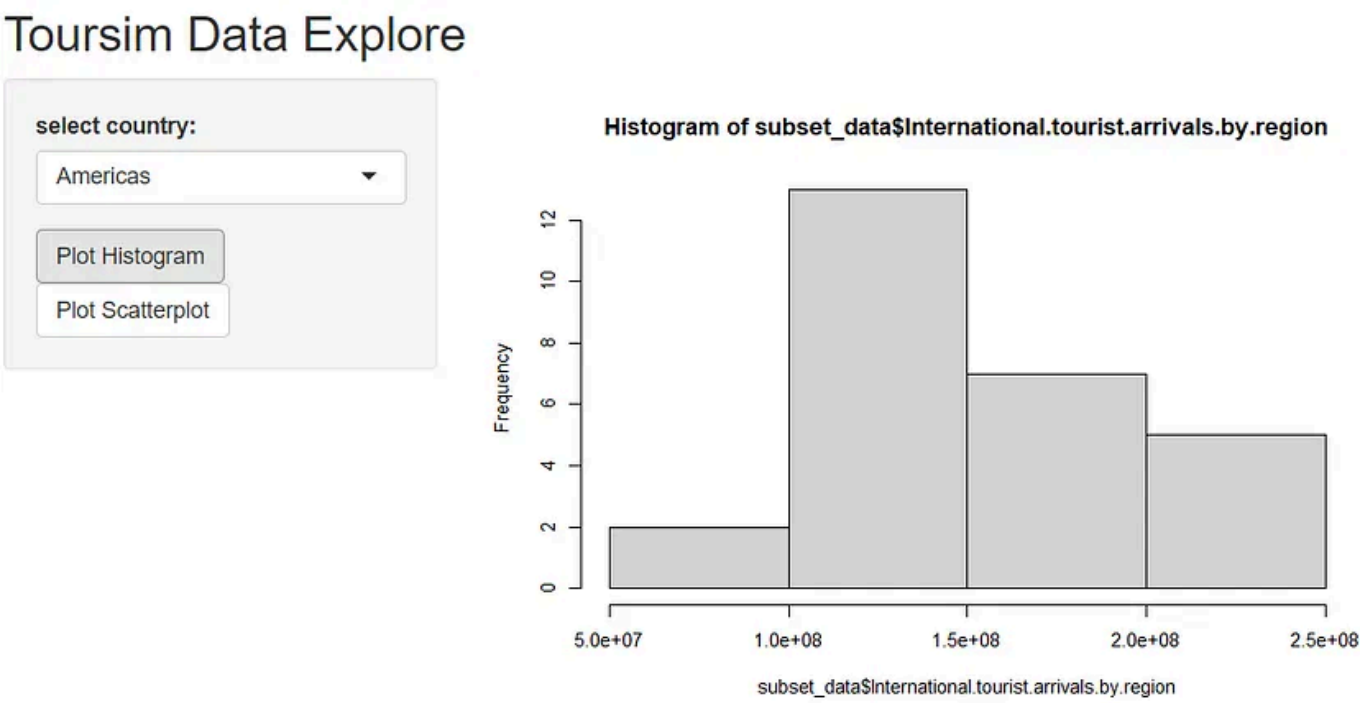
```
# Define server logic required to draw a histogram
server <- function(input, output) {
  observeEvent(input$plot_hist,{
    subset_data<-subset(data,country==input$Country)
    output$Plot <- renderPlot({
        hist(subset_data$International.tourist.arrivals.by.region)
    })
  })


    output$plot<-renderPlot({
    # generate bins based on input$bins from ui.R
        x    <- faithful[, 2]
        bins <- seq(min(x), max(x), length.out = input$bins + 1)

        # draw the histogram with the specified number of bins
        hist(x, breaks = bins, col = 'darkgray', border = 'white',
             xlab = 'Waiting time to next eruption (in mins)',
             main = 'Histogram of waiting times')
    })
}

# Run the application
shinyApp(ui = ui, server = server)
```

By following these steps, I created an interactive web application using R Shiny to visualize data with interactive graphs and charts. Like this,



Important of R Shiny

In an article featuring a Shiny app, readers can engage with various interactive elements to explore data, visualize trends, and gain insights. Here are some of the things readers can typically do with a Shiny app embedded in an article:

- **Data Exploration**: Readers can interact with the app to explore datasets by adjusting filters, selecting variables of interest, and drilling down into specific data subsets.
- **Data Visualization:** Readers can create interactive plots, charts, graphs, and maps using libraries like ggplot2
- **Statistical Analysis**: Perform statistical analysis, hypothesis testing, regression, clustering, and other analytical tasks within the app.

Seeing Patterns

Identifying Patterns

Identifying patterns in a tourism dataset can reveal valuable insights about traveler behavior, preferences, and trends. Pattern can vary greatly depending on the type of data and the context in which it's analyzed. During exploring the tourism dataset,

- I identified patterns in destination preferences among travelers. Certain regions or countries may consistently attract higher numbers of tourists, while others experience fluctuations based on factors like political stability, economic conditions, or natural disasters.
- Overall, international tourist arrivals increased by region during the period 1995 to 2021.

Importance of pattern

Identifying patterns in tourism datasets provides invaluable insights into traveler behavior, preferences, and trends. By leveraging these insights, tourism stakeholders can make informed decisions, optimize resource allocation, and drive sustainable growth in the tourism industry and also its importance to the country's economy .

For instance:

- Identifying areas of environmental impact and potential overcrowding can inform sustainable tourism practices.
- Recognizing seasonal fluctuations can help optimize marketing campaigns and infrastructure investments to increase tourism.

- By understanding which attractions and activities are driving visitor demand, can increase tourism attraction

Making Predictions

In my exploration of the tourism dataset, I used R to make predictions by using **Regression Models** to predictions such as tourist arrivals rates. Here's a step-by-step how I used R to make predictions using regression model

```
Regression Model
```{r}
model=lm(International.tourist.arrivals.by.region~Year,data=data)
summary(model)
```
```

> *As shown below ,I found these data after the predictions*

- The differences between the actual measured values and the corresponding values on the fitted regression line. (Min,1Q,Median,3Q,Max)
- The statistical standard errors for each of the coefficients.
- Significance of the coefficients of the model
- The reported R2 of 0.2324 means that the model explains 23.24% of the data's variation(used to test how good is your model)
- significance of the model to check whether there is a relationship between International tourist arrivals by region and Year.(F-statistic: 4.448 on 1 and 187 DF, p-value: 0.03626)

```
Call:
lm(formula = International.tourist.arrivals.by.region ~ Year,
    data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-196786745 -132934403 -104491301   47384219  680445992

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.965e+09  3.849e+09  -2.069   0.0399 *
Year         4.043e+06  1.917e+06   2.109   0.0363 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 205300000 on 187 degrees of freedom
Multiple R-squared:  0.02324,   Adjusted R-squared:  0.01801
F-statistic: 4.448 on 1 and 187 DF,  p-value: 0.03626
```

> *How I combined my predictions with my graphs?*

- By combining predictions with graphs such as “Plot Histogram, Scatter Plot, Box Plot " we can identify about patterns that so far how tourism going on
- By combining prediction with regression model we can identify numerical data such as Min,1Q,Median,3Q,Max

Learning from Analysis

I learned that by doing predictive analysis, data can be deeply uncovered and data analysis can be made more interesting and attractively.

Conclusion,

In this article, I explored the Tourism Dataset, Create an interactive Shiny app, Identify patterns and Use regression model to make predictions. Analyzing a tourism dataset can reveal insightful patterns and trends, offering valuable information for various stakeholders like government bodies, tourism agencies, and businesses. This article was not only educational project but also demonstrated the journey of R .I encourage you to try this dataset

As we embark on a journey towards a more resilient and equitable tourism future, let data be our guiding star!!

Written by Shainyravishika

0 Followers

Edit profile

Recommended from Medium

Alexander Nguyen in Level Up Coding

The resume that got a software engineer a $300,000 job at Google.

1-page. Well-formatted.

Jun 1 · 12.1K · 162

Nigel Stanley

How in 2024 Labour won a big majority on a similar share of the...

Some argue that Labour could have won this time with 2019 leaders and policies. They are...

3d ago · 1.4K · 38

Lists

Staff Picks
685 stories · 1129 saves

Stories to Help You Level-Up at Work
19 stories · 687 saves

Self-Improvement 101
20 stories · 2287 saves

Productivity 101
20 stories · 2017 saves

John Gorman

Matthew Gazzano in Towards Data Science

Stop Wasting Your Time

A Simple Framework for Making Better Decisions

Jun 4 16.4K 308

AI-Proof Your Data Science Skill Set by Embracing Four Timeless...

And stay competitive in a saturated job market

2d ago 46 1

Jan Kammerath

Why Tech Workers Are Fleeing Germany—A Reality Check

Over the past months and years I have seen a number of friends and colleagues leave...

Jun 23 5K 176

Liu Zuo Lin

You're Decent At Python If You Can Answer These 7 Questions...

No cheating pls!!

Mar 6 6K 29

See more recommendations