Homework 2

Ola Tranum Arnegaard - 12/08/2019

---

1) *Read Chapters 4*

2) *Please use the text book "Modern Data Science with R" and complete the following exercises: Chapter 4 Page 88 4.1,4.2,4.3,4.4*

Dataset:

```
head(flights,3)
```

```
## # A tibble: 3 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>
```

```
head(planes,3)
```

```
## # A tibble: 3 x 9
##   tailnum  year type       manufacturer   model  engines seats speed engine
##   <chr>   <int> <chr>      <chr>          <chr>    <int> <int> <int> <chr>
## 1 N10156   2004 Fixed win~ EMBRAER        EMB-1~       2    55    NA Turbo~
## 2 N102UW   1998 Fixed win~ AIRBUS INDUST~ A320-~       2   182    NA Turbo~
## 3 N103US   1999 Fixed win~ AIRBUS INDUST~ A320-~       2   182    NA Turbo~
```

Exercise 4.1

Each of these tasks can be performed using a single data verb. For each task, say which verb it is:

1. Find the average of one of the variables:

```
mean(var1)
```

2. Add a new column that is the ratio between two variables:

```
df %>% mutate(newCol, var1/var2)
```

3. Sort the cases in descending order of a variable.

```
df %>% arrange(desc(var1))
```

4. Create a new data table that includes only those cases that meet a criterion.

```
df %>% filter(var1 == criterion)
```

5. From a data table with three categorical variables A, B, and C, and a quantitative variable X, produce a data frame that has the same cases but only the variables A and X.

```
df %>% select(A,X)
```

Exercise 4.2

Use the `nycflights13` package and the `flights` data frame to answer the following questions: What month had the highest proportion of cancelled flights? What month had the lowest? Interpret any seasonal patterns.

```
# Can't find any data for "cancelled flights"?
# flights_sum <- flights %>% group_by(month) %>%
#                 summarize(prop_cancelled = {numer of cancelled flights each month}/n())
```

Exercise 4.3

Use the `nycflights13` package and the flights data frame to answer the following question: What plane (specified by the `tailnum` variable) traveled the most times from New York City airports in 2013? Plot the number of trips per week over the year.

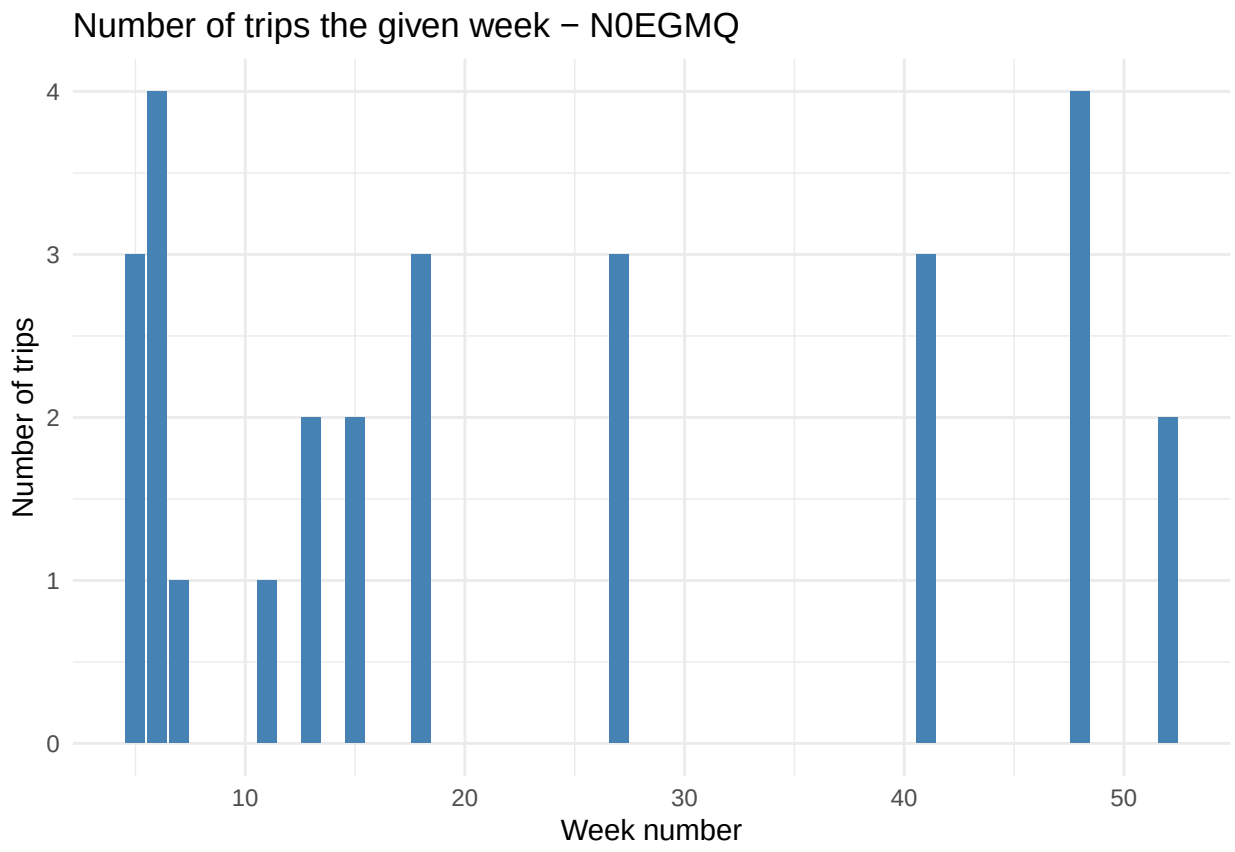
```
flights %>% filter(year == 2013 & origin == 'JFK') %>%
  group_by(tailnum) %>% summarize(number_of_trips = n())
```

```
## # A tibble: 1,958 x 2
##    tailnum number_of_trips
##    <chr>             <int>
##  1 D942DN                1
##  2 N0EGMQ               28
##  3 N102UW               15
##  4 N103US               15
##  5 N104UW               12
##  6 N105UW               13
##  7 N107US               14
##  8 N108UW               15
##  9 N109UW               10
## 10 N110UW               10
## # ... with 1,948 more rows
```

Here we see that it is `tailnum` **N0EGMQ**

```
flights_43 <- flights %>% filter(year == 2013 & origin == 'JFK' & tailnum == 'N0EGMQ')
flights_43 <- flights_43 %>% mutate(week = as.numeric(strftime(time_hour, format = "%V")))
flights_43_per_week <- flights_43 %>% group_by(week) %>% summarize(number_of_trips = n())

plot <- ggplot(flights_43_per_week, aes(x = flights_43_per_week$week, y = flights_43_per_week$number_of_
plot <- plot + geom_bar(stat="identity", fill="steelblue") + xlab("Week number")
plot <- plot + ylab("Number of trips") + ggtitle("Number of trips the given week - N0EGMQ")
plot + theme_minimal()
```



Exercise 4.4

Use the `nycflights13` package and the flights and planes tables to answer the following questions: What is the oldest plane (specified by the tailnum variable) that flew from New York City airports in 2013? How many airplanes that flew from New York City are included in the planes table?

```
flights_44 <- flights %>% filter(year == 2013 & origin == 'JFK')
flights_44 <- flights_44 %>% inner_join(planes, by = c("tailnum" = "tailnum")) %>% arrange(year.y)
head(select(flights_44, year.y, tailnum),1)
```

```
## # A tibble: 1 x 2
##   year.y tailnum
##    <int> <chr>
## 1   1956 N381AA
```

We see that the oldest plane that flew from JFK in 2013 was **N381AA** which is from **1956**

```
planes_44 <- planes
flights_44_2 <- flights %>% filter(origin == 'JFK') %>% left_join(planes_44, by = c("tailnum" = "tailnu
flights_44_2_inc <- flights_44_2 %>% filter(!is.na(model) | !is.na(manufacturer)) #Are the joined colum
flights_44_2_not_inc <- flights_44_2 %>% filter(is.na(model) | is.na(manufacturer))

nrow(flights_44_2_inc %>% group_by(tailnum)) #How many distinct different airplanes
```

```
## [1] 94142
```

```
nrow(flights_44_2_not_inc %>% group_by(tailnum))
```

```
## [1] 17137
```

We see that a total of **94142** different planes that flew from `JFK` are included in the planes table and **17137** are not.

```
x_flip = fliplr(x)
r = conv(x_flip,y)
```