EDA Beer & Alien Sightings

Fanny Chow

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```r
library(tidyr)
```

```r
# set paths to data source & read in files

setwd("~/Google Drive/stat/UFOTracker")
my.path <- "~/Google Drive/stat/UFOTracker"
beer.path <- "data/raw/brew_count_by_state_1984_2017.csv"
sightings.path <- "data/raw/ufo_sightings.csv"

beer.raw <- fread(file.path(my.path, beer.path), header=TRUE, na.strings=c("*", ""))
sightings.raw <- fread(file.path(my.path, sightings.path), header = TRUE, na.strings = c("", "Unknown",
```

```r
# clean up junk at bottom file
beer.db <- beer.raw %>%
  filter(!is.na(STATE)) %>%
  filter(STATE != "Total") %>%
  filter(STATE != "Other") %>%
  filter(STATE != "* No reportable data") %>%
  filter(STATE != "«This list will be updated quarterly.")
```

```r
# create proper data types
beer.db$STATE <- as.factor(beer.db$STATE)

# wide to long format
olddata_wide <- beer.db
keycol <- "year"
valuecol <- "breweries"
gathercols <- as.character(seq(1984, 2017))
```

```
beer.df <- gather_(olddata_wide, keycol, valuecol, gathercols)
```

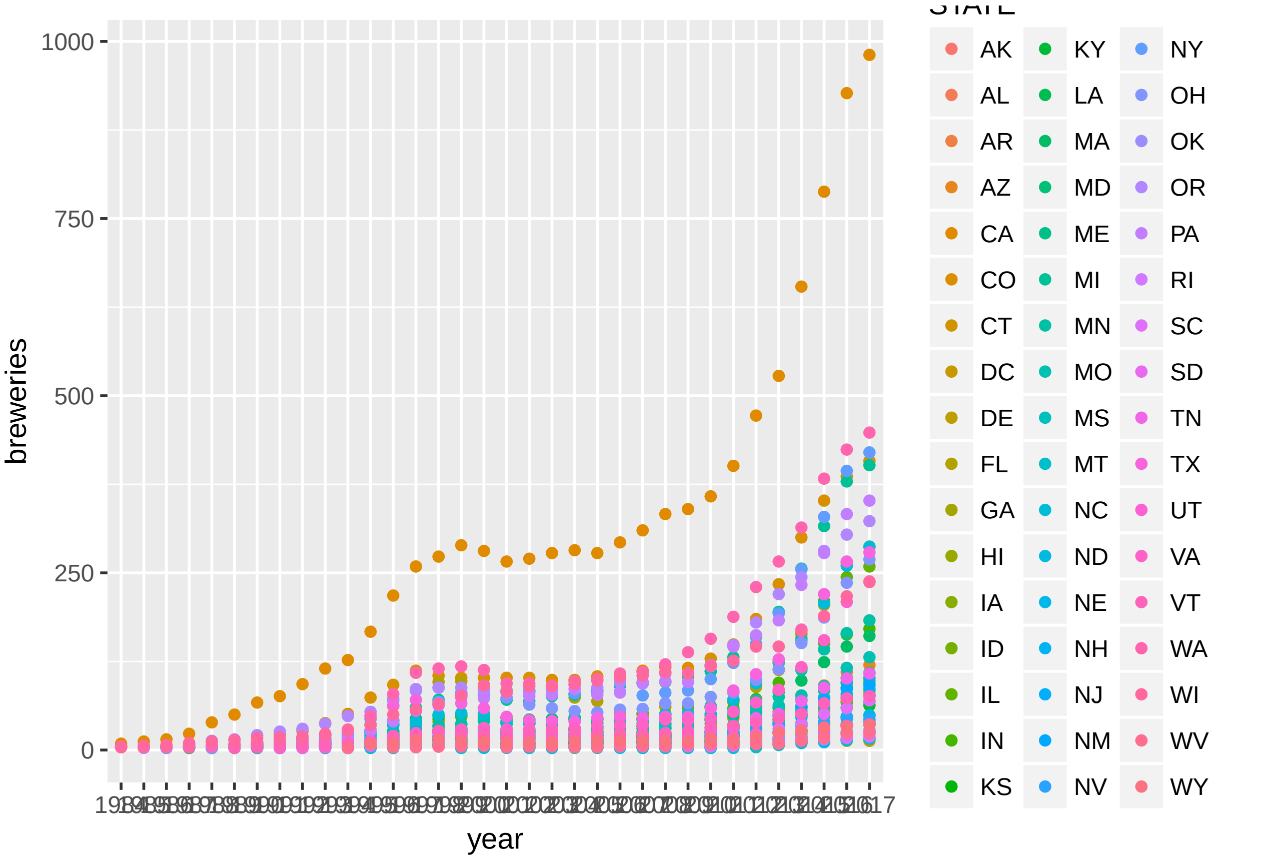
At a high-level glance, the general trend is increase number of breweries through the years for each s tate. Note that the number of breweries in 2005 will be contingent on the number of breweries in 2004, and there will be autocorrolation through years.

```
beer.df$breweries <- as.numeric(beer.df$breweries)
```

```
#breweries.year <- ggplot(beer.df, aes(x = year, y = breweries, group=1))
#breweries.year + geom_point() + geom_line(aes(color = STATE))

ggplot() +
  geom_point(data=beer.df, aes(year, breweries, color=STATE))
```

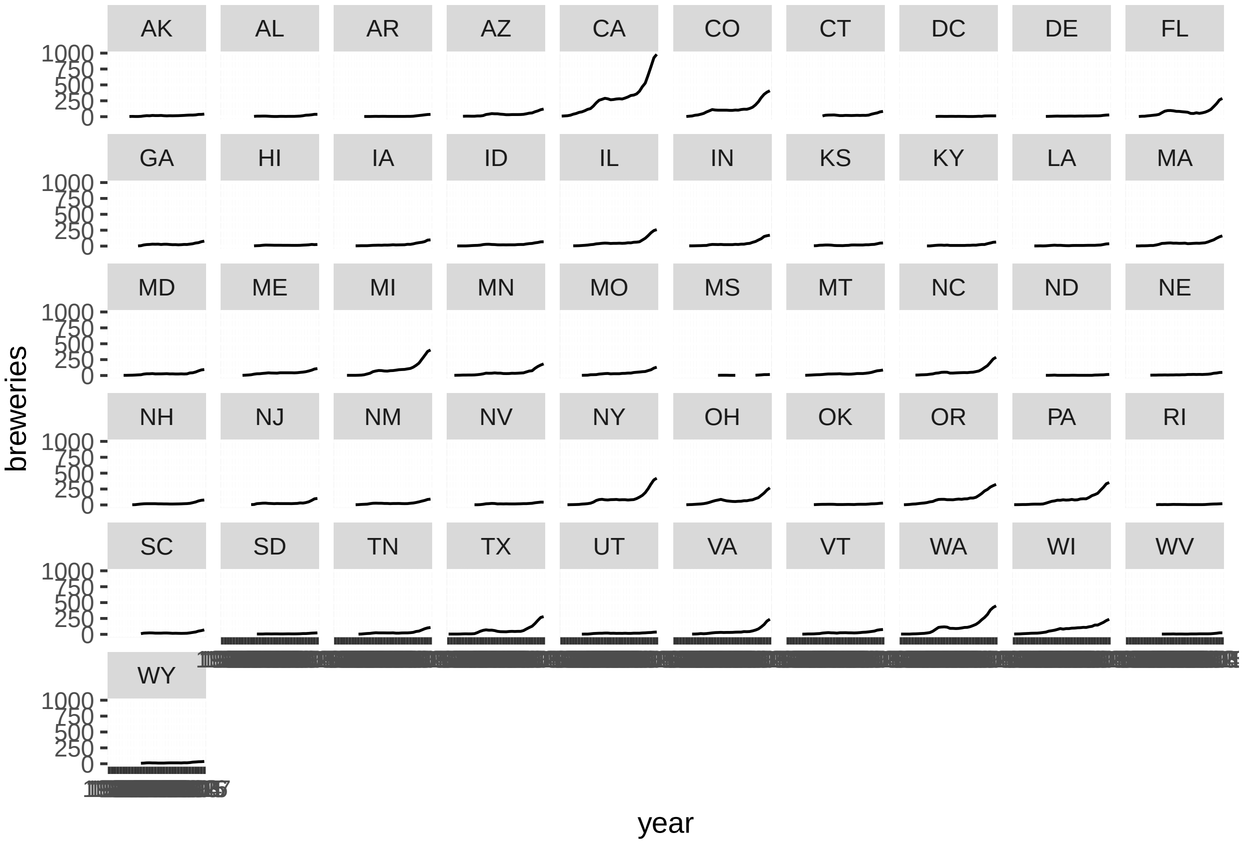
```
## Warning: Removed 350 rows containing missing values (geom_point).
```



Let's take a look at the number of breweries in each state through the years sorted by states. Since there's over 50 states we're looking at, it's challenging to discern trends from looking at all the states at once.

```
breweries.year <- ggplot(beer.df, aes(x = year, y = breweries, group=1))
(p2 <- breweries.year + geom_line() +
   facet_wrap(~STATE, ncol = 10))
```

```
## Warning: Removed 7 rows containing missing values (geom_path).
```

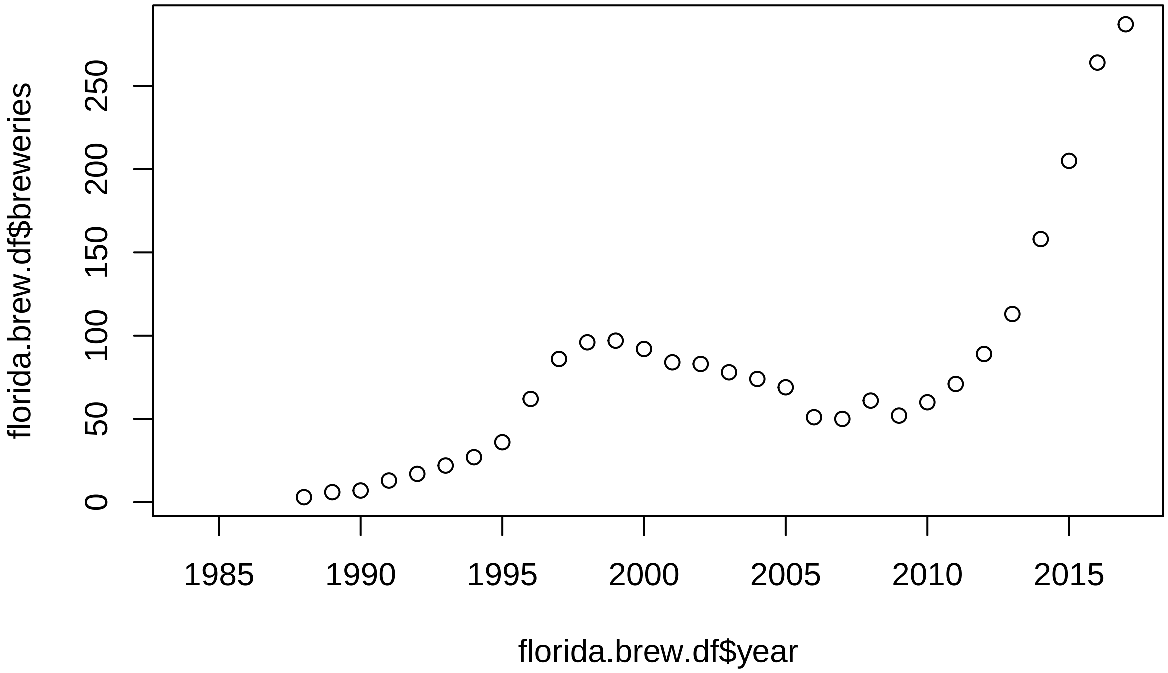


Let's focus on the state of Florida through the years. We observe an upward trend and then a sudden dip from the late 90's to 2010.

```
florida.brew.df <- beer.df %>%
  filter(STATE == "FL")
florida.brew.df
```

```
##    STATE year breweries
## 1     FL 1984        NA
## 2     FL 1985        NA
## 3     FL 1986        NA
## 4     FL 1987        NA
## 5     FL 1988         3
## 6     FL 1989         6
## 7     FL 1990         7
## 8     FL 1991        13
## 9     FL 1992        17
## 10    FL 1993        22
## 11    FL 1994        27
## 12    FL 1995        36
## 13    FL 1996        62
## 14    FL 1997        86
## 15    FL 1998        96
## 16    FL 1999        97
## 17    FL 2000        92
## 18    FL 2001        84
## 19    FL 2002        83
## 20    FL 2003        78
## 21     FL 2004        74
## 22     FL 2005        69
## 23     FL 2006        51
## 24     FL 2007        50
## 25     FL 2008        61
## 26     FL 2009        52
## 27     FL 2010        60
## 28     FL 2011        71
## 29     FL 2012        89
## 30     FL 2013       113
## 31     FL 2014       158
## 32     FL 2015       205
## 33     FL 2016       264
## 34     FL 2017       287
```

```
plot(florida.brew.df$year, florida.brew.df$breweries)
```



Since the range of time in the breweries data is from 1984-2017, let's subset the equivalent years from the sightings data.

```
# clean up sightings data
fl.sightings <- sightings.raw %>%
  filter(state == 'FL') %>%
  mutate(year = as.numeric(format(as.Date(date_time, format="%m/%d/%y"),"%Y"))) %>%
  filter(year >= 1984) %>%
  filter(year <= 2017) %>%
  #group_by(year) %>%
  count(year) %>%
  rename(sightings_year = n)
```

```
## Warning in strptime(x, format, tz = "GMT"): unknown timezone 'zone/tz/
## 2017c.1.0/zoneinfo/America/Los_Angeles'
```

```
 # mutate(sightings_year = n())
#group_by(`Student ID`) %>%
 # mutate(`Dupe Check`= n())

fl.sightings
```

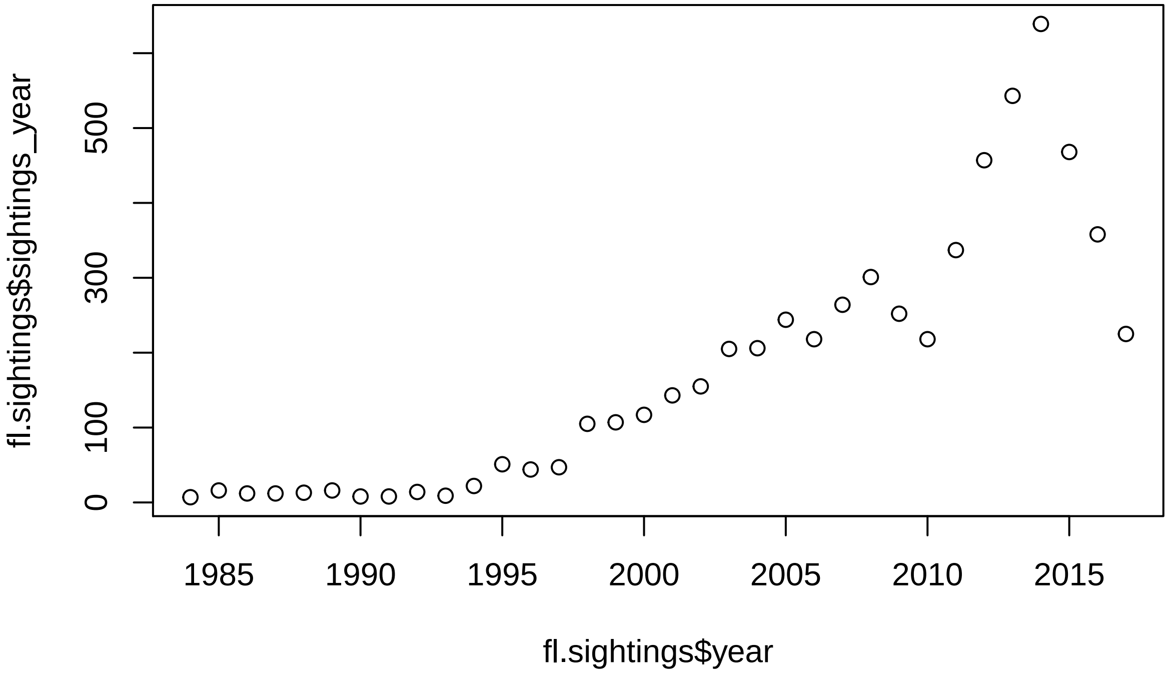
```
## # A tibble: 34 x 2
##     year sightings_year
##    <dbl>          <int>
##  1  1984              7
##  2  1985             16
##  3  1986             12
##  4  1987             12
##  5  1988             13
##  6  1989             16
##  7  1990              8
##  8  1991              8
##  9  1992             14
## 10  1993              9
## # ... with 24 more rows
```

```
#fl.sightings$year <- as.numeric(format(as.Date(fl.sightings$date_time, format="%m/%d/%y"),"%Y"))
#filter(fl.sightings, year > 1983)


#fl.sightings.count <- as.data.frame(table(fl.sightings$year))
#colnames(fl.sightings.count) <- c("year","sightings_per_year")
#as.numeric(fl.sightings.count$year)
#str(fl.sightings.count)
```

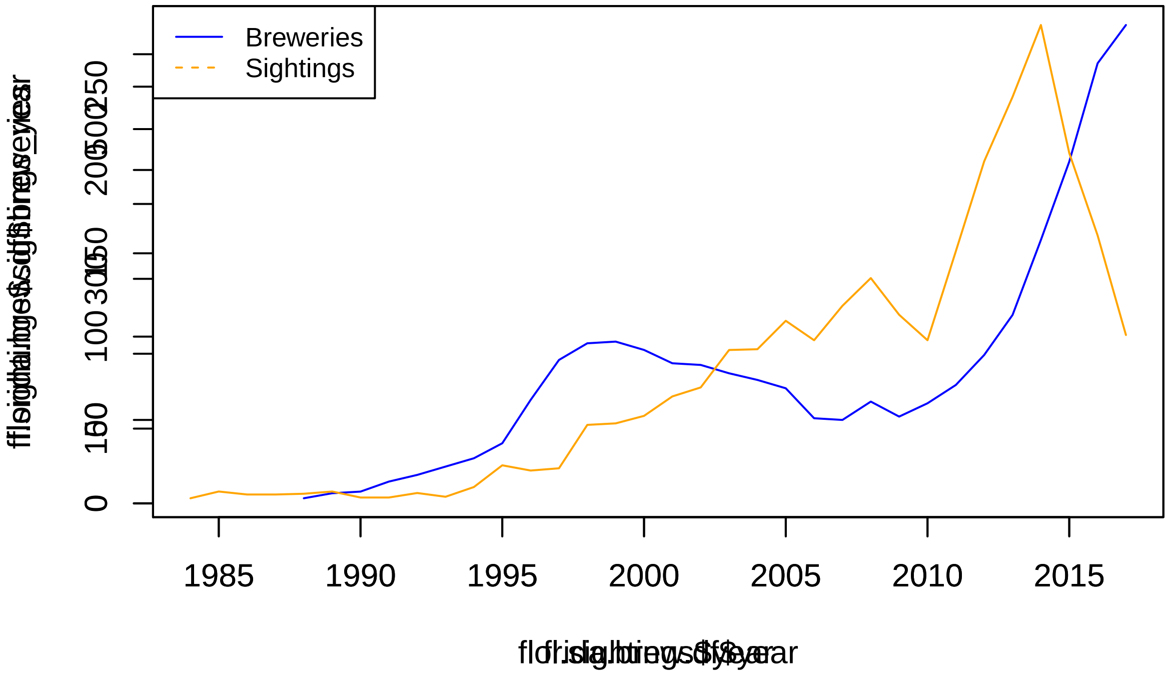
Let's take a snapshot of sightings per year in Florida.

```
plot(fl.sightings$year, fl.sightings$sightings_year)
```



Let's compare the 2 plots at once. Interesting how the 2 plots follow the same shape until the early 2000s and then diverge drastically aftewards.

```r
plot(florida.brew.df$year, florida.brew.df$breweries, type = "l", col="blue")
par(new=TRUE)
plot(fl.sightings$year, fl.sightings$sightings_year, type="l", col="orange")
legend("topleft",legend=c("Breweries", "Sightings"),
       col=c("blue", "orange"), lty=1:2, cex=0.8)
```


Breweries
Sightings
florida.brew.df$year
fl.sightings$year
florida.brew.df$breweries
fl.sightings$sightings_year