Lab 2 SOLUTION

Grader

3/25/2020

Question 1

What is the range of years of production of the movies of this data set?

```
data(movies)
range(movies$year)
```

```
[1] 1893 2005
```

The production years of the oldest and most recent movie are 1893 and 2005 respectively.

Question 2

What proportion of movies have their budget included in this data base, and what proportion doesn't? What are top 5 most expensive movies in this data set?

```
nobug = sum(is.na(movies$budget))
N = dim(movies)[1]
prop = (N-nobug)/N
print(prop)
```

```
[1] 0.08870858
```

The proportion of movies have their budget included in this data is 8.87%, and the oppersite proportion is 91.13%.

The top 5 most expensive movies are

```
movies %>% top_n(5,budget) %>% select(title)
```

```
# A tibble: 5 x 1
  title
  <chr>
1 Spider-Man 2
2 Terminator 3: Rise of the Machines
3 Titanic
4 Troy
5 Waterworld
```

Question 3

What are top 5 longest movies?

```
movies %>% top_n(5,length) %>% select(title,length)
```

```
# A tibble: 5 x 2
  title                                             length
  <chr>                                              <int>
1 Cure for Insomnia, The                              5220
2 Four Stars                                          1100
3 Longest Most Meaningless Movie in the World, The    2880
4 Out 1                                                773
```

```
5 Resan                                                    873
```

Question 4

** Of all short movies, which one is the shortest? Which one is the longest? How long are the shortest and the longest short movies? **

The longest shore movie is

```
movies %>% filter(Short==1) %>% top_n(1,length) %>% select(title,length)
```

```
# A tibble: 1 x 2
  title               length
  <chr>                <int>
1 10 jaar leuven kort    240
```

The shortest movies are

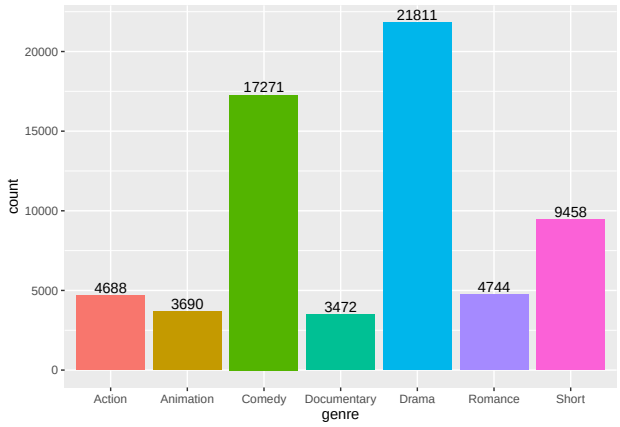
```
movies %>% filter(Short==1) %>% top_n(-1,length) %>% select(title,length)
```

```
# A tibble: 165 x 2
   title                                                 length
   <chr>                                                  <int>
 1 17 Seconds to Sophie                                       1
 2 2 A.M. in the Subway                                       1
 3 Admiral Cigarette                                          1
 4 Admiral Dewey Leading Land Parade                          1
 5 Alphonse and Gaston, No. 3                                 1
 6 Ameta                                                      1
 7 Amy Muller                                                 1
 8 Arabian Gun Twirler                                        1
 9 Arrival of McKinley's Funeral Train at Canton, Ohio        1
10 As Seen Through a Telescope                                1
# ... with 155 more rows
```

Question 5

How many movies of each genre (action, animation, comedy, drama, documentary, romance, short) are there in this data base?

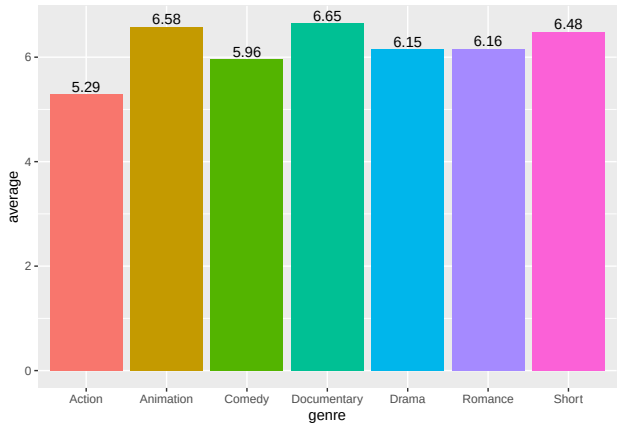
```
tep=movies %>% select(Action:Short) %>% colSums
data.frame(genre = names(tep),count = tep) %>% ggplot(aes(x=genre,y=count,fill=genre)) +
  geom_bar(stat="identity") + theme(legend.position = "none") +
  geom_text(aes(label=count),position=position_dodge(width=0.9), vjust=-0.25)
```



Question 6

What is the average rating of all movies within each genre?

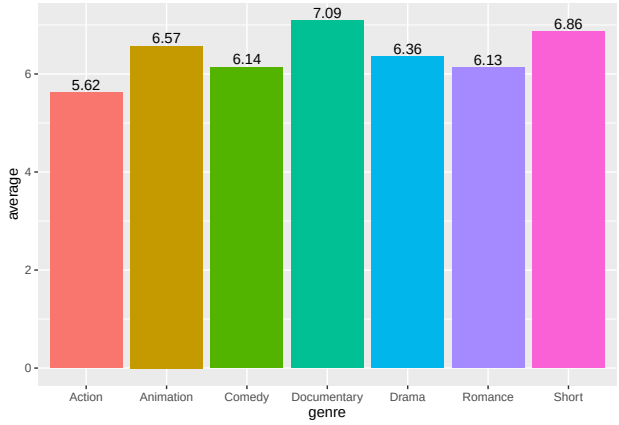
```
movies %>% tidyr::gather(key="genre",value="identity","Action","Animation","Comedy","Drama",
                  "Documentary", "Romance", "Short") %>%
  filter(identity == 1) %>%
  group_by(genre)  %>%
  summarise(average = mean(rating)) %>% ggplot(aes(x=genre,y=average,fill=genre)) +
  geom_bar(stat="identity") + theme(legend.position = "none") +
  geom_text(aes(label=round(average,2)),position=position_dodge(width=0.9), vjust=-0.25)
```



Question 7

What is the average rating of all movies within each genre that were produced in the years 2000-2005?

```
movies %>% tidyr::gather(key="genre",value="identity","Action","Animation","Comedy","Drama",
                  "Documentary", "Romance", "Short") %>%
  filter(identity == 1) %>%
  filter(year >= 2000 & year < 2005) %>%
  group_by(genre)  %>%
  summarise(average = mean(rating)) %>% ggplot(aes(x=genre,y=average,fill=genre)) +
  geom_bar(stat="identity") + theme(legend.position = "none") +
  geom_text(aes(label=round(average,2)),position=position_dodge(width=0.9), vjust=-0.25)
```



Question 8

For each of the first 6 genres (not including short movies) consider only movies from 1990 until the last year recorded and plot a function of the number of movies in this data base of corresponding genre produced by year, for years from 1990 until the last year recorded.

```
movies %>%
  tidyr::gather(key = "genre", value =  "identity", "Action", "Animation", "Comedy",
         "Drama", "Documentary", "Romance") %>%
  filter(identity == 1) %>%
  filter (year >= 1990) %>%
  group_by(genre, year) %>%
  add_tally(name = "Movies_per_Year") %>%
  group_by(year, genre, Movies_per_Year) %>%
  summarize() %>%
  ggplot(aes(x = year, y = Movies_per_Year, colour = genre)) +
  geom_line(size=1.05) + geom_point() + ggtitle('Total Movies Produced by Genre from 1990-2005')
```