Lab 1: NYC Flights

October 5, 2017

Objectives

In this first lab you will:

- use an established r dataset from `nycflights13`
- explore the dataset
- create graphical summaries

Most of this lab comes directly from the R for Data Science book by Hadley Wickham and Garrett Grolemund, available at http://r4ds.had.co.nz/.

Load Data

Load the data and check out the structure of the dataset.

```
library(nycflights13)
data(flights)
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>
```

Questions

How many rows are in the dataset? `nrow(DATA)`

How many variables? `ncol(DATA)`

What type (class) of variable is the time_hour variable? `class(DATA$VARIABLE)`

> Hint: ?class to get more help.

Exploring the data

Let's dig into the dataset and explore it.

```r
# using dplyr's n_distinct is the same as length(unique(x)) from base R
# but faster and easier to use
n_distinct(flights$tailnum)
```

```
## [1] 4044
```

```r
# how many origin airports in NY?
n_distinct(flights$origin)
```

```
## [1] 3
```

```r
# what are the origin airports?
levels(as.factor(flights$origin))
```

```
## [1] "EWR" "JFK" "LGA"
```

```r
# how many destinations?
# (note: way more, probably don't want to list them all)
n_distinct(flights$dest)
```

```
## [1] 105
```

How many total flights departed from just JFK? This will use the `filter` command from `dplyr`.

Note the use of the double equals sign == which is the equivalent of "is equal to"

```r
flights %>%
  filter(origin == 'JFK') %>%
  nrow() # nrow() gives the number of rows
```

```
## [1] 111279
```

What was the average departure delay? What is the standard deviation?

```r
summarize(flights, mean = mean(dep_delay, na.rm=T), sd = sd(dep_delay, na.rm=T))
```

```
## # A tibble: 1 x 2
##       mean       sd
##      <dbl>    <dbl>
## 1 12.63907 40.21006
```

```r
# note the na.rm=T argument. This tells R to ignore (remove) NA
# values when calculating the mean and standard deviation
```

Questions

1. How many flights departed from LaGuardia (LGA) for Portland (PDX)

   Hint: You can combine filtered terms with & (and) or | (or).

```r
flights %>%
  filter(origin == 'JFK' & dest == 'PDX') %>%
  nrow()
```

```
## [1] 783
```

2. What was the average `air_time` for these flights?

```r
flights %>%
  filter(origin == 'JFK' & dest == 'PDX') %>%
  summarise(mean_air_time = mean(air_time, na.rm=T), sd_air_time = sd(air_time, na.rm=T))
```

```
## # A tibble: 1 x 2
##   mean_air_time sd_air_time
##           <dbl>       <dbl>
## 1      330.9101    16.00738
```

3. How many flights from each airport happened in July?

```
flights %>%
  filter(month == 7) %>%
  group_by(month, origin) %>%
  summarise(count = n())
```

```
## # A tibble: 3 x 3
## # Groups:   month [?]
##   month origin count
##   <int>  <chr> <int>
## 1     7    EWR 10475
## 2     7    JFK 10023
## 3     7    LGA  8927
```

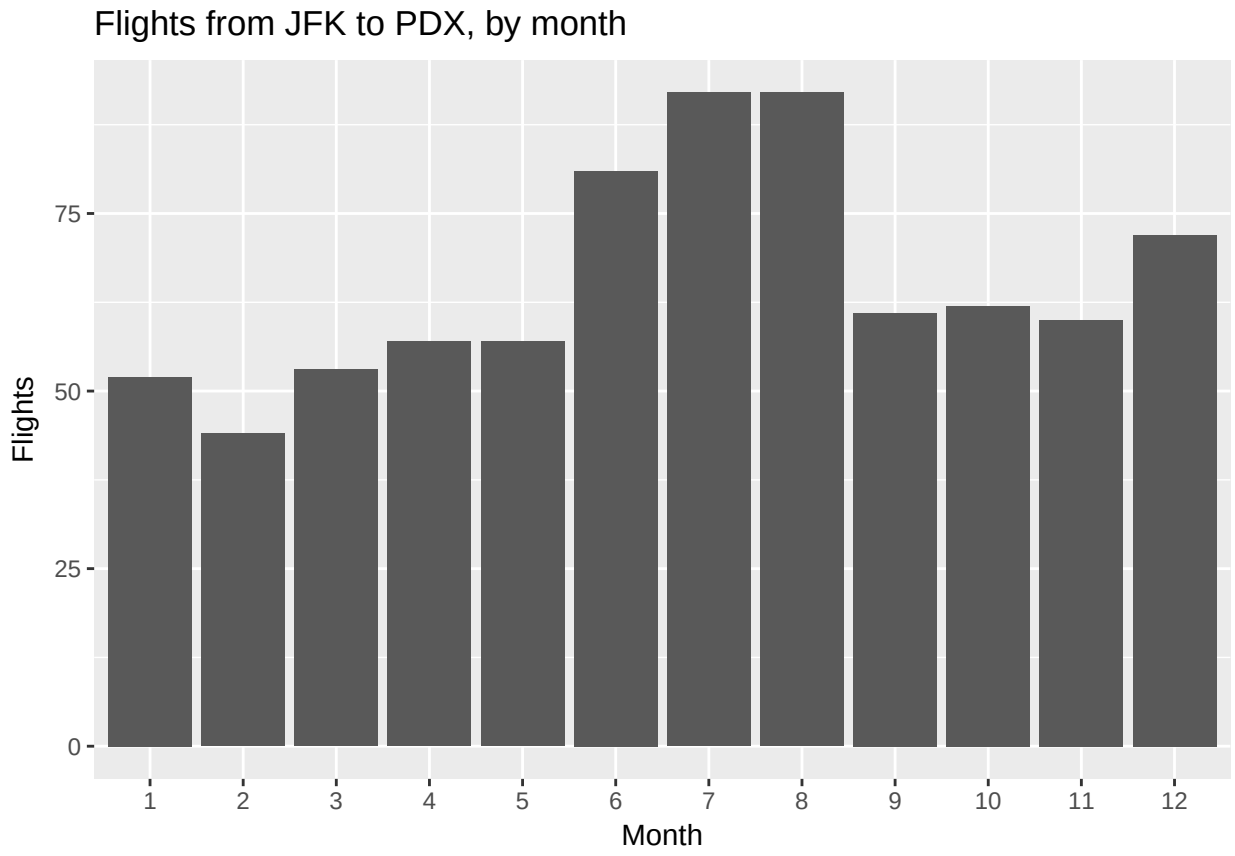
Visualizing the data

geom_histogram()

Using `ggplot2` let's create some graphical summaries of the data.

```
flights %>% filter(origin == 'JFK' & dest == 'PDX') %>%
  # note that ggplot2 doesn't use the %>% pipe operator
  # it was written before that was adopted
  # so it still uses a + sign.
  ggplot() +
  # define the geom
  geom_histogram(aes(x=factor(month)), stat='count') +
  # x-axis title
  scale_x_discrete("Month") +
  # y-axis title
  scale_y_continuous('Flights') +
  # graph title
  ggtitle('Flights from JFK to PDX, by month')
```

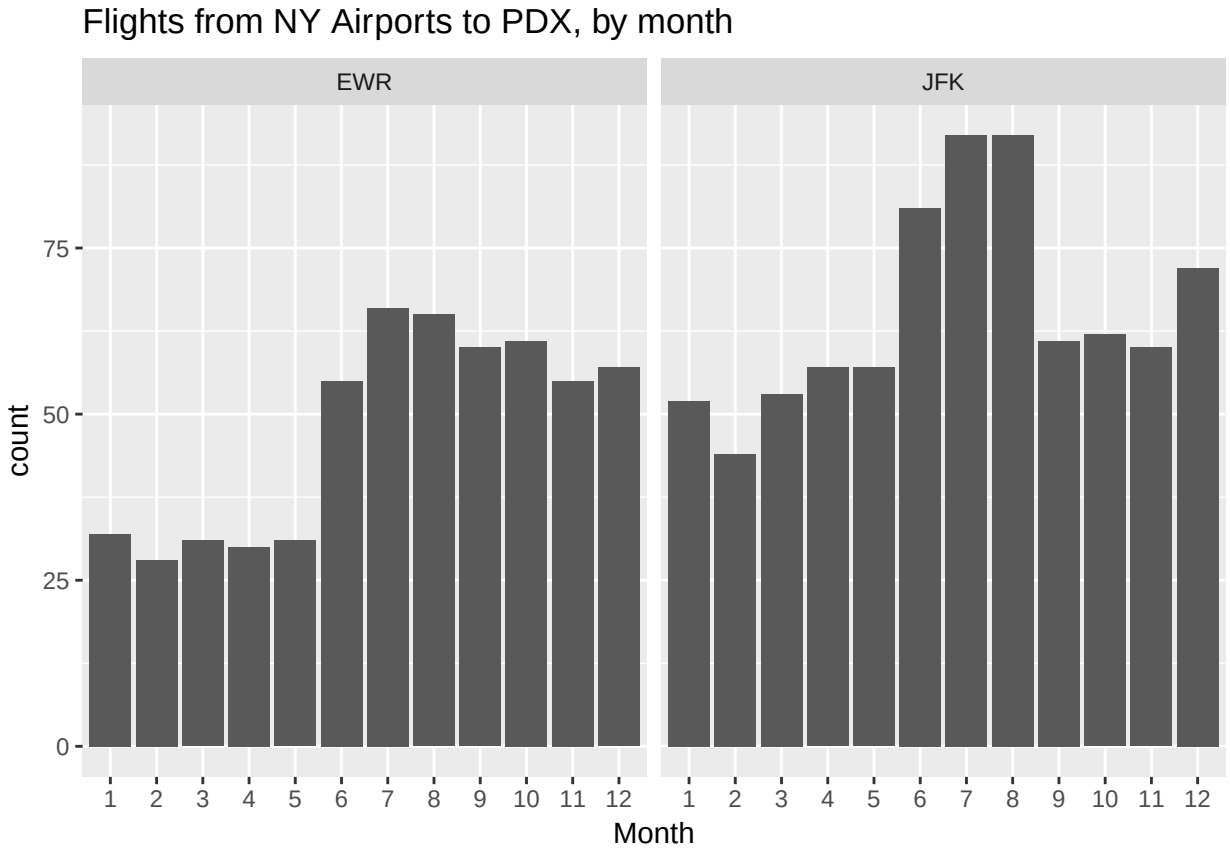
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



facet_wrap()

```
# all airports to PDX as a facet
flights %>% filter(dest=='PDX') %>%
  # note that LGA does not fly to PDX so it is automatically filtered out
  ggplot() +
  geom_histogram(aes(x=factor(month)), stat='count') +
  # facet_wrap splits the graphs up by the specified variable
  facet_wrap(~origin) +
  # title and labels as before
  scale_x_discrete('Month') +
  ggtitle('Flights from NY Airports to PDX, by month')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



geom_density()

`geom_density()` creates density plots, which can be easily overlayed to show distributions between groups.

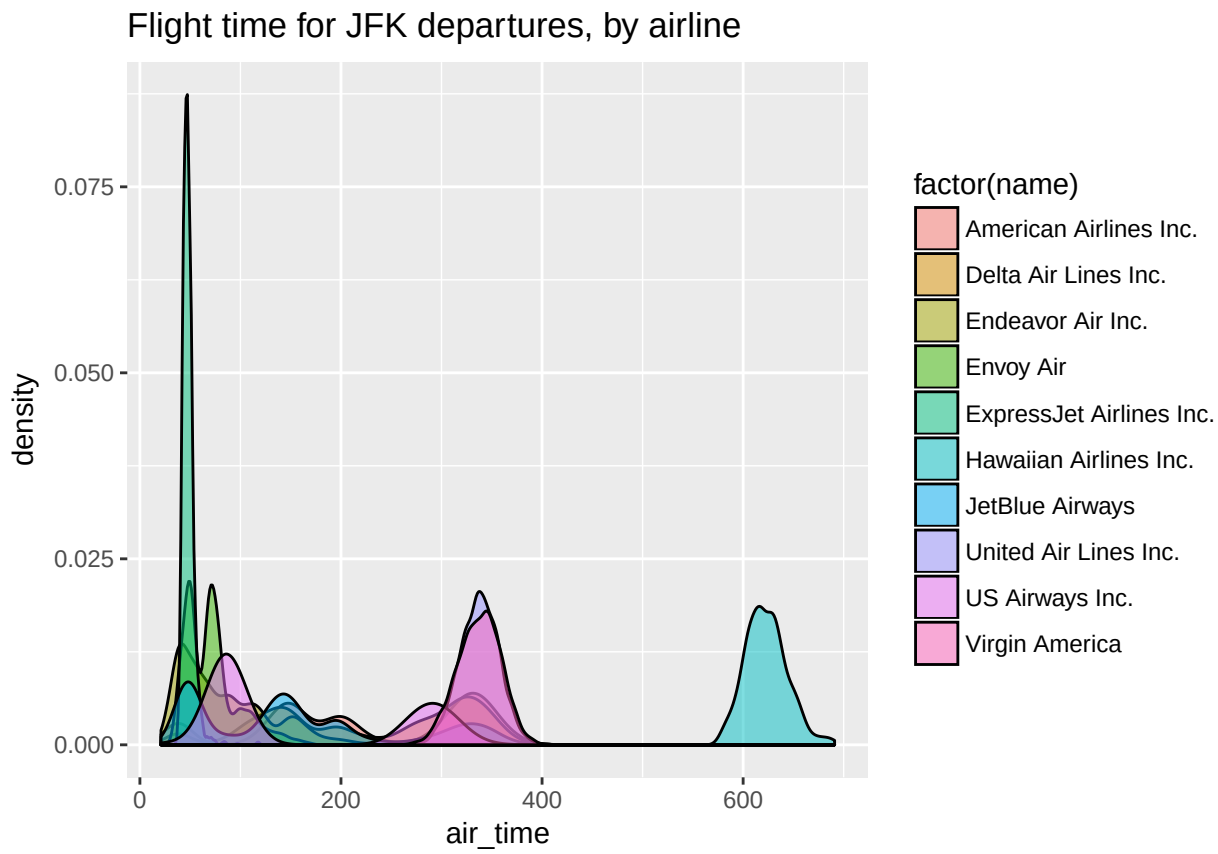
Example:

Do certain carriers fly more long or short routes? Let's look just at flights leaving `JFK`, by carrier, with respect to total `air_time`. For this, we'll need to use another dataset in the `nycflights13` package that contains the names of the airlines, instead of just the carrier codes. This will require a `left_join` (SQL users will recognize this term) on a common column between the two datasets.

```
data(airlines) # loaded from nycflights13 package

flights  %>% # start with flight data
  left_join(airlines, by="carrier") %>% # join with airlines data to get names of airlines
  subset(origin == 'JFK') %>% # only flights leaving from JFK
  ggplot(aes(x=air_time)) +  # generate ggplot object
  geom_density(stat='density', # add density plot geom
               aes(fill=factor(name)), # name is the new column we joined to our data
               alpha = .5) + # alpha sets transparency
  ggtitle('Flight time for JFK departures, by airline') # title
```

```
## Warning: Removed 2200 rows containing non-finite values (stat_density).
```

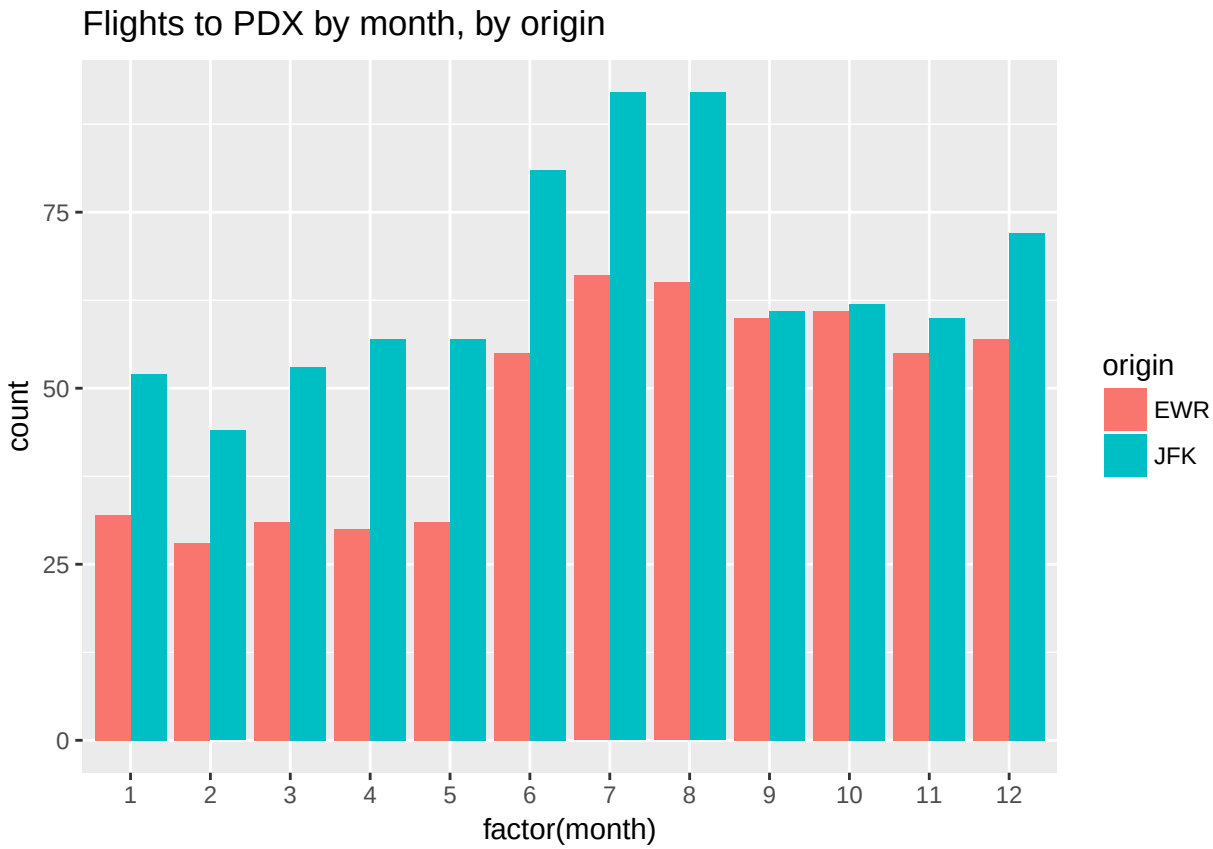


Exercises

1. Create a graph showing flights to `PDX` by month from both `EWR` and `JFK`, but as a dodged bar chart instead of a faceted one.

   Hint: position = "dodge" is the option for geom_bar()

```
flights %>% filter(dest=='PDX') %>%
  ggplot() +
  geom_bar(position = "dodge", stat = "count", aes(x=factor(month), fill=origin)) +
  ggtitle('Flights to PDX by month, by origin')
```

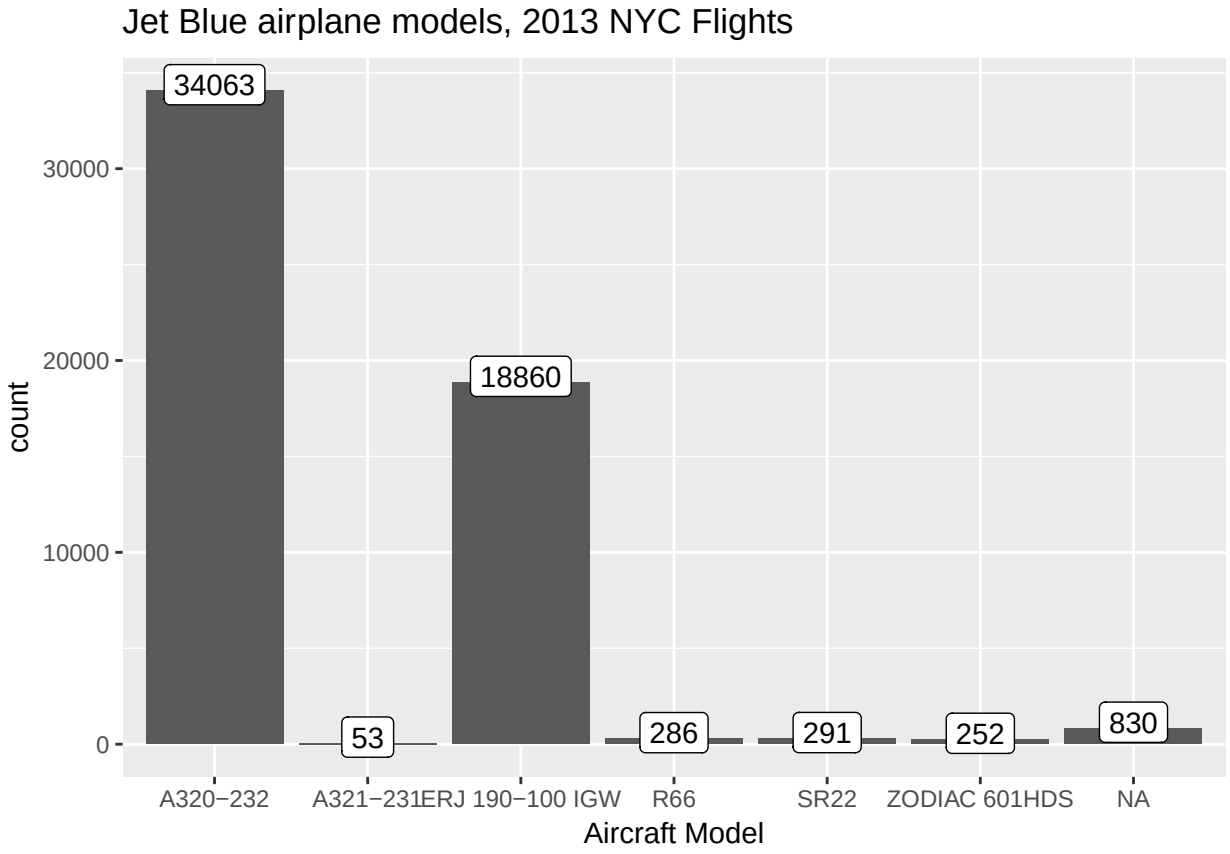


2. What types of planes, and how many of each does Jet Blue (`carrier == B6`) fly?

   Hint: You'll need to join with the `planes` dataset in the `nycflights13` package for this one.

```
data(planes) # load other dataset from nycflights13

flights %>%
  left_join(planes, by="tailnum") %>% # join with planes by tailnum
  filter(carrier == 'B6') %>% # filtered to just include Jet Blue (B6)
  ggplot() + geom_bar(stat = 'count', aes(x=factor(model))) +
  geom_label(stat='count', aes(label = ..count.., x=factor(model)), vjust=.35) +
  xlab('Aircraft Model') +
  ggtitle('Jet Blue airplane models, 2013 NYC Flights')
```


Jet Blue airplane models, 2013 NYC Flights
34063
53
18860
286
291
252
830
30000
20000
10000
0
count
A320−232
A321−231
ERJ 190−100 IGW
R66
SR22
ZODIAC 601
HDS
NA
Aircraft Model