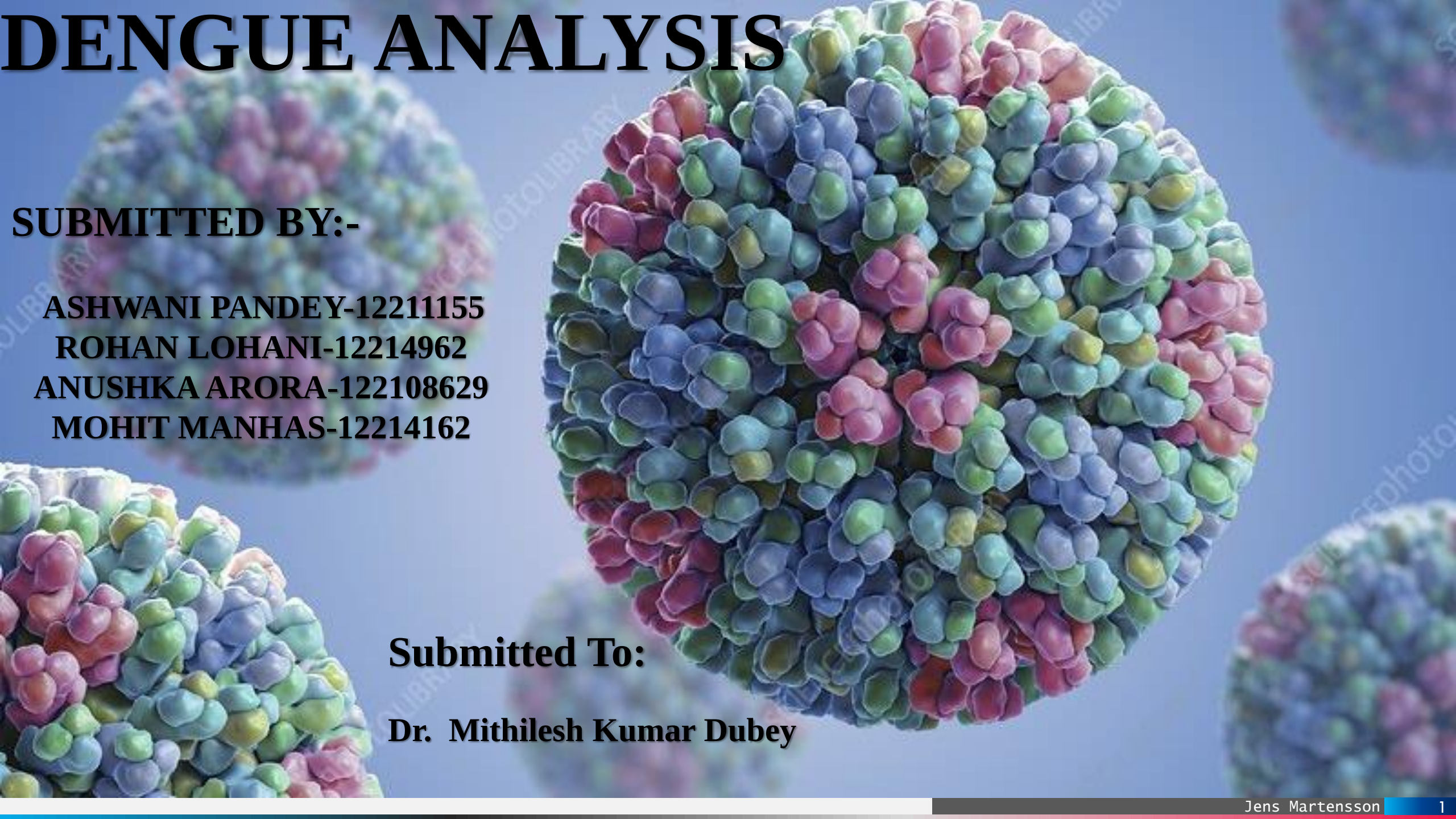
DENGUE ANALYSIS



SUBMITTED BY:-

ASHWANI PANDEY-12211155
ROHAN LOHANI-12214962
ANUSHKA ARORA-122108629
MOHIT MANHAS-12214162

Submitted To:

Dr. Mithilesh Kumar Dubey

Why this project ?

- Need for such survey
- Public health Impact
- Data availability
- Interdisciplinary approaches

INTRODUCTION

- Dengue fever, a mosquito-borne viral infection
- The World Health Organization estimates that around 390 million dengue infections occur globally each year, of which approximately 96 million manifest clinically.
- This study investigates dengue prevalence in LPU hostel and village near LPU. We analyze a dataset from google form containing numeric and character data points with features like [name, age, suffered or not, duration, RBC/WBC level, hemoglobin level, platelets count and PCV count]. We employ machine learning algorithms including Logistic Regression to classify dengue cases. We evaluate the models using metrics like accuracy, precision, recall, and F1-score.

SOFTWARE REQUIREMENT

1. R Studio :

R Studio is an integrated development environment (IDE) for R, a programming language and environment for statistical computing and graphics. It provides a user-friendly interface for writing, executing, and debugging R code, along with features like data visualization tools and package management, making it popular among statisticians and data scientists.

2. Excel:

We use excel to organize, analyze, and visualize data using rows and columns. It offers functions for arithmetic, statistical, and financial calculations, as well as graphing tools for creating charts and graphs. Excel is widely used in various industries for data management, analysis, and reporting tasks.

3. Google form:

We use the google form for data collection. It enables to gather relevant information from users regarding their accommodation preferences, contributing to the dataset essential for training and validating your prediction model on housing choices.

DATA COLLECTION PROCESS

1. Our beautiful Form
2. https://forms.office.com/r/G3UT5FGR5X
3. We faced less challenges as compared to other collectors. Because we heavily invested our time in form creation, though can’t ignore the financial ones

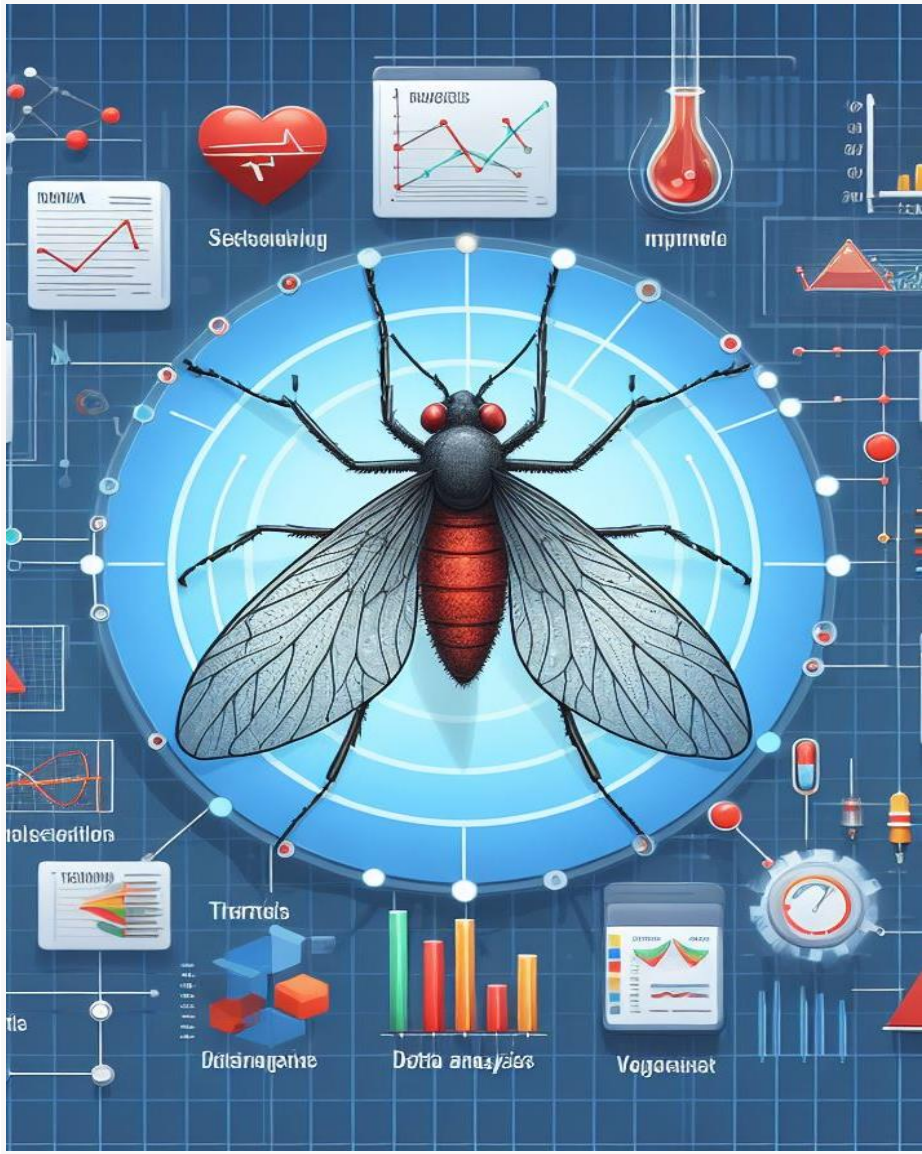
DATA ANALYSIS

Data Cleaning and Preprocessing-

- Identifying missing data points and choosing appropriate strategies for handling with missing value
- Review the data for inconsistencies or outliers and if error found then correct it
- Ensure consistency in units across different variables and converts units if necessary to maintain uniformity.

Data Transformation and Feature Engineering-

- Derive new features from existing data to capture additional information like calculating ratios, percentages, or combining existing features.
- Convert categorical data into numerical formats suitable for machine learning algorithms.



Data pre-processing

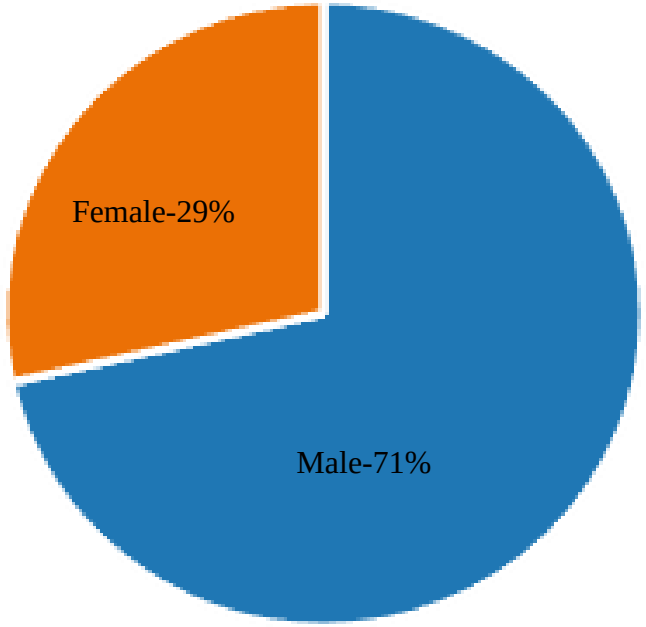
- We deleted unwanted columns from the datasets.
- We filled all the missing data in the We filled in all the missing values in the dataset.
- We did some changes in the datasets as required.

Gender Participation

Gender participation in data collection

Male-238

Female-95

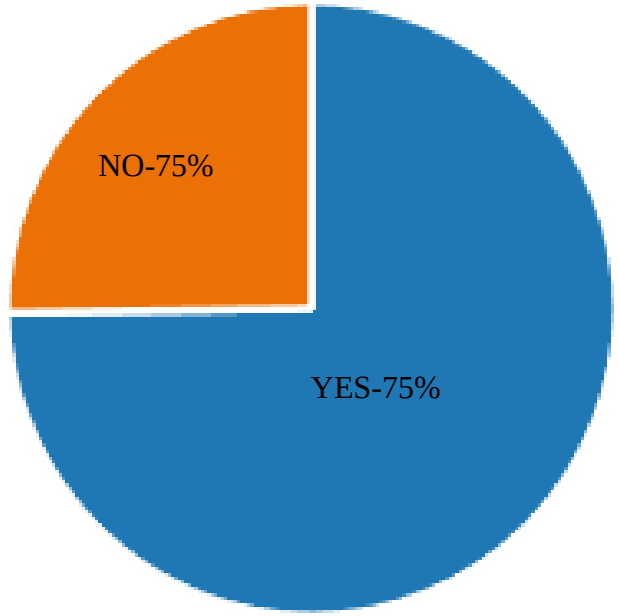


SUFFERING FROM DENGUE

How many people who were suffered from dengue ?

Conclusion:

75% male and 25% female suffered from dengue.



DURATION OF FEVER

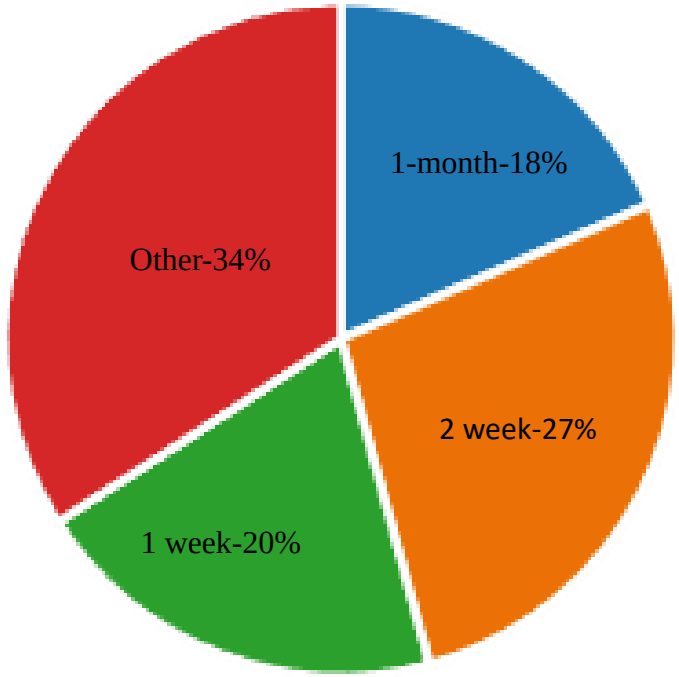
Graph for the duration of fever of suffered people means how long they suffered from dengue-

1 month-46

2 weeks-68

1 weeks-50

Other-85



DESCRIPTIVE STATISTICS

We find:

- Mean
- Median
- Mode
- Standard Deviation

```
> library(readxl)
> Denguefever <- read_excel("F:/PROJECTS/Data Science Projects/Denguefever.xlsx")
> View(Denguefever)
> mean(Denguefever$Age)
[1] 22.58841
> mean(Denguefever$`Duration of fever`)
[1] 11.92683
> mean(Denguefever$`Haemoglobin Level`)
[1] 8.820335
> mean(Denguefever$`PCV count`)
[1] 26.90143
> mean(Denguefever$`Platelets Count`)
[1] 94054.78
> mean(Denguefever$`Leukocytes Count`)
[1] 6.036463
> mean(Denguefever$`RBC Count`)
[1] 3.640976
> median(Denguefever$Age)
[1] 20
> median(Denguefever$`Duration of fever`)
[1] 12
> median(Denguefever$`Haemoglobin Level`)
[1] 10.775
> median(Denguefever$`PCV count`)
[1] 32.4
> median(Denguefever$`Platelets Count`)
[1] 70000
> median(Denguefever$`Leukocytes Count`)
[1] 6
> median(Denguefever$`RBC Count`)
[1] 4.7
> sd(Denguefever$Age)
[1] 8.133338
> sd(Denguefever$`Duration of fever`)
[1] 9.927296
> sd(Denguefever$`Haemoglobin Level`)
[1] 5.589481
> sd(Denguefever$`PCV count`)
[1] 16.59854
> sd(Denguefever$`Platelets Count`)
[1] 220305.5
```

DESCRIPTIVE STATISTICS

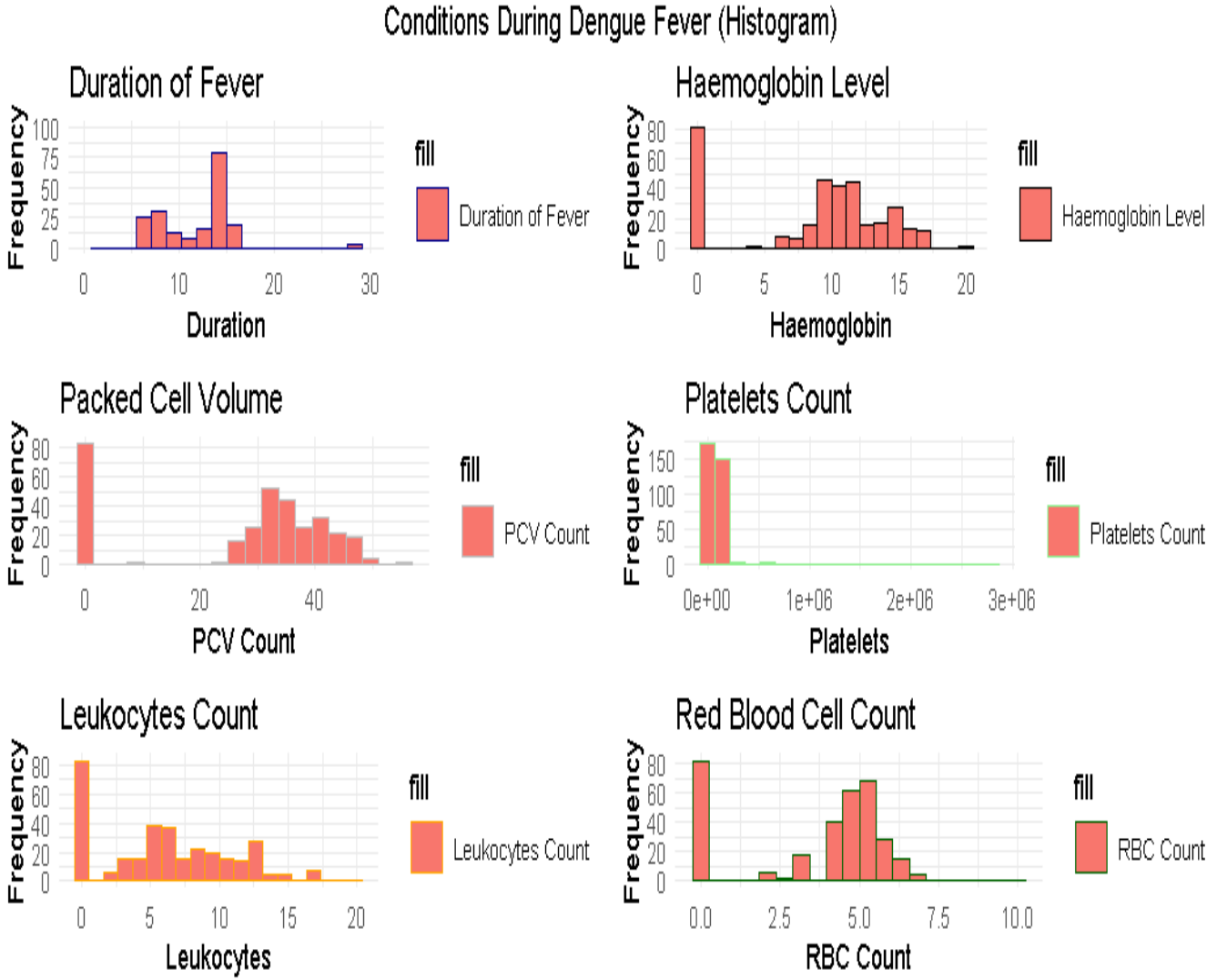
We find:

- Co-variance
- Cor-relation

```
> sd(Denguefever$`Leukocytes Count`)
[1] 4.799808
> sd(Denguefever$`RBC Count`)
[1] 2.27507
> library(readxl)
> Denguefever <- read_excel("F:/PROJECTS/Data Science Projects/Denguefever.xlsx")
> View(Denguefever)
> cov(Denguefever$Age,Denguefever$`Duration of fever`)
[1] -1.862012
> cov(Denguefever$Age,Denguefever$`Haemoglobin Level`)
[1] -0.3175985
> cov(Denguefever$Age,Denguefever$`PCV count`)
[1] -4.027757
> cov(Denguefever$Age,Denguefever$`Platelets Count`)
[1] -56990.99
> cov(Denguefever$Age,Denguefever$`Leukocytes Count`)
[1] 0.7171638
> cov(Denguefever$Age,Denguefever$`RBC Count`)
[1] -0.3891385
> cor(Denguefever$Age,Denguefever$`Duration of fever`)
[1] -0.02306125
> cor(Denguefever$Age,Denguefever$`Haemoglobin Level`)
[1] -0.006986155
> cor(Denguefever$Age,Denguefever$`PCV count`)
[1] -0.02983491
> cor(Denguefever$Age,Denguefever$`Platelets Count`)
[1] -0.03180623
> cor(Denguefever$Age,Denguefever$`Leukocytes Count`)
[1] 0.0183707
> cor(Denguefever$Age,Denguefever$`RBC Count`)
[1] -0.02103006
```

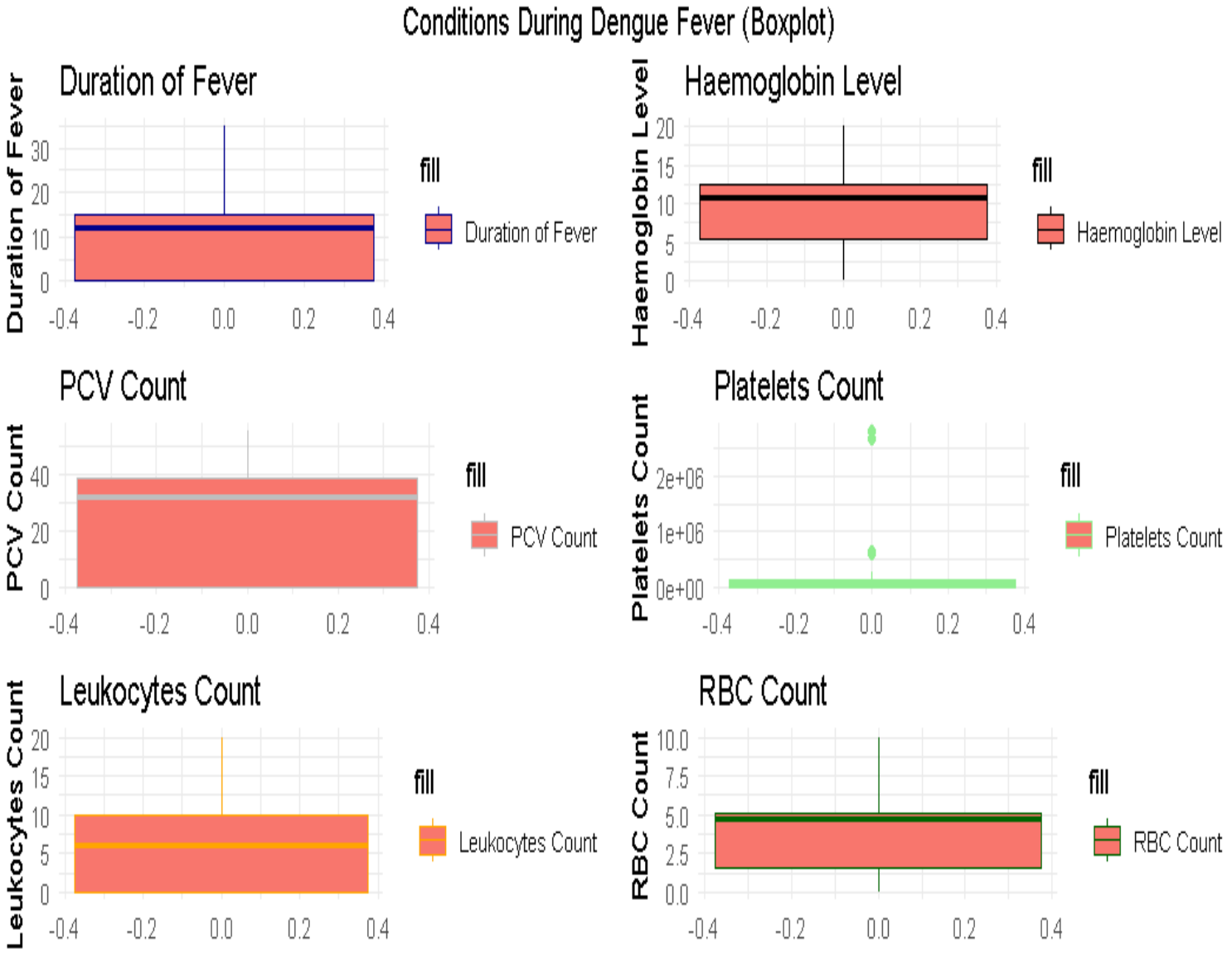
VISUALIZATION IN HISTOGRAMS

A histogram is a graphical representation of the distribution of numerical data. It consists of a series of adjacent rectangles, or bins, where the width of each bin represents a range of values, and the height of each bin represents the frequency or count of data points within that range.



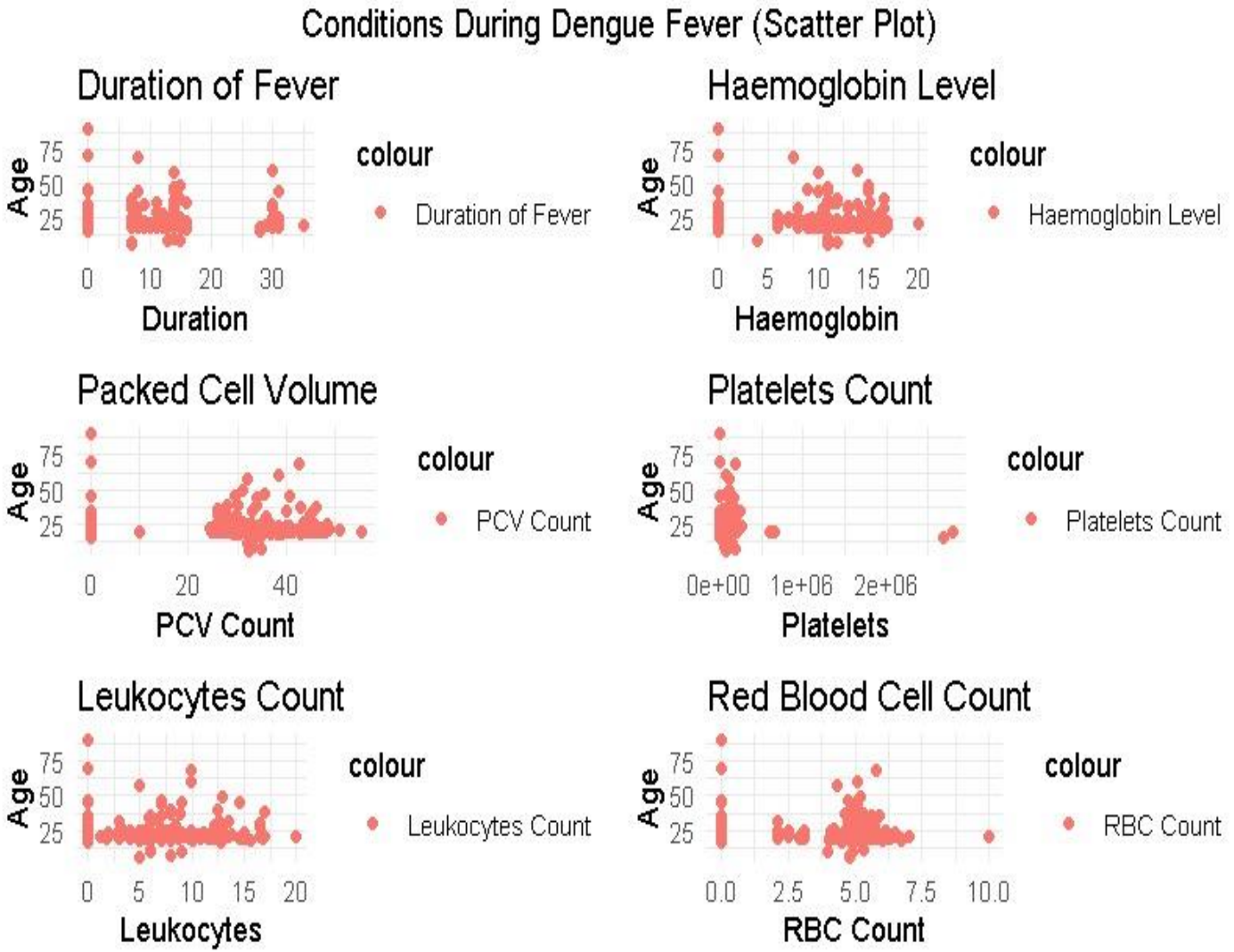
VISUALIZATION IN BOXPLOT

In R, a boxplot is a graphical representation of the distribution of numerical data through their quartiles, as described earlier. The boxplot() function in R is used to create boxplots, and it accepts one or more numerical vectors as input.



VISUALIZATION IN SCATTERPLOT

A scatterplot is a type of graphical representation used to display the relationship between two continuous variables. It consists of a plot where each data point is represented by a dot or marker, positioned according to its values on the two variables being compared.



CONCLUSION

The conclusion for the presentation would be a summary of the key findings and insights from data collection and analysis process which involves visualization of gender participation, number of people suffered, age of people suffered from dengue and the distribution of disease in LPU Hostel and nearby village. It would also highlight the challenges faced during data collection, the use of Google forms and excel sheet for data collection and analysis and importance of interdisciplinary approaches in public health impact studies. At last it emphasizes the significance of the study in identifying high-risk populations, informing, targeted prevention and control strategies, and the overall public health impact of the research.

Thank You

Ashwani Pandey

ashwanipandey4545@gmail.com