R functions

Math 241, Week 3

```r
# it's good practice to check that all the packages required are loaded and installed
libs <- c('tidyverse','knitr','viridis', 'mosaic','mosaicData','babynames', 'Lahman','nycflights13')
for(l in libs){
  if(!require(l,character.only = TRUE, quietly = TRUE)){
    message( sprintf('Did not have the required package << %s >> installed. Downloading now ... ',l))
    install.packages(l)
  }
  library(l, character.only = TRUE, quietly = TRUE)
}
```

Goals of this in-class activity:

- Practice creating functions in R.

Notes:

- When creating your graphs, consider context (i.e. axis labels, title, ...)!
- If I provide partially completed code, I will put `eval = FALSE` in the chunk. Make sure to change that to `eval = TRUE` once you have completed the code in the chunk.
- Be prepared to ask for help from me, Tory, and your classmates!

Problem 1 (Medium):

Write a function called `count_name` that, when given a name as an argument, returns the total number of births by year from the `babynames` data frame in the `babynames` package that match that name. The function should return one row per year that matches (and generate an error message if there are no matches). Run the function once with the argument Ezekiel and once with Ezze.

```r
data(babynames) # this will explicitly ask R to load the babynames dataset to your environment

count_name <- function(x) {
  if (is.element(x, babynames$name)) { babynames %>%
      filter(name == x) %>%
      group_by(year) %>%
      summarize(total = sum(n)) %>%
      return()
    }
  else {
    stop("Name not found")
  }
}
```

```
count_name("Ezekiel") %>%
head()
```

```
## # A tibble: 6 x 2
##    year total
##   <dbl> <int>
## 1  1880    16
## 2  1881    22
## 3  1882    11
## 4  1883    14
## 5  1884    13
## 6  1885    10
```

Problem 2 (Medium):

1. Write a function called `count_na` that, when given a vector as an argument, will count the number of NA's in that vector. Count the number of missing values in the `SEXRISK` variable in the `HELPfull` data frame in the `mosaicData` package.

```
data(HELPfull) # this will explicitly ask R to load the HELPfull dataset to your environment
count_na <- function(x){
  return(sum(is.na(x)))
}
HELPfull %>%
  pull(SEXRISK) %>%
  count_na()
```

```
## [1] 19
```

2. Apply `count_na` to the columns of the `Teams` data frame from the `Lahman` package. How many of the columns have missing data?

```
library(Lahman) # this will explicitly ask R to load the Teams dataset to your environment

#' There are several ways to apply a function to multiple columns of a data frame.
#' Here we first use the map function from the purrr package to apply the function to each column.
missvars <- Teams %>%
  map_int(count_na)

#' Equivalently, you can use the apply function from base R to apply the function to each column.
missvars <- Teams %>%
  apply(2, count_na)

mosaic::tally(~missvars)
```

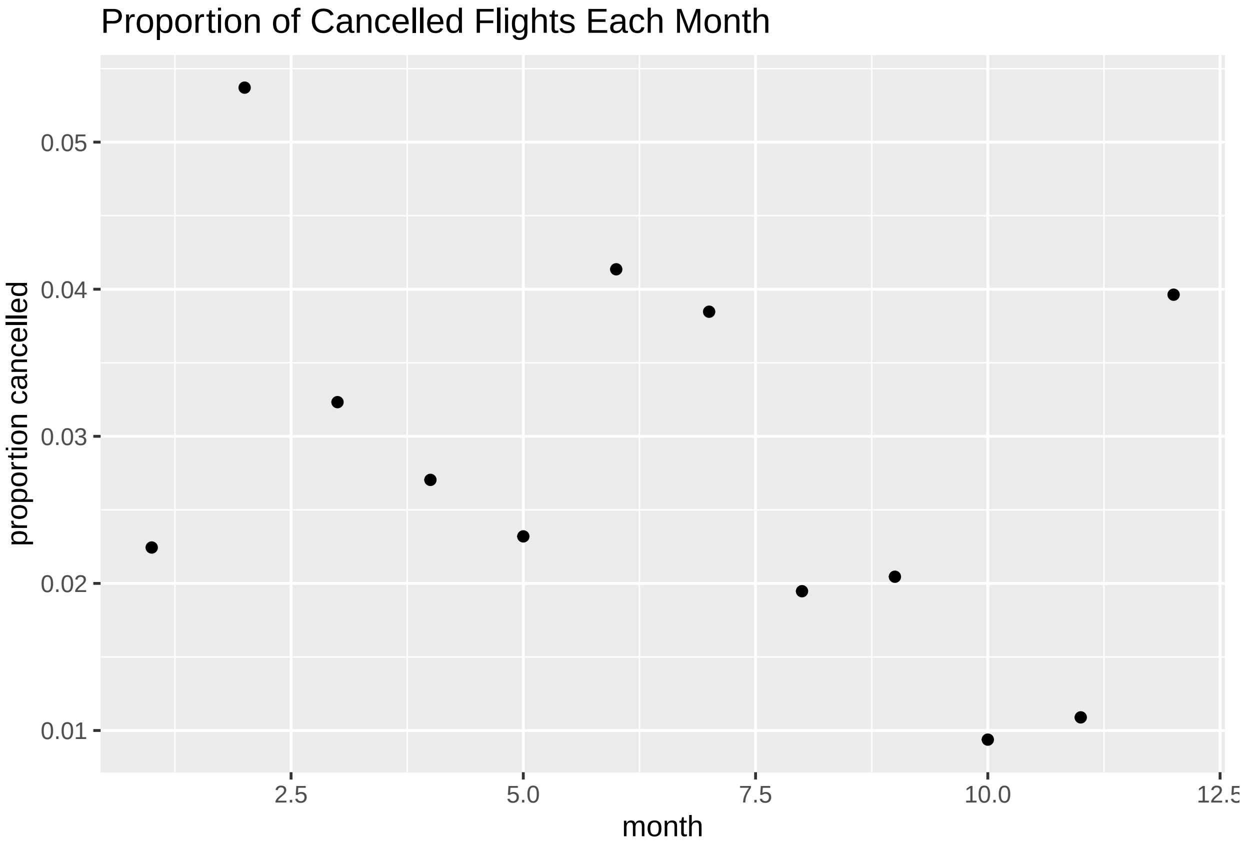
```
## missvars
##    0   16   28   34  125  279  357  399  831 1158 1517 1541 1545 2181
##   35    1    1    1    1    1    1    1    1    1    1    1    1    1
```

Problem 3 (Medium):

Write a function called `prop_cancel` that takes as arguments a month number and destination airport and returns the proportion of flights missing arrival delay for each day to that destination. Apply this function to the `nycflights13` package for February and Atlanta airport ATL and again with an invalid month number.

```r
flights2 <- flights %>%
  group_by(month) %>%
  summarize(
    cancelled = sum(is.na(arr_delay)),
    total = n(),
    prop_cancelled = cancelled / total
  )

ggplot(data = flights2, aes(x = month, y = prop_cancelled)) +
  geom_point() +
  labs(
    title = "Proportion of Cancelled Flights Each Month",
    y = "proportion cancelled"
  )
```



February had the highest proportion of cancelled flights while October had the lowest. The data shows that February, December, and summer are the periods with the greatest number of cancellations, which may be because they are often also the stormiest and snowiest periods of the year, with severe weather likely to be causing cancellations.