Take Home Quiz 2

Robert Campbell

Due 1:00pm Monday, March 29

```
library(dplyr)
library(ggplot2)
library(babynames)
library(stringr)
```

This quiz should take you approximately 25 minutes. Place your answers into this markdown document, knit it, and hand in the result as a PDF. Just answering is not enough - you need to include the R code that produces your answer.

You may use R, the internet, and any reference material, but do not work together and do not get help (except from Dr. Clair).

Problem 1

This problem uses the `babynames` data from the `babynames` library.

a. Find the most popular girl's name in the year 2000. Emily

```
babynames %>% filter(sex=="F", year==2000) %>% arrange(desc(n)) %>% head()
```

```
## # A tibble: 6 x 5
##    year sex   name        n    prop
##   <dbl> <chr> <chr>   <int>   <dbl>
## 1  2000 F     Emily   25953 0.0130
## 2  2000 F     Hannah  23080 0.0116
## 3  2000 F     Madison 19967 0.0100
## 4  2000 F     Ashley  17997 0.00902
## 5  2000 F     Sarah   17697 0.00887
## 6  2000 F     Alexis  17629 0.00884
```

b. Find the most popular girl's name in the year 2000 that starts with "Q". Quinn

```
babynames %>% filter(sex=="F", year==2000) %>% filter(str_detect(name,"^Q"))%>%
  arrange(desc(n)) %>% head()
```

```
## # A tibble: 6 x 5
##    year sex   name         n      prop
##   <dbl> <chr> <chr>    <int>     <dbl>
## 1  2000 F     Quinn      297 0.000149
```

```
## 2  2000 F     Quincy      73 0.0000366
## 3  2000 F     Quiana      61 0.0000306
## 4  2000 F     Queen       59 0.0000296
## 5  2000 F     Quanisha    37 0.0000186
## 6  2000 F     Quianna     27 0.0000135
```

Problem 2

Continue using `babynames`. Not all babies are counted in this data set - it only includes names that are given to five or more babies. The `prop` variable gives the percentage of all babies born that year with the given name.

a. What percentage of all female babies born in 2000 are included in this data? (Add up the `prop` variable for all female babies born in 2000.) 91%

```
babynames %>% filter(sex=="F", year==2000) %>% summarise(population = sum(prop))
```

```
## # A tibble: 1 x 1
##   population
##        <dbl>
## 1      0.910
```

b. How many total female babies born in 2000 are included in this data? 1,815,110

```
babynames %>% filter(sex=="F", year==2000) %>% summarise(population = sum(n))
```

```
## # A tibble: 1 x 1
##   population
##        <int>
## 1    1815110
```

c. Use parts a and b to estimate the total number of female babies born in 2000 in the U.S. 1,994,626

```
babynames %>% filter(sex=="F", year==2000) %>% summarise(population = (sum(n)/sum(prop)))
```

```
## # A tibble: 1 x 1
##   population
##        <dbl>
## 1   1994855.
```

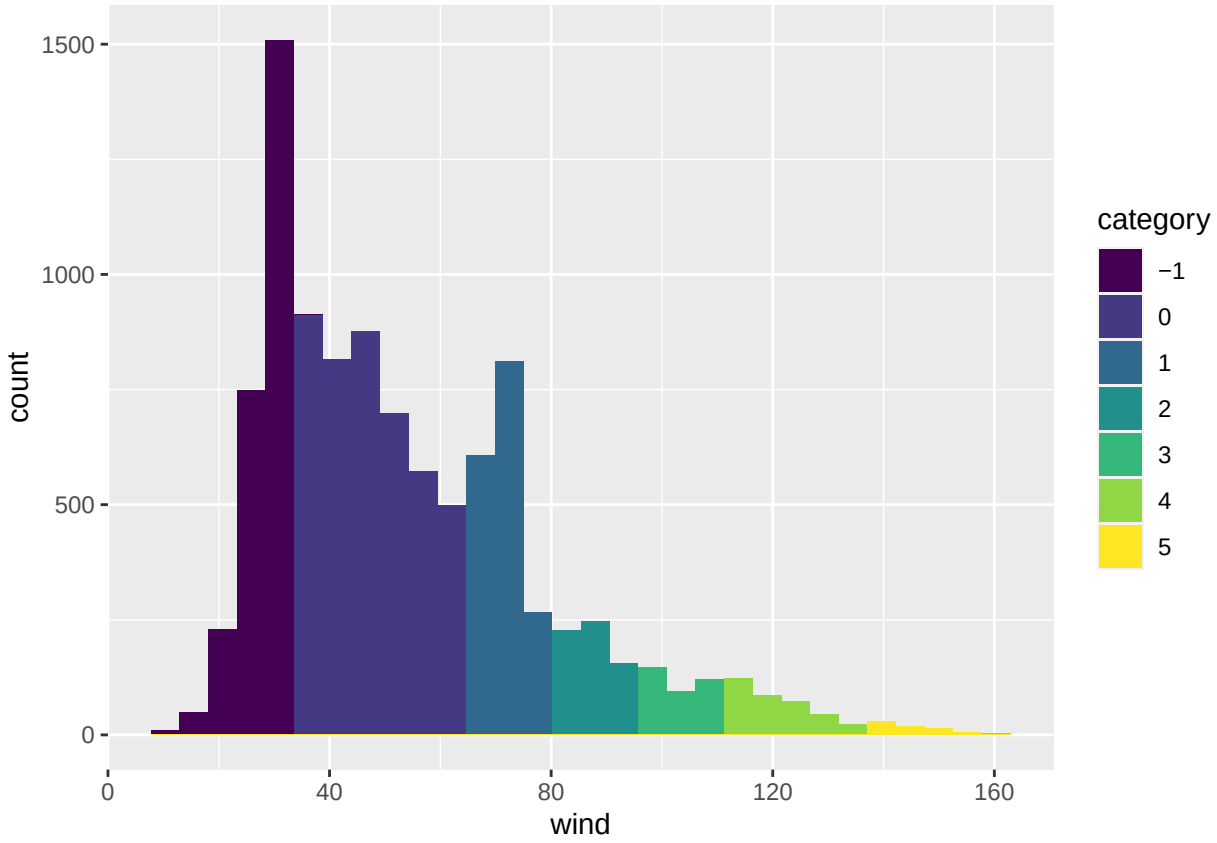
Problem 3

The data set `storms` is included in the `dplyr` package. It contains information about 198 tropical storms.

a. Use ggplot to produce a histogram of the `wind` speeds in this data set. Fill your bars using the category variable so you can see the bands of color corresponding to the different storm categories.

```
storms <- dplyr::storms
storms %>% ggplot(aes(x=wind,fill=category)) + geom_histogram()
```

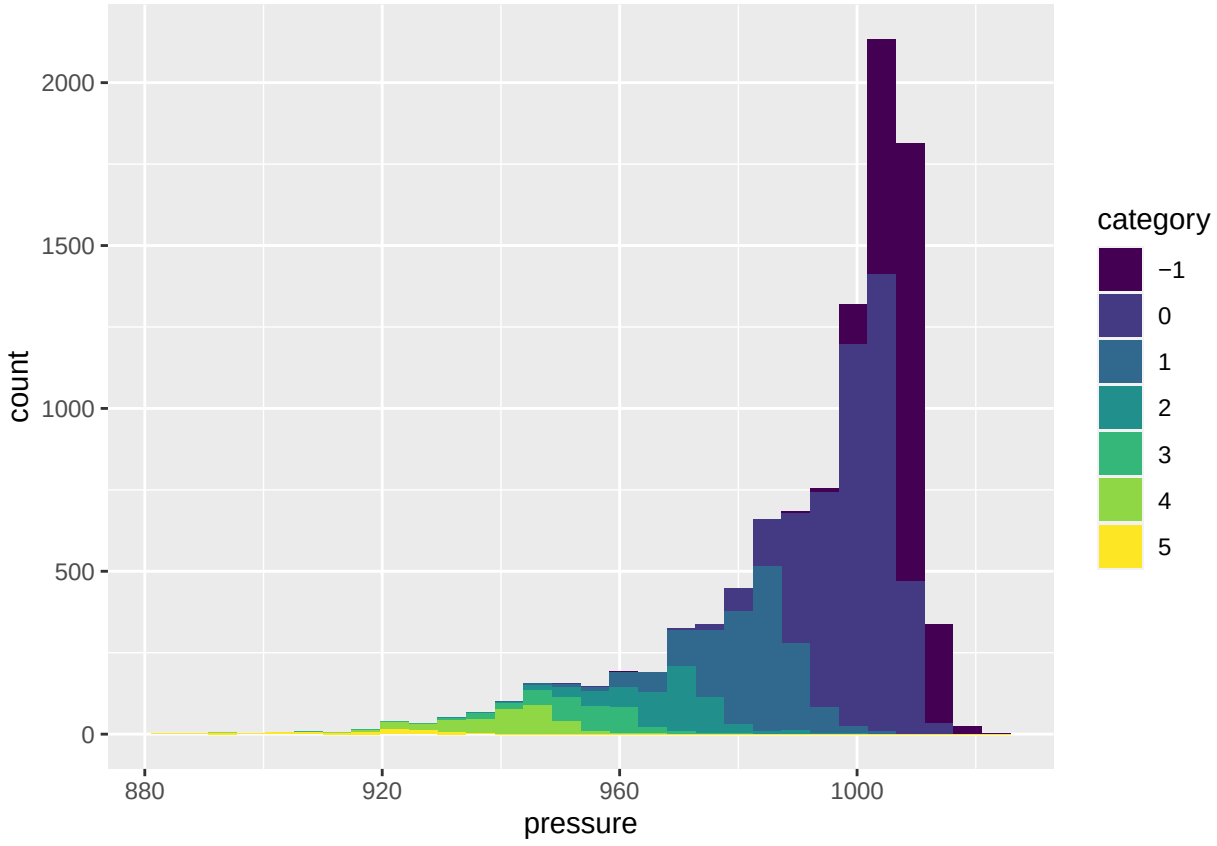
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



b. Repeat part (a) but make a histogram of the `pressure` variable. You should observe that high category storms have low pressure.

```
storms %>% ggplot(aes(x=pressure,fill=category)) + geom_histogram()
```

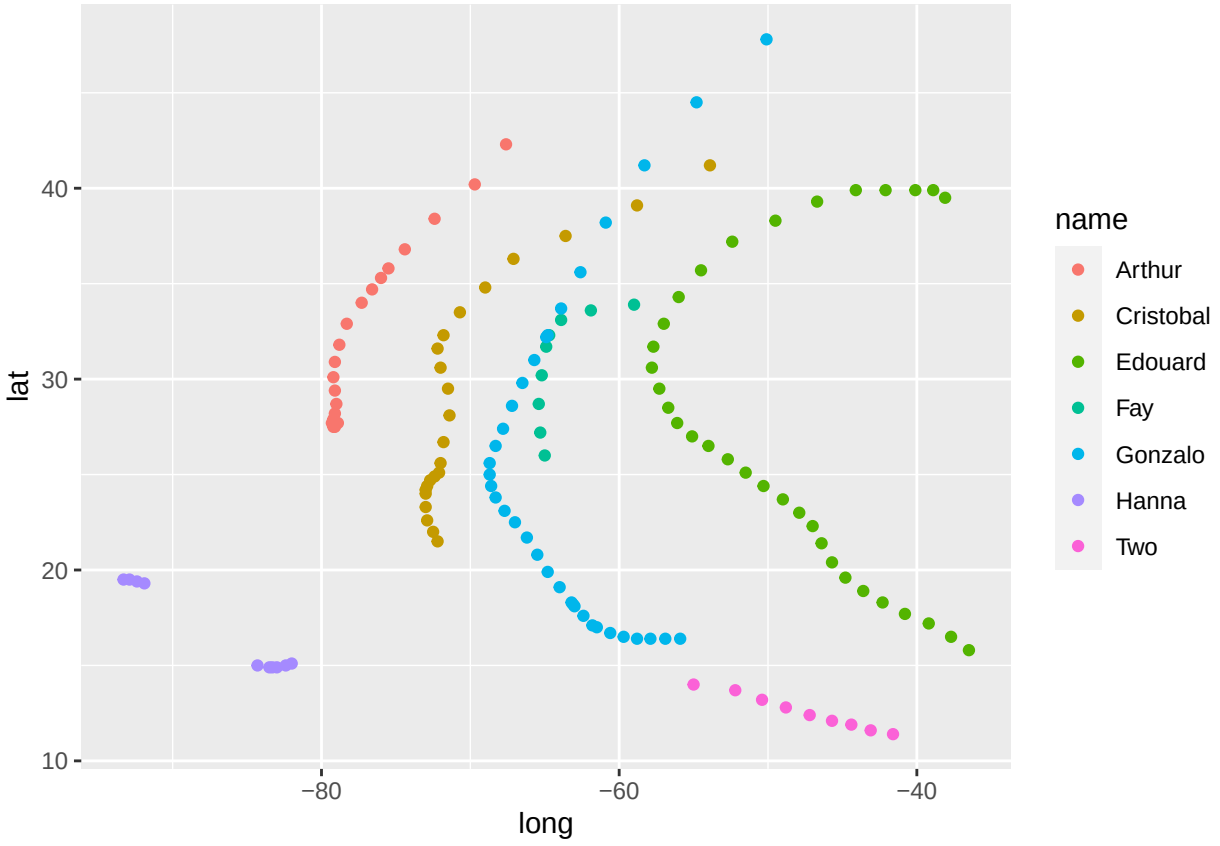
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Problem 4

Use ggplot to produce a plot showing the position track of each storm from 2014 (use `long` for x and `lat` for y). Color your points by the name of the storm so you can distinguish the seven storm tracks. Which storm in 2014 made it the furthest North? Gonzalo

```
storms %>%filter(year==2014) %>% ggplot(aes(x=long, y=lat,color=name)) + geom_point()
```



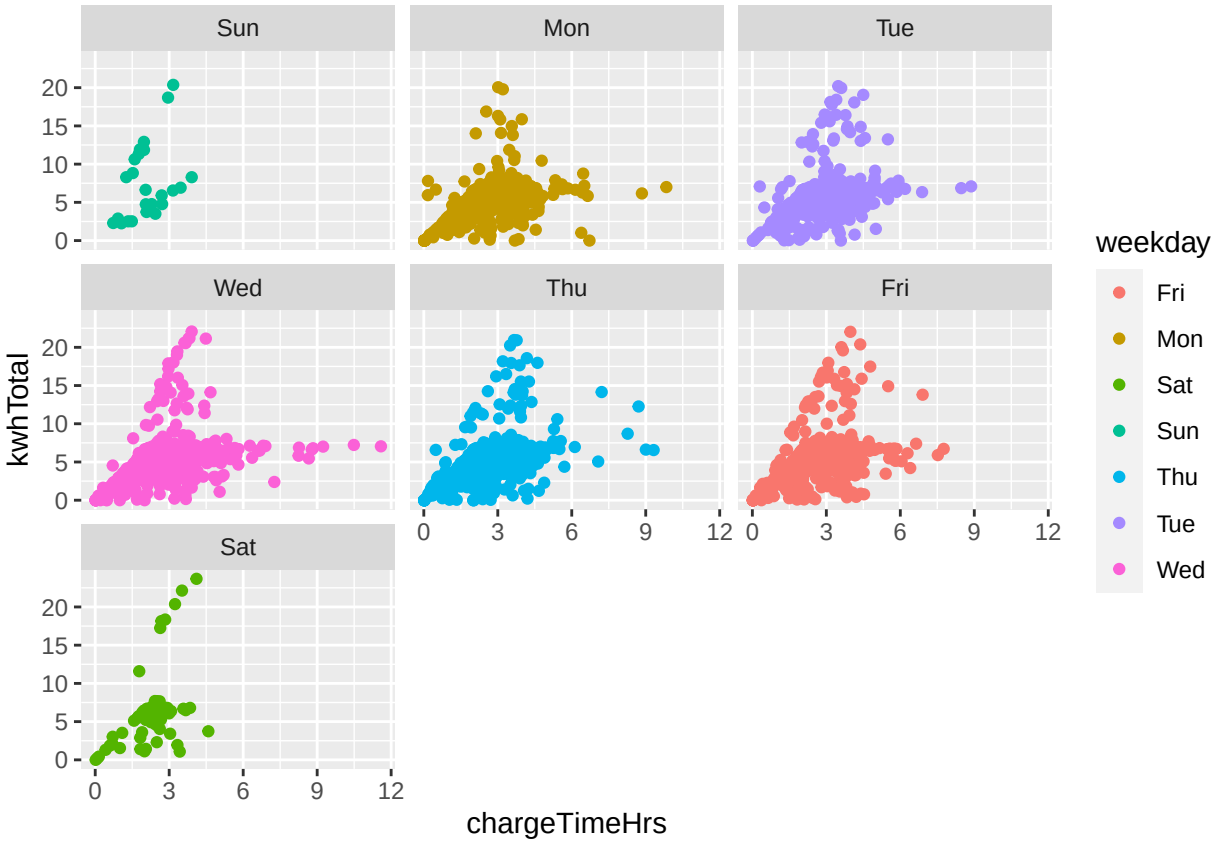
Problem 5

The `ecars` data set from `fosdata` gives information about electric car charging sessions.

Create a visualization showing seven scatterplots with the `chargeTimeHrs` variable on the x axis and the `kwhTotal` variable on the y axis. Facet your visualization with one plot per day of week, in the correct day order.

There is one outlier with a very high charge time that you should remove.

```
ecars <- fosdata::ecars
ecars %>% filter(chargeTimeHrs != max(chargeTimeHrs)) %>%
  ggplot(aes(x=chargeTimeHrs,y=kwhTotal, color=weekday)) + geom_point() +
  facet_wrap(~factor(weekday,levels=c("Sun","Mon","Tue","Wed","Thu","Fri","Sat")))
```


Sun
Mon
Tue
Wed
Thu
Fri
Sat
kwhTotal
chargeTimeHrs
0
5
10
15
20
0
3
6
9
12
weekday
Fri
Mon
Sat
Sun
Thu
Tue
Wed