Homework_3

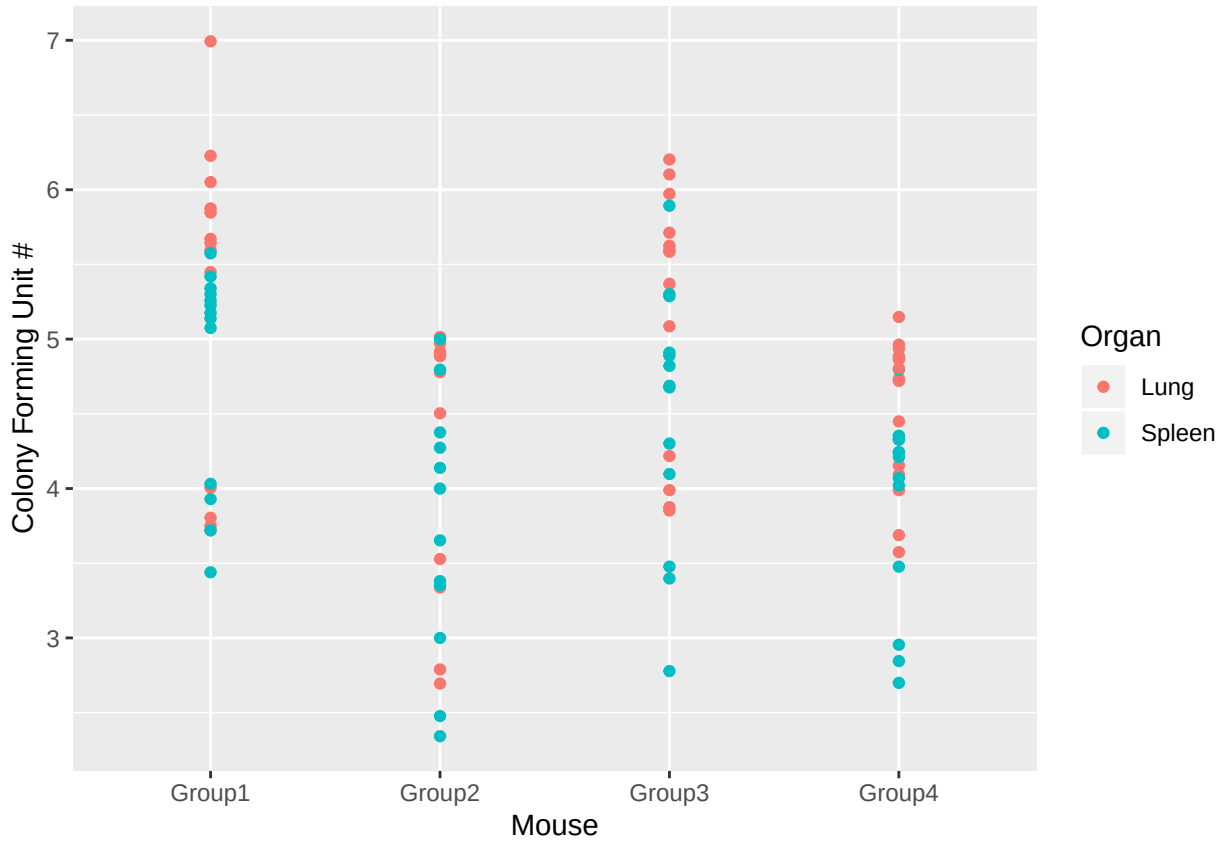
Burton Karger

10/14/2019

```
library(readr)
library(tidyverse)
library(knitr)
library(magrittr)
library(rmarkdown)
library(titanic)
library(purrr)
library(stringr)
library(forcats)
library(stringi)
library(listviewer)
```

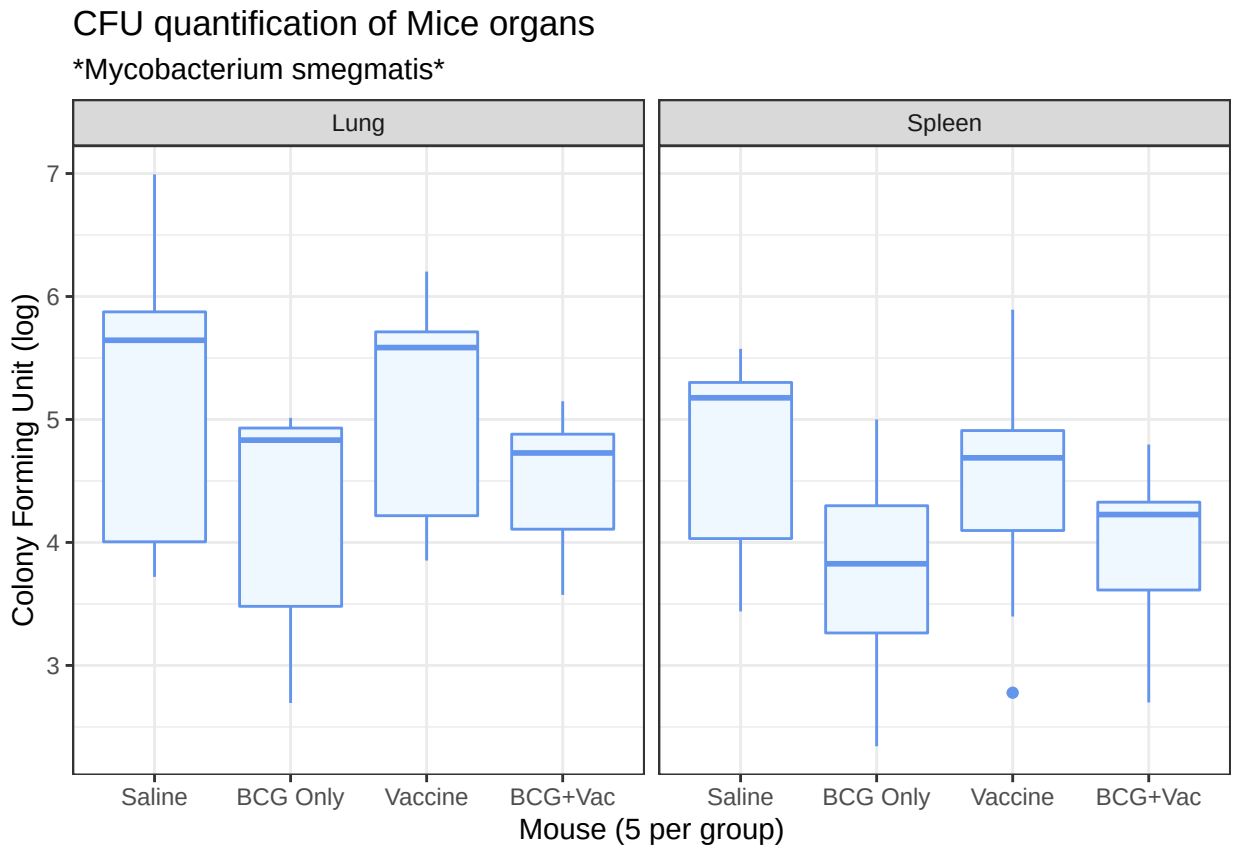
Quick and dirty plot of Colony forming units from *Mycobacterium smegmatis* per experimental mouse group (1-4).
This graph has clear labels to denote the x and y axes and a legend that denotes organ by color. There is a fair proportion of data to ink used in this plot. Though due to overlap it is difficult to distinguish where some points fall even though they are colored quite clearly. The geom_point option allows for clear visual of the max and minimums for each group. There is order in the graph by going in ascending group order since we are use to thinking about what each group is testing for, Gp 1 being saline, Gp2 being standard BCG, Gp3 being the test vaccine, and Gp4 being BCG+test vaccine.

```
read_csv("CFU Counts - Sheet1.csv", col_names = TRUE) %>%
  as_tibble %>%
  group_by(Mouse) %>%
  ggplot(aes(x = Mouse, y = CFU, color = Organ)) +
  geom_point() +
  ylab("Colony Forming Unit #")
```



This is graph of the same data as above but with much greater detail. Increased data density by using a facet_wrap call to graph both organs (lung and spleen) in the same area occupied by the overlapping graph which used color to denote organ infiltration by bacilli, instead here we used small multiples. Using a minimal, theme_bw, we are able to clearly see an outlier in Group3 for the spleen organ that has a substantial lower CFU burden compared to the rest of it's group.This theme option removes things like a dark background not necessarily needed when viewing data in Rstudio or other forms on the computer. A title of the overall data was added using ggtitle, along with a subtitile to provide some more information about the bacteria that we are referring to with "CFUs". Porbably the most important alteration was the use of a boxplot graph instead of scatter plot to denote the data as it clusters the groups for easier visualization.

```
read_csv("CFU Counts - Sheet1.csv", col_names = TRUE) %>%
  as.tibble() %>%
  mutate(Mouse = str_replace(Mouse, "Group1", "Saline"),
         Mouse = str_replace(Mouse, "Group2", "BCG Only"),
         Mouse = str_replace(Mouse, "Group3", "Vaccine"),
         Mouse = str_replace(Mouse, "Group4", "BCG+Vac")) %>%
  group_by(Mouse) %>%
  ggplot(aes(x = factor(Mouse, levels = c("Saline", "BCG Only", "Vaccine", "BCG+Vac")), y = CFU)) +
  geom_boxplot(color = "cornflowerblue", fill = "aliceblue") +
  facet_wrap("Organ") +
  theme_bw() +
  ggtitle("CFU quantification of Mice organs", subtitle = ("*Mycobacterium smegmatis*")) +
  labs(x = "Mouse (5 per group)", y = "Colony Forming Unit (log)")
```



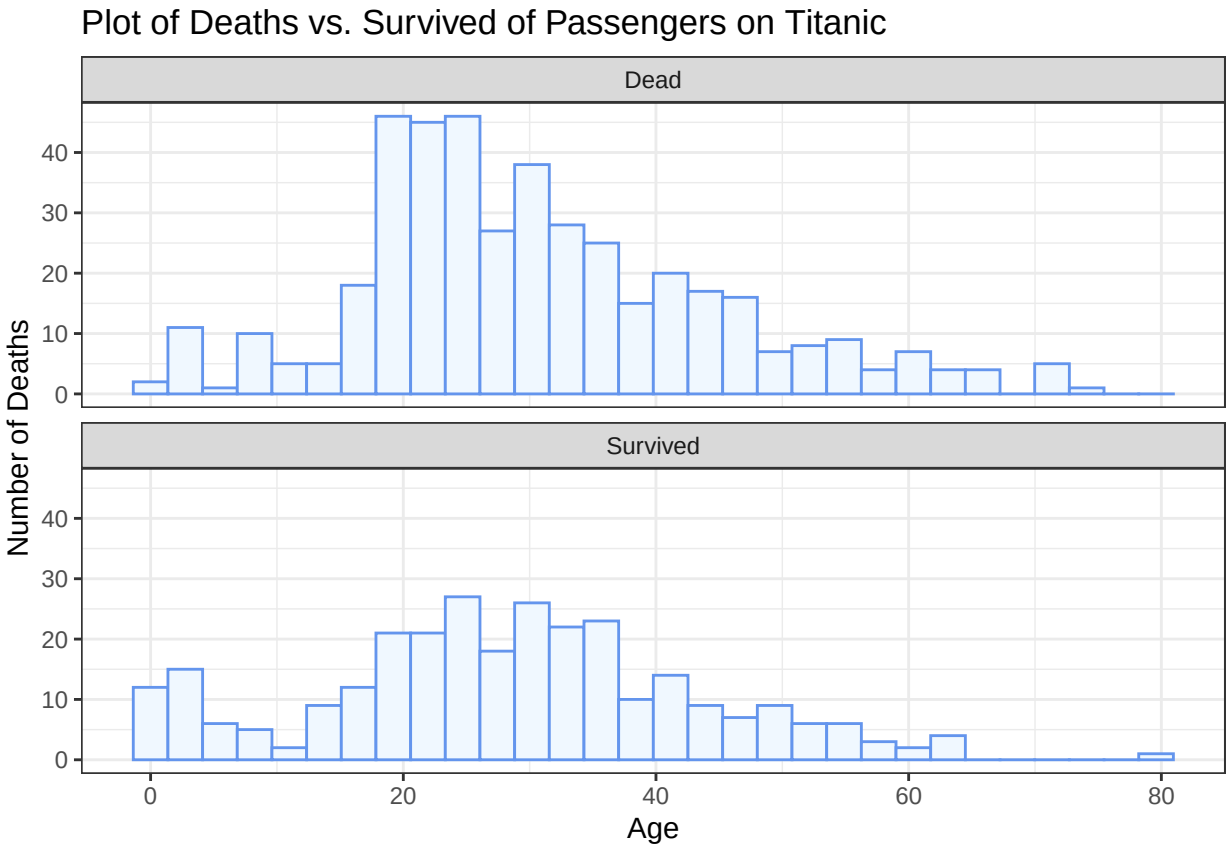
Top 5 Dead/Survived.

```
data("titanic_train")
titanic_train %>%
  select(Survived, Age) %>%
  mutate(Survived = str_replace(Survived, "0", "Dead"),
         Survived = str_replace(Survived, "1", "Survived")) %>%
  group_by(Survived) %>%
  arrange(Survived, Age) %>%
  top_n(5, Age)
```

```
## # A tibble: 11 x 2
## # Groups:   Survived [2]
##    Survived   Age
##    <chr>    <dbl>
##  1 Dead      70
##  2 Dead      70
##  3 Dead      70.5
##  4 Dead      71
##  5 Dead      71
##  6 Dead      74
##  7 Survived  62
##  8 Survived  62
##  9 Survived  63
## 10 Survived  63
## 11 Survived  80
```

Histogram

```
titanic_train %>%
  select(Survived, Age) %>%
  mutate(Survived = str_replace(Survived, "0", "Dead"),
         Survived = str_replace(Survived, "1", "Survived")) %>%
  filter(!is.na(Age), !is.na(Survived)) %>%
  ggplot(aes(x = Age)) +
  geom_histogram(color = "cornflowerblue", fill = "aliceblue", bins = 30) +
  ggtitle("Plot of Deaths vs. Survived of Passengers on Titanic") +
  ylab("Number of Deaths") +
  facet_wrap(. ~Survived, ncol = 1) +
  theme_bw()
```



Mean Age groups - NA values.

```
df <- titanic_train %>%
  select(Survived, Age) %>%
  mutate(Survived = str_replace(Survived, "0", "Dead"),
         Survived = str_replace(Survived, "1", "Survived")) %>%
  group_by(Survived) %>%
  mutate(NA_Values = is.na(Age)) %>%
  summarize(Mean_Age = mean(Age, na.rm = TRUE), total_number = n(), NA_Values = sum(NA_Values))
df
```

```
## # A tibble: 2 x 4
```

```
##   Survived Mean_Age total_number NA_Values
##   <chr>       <dbl>        <int>     <int>
## 1 Dead         30.6          549       125
## 2 Survived     28.3          342        52
```

Testing difference in means using a ttest.

```
t.test(Age ~ Survived, titanic_train)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Survived
## t = 2.046, df = 598.84, p-value = 0.04119
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.09158472 4.47339446
## sample estimates:
## mean in group 0 mean in group 1
##        30.62618        28.34369
```

Questions 4.

```
whole_words <- read_delim("words.txt", delim = " ")

df <- whole_words %>%
  as_tibble() %>%
  rename(words = `2`) %>%
  mutate(words = str_to_lower(words)) %>%
  filter(words == str_extract_all(words, pattern = "^m.{7}")) %>%
  mutate(first_word = stri_sub(words, from = 1, to = 4),
         second_word = stri_sub(words, from = 5, to = 8),
         first_letter = stri_sub(second_word, from = 4, to = 4),
         restofword = stri_sub(second_word, from = 1, to = 3),
         pasted_2ndword = paste( first_letter, restofword, sep = "")) %>%
  rename(original = words) %>%
  distinct() %>%
  select(first_word, pasted_2ndword) %>%
  semi_join(whole_words, by = c("first_word" = "2")) %>%
  semi_join(whole_words, by = c("pasted_2ndword" = "2"))

df
```

```
## # A tibble: 114 x 2
##    first_word pasted_2ndword
##    <chr>      <chr>
##  1 made       alen
##  2 made       alin
##  3 maha       amay
##  4 mail       sing
##  5 mail       slot
##  6 majo       erat
##  7 majo       grin
```

```
##  8 malm       ties
##  9 malt       ties
## 10 mand       sala
## # ... with 104 more rows
```