DEGGENDORF INSTITUTE OF TECHNOLOGY

Analysis of Malaria Disease

Final assignment for the course “Data Visualization” - SS23

Submitted by:

Tushar Rewatkar (12202513)

Siddhi Gunaji (12201642)

tushar.rewatkar@stud.th-deg.de

siddhi.gunaji@stud.th-deg.de

Under the supervision of

Prof. Dr Javier Valdes

Contents

List of Figures

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

USA : United States of America

CA : California

MD : Maryland

EDA : Exploratory Data Analysis

TX : Texas

FL : Florida

Chapter 1

Overview

In the United States, the most common vector-borne pathogens are transmitted by ticks or mosquitoes, including those causing Malaria. In this report, we will see how the cases of Malaria diseases has been increased in USA over the period from 2012-2017. Also we will see how environmental Factors and Rise in Population has affected the rising number of cases.

Keywords: Malaria, Vector-borne diseases, Rainfall, Palmer Drought Index, Temperature, Population

1.1 Introduction

A female Anopheles mosquito that has contracted malaria will bite you. These mosquitoes like warm, humid temperatures, which can alter as a result of environmental changes. For instance, variations in temperature and rainfall can change the habitats and breeding habits of mosquitoes, affecting how malaria is transmitted.

We go into the issue at hand in the following section, looking at the Cases of malaria in the US from 2012 to 2017 in connection to environmental conditions and population density.

1.2 Problem Definition

In this study, We took cases of malaria in the US between 2012 and 2017 is studied in relation to environmental variables such as rainfall, temperature, and population. We analyze these variables in order to find trends which can help us to prevent this disease in future.

1.3Objective

The main objective of this project is to perform Data Visualization of Malaria Cases from the ”Project Tycho” data-sets and observe how Environmental factors and population has affect this disease. This EDA will help researchers to compare and see if any factors caused the increase in number of vector-borne diseases in USA.

Chapter 2

Methodology

Various EDA has been performed on the data-sets by plotting graphs and charts in this assignment shown as follows:

2.1 Cases of Malaria in USA(2000- 2017)

Let's begin with visualizing the Number of cases each year from year 2000-2017 of Malaria in US.

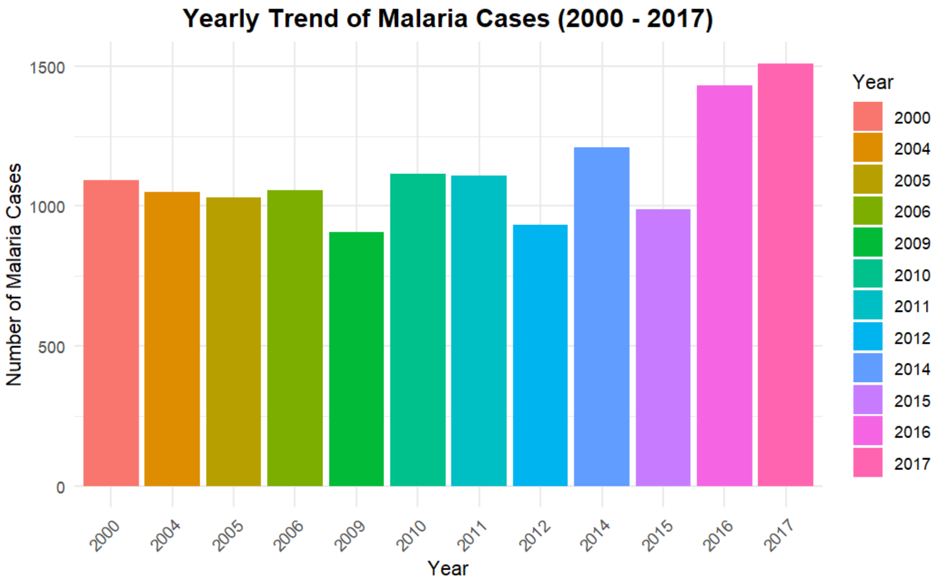
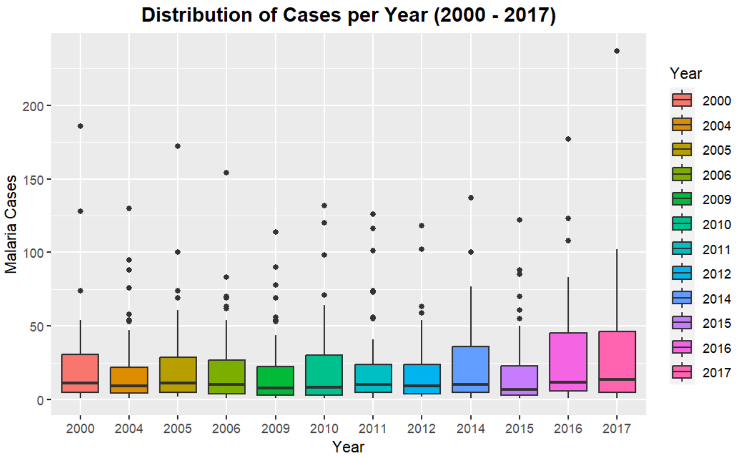
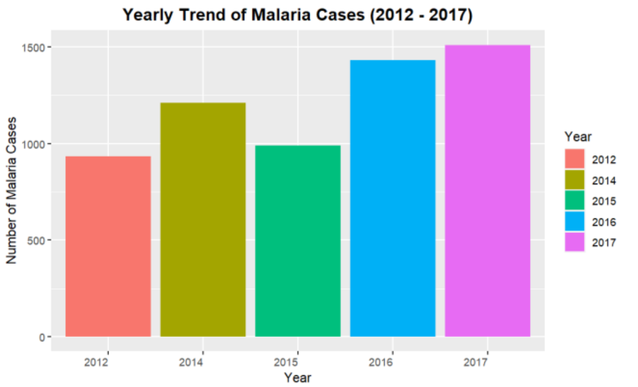


Fig 2.1 Yearly Trend of Malaria Cases (2000 - 17)

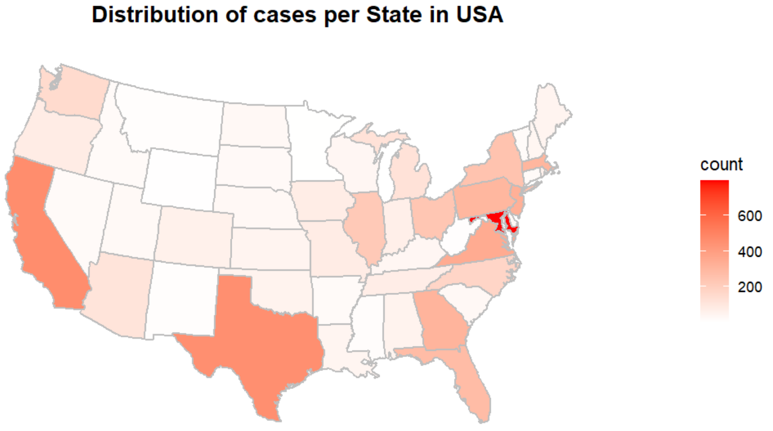
You can see from the trend from 2000 to 2012 there is not much variation in number of cases. But if you Observe closely from 2012 to 2017 you can see the increase in the number of cases by each year.



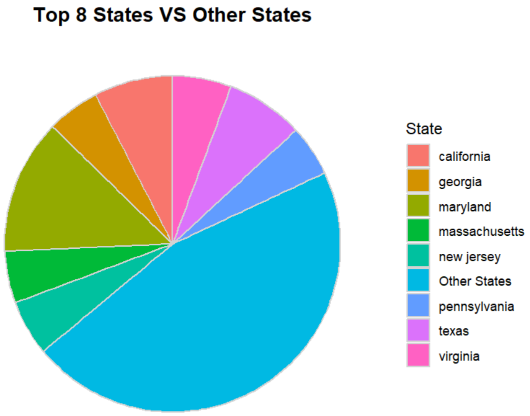
In this graph you can see the how the Cases are distributed from 2000-2017. You can see the from above plot that year 2017 & 2016 , where number cases are more distributed.



As we discussed, we will take the data from 2012-2017 as you can see steep increases in number of cases year by year. Here Point to be noted, we have omitted 2013 data as it only had 3 cases. So that year was a outlier for us.

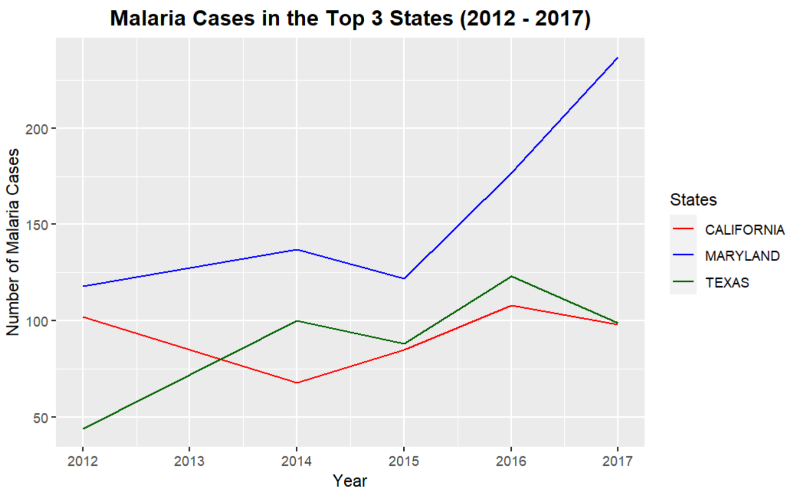


Above map plot shows the number of cases in each state of USA. Maryland , California and Texas are having the top 3 states who have most number of cases. **You can observe that most of the cases occurred in coastal region of USA. Its an interesting fact to be noted.**



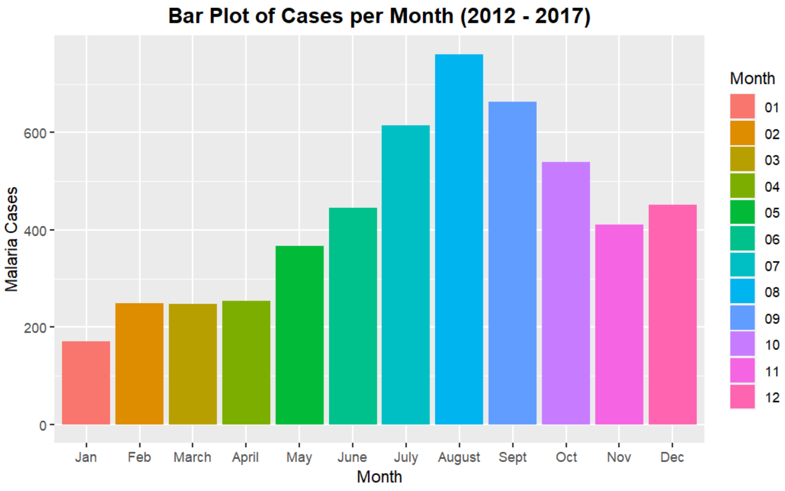
Above Plot shows Distribution of Top 8 States vs Other States.

You can see out of 52 states of USA, Distribution of Cases of top 8 states which are costal states of USA is more than the rest of US states combined.

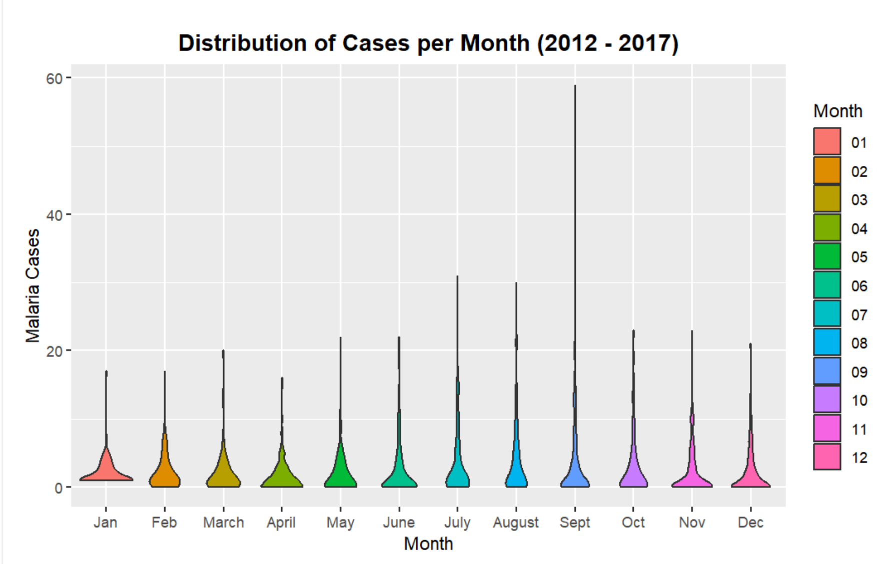


The above line graph shows the Number of cases in California, Maryland and Texas from 2012-2017.You can clearly see Maryland is having the highest number of cases from 2012 to 2017.

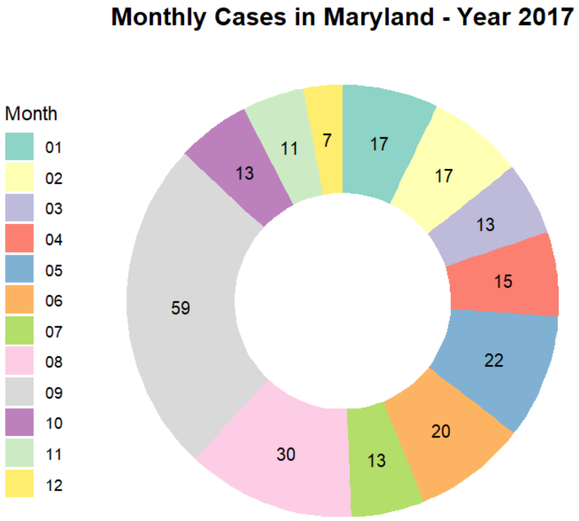
2.2 Visualizing Monthly Malaria disease(2012-2017)



The plot shows the Cases in each month from year 2012-2017. From this graph we can clearly observe that Cases happened in this due course is clearly seasonal as in July, August and September having the highest number of cases. **So clearly some Seasonal Environment changes are affecting the disease.**



Above plot shows Monthly Distribution of range of the cases from 2012-2017. July, August and September having the maximum spread. This clearly shows that data have seasonality.

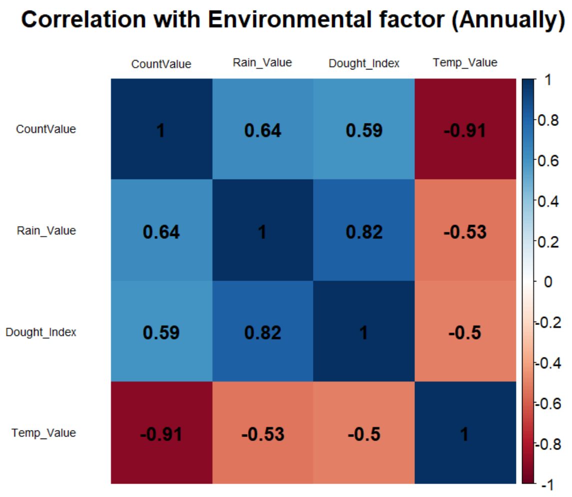


So, we took one a state who has highest number of cases which is Maryland in a year which has highest number of cases i.e. 2017. Above is doughnut plot for Maryland in year 2017.

2.3 Impact of Environmental Factors

So from above oberservation, we conclude that we number of cases were incresease due to seaonlaity and in the area of coastal regions.

We will now take the environmental factors into account i.e Rainfall, Palmer Dought Index and Temprature value of California State. We took the data from the National centres for Environmental Information.



We already have case count values of California State Annually i.e (2012-17) and we have Envirmomental factors value from above website.
The above plot shows the correaltion between Cases count with Rain values, Palmer Drought Index and Temperature value.

The Palmer Drought Severity Index (PDSI)- It is a standardized index that generally spans -10 (dry) to +10 (wet)

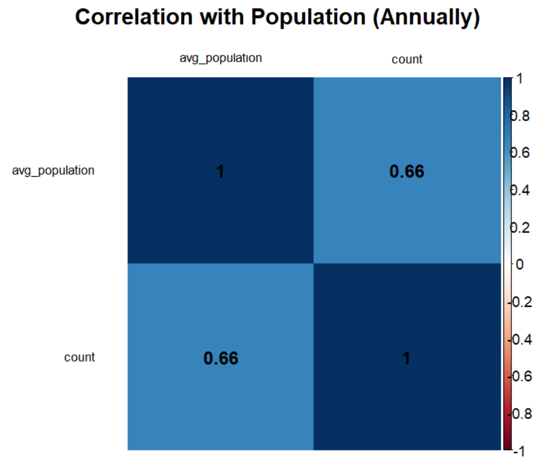
From above plot we can conclude that the count value have highly correlated with annual rainfall of California state and also with dought Index value with the p-value of 0.59.

From this we can conclude that a state who have a having more than average rainfall annually, also having the considerable positive Dought factor value are responsible for the high number of cases in a state from year (2012-2017).

Interesently one point to be noted here, Temperature value is negatively correlated with Count values. And drought value is positivily correlated. Therefore, it is essential to thoroughly examine the temperature data.We will discuss this in last part of the report (Monthly analysis).

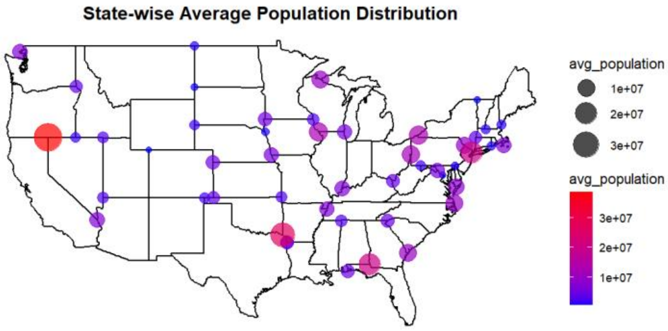
2.4 Impact of Population (Annually)

Lets see how population can explain the number of Malaria-cases diseases in the states of USA.

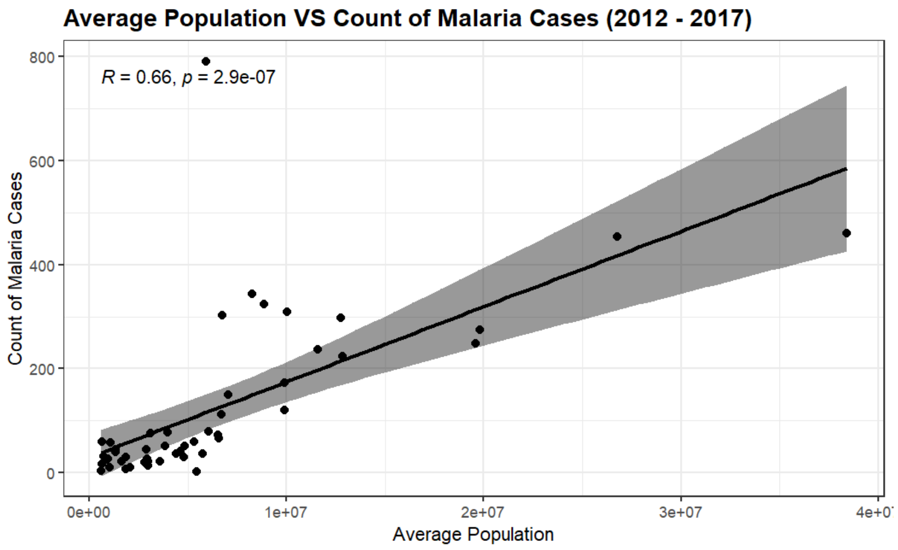


We took the Average Annual population of each state and Malaria Cases count of each state from year 2012-17, and the results were quite Interesting.

From above heatmap we can clearly check the correlation between the count values and Average population is positive with p-value of 0.66



From this graph it clearly shows a state having high population density is more vulnerable to Malaria.



This is Scatter plot with Malaria cases vs Average population, this clearly show how linearly Malaria count value is dependent on Population.

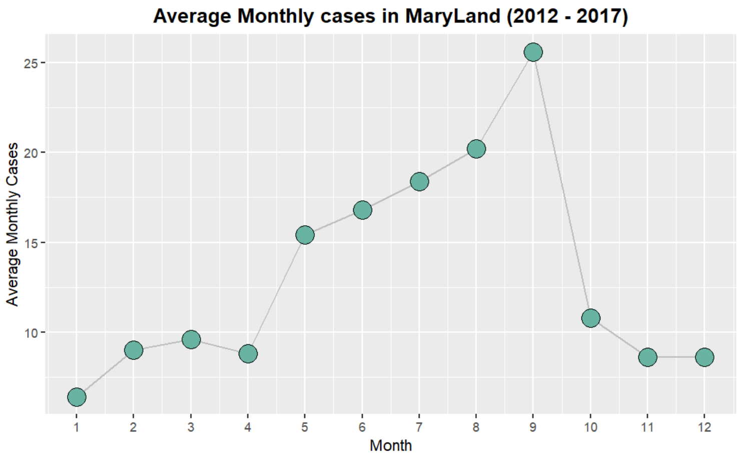
So, from above observations we can conclude that, a state with high population value will increase the impact of number of Malaria cases.

2.5 Impact of Temperature (Monthly)

In Section 2.3 we have discuss the Environmental factors Annually, but we haven’t discuss the Temperature effect Monthly.

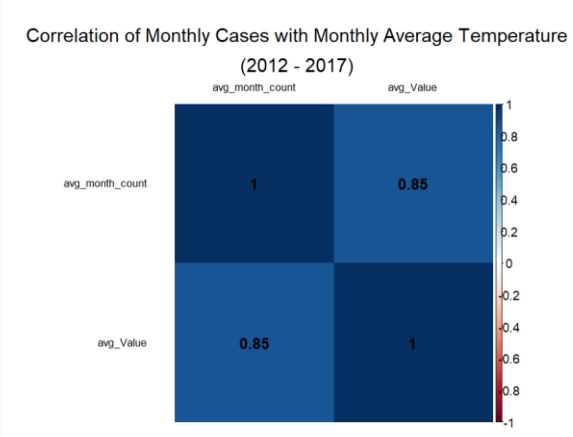
Now, Lets consider a state which having the highest number of Cases i.e Maryland from year (2012-17). We took the data monthly that means we combined all the data of January from 2012-17, same as follows.

We then took the monthly average temperature from year 2012-17 which we took it from National centres for Environmental Information.



We plotted the graph between the months and Average monthly cases. As you can see in Maryland state, we can observe seasonality from June to October.

Because of this seasonality we plotted the heatmap between the Average monthly cases vs Average monthly Temperature of State Maryland from year to year.



As you can see from above graph the Average monthly temperature have positive correlation with monthly cases. This concludes that temperature is leading to increase in number of cases from month June to September. This shows that data have seasonality in it and temperature is affecting it

Chapter 3

RESULTS

From year 2012 to 2017 we can find the increase in number of cases year by year all over USA. In 2017 we observed highest number of Malaria cases.

It can be observed that Maryland, California and Texas has been reported highest number of Malaria disease cases and also we have seen how environmental factors are affecting it as well as well Population of state.

We also observed seasonality in a year which were proven with the heatmap of temperature with the number of cases.

Chapter 4

DISCUSSION AND CONCLUSION

- Our Analysis has revealed how Environmental factors are affecting the malaria cases in USA from 2012-17.
- We discussed that the number of cases are high in US coastal regions.
- A positive correlation between the rainfall and Palmer drought index factor show how both of these factors leading to increase in number of cases.
- Population density is another reason for increase in the number of malaria cases
- We also discussed the impact of seasonality, this was proven with positively correlation of Average monthly temperature with average monthly malaria count from year 2012-17.

From all these discussion, we can clearly conclude that the increase in number of malaria cases is highly dependent on Environmental factors such as annual rainfall, drought factor and Temperature(Seasonal) with affect of high population density. All these condition leads to humid conditions in a populated area, which can help malaria mosquito to bread faster and help affecting the spread of the disease.

Appendix

Source code for the plots

2.5.1 Bar plot

```
# BAR PLOT: Total malaria cases per year (2000 to 2017)


ggplot(monthly_sum_fltr, aes(x = as.character(start_year), y = CountValue, fill =
as.character(start_year))) +
  geom_bar(stat = "identity") +
  labs(x = "Year", y = "Number of Malaria Cases") +
  ggtitle("Yearly Trend of Malaria Cases (2000 - 2017)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

2.5.2 Box Plot

```
# BOX PLOT: Malaria cases per Year (2000 - 2017)


box_plot <- ggplot(malaria_yearly_cases_old, aes(x = start_year, y = CountValue, fill =
start_year)) +
  geom_boxplot() + labs(x = "Year", y = "Malaria Cases") +
  ggtitle("Box Plot: Cases per Year (2000 - 2017)") + theme_minimal()

print(box_plot)
```

2.5.3 Voilin Plot

```
# VIOLIN PLOT: Malaria cases month wise (2012 - 2017)


violin_plot <- ggplot(fltr_mnth, aes(x = end_month_name, y = month_count, fill = 
end_month)) +
  geom_violin() +
  labs(x = "Month", y = "Malaria Cases") +
  ggtitle("Violin Plot: Cases per Month (2012 - 2017)") +
  theme_minimal()

# Display the violin plot
print(violin_plot)
```

2.5.4 Donut plot

```
donut_plot <- ggplot(mary_2017, aes(ymax = y_max, ymin = y_min, xmax = 4, xmin = 3, 
fill = end_month)) +
  geom_rect() +
  geom_text(x = 3.5, aes(y = lbl_pos, label = lbl), size = 3.6) +
  scale_fill_brewer(palette = "Set3", name="Months") +
  coord_polar(theta = "y") +
  xlim(c(2, 4)) +
  theme_void() +
  theme(legend.position = "left", plot.title = element_text(hjust = 0.5, size = 14, face = 
"bold")) +
  ggtitle("Monthly Cases in Maryland - Year 2017")  # Add a title using ggtitle()
```

2.5.5 Scatter plot

```
ggscatter(state_population_with_countValue, x = "avg_population", y = "count",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson", title = "Scatter Plot with Fitted Line",
          xlab = "Average Population", ylab = "Count of Malaria Cases") +
          theme_bw() + theme(plot.title = element_text(size = 14, face = "bold"))
```

2.5.6 LINE PLOT WITH POINTS

```
ggplot(avg_data, aes(x=end_month, y=avg_month_count)) +
  geom_line(color="grey") +
  geom_point(shape=21, color="black", fill="#69b3a2", size=6) +
  scale_x_continuous(breaks = seq(min(avg_data$end_month), max(avg_data$end_month), 1), labels = as.character(seq(min(avg_data$end_month), max(avg_data$end_month), 1))) +
  ggtitle("Evolution of count per month of MaryLand from 2012-2017")
```

2.5.7 Map Plot

```
ggplot(malaria_map, aes(x = long, y = lat, group = group, fill = count)) + geom_polygon(color = 'gray') + scale_fill_gradient2(low = 'white', high='red')+ theme_void()+ ggtitle('Malaria cases in USA') + coord_map('polyconic')
```

Bibliography

[1] Project Tycho unlocks global health data to a rapidly growing user community of over 3,000 researchers, students, journalists, officials, and others in over 90 countries.

[2] USA State Population Totals: 2010-2019.

[3] USA Climate dataset – National Centers for Environmental Information

[4] Wikipedia