

Lab 1: NYC Flights

Aaron C Cochran

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 Golemund, 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 <dtm>
```

Questions

How many rows are in the dataset?

How many variables?

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

Exploring the data

Let's dig into the dataset and explore it.

```
# 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
```

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

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

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

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

```
# 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”

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

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

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

```
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
```

```
# 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).
2. What was the average `air_time` for these flights?
3. How many flights from each airport happened in July?

Visualizing the data

`geom_histogram()`

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

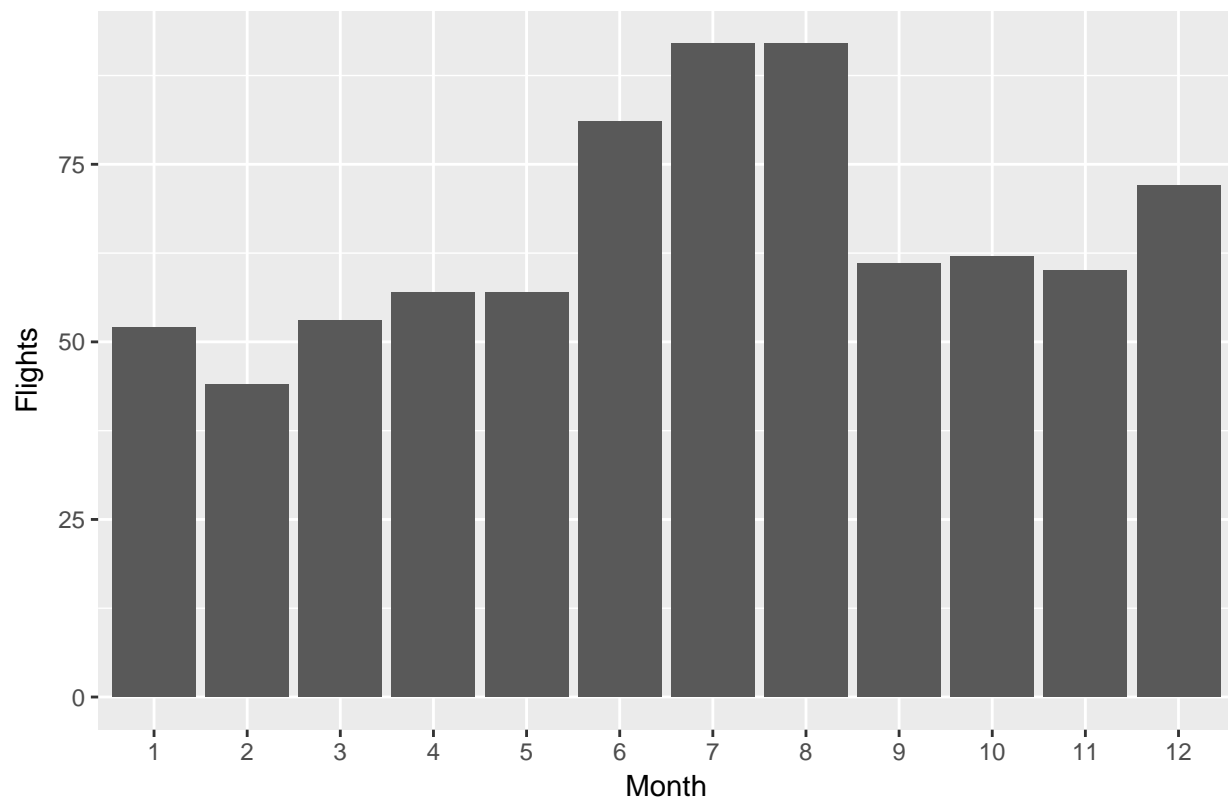
```

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

Flights from JFK to PDX, by month



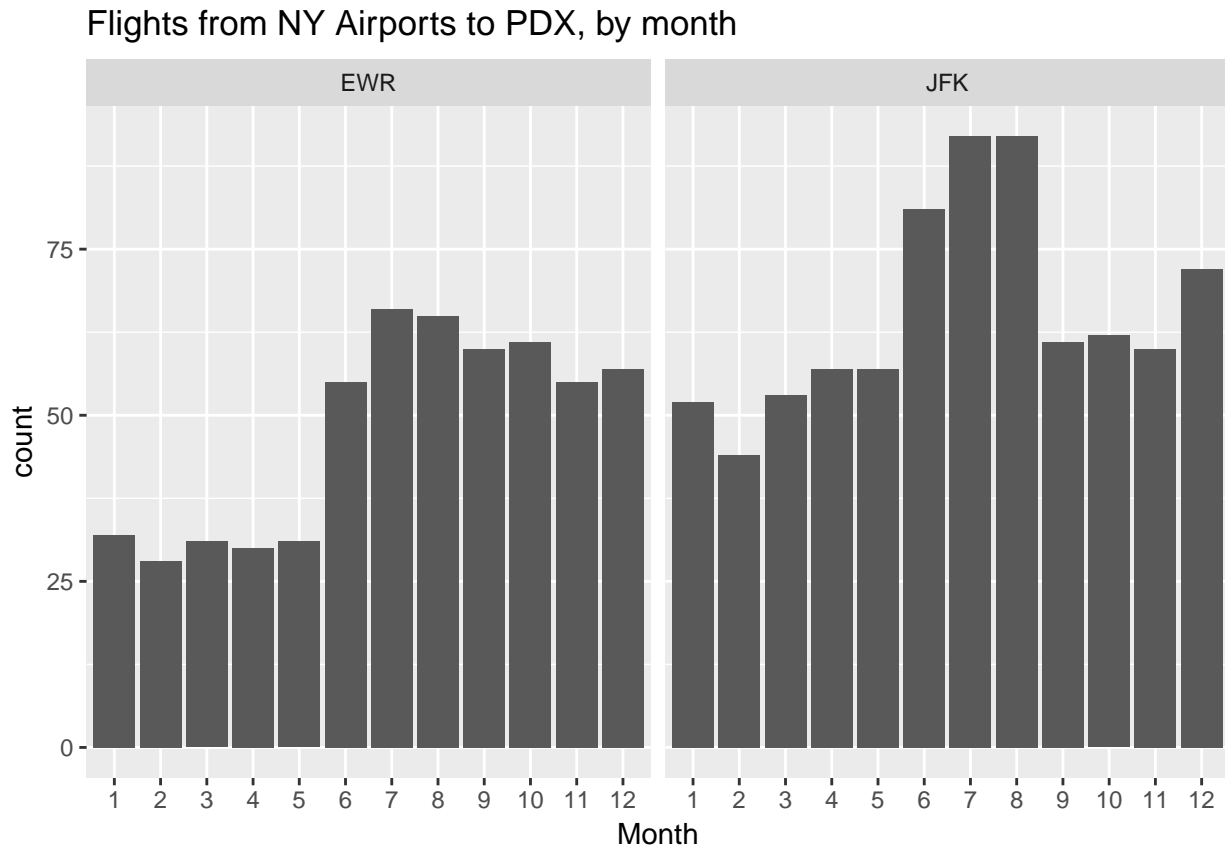
```

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:

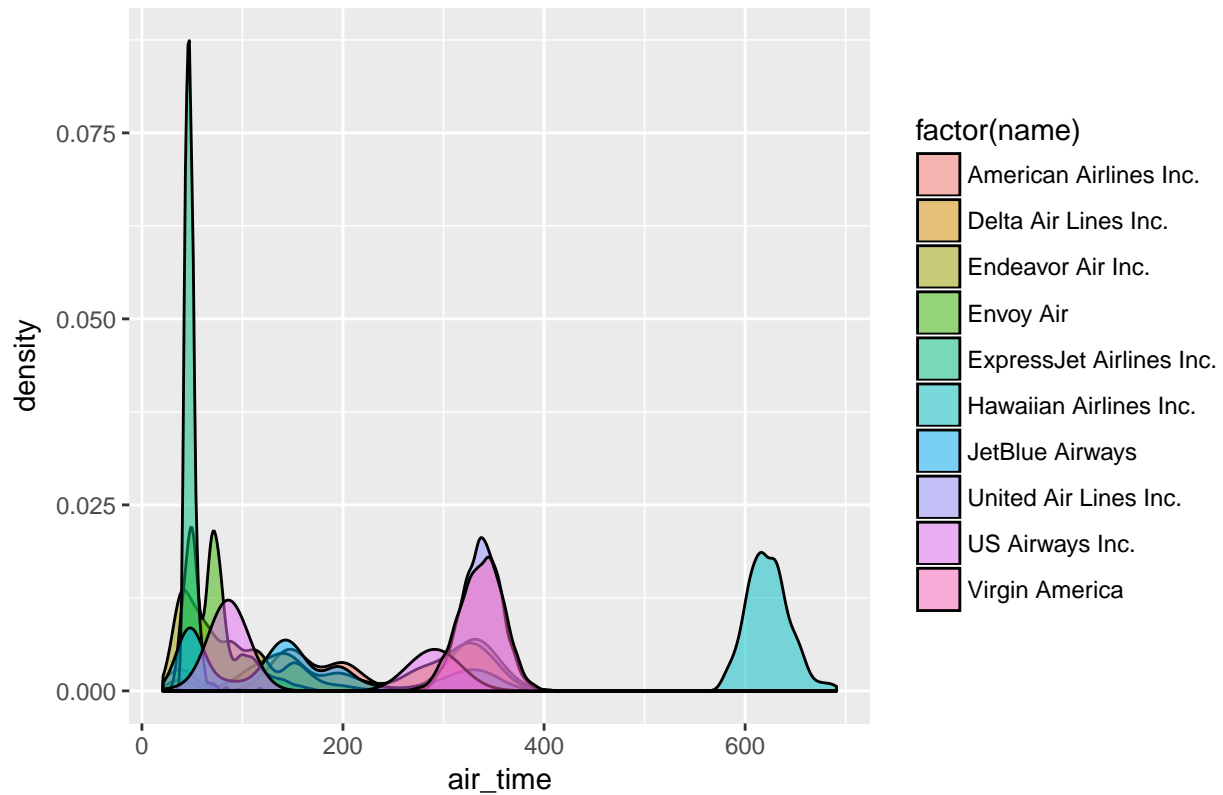
Do certain carriers fly more long or short routes? Let's look just at flights leaving JFK, by carrier, with respect to total `air_time`. For this, we'll need to use another dataset in the `nycflights13` package that contains the names of the airlines, instead of just the carrier codes. This will require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

Flight time for JFK departures, by airline



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()`

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.