homework_4_kanutala_santosh

Santosh Kanutala

7/1/2018

****My Github repository for my assignments can be found at below URL: (https://github.com/santumagic/compscix-415-2assignments.git)****

```
library(tidyverse)
library(mdsr)
library(nycflights13)
```

Section 5.6.7: #2, #4 and #6 only. Extra Credit: Do #5

QUESTION 2:

```
# First lets find the not cancelled flights.
not_cancelled <- flights %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))

# given code for not_cancelled %>% count(dest)
not_cancelled %>%
  count(dest)
```

```
## # A tibble: 104 x 2
##    dest      n
##    <chr> <int>
##  1 ABQ     254
##  2 ACK     264
##  3 ALB     418
##  4 ANC       8
##  5 ATL   16837
##  6 AUS    2411
##  7 AVL     261
##  8 BDL     412
##  9 BGR     358
## 10 BHM     269
## # ... with 94 more rows
```

```
# new code for not_cancelled %>% count(dest) by group by and summarise
not_cancelled %>%
  group_by(dest) %>%
  summarise(n = n())
```

```
## # A tibble: 104 x 2
##    dest      n
##    <chr> <int>
##  1 ABQ     254
##  2 ACK     264
##  3 ALB     418
##  4 ANC       8
```

```
##  5 ATL   16837
##  6 AUS    2411
##  7 AVL     261
##  8 BDL     412
##  9 BGR     358
## 10 BHM     269
## # ... with 94 more rows
```

```
# given code for not_cancelled %>% count(tailnum, wt = distance)
not_cancelled %>%
  count(tailnum, wt = distance)
```

```
## # A tibble: 4,037 x 2
##    tailnum      n
##    <chr>    <dbl>
##  1 D942DN    3418
##  2 N0EGMQ  239143
##  3 N10156  109664
##  4 N102UW   25722
##  5 N103US   24619
##  6 N104UW   24616
##  7 N10575  139903
##  8 N105UW   23618
##  9 N107US   21677
## 10 N108UW   32070
## # ... with 4,027 more rows
```

```
# new code for not_cancelled %>% count(tailnum, wt = distance) group by and summarise
not_cancelled %>%
  group_by(tailnum) %>%
  summarize(n = sum(distance, na.rm = TRUE))
```

```
## # A tibble: 4,037 x 2
##    tailnum      n
##    <chr>    <dbl>
##  1 D942DN    3418
##  2 N0EGMQ  239143
##  3 N10156  109664
##  4 N102UW   25722
##  5 N103US   24619
##  6 N104UW   24616
##  7 N10575  139903
##  8 N105UW   23618
##  9 N107US   21677
## 10 N108UW   32070
## # ... with 4,027 more rows
```

QUESTION 4:

```
# number of cancelled flights per day
(cancelled_flights <- flights %>%
  filter(is.na(dep_time)) %>%
  count(day))
```

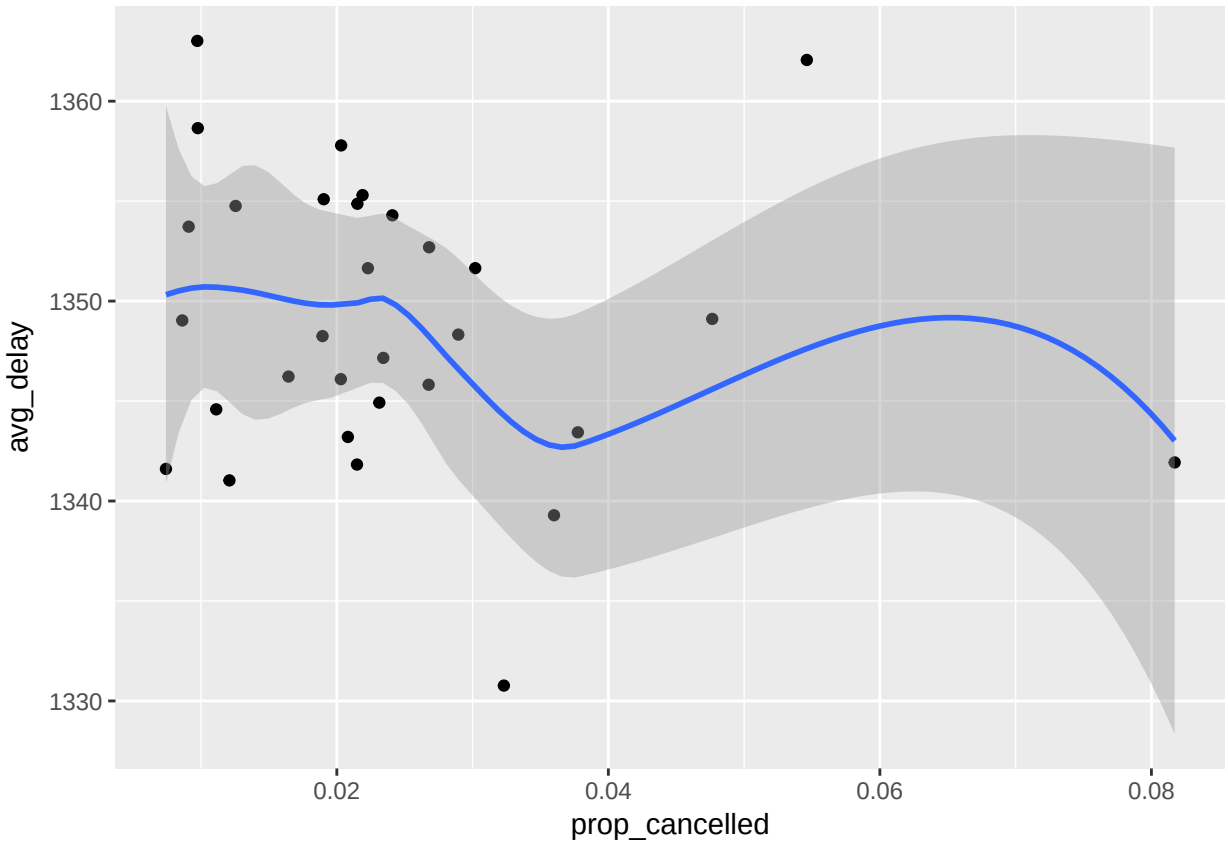
```
## # A tibble: 31 x 2
```

```
##      day     n
##    <int> <int>
##  1     1   246
##  2     2   250
##  3     3   109
##  4     4    82
##  5     5   226
##  6     6   296
##  7     7   318
##  8     8   921
##  9     9   593
## 10    10   535
## # ... with 21 more rows
```

```
# proportion of day cancelled flights vs aerage delays
(cancelled_flights <- flights %>%
  group_by(day) %>%
  summarize(prop_cancelled = sum(is.na(dep_time)) / n(),
            avg_delay = mean(dep_time, na.rm = TRUE)))
```

```
## # A tibble: 31 x 3
##      day prop_cancelled avg_delay
##    <int>          <dbl>     <dbl>
##  1     1        0.0223      1352.
##  2     2        0.0231      1345.
##  3     3        0.00972     1363.
##  4     4        0.00741     1342.
##  5     5        0.0208      1343.
##  6     6        0.0268      1346.
##  7     7        0.0289      1348.
##  8     8        0.0817      1342.
##  9     9        0.0546      1362.
## 10    10        0.0477      1349.
## # ... with 21 more rows
```

```
# plot for the relationship
ggplot(cancelled_flights, aes(x = prop_cancelled, y = avg_delay)) +
  geom_point() +
  geom_smooth()
```


1360
1350
1340
1330
avg_delay
0.02
0.04
0.06
0.08
prop_cancelled