SaahilDeshpande_warmup

Saahil Deshpande

November 18, 2017

Question 1

How many total flights are there in the data set?

```
require(ggplot2)

require(dplyr)

load(file="2014flights.Rdata")
format(object.size(df),units='MiB')#get the size in memory

## [1] "821.8 MiB"

summarise(df,count = n())

##     count
## 1 5819811
```

There are a total of 5819811 flights in the data set

Question 2

How many flights were there for each day of the week?

```
by_DOW<- df %>% group_by(DAY_OF_WEEK) %>% summarise(count = n())
by_DOW

## # A tibble: 7 x 2
##   DAY_OF_WEEK  count
##         <int>  <int>
## 1           1 867299
## 2           2 845320
## 3           3 864033
## 4           4 862984
## 5           5 863423
## 6           6 698712
## 7           7 818040
```

The week begins from Monday, hence 1 corresponds to a Monday, and ends on a Sunday, hence 7 corresponds to a Sunday.

Question 3

Which month has the greatest proportion of late flights? Formally test the question: is there any difference in the true proportion of late flights across months?

```
by_mon<- df %>% filter(CANCELLED==0,!is.na(ARR_DELAY))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.1
```

```
by_mon<- by_mon %>% group_by(MONTH) %>% summarise(count = n(), delayed =
length(which(ARR_DELAY>0)), proportion=length(which(ARR_DELAY>0))/n())
by_mon
```

```
## # A tibble: 12 x 4
##    MONTH  count delayed proportion
##    <int>  <int>   <int>      <dbl>
##  1     1 439620  206492  0.4697057
##  2     2 405741  184487  0.4546915
##  3     3 493043  207478  0.4208112
##  4     4 476881  185651  0.3893026
##  5     5 488073  200422  0.4106394
##  6     6 490716  230496  0.4697136
##  7     7 511003  220419  0.4313458
##  8     8 500117  204515  0.4089343
##  9     9 461725  172586  0.3737853
## 10    10 485013  190567  0.3929111
## 11    11 457046  175173  0.3832721
## 12    12 469400  207445  0.4419365
```

June, that is the 6th month seems to have the maximum number of delayed flights with 230496 late flights. July has the next maximum nuber of delayed flights but they are nearly 10000 flights lesser than june. But this data could be misleading as we are not considering the total number of flights that actually did travel every month. If we check the proportion of flights that were late every month, we can see that even though june still have the maximum proportion of late flights, the other months have nearly the same proportion. Based on this data it wouldn't be appropriate to say that there exists a difference in the true proportion of late flights across months.

Question 4

Which day is best for minimizing average departure delays?

```
by_day<- df %>% filter(CANCELLED==0,!is.na(DEP_DELAY))
by_day<- by_day %>% group_by(DAY_OF_WEEK) %>% summarise(avg_dep_delay =
mean(DEP_DELAY))
by_day[which.min(by_day$avg_dep_delay),]
```

```
## # A tibble: 1 x 2
##   DAY_OF_WEEK avg_dep_delay
##         <int>         <dbl>
## 1           6      8.526104
```

Based on our data we can say that saturday is the best day for minimizing the average departure delays.

Question 5

Which departure and arrival airport combination is associated with the worst median delay?

```
by_arpt<- df%>% filter(CANCELLED==0,!is.na(ARR_DELAY))
by_arpt<- by_arpt %>% group_by(ORIGIN_AIRPORT_ID,DEST_AIRPORT_ID) %>%
summarise(delay = median(ARR_DELAY))
by_arpt[which.max(by_arpt$delay),]
```

```
## # A tibble: 1 x 3
## # Groups:   ORIGIN_AIRPORT_ID [1]
##   ORIGIN_AIRPORT_ID DEST_AIRPORT_ID delay
##               <int>           <int> <dbl>
## 1             10693           10599   399
```

flights departing from Nashville, TN: Nashville International and arriving at Birmingham, AL: Birmingham-Shuttlesworth International would have the worst median delay of 399 minutes.