2020 has been a tough year, and I've been doing my best to keep busy (and distracted from all the insanity – both at the personal and worldwide levels). Earlier this year, I took a course in machine learning techniques and have been working on applying those techniques to work datasets, as well as fun sets through Kaggle.com.

Today, I thought I'd share another dataset I discovered through Kaggle: TV shows available on one or more streaming service (Netflix, Hulu, Prime, and Disney+). There are lots of fun things we could do with this dataset. Let's start with some basic visualization and summarization.

```
setwd("~/Dropbox")

library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────
──────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.0     ✓ purrr   0.3.4
## ✓ tibble  3.0.0     ✓ dplyr   0.8.5
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

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

```
Shows <- read_csv("tv_shows.csv")
```

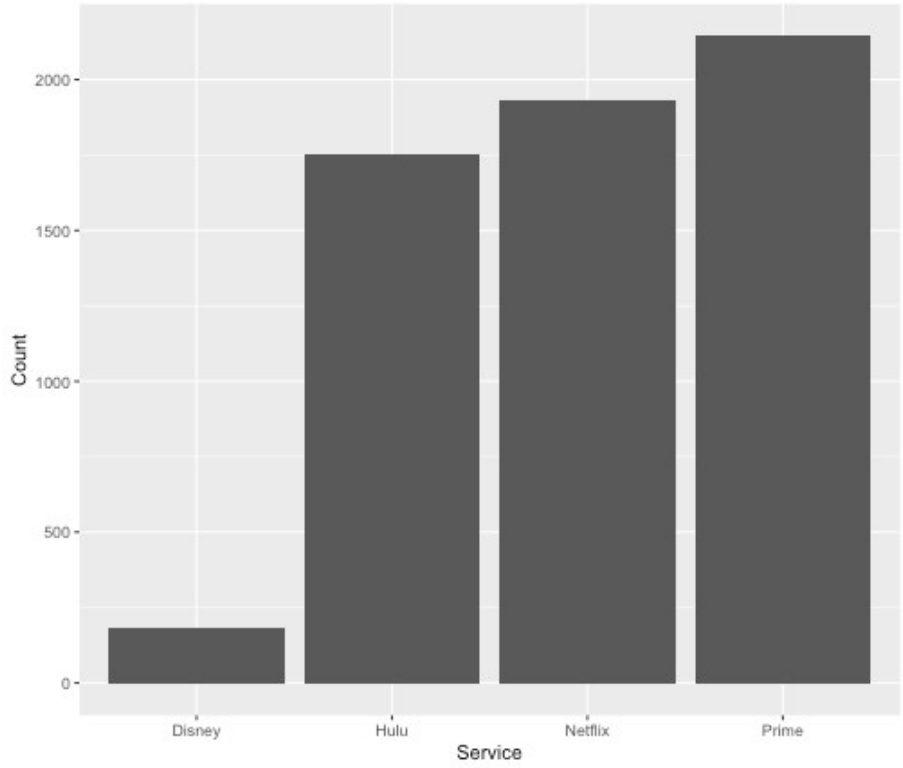
```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   Title = col_character(),
##   Year = col_double(),
##   Age = col_character(),
##   IMDb = col_double(),
##   `Rotten Tomatoes` = col_character(),
##   Netflix = col_double(),
##   Hulu = col_double(),
##   `Prime Video` = col_double(),
##   `Disney+` = col_double(),
##   type = col_double()
## )
```

First, we can do some basic summaries, such as how many shows in the dataset are on each of the streaming services.

```
Counts <- Shows %>%
  summarise(Netflix = sum(Netflix),
            Hulu = sum(Hulu),
            Prime = sum(`Prime Video`),
            Disney = sum(`Disney+`)) %>%
  pivot_longer(cols = Netflix:Disney,
               names_to = "Service",
               values_to = "Count")

Counts %>%
  ggplot(aes(Service,Count)) +
```

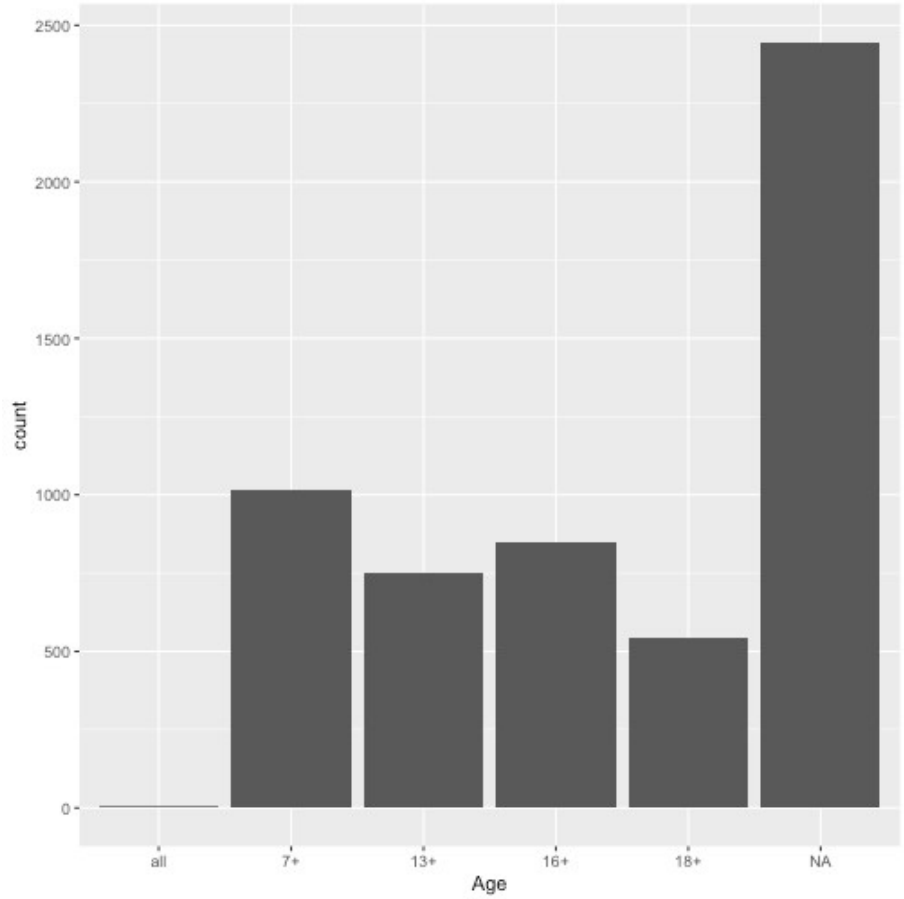
```
  geom_col()
```




The biggest selling point of Disney+ is to watch their movies, though the few TV shows they offer can't really be viewed elsewhere (e.g., The Mandalorian). For the sake of simplicity, we'll drop Disney+, and focus on the big 3 services for TV shows.

The dataset also contains an indicator of recommended age, which we can plot.

```
Shows <- Shows %>%
  mutate(Age = factor(Age,
                      labels = c("all",
                                 "7+",
                                 "13+",
                                 "16+",
                                 "18+"),
                      ordered = TRUE))

Shows %>%
  ggplot(aes(Age)) +
  geom_bar()
```




Many are 'NA' for age, though it isn't clear why. Are these older shows, added before these streaming services were required to add guidance on these issues? Is this issue seen more for a particular streaming site? Let's find out

```
Shows %>%
  group_by(Age) %>%
  summarise(Count = n(),
            Year_min = min(Year),
            Year_max = max(Year),
            Prime = sum(`Prime Video`)/2144,
            Netflix = sum(Netflix)/1931,
            Hulu = sum(Hulu)/1754)
```

```
## Warning: Factor `Age` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## # A tibble: 6 x 7
##   Age   Count Year_min Year_max    Prime Netflix   Hulu
##
## 1 all       4     1995     2003 0.000466 0.00155 0
## 2 7+     1018     1955     2020 0.0975   0.206   0.293
## 3 13+     750     1980     2020 0.0849   0.186   0.136
## 4 16+     848     1943     2020 0.104    0.155   0.208
## 5 18+     545     1932     2020 0.0896   0.0886  0.0906
## 6    2446     1901     2020 0.623    0.363   0.272
```

It seems the biggest "offender" for missing age information is Prime – about 62% of the shows don't have an age indicator. More surprising, though, is the minimum year for some of these categories. I'm no expert in the history of TV, but I don't think any shows were being broadcast in 1901. What are these outliers?

```
YearOutliers <- Shows %>%
  filter(Year < 1940)
```

```
list(YearOutliers$Title)
```

```
## [[1]]
## [1] "Born To Explore"                    "The Three Stooges"
## [3] "The Little Rascals Classics"        "Space: The New Frontier"
## [5] "Gods & Monsters with Tony Robinson" "History of Westinghouse"
## [7] "Betty Boop"
```

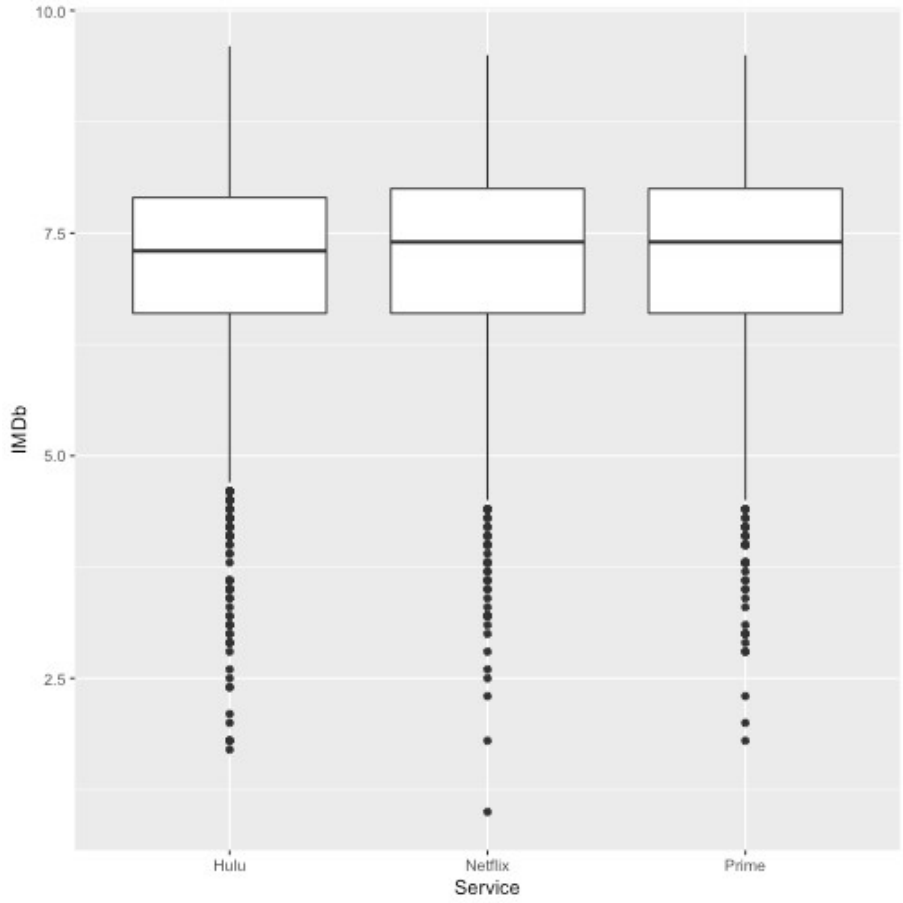
Four of these entries are clearly in error – these are newer shows. This isn't important at the moment, but it's interesting nonetheless.

In terms of getting the most "bang for your buck," Amazon Prime has the most shows to offer (though if you're looking for data on recommended age for the TV show, Prime has the most missingness). But Hulu and Netflix, in terms of volume, are pretty comparable to Prime. What can be said about the quality of content on each of the 3?

The dataset offers some indicators of quality: IMDb rating and Rotten Tomatoes score. How do the 3 services measure up on these indicators?

```
Netflix <- Shows %>%
  filter(Netflix == 1) %>%
  select(IMDb, `Rotten Tomatoes`) %>%
  mutate(Service = "Netflix")

Hulu <- Shows %>%
  filter(Hulu == 1) %>%
  select(IMDb, `Rotten Tomatoes`) %>%
  mutate(Service = "Hulu")

Prime <- Shows %>%
  filter(`Prime Video` == 1) %>%
  select(IMDb, `Rotten Tomatoes`) %>%
  mutate(Service = "Prime")

BigThree <- rbind(Netflix, Hulu, Prime)

BigThree <- BigThree %>%
  mutate(RotTom = as.numeric(sub("%","",`Rotten Tomatoes`))/100)

BigThree %>%
  ggplot(aes(Service, IMDb)) +
  geom_boxplot()
```

```
## Warning: Removed 1194 rows containing non-finite values (stat_boxplot).
```




```
library(scales)
```

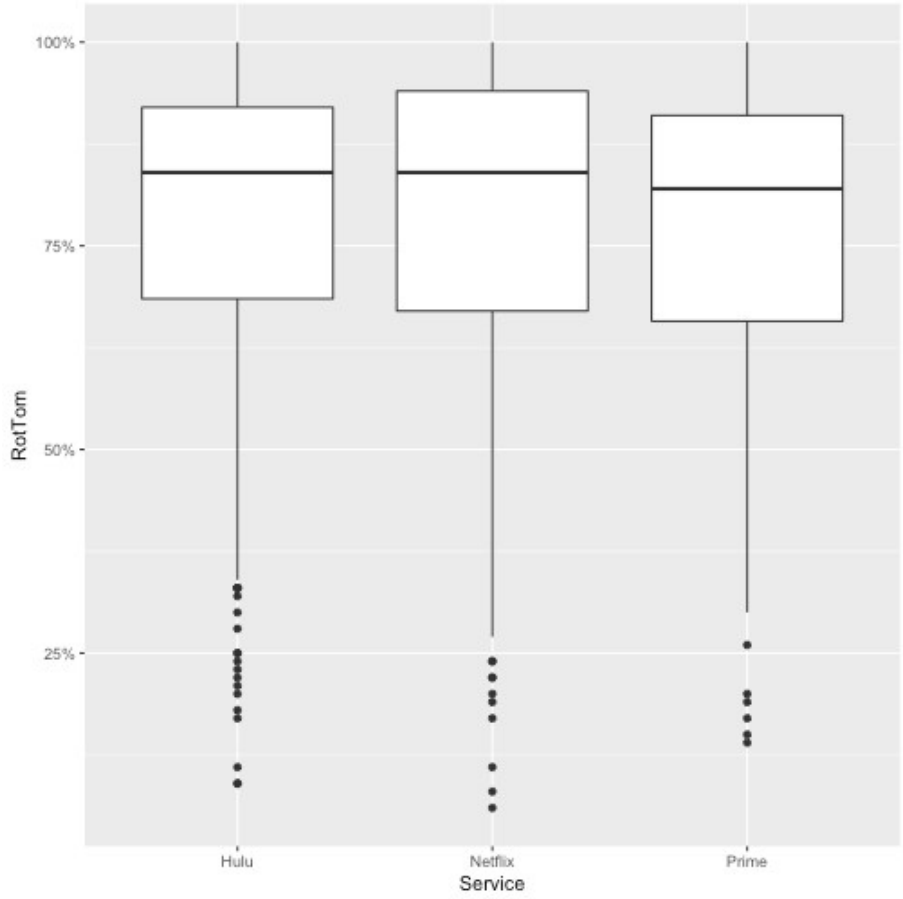
```
##
## Attaching package: 'scales'
```

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

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

```
BigThree %>%
  ggplot(aes(Service, RotTom)) +
  geom_boxplot() +
  scale_y_continuous(labels = percent)
```

```
## Warning: Removed 4772 rows containing non-finite values (stat_boxplot).
```




It doesn’t appear the 3 streaming services differ too much in terms of quality. But there’s more analysis we can do of this dataset. More later.