Analysis of Species

Scott White

April 8, 2019

```
library(here)
```

```
## Warning: package 'here' was built under R version 3.5.3
```

```
## here() starts at C:/Users/Scott/Dropbox/Masters/STAT 7350 - Visualization of Biological Data/STAT735
```

```
library(tidyverse)
```

```
## -- Attaching packages ----------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## -- Conflicts -------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 3.5.2
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 3.5.2
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 3.5.2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(ggmosaic)
```

```
## Warning: package 'ggmosaic' was built under R version 3.5.3
```

Read in data

```
bees <- read_csv(here("Assignments/Assignment3/data_output/cleaned_bees_columns_removed.csv"),
                 col_types = cols(Division = col_character()))
```

```
bees <- bees %>% mutate(Month = as_factor(Month))
bees <- bees %>% mutate(Month = fct_relevel(Month, levels = c("May", "Jun", "Jul", "Aug", "Sept")))
```

```
## Warning: Outer names are only allowed for unnamed scalar atomic inputs
```

```
bees <- bees %>% mutate(State = as_factor(State))
bees <- bees %>% mutate(State = fct_relevel(State, levels = c("Wisconsin", "Illinois", "Minnesota", "Iou
```

```
## Warning: Outer names are only allowed for unnamed scalar atomic inputs
```

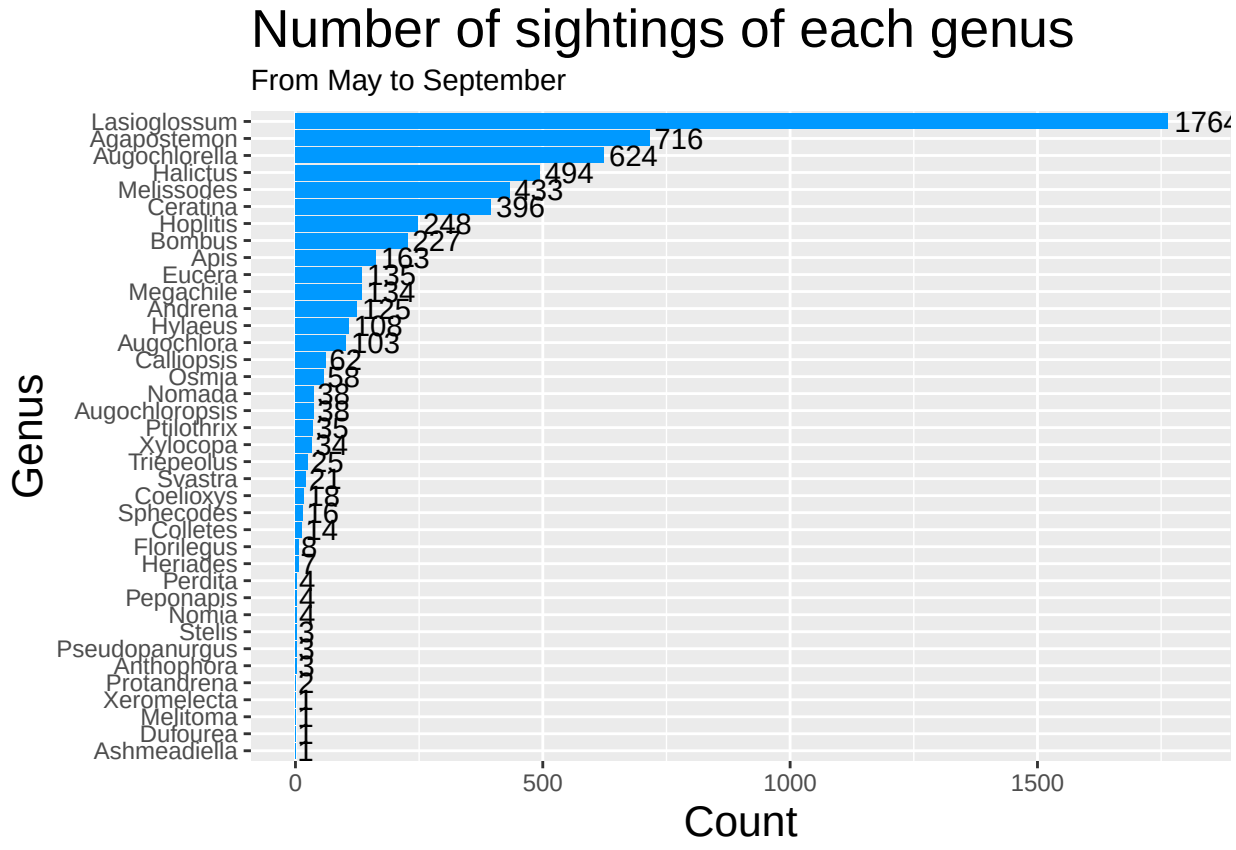
```
bees
```

```
## # A tibble: 6,071 x 12
##       ID Genus Gender Species State County Refuge Division `Sample Site`
##    <dbl> <chr> <chr>  <chr>   <fct> <chr>  <chr>  <chr>    <chr>
##  1     1 Agap~ Female serice~ Miss~ Lafay~ Big M~ <NA>     <NA>
##  2     2 Agap~ Female serice~ Miss~ Lafay~ Big M~ <NA>     BB
##  3     3 Agap~ Female serice~ Miss~ Lafay~ Big M~ <NA>     BB
##  4     4 Agap~ Female serice~ Miss~ Ray    Big M~ <NA>     BB
##  5     5 Agap~ Male   serice~ Miss~ Lafay~ Big M~ <NA>     BB
##  6     6 Agap~ Female serice~ Miss~ Ray    Big M~ <NA>     JAB
##  7     7 Agap~ Female serice~ Miss~ Ray    Big M~ <NA>     JAB
##  8     8 Agap~ Female serice~ Miss~ Ray    Big M~ <NA>     JAB
##  9     9 Agap~ Female serice~ Miss~ Ray    Big M~ <NA>     JAB
## 10    10 Agap~ Male   serice~ Miss~ Ray    Big M~ <NA>     JAB
## # ... with 6,061 more rows, and 3 more variables: Habitat <chr>,
## #   Collector <chr>, Month <fct>
```

```
apply(bees, 2, function(x) length(unique(x)))
```

```
##          ID       Genus      Gender     Species       State      County
##        6071          38           2         197           5          15
##      Refuge    Division Sample Site     Habitat   Collector       Month
##          15          40          78          30          23           6
```

```
genus_counts <- bees %>% group_by(Genus) %>% summarise(Count = n()) %>%
  ggplot(aes(reorder(Genus, Count), Count)) +
  geom_col(fill = "#0099FF")  +
  geom_text(aes(label = Count, hjust = -0.1)) +
  coord_flip() +
  ggtitle("Number of sightings of each genus", "From May to September") +
  xlab("Genus") +
  ylim(c(0, 1800)) +
  theme(axis.title = element_text(size = 16),
        plot.title = element_text(size = 20))
genus_counts
```

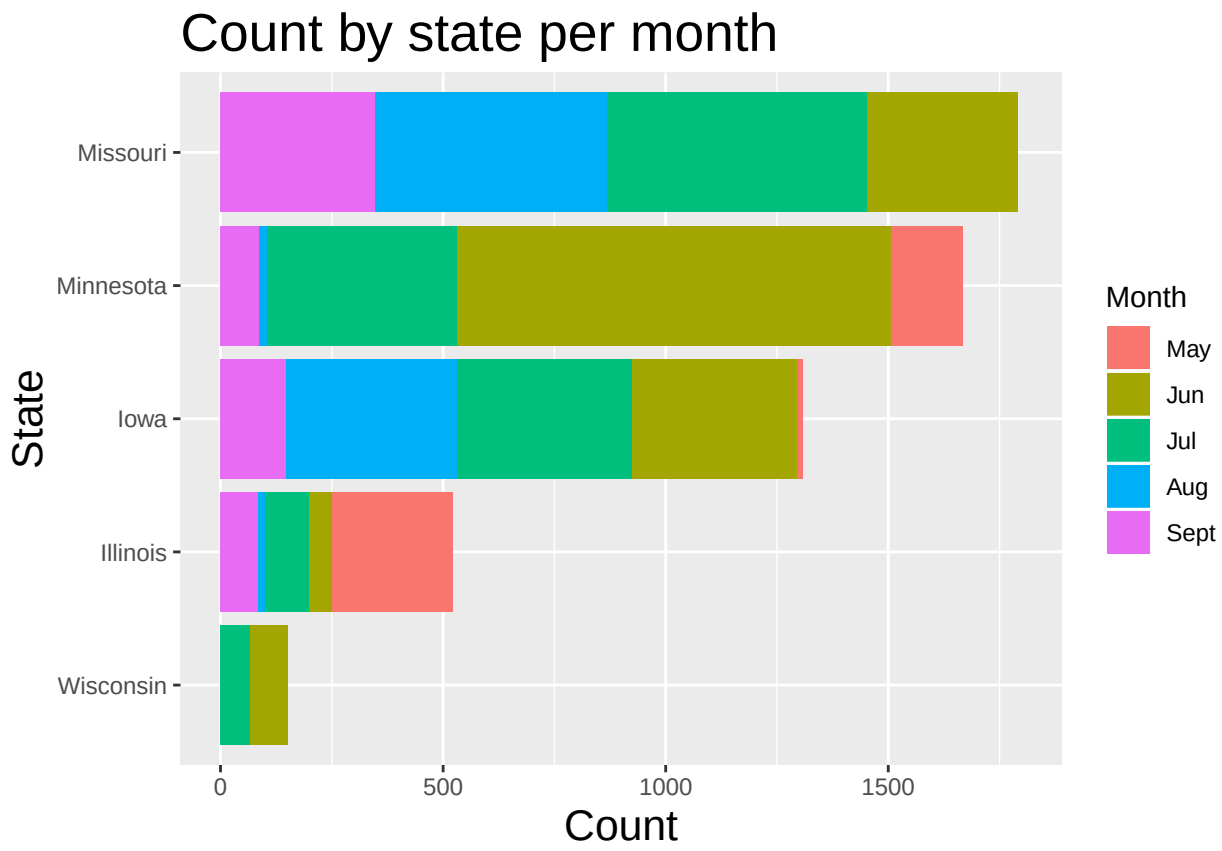



```
# ggsave("genus_counts.png", genus_counts)
```

```
bees %>% group_by(State, Month, Genus) %>% summarise(Count = n()) %>% drop_na() %>%
  ggplot(aes(reorder(State, Count, FUN = function(x)sum(x)), Count)) +
  geom_col(aes(fill = Month))  +
  # geom_text(aes(label = Count, hjust = -0.1)) +
  coord_flip() +
  ggtitle("Count by state per month") +
  xlab("State") +
  ylim(c(0, 1800)) +
  theme(axis.title = element_text(size = 16),
        plot.title = element_text(size = 20))
```

```
## Warning: Factor `Month` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## Warning: Removed 30 rows containing missing values (geom_col).
```

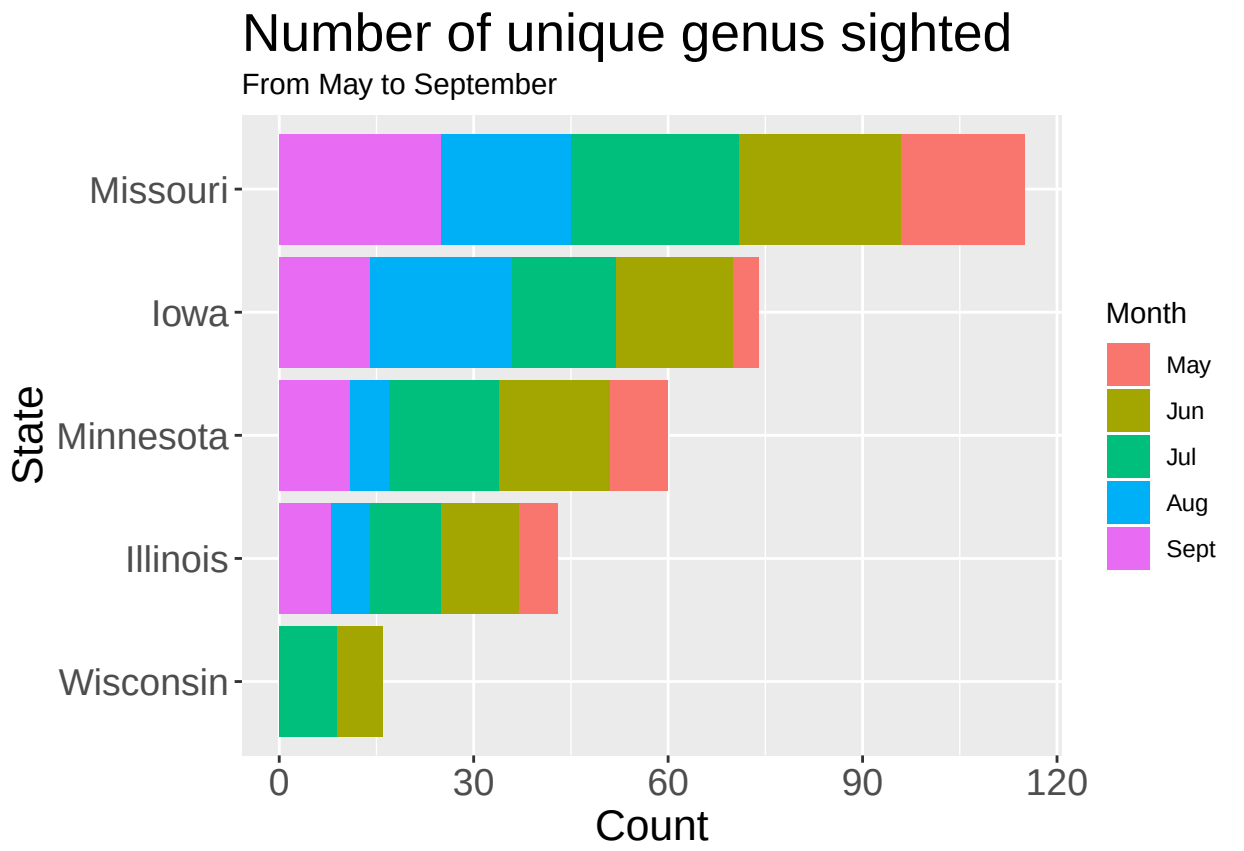


Question: Is the distribution of the number of unique sightings equal over month and state?

```
genus_state_month <- bees %>% group_by(Genus, State, Month) %>% summarise(Unique = n_distinct(Genus)) %
  drop_na()
```

```
## Warning: Factor `Month` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

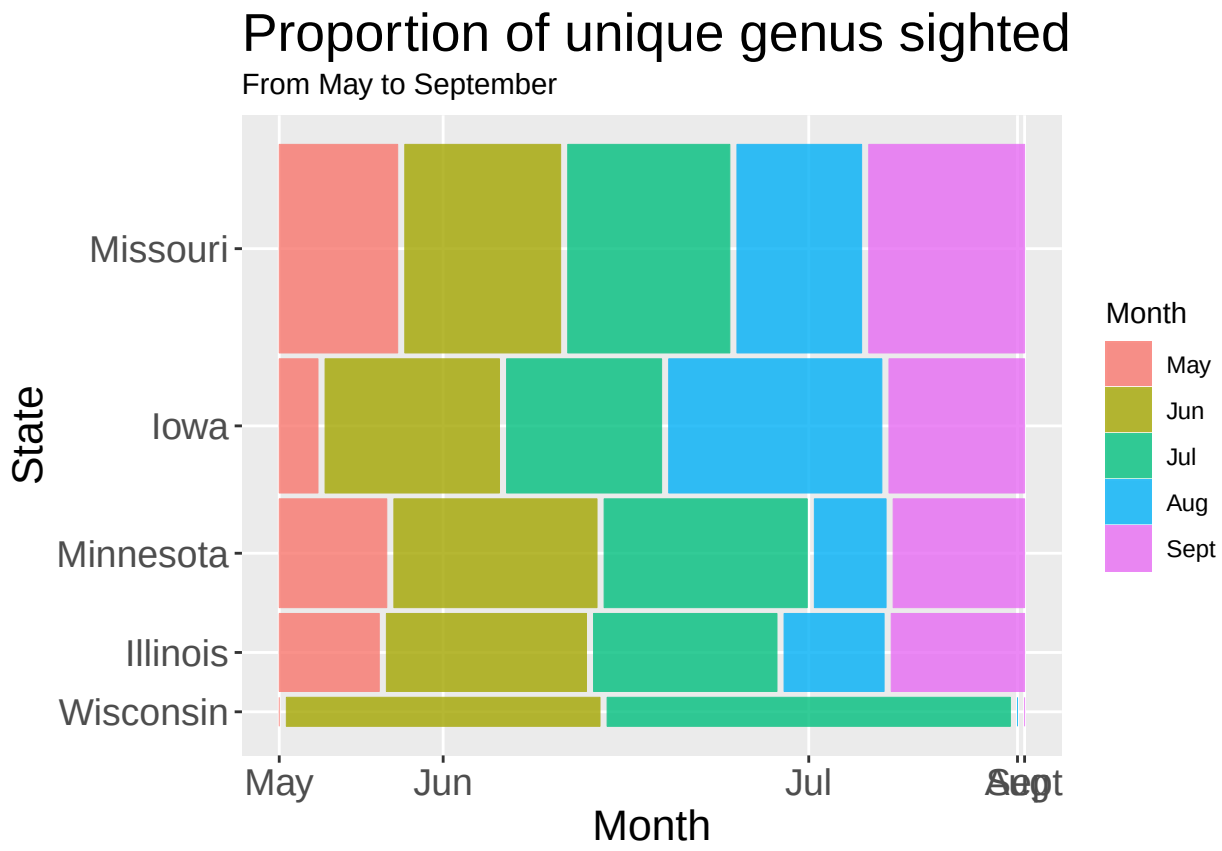
```
genus_state_month_barplot <- genus_state_month %>%
  ggplot(aes(reorder(State, Unique, FUN = function(x)sum(x)), Unique)) +
  geom_col(aes(fill = Month)) +
  ggtitle("Number of unique genus sighted",
          subtitle = "From May to September") +
  coord_flip() +
  xlab("State") +
  ylab("Count") +
  theme(axis.title = element_text(size = 16),
        plot.title = element_text(size = 20),
        axis.text = element_text(size = 14))
genus_state_month_barplot
```




```
# ggsave("genus_state_month_barplot.png", plot = genus_state_month_barplot)

# ggsave("genus_state_month_barplot.png", plot = genus_state_month_barplot,
#        path = here("Assignments", "Assignment3", "figures/genus_state_month_barplot.png"))
```

```
# genus_state_month <- bees %>% group_by(State, Month) %>% summarise(Unique = n_distinct(Genus)) %>%
#   drop_na()
genus_state_month_mosaicplot <-genus_state_month %>%
  ggplot() +
  geom_mosaic(aes(weight = Unique, x = product(State), fill = Month)) +
  ggtitle("Proportion of unique genus sighted",
          subtitle = "From May to September") +
  xlab("State") +
  ylab("Month") +
  coord_flip() +
  theme(axis.title = element_text(size = 16),
        plot.title = element_text(size = 20),
        axis.text = element_text(size = 14))
genus_state_month_mosaicplot
```




```
# ggsave("genus_state_month_mosaicplot.png", plot = genus_state_month_mosaicplot)
```

```
ct_gsm<- genus_state_month %>% group_by(State, Month) %>% summarise(Unique = n_distinct(Genus)) %>%
  spread(key = Month, value = Unique)
ct_gsm
```

```
## # A tibble: 5 x 6
## # Groups:   State [5]
##   State       May   Jun   Jul   Aug  Sept
##   <fct>     <int> <int> <int> <int> <int>
## 1 Wisconsin     0     7     9     0     0
## 2 Illinois      6    12    11     6     8
## 3 Minnesota     9    17    17     6    11
## 4 Iowa          4    18    16    22    14
## 5 Missouri     19    25    26    20    25
```

```
ungroup(ct_gsm) %>%
  select(-State) %>% chisq.test()
```

```
## Warning in chisq.test(.): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 31.731, df = 16, p-value = 0.01084
```

```
ungroup(ct_gsm) %>%
filter(State != c("Wisconsin"))%>%
  select(-State) %>% chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 14.575, df = 12, p-value = 0.2655
```

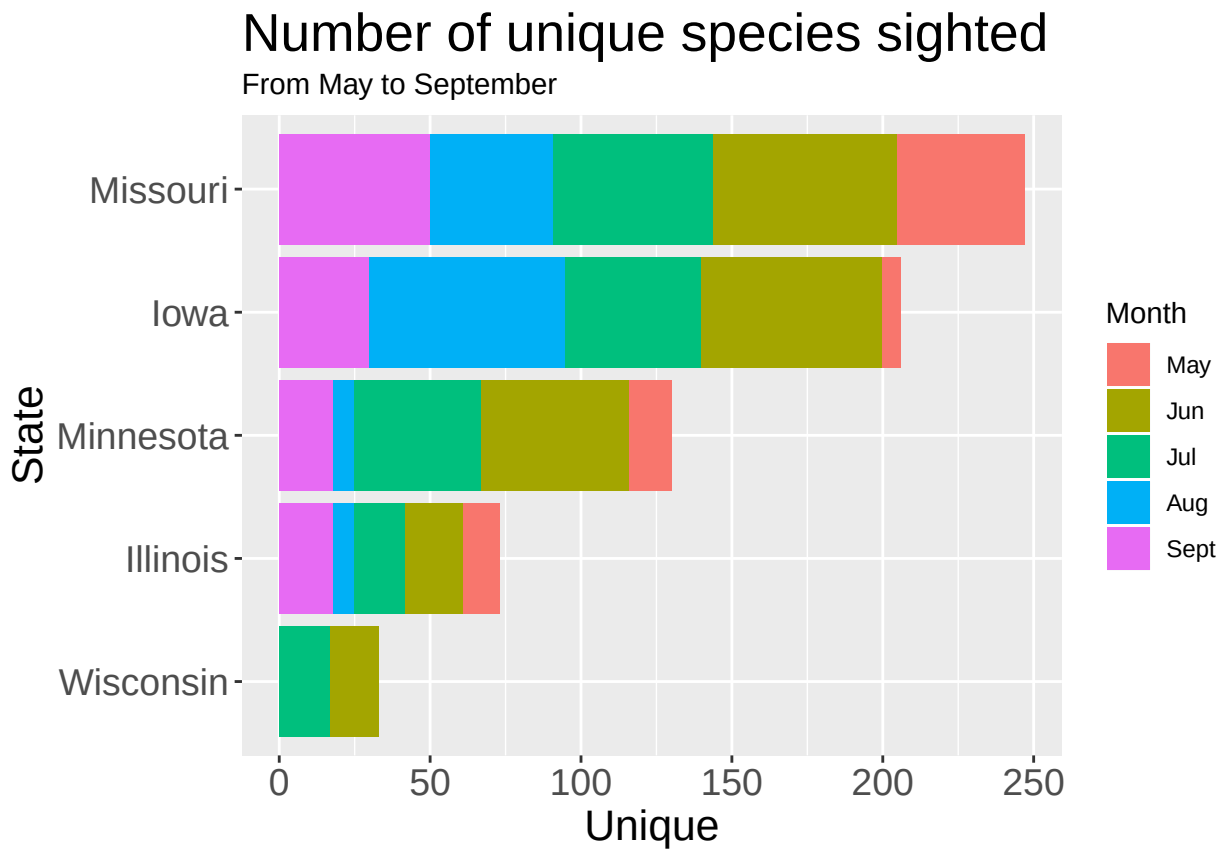
```
ungroup(ct_gsm) %>%
filter(State %nin% c("Wisconsin", "Illinois")) %>%
  select(-State) %>% chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 13.447, df = 8, p-value = 0.09736
```

```
species_state_month <- bees %>% group_by(Species, State, Month) %>% summarise(Unique = n_distinct(Speci
  drop_na()
```

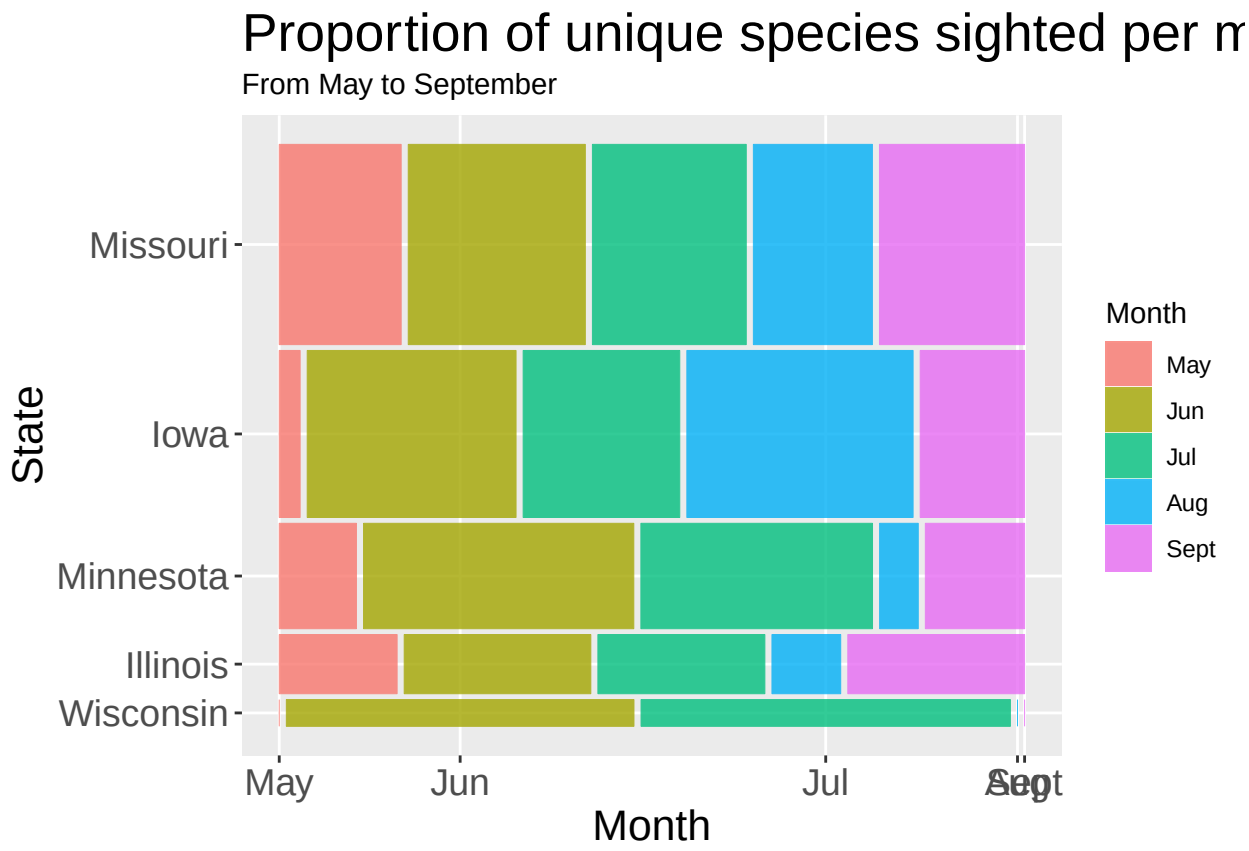
```
## Warning: Factor `Month` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
species_state_month_bargraph <- species_state_month %>%
  ggplot(aes(reorder(State, Unique, FUN = function(x)sum(x)), Unique)) +
  geom_col(aes(fill = Month)) +
  coord_flip() +
  ggtitle("Number of unique species sighted",
          "From May to September") +
  xlab("State") +
  theme(axis.title = element_text(size = 16),
        plot.title = element_text(size = 20),
        axis.text = element_text(size = 14))
species_state_month_bargraph
```




```
# ggsave("species_state_month_bargraph.png", species_state_month_bargraph)
```

```
species_state_month_mosaicplot <- species_state_month %>%
  ggplot() +
  geom_mosaic(aes(weight = Unique, x = product( State  )   , fill = Month)) +
  coord_flip() +
  ggtitle("Proportion of unique species sighted per month",
          "From May to September") +
  xlab("State") +
  ylab("Month") +
  theme(axis.title = element_text(size = 16),
        plot.title = element_text(size = 20),
        axis.text = element_text(size = 14))
species_state_month_mosaicplot
```




```
# ggsave("species_state_month_mosaicplot.png", species_state_month_mosaicplot)
```

```
ct_ssm<- species_state_month %>% group_by(State, Month) %>% summarise(Unique = n_distinct(Species)) %>%
  spread(key = Month, value = Unique)
ct_ssm
```

```
## # A tibble: 5 x 6
## # Groups:   State [5]
##   State       May   Jun   Jul   Aug  Sept
##   <fct>     <int> <int> <int> <int> <int>
## 1 Wisconsin     0    16    17     0     0
## 2 Illinois     12    19    17     7    18
## 3 Minnesota    14    49    42     7    18
## 4 Iowa          6    60    45    65    30
## 5 Missouri     42    61    53    41    50
```

```
ungroup(ct_ssm) %>%
  select(-State) %>% chisq.test()
```

```
## Warning in chisq.test(.): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 103.79, df = 16, p-value = 6.709e-15
```

```r
ungroup(ct_ssm) %>%
filter(State != c("Wisconsin"))%>%
  select(-State) %>% chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 71.95, df = 12, p-value = 1.38e-10
```

```r
ungroup(ct_ssm) %>%
filter(State %nin% c("Wisconsin", "Illinois")) %>%
  select(-State) %>% chisq.test()
```

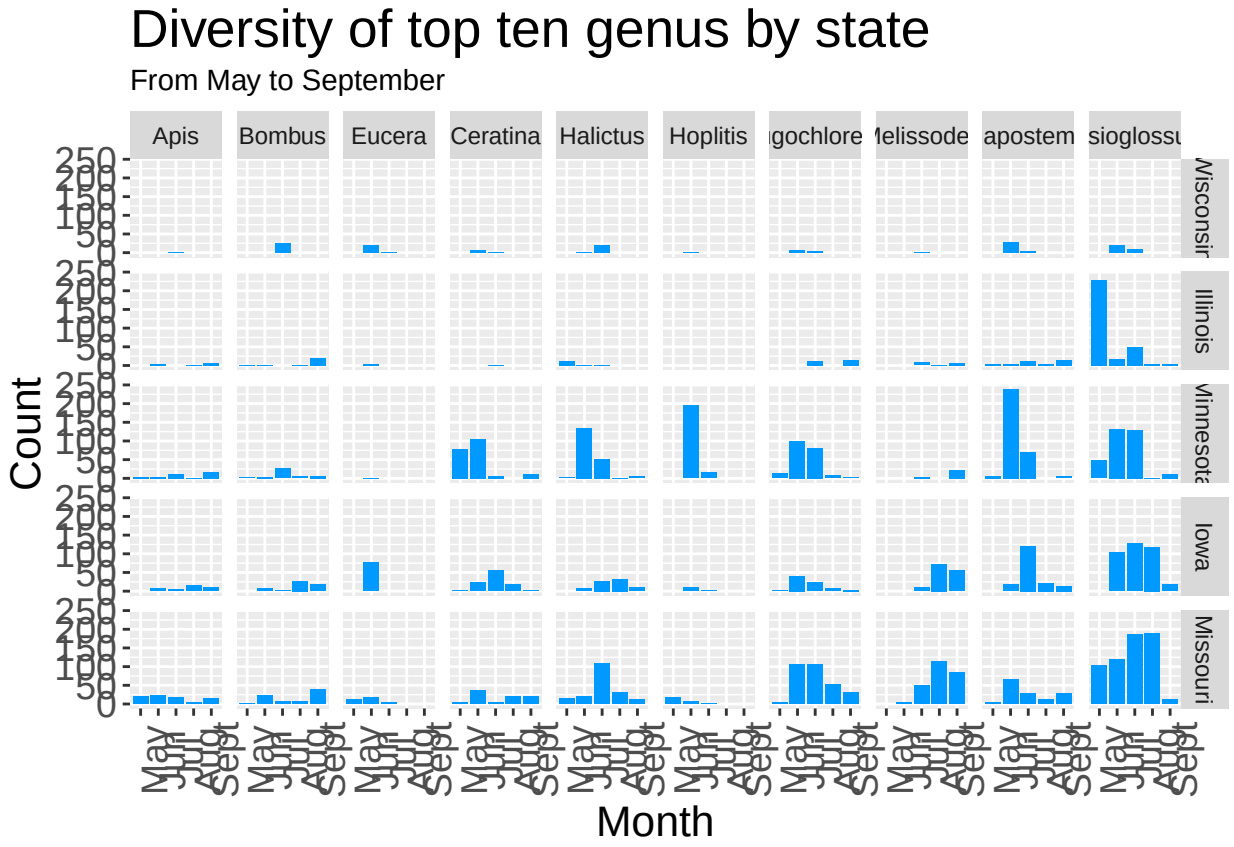
```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 63.611, df = 8, p-value = 9.078e-11
```

```r
top10_genus <- bees %>% group_by(Genus) %>% summarise(Count = n()) %>% top_n(10, Count) %>% arrange(des

state_month_genus_count <- bees %>% group_by(State, Month, Genus) %>% summarise(Count = n())
```

```
## Warning: Factor `Month` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```r
state_month_genus_count_facetplot <- state_month_genus_count %>% drop_na %>% filter(Genus %in% top10_ge
  ggplot(aes(Month, Count)) +
  geom_col(fill = "#0099FF") +
  ggtitle("Diversity of top ten genus by state",
          "From May to September")+
  facet_grid(State~reorder(Genus, Count)) +
  theme(axis.title = element_text(size = 16),
        plot.title = element_text(size = 20),
        axis.text = element_text(size = 14),
        axis.text.x = element_text(angle = 90))
state_month_genus_count_facetplot
```



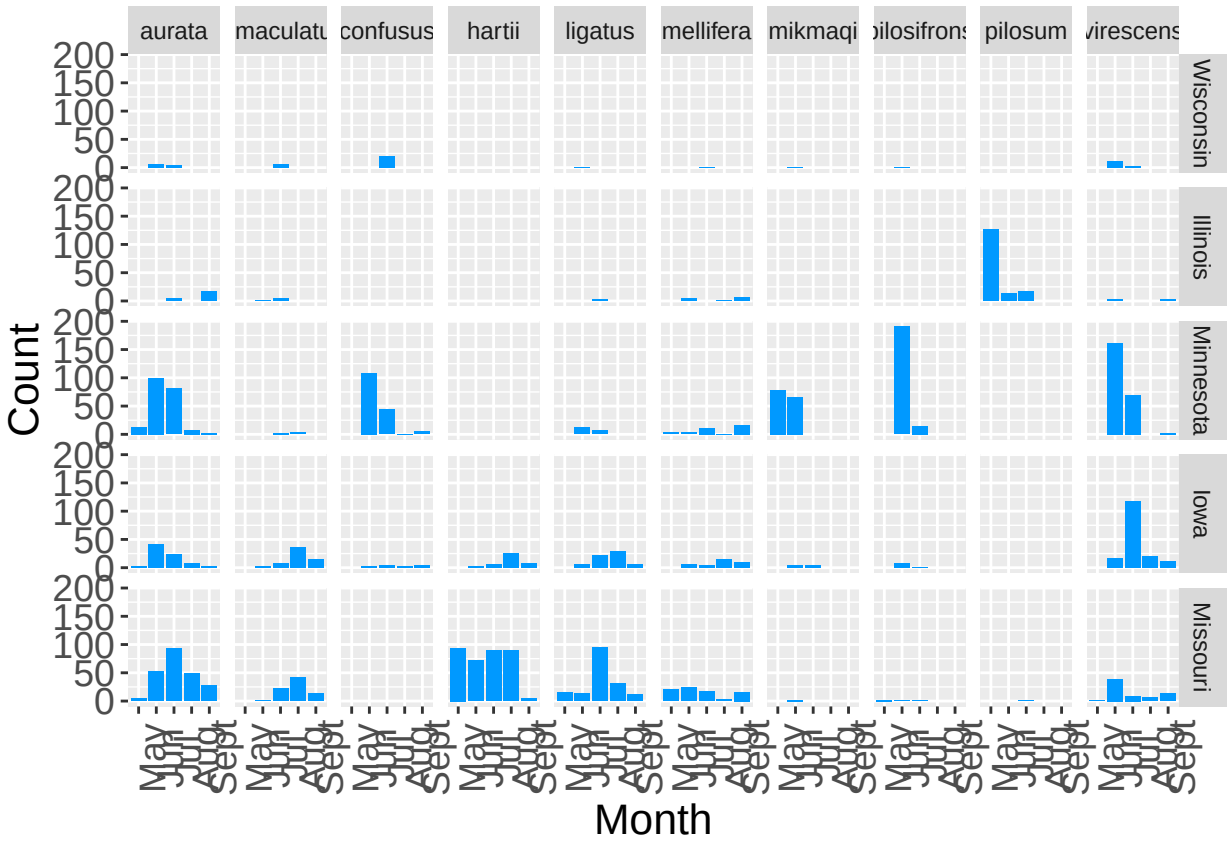

```
ggsave("state_month_genus_count_facetplot.png", state_month_genus_count_facetplot, height = 8, width =
```

```
top10_species <- bees %>% group_by(Species) %>% summarise(Count = n()) %>% top_n(10, Count) %>% arrange

state_month_species_count <- bees %>% group_by(State, Month, Species) %>% summarise(Count = n())
```

```
## Warning: Factor `Month` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
state_month_species_count_facetplot <- state_month_species_count %>% drop_na %>% filter(Species %in% to
  ggplot(aes(Month, Count)) +
  geom_col(fill = "#0099FF") +
  facet_grid(State~Species) +
  theme(axis.title = element_text(size = 16),
        plot.title = element_text(size = 20),
        axis.text = element_text(size = 14),
        axis.text.x = element_text(angle = 90))
state_month_species_count_facetplot
```


aurata
maculatu
confusus
hartii
ligatus
mellifera
mikmaqi
ilosifrons
pilosum
irescens
Wisconsin
Illinois
Minnesota
Iowa
Missouri
Count
0
50
100
150
200
May
July
Aug
Sept
Month