STOR 320 Tutorial on Data Visualization

January 15, 2021

Introduction to RMarkdown and ggplot2

This is the default R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

For **our** class we will always, **Knit to PDF**!!!

For more assistance with RMarkdown, see Chapter 21 in *R for Data Science* and the RMarkdown cheat sheet at https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf, which link is also found on the course website.

Overview of the Mammals Sleep Dataset from the Tidyverse

```
msleep #Prints the data but takes up a lot of space
```

```
## # A tibble: 83 x 11
##    name  genus vore  order conservation sleep_total sleep_rem sleep_cycle awake
##    <chr> <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl> <dbl>
##  1 Chee~ Acin~ carni Carn~ lc                  12.1      NA        NA      11.9
##  2 Owl ~ Aotus omni  Prim~ <NA>                17         1.8      NA       7
##  3 Moun~ Aplo~ herbi Rode~ nt                  14.4       2.4      NA       9.6
##  4 Grea~ Blar~ omni  Sori~ lc                  14.9       2.3       0.133   9.1
##  5 Cow   Bos   herbi Arti~ domesticated         4         0.7       0.667  20
##  6 Thre~ Brad~ herbi Pilo~ <NA>                14.4       2.2       0.767   9.6
##  7 Nort~ Call~ carni Carn~ vu                   8.7       1.4       0.383  15.3
##  8 Vesp~ Calo~ <NA>  Rode~ <NA>                 7        NA        NA      17
##  9 Dog   Canis carni Carn~ domesticated        10.1       2.9       0.333  13.9
## 10 Roe ~ Capr~ herbi Arti~ lc                   3        NA        NA      21
## # ... with 73 more rows, and 2 more variables: brainwt <dbl>, bodywt <dbl>
```

```
head(msleep,5) #Prints the first 5 rows
```

```
## # A tibble: 5 x 11
##   name  genus vore  order conservation sleep_total sleep_rem sleep_cycle awake
##   <chr> <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl> <dbl>
## 1 Chee~ Acin~ carni Carn~ lc                  12.1      NA        NA      11.9
## 2 Owl ~ Aotus omni  Prim~ <NA>                17         1.8      NA       7
## 3 Moun~ Aplo~ herbi Rode~ nt                  14.4       2.4      NA       9.6
## 4 Grea~ Blar~ omni  Sori~ lc                  14.9       2.3       0.133   9.1
## 5 Cow   Bos   herbi Arti~ domesticated         4         0.7       0.667  20
## # ... with 2 more variables: brainwt <dbl>, bodywt <dbl>
```

```
str(msleep) #Lists all variables and the type of variable
```

```
## tibble [83 x 11] (S3: tbl_df/tbl/data.frame)
##  $ name        : chr [1:83] "Cheetah" "Owl monkey" "Mountain beaver" "Greater short-tailed shrew" ..
##  $ genus       : chr [1:83] "Acinonyx" "Aotus" "Aplodontia" "Blarina" ...
##  $ vore        : chr [1:83] "carni" "omni" "herbi" "omni" ...
##  $ order       : chr [1:83] "Carnivora" "Primates" "Rodentia" "Soricomorpha" ...
##  $ conservation: chr [1:83] "lc" NA "nt" "lc" ...
##  $ sleep_total : num [1:83] 12.1 17 14.4 14.9 4 14.4 8.7 7 10.1 3 ...
##  $ sleep_rem   : num [1:83] NA 1.8 2.4 2.3 0.7 2.2 1.4 NA 2.9 NA ...
##  $ sleep_cycle : num [1:83] NA NA NA 0.133 0.667 ...
##  $ awake       : num [1:83] 11.9 7 9.6 9.1 20 9.6 15.3 17 13.9 21 ...
##  $ brainwt     : num [1:83] NA 0.0155 NA 0.00029 0.423 NA NA NA 0.07 0.0982 ...
##  $ bodywt      : num [1:83] 50 0.48 1.35 0.019 600 ...
```

```
summary(msleep) #Provides summary statistics for all variables in dataset
```

```
##      name              genus               vore              order          
##  Length:83          Length:83          Length:83          Length:83         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  conservation        sleep_total      sleep_rem      sleep_cycle    
##  Length:83          Min.   : 1.90   Min.   :0.100   Min.   :0.1167  
##  Class :character   1st Qu.: 7.85   1st Qu.:0.900   1st Qu.:0.1833  
##  Mode  :character   Median :10.10   Median :1.500   Median :0.3333  
##                     Mean   :10.43   Mean   :1.875   Mean   :0.4396  
##                     3rd Qu.:13.75   3rd Qu.:2.400   3rd Qu.:0.5792  
##                     Max.   :19.90   Max.   :6.600   Max.   :1.5000  
##                                     NA's   :22      NA's   :51      
##      awake          brainwt            bodywt        
##  Min.   : 4.10   Min.   :0.00014   Min.   :   0.005  
##  1st Qu.:10.25   1st Qu.:0.00290   1st Qu.:   0.174  
##  Median :13.90   Median :0.01240   Median :   1.670  
##  Mean   :13.57   Mean   :0.28158   Mean   : 166.136  
##  3rd Qu.:16.15   3rd Qu.:0.12550   3rd Qu.:  41.750  
##  Max.   :22.10   Max.   :5.71200   Max.   :6654.000  
##                  NA's   :27                          
```

```
summary(msleep$awake) #Provides summary statistics for the awake variable in dataset msleep
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.10   10.25   13.90   13.57   16.15   22.10 
```

```
dim(msleep) #Outputs a Vector Giving the Number of Rows and Columns
```

```
## [1] 83 11
```

```
unique(msleep$vore) #Lists all the unique values for a categorical variable Animals are Classified as C
```

```
## [1] "carni"   "omni"    "herbi"   NA        "insecti"
```

```
which(is.na(msleep$vore)) #Returns the Observation index where missing values exist
```

```
## [1]  8 55 57 58 63 69 73
```

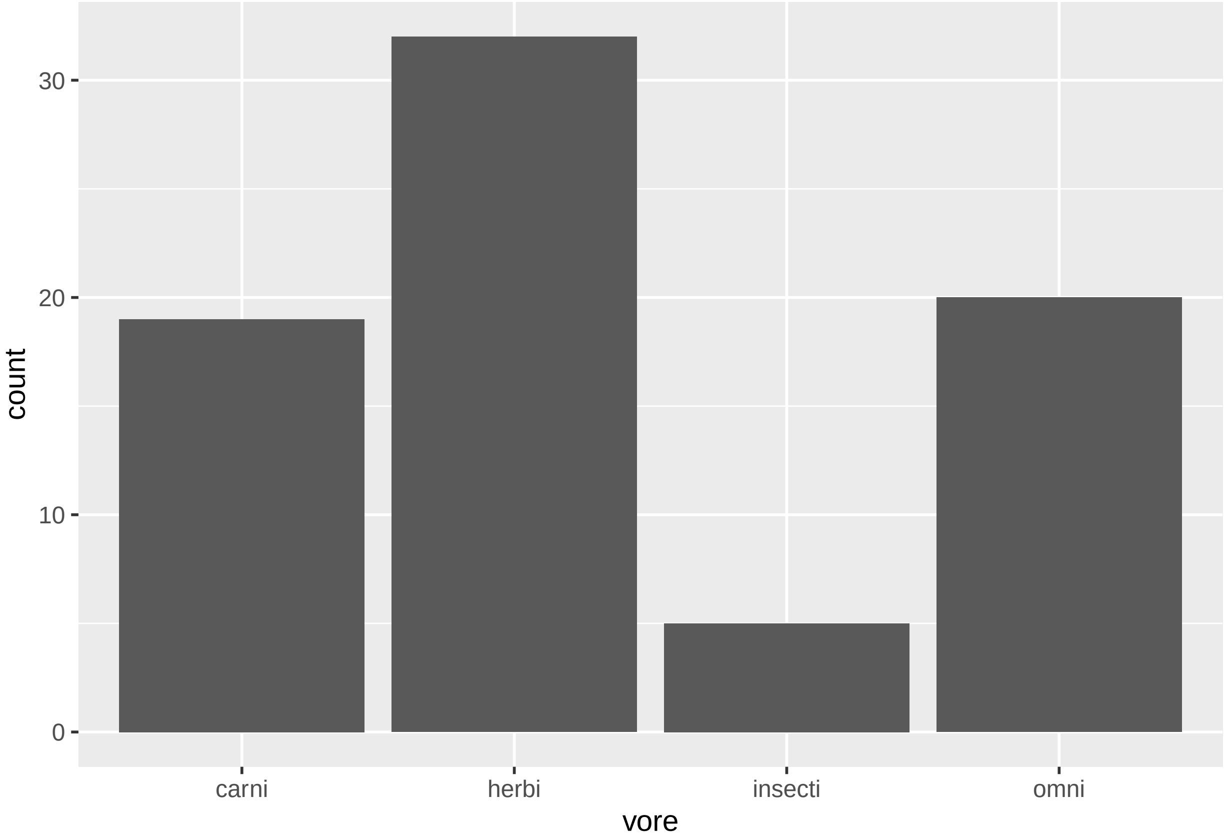
```
msleep2=msleep[-which(is.na(msleep$vore)),] #Removes the 7 Observations that are missing a vore-specifi
```

In this dataset, there are 83 observations and 11 variables.

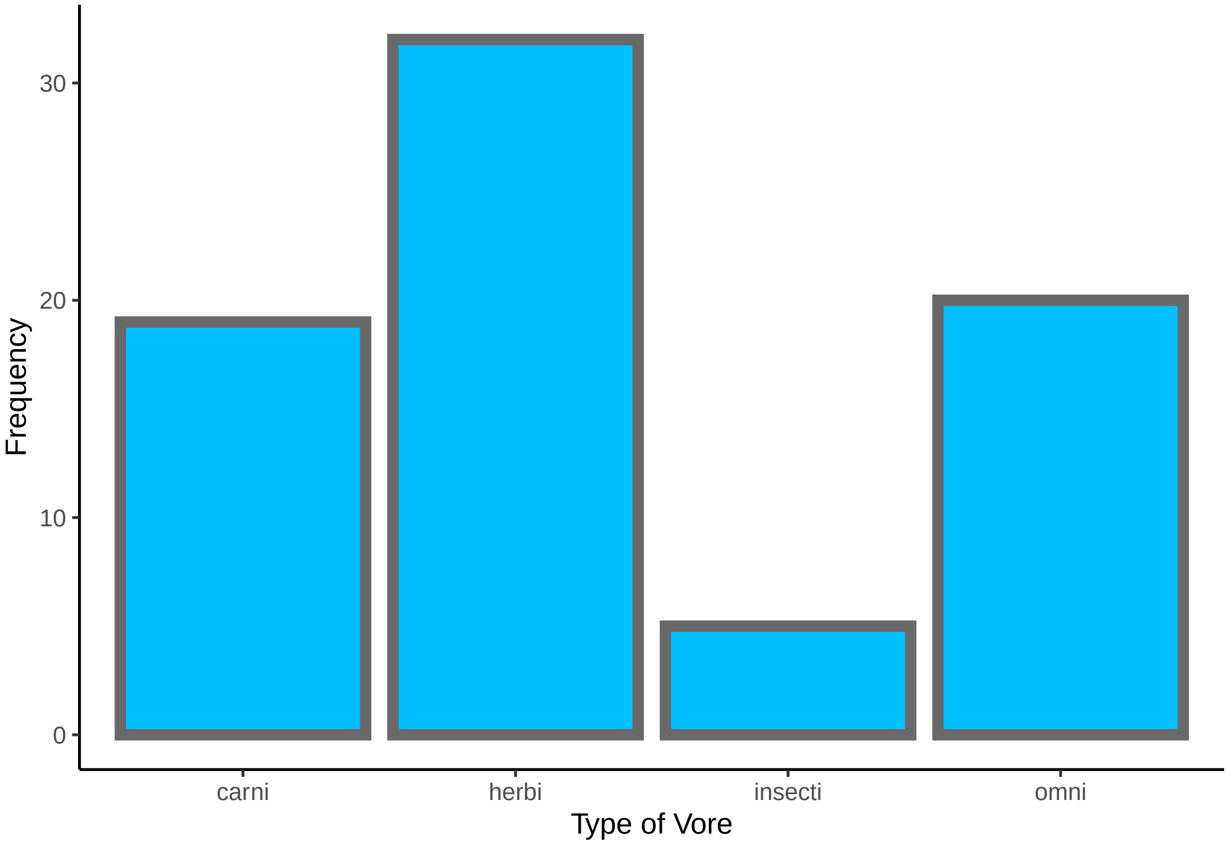
ggplot Discovery

##Barplot Examples

```
ggplot(data=msleep2) +
  geom_bar(aes(x=vore))
```



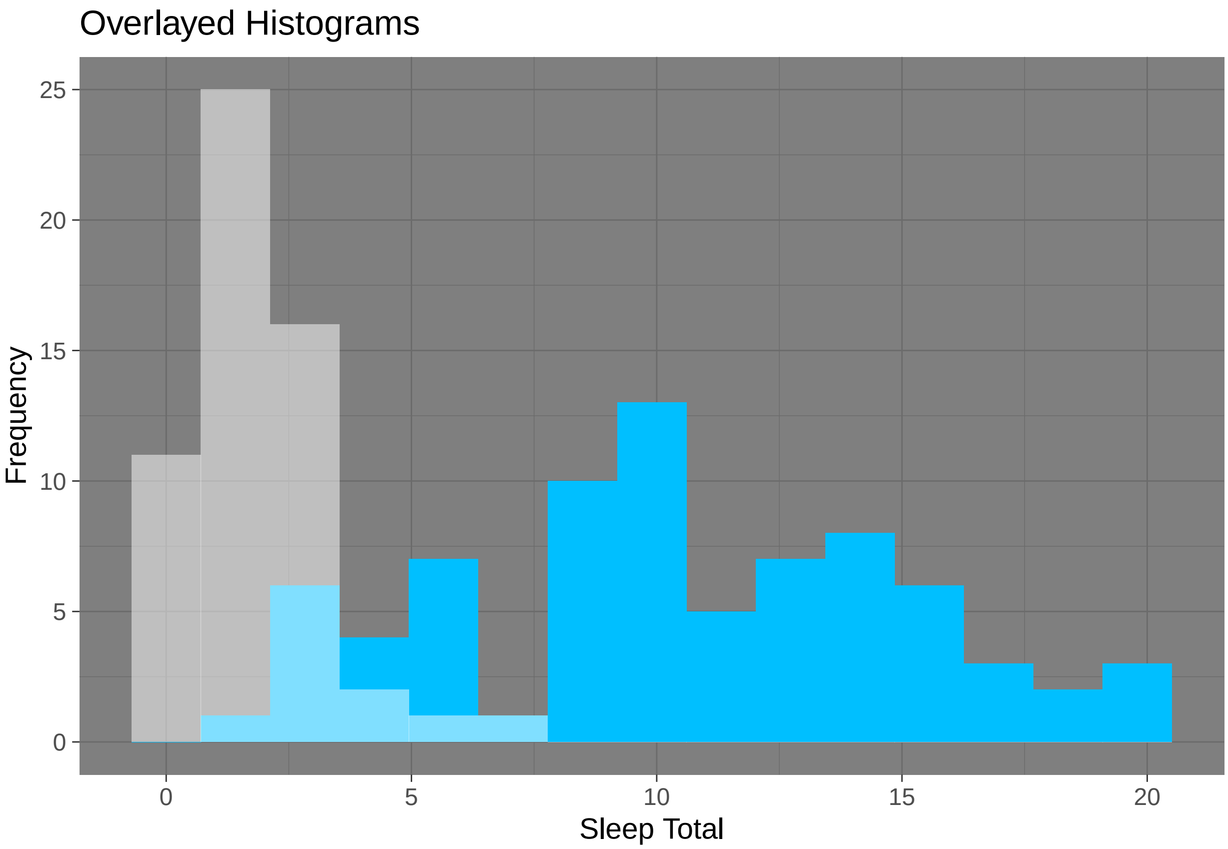
```
ggplot(data=msleep2) +
  geom_bar(aes(x=vore),color="dimgrey",fill="deepskyblue1",size=2) +
  xlab("Type of Vore") + ylab("Frequency") +
  theme_classic()
```



Histogram and Boxplot Examples

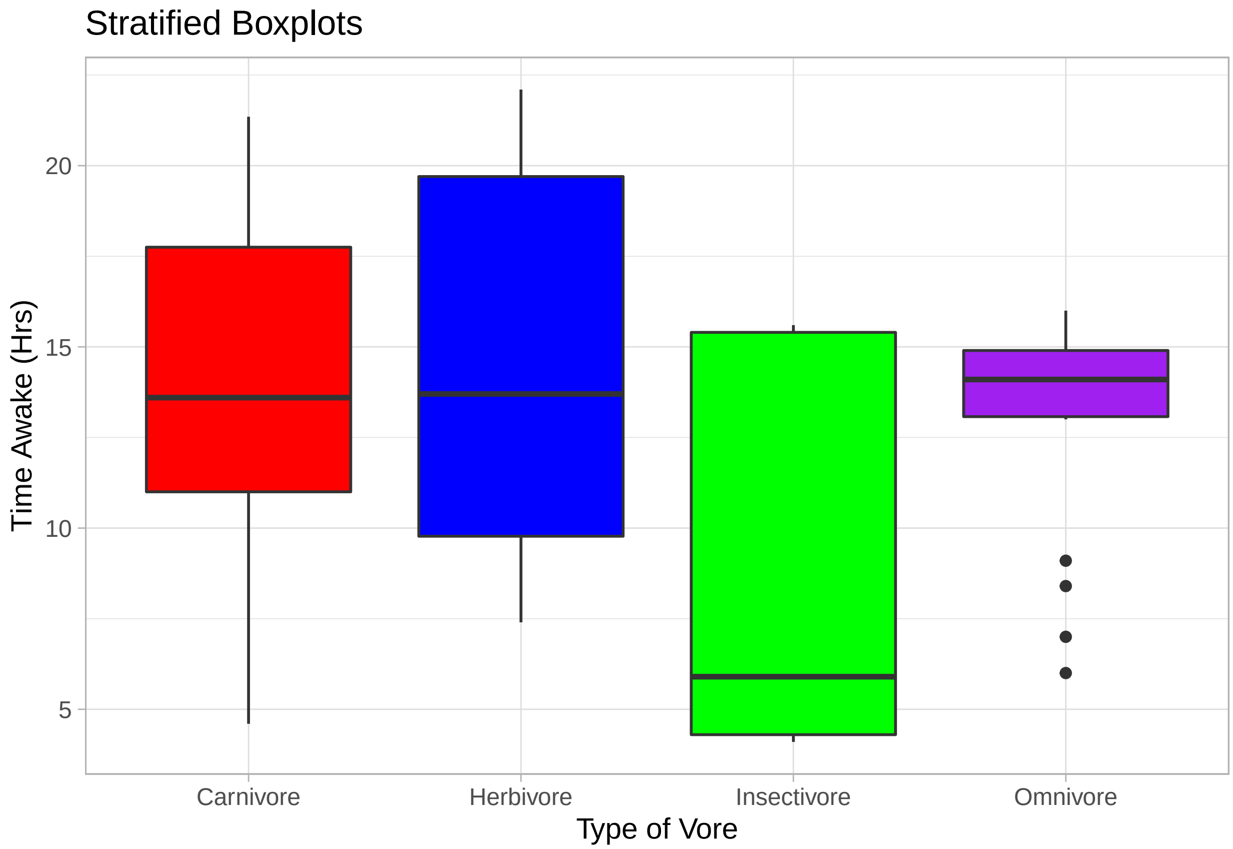
```
ggplot(data=msleep2) +
  geom_histogram(mapping=aes(x=sleep_total),bins=15,fill="deepskyblue1") +
  geom_histogram(mapping=aes(x=sleep_rem),bins=15,fill="white",alpha=0.5) +
  labs(x="Sleep Total",y="Frequency",title="Overlayed Histograms") + theme_dark()
```

```
## Warning: Removed 20 rows containing non-finite values (stat_bin).
```



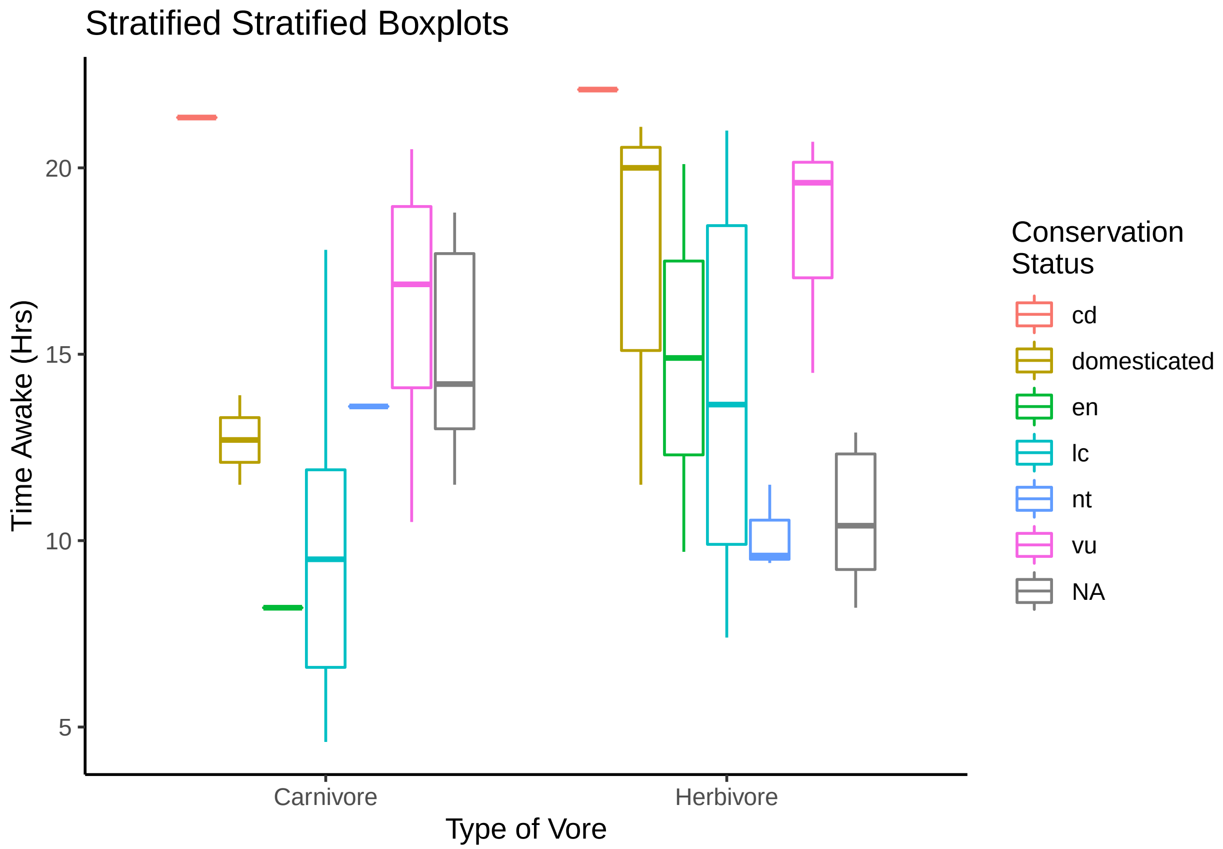
```
#Warning due to NA
```

```
ggplot(data=msleep2) +
  geom_boxplot(aes(x=vore,y=awake),fill=c("red","blue","green","purple")) +
  xlab("Type of Vore") + ylab("Time Awake (Hrs)") +
  theme_light()+ggtitle("Stratified Boxplots") +
  scale_x_discrete(labels=c("Carnivore","Herbivore","Insectivore","Omnivore"))
```



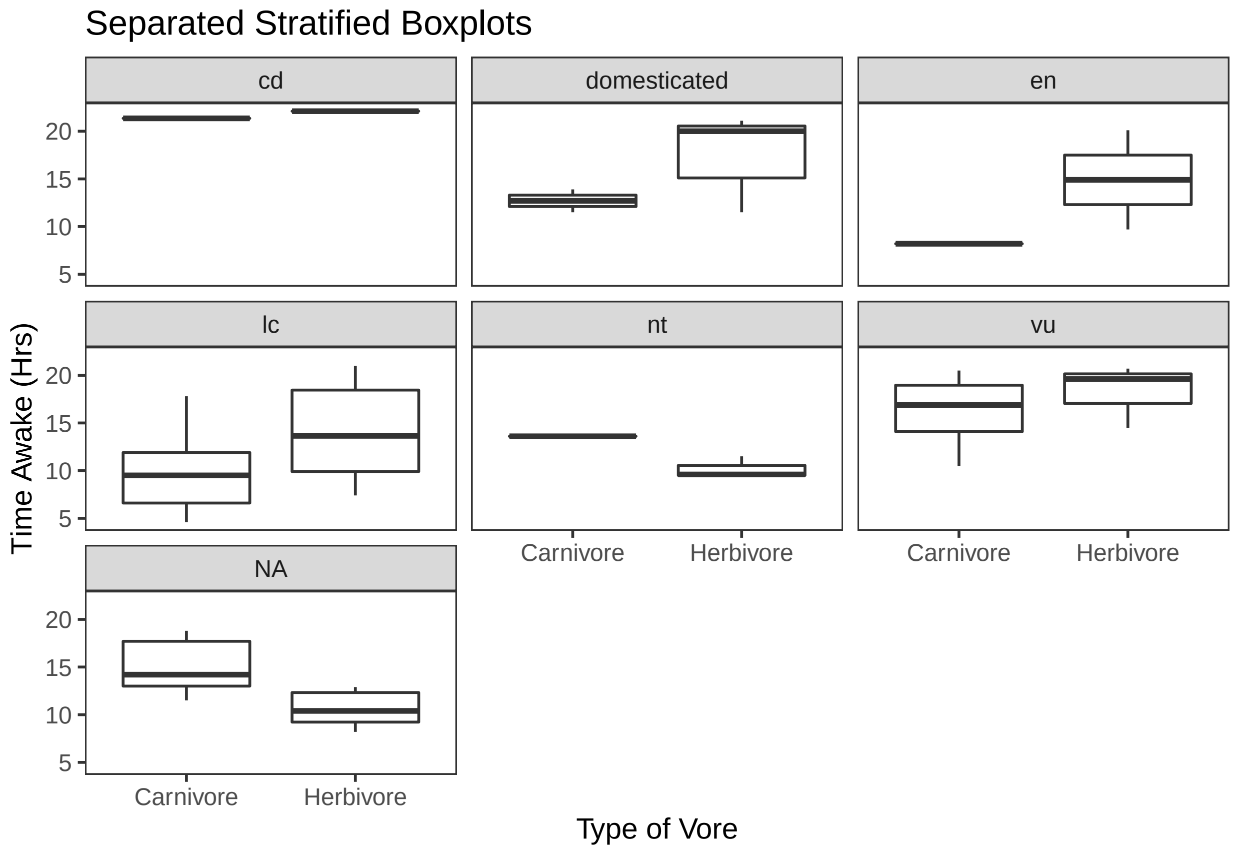
```
ggplot(data=msleep2) +
  geom_boxplot(aes(x=vore,y=awake,color=conservation)) +
  xlab("Type of Vore") + ylab("Time Awake (Hrs)") +
  theme_light()+ggtitle("Stratified Stratified Boxplots") +
  scale_x_discrete(limits=c("carni","herbi"),labels=c("Carnivore","Herbivore")) +
  guides(color=guide_legend(title="Conservation \nStatus")) +
  theme_classic()
```

```
## Warning: Removed 25 rows containing missing values (stat_boxplot).
```



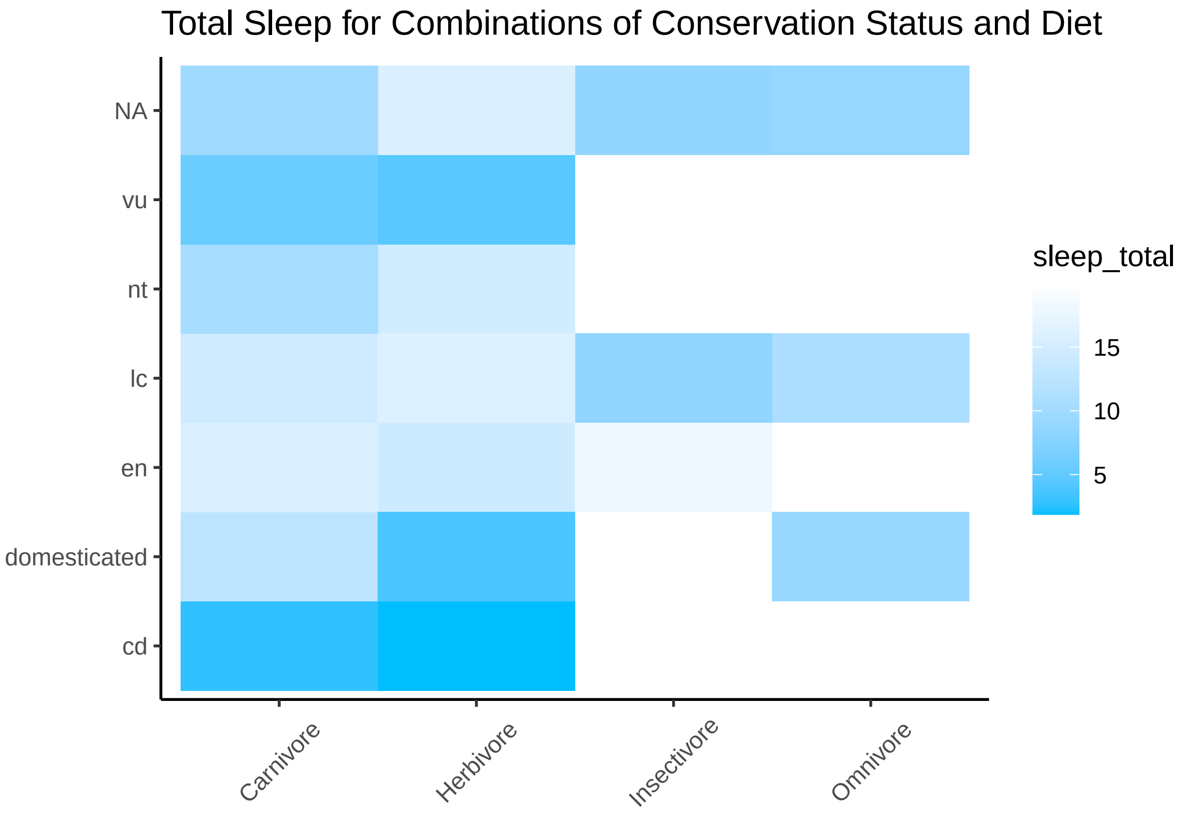
```
ggplot(data=msleep2) +
  geom_boxplot(aes(x=vore,y=awake)) +
  facet_wrap(conservation~.) +
  xlab("Type of Vore") + ylab("Time Awake (Hrs)") +
  theme_light()+ggtitle("Separated Stratified Boxplots") +
  scale_x_discrete(limits=c("carni","herbi"),labels=c("Carnivore","Herbivore")) +
  theme_test()
```

```
## Warning: Removed 25 rows containing missing values (stat_boxplot).
```



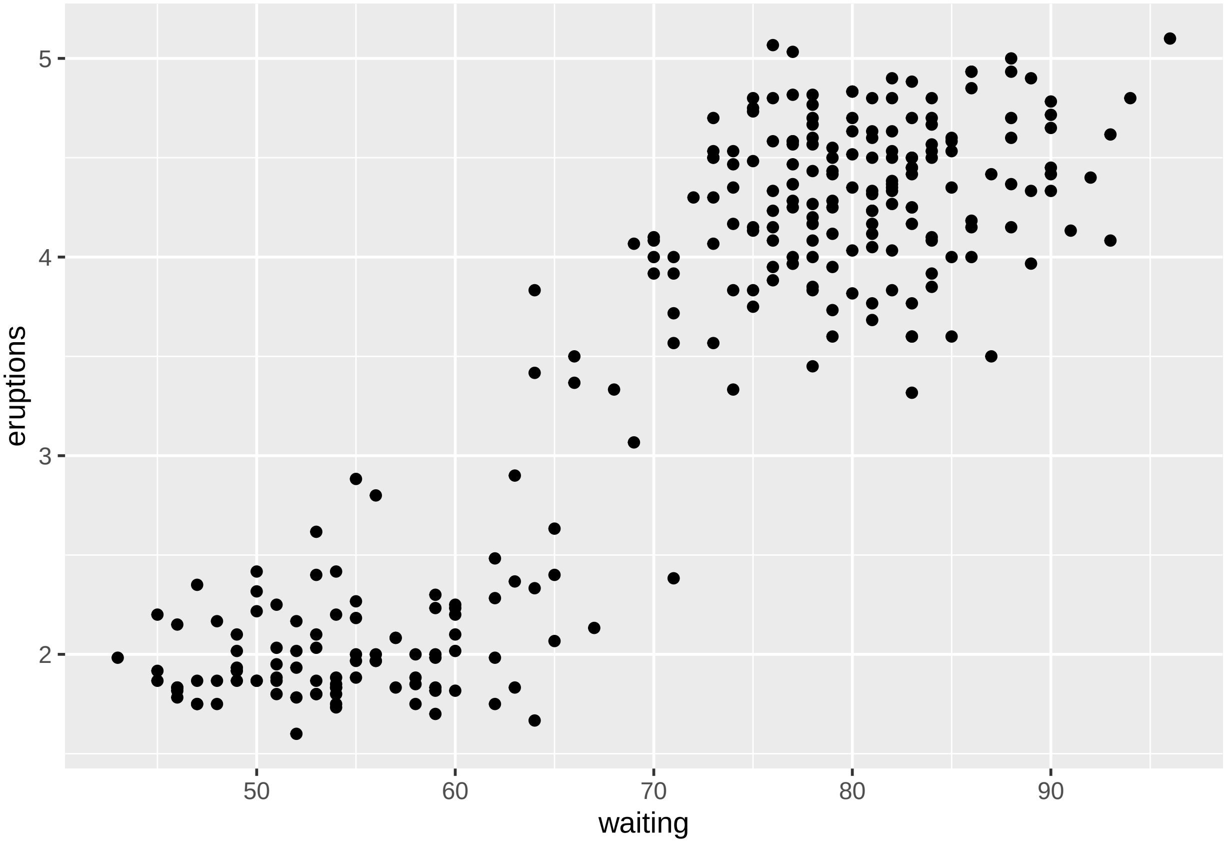
Heatmap Examples (Triple Variables)

```
ggplot(data=msleep2,aes(x=vore,y=conservation)) +
  geom_tile(aes(fill=sleep_total)) +
  scale_fill_gradient(low="deepskyblue1",high="white")+
  theme_classic() +
  scale_x_discrete(label=c("Carnivore","Herbivore","Insectivore","Omnivore")) +
  theme(axis.text.x=element_text(angle=45,vjust=0.5))+
  xlab("")+ylab("") +
  ggtitle("Total Sleep for Combinations of Conservation Status and Diet")
```

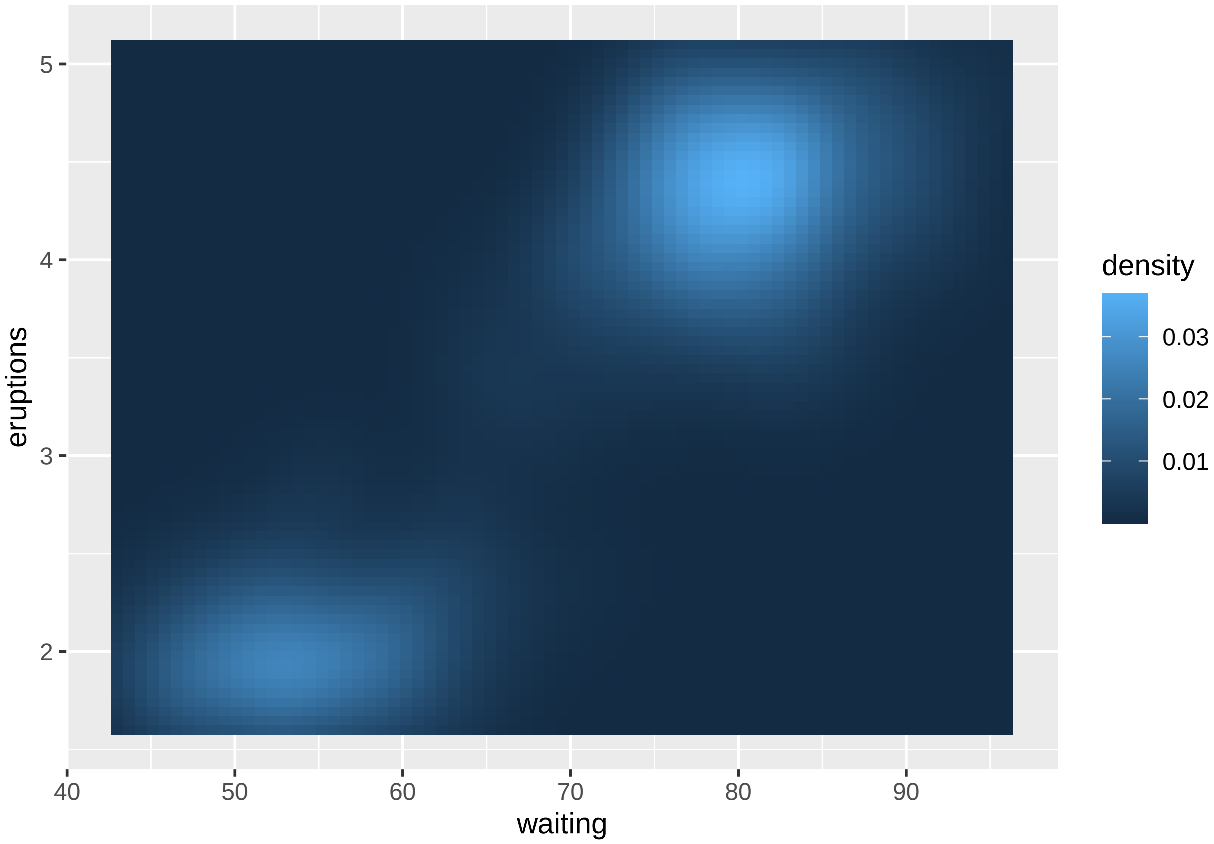


The next example can be found at https://ggplot2.tidyverse.org/reference/scale_brewer.html. These examples are based on the classic Old Faithful data set. The data set provides the joint probability distribution of waiting time between eruptions and the duration of the eruptions. The original data set `faithful`contains sample data from monitoring the famous geyser Old Faithful. The data set `faithfuld` from **ggplot2** provides emperical joint density estimates for relationship between these two variables.

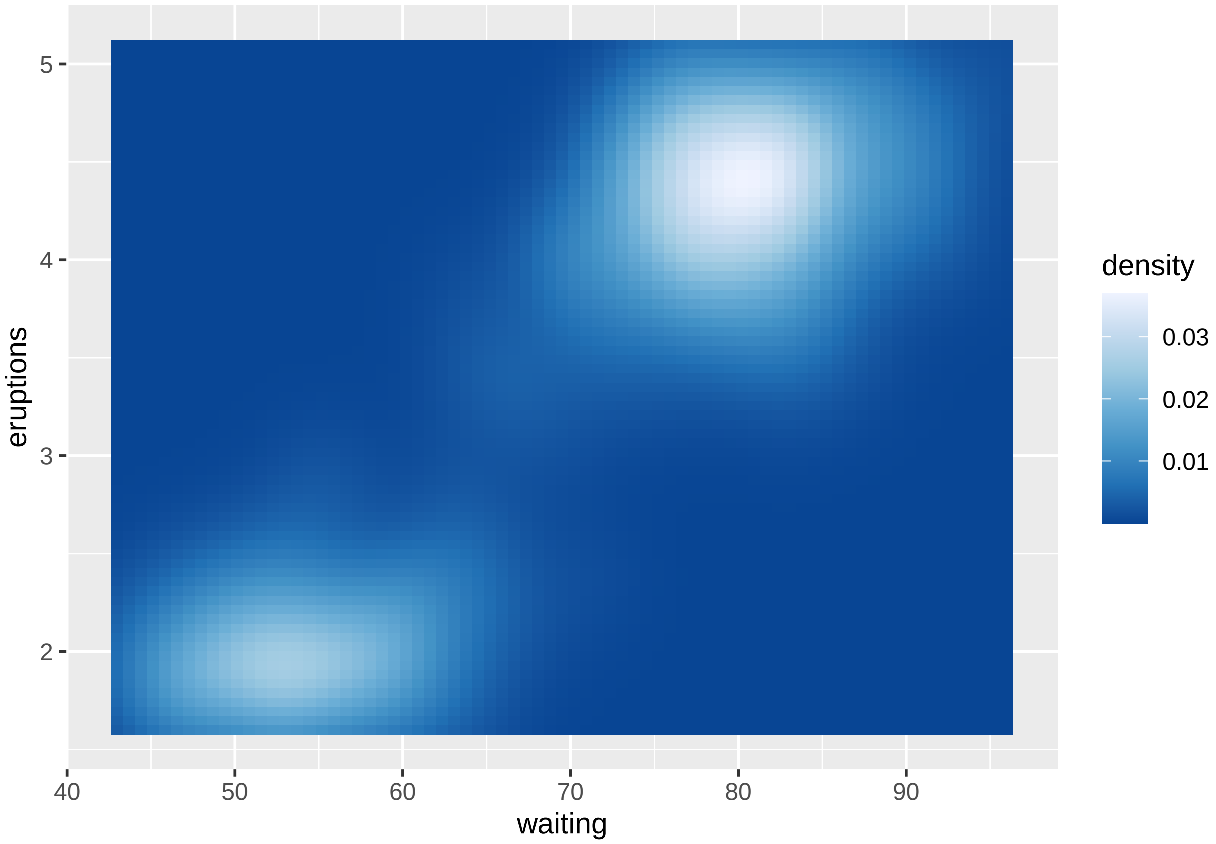
```
#First Notice from Original Old Faithful Data Sets
ggplot(faithful) +
  geom_point(aes(x=waiting,y=eruptions),col="black")
```



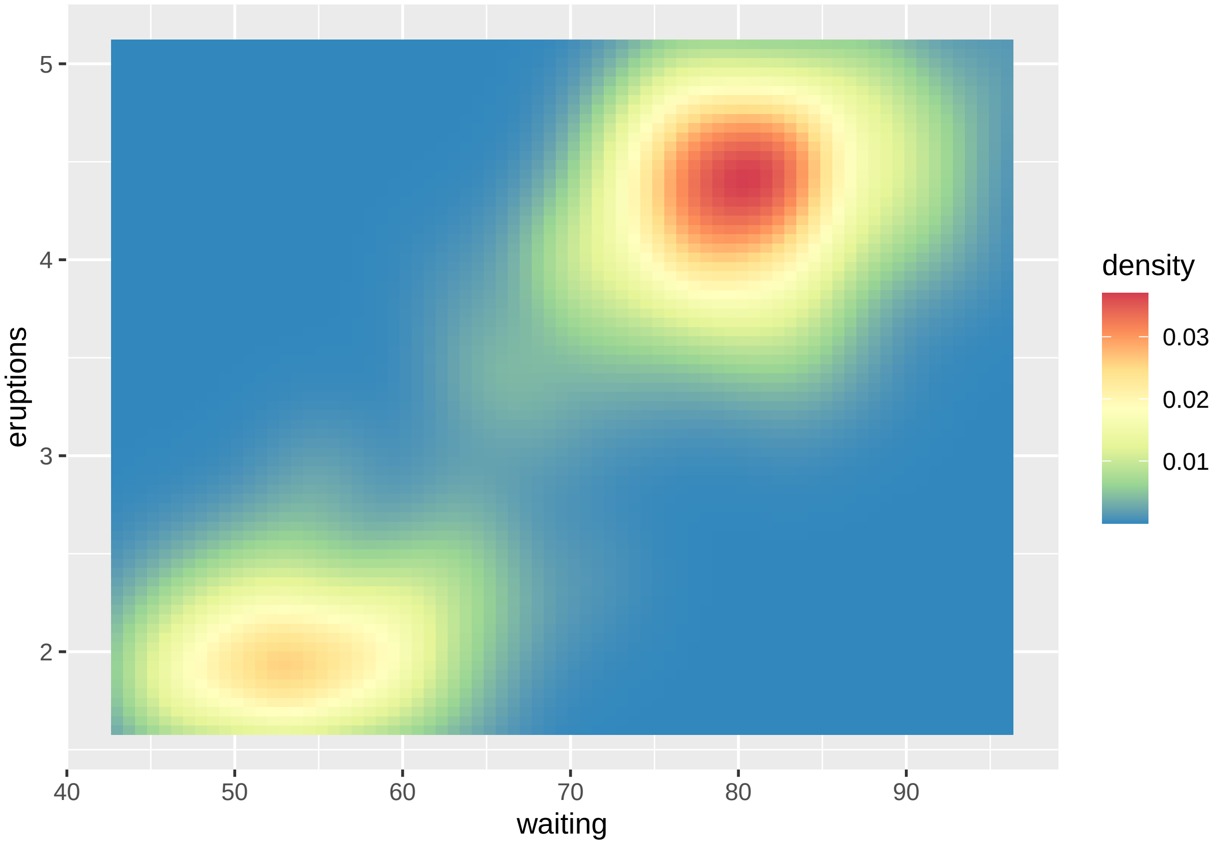
```
#Now we Construct a Heatmap Showing the
v <- ggplot(faithfuld) +
  geom_tile(aes(waiting, eruptions, fill = density))
v
```



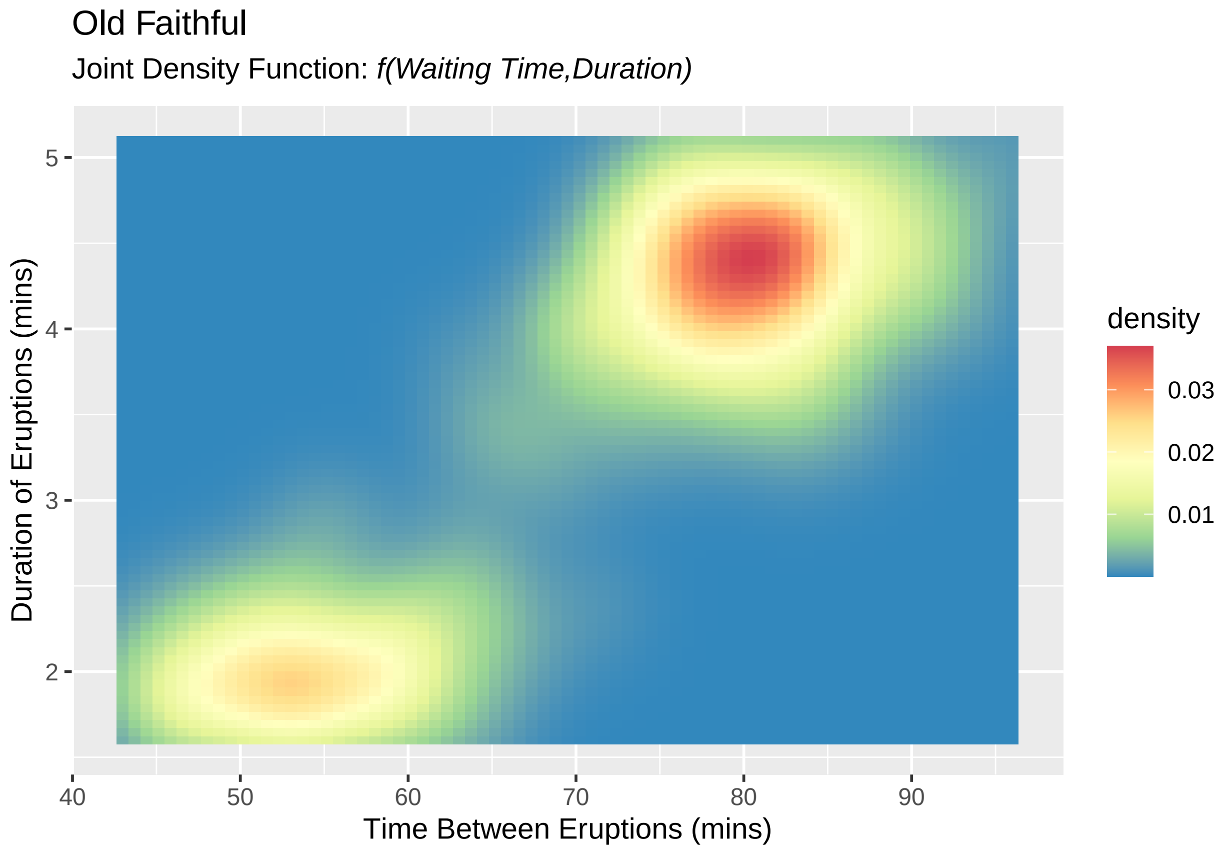
```
v2=v + scale_fill_distiller()
v2
```



```
v3=v+scale_fill_distiller(palette = "Spectral")
v3
```



```
v4=v3 + xlab("Time Between Eruptions (mins)") + ylab("Duration of Eruptions (mins)") +
  ggtitle("Old Faithful") + labs(subtitle=expression(paste("Joint Density Function: ",italic("f(Waiting
v4
```


Old Faithful
Joint Density Function: *f(Waiting Time,Duration)*
density
0.03
0.02
0.01
5
4
3
2
Duration of Eruptions (mins)
40
50
60
70
80
90
Time Between Eruptions (mins)