Homework #03: The Joy of Probability

due February 24 11:59 PM

Ayush Jain

02/19

Load Packages and Data

```
library(tidyverse)
library(fivethirtyeight)
library(viridis)
```

Exercise 1

```
bob_ross %>%
  filter(tree == '1') %>%
  summarise(n())
```

```
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1   361
```

Answer: There are 403 episodes, out of which 361 have trees in them. There is a 89.58% probability that a randomly selected episode has a tree in it.

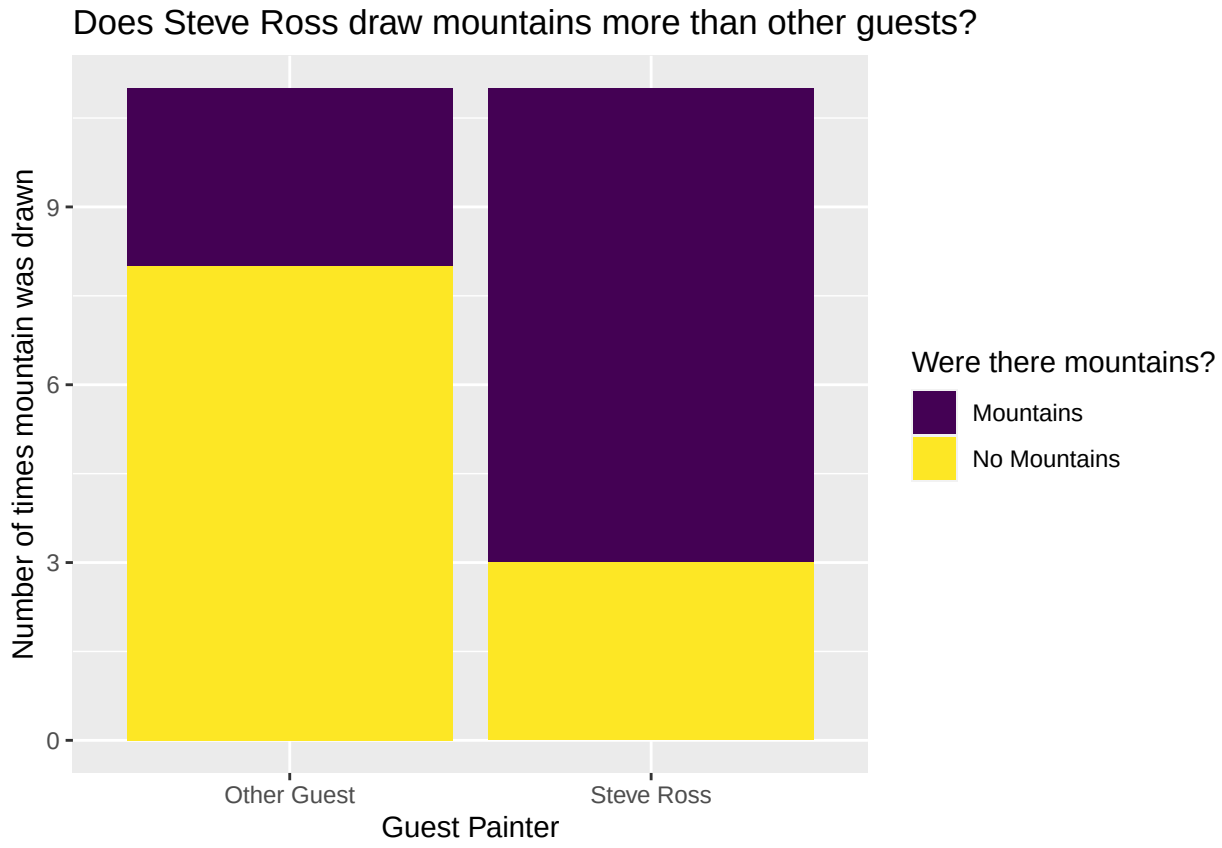
Exercise 2

```
bob_ross %>%
  filter(guest == '1') %>%
  summarise(mean = mean(steve_ross))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1   0.5
```

```
guesteps <- bob_ross %>%
  filter(guest == '1') %>%
  mutate(stevey = ifelse(steve_ross == '1',
                         'Steve Ross',
                         'Other Guest')) %>%
  mutate(mountainyes = ifelse(mountain == '1',
                         'Mountains',
                         'No Mountains'))

ggplot(data = guesteps,
       mapping = aes(x = stevey, fill = mountainyes)) +
geom_bar() +
  labs(title =
         "Does Steve Ross draw mountains more than other guests?",
       x = "Guest Painter",
       y = "Number of times mountain was drawn") +
  scale_fill_viridis(discrete = TRUE,
                     name = "Were there mountains?")
```



Answer: Given there was a guest painter, there was a 50% chance it was Steve Ross. Also, Steve Ross liked drawing mountains more than other guests.

Exercise 3

```
ross_paintings <- bob_ross %>%
  filter(guest == '0')
```

Exercise 4

```
ross_paintings %>%
  mutate(cirr_cum = ifelse((cirrus == '1') &
                             (cumulus == '1'),
                           1, 0)) %>%
summarise(mean = mean(cirr_cum))
```

```
## # A tibble: 1 x 1
##      mean
##     <dbl>
## 1 0.00262
```

Answer: The events are not disjoint because the mean of the number of times that Ross drew a cirrus cloud and a cumulus cloud is not 0, which means there are times where both were drawn

Exercise 5

```
ross_paintings %>%
  count(cabin)
```

```
## # A tibble: 2 x 2
##   cabin     n
##   <int> <int>
## 1     0   313
## 2     1    68
```

```
M = 68

ross_paintings %>%
  filter(cabin == '1') %>%
  count(lake)
```

```
## # A tibble: 2 x 2
##    lake     n
##   <int> <int>
## 1     0    44
## 2     1    24
```

```
X = 24

set.seed(2182022) # don't change the seed
 num_lakes = rbinom(100000, M, prob = 0.5)
 cabin_lakes = data.frame(num_lakes)

 cabin_lakes%>%
   mutate(both = ifelse(num_lakes > 24, 1, 0)) %>%
   summarise(mean = mean(both))
```

```
##      mean
## 1 0.99001
```

Answer: 68 of Ross' paintings featured a cabin. Given they featured a cabin, 24 also featured a lake. Lastly, if Ross flipped a coin every time he drew a cabin to decide whether he should draw a lake, only 1% of the time would he draw 24 or less cabins. This makes sense because the probability of a coin flip telling him to draw a lake (50%) is much greater than the % of times he actually drew a lake (35%), which is why most observations are higher than 24

Exercise 6

```
b_given_a <- ross_paintings %>%
  filter(river == "1") %>%
  count(mountain) %>%
  mutate(prob = n / sum(n)) %>%
  filter(mountain == "1")

a <- ross_paintings %>%
  count(river) %>%
  mutate(prob = n / sum(n)) %>%
  filter(river == "1")

b <- ross_paintings %>%
  count(mountain) %>%
  mutate(prob = n / sum(n)) %>%
  filter(mountain == "1")

(b_given_a$prob*a$prob)/b$prob
```

```
## [1] 0.3221477
```

```
b_given_a$prob
```

```
## [1] 0.3870968
```

```
a$prob
```

```
## [1] 0.3254593
```

```
b$prob
```

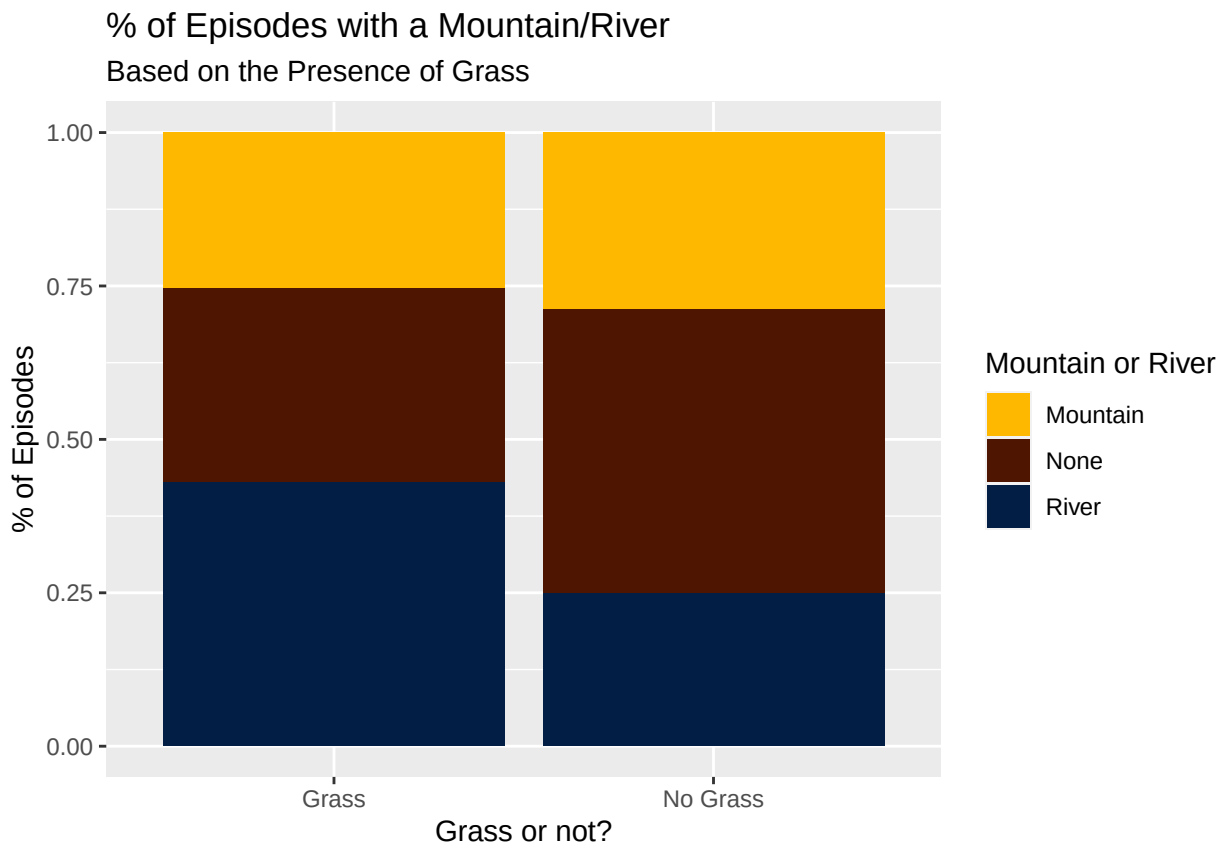
```
## [1] 0.3910761
```

Answer: There is a 32.21477% chance it also features a river. Also, they are not independent events because the probability of him painting a river given that he painted a mountain is not the same as the probability of him painting a river, which means that painting a mountain (Event A) does affect painting a river (Event B)

Exercise 7

Research Question: Did Bob Ross paint more mountains and/or rivers when he drew grass?

```
bob_ross %>%
mutate(grassy = if_else(grass == 1, "Grass", "No Grass"),
mtn_or_rvr = case_when(river == 1 ~ "River",
mountain == 1 ~ "Mountain",
TRUE ~ "None")) %>%
ggplot(mapping = aes(x = grassy, fill = mtn_or_rvr)) +
geom_bar(position = "fill") +
scale_fill_manual(values= c("#FFB800", "#4E1500", "#021E44"))+
labs(x= "Grass or not?",
     y = "% of Episodes",
fill = "Mountain or River",
title = "% of Episodes with a Mountain/River",
subtitle = "Based on the Presence of Grass")
```



Answer: We can see that the % of episodes where Ross did not draw a mountain or a river drastically increased when there was no grass. Interestingly though, he drew more mountains when there was no grass and more rivers when there was grass