While working with the dataset to plan my learning sessions, I started playing around and thought it would be fun to show the various shapes of UFOs reported over time, to see if there were any shifts. Spoiler: There were. But I needed to clean the data a bit first.

```
setwd("~/Downloads/UFO Data")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --
```

```
## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
options(scipen = 999)
```

```
UFOs <- read_csv("UFOsightings.csv", col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   datetime = col_character(),
##   city = col_character(),
##   state = col_character(),
##   country = col_character(),
##   shape = col_character(),
##   `duration (seconds)` = col_double(),
##   `duration (hours/min)` = col_character(),
##   comments = col_character(),
##   `date posted` = col_character(),
##   latitude = col_double(),
##   longitude = col_double()
## )
```

```
## Warning: 4 parsing failures.
##   row                col               expected   actual               file
## 27823 duration (seconds) no trailing characters `        'UFOsightings.csv'
## 35693 duration (seconds) no trailing characters `        'UFOsightings.csv'
## 43783 latitude           no trailing characters q.200088 'UFOsightings.csv'
## 58592 duration (seconds) no trailing characters `        'UFOsightings.csv'
```

There are 30 shapes represented in the data. That's a lot to show in a single figure.

```
UFOs %>%
  summarise(shapes = n_distinct(shape))
```

```
## # A tibble: 1 x 1
##   shapes
##
## 1     30
```

If we look at the different shapes in the data, we can see some overlap, as well as shapes with low counts.

```
UFOs %>%
```

```
  group_by(shape) %>%
  summarise(count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 30 x 2
##    shape    count
##
##  1 changed      1
##  2 changing  1962
##  3 chevron    952
##  4 cigar     2057
##  5 circle    7608
##  6 cone       316
##  7 crescent     2
##  8 cross      233
##  9 cylinder  1283
## 10 delta        7
## # ... with 20 more rows
```

For instance, “changed” only appears in one record. But “changing,” which appears in 1,962 records should be grouped with “changed.” After inspecting all the shapes, I identified the following categories that accounted for most of the different shapes:

- changing, which includes both changed and changing
- circles, like disks, domes, and spheres
- triangles, like deltas, pyramids, and triangles
- four or more sided: rectangles, diamonds, and chevrons
- light, which counts things like flares, fireballs, and lights

I also made an “other” category for shapes with very low counts that didn’t seem to fit in the categories above, like crescents, teardrops, and formations with no further specification of shape. Finally, shape was blank for some records, so I made an “unknown” category. Here’s the code I used to recategorize shape.

```
changing <- c("changed", "changing")
circles <- c("circle", "disk", "dome", "egg", "oval","round", "sphere")
triangles <- c("cone","delta","pyramid","triangle")
fourormore <- c("chevron","cross","diamond","hexagon","rectangle")
light <- c("fireball","flare","flash","light")
other <- c("cigar","cylinder","crescent","formation","other","teardrop")
unknown <- c("unknown", 'NA')

UFOs <- UFOs %>%
  mutate(shape2 = ifelse(shape %in% changing,
                         "changing",
                         ifelse(shape %in% circles,
                                "circular",
                                ifelse(shape %in% triangles,
                                       "triangular",
                                       ifelse(shape %in% fourormore,
                                              "four+-sided",
                                              ifelse(shape %in% light,
                                                     "light",
                                                     ifelse(shape %in% other,

"other","unknown")))))))
```

My biggest question mark was cigar and cylinder. They’re not really circles, nor do they fall in the four or more sided category. I could create another category called “tubes,” but ultimately just put them in other.

Using the code above as an example, you could see what happens to the chart if you put them in another category or create one of their own.

For the chart, I dropped the unknowns.

```
UFOs <- UFOs %>%
  filter(shape2 != "unknown")
```

Now, to plot shapes over time, I need to extract data information. The “datetime” variable is currently a character, so I have to convert that to a date. I then pulled out year, so that each point on my figure was the count of that shape observed during a given year.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
UFOs <- UFOs %>%
  mutate(Date2 = as.Date(datetime, format = "%m/%d/%Y"),
         Year = year(Date2))
```

Now we have all the information we need to plot shapes over time, to see if there have been changes. We’ll create a summary dataframe by Year and shape2, then create a line chart with that information.

```
Years <- UFOs %>%
  group_by(Year, shape2) %>%
  summarise(count = n())
```

```
## `summarise()` regrouping output by 'Year' (override with `.groups` argument)
```

```
library(scales)
```

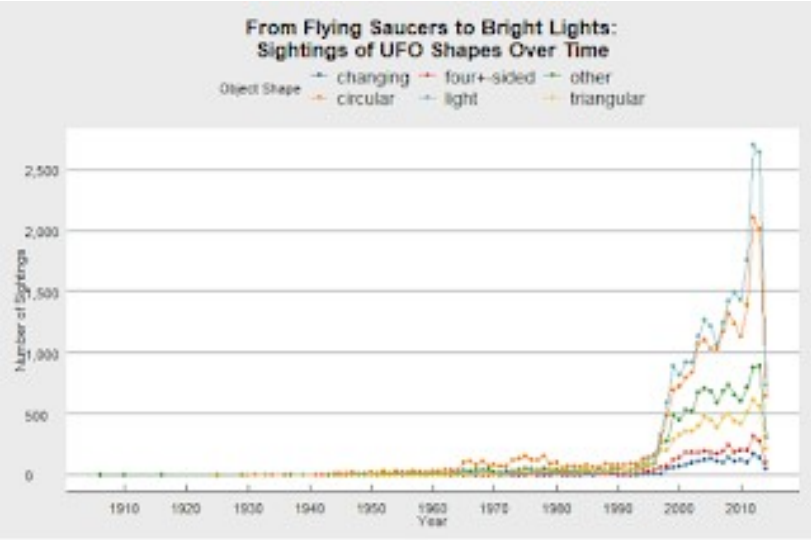
```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(ggthemes)
```

```
Years %>%
  ggplot(aes(Year, count, color = shape2)) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(1910,2020,10)) +
  scale_y_continuous(breaks = seq(0,3000,500), labels = comma) +
  labs(color = "Object Shape", title = "From Flying Saucers to Bright
Lights:\nSightings of UFO Shapes Over Time") +
  ylab("Number of Sightings") +
  theme_economist_white() +
  scale_color_tableau() +
  theme(plot.title = element_text(hjust = 0.5))
```



Until the mid-90s, the most commonly seen UFO was circular. After that, light shapes became much more common. I'm wondering if this could be explained in part by UFOs in pop culture, moving from the flying saucers of earlier sci-fi to the bright lights without discernible shape in the more recent sci-fi. The third most common shape is our "other" category, which suggests we might want to rethink that one. It could be that some of the shapes within that category are common enough to warrant their own category, while receiving other for those that don't have a good category of their own. Cigar and cylinder, for instance, have high counts and could be put in their own category. Feel free to play around with the data and see what you come up with!