Capstone_Statistics

Jennifer Klein

3/13/2018

Now that my dataset is all cleaned up and properly wrangled, it's now ready for analysis.

Some background on my dataset: it features 20 podcasts and 199 episodes. The shortest episode clocked in at a mere 15 minutes, while the longest was a whopping 6 hours. The most recent podcast episode was published August 31, 2017 (the day I pulled my data) and an episode of Dan. Carlin's Hardcore History from October 30, 2013 was still hanging in there with a 3 ranking.

I decided to create a histogram, a scatterplot and a time-series plot to try to gain insight into my data.

Creating The Histogram

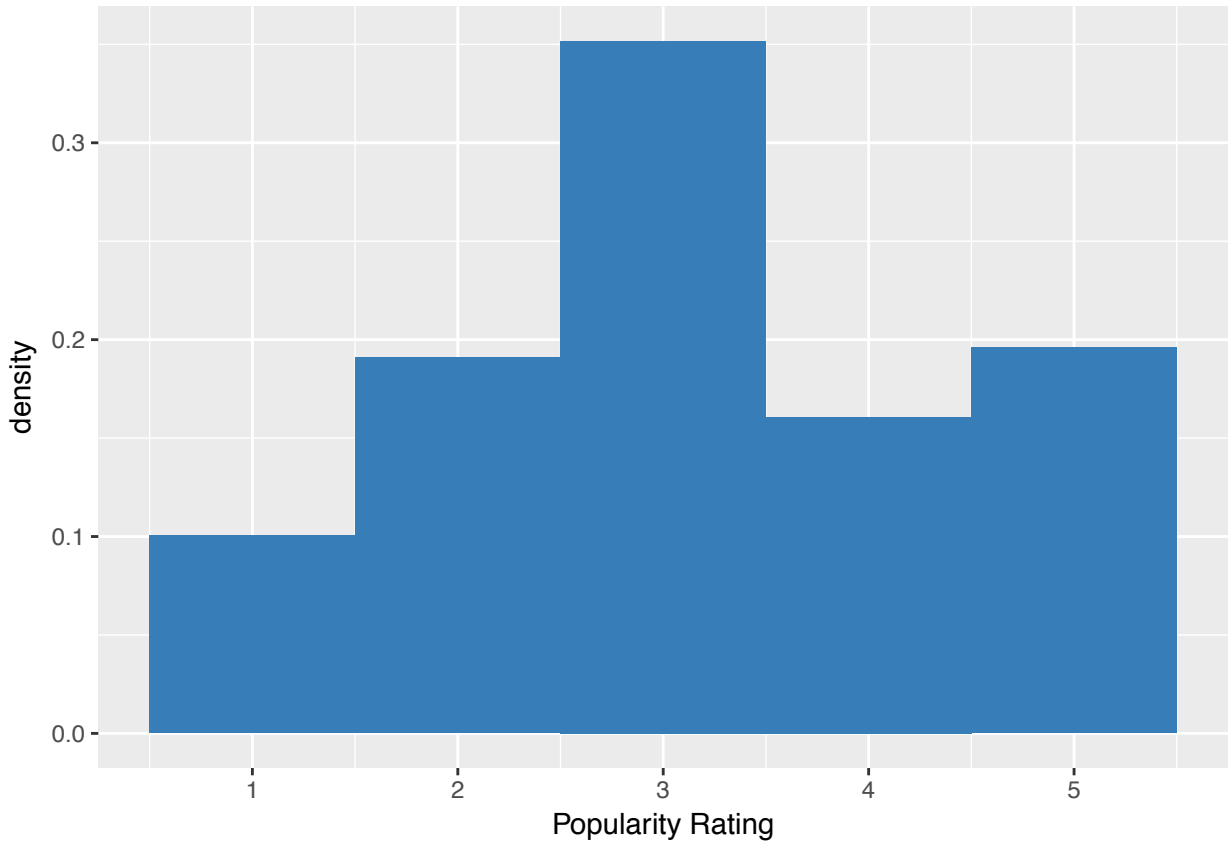
For my histogram, I decided to look into the "Episode Popularity" column from my dataset.

As I outlined in the Data Wrangling section, I used Apple Podcasts' popularity rankings to assign a numerical rating (1-5, with 5 being the highest) to each episode from the top 50 most listened to podcasts in the "Society & Culture" section on Apple's site.

```
library(ggplot2)
library(readr)
Podcast_Dataset <- read_csv("~/Documents/Podcast_Dataset.csv")
```

```
## Parsed with column specification:
## cols(
##   Podcast = col_character(),
##   `Episode Name` = col_character(),
##   `Running Time` = col_integer(),
##   `Release Date` = col_integer(),
##   `Episode Description` = col_character(),
##   `Popularity Rating` = col_integer()
## )
```

```
ggplot(Podcast_Dataset, aes(`Popularity Rating`)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill = "#377EB8")
```



Here we can see that the majority of podcast episodes are only somewhat popular, with most being ranked at around 3. Interestingly, we can also see that popular podcasts ranked at a 5 pop up only slightly more frequently than those that are slightly unpopular, ranked at a 2.

As for the Release Date data, we can see that podcast episodes from the summer months appear with the highest density. This makes sense, as I pulled my data in the summer

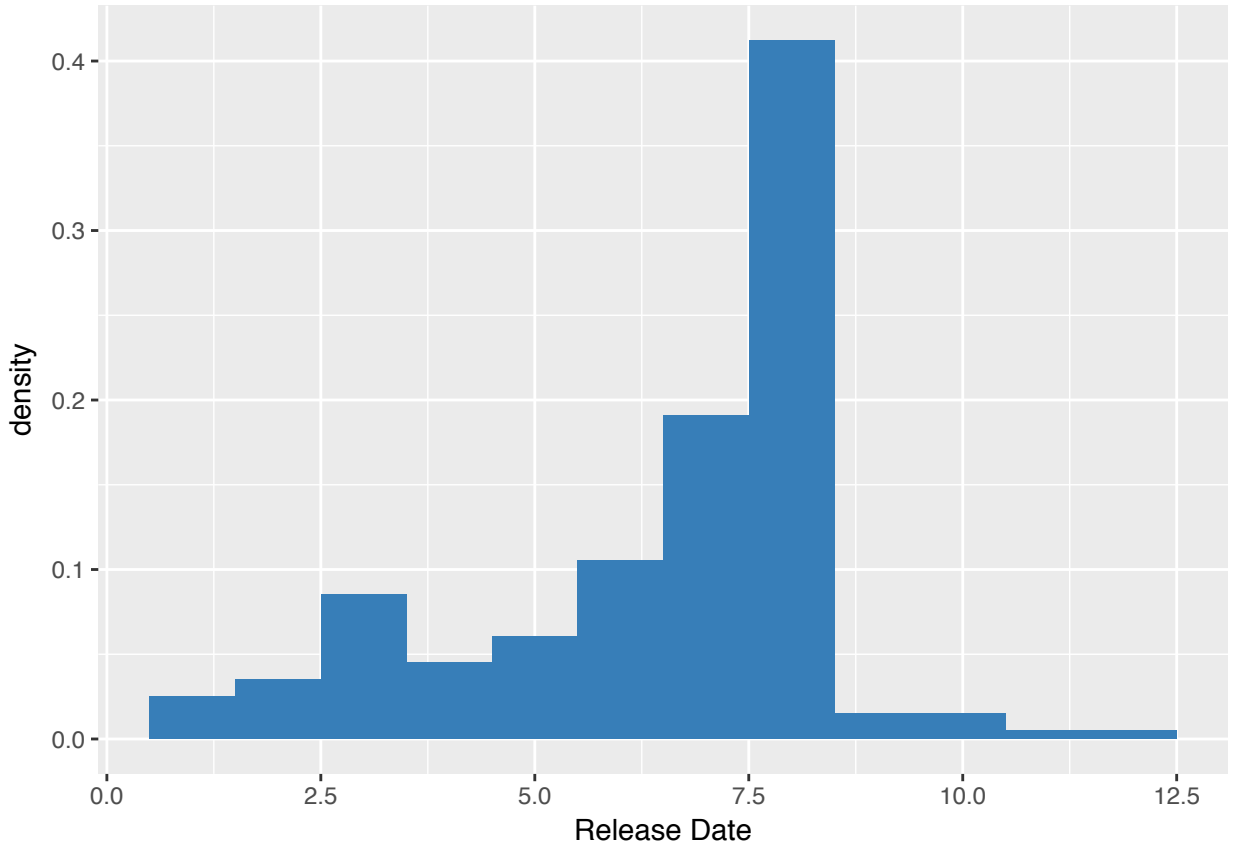
Creating The Scatterplot

Since I collected my data over the summer, I wondered if perhaps the popularity ratings were showing a recency bias. As you can see in the histogram below, podcast episodes released in July and August appear with by far the highest density.

```
library(ggplot2)
library(readr)
Podcast_Dataset <- read_csv("~/Documents/Podcast_Dataset.csv")
```

```
## Parsed with column specification:
## cols(
##   Podcast = col_character(),
##   `Episode Name` = col_character(),
##   `Running Time` = col_integer(),
##   `Release Date` = col_integer(),
##   `Episode Description` = col_character(),
##   `Popularity Rating` = col_integer()
## )
```

```
ggplot(Podcast_Dataset, aes(`Release Date`)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill = "#377EB8")
```

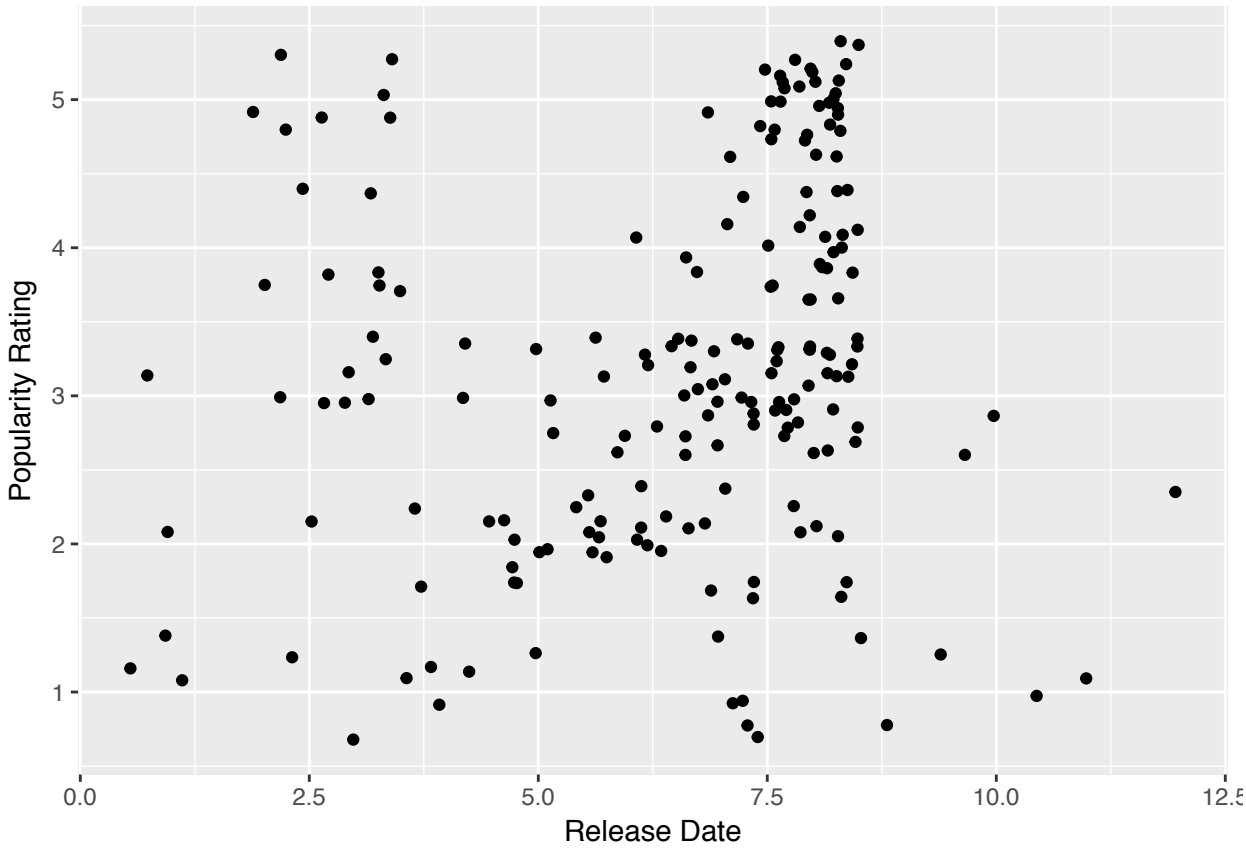


My thought was that perhaps the most popular podcasts were the ones most recently listened to by subscribers who downloaded new episodes every week. These summer episodes would be the most popular, receiving ratings of 4 and 5, while podcasts with earlier release dates would have ratings of 3 or lower.

I decided to make a scatterplot of the "Release Date" and "Popularity Rating" variables. I assigned each month their numeric value (1 for January, 2 for February, etc.) to make my data easier to plot.

```
library(ggplot2)

ggplot(Podcast_Dataset, aes(x = `Release Date`, y = `Popularity Rating`)) +
  geom_point(position = position_jitter(0.5))
```



It turns out, there is somewhat of a correlation there. There are definitely more podcast episodes pulled from the summer months of July and August, and their popularity ratings are pretty high - we can see a lot of 3s, 4s and 5s. Of course, there are some outliers there: what is it about those episodes from January and February that remained so popular months later? I wanted to see what else I could analyze in relation to popularity rating.

Creating The Time-series Plot

I decided to make a time-series plot for each Podcast's Popularity Rating and Running Time. Maybe shorter podcasts are more popular?

```
ggplot(Podcast_Dataset, aes(x = `Popularity Rating`, y = `Running Time`, color = `Podcast`)) +
  geom_line(position = position_fill(0.1))
```

```
## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for Unicode character U+2019

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for Unicode character U+2019

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)):
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)):
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <80>
```

```
## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)):
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <99>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for Unicode character U+2019

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for Unicode character U+2019

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs': dot
## substituted for <99>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for Unicode character U+2019

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for Unicode character U+2019

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs': dot
## substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs':
## dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs':
## dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
```

```
## conversion failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs':
## dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs':
## dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs':
## dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs':
## dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Dan. Carlin's Hardcore History' in 'mbcsToSbcs': dot
## substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs':
## dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs':
## dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Oprah's SuperSoul Conversations' in 'mbcsToSbcs':
## dot substituted for <99>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs':
## dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs':
## dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'What's Good with Stretch & Bobbito' in 'mbcsToSbcs':
## dot substituted for <99>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x
## $x, x$y, : conversion failure on 'Dan. Carlin's Hardcore History' in
## 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x
## $x, x$y, : conversion failure on 'Dan. Carlin's Hardcore History' in
## 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x
## $x, x$y, : conversion failure on 'Dan. Carlin's Hardcore History' in
```

```
## 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x
## $x, x$y, : conversion failure on 'Oprah's SuperSoul Conversations' in
## 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x
## $x, x$y, : conversion failure on 'Oprah's SuperSoul Conversations' in
## 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x
## $x, x$y, : conversion failure on 'Oprah's SuperSoul Conversations' in
## 'mbcsToSbcs': dot substituted for <99>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'What's Good with Stretch & Bobbito' in
## 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'What's Good with Stretch & Bobbito' in
## 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'What's Good with Stretch & Bobbito' in
## 'mbcsToSbcs': dot substituted for <99>
```

According to this time-series plot, there isn't really a correlation between episode length and popularity. The most popular podcasts come in all different lengths: short, long and in between.

So outside of recency, what makes a podcast popular? What do listeners want and how can that translate for advertisers and publishers? My guess is that content is king - it's time to get into some text analysis.