Lab 2: Exploration by visualization: the streaming movies dataset

Jordan Hutcherson

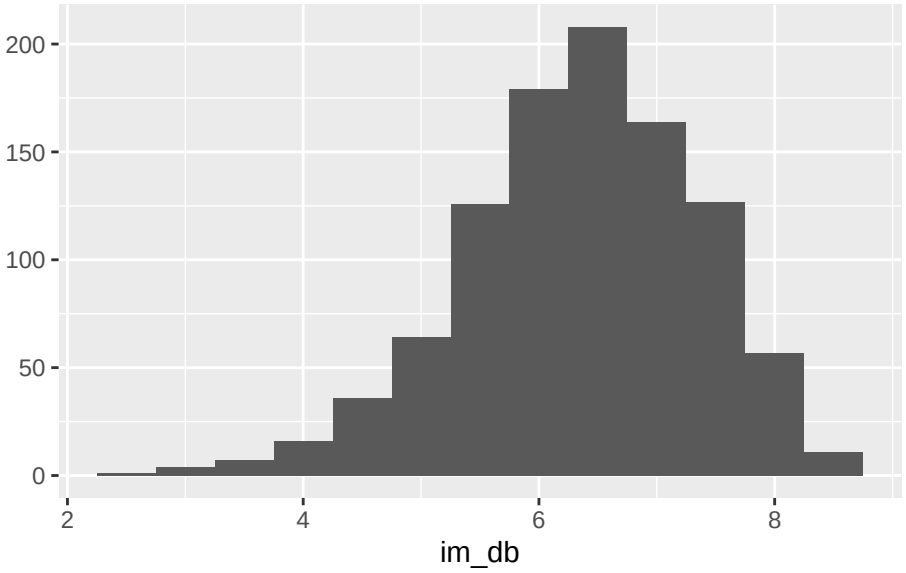
2024-02-06

Visualization by example

Exercise 1

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

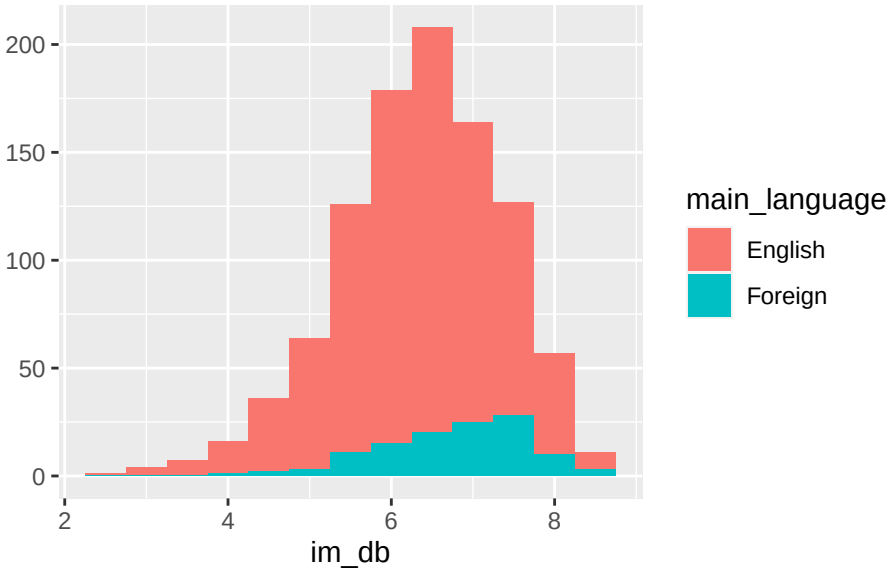
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



- The histogram is visually representing the distribution of "im_db" ratings of movies. Most of the movies are clustered around the central range.

Exercise 2

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



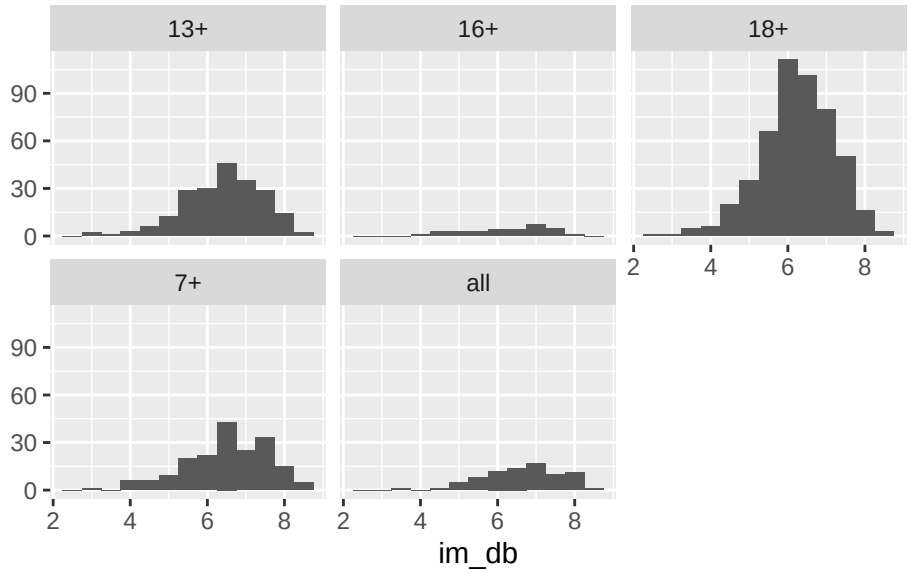
- Fill main language is coloring the bars in the histogram based on the variable "main language". It will show the distribution of the another variable while also providing visuals of that distribution by the different languages .
- Most of the movies in this dataset are english.

Exercise 3

- Uni modal with english being slight skew to the right and roughly symmetric. For foreign it is slightly skew to the right as well. English seems to be centered around a higher im db rating compared to the foreign language.

Exercise 4

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

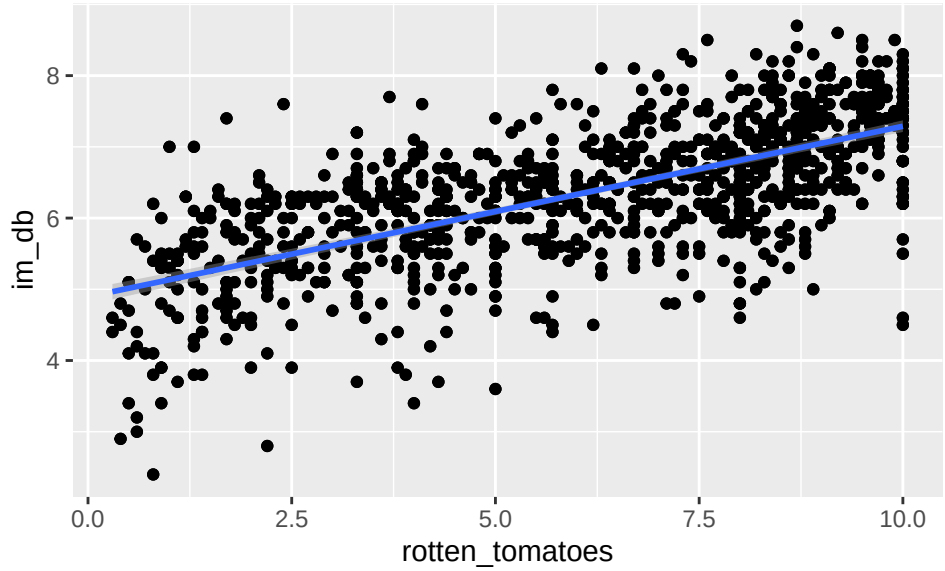


- There are 5 facets.
- Each faceted sup-plot represents the distribution of im db ratings for movies with different age categories.
- I would say plot with 18+ contains the most movies.

Exercise 5

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

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

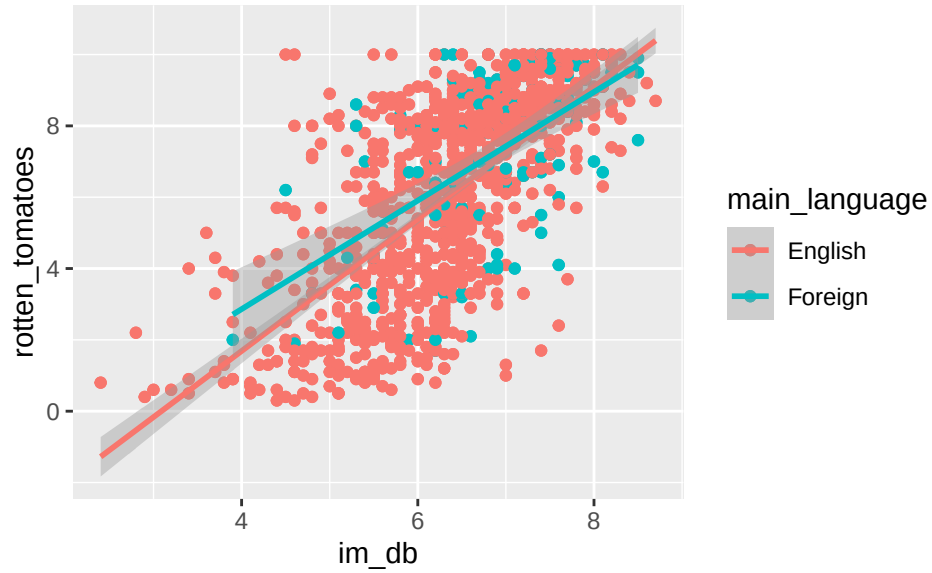


I see a positive linear association, the points are clustered, especially around the middle range. There are outliers that fall from the trend line and do not follow the general pattern of the rest of the data.

Exercise 6

```
qplot(x = im_db, y = rotten_tomatoes, color = main_language, data = streaming) +
 geom_point() +
 geom_smooth(method="lm")
```

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

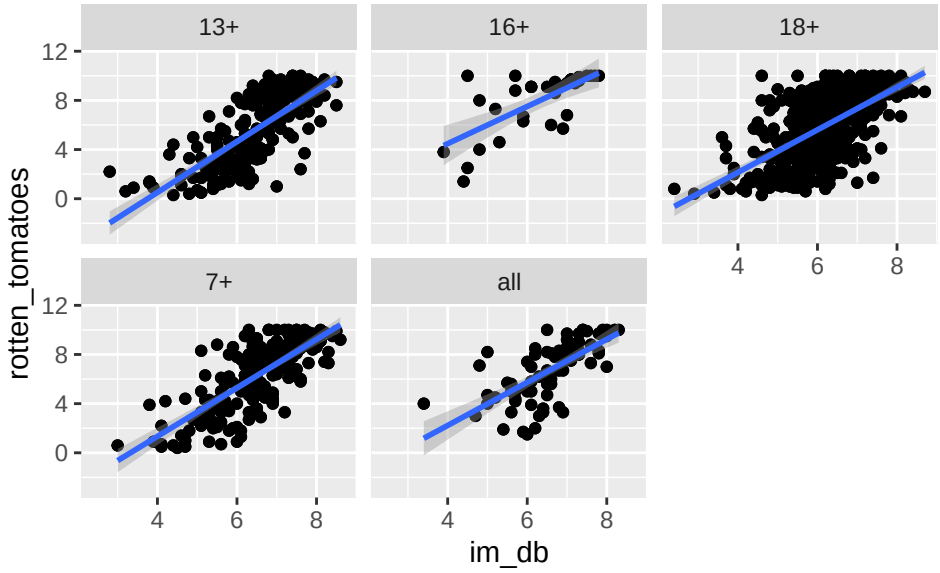


- Both english and foreign movies have a positive relationship between im db and rotten tomatoes rating, with no big significant difference in the trend based on language.

Exercise 7

```
qplot(x = im_db, y = rotten_tomatoes, facets = ~ age, data = streaming) +
 geom_point() +
 geom_smooth(method="lm")
```

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



- This plot differentiates the relationship between im db and rotten tomatoes ratings across the different age groups, offering a better view that was not in the last scatter plot.

Exercise 8

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

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

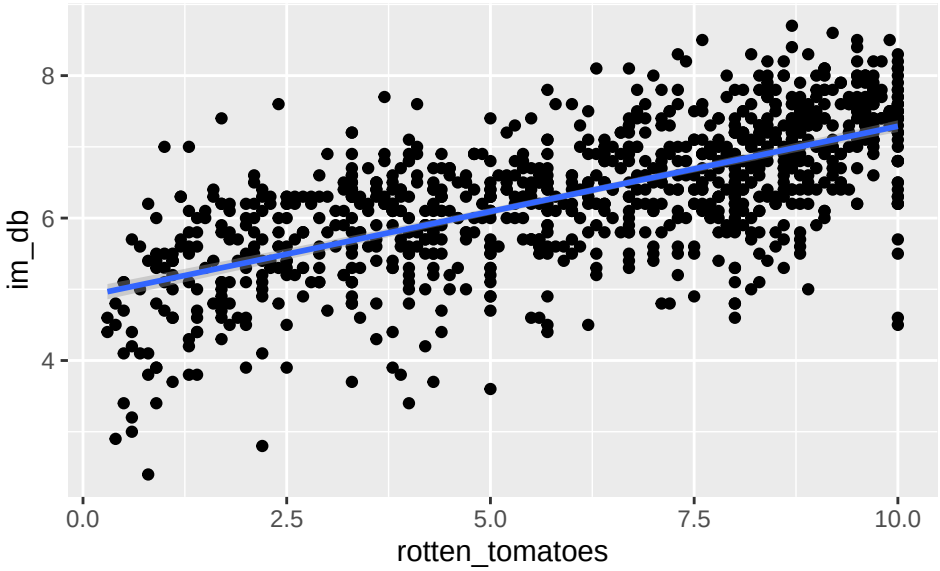
- this shows a consistent positive linear trend which coincides with all the previous data.

Exercise 9

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

```
## Warning in geom_point(method = "lm"): Ignoring unknown parameters: ‘method‘
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```


8
6
4
im_db
0.0
2.5
5.0
7.5
10.0
rotten_tomatoes