Lab 2: Exploration by visualization: the streaming movies dataset

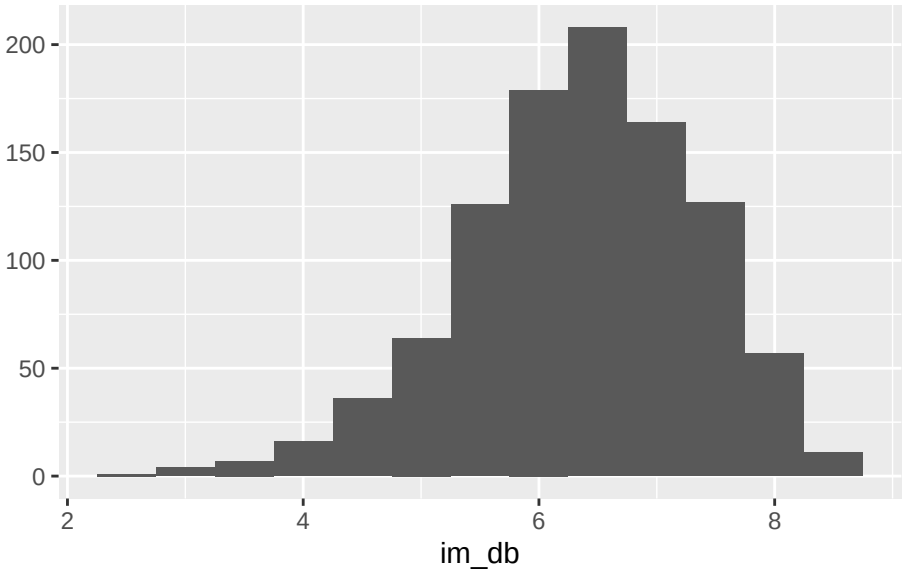
Eric Simo

2023-07-02

Visualization by example

Exercise 1

```
qplot(x = im_db, binwidth = 0.5, data = streaming)
```

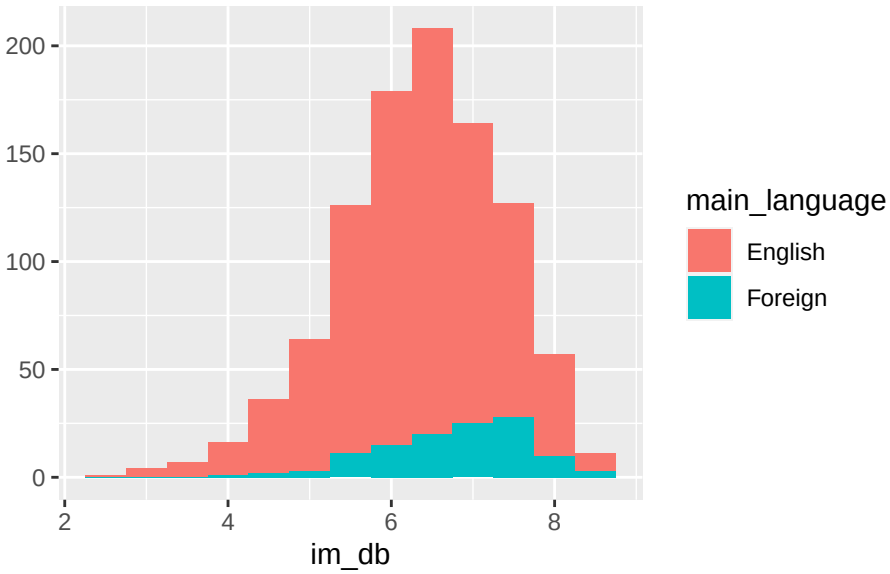


- The histogram is showing the im_bd average height rating across the streaming table.

Exercise 2

```
qplot(
  x = im_db,
  binwidth = 0.5,
```

```
  fill = main_language,
  data = streaming
  )
```



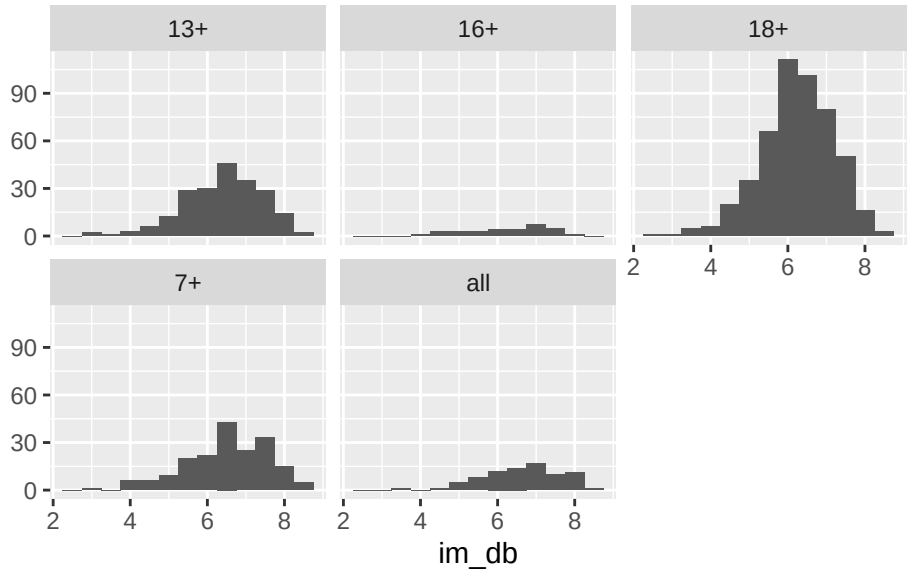
- By Adding the fill() function, It used the value provided (main_language) to fill in the missing entries in the language column(English and Foreign).
- Most of the movies in this database are in English.

Exercise 3

Based on the histogram. The modality will be unimodal and the skewness will be left-skew for the English distribution. The modality will be unimodal and the skewness will be left-skew for the Foreign distribution.
The tangible difference between the 2 distributions is that the entire databse barely has movies in forgein languages.
Based on my experience of watching movies in recent years, I would have definitely expected that result.

Exercise 4

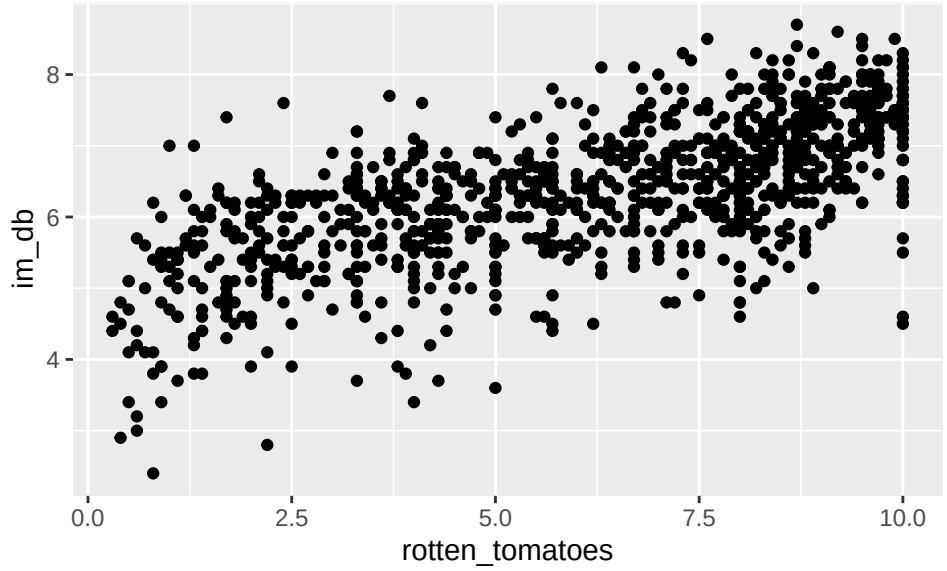
```
qplot(
  x = im_db,
  binwidth = 0.5,
  facets = ~ age,
  data = streaming
  )
```



- There are 5 facets.
- Each faceted sup-plot represent age.
- The facet's distribution with the most movies is the 18+ distribution.

Exercise 5

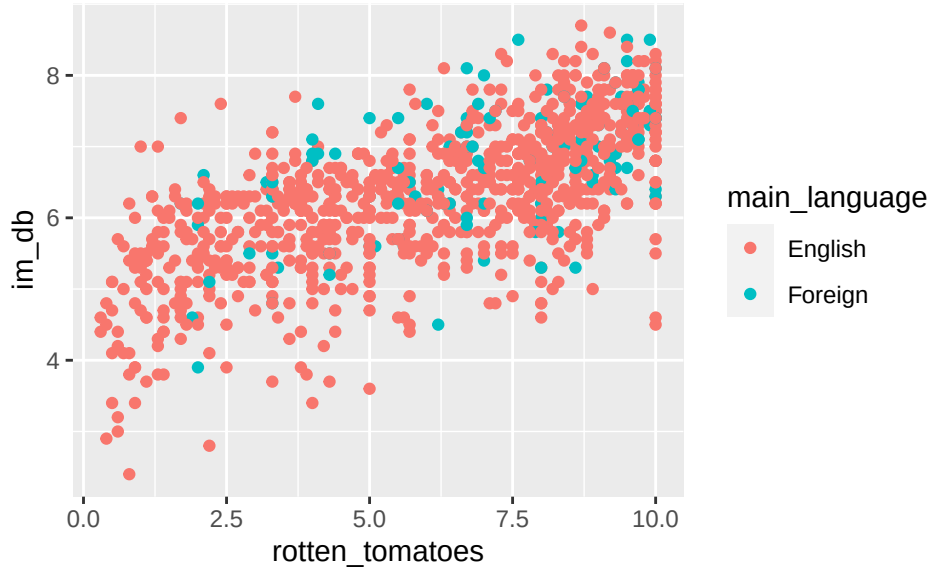
```
qplot(x = rotten_tomatoes, y = im_db, data = streaming)
```



The linear relationship between the two variables is positive. The correlation is high enough to draw a line but a non-linear one since we can a slight curve.

Exercise 6

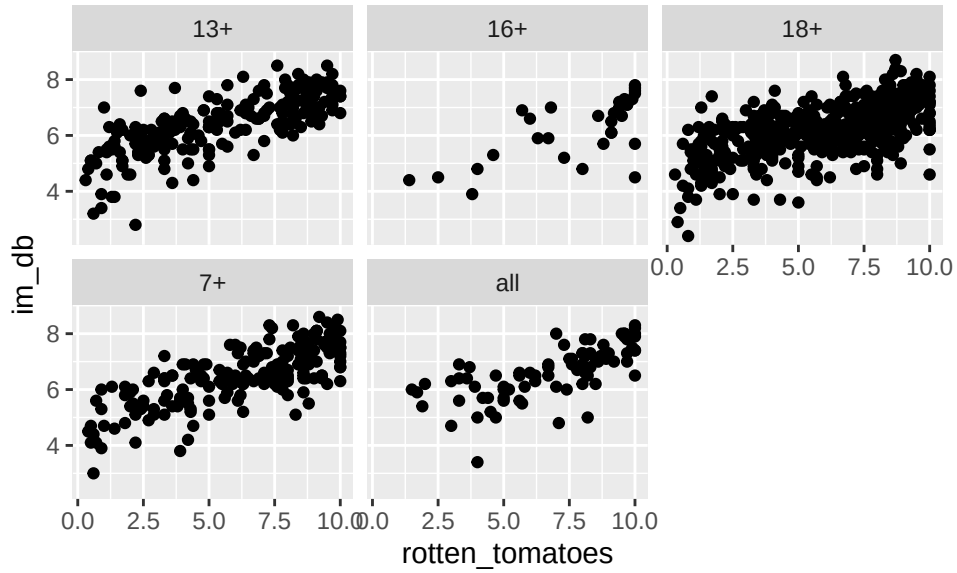
```
qplot(
  x = rotten_tomatoes,
  y = im_db,
  data = streaming,
  color = main_language
  )
```



The relationship between ratings on IMDB and Rotten Tomatoes look alike for English and foreign language movies because of the both follow the same direction of slope, contain both high level of correlation in the same are and one line could be drawn to represent both English and Foreign.

Exercise 7

```
qplot(
  x = rotten_tomatoes,
  y = im_db,
  facets = ~ age,
  data = streaming
  )
```



The information presented is not different from the information in Exercise 5 because every age categories show the same relationship between their ratings on IMDB vs Rotten Tomatoes.

Exercise 8

```
qplot(
  x = rotten_tomatoes,
  y = im_db,
  geom = "smooth",
  method = "lm",
  data = streaming
  )
```

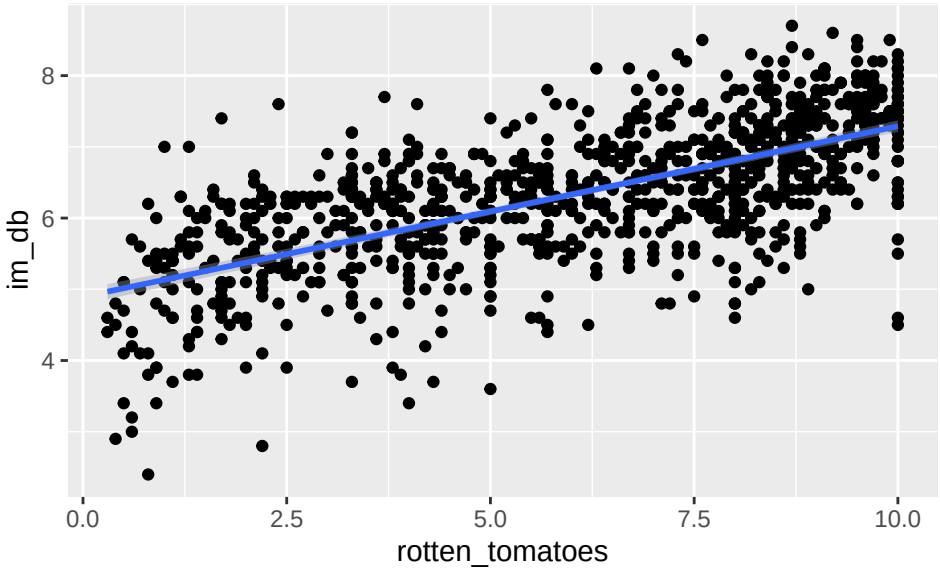
```
## `geom_smooth()` using formula = 'y ~ x'
```

The model definitely follows the trends previously described in the data.

Exercise 9

```
qplot(
  x = rotten_tomatoes,
  y = im_db,
  geom = c("point", "smooth"),
  method = "lm",
  data = streaming
  )
```

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


8
6
4
im_db
0.0
2.5
5.0
7.5
10.0
rotten_tomatoes