Exploration by Visualization: The Streaming Movies Dataset

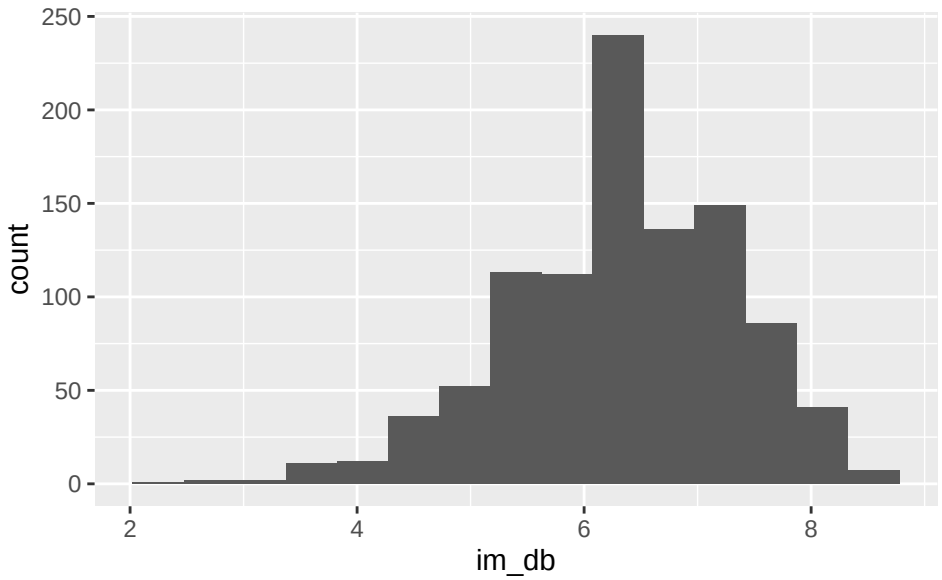
Marina Huang

2022-05-26

Visualization by example

Histogram of IMDB Scores

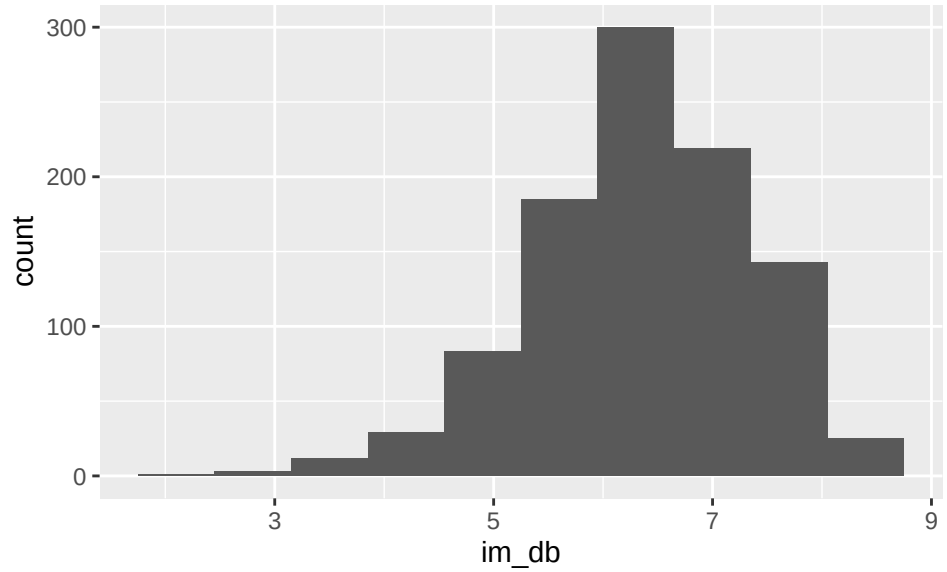
```
ggplot(data = streaming) +
  geom_histogram(mapping = aes(x = im_db), bins = 15)
```



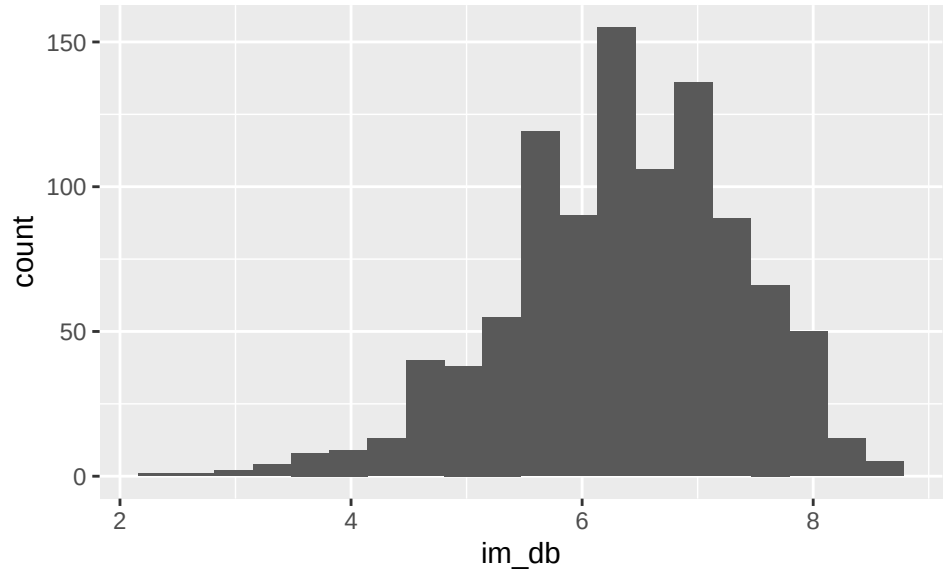
This histogram is counting the number of times there is a certain IMDB score. Based on this graph, the score of roughly 6 to 6.5 has the highest number of counts.

The Difference Between bins and binwidth

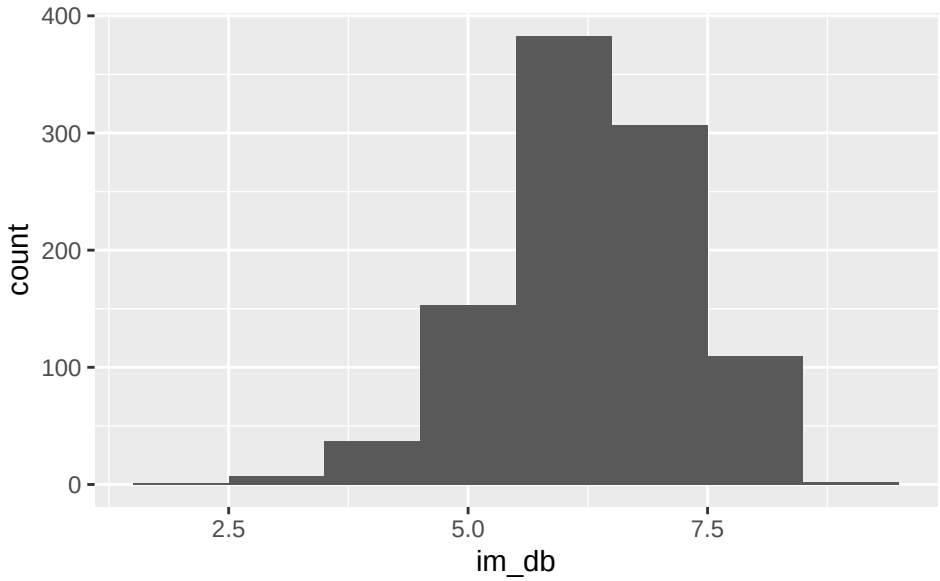
```
ggplot(data = streaming) +
  geom_histogram(mapping = aes(x = im_db), bins = 10)
```




```
ggplot(data = streaming) +
  geom_histogram(mapping = aes(x = im_db), bins = 20)
```



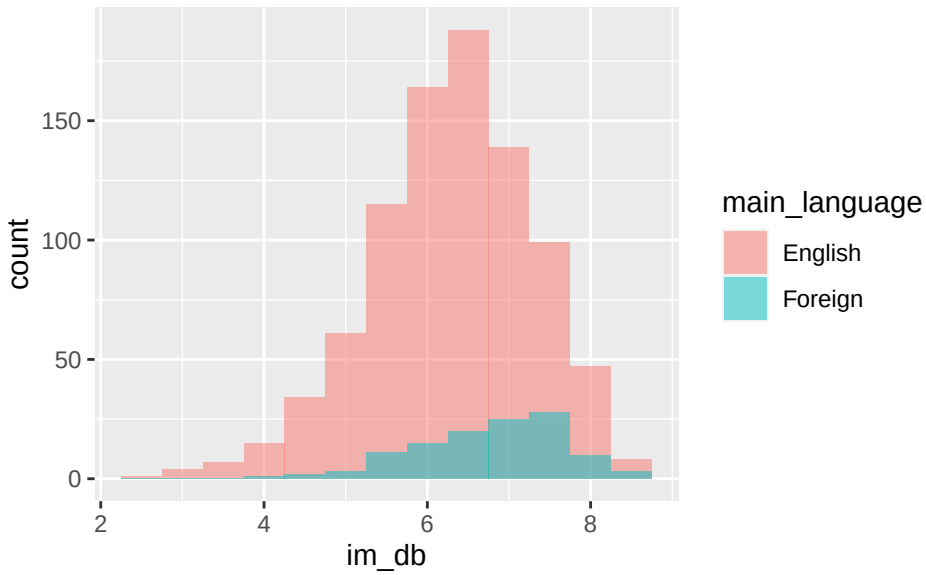

```
ggplot(data = streaming) +
  geom_histogram(mapping = aes(x = im_db), binwidth = 1)
```



`bins` determines how many bars, or equally spaced intervals there are in the histogram. Meanwhile, `binwidth` determines the range of the intervals.

Histogram of Movies in English vs. Foreign Languages

```
ggplot(data = streaming) +
  geom_histogram(
    mapping=aes(x = im_db, fill = main_language), binwidth = 0.5, alpha = 0.5,
    position = "identity"
  )
```



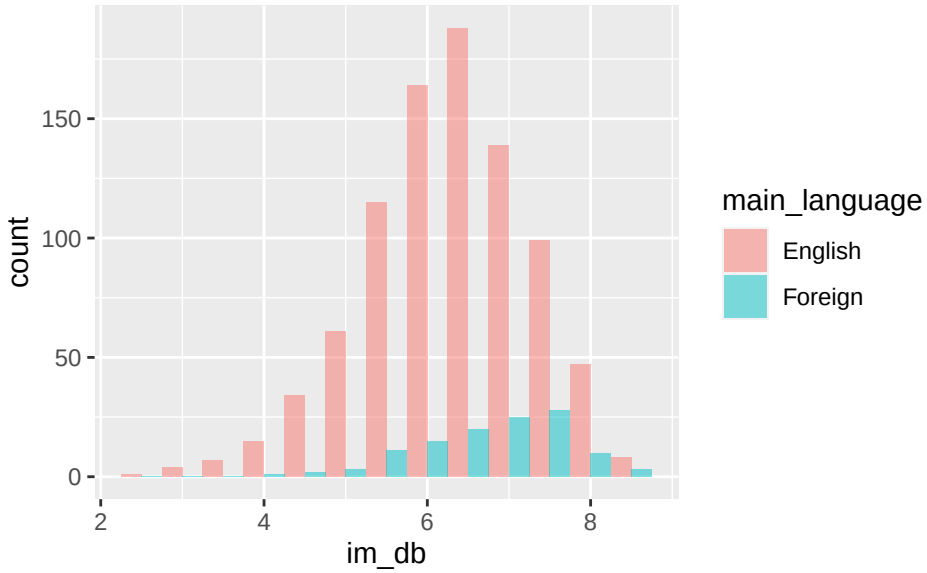
Adding `fill = main_language` filled the histogram in different colors, distinguished by `main_language`. In other words, there are actually 2 histograms shown in one graph, but they two histograms are distinguished by different colors. This definitely changes the way we might interpret the visualization because without the `fill = main_language`, we would not have been able to see that movies with a foreign main language would tend to have higher IMDB scores than movies with English as the main language.

Analysis

The shape of the **English** IMDB rating distribution is slightly left skewed with the center being around 6 to 6.5. The shape of the **Foreign** IMDB rating distribution has very left skewed and has the center at around 7. This is a distinguishable difference. This was something I was expecting to see because there have been a lot of foreign movies in the recent years being featured on popular streaming websites, like Netflix. Also, if the ratings of these foreign movies were low, I think American streaming websites would be less inclined to buy the licences to stream the movies.

Use of alpha = 0.5 and position = "dodge":

```
ggplot(data = streaming) +
  geom_histogram(
    mapping=aes(x = im_db, fill = main_language), binwidth = 0.5, alpha = 0.5,
    position = "dodge"
  )
```



i. Alpha changes the opacity of the histogram.
ii. `position = "identity"` changed the order of the languages displayed on the histogram. The **Foreign** part is now moved to the front instead of behind **English**.
iii. The two language categories do not overlap with `position = "dodge"`.

Scatterplots

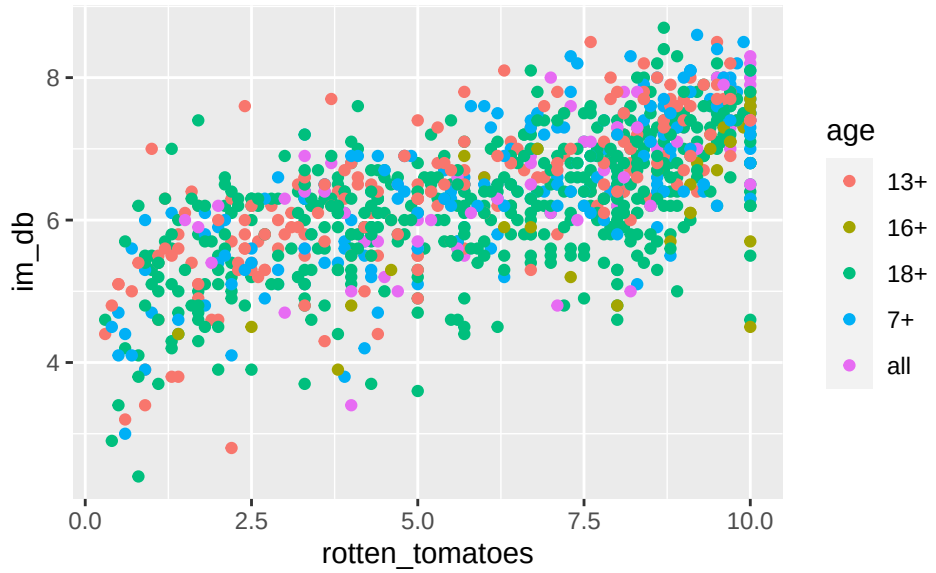
Correlation Between Rotten Tomato and IMDB Scores

```
ggplot(data = streaming) +
  geom_point(mapping = aes(x = rotten_tomatoes, y = im_db))
```

This plot has a positive trend. This implies that if the Rotten Tomato score of a film is high, then so is the IMDB score of that film.

Coloring Data Points By Age Rating

```
ggplot(data = streaming) +
  geom_point(mapping = aes(x = rotten_tomatoes, y = im_db, color = age))
```

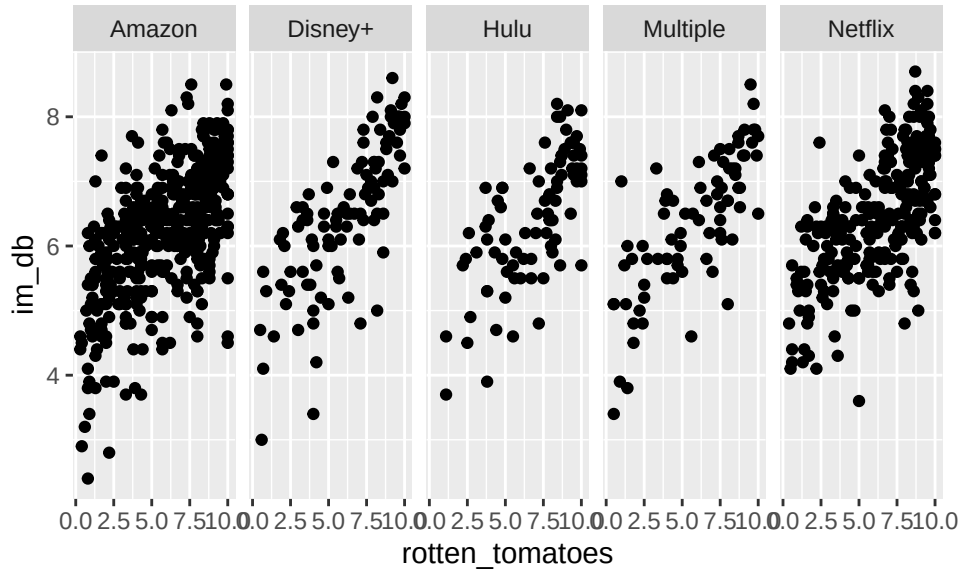


All of the colors seem to spread out across the plot. There is no age category that would only appear with a specific rating. This tells us that the age variable does not have an effect on how IMBD ratings depend on Rotten Tomatoes ratings.

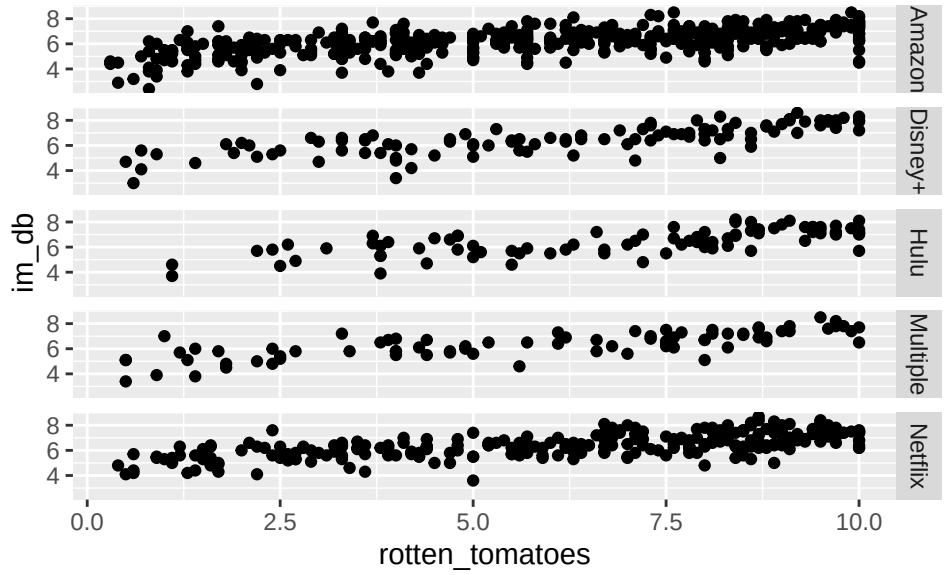
Faceting

IMDB and Rottem Tomato Scores By Streaming Services

```
ggplot(data = streaming) +
  geom_point(mapping = aes(x = rotten_tomatoes, y = im_db)) +
  facet_grid(. ~ streaming)
```



```
ggplot(data = streaming) +
  geom_point(mapping = aes(x = rotten_tomatoes, y = im_db)) +
  facet_grid(streaming ~ .)
```



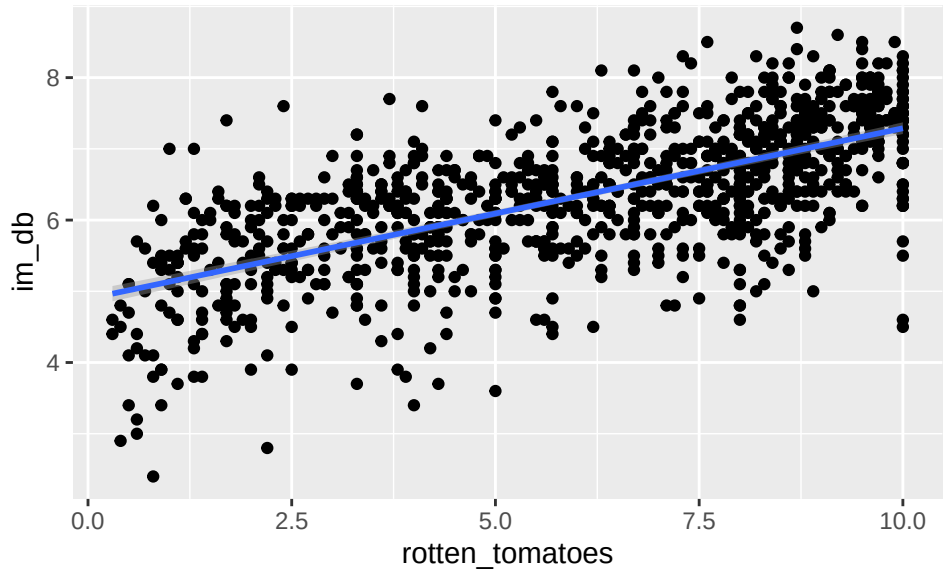
Adding `facet_grid` divided the plot based on the variable `streaming`. There is a plot for each category under `streaming`. Having `. ~ streaming` displays the category labels horizontally and having `streaming ~ .` displays the category labels aligned vertically.

Modeling in ggplot2

Using geom_point and geom_smooth

```
ggplot(data = streaming) +
  geom_point(mapping = aes(x = rotten_tomatoes, y = im_db)) +
  geom_smooth(mapping = aes(x = rotten_tomatoes, y = im_db), method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

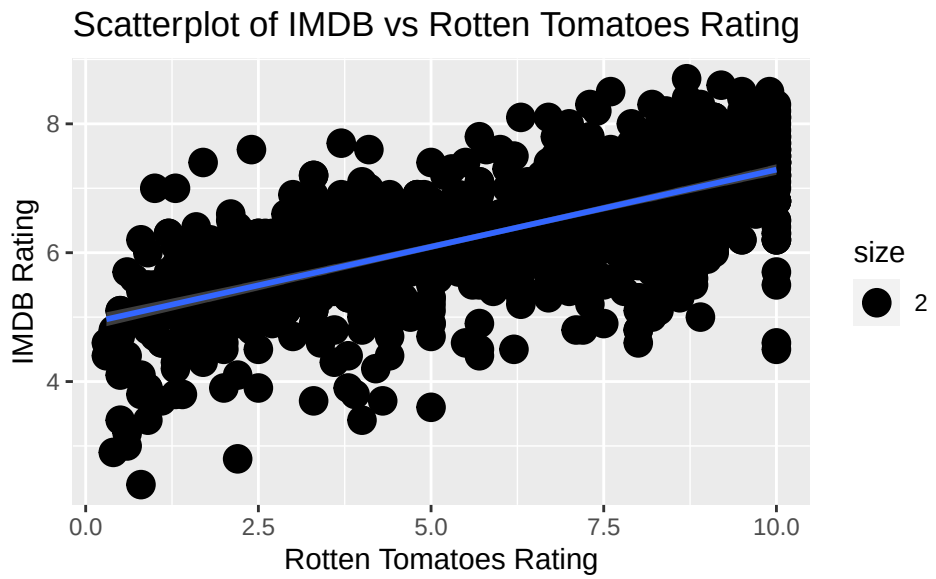


The plot follows a positive blue trendline, which matches what I've previously described about the data. The semi-transparent gray region probably represents how well the data points fit the trend. The smaller the gray area is, the better the points in the region fits the trend.

Adding Axis Titles

```
ggplot(data = streaming) +
  geom_point(mapping = aes(x = rotten_tomatoes, y = im_db, size = 2)) +
  geom_smooth(mapping = aes(x = rotten_tomatoes, y = im_db), method = "lm") +
labs(
  title = "Scatterplot of IMDB vs Rotten Tomatoes Rating",
  x = "Rotten Tomatoes Rating",
  y = "IMDB Rating"
)
```

```
## `geom_smooth()` using formula 'y ~ x'
```


Scatterplot of IMDB vs Rotten Tomatoes Rating
IMDB Rating
8
6
4
0.0
2.5
5.0
7.5
10.0
Rotten Tomatoes Rating
size
2