Week 1 Statistical computing

MS 276

September, 2017

The airlines data set

Today's class is going to focus on a data set containing all flights departing the three primary New York City airports in 2013. After installing the package, we take a look at the data set below.

```
library(nycflights13)
library(tidyverse)
head(flights) %>% print.data.frame()
```

```
##   year month day dep_time sched_dep_time dep_delay arr_time sched_arr_time
## 1 2013     1   1      517            515         2      830            819
## 2 2013     1   1      533            529         4      850            830
## 3 2013     1   1      542            540         2      923            850
## 4 2013     1   1      544            545        -1     1004           1022
## 5 2013     1   1      554            600        -6      812            837
## 6 2013     1   1      554            558        -4      740            728
##   arr_delay carrier flight tailnum origin dest air_time distance hour
## 1        11      UA   1545  N14228    EWR  IAH      227     1400    5
## 2        20      UA   1714  N24211    LGA  IAH      227     1416    5
## 3        33      AA   1141  N619AA    JFK  MIA      160     1089    5
## 4       -18      B6    725  N804JB    JFK  BQN      183     1576    5
## 5       -25      DL    461  N668DN    LGA  ATL      116      762    6
## 6        12      UA   1696  N39463    EWR  ORD      150      719    5
##   minute           time_hour
## 1     15 2013-01-01 05:00:00
## 2     29 2013-01-01 05:00:00
## 3     40 2013-01-01 05:00:00
## 4     45 2013-01-01 05:00:00
## 5      0 2013-01-01 06:00:00
## 6     58 2013-01-01 05:00:00
```

```
dim(flights)
```

```
## [1] 336776     19
```

```
set.seed(0)
flights <- flights %>% sample_n(10000)
```

The **codebook** (description of the variables) can be accessed by pulling up the help file:

```
?flights
```

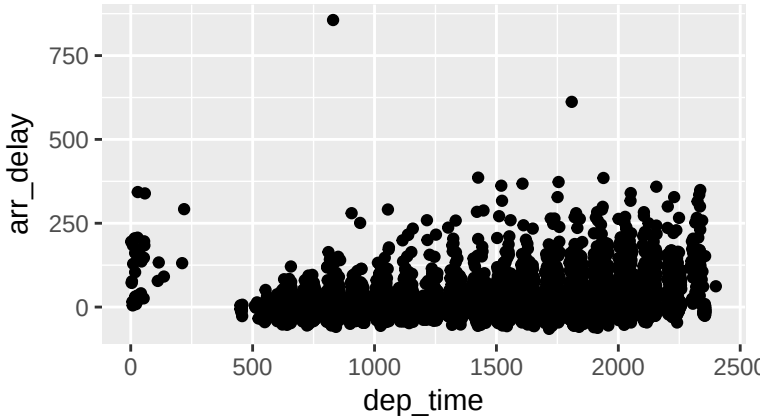
The `flights` data frame is a massive trove of information. Let's think about some graphs that we could make:

- Do later flights (later in the day) have longer or shorter arrival delays?
- How do departure delays vary by month?
- Which NYC airport is the most likely to have a delay of any kind?

Grammar of graphics

To plot `flights`, the following code puts `dep_time` on the x-axis and `arr_delay` on the y-axis.

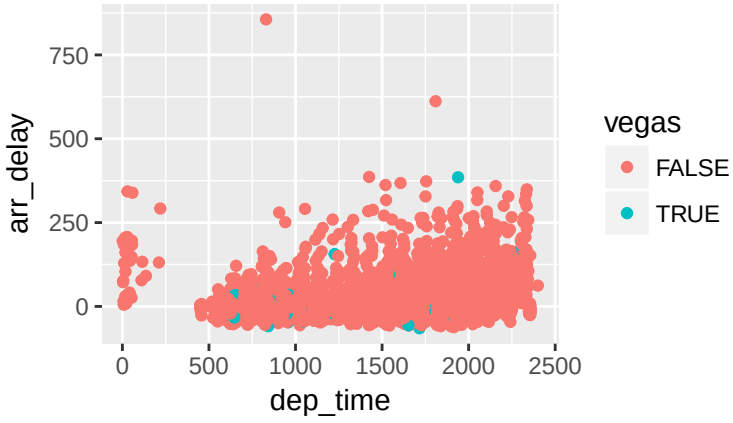
```
ggplot(data = flights) +
  geom_point(aes(x = dep_time, y = arr_delay))
```



- Describe the relationship between departure time and arrival delay.
- Identify the components of a ggplot graph

Mappings

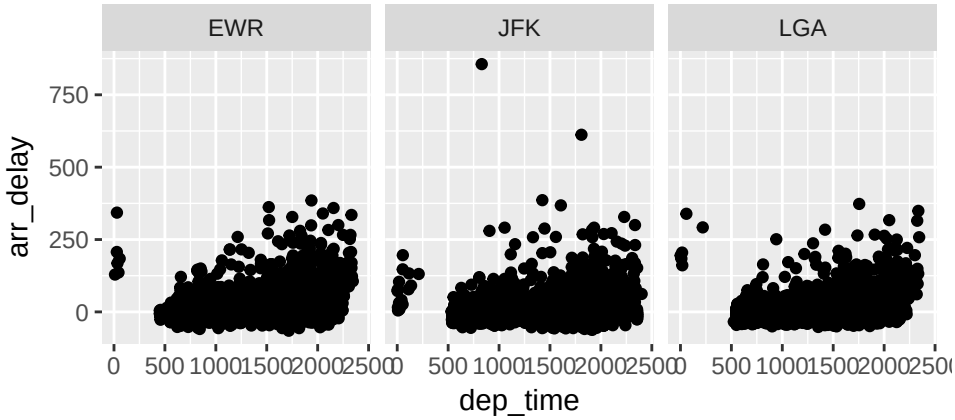
```
flights <- flights %>% mutate(vegas = (dest == "LAS"))
ggplot(data = flights) +
  geom_point(aes(x = dep_time, y = arr_delay, colour = vegas))
```



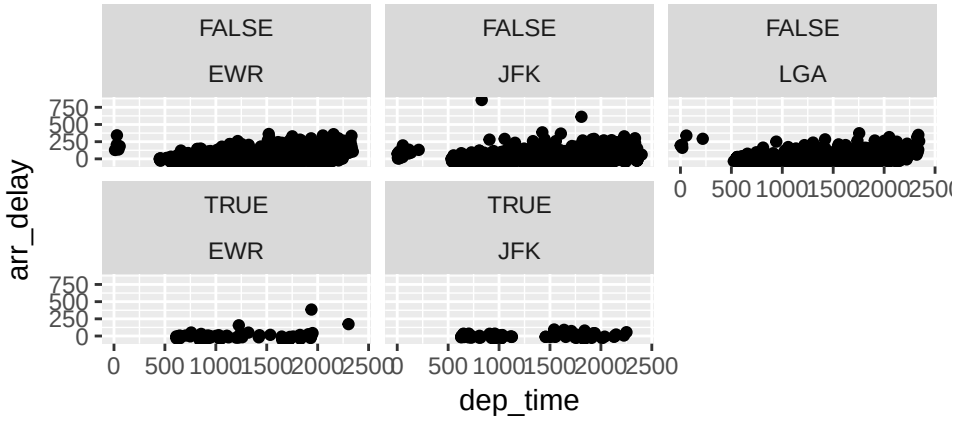
What are other mappings that we could have used?

Facets

```
ggplot(data = flights) +
  geom_point(aes(x = dep_time, y = arr_delay)) +
  facet_wrap(~origin)
```

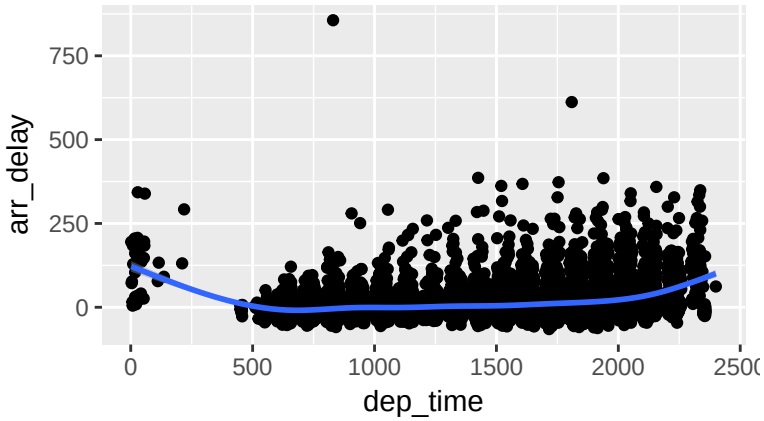



```
ggplot(data = flights) +
  geom_point(aes(x = dep_time, y = arr_delay)) +
  facet_wrap(vegas ~ origin)
```



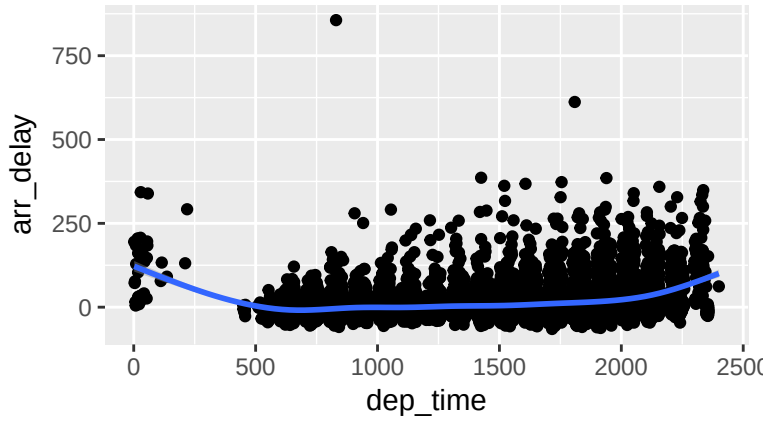
Geometric objects

```
ggplot(data = flights) +
  geom_point(aes(x = dep_time, y = arr_delay)) +
  geom_smooth(aes(x = dep_time, y = arr_delay))
```

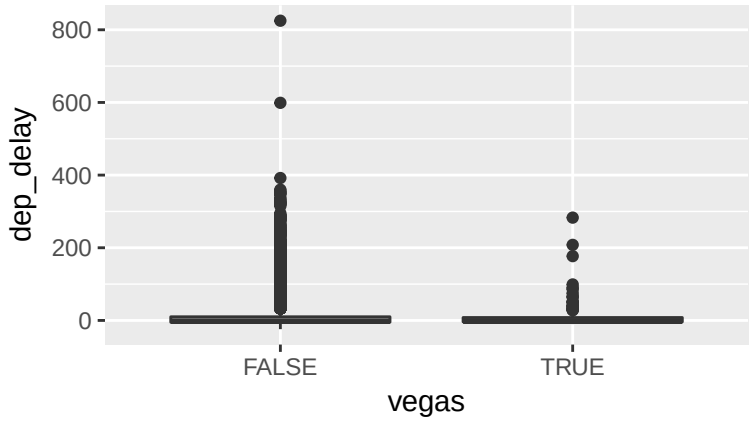



```
ggplot(data = flights, aes(x = dep_time, y = arr_delay)) +
  geom_point() +
```

```
  geom_smooth()
```




```
ggplot(data = flights, aes(x = vegas, y = dep_delay)) +
  geom_boxplot()
```




```
ggplot(data = flights, aes(x = carrier)) +
  geom_bar()
```