Exploring Books

April 27, 2022

The dataset used for this challenge is titled 'books.csv'.

What topic does the dataset cover?

Dataset on Amazon's Top 50 bestselling books from 2009 to 2019. Contains 550 books, data has been categorized into fiction and non-fiction using Goodreads

Acknowledgements

The dataset can be found at https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019.

Variables

- Name
- Author
- User_rating
- Reviews
- Price
- Year
- Genre

Assignment

Complete each of the following questions using R Studio and submit your answers as a detailed report.

1. Visualize and describe the distribution of book prices.
2. Visualize and describe the distribution for book user ratings
3. What is the highest user rating a book received in 2019? Which books received this rating and which would you be most likely to read?
4. Which authors has the highest average price for their books in 2017? (Top 10)
5. Which authors have the highest average user review rating for their books? (Top 10)
6. Does genre appear to have a significant effect on price? Use pairwise boxplots to support your answer.
7. Are user rating and price linearly related? Use a scatter plot to support your answer.

Solution:

```r
### Load packages and dataset
library(tidyverse)
library(ggplot2)
book <- read.csv('data/datasets/book.csv')

# First just take a glimpse
glimpse(book)

### Question 1: Distribution for the price of a book
book %>% ggplot(aes(x=Price)) +
  geom_histogram(fill="green", color="black") +
  theme_classic() +
  labs(title="Distribution for book prices")

### Question 2: Distribution for book ratings
book %>% ggplot(aes(x=User_rating)) +
  geom_histogram(fill="purple", color="black") +
  theme_classic() +
  labs(x="User Rating", title="Distribution for book ratings")

### Question 3: 52 Top rated books in 2019
# There are 52 books with user ratings of 4.9, none with 5
sum(book$User_rating == 4.9)
# Let's look at books that recieved 4.9
book %>% filter(Year==2019) %>%
  arrange(desc(User_rating)) %>%
  select(Name) %>%
  head(50)

### Question 4: Authors with most expensive books 2017
book %>% group_by(Author) %>% filter(Year == 2017) %>%
  summarise(mean=mean(Price)) %>%
  arrange(desc(mean)) %>%
  head(10)

### Question 5: Top Author based on reviews
book %>% group_by(Author) %>% summarise(mean=mean(User_rating)) %>%
  arrange(desc(mean)) %>% filter(mean==4.9)

### Question 6: Does genre have an effect on price?
book %>% ggplot(aes(y=User_rating, fill=Genre))+
  geom_boxplot() +
  theme_classic() +
  labs(x="Genre", y="User Rating")

### Question 7: User_rating vs Price. No relationship
book %>% ggplot(aes(x=User_rating, y=Price)) +
  geom_point() +
  theme_classic()

```