HW2

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. summarize() #### 2. Add a new column that is the ratio between two variables. mutate() #### 3. Sort the cases in descending order of a variable. arrange() #### 4. Create a new data table that includes only those cases that meet a criterion. filter() #### 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. select()

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 fights? What month had the lowest? Interpret any seasonal patterns.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(nycflights13)
library(hflights)
str(hflights)
```

```
## 'data.frame':    227496 obs. of  21 variables:
##  $ Year             : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ Month            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ DayofMonth       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ DayOfWeek        : int  6 7 1 2 3 4 5 6 7 1 ...
##  $ DepTime          : int  1400 1401 1352 1403 1405 1359 1359 1355 1443 1443 ...
##  $ ArrTime          : int  1500 1501 1502 1513 1507 1503 1509 1454 1554 1553 ...
##  $ UniqueCarrier    : chr  "AA" "AA" "AA" "AA" ...
##  $ FlightNum        : int  428 428 428 428 428 428 428 428 428 428 ...
##  $ TailNum          : chr  "N576AA" "N557AA" "N541AA" "N403AA" ...
##  $ ActualElapsedTime: int  60 60 70 70 62 64 70 59 71 70 ...
##  $ AirTime          : int  40 45 48 39 44 45 43 40 41 45 ...
##  $ ArrDelay         : int  -10 -9 -8 3 -3 -7 -1 -16 44 43 ...
##  $ DepDelay         : int  0 1 -8 3 5 -1 -1 -5 43 43 ...
##  $ Origin           : chr  "IAH" "IAH" "IAH" "IAH" ...
##  $ Dest             : chr  "DFW" "DFW" "DFW" "DFW" ...
##  $ Distance         : int  224 224 224 224 224 224 224 224 224 224 ...
```

```
##  $ TaxiIn           : int  7 6 5 9 9 6 12 7 8 6 ...
##  $ TaxiOut          : int  13 9 17 22 9 13 15 12 22 19 ...
##  $ Cancelled        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CancellationCode : chr  "" "" "" "" ...
##  $ Diverted         : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
cancelled_by_month <- hflights %>% group_by(Month) %>% summarize(n_cancelled = sum(Cancelled))

max_min <- cancelled_by_month %>% summarize(max_cancelled=max(n_cancelled), min_cancelled = min(n_cancel

months_hi_lo <- cancelled_by_month %>%filter(n_cancelled == max_min$max_cancelled | n_cancelled == max_

months_hi_lo
```

```
## # A tibble: 2 x 3
##   Month n_cancelled label
##   <int>       <int> <chr>
## 1     2        1108 highest proportion of cancelled fights
## 2    11          56 lowest proportion of cancelled fights
```

It could seem like February has the highest proportion of cancelled flights because usually early in the year there's more chance of snow than in November, which is the month with the lowest proportion of cancelled flights.

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?

```
str(flights)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  19 variables:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int  517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num  2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int  830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num  11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr  "UA" "UA" "AA" "B6" ...
##  $ flight        : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num  227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num  1400 1416 1089 1576 762 ...
##  $ hour          : num  5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num  15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

Considering “EWR” is not in NYC, because it is actually in New Jersey and “LGA” is in Queens.

The plane that traveled the most times from New York City in 2013:

```
x<-flights %>% filter(origin == "JFK" & year == 2013 & is.na(tailnum) == FALSE) %>% group_by(tailnum) %
```

Plot the number of trips per week over the year.

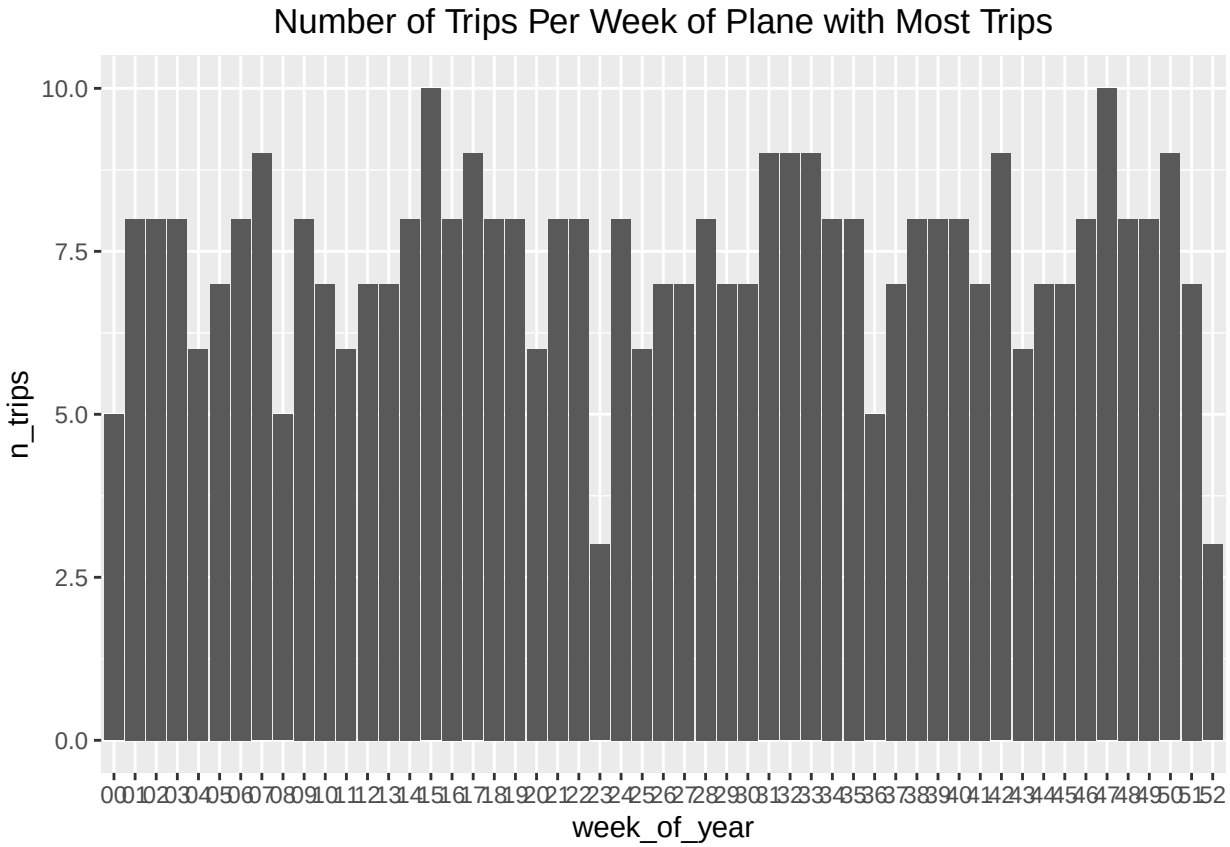
I’m assuming it refers to the plane that had the most number of trips. The week of the year can be determined by using format %U from the timestamp of the flight.

```
trips<-flights %>% filter(tailnum==x$tailnum) %>% mutate(week_of_year = format(as.Date(as.character(as.

trips
```

```
## # A tibble: 53 x 2
##    week_of_year n_trips
##    <chr>          <int>
##  1 00                 5
##  2 01                 8
##  3 02                 8
##  4 03                 8
##  5 04                 6
##  6 05                 7
##  7 06                 8
##  8 07                 9
##  9 08                 5
## 10 09                 8
## # ... with 43 more rows
```

```
library(ggplot2)
ggplot(trips, aes(x = week_of_year, y = n_trips)) + geom_bar(stat = "identity") + ggtitle("Number of Tr
```



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?

```
str(planes)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    3322 obs. of  9 variables:
##  $ tailnum     : chr  "N10156" "N102UW" "N103US" "N104UW" ...
##  $ year        : int  2004 1998 1999 1999 2002 1999 1999 1999 1999 1999 ...
##  $ type        : chr  "Fixed wing multi engine" "Fixed wing multi engine" "Fixed wing multi engine" "
##  $ manufacturer: chr  "EMBRAER" "AIRBUS INDUSTRIE" "AIRBUS INDUSTRIE" "AIRBUS INDUSTRIE" ...
##  $ model       : chr  "EMB-145XR" "A320-214" "A320-214" "A320-214" ...
##  $ engines     : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ seats       : int  55 182 182 182 55 182 182 182 182 182 ...
##  $ speed       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ engine      : chr  "Turbo-fan" "Turbo-fan" "Turbo-fan" "Turbo-fan" ...
```

```
flights%>%select(tailnum)%>%left_join(select(planes, tailnum, year), by = c("tailnum" = "tailnum"))%>%s
```

```
## # A tibble: 1 x 2
```

```
##   tailnum  year
##   <chr>   <int>
## 1 N381AA   1956
```

How many airplanes that flew from New York City are included in the planes table?

An inner join will give us all the planes that are in the flights table. We filter by the one airport in NYC. The total number of airplanes are not necessarily the number of rows (flights), we have to count the unique number of planes.

```
flights%>%filter(origin=='JFK')%>%select(tailnum)%>%inner_join(select(planes, tailnum), by = c("tailnum
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1  1381
```