Data Management With R: Exploratory Data Analysis

Matthias Haber

25 September 2017

Prerequisites

Last week's homework

Exploratory Data Analysis

Homework Exercises

Prerequisites

Packages

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3
```

Data

336,776 flights that departed from New York City in 2013

```
# install.packages("nycflights13")
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R versio
```

year	month	day	dep_time	sched_dep_time	dep_delay
2013	1	1	517	515	2
2013	1	1	533	529	4
2013	1	1	542	540	2
2013	1	1	544	545	-1

Last week's homework

Homework Solutions

```
library(tidyverse)
library(nycflights13)
data("flights")
```

Homework Question 1

```
# Which destination has the most carriers?
flights %>%
  filter(!is.na(dep_delay), !is.na(arr_delay)) %>%
  group_by(dest) %>%
  summarise(carriers = n_distinct(carrier)) %>%
  arrange(desc(carriers)) %>%
  head(n = 4)
```

```
## # A tibble: 4 × 2
##    dest carriers
##   <chr>    <int>
## 1   ATL        7
## 2   BOS        7
## 3   CLT        7
## 4   ORD        7
```

Homework Question 2

```
# Which destination has the largest spread (standard
# deviation) in terms of distance that planes traveled
# to get to it.
flights %>%
  group_by(dest) %>%
  summarise(spread = sd(distance)) %>%
  arrange(desc(spread)) %>%
  head(n = 1)
```

```
## # A tibble: 1 × 2
##    dest   spread
##   <chr>    <dbl>
## 1   EGE 10.54907
```

Homework Question 3

```
# What is the average (mean) departure delay of United
# Airlines? Round to the nearest integer.
flights %>%
  filter(carrier == "UA") %>%
  summarise(delay = round(mean(dep_delay, na.rm =TRUE)))
```

```
## # A tibble: 1 × 1
##   delay
##   <dbl>
## 1    12
```

Homework Question 4

```
# How many flights were delayed by at least an hour,
# but made up over 45 minutes in flight?
flights %>%
  filter(dep_delay >= 60, dep_delay-arr_delay > 45) %>%
  n_distinct()
```

```
## [1] 245
```

Homework Question 5

```
# At what time (minutes after midnight) did the first
# flight leave on September 18, 2013?
flights %>%
  filter(month == 9, day == 18) %>%
  mutate(dep_time2 = dep_time %/% 100 * 60 +
           dep_time %% 100) %>%
  select(dep_time2) %>%
  arrange(dep_time2) %>%
  head(n=1)
```

```
## # A tibble: 1 × 1
##   dep_time2
##       <dbl>
## 1       290
```

Homework Question 6

```
# How many flights left before 5am in September (including
# # of delayed flights from the previous day)?
flights %>%
  filter(!is.na(dep_delay)) %>%
  filter(month == 9, dep_time < 500) %>%
  n_distinct()
```

```
## [1] 66
```

Homework Question 7

```
# Which departure airport (FAA airport code) has the
# highest number of departure delays that are longer
# than 2 hours?
flights %>%
  filter(dep_delay > 120) %>%
  group_by(origin) %>%
  summarise(delay = n())
```

```
## # A tibble: 3 × 2
##   origin delay
##    <chr> <int>
## 1    EWR  3884
## 2    JFK  3048
## 3    LGA  2791
```

Homework Question 8

```
# Which departure airport (FAA airport code) has
# the longest mean departure delay in September?
flights %>%
  filter(month == 9) %>%
  group_by(origin) %>%
  summarise(delay = mean(dep_delay, na.rm =TRUE))
```

```
## # A tibble: 3 × 2
##   origin    delay
##    <chr>    <dbl>
## 1    EWR 7.290954
## 2    JFK 6.635776
## 3    LGA 6.207439
```

Homework Question 9

```
# Which carrier (two letter abbreviation) has the
# shortest average (mean) departure delay when you
# take into account the distance that carrier traveled?
flights %>%
  group_by(carrier) %>%
  mutate(delay = dep_delay / distance) %>%
  summarise(delay = mean(delay, na.rm =T)) %>%
  arrange(delay) %>%
  head(n=1)
```

```
## # A tibble: 1 × 2
##   carrier         delay
##     <chr>         <dbl>
## 1      HA 0.0009834607
```

Homework Question 10

```
# Which plane (tailnum) has the worst on-time
# record in terms of arrival delay?
flights %>%
  group_by(tailnum) %>%
  summarise(delay = mean(arr_delay, na.rm =T)) %>%
  arrange(desc(delay)) %>%
  head(n=1)
```

```
## # A tibble: 1 × 2
##   tailnum delay
##     <chr> <dbl>
## 1  N844MH   320
```

Exploratory Data Analysis

Goals

> *"There are no routine statistical questions, only questionable statistical routines." — Sir David Cox*
>
> *"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." — John Tukey*

Learn how to use visualization and transformation to systematically explore your data to answer or generate questions about your data.

Homework Exercises

Homework Exercises

For this week's homework exersises go to Moodle and answer the Quiz posted in the Data Transformation section. You will be asked a number of questions randomly selected from a question pool. If you work in pairs, then you might get two different sets of questions.

Deadline: Sunday, October 1 before midnight.

That's it for today. Questions?