Week 5 homework assignment

Mubarak Ganiyu

09/29/2021

DSI-EDA

Professor Michael Shepherd

Part 1

The dataset chosen is the one about every obsenity or death in Tarantino movies. It has been loaded in the code chunk below.

Part 2

The dataset was collected by Oliver Roeder from FiveThirtyEight. The purpose behind creating the dataset was to create a catalog of all the swear words and deaths that has occurred in every Tarantino movie. This is a great dataset because I am a big fan of Quentin Tarantino. I have seen every single one of his movies from Reservoir Dogs to Once Upon a Time in Hollywood, and coming across this dataset has inspired me to see them again. Each row details the movie name, the type of event (swear word or death), a word for swear word or a missing value for death, and the minute at which the swear word occured.

```
dim(df)
```

```
## [1] 1894    4
```

```
names(df)
```

```
## [1] "movie"      "type"       "word"       "minutes_in"
```

```
str(df)
```

```
## 'data.frame':    1894 obs. of  4 variables:
##  $ movie     : chr  "Reservoir Dogs" "Reservoir Dogs" "Reservoir Dogs" "Reservoir Dogs" ...
##  $ type      : chr  "word" "word" "word" "word" ...
##  $ word      : chr  "dick" "dicks" "fucked" "fucking" ...
##  $ minutes_in: num  0.4 0.43 0.55 0.61 0.61 0.66 0.9 1.43 1.56 1.66 ...
```

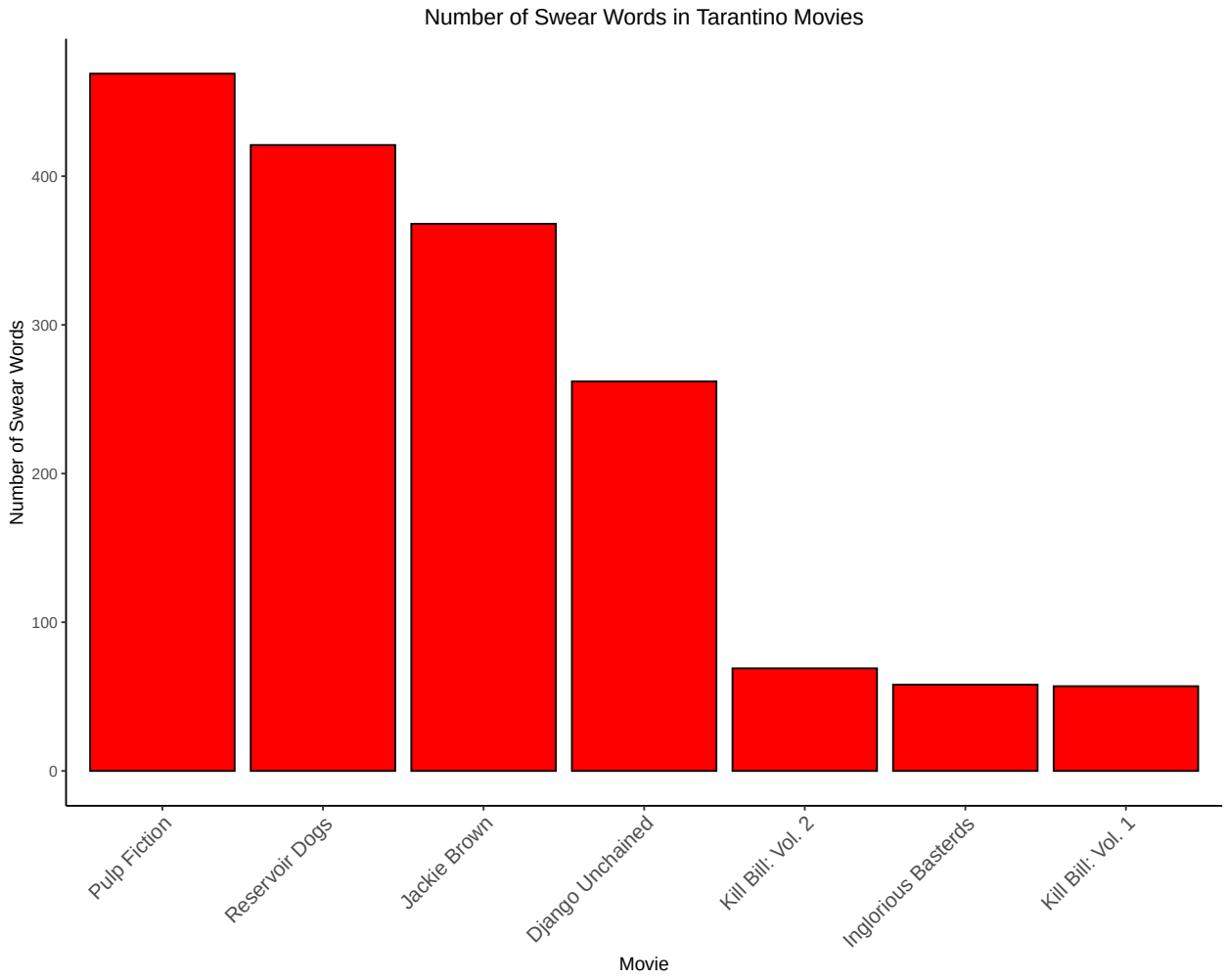
There are 1894 rows and 4 columns. The movie, type and word columns are character types and the minutes_in is a numerical column.

Part Three

Which movie has the most swear words? The objective is to figure out which movie displays the most profanity, and figure out the swear words that occurred the most in that movie.

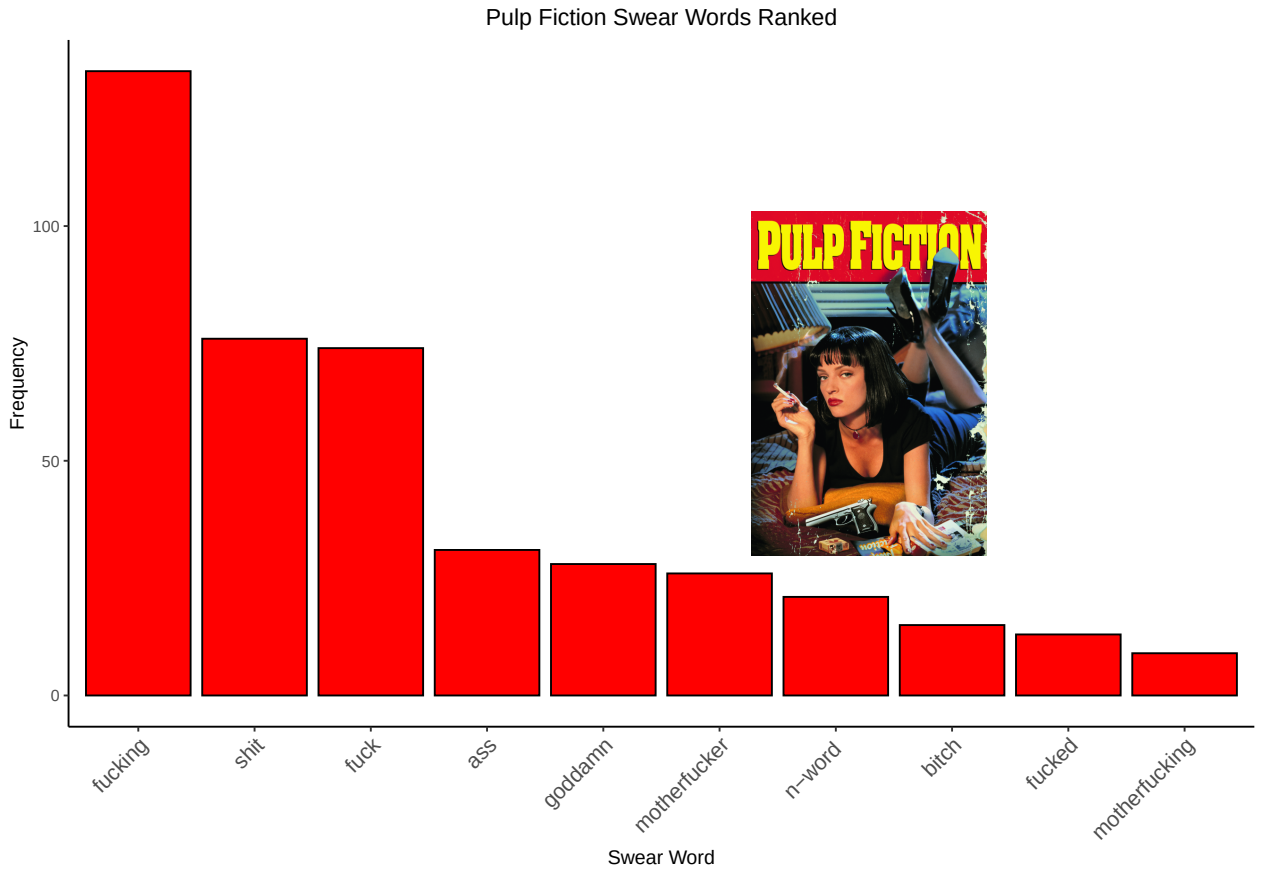
```
df_movies_word <- df %>%
  filter(type == "word") %>%
  count(movie)
ggplot(df_movies_word) +
  geom_col(aes(x = reorder(movie, -n), y = n), color = "black", fill = "red") +
```

```
  labs(title = "Number of Swear Words in Tarantino Movies",
       x = "Movie",
       y = "Number of Swear Words") +
  theme_classic() +
  theme(plot.title = element_text(hjust=0.5),
        axis.text.x= element_text(size=12, angle =45, hjust = 1))
```




```
df_movies_curses <- df %>%
  filter(type == "word", movie == "Pulp Fiction") %>%
  count(word) %>%
  arrange(desc(n)) %>%
  head(10)
ggplot(df_movies_curses) +
  geom_col(aes(x = reorder(word, -n), y = n), color = "black", fill = "red") +
  labs(title = "Pulp Fiction Swear Words Ranked",
       x = "Swear Word",
       y = "Frequency") +
  theme_classic() +
  theme(plot.title = element_text(hjust=0.5),
        axis.text.x= element_text(size=12, angle =45, hjust = 1)) +
  inset_element(p = img_n,
                left = 0.5,
```

```
                bottom = 0.25,
                right = 0.85,
                top = 0.75)
```



The first graphic was used to figure out the Tarantino movie that had the highest number of curse words. It turns out it is the movie that most people would regard as his magnum opus: **Pulp Fiction**. The second one clearly highlights which swear word was utilized the most in Pulp Fiction. In this case, **fucking** is the swear word that was cursed out the most in Pulp Fiction. These two graphics were chosen because they address the question that was brought up earlier about which movie had the most swear words.

Part Four

It took me about 30 seconds. As soon as I saw Tarantino, I clicked on the github repo and went straight into the .csv file. I did a little bit of data wrangling to extract the rows containing swear words so as to count the number of swear words per movie. This made it possible to create the first graphic showcasing the number of swear words in each Tarantino movies. For the second graphic, I had to extract the rows containing Pulp Fiction and swear words. Then, I had to count the frequency of the swear words. This was crucial towards building the second graphic. This was an interesting assignment as I learnt that the earlier Tarantino movies contained more swear words than the most recent ones.