data_visualization_hackathon

Sohrab Khan

2024-07-26

Hackathon: Data Visualization in R

Load and Install Repositories and Packages

```
        library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
        library(ggplot2)
        library(readxl)
        library(dplyr)
```

Create a data folder manually or use R code mentioned below

```
        dir.create("hackathon_data_visualization")
```

```
## Warning in dir.create("hackathon_data_visualization"):
## 'hackathon_data_visualization' already exists
```

Download example data

```
         url <- "https://raw.githubusercontent.com/AMMnet/AMMnet-Hackathon/main/01_data-vis/data/"
         download.file(paste0(url,"mockdata_cases.csv"),
                       destfile = "hackathon_data_visualization/mockdata_cases.csv")
         download.file("https://raw.githubusercontent.com/AMMnet/AMMnet-Hackathon/main/01_data-vis/data/mosq_mock.csv", destfile = "hackathon_data_visualization/mosq_mock.csv")
```

Load example data

```
         library(readr)
         malaria_data <- read_csv("hackathon_data_visualization/mockdata_cases.csv")
```

```
## Rows: 514 Columns: 10
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): location, ages
## dbl (8): month, year, total, positive, xcoord, ycoord, prev, time_order_loc
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
         View(malaria_data)
         mosq_data <- read_csv("hackathon_data_visualization/mosq_mock.csv")
```

```
## Rows: 104 Columns: 19
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (4): Village, Method, Location, hour
## dbl (15): session, Compound.ID, ag.Male, Ag.unfed, Ag.halffed, Ag.fed, Ag.gr...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
         View(mosq_data)
```

Characterize means Exploring our data

How many rows and columns

```
         dim(malaria_data)
```

```
## [1] 514  10
```

General information

```
head(malaria_data)
```

```
## # A tibble: 6 × 10
##   location month  year ages    total positive xcoord ycoord  prev time_order_loc
##   <chr>    <dbl> <dbl> <chr>   <dbl>    <dbl>  <dbl>  <dbl> <dbl>          <dbl>
## 1 mordor       1  2018 15_abo…   122       30  -20.5   30.2 0.246              1
## 2 mordor       2  2018 15_abo…   168       57  -20.1   30.3 0.339              2
## 3 mordor       3  2018 15_abo…    97       20  -20.1   30.4 0.206              3
## 4 mordor       4  2018 15_abo…    91       23  -20.0   30.5 0.253              4
## 5 mordor       5  2018 15_abo…    67       19  -20.7   30.7 0.284              5
## 6 mordor       6  2018 15_abo…   107       25  -19.2   30.5 0.234              6
```

```
summary(malaria_data)
```

```
##    location             month             year          ages          
##  Length:514         Min.   : 1.000   Min.   :2018   Length:514        
##  Class :character   1st Qu.: 4.000   1st Qu.:2018   Class :character  
##  Mode  :character   Median : 7.000   Median :2019   Mode  :character  
##                     Mean   : 6.486   Mean   :2019                     
##                     3rd Qu.: 9.000   3rd Qu.:2020                     
##                     Max.   :12.000   Max.   :2020                     
##      total          positive          xcoord           ycoord     
##  Min.   : 20.0   Min.   : -1.00   Min.   :-21.84   Min.   :28.52  
##  1st Qu.: 46.0   1st Qu.: 14.00   1st Qu.:-20.39   1st Qu.:29.64  
##  Median :103.0   Median : 33.00   Median :-20.06   Median :29.99  
##  Mean   :141.5   Mean   : 47.81   Mean   :-20.04   Mean   :30.00  
##  3rd Qu.:206.0   3rd Qu.: 67.00   3rd Qu.:-19.71   3rd Qu.:30.32  
##  Max.   :611.0   Max.   :264.00   Max.   :-18.79   Max.   :31.81  
##       prev          time_order_loc 
##  Min.   :-0.04545   Min.   : 1.00  
##  1st Qu.: 0.24615   1st Qu.: 9.00  
##  Median : 0.33016   Median :18.00  
##  Mean   : 0.31518   Mean   :17.65  
##  3rd Qu.: 0.39024   3rd Qu.:26.00  
##  Max.   : 0.53488   Max.   :35.00  
```

Exploring individual columns of the data

```
malaria_data$location # values for a single column
```

```
##   [1] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##   [6] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [11] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [16] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [21] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [26] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [31] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [36] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [41] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [46] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [51] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [56] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [61] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [66] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [71] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [76] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [81] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [86] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [91] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
##  [96] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
## [101] "mordor"     "mordor"     "mordor"     "mordor"     "mordor"
## [106] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [111] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [116] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [121] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [126] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [131] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [136] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [141] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [146] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [151] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [156] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [161] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [166] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [171] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [176] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [181] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [186] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [191] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [196] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [201] "narnia"     "narnia"     "narnia"     "narnia"     "narnia"
## [206] "narnia"     "narnia"     "narnia"     "narnia"     "neverwhere"
## [211] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [216] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [221] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [226] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [231] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [236] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [241] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [246] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [251] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [256] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [261] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [266] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [271] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [276] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [281] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [286] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [291] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [296] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [301] "neverwhere" "neverwhere" "neverwhere" "neverwhere" "neverwhere"
## [306] "oz"         "oz"         "oz"         "oz"         "oz"
## [311] "oz"         "oz"         "oz"         "oz"         "oz"
## [316] "oz"         "oz"         "oz"         "oz"         "oz"
## [321] "oz"         "oz"         "oz"         "oz"         "oz"
## [326] "oz"         "oz"         "oz"         "oz"         "oz"
## [331] "oz"         "oz"         "oz"         "oz"         "oz"
## [336] "oz"         "oz"         "oz"         "oz"         "oz"
## [341] "oz"         "oz"         "oz"         "oz"         "oz"
## [346] "oz"         "oz"         "oz"         "oz"         "oz"
## [351] "oz"         "oz"         "oz"         "oz"         "oz"
## [356] "oz"         "oz"         "oz"         "oz"         "oz"
## [361] "oz"         "oz"         "oz"         "oz"         "oz"
## [366] "oz"         "oz"         "oz"         "oz"         "oz"
## [371] "oz"         "oz"         "oz"         "oz"         "oz"
## [376] "oz"         "oz"         "oz"         "oz"         "oz"
## [381] "oz"         "oz"         "oz"         "oz"         "oz"
## [386] "oz"         "oz"         "oz"         "oz"         "oz"
## [391] "oz"         "oz"         "oz"         "oz"         "oz"
## [396] "oz"         "oz"         "oz"         "oz"         "oz"
## [401] "oz"         "oz"         "oz"         "oz"         "oz"
## [406] "oz"         "oz"         "oz"         "oz"         "wonderland"
## [411] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [416] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [421] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [426] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [431] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [436] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [441] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [446] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [451] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [456] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [461] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [466] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [471] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [476] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [481] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [486] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [491] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [496] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [501] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [506] "wonderland" "wonderland" "wonderland" "wonderland" "wonderland"
## [511] "wonderland" "wonderland" "wonderland" "wonderland"
```

```
         unique(malaria_data$location) # unique values for a single column
```

```
## [1] "mordor"     "narnia"     "neverwhere" "oz"         "wonderland"
```

```
         table(malaria_data$location) # frequencies for a single column
```

```
## 
##     mordor     narnia neverwhere         oz wonderland 
##        105        104         96        104        105
```

```
         table(malaria_data$location, malaria_data$ages) # frequencies for multiple column
```

```
## 
##              15_above 5_to_14 under_5
##   mordor           35      35      35
##   narnia           35      35      34
##   neverwhere       32      32      32
##   oz               35      35      34
##   wonderland       35      35      35
```

Check for if there are any missing values- NA

```
         sum(is.na(malaria_data))
```

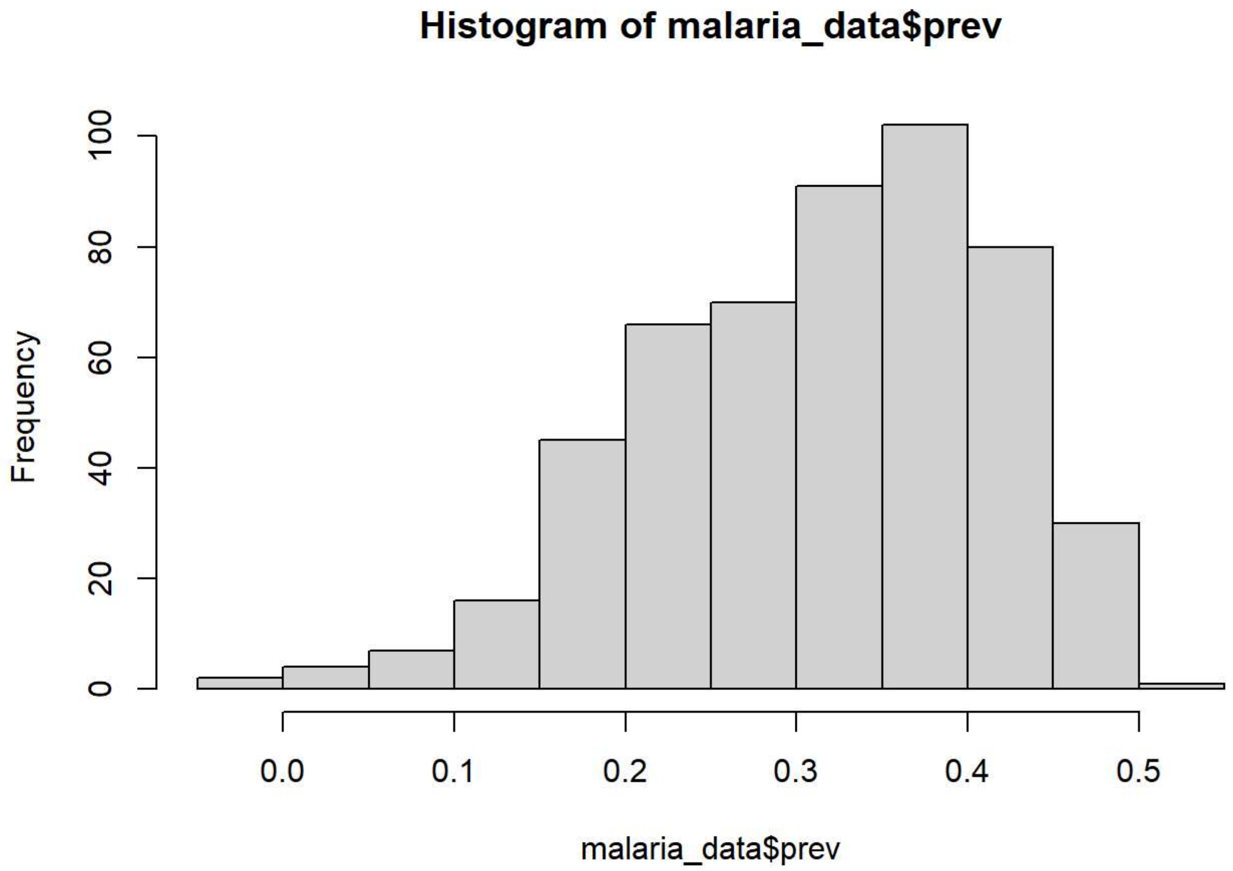
```
## [1] 0
```

Exploratory Visualizations Using Base R Functions

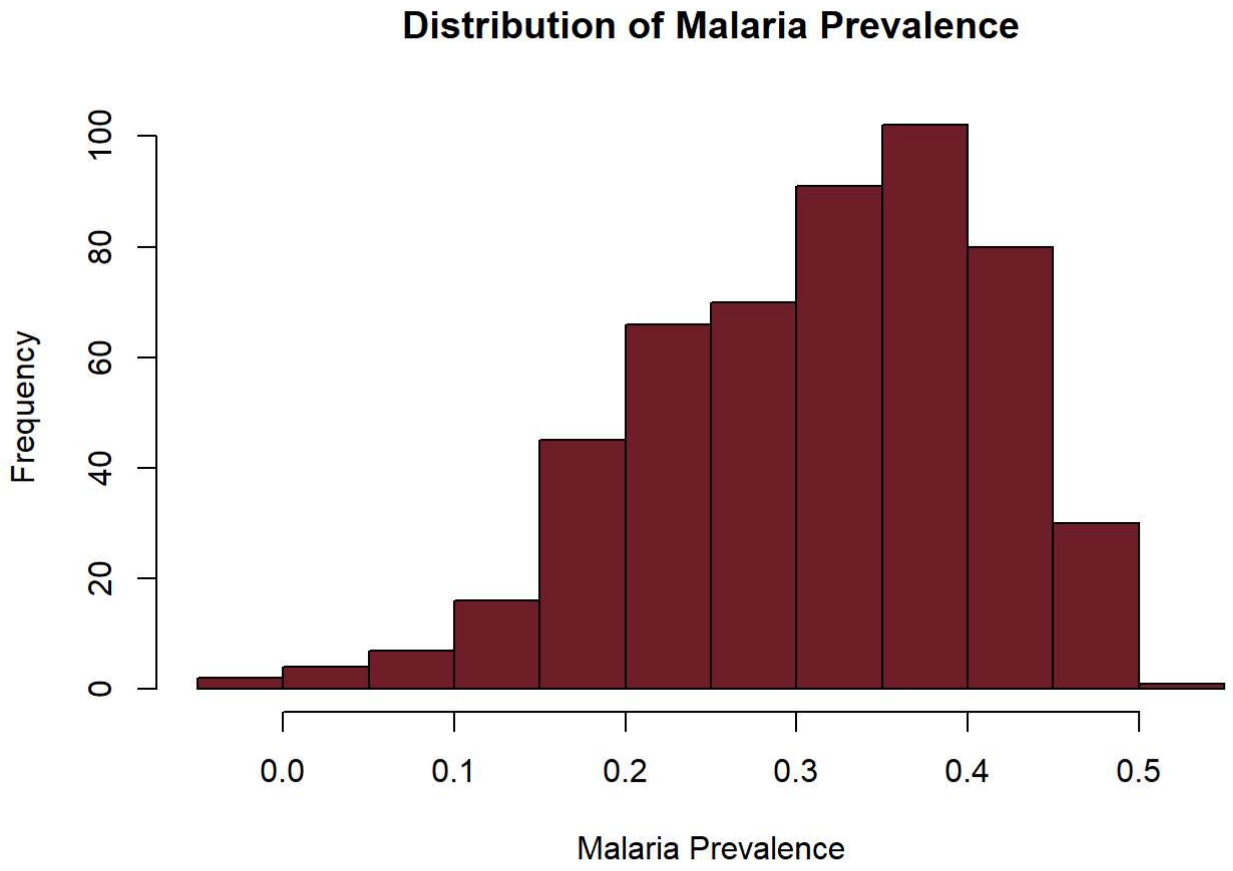
Single variable or column comparison

**Histogram or Frequency Chart

```
         hist(malaria_data$prev)
```

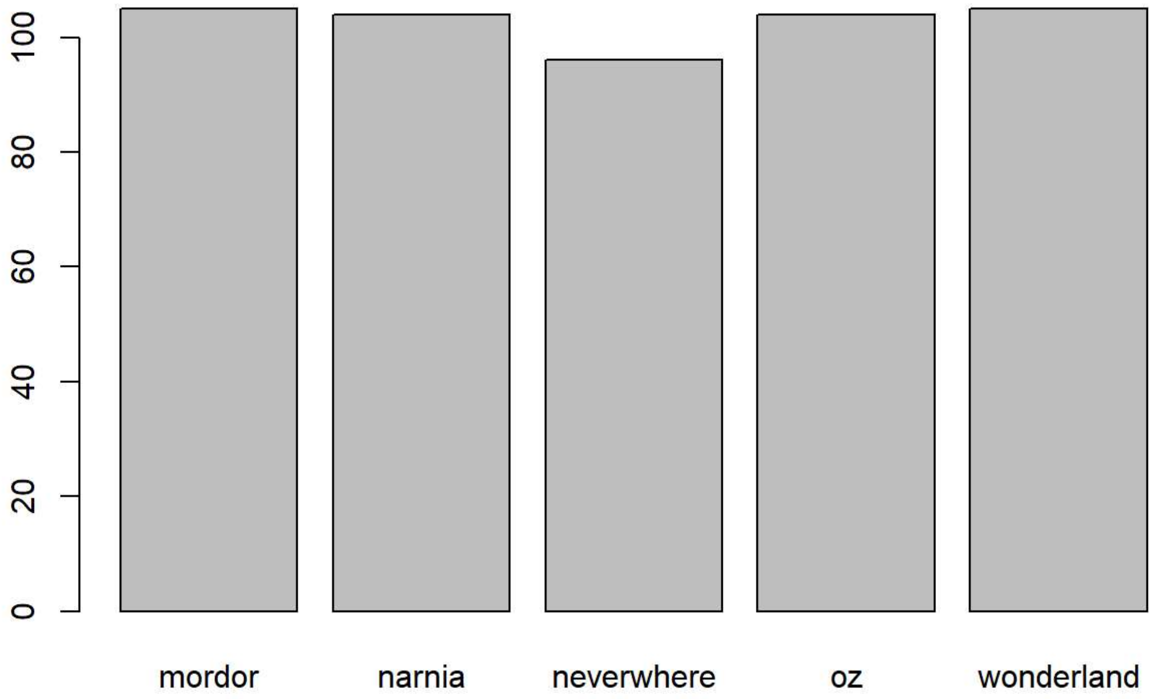


```r
hist(malaria_data$prev,
     breaks = 10, # breaks mean how many individual bars do we need to group
     main = "Distribution of Malaria Prevalence",
     xlab = "Malaria Prevalence",
     ylab = "Frequency",
     col = "#701f28",
     border = "black")
```

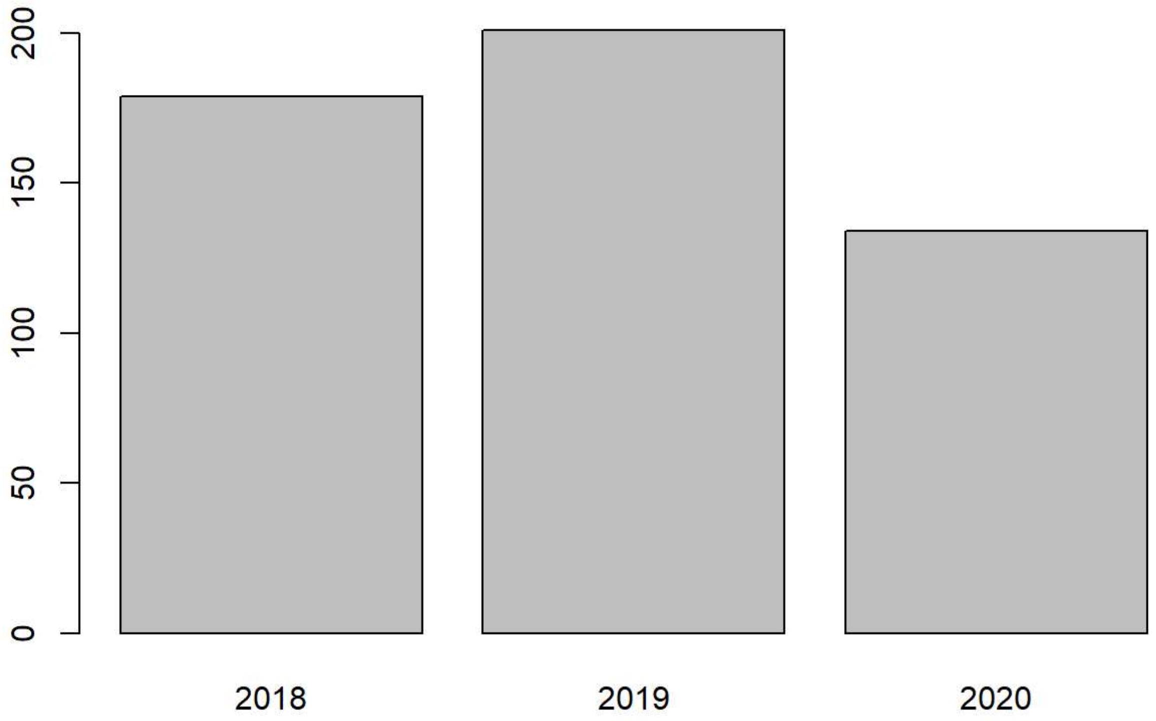


Barplot- tell us the number of counts within a categorical variable/ column

```
barplot(table(malaria_data$location))
```



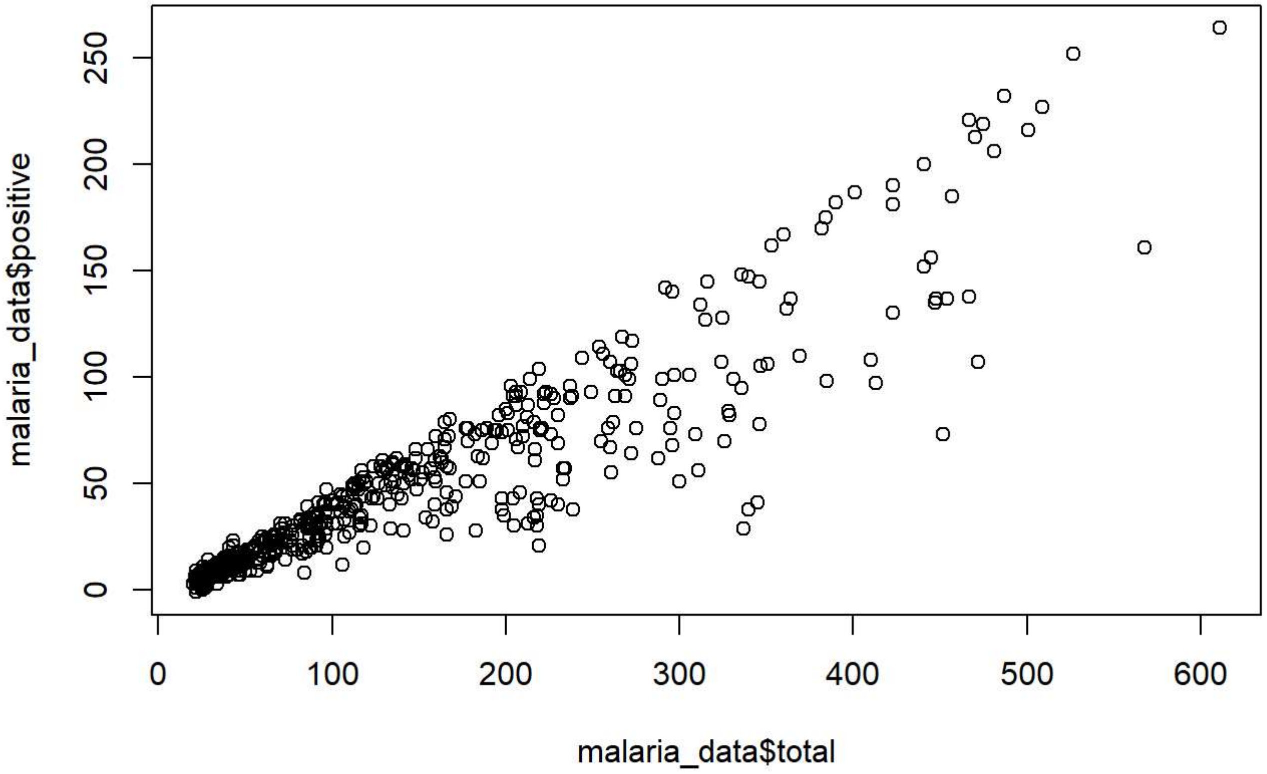
```r
barplot(table(malaria_data$year))
```



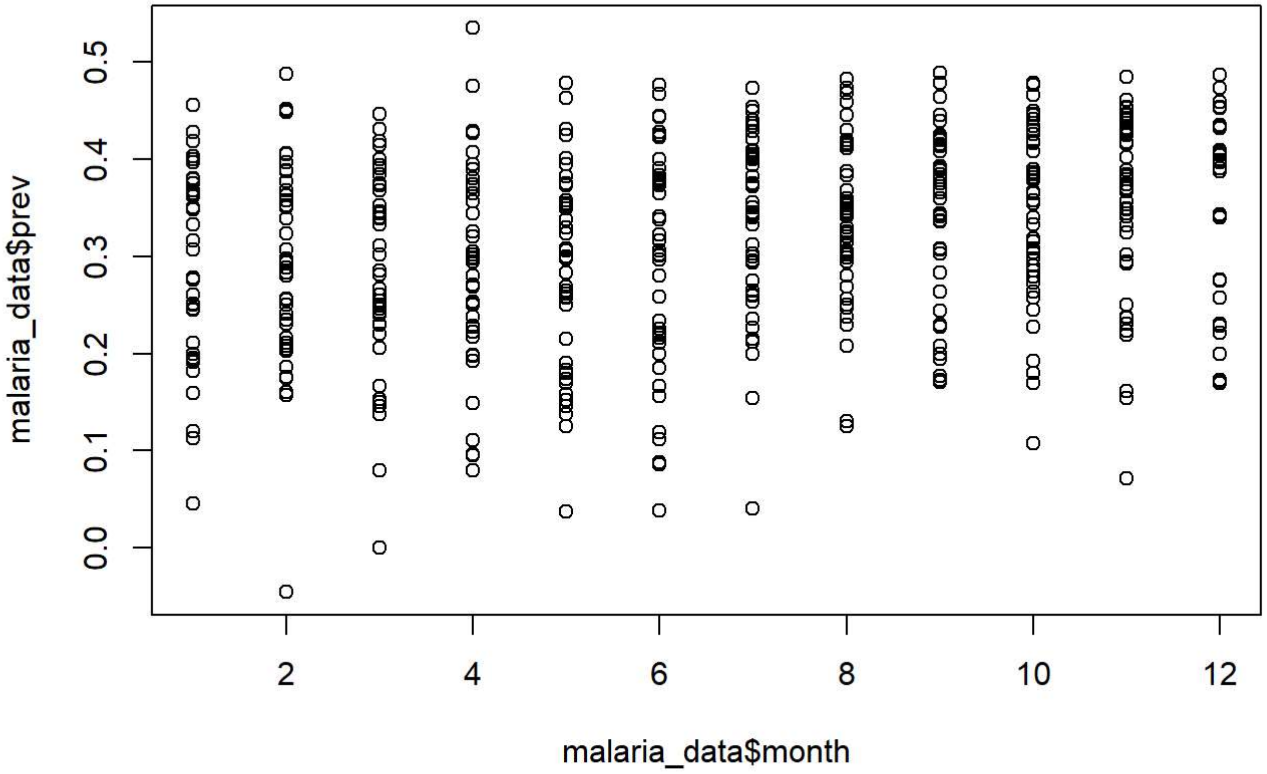
Plotting Multiple column/ variable

Scatterplot using R code - plot

```
        plot(x = malaria_data$total, y = malaria_data$positive)
```

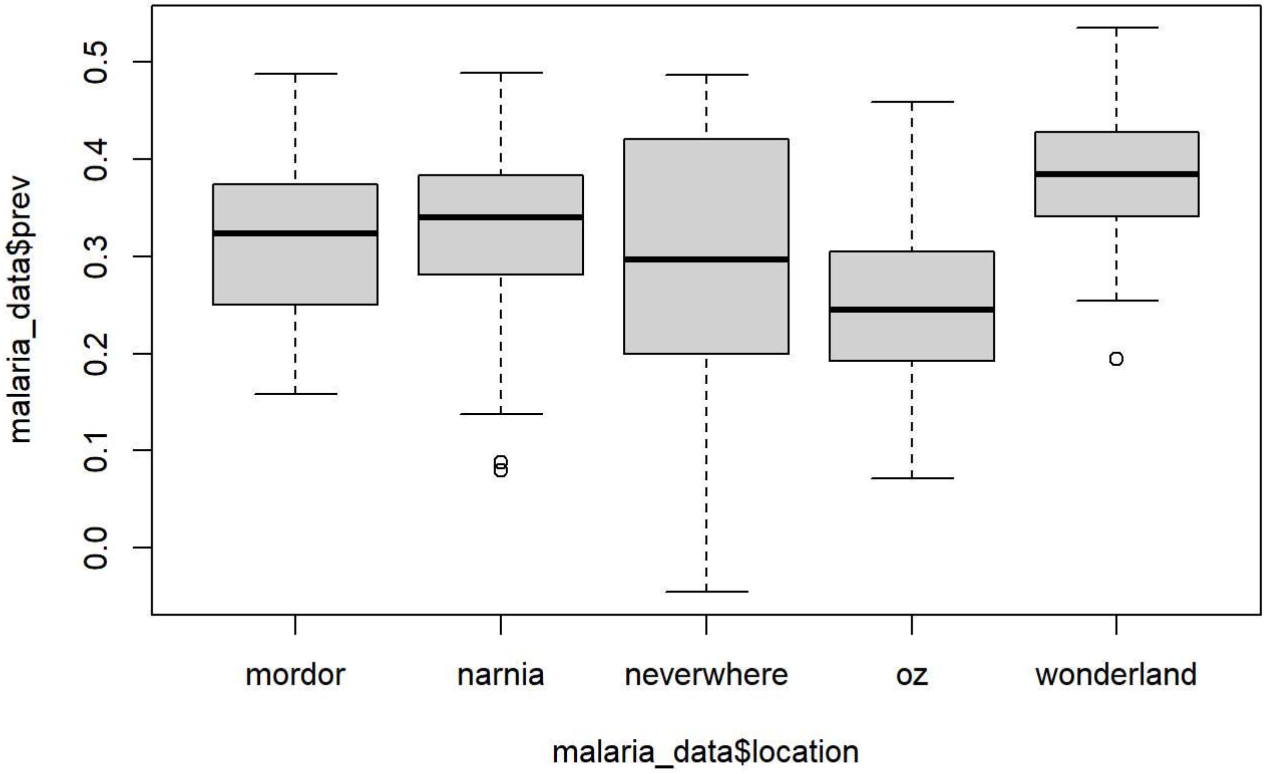


```
plot(x = malaria_data$month, y = malaria_data$prev)
```

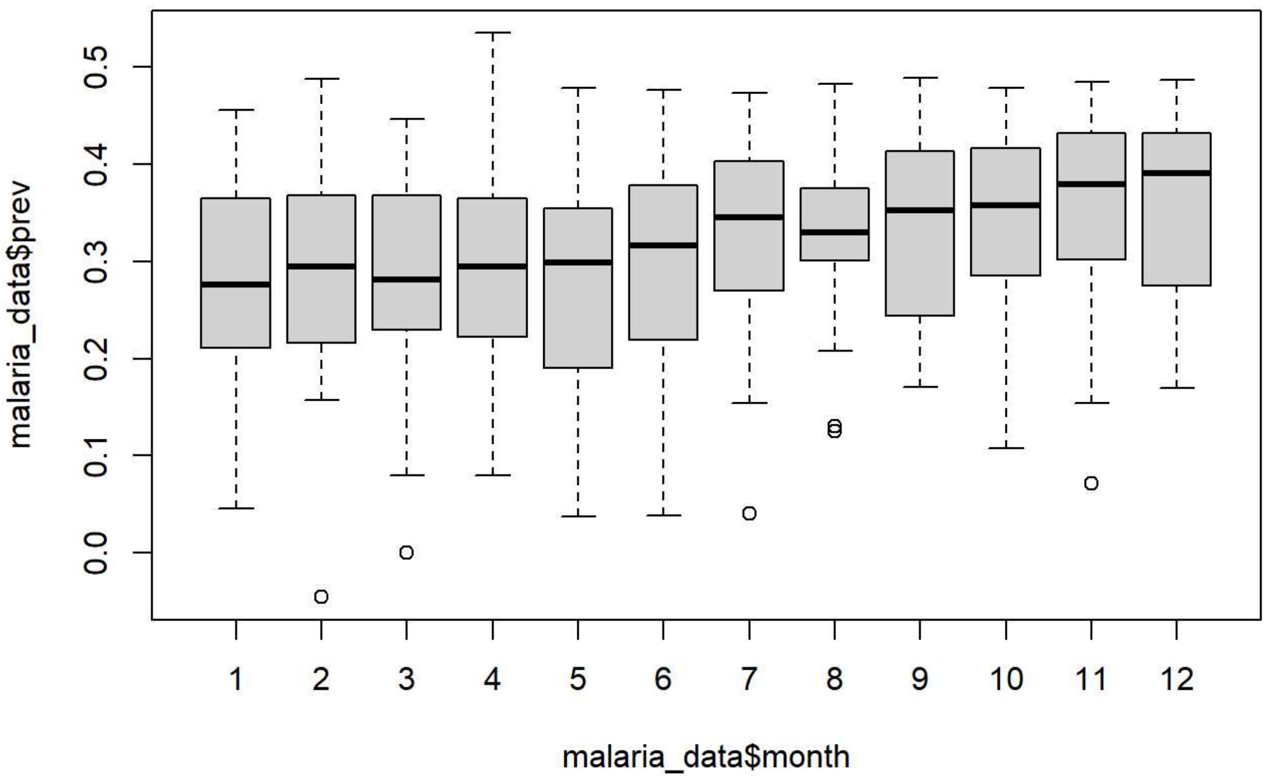


Boxplot using R code - boxplot

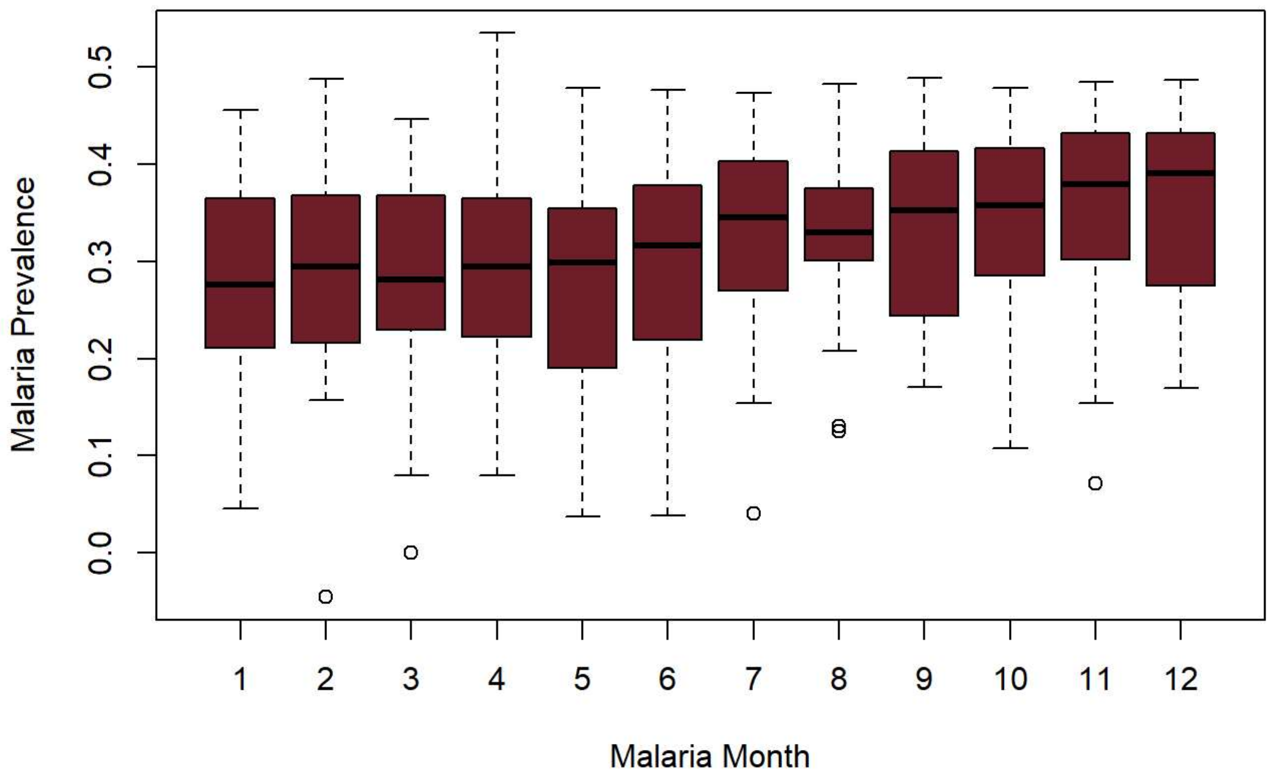
```
boxplot(malaria_data$prev ~ malaria_data$location)
```



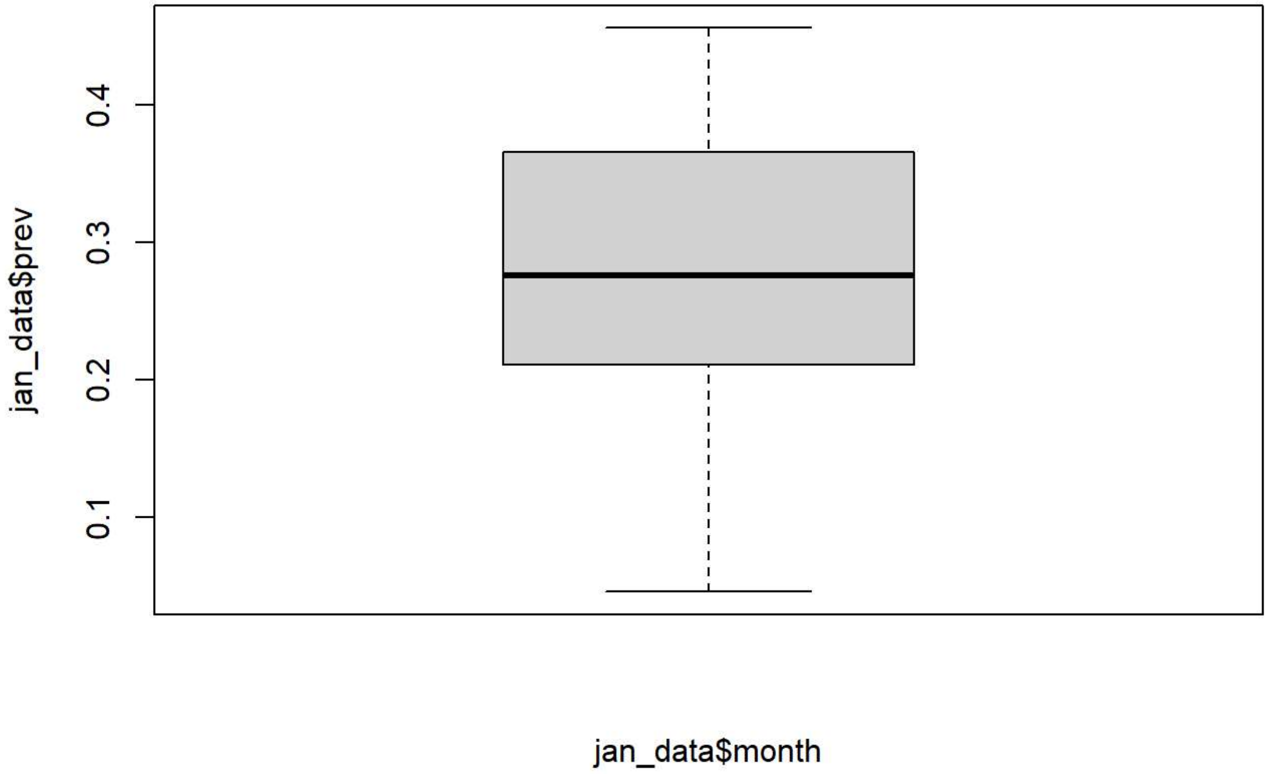
```
boxplot(malaria_data$prev ~ malaria_data$month)
```



```r
boxplot(malaria_data$prev ~ malaria_data$month,
        data = malaria_data,
        xlab = "Malaria Month",
        ylab = "Malaria Prevalence",
        col = "#701f28",
        border = "black")
```



```
jan_data <- filter(malaria_data, month==1)
boxplot(jan_data$prev ~ jan_data$month)
```



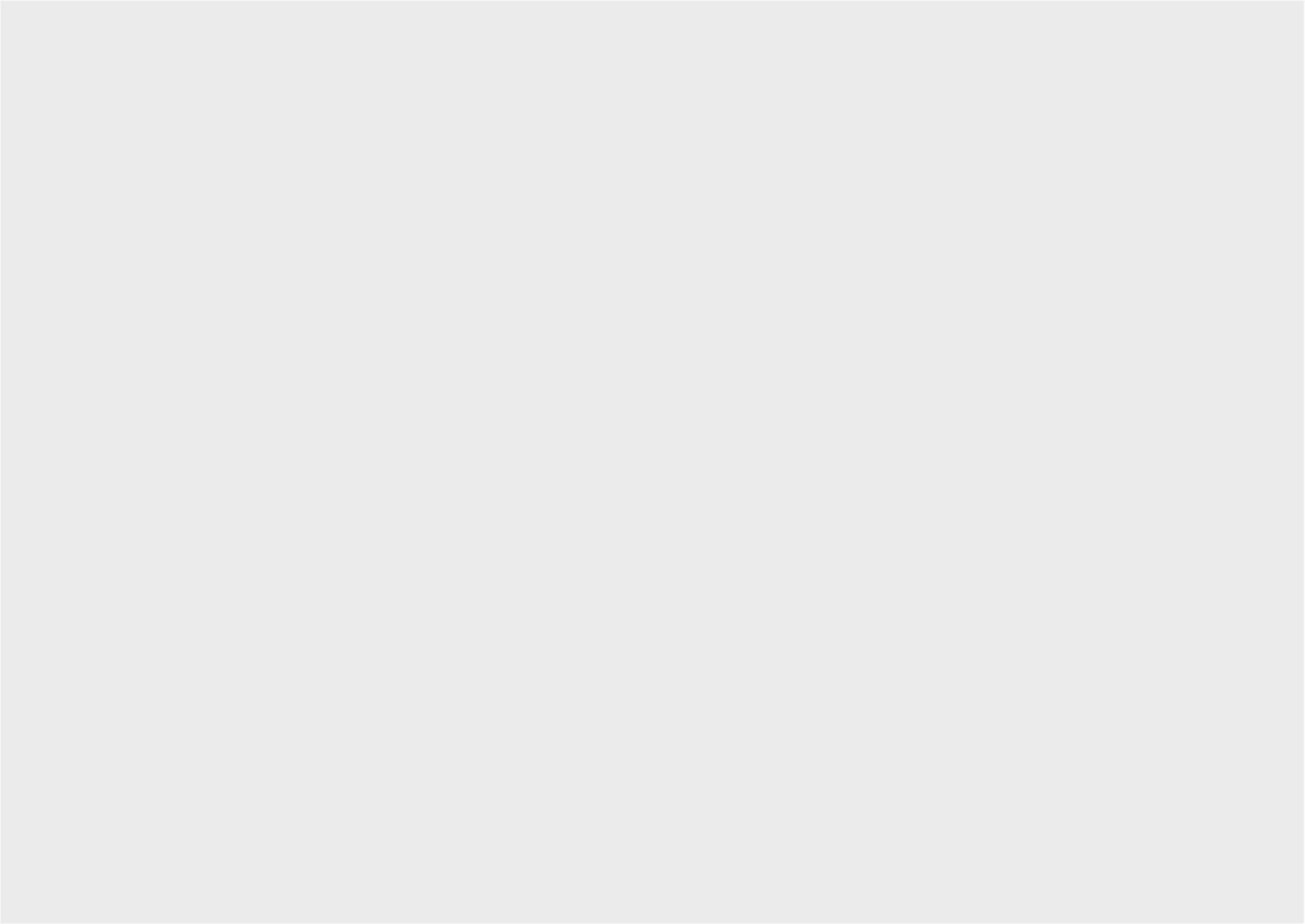
```
boxplot
```

```
## function (x, ...)
## UseMethod("boxplot")
## <bytecode: 0x0000021c4ef14b58>
## <environment: namespace:graphics>
```

Step by Step Data Visualization using ggplot2

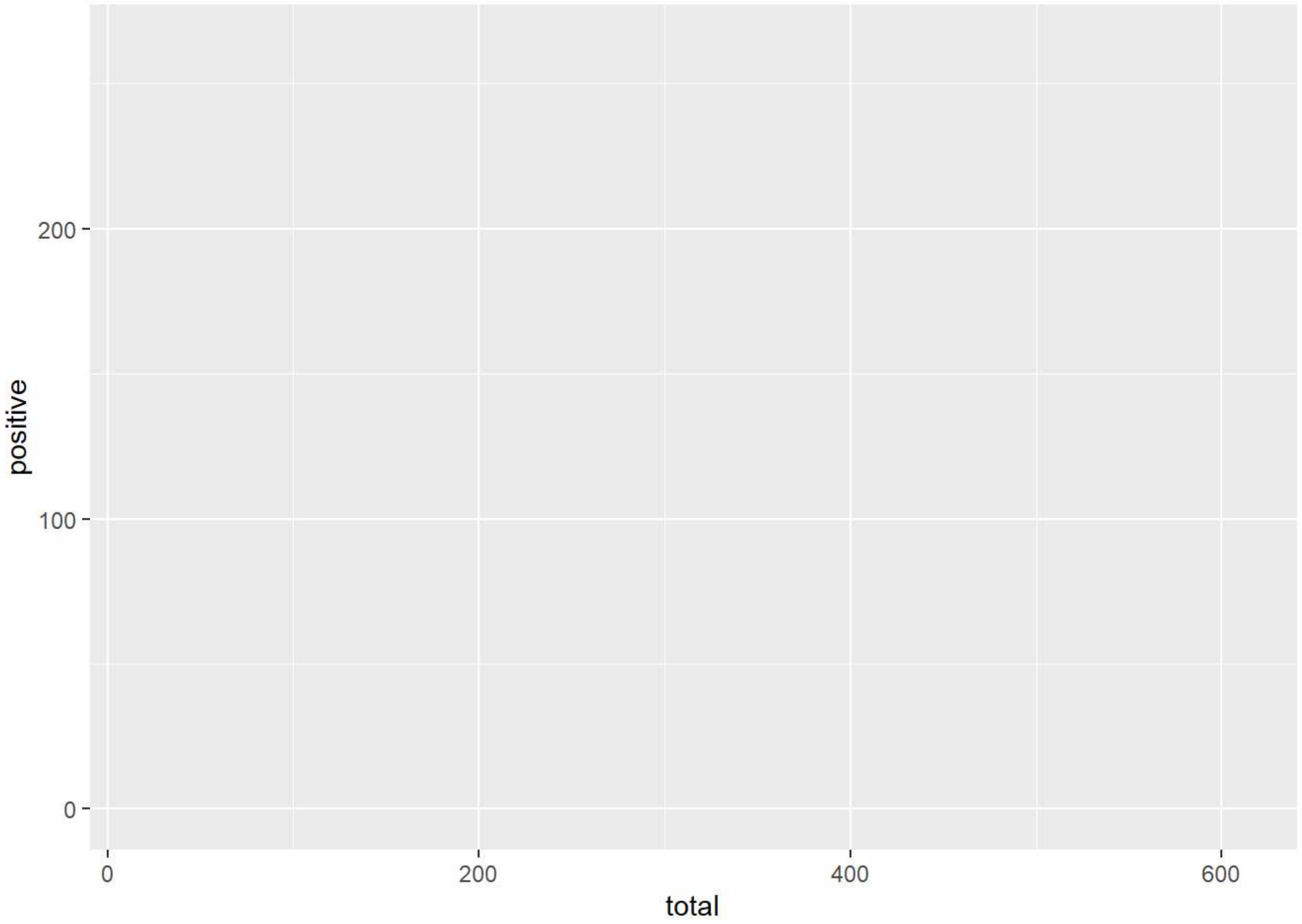
Data

```
ggplot(data = malaria_data)
```



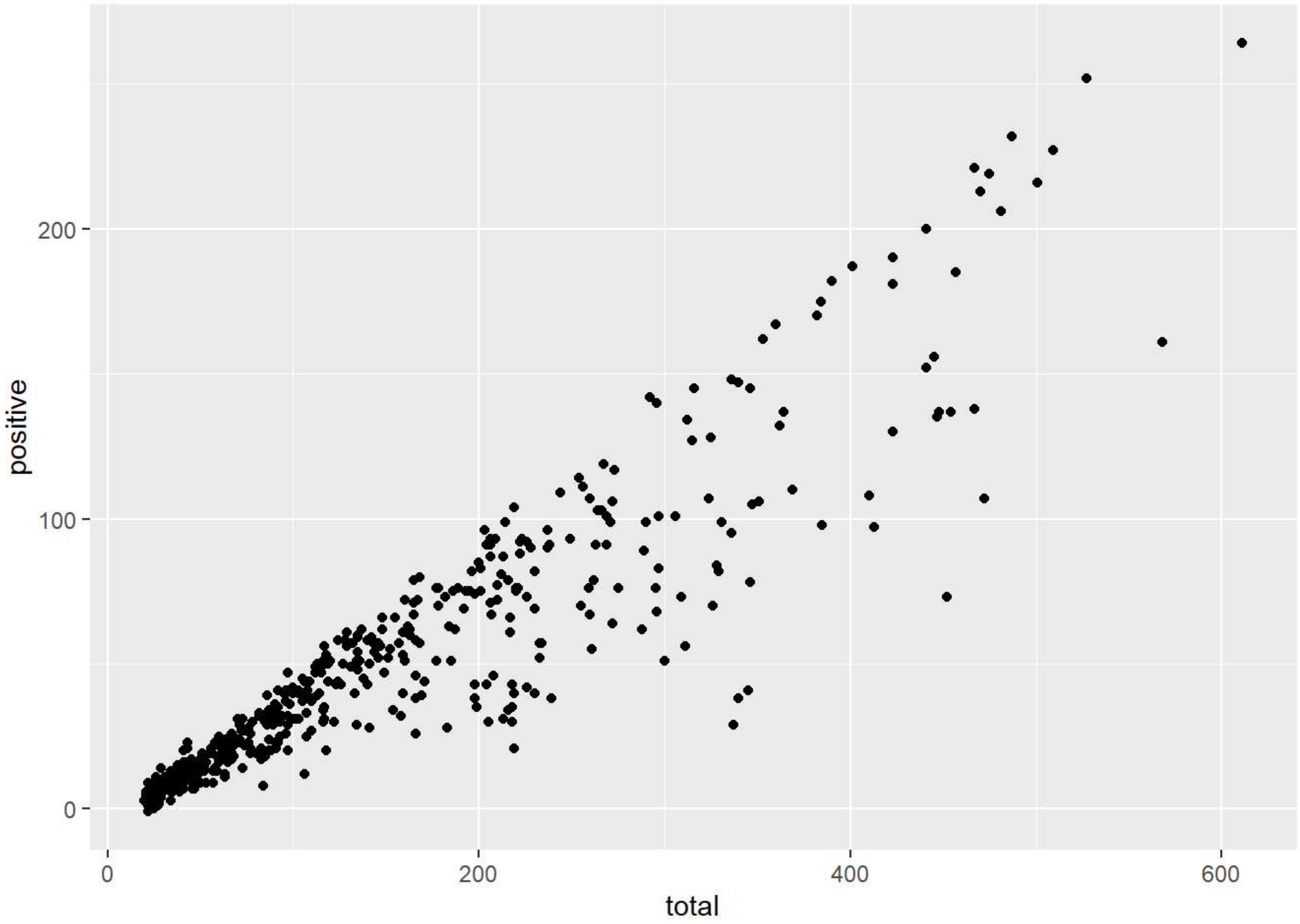
Organize Data(aes())

```
        ggplot(data = malaria_data, aes(x = total, y = positive))
```

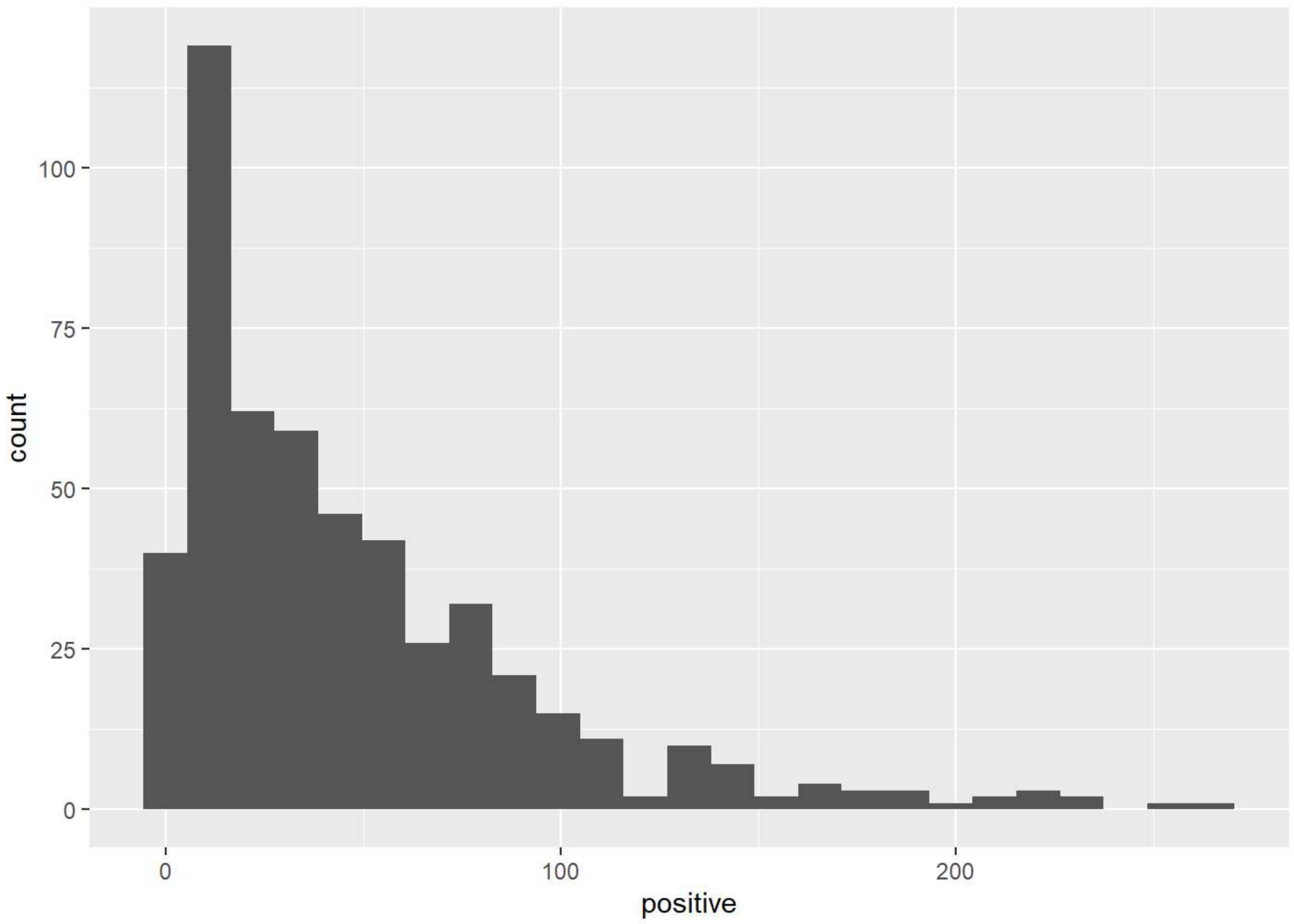


Visualizing Data(geom_)

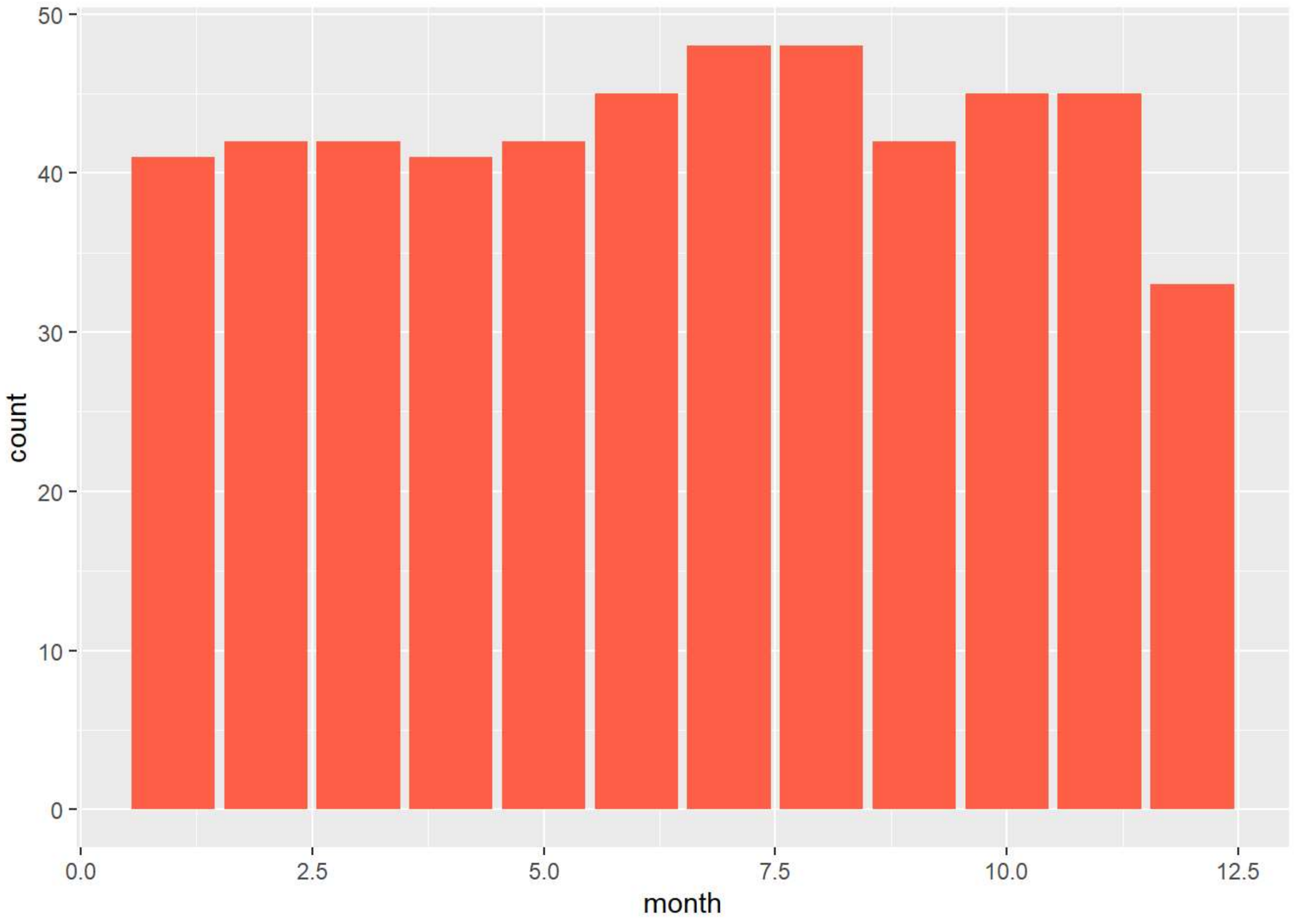
```
        ggplot(data = malaria_data, aes(x = total, y = positive)) + geom_point()
```



```
ggplot(data = malaria_data, aes(x = positive)) + geom_histogram(bins = 25)
```

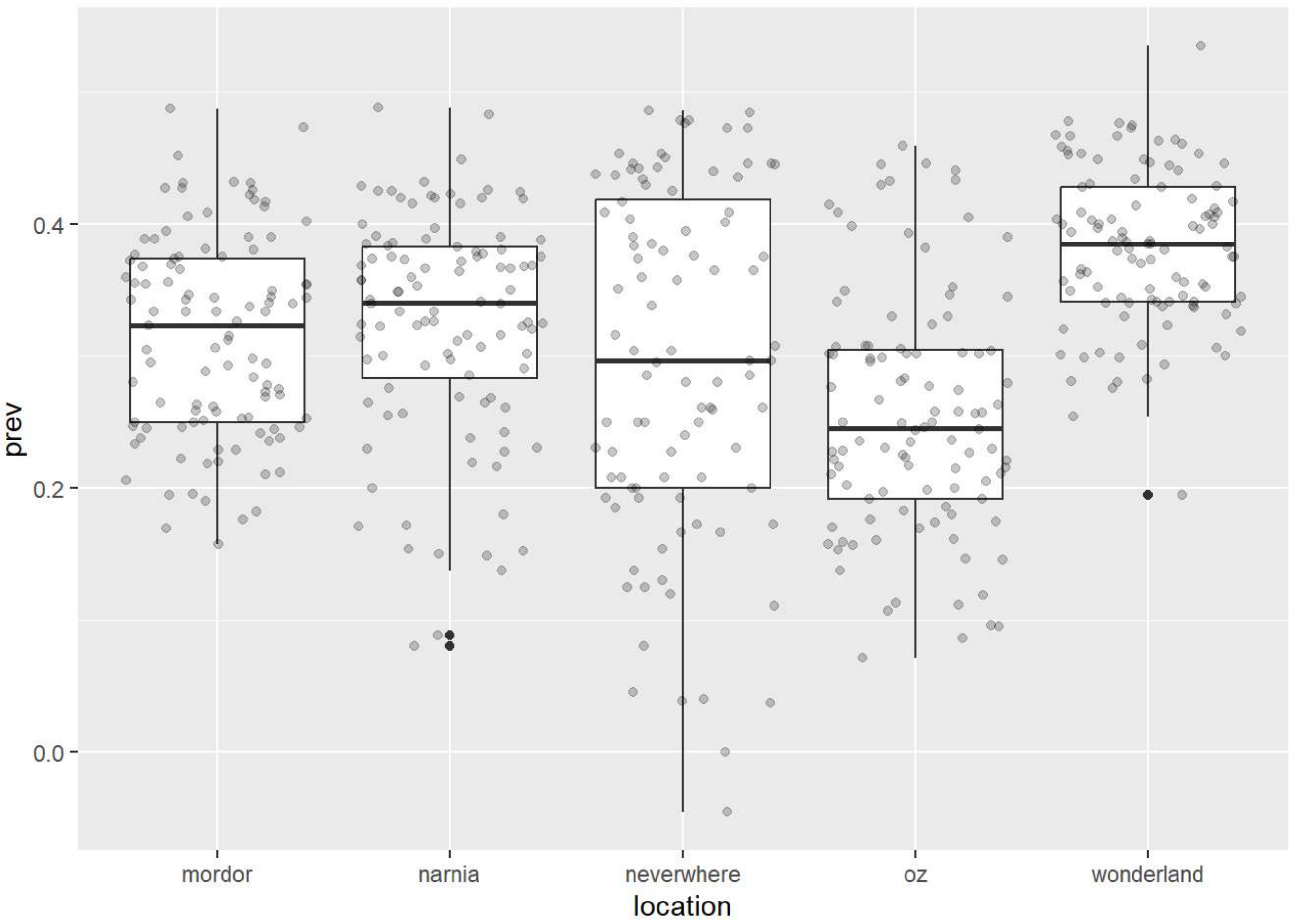


```
ggplot(data = malaria_data, aes(x = month)) + geom_bar(fill = "tomato")
```

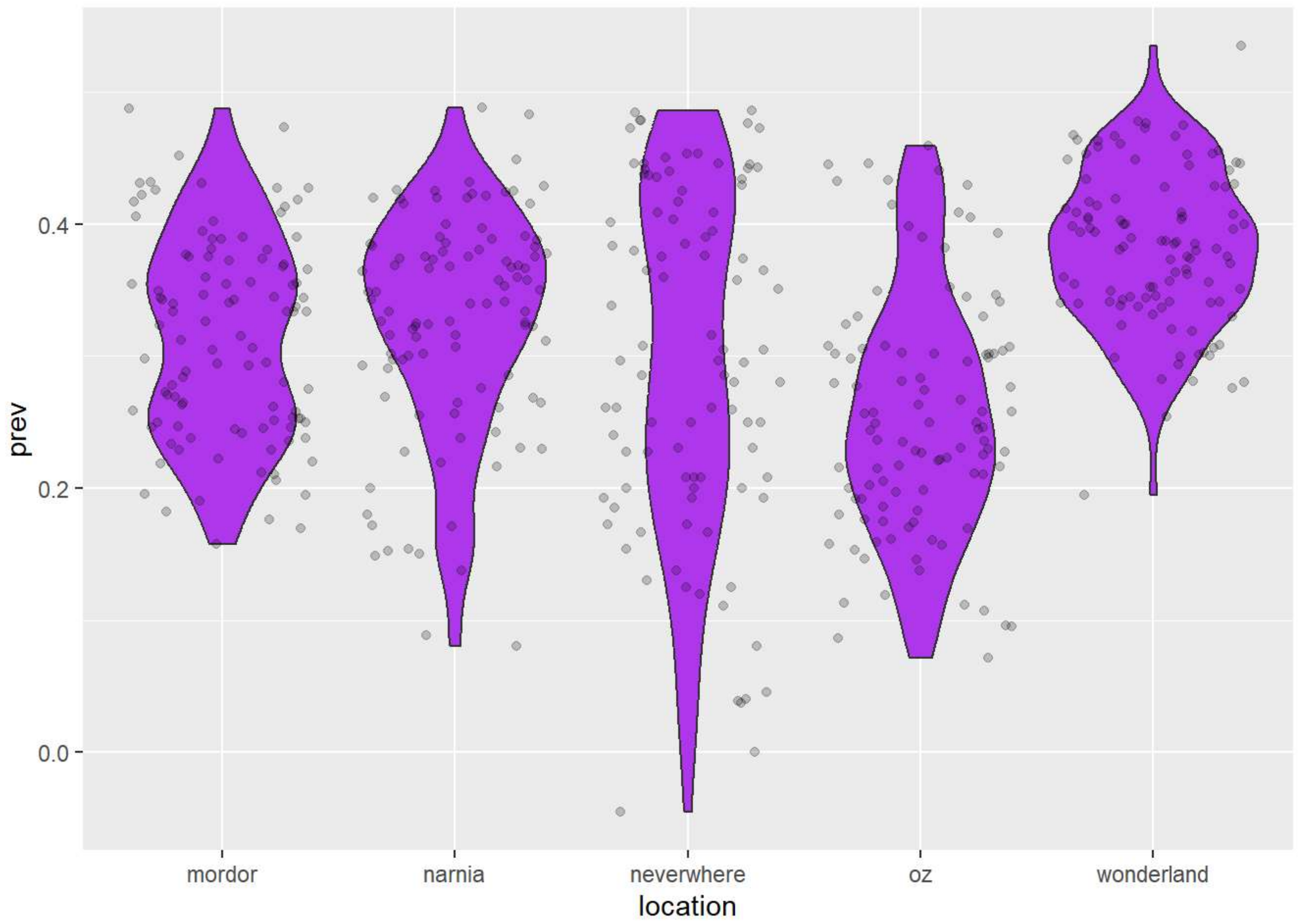


Add multiple geoms_

```
         ggplot(data = malaria_data, aes(x = location, y = prev)) + geom_boxplot()+
           geom_jitter(alpha = 0.2) #Boxplot shows value for the range but they don't show distribution
```

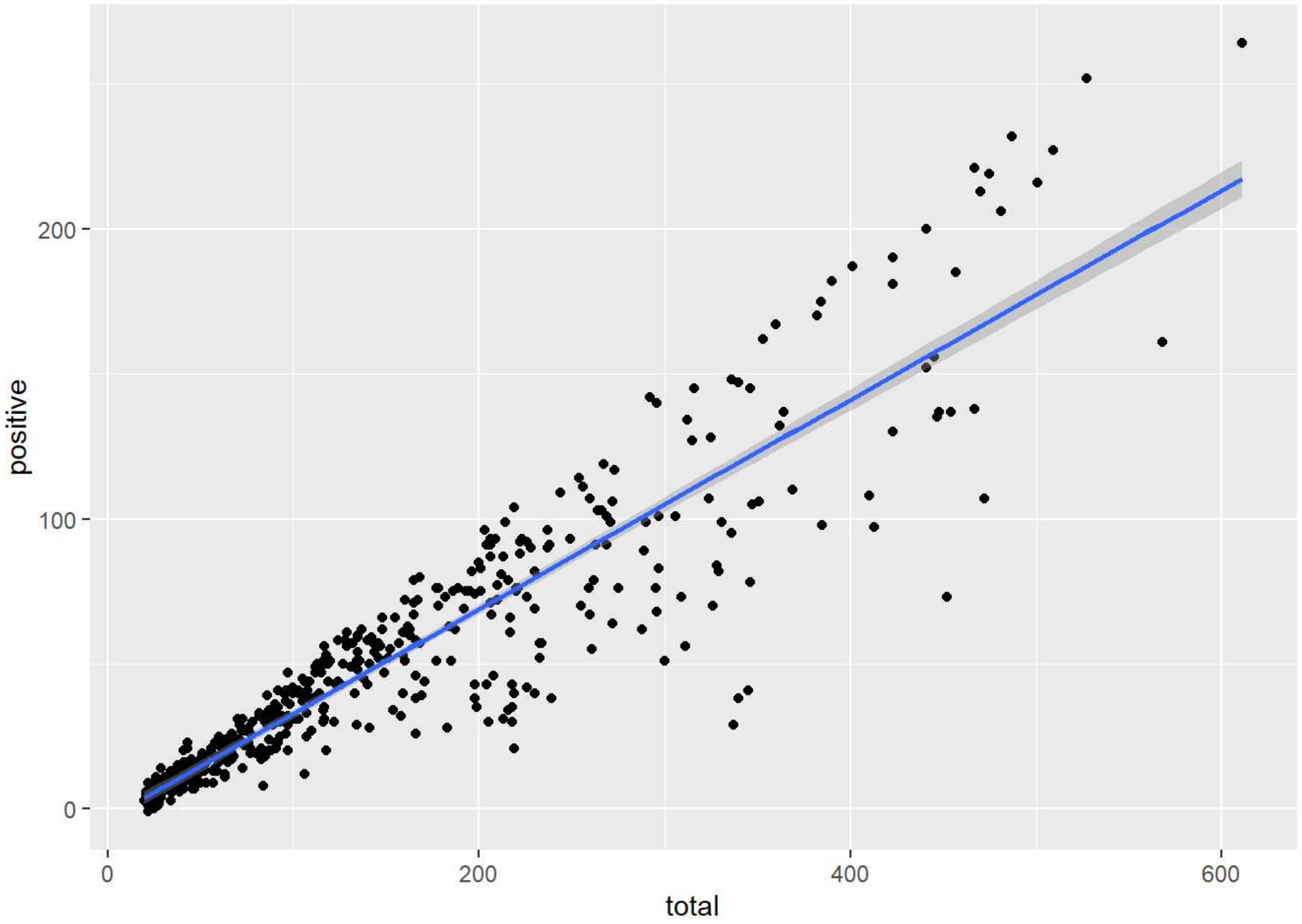


```
       ggplot(data = malaria_data, aes(x = location, y = prev)) + geom_violin(fill= "darkorchid2")+
         geom_jitter(alpha = 0.2) #Similar to the boxplot but the shape is different shows distribution
```



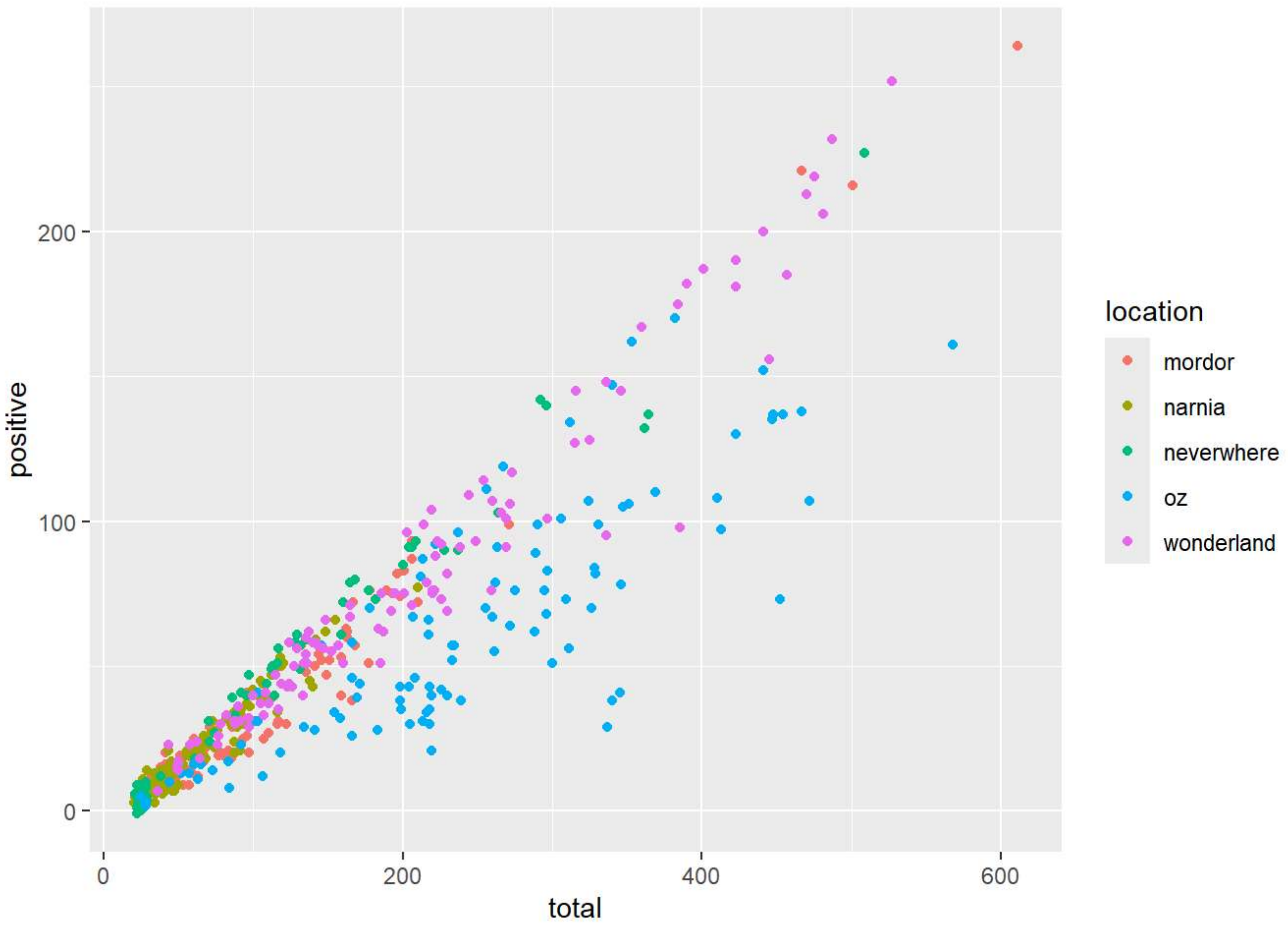
```
        ggplot(data = malaria_data, aes(x = total, y = positive)) + geom_point() +  geom_smooth
(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

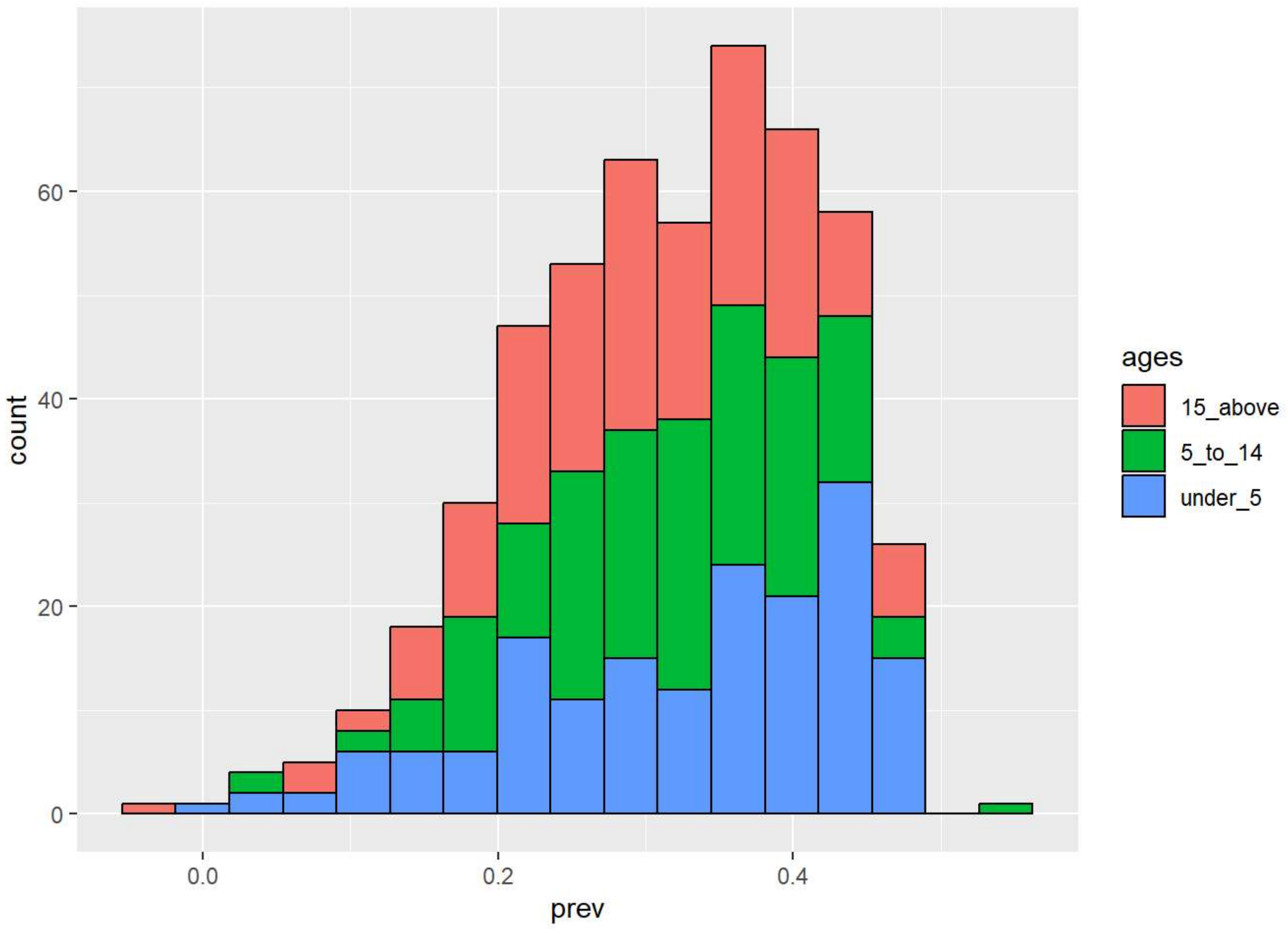


Extending the aesthetic

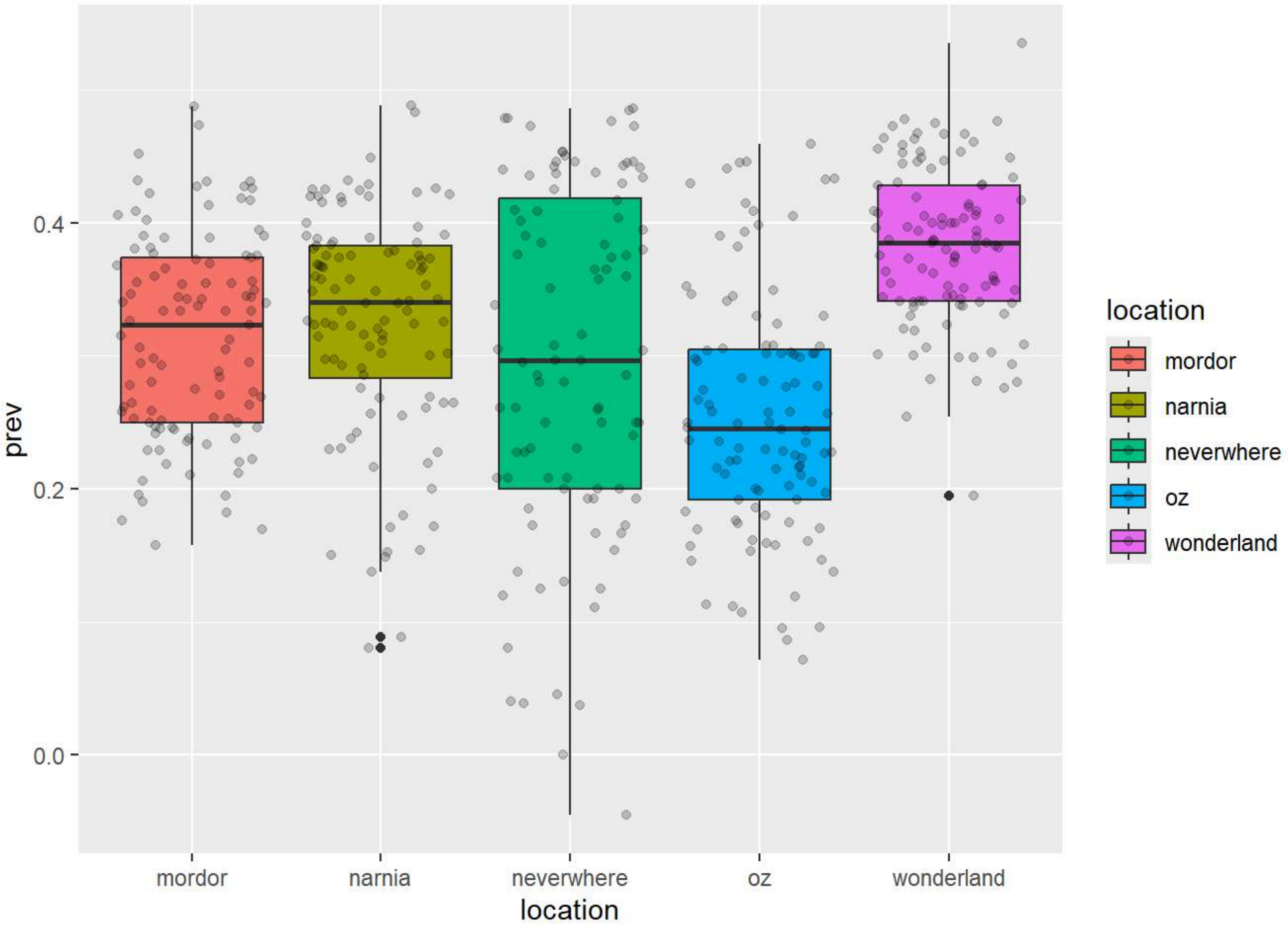
```
        ggplot(data = malaria_data, aes(x = total, y = positive, colour = location)) + geom_poi
nt()
```



```
        ggplot(data = malaria_data, aes(x = prev, fill = ages)) + geom_histogram(colour = "blac
k", bins = 17)
```

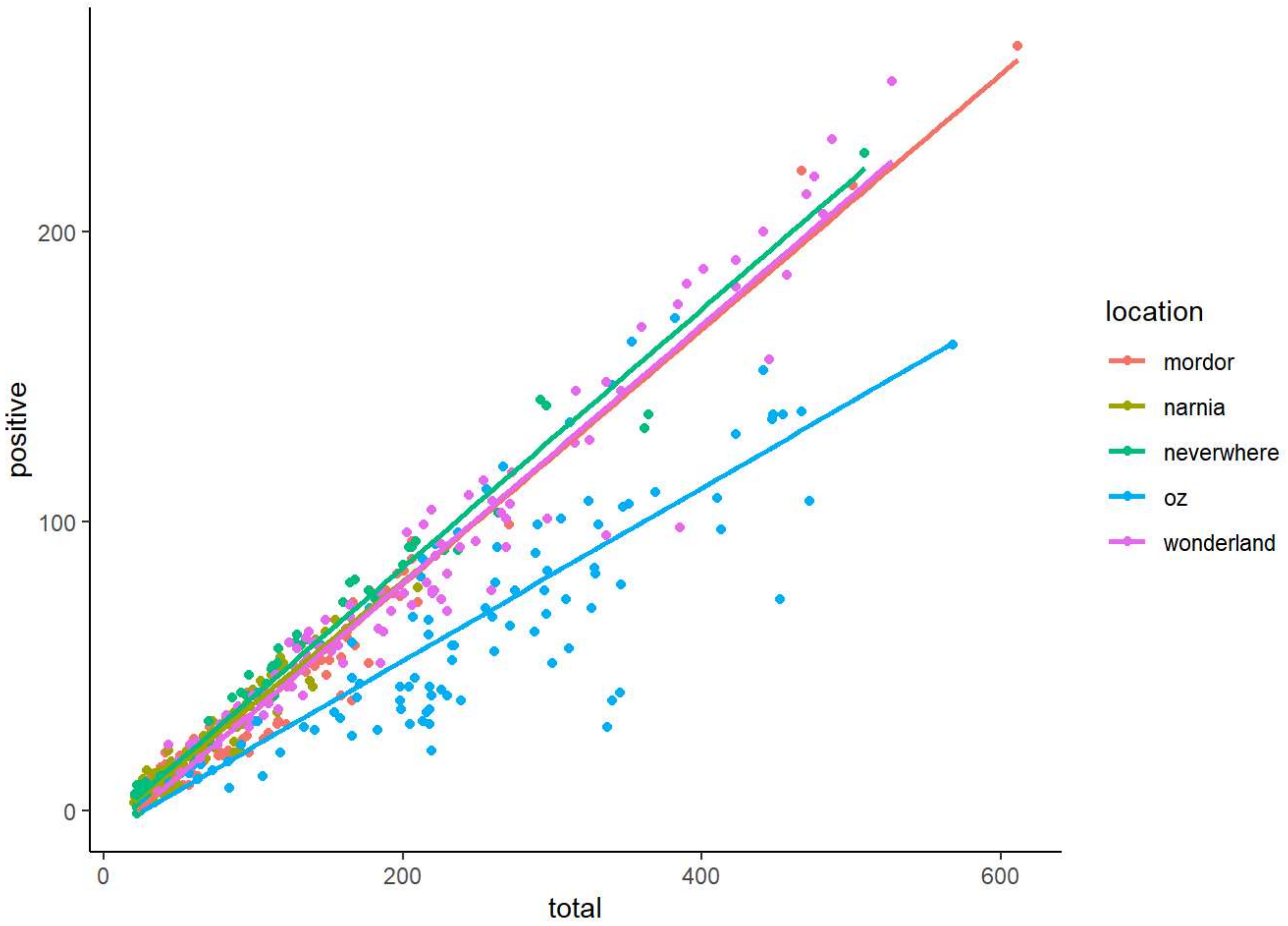


```
        ggplot(data = malaria_data, aes(x = location, y = prev, fill = location)) + geom_boxplo
t()+
          geom_jitter(alpha = 0.2)
```

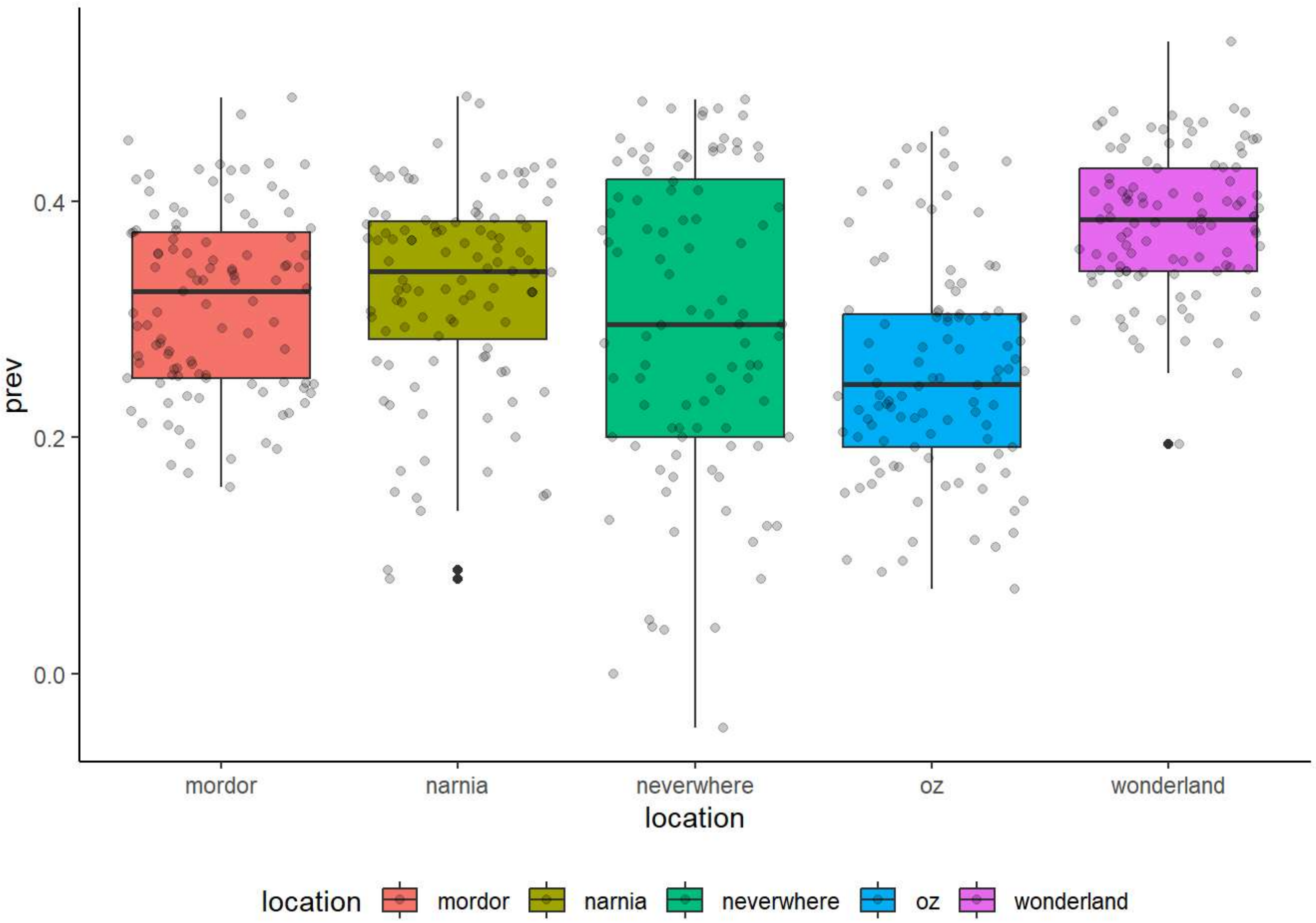


```
        ggplot(data = malaria_data, aes(x = total, y = positive, colour = location)) + geom_poi
nt()+ geom_smooth(method = "lm", se= FALSE) + theme_classic()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

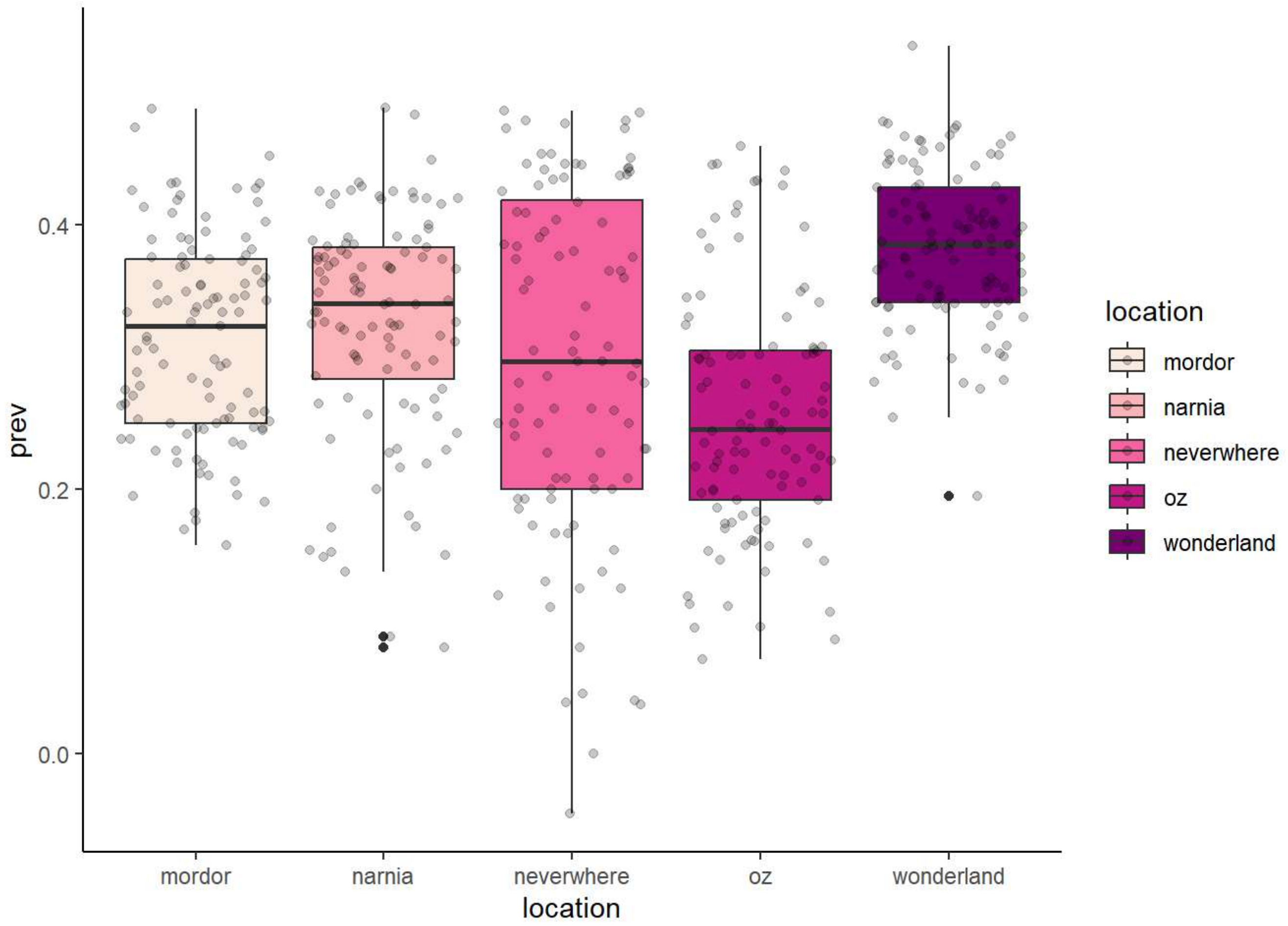


```
        ggplot(data = malaria_data, aes(x = location, y = prev, fill = location)) + geom_boxplo
t()+
          geom_jitter(alpha = 0.2) +theme_classic() + theme(legend.position = "bottom")
```



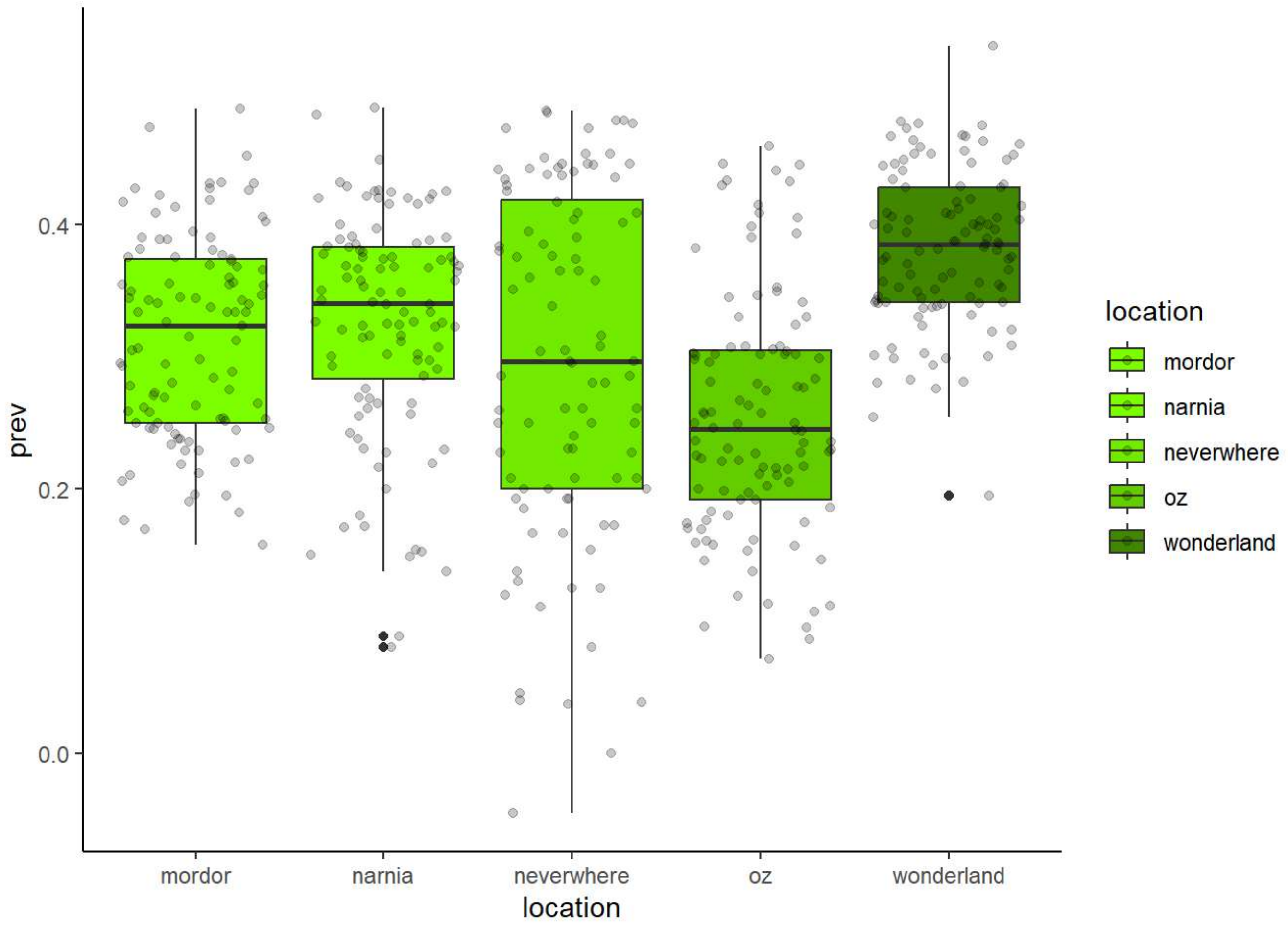
Add colour Palette automatic using scale fill brewer

```
        ggplot(data = malaria_data, aes(x = location, y = prev, fill = location)) + geom_boxplo
t()+
          geom_jitter(alpha = 0.2) +theme_classic() + scale_fill_brewer(palette = "RdPu")
```



Add colour manually

```
        ggplot(data = malaria_data, aes(x = location, y = prev, fill = location)) + geom_boxplo
t()+
          geom_jitter(alpha = 0.2) +theme_classic() + scale_fill_manual(values = c("chartreus
e", "chartreuse1", "chartreuse2","chartreuse3", "chartreuse4"))
```

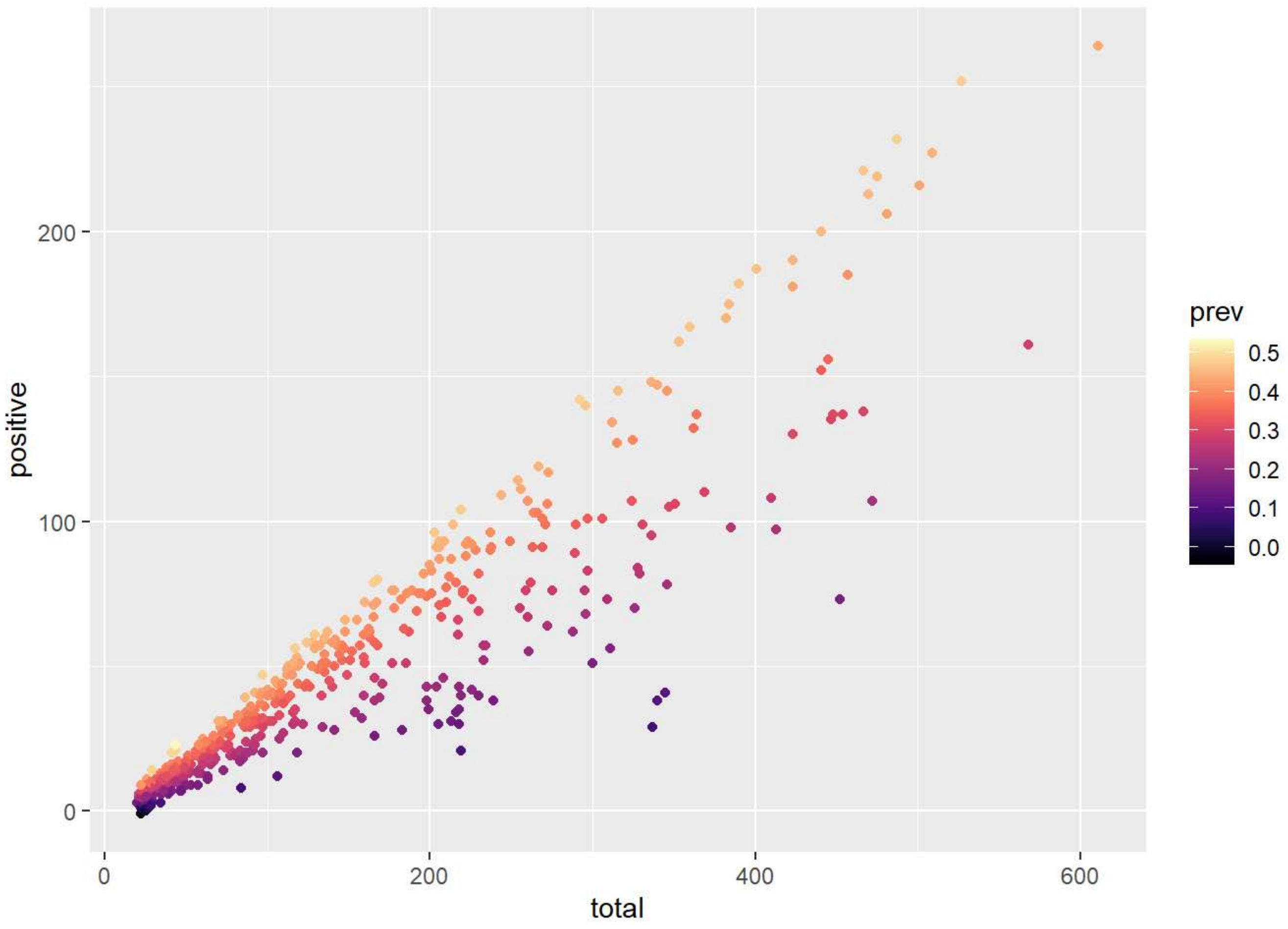


Use viridis package to create custom color palettes

```
        library(viridis)
```

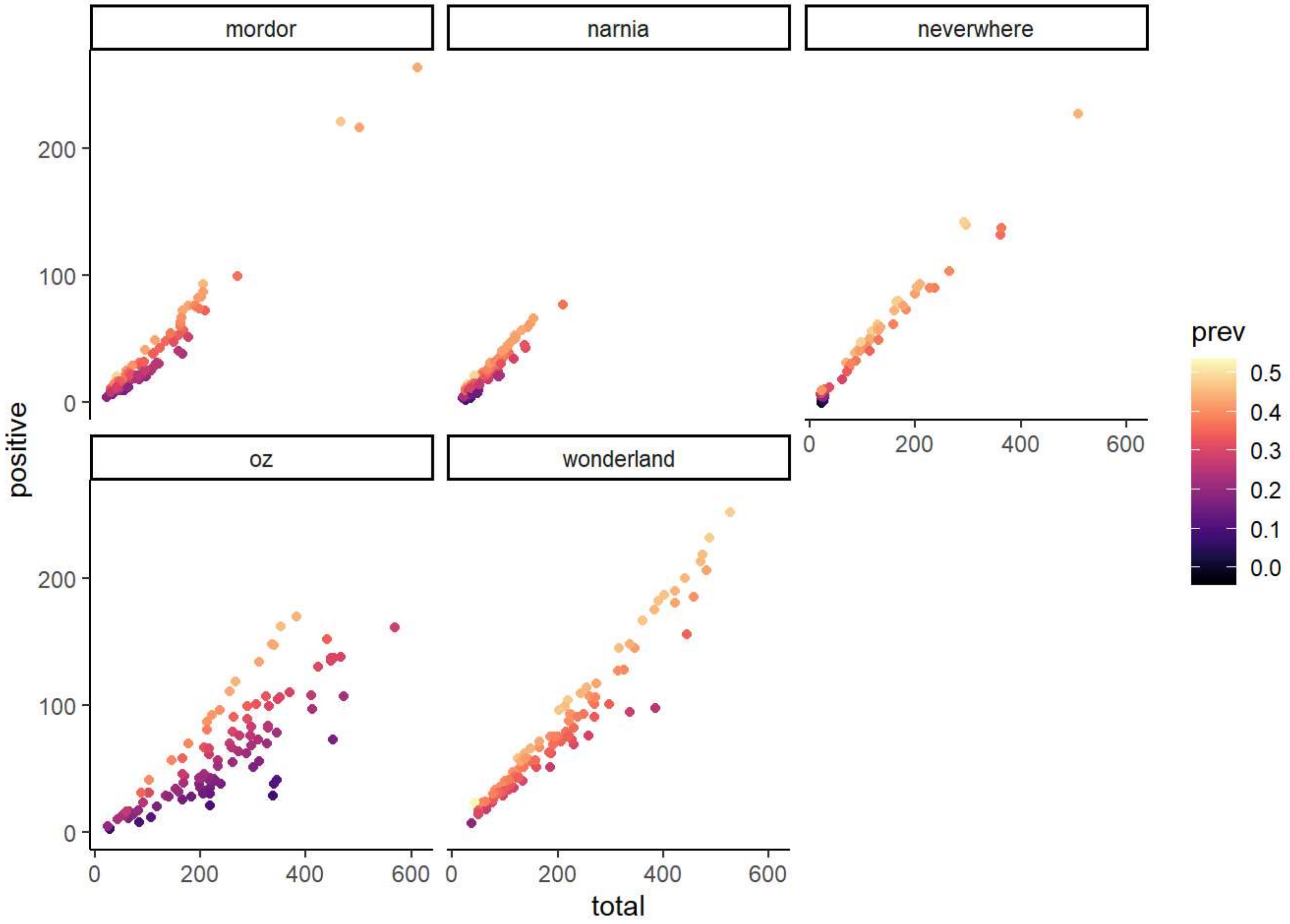
```
## Loading required package: viridisLite
```

```
        ggplot(data = malaria_data, aes(x = total, y = positive, colour = prev)) + geom_point()
+
          scale_color_viridis(option = "magma")
```



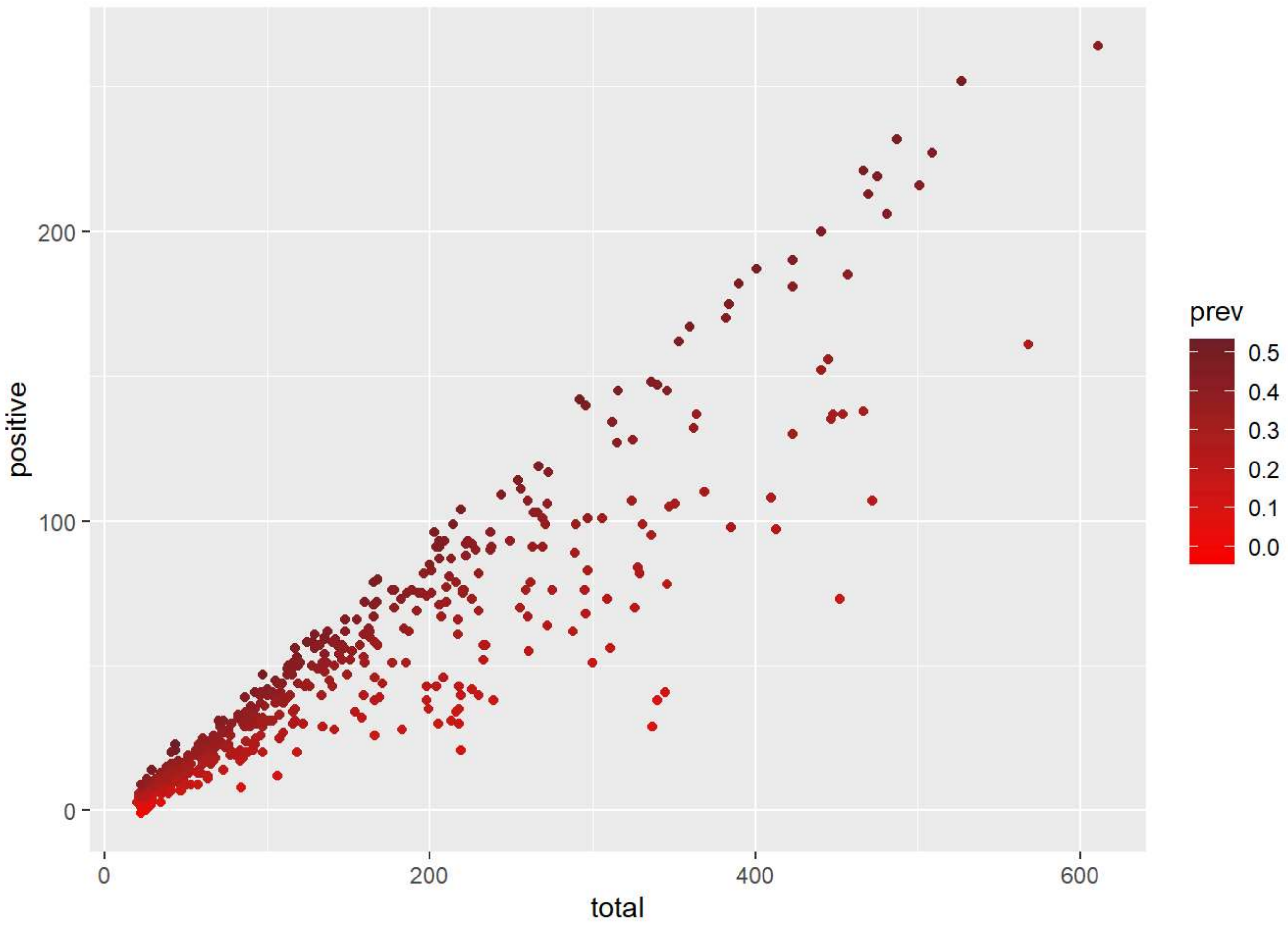
Small multiple plots using facet

```
         ggplot(data = malaria_data, aes(x = total, y = positive, colour = prev)) + geom_point()
+
           scale_color_viridis(option = "magma") +
           facet_wrap(~location)+theme_classic()
```



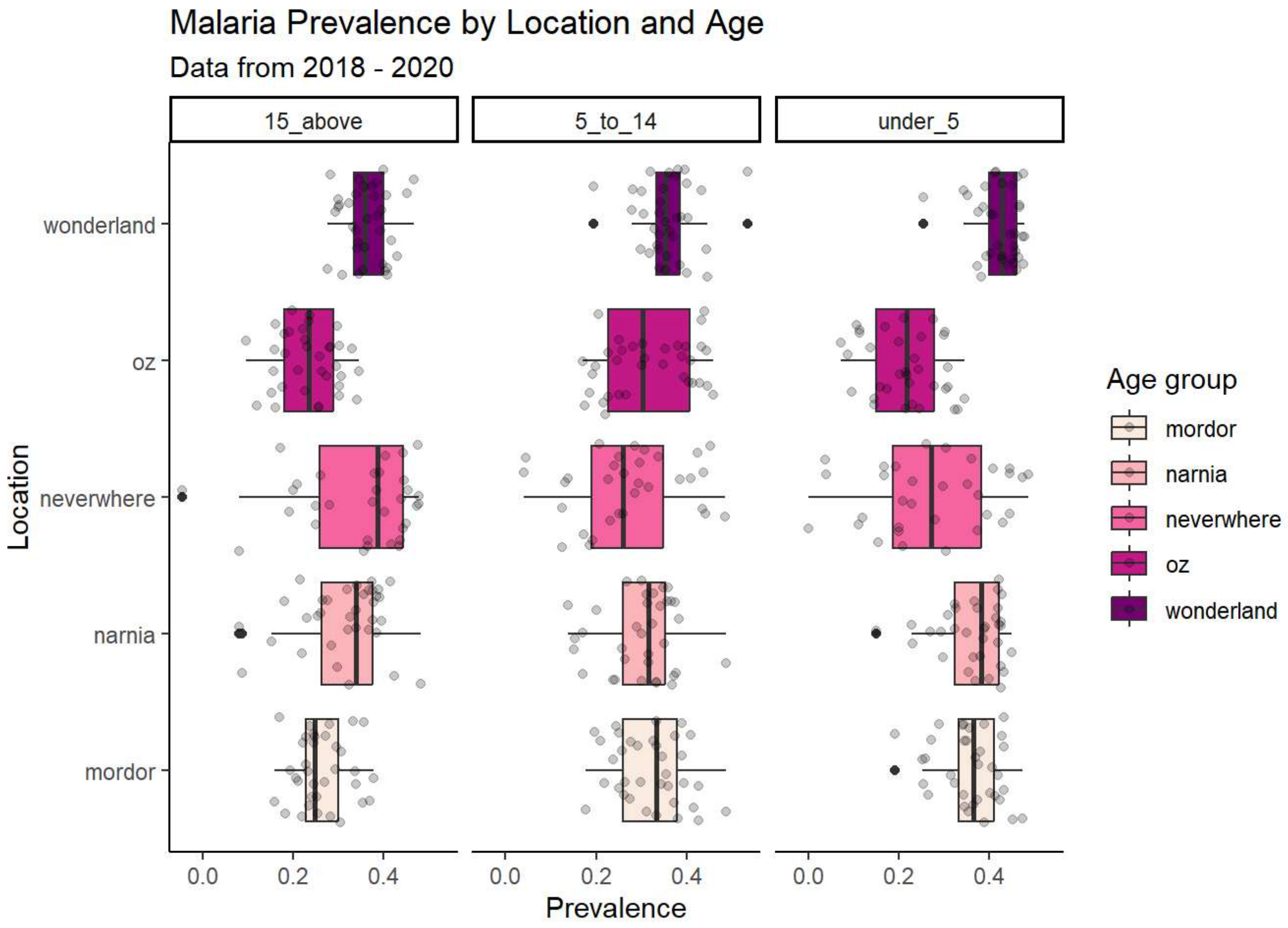
Visualize continous data on a spectrum using scale color gradient

```
        ggplot(data = malaria_data, aes(x = total, y = positive, colour = prev)) + geom_point()
+
          scale_color_gradient(low = "red", high = "#701f28")
```



Adding title, subtitle and flipping x and y axis

```
         ggplot(data = malaria_data, aes(x = location, y = prev, fill = location)) + geom_boxplot()+
           geom_jitter(alpha = 0.2) +
           facet_wrap(~ages) +
           coord_flip() + # flips the x and y axis
           theme_classic() + scale_fill_brewer(palette = "RdPu")+
           labs(title = "Malaria Prevalence by Location and Age",
                subtitle = "Data from 2018 - 2020",
                x = "Location",
                y = "Prevalence",
                fill = "Age group")
```



```
        ggsave("malaria_age_prevalenceboxplot.png", width = 10, height = 6, units = "in", dpi =
300)
```