EDA_Qiang

Qiang Fang

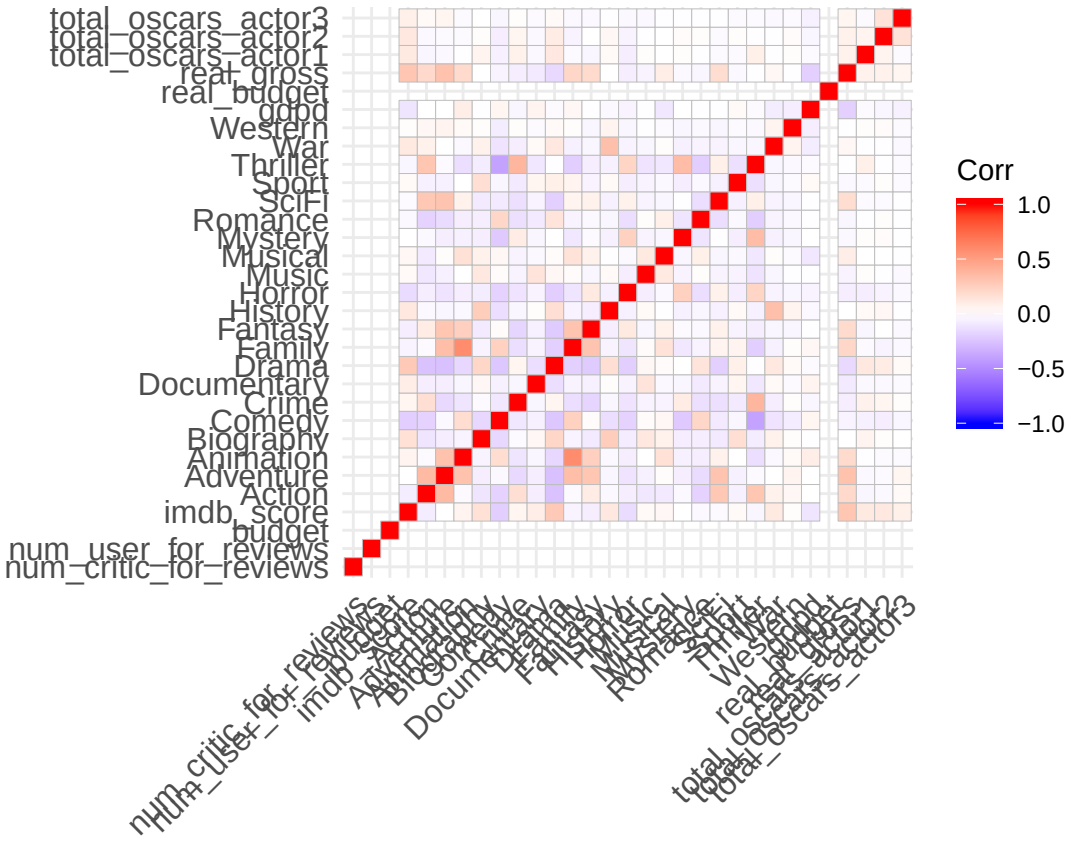
March 17, 2019

```
#load("D:/academic/DS 5110 Introduction to Data Management and Processing/project/katrina/proj_cleaned_
load(file = '~/DS5110/data/proj_cleaned_dta.RData')
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 3.5.3
```

```
train_num<- select(train,num_critic_for_reviews,num_user_for_reviews,budget,imdb_score,
       Action,Adventure,Animation,Biography,Comedy,Crime,Documentary,Drama,Family,Fantasy,History,Horro
corr <- cor(train_num)
ggcorrplot(corr)
```

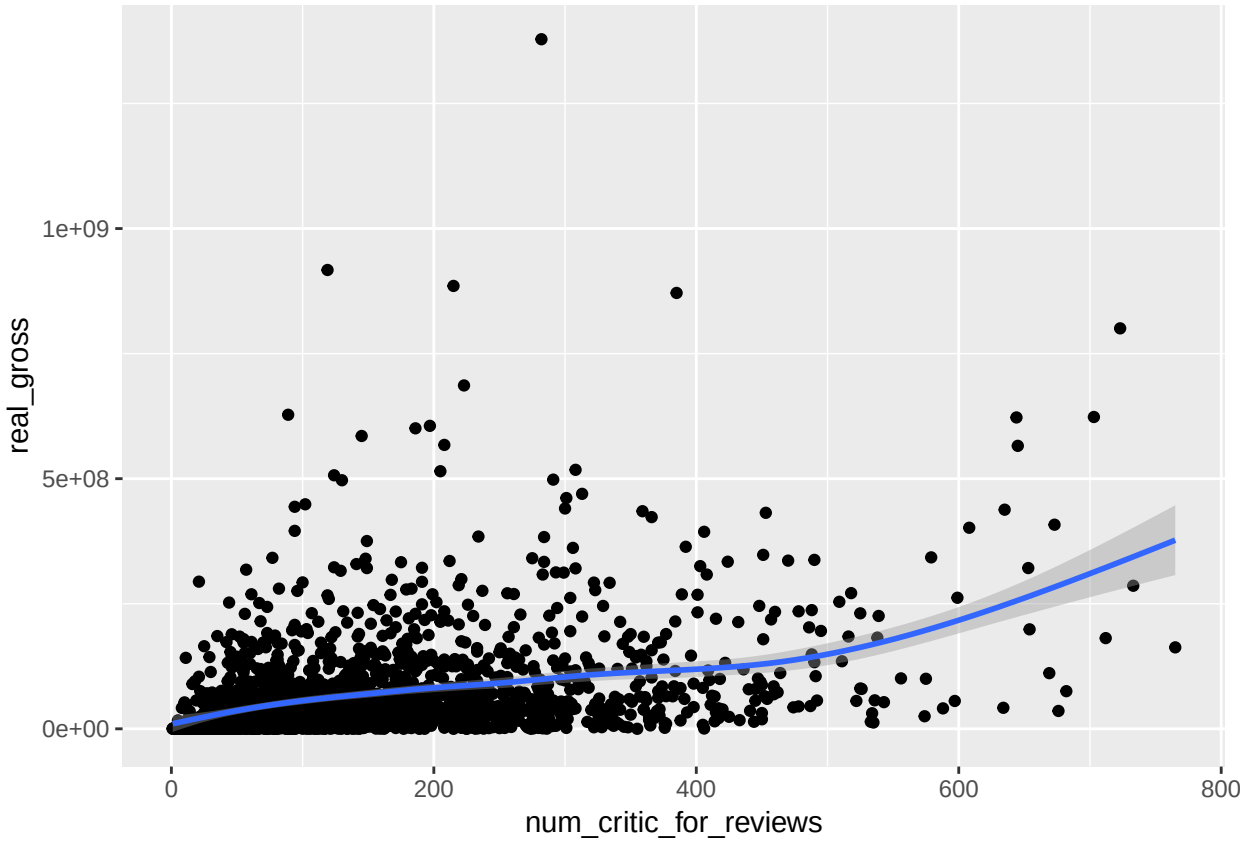



```
ggplot(train, aes(x=num_critic_for_reviews,y=real_gross)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

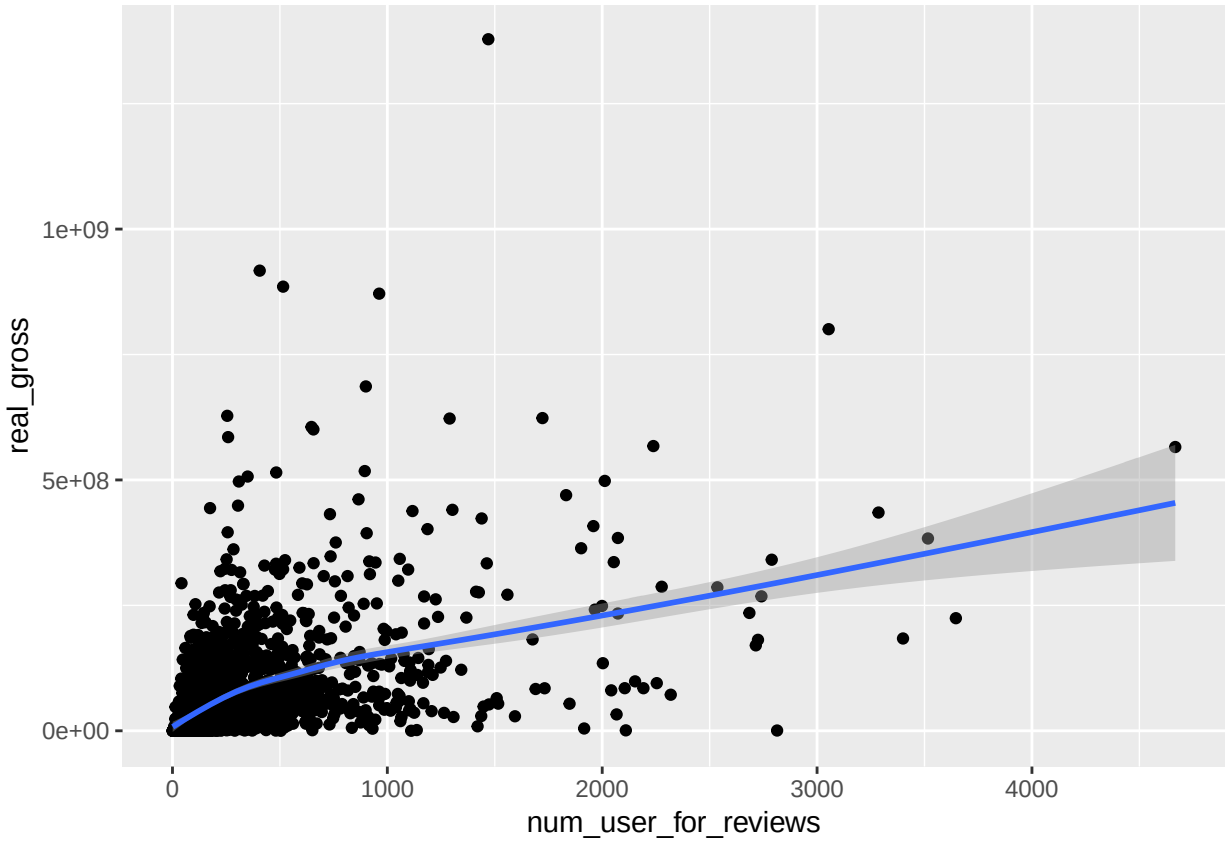
```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



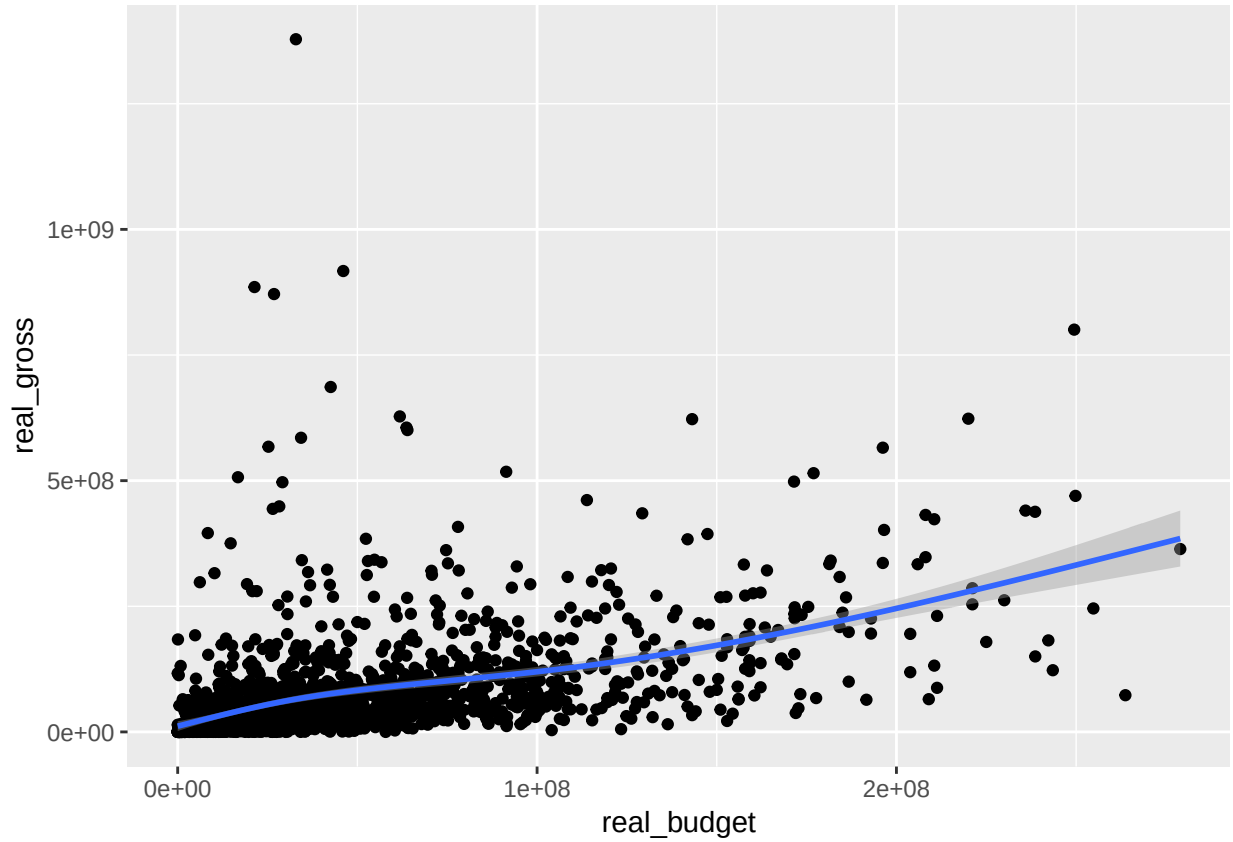

```
ggplot(train, aes(x=num_user_for_reviews,y=real_gross)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
## Warning: Removed 1 rows containing missing values (geom_point).
```



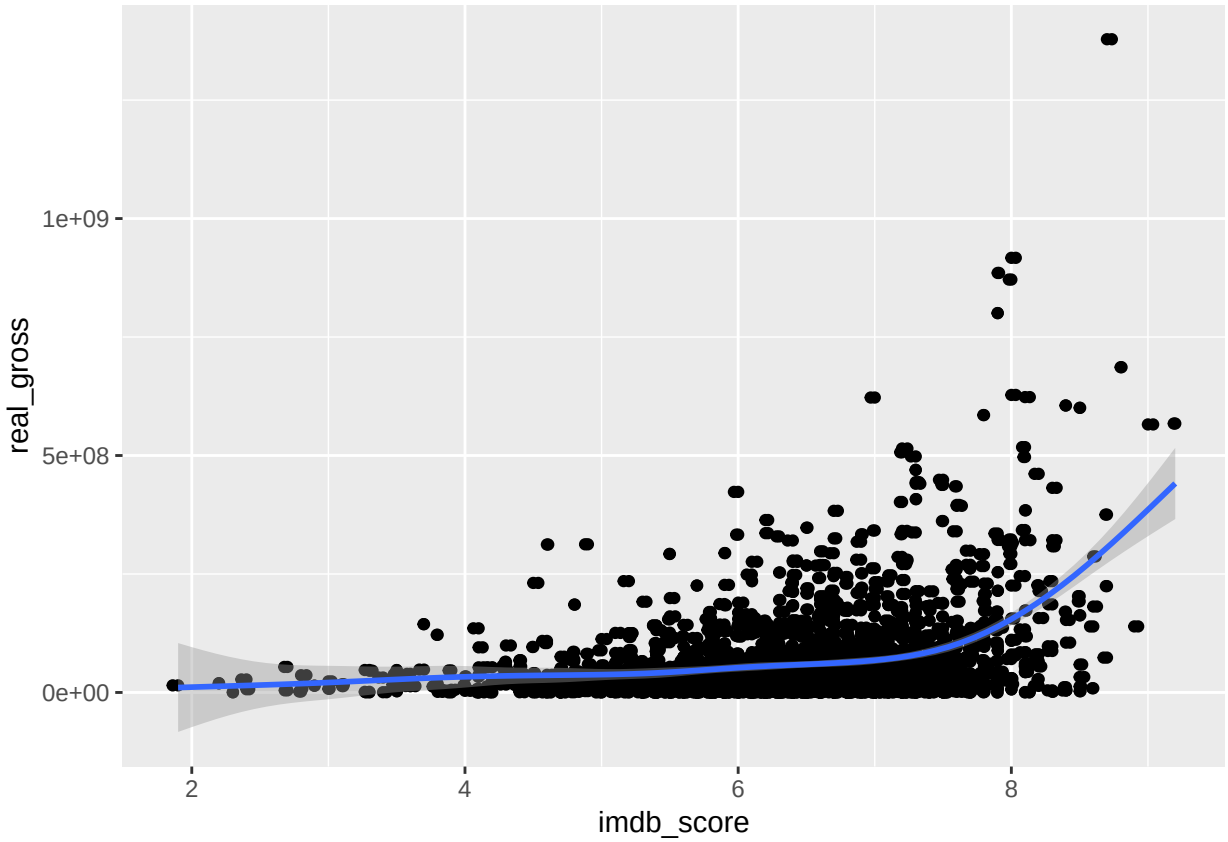

```
ggplot(train, aes(x=real_budget,y=real_gross)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## Warning: Removed 102 rows containing non-finite values (stat_smooth).
## Warning: Removed 102 rows containing missing values (geom_point).
```

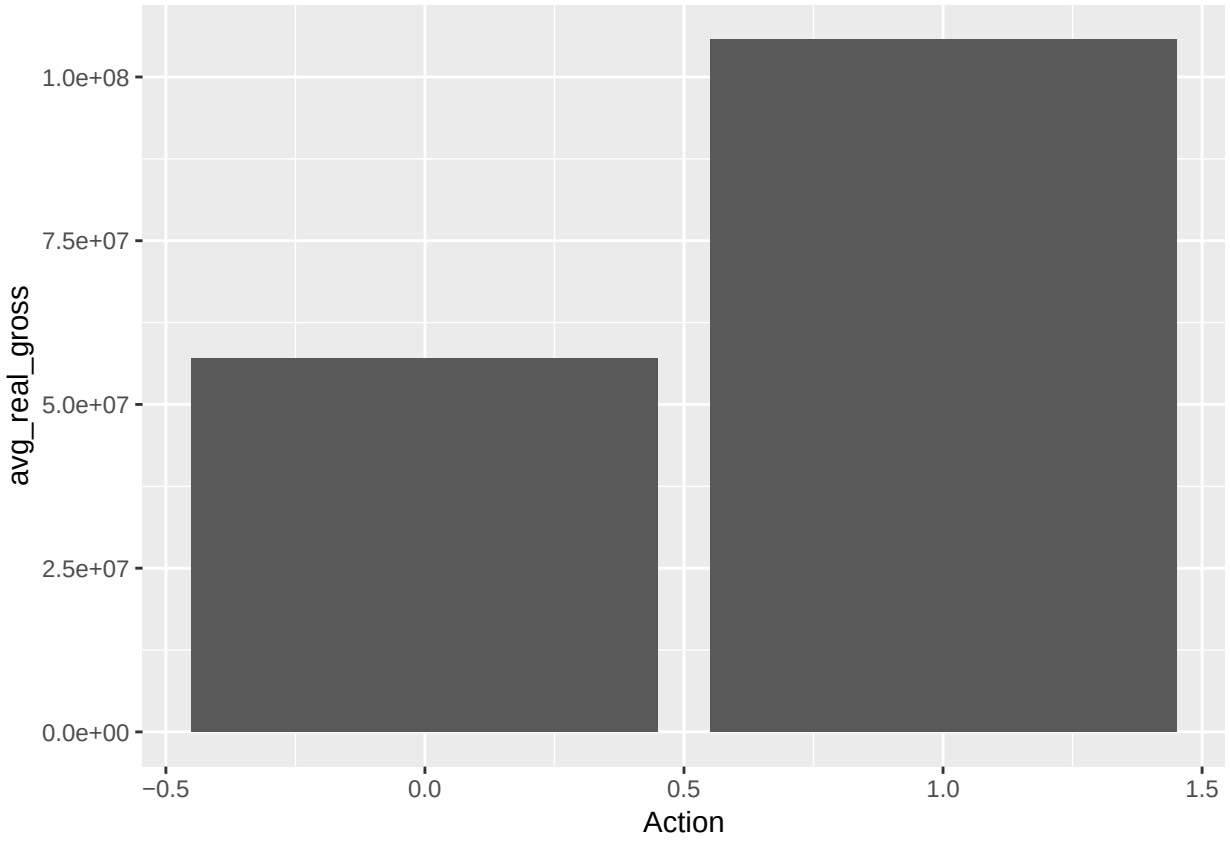


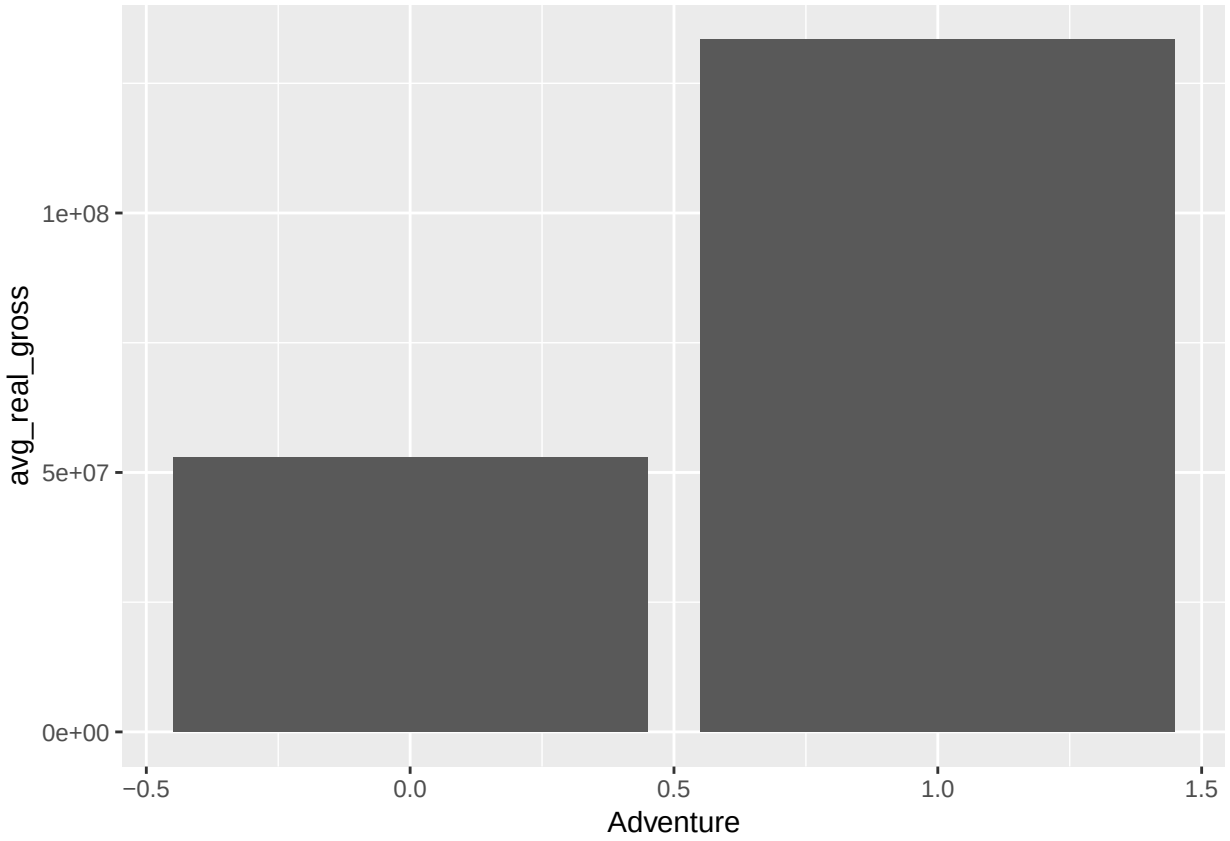

```
ggplot(train, aes(x=imdb_score,y=real_gross)) +
  geom_point() +
  geom_jitter() +
  geom_smooth()
```

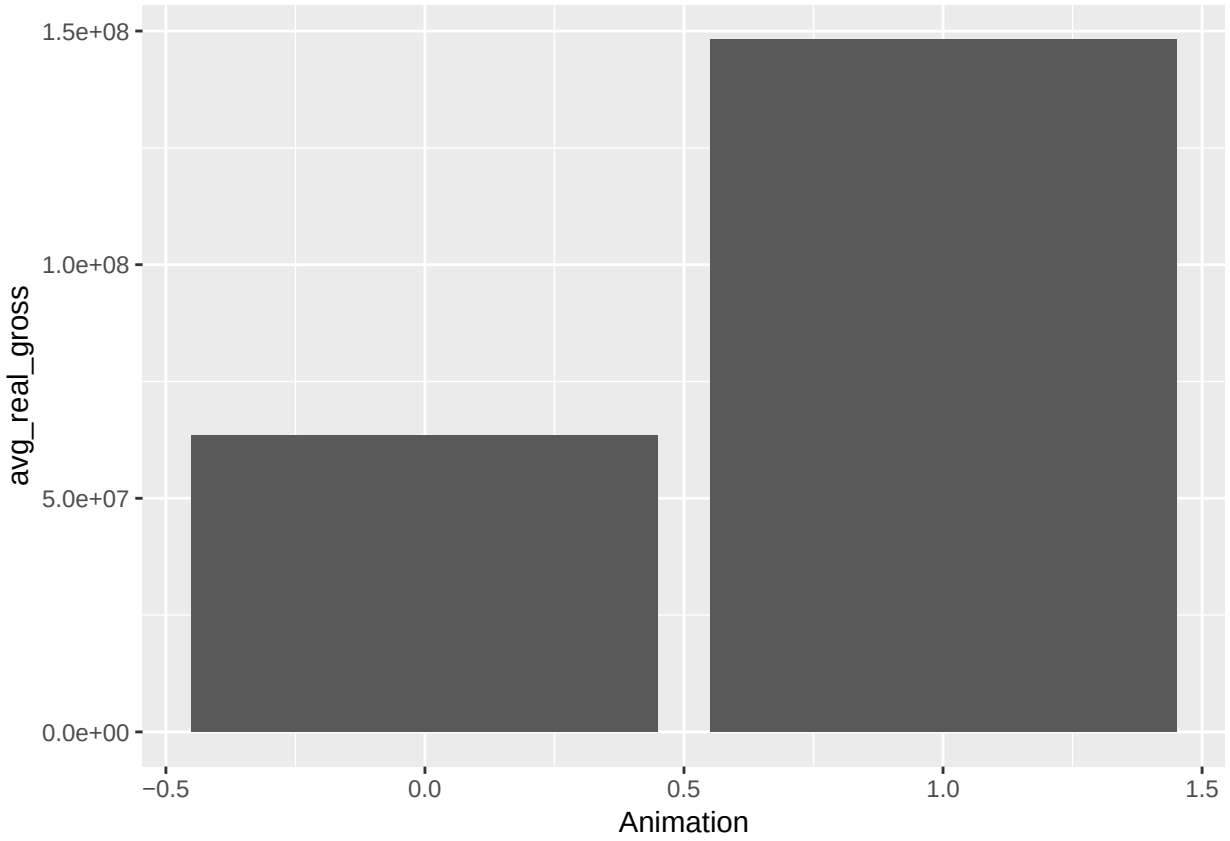
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

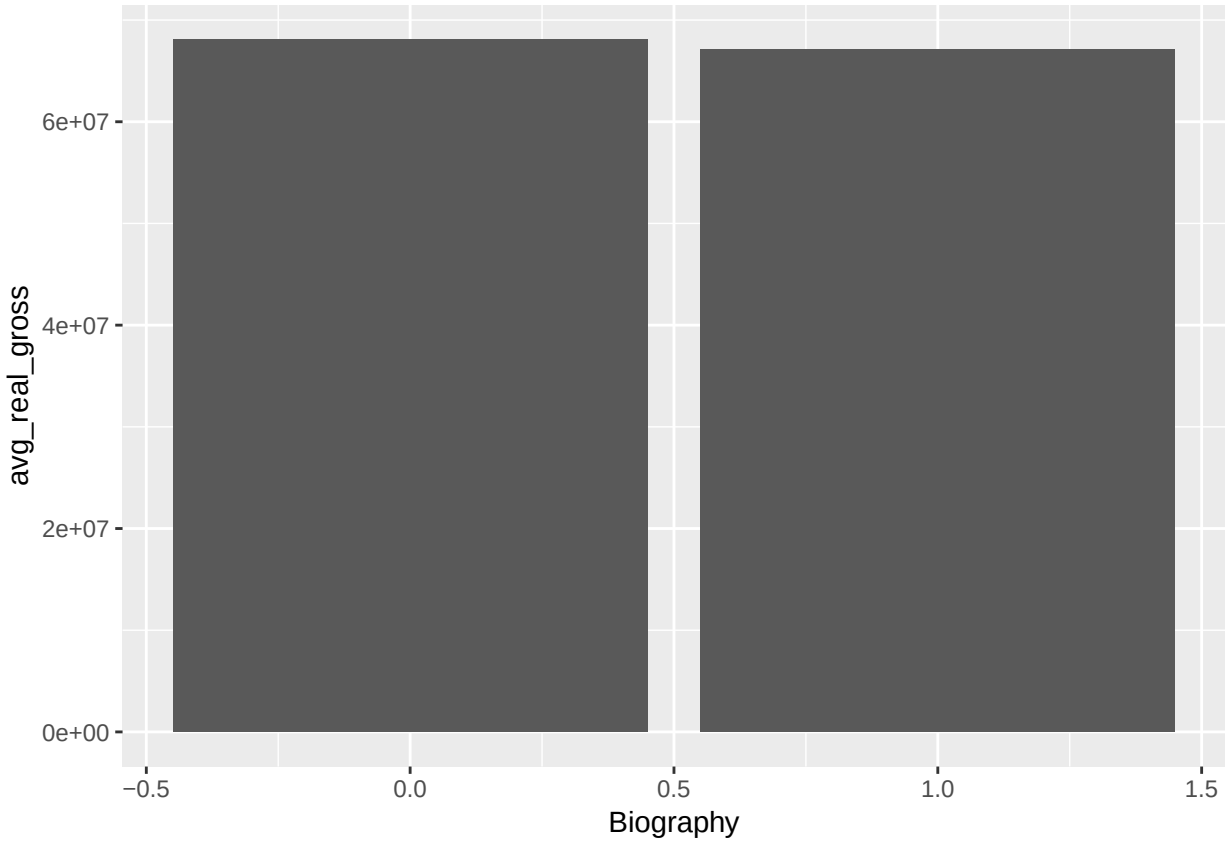


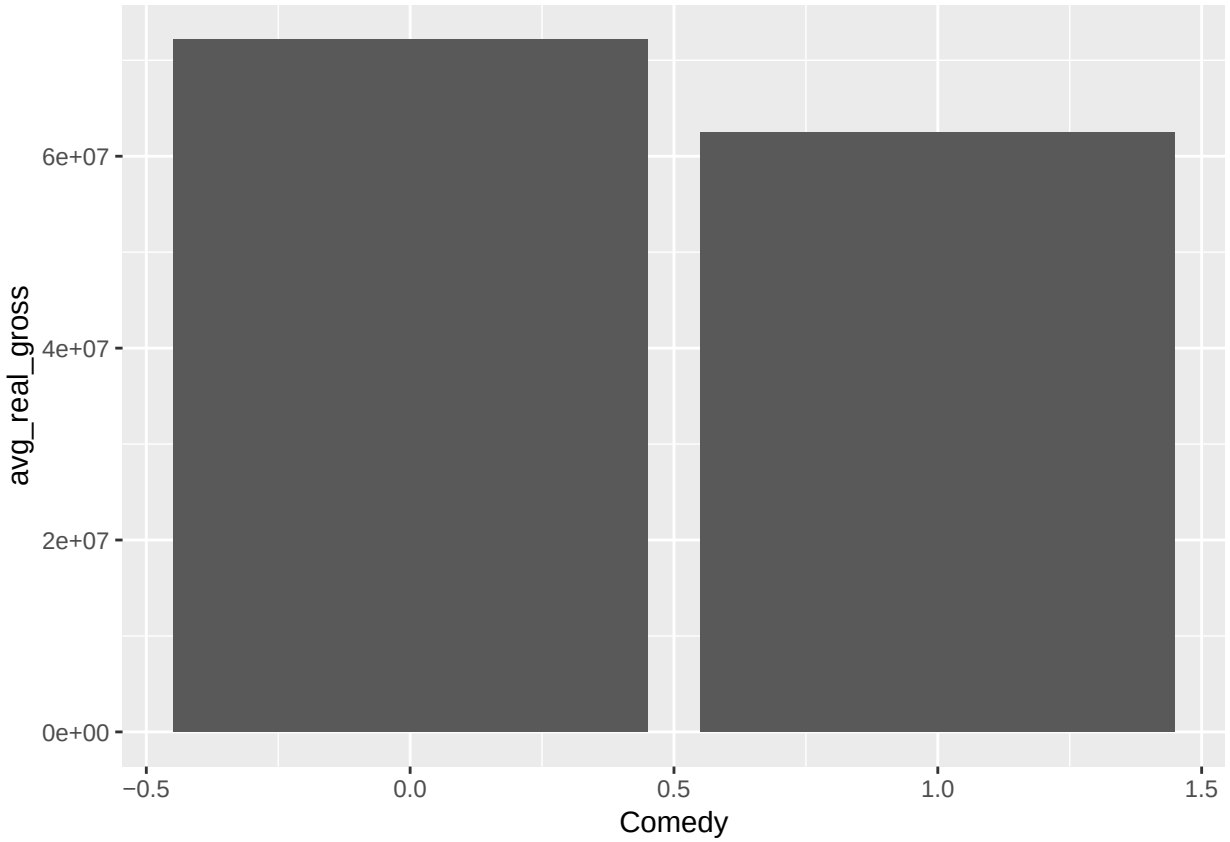

```
plotGenre <- function(df){
    for(i in 22:42){
       col_name <- colnames(train)[[i]]
       g <- df %>%
        group_by_(col_name) %>%
        summarize(avg_real_gross = mean(real_gross)) %>%
         ggplot(df,mapping = aes_string(x=col_name,y="avg_real_gross")) +
         geom_col()
       #print(i)
       print(g)
      }
}
plotGenre(train)
```

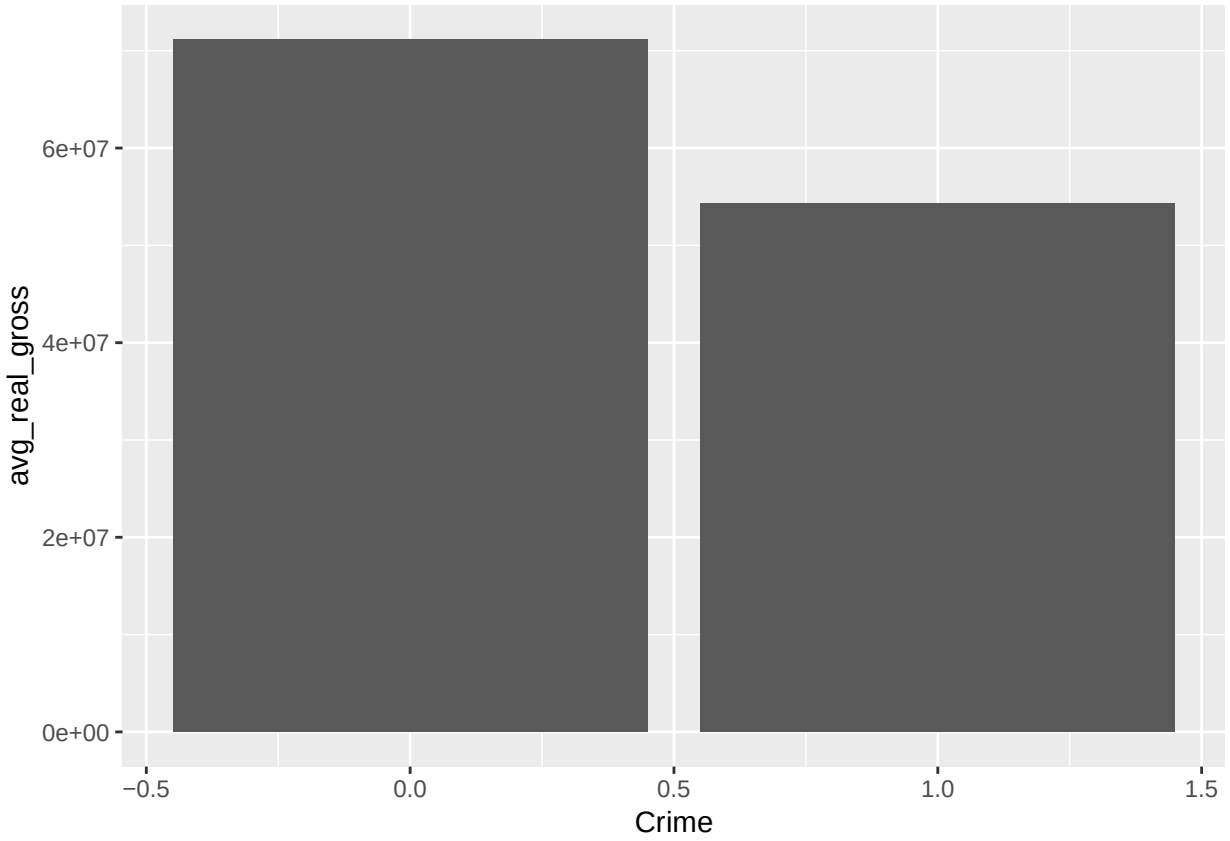

1.0e+08
7.5e+07
5.0e+07
2.5e+07
0.0e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Action

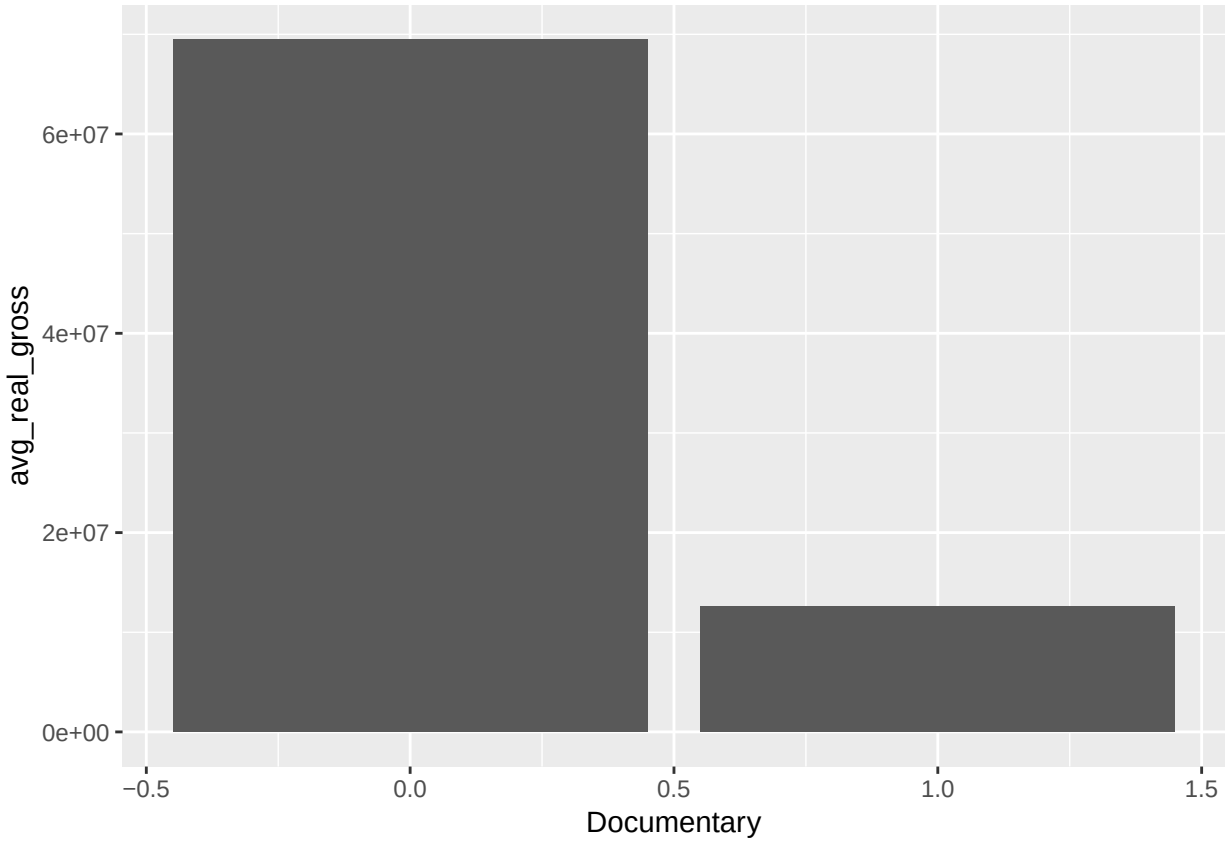

avg_real_gross
1e+08
5e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Adventure

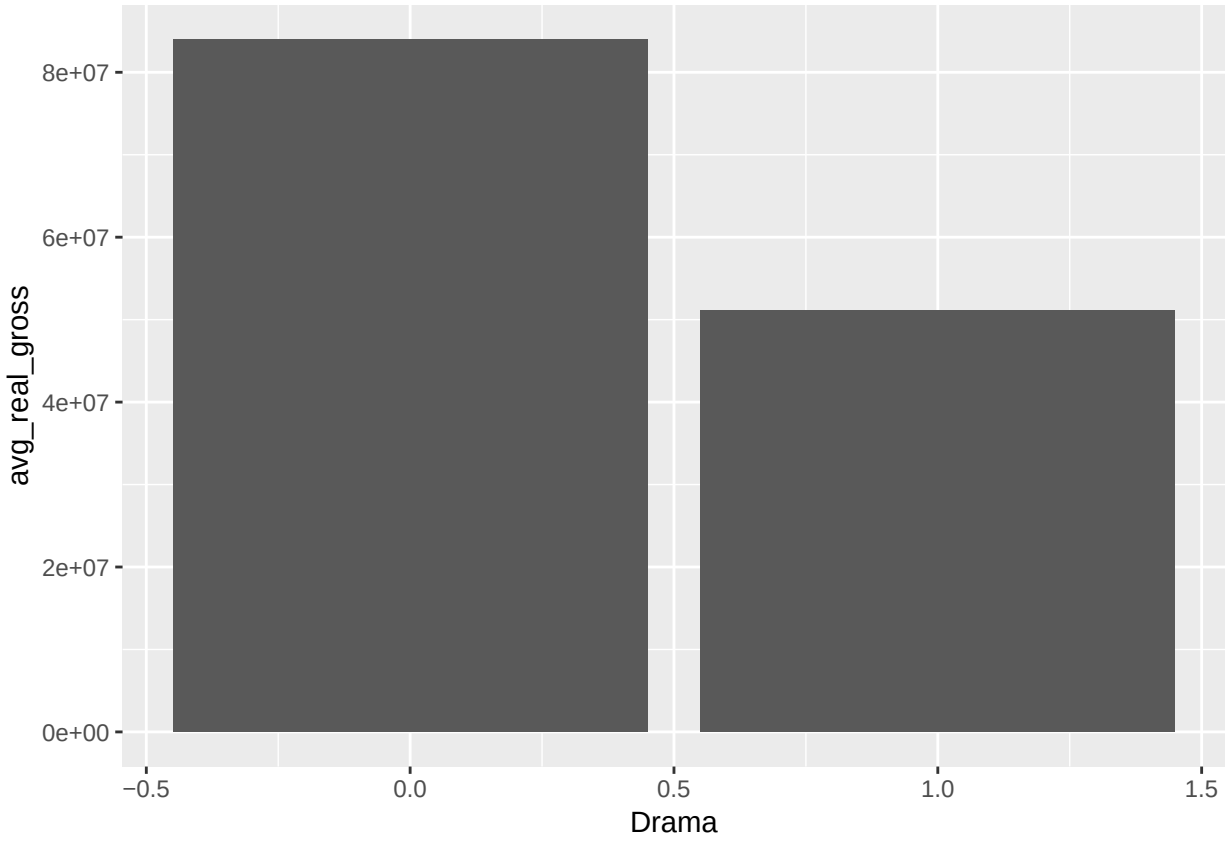

1.5e+08
1.0e+08
5.0e+07
0.0e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Animation

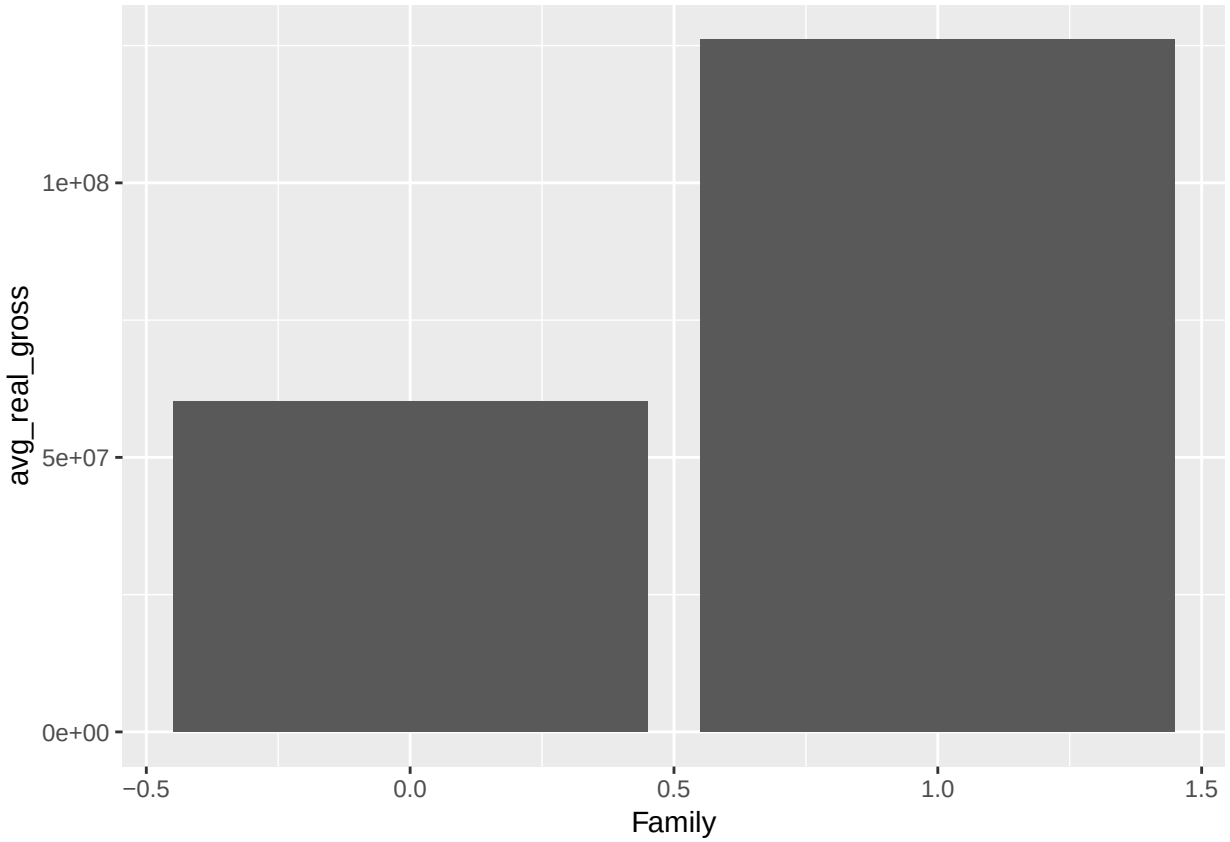

6e+07
4e+07
2e+07
0e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Biography

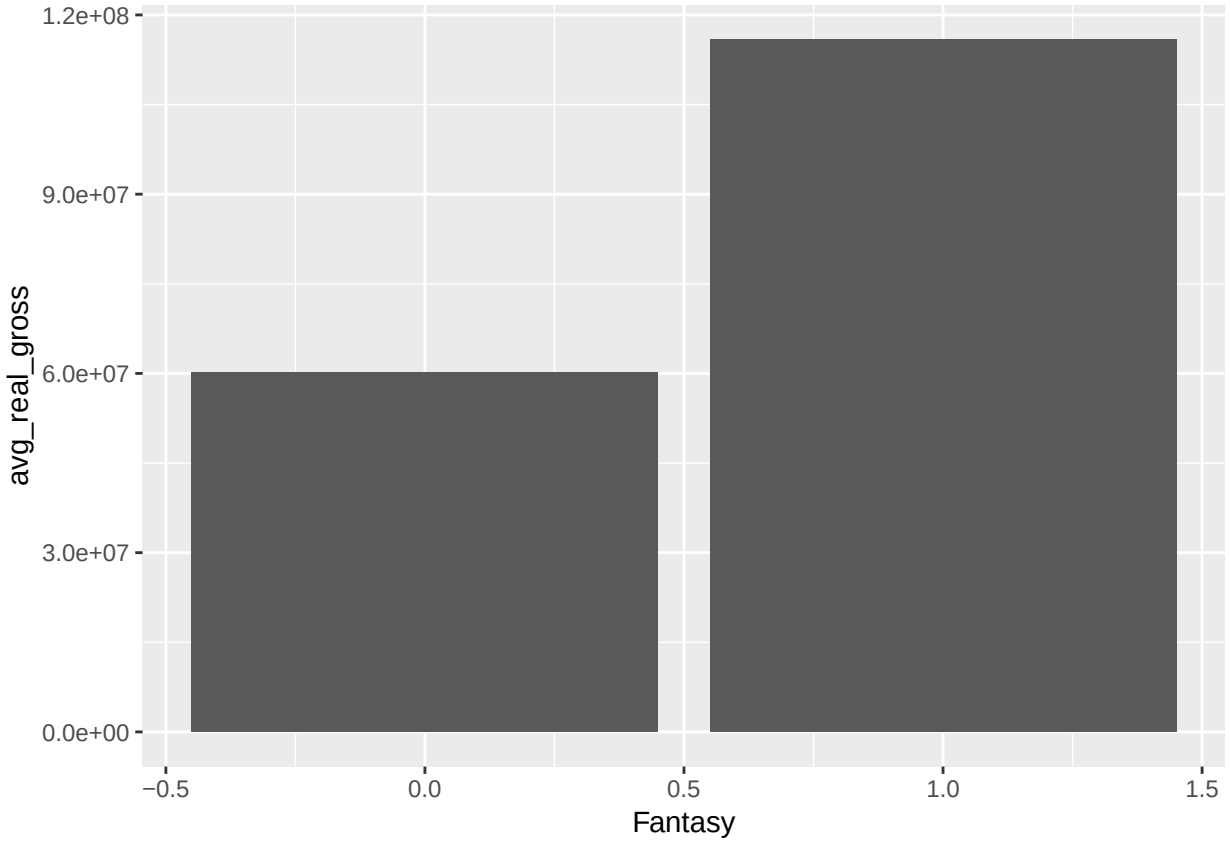

avg_real_gross
6e+07
4e+07
2e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Comedy

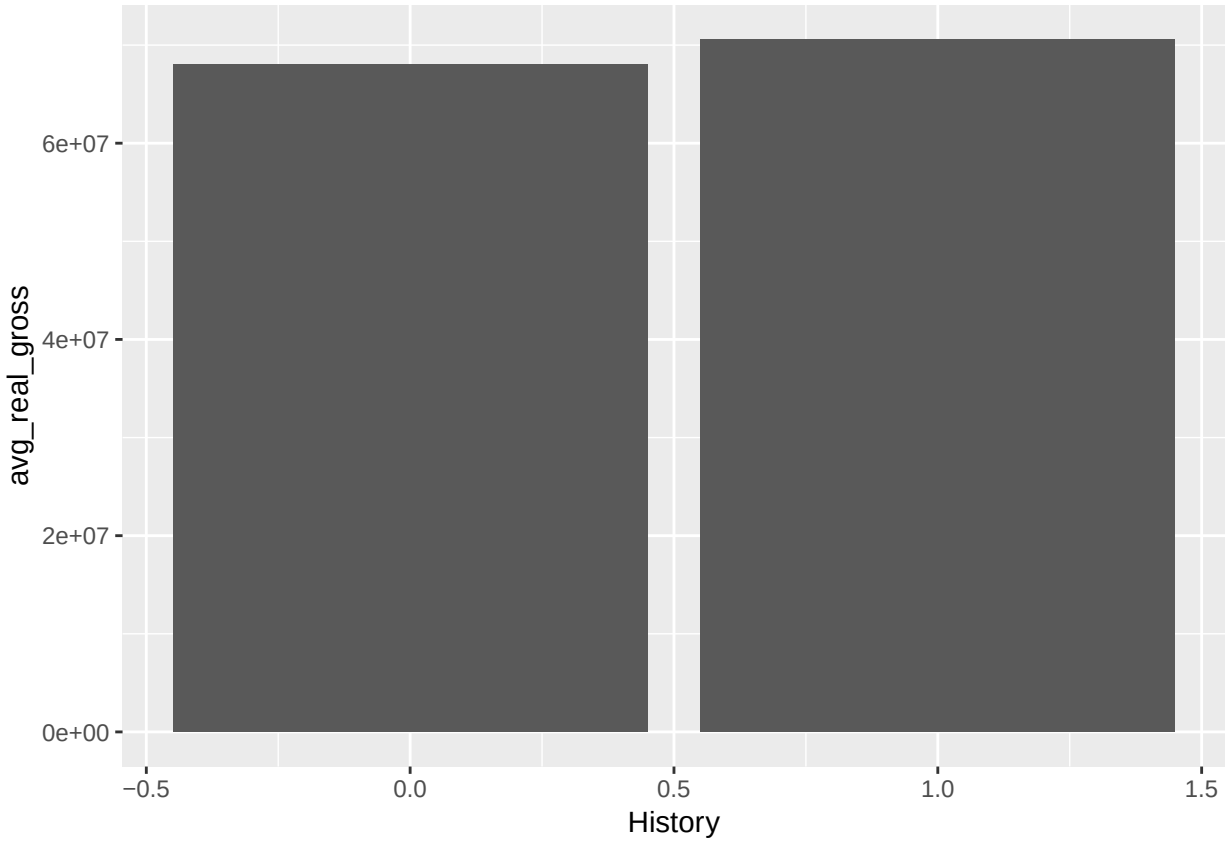

avg_real_gross
6e+07
4e+07
2e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Crime

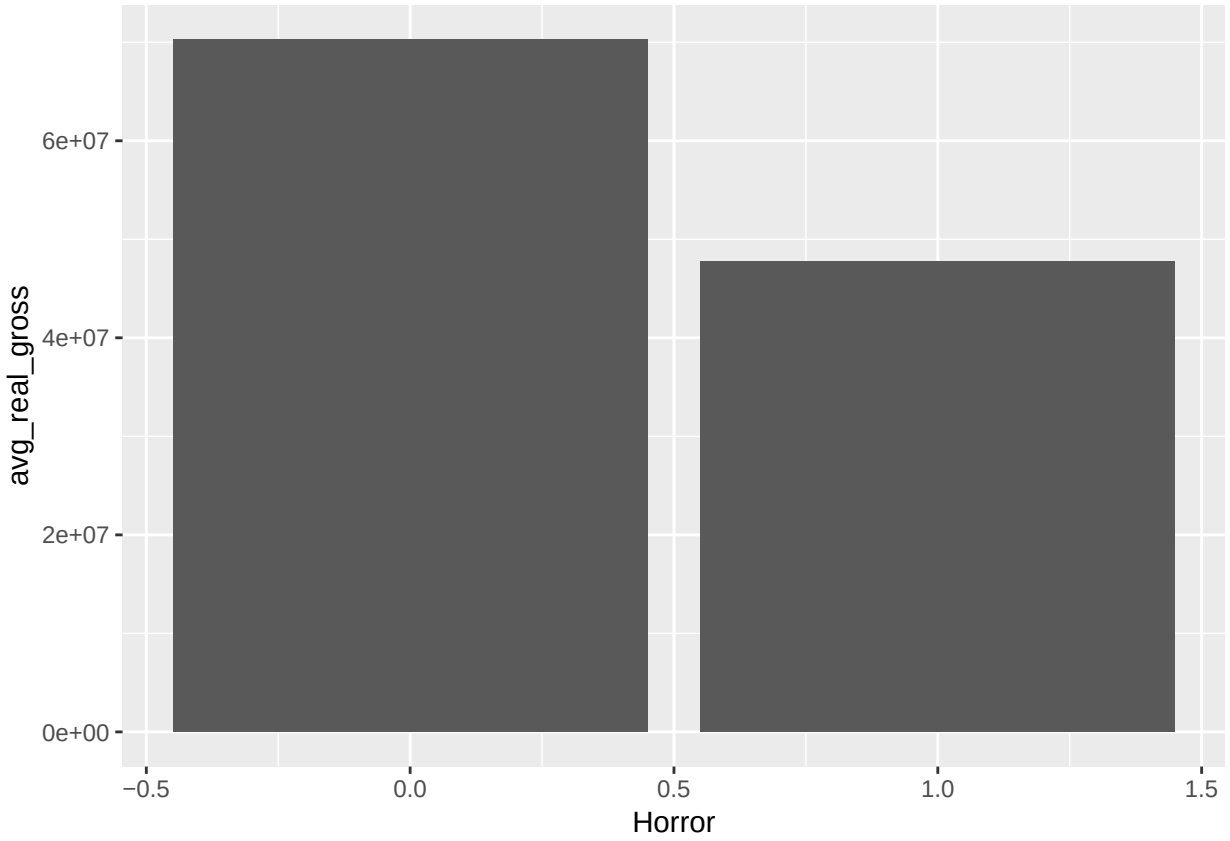

avg_real_gross
6e+07
4e+07
2e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Documentary

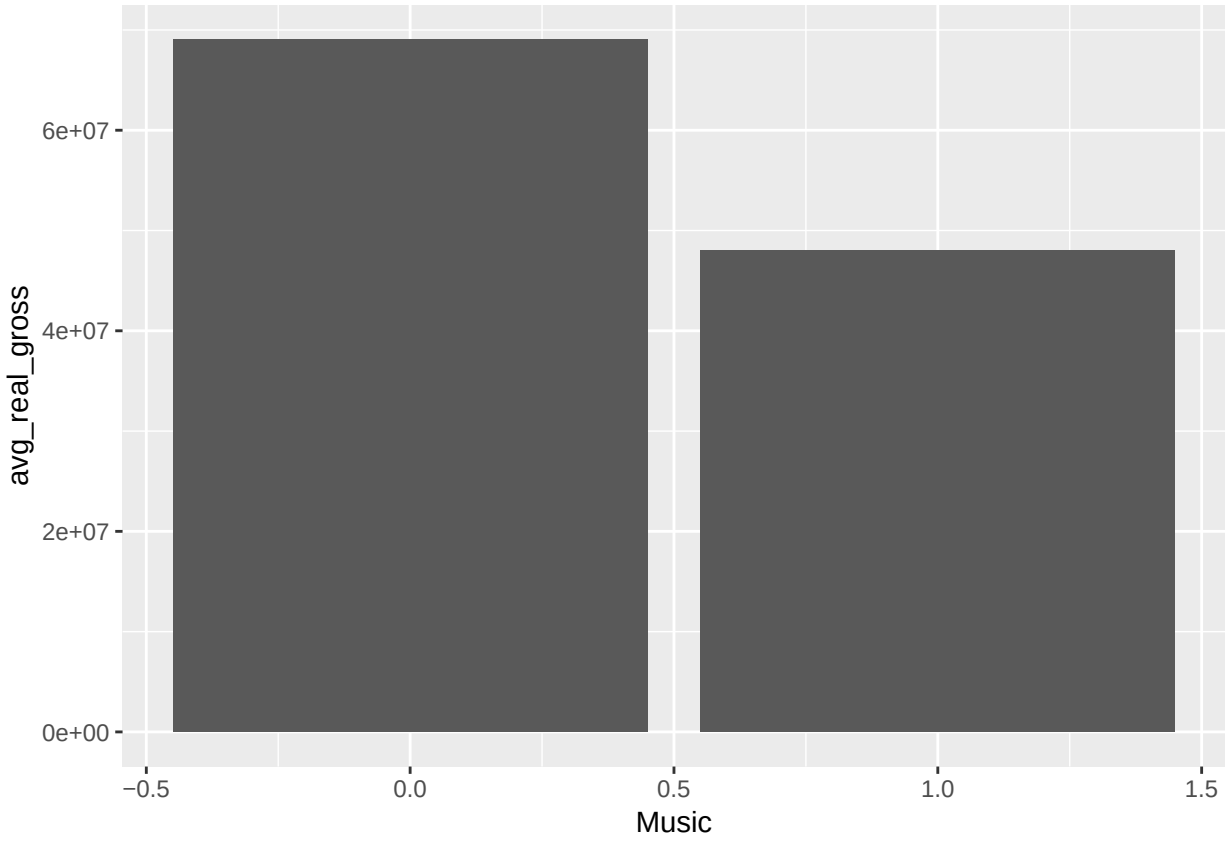

8e+07
6e+07
4e+07
2e+07
0e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Drama

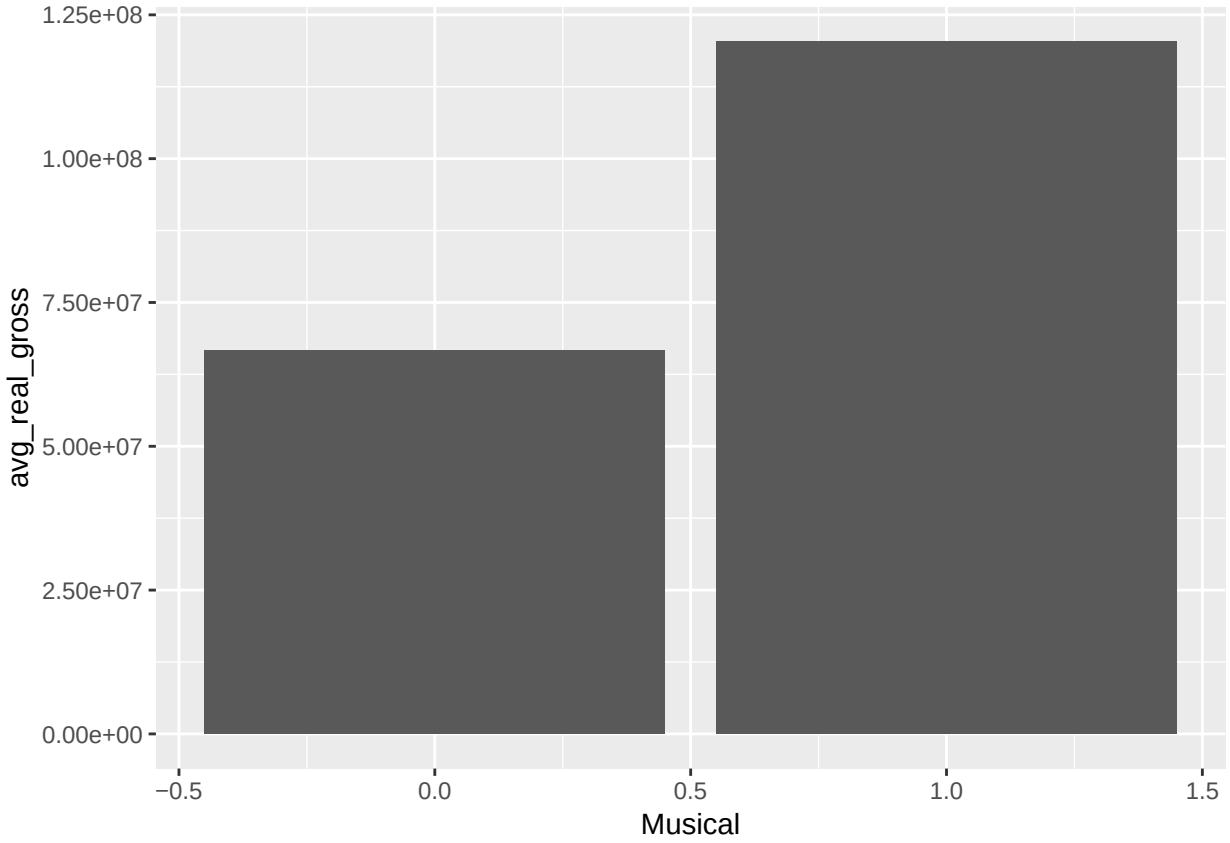

avg_real_gross
1e+08
5e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Family

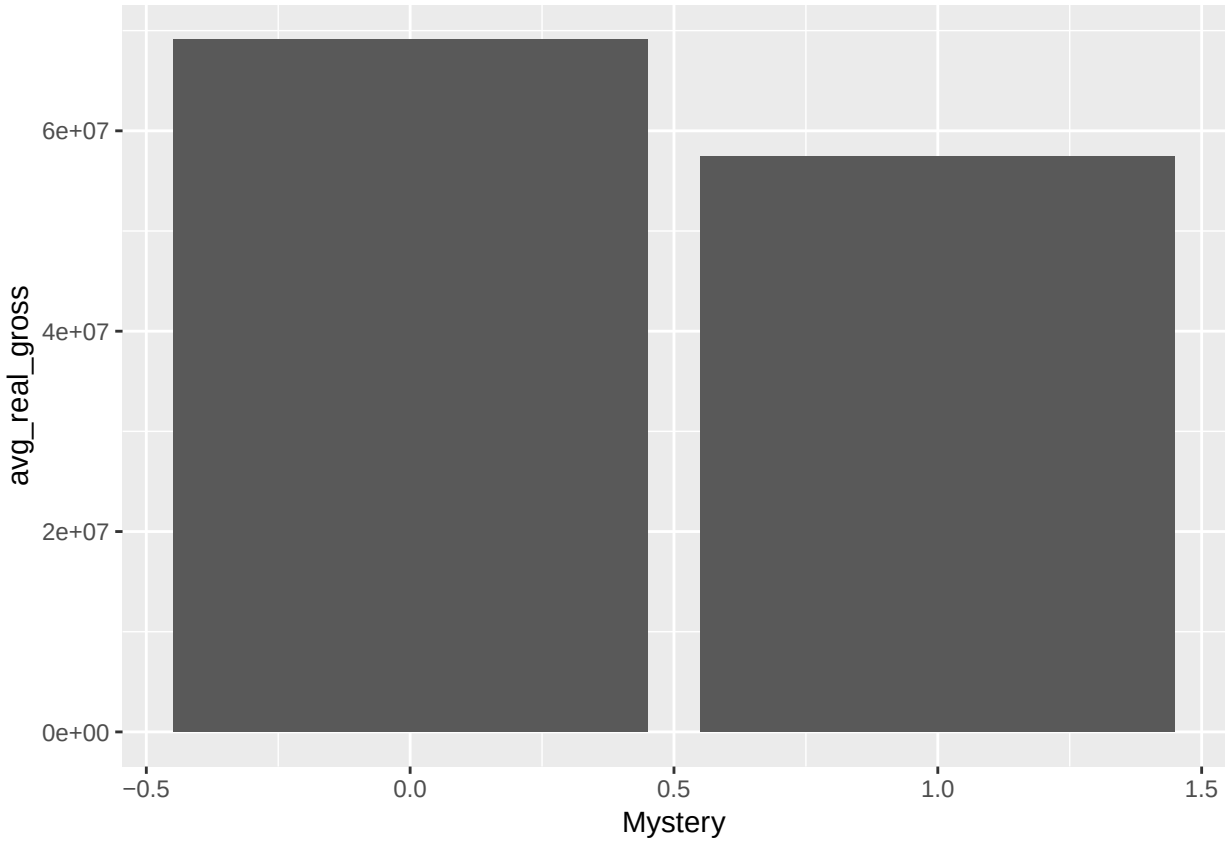

1.2e+08
9.0e+07
6.0e+07
3.0e+07
0.0e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Fantasy

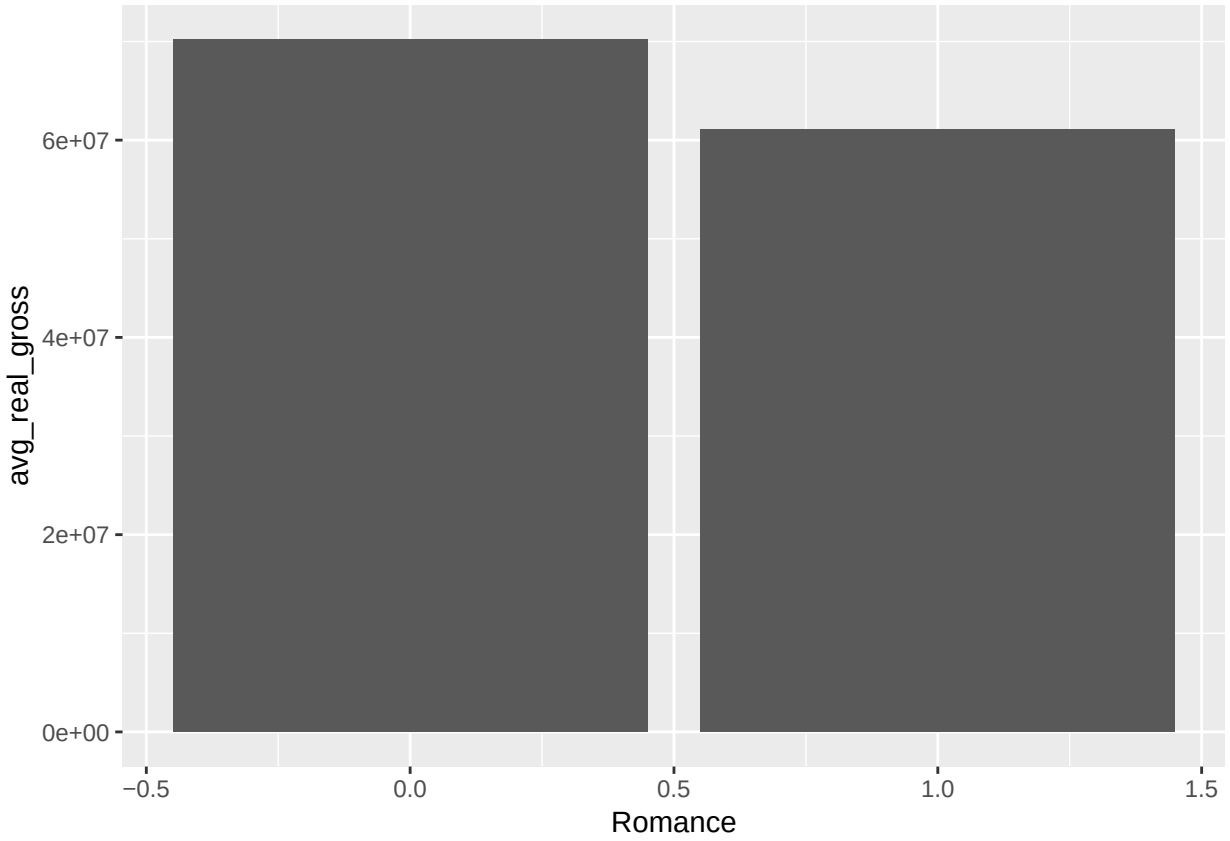

avg_real_gross
6e+07
4e+07
2e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
History

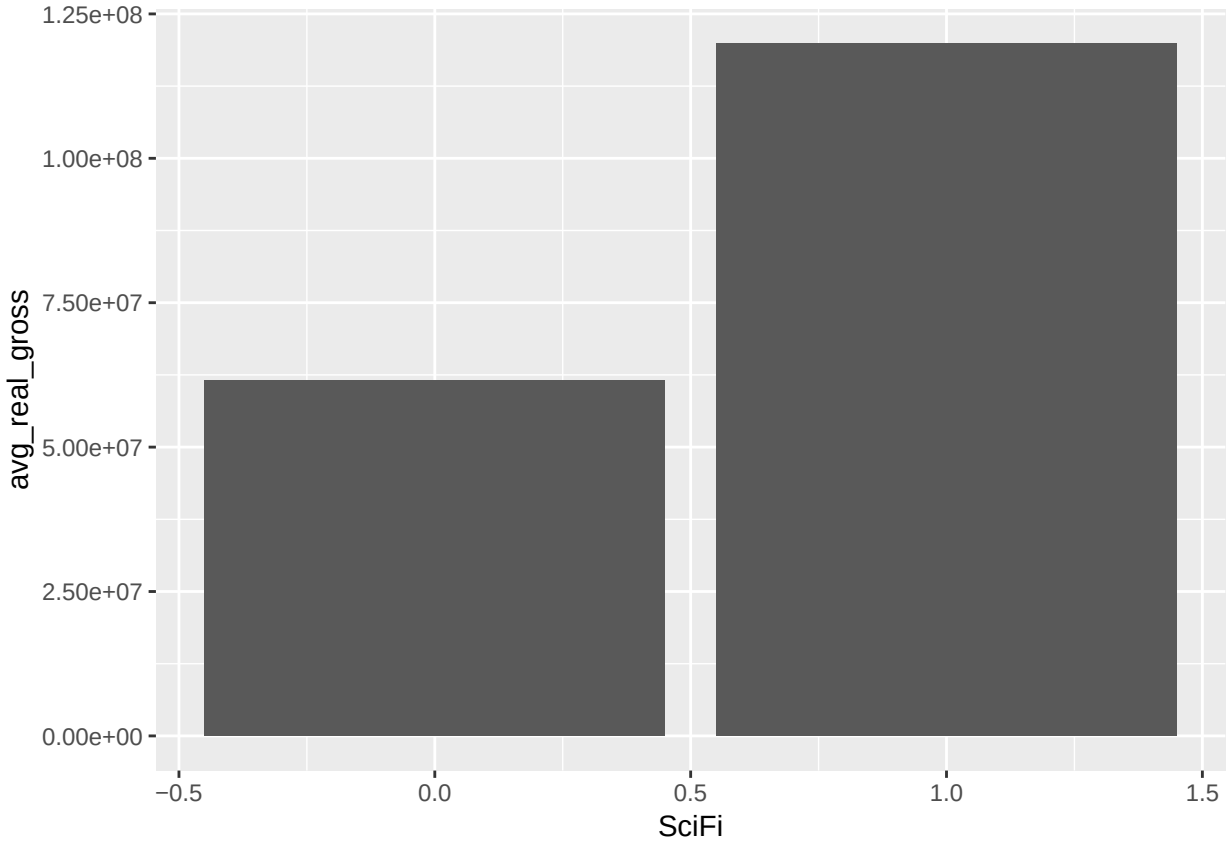

avg_real_gross
6e+07
4e+07
2e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Horror

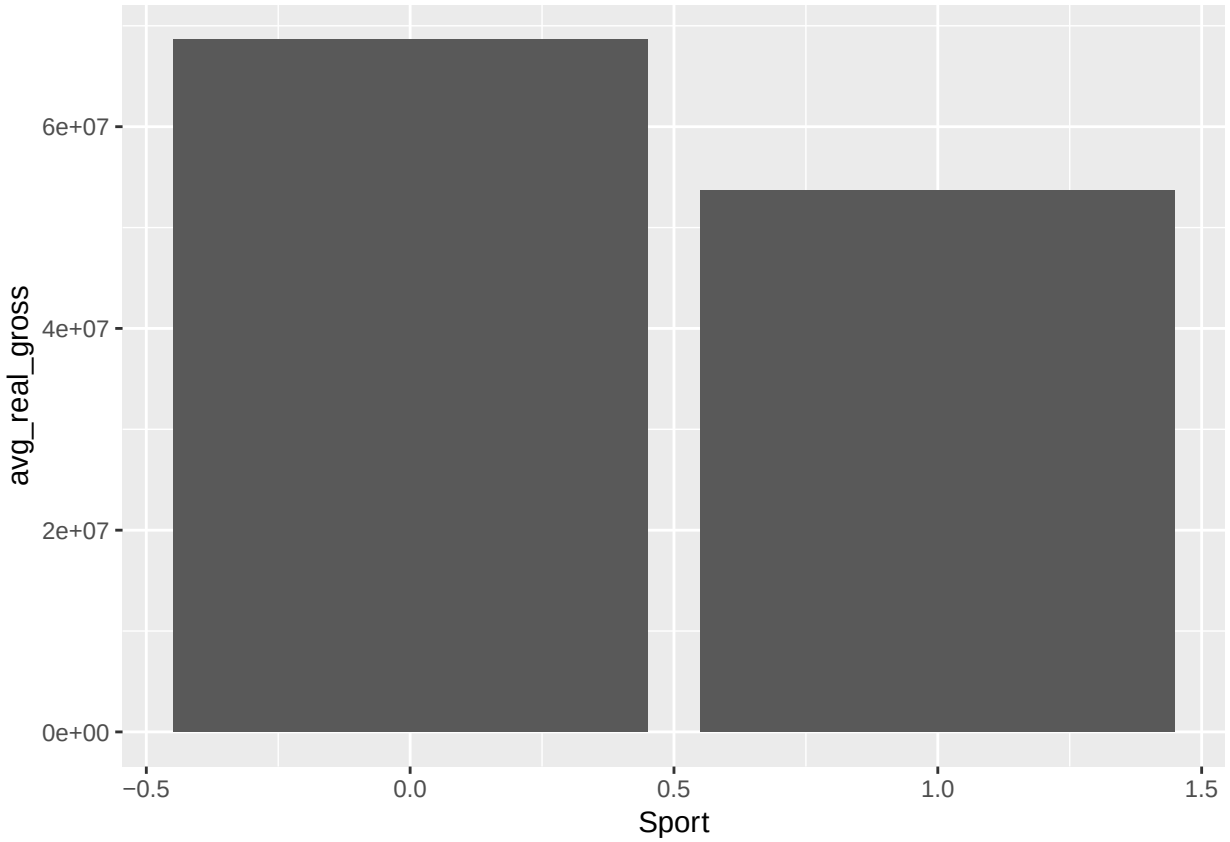

6e+07
4e+07
2e+07
0e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Music

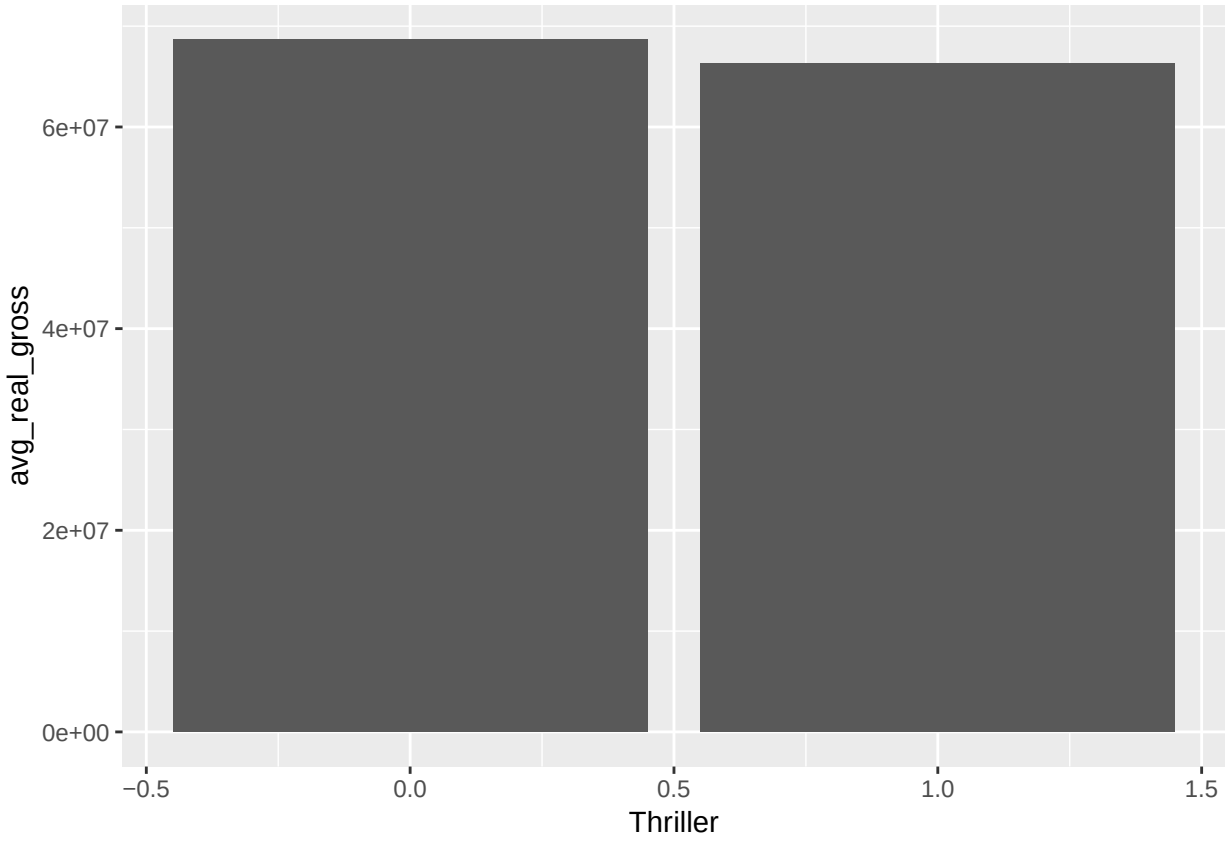

1.25e+08
1.00e+08
7.50e+07
5.00e+07
2.50e+07
0.00e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Musical

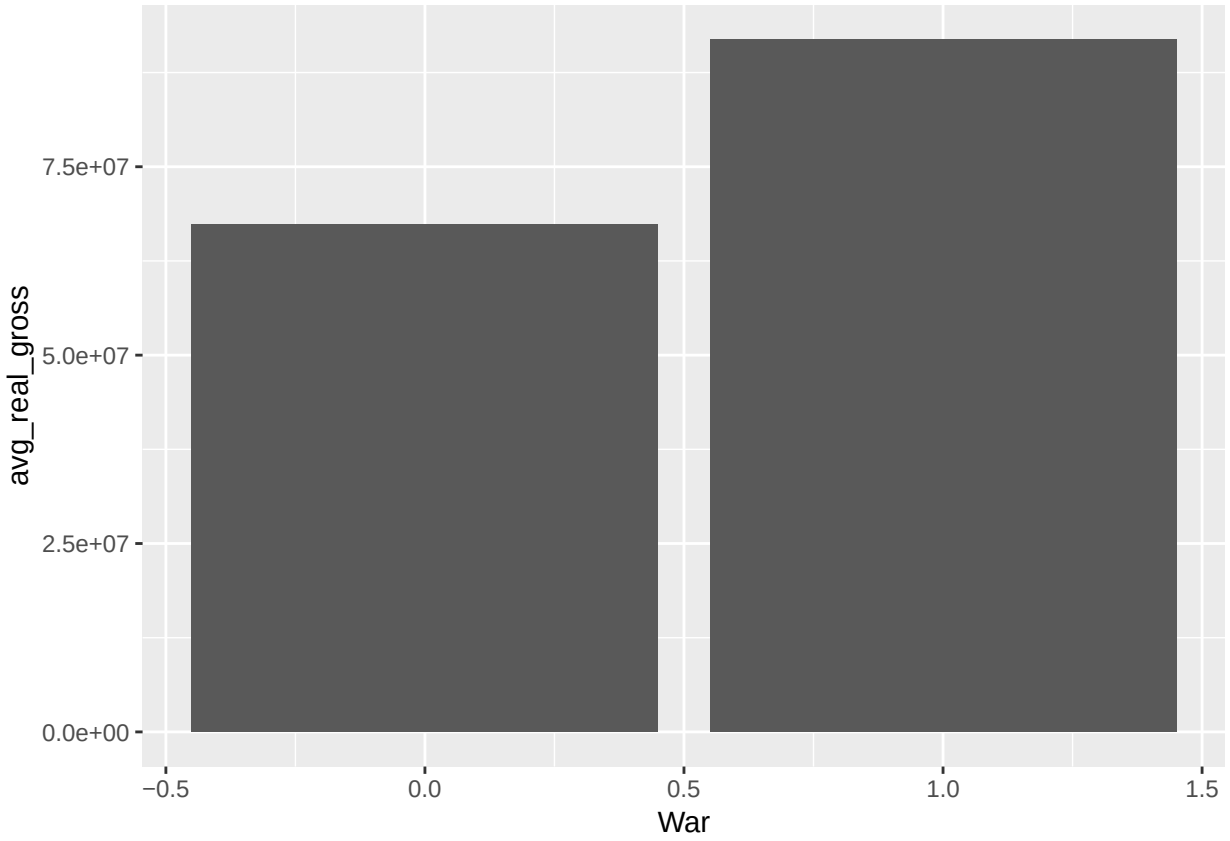

avg_real_gross
6e+07
4e+07
2e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Mystery

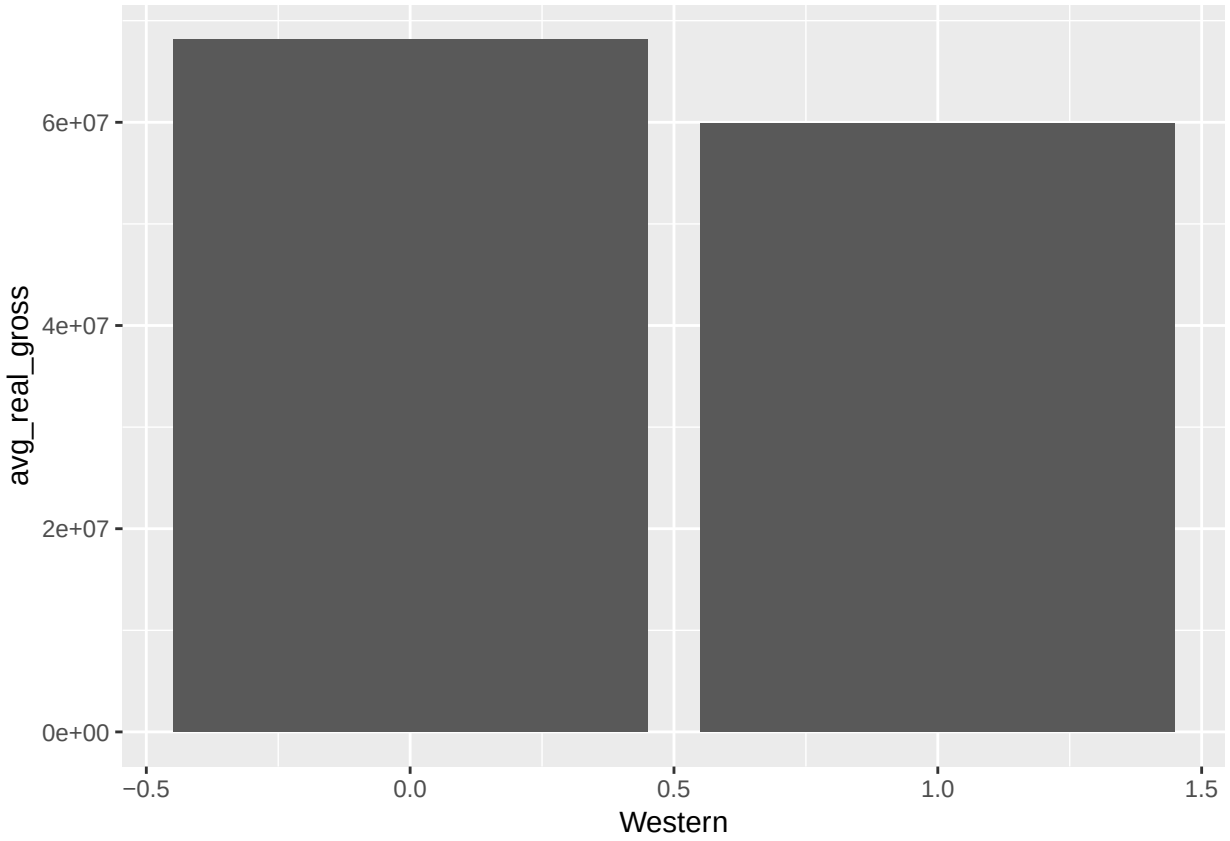

avg_real_gross
6e+07
4e+07
2e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Romance


1.25e+08
1.00e+08
7.50e+07
5.00e+07
2.50e+07
0.00e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
SciFi


6e+07
4e+07
2e+07
0e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Sport


avg_real_gross
6e+07
4e+07
2e+07
0e+00
−0.5
0.0
0.5
1.0
1.5
Thriller


avg_real_gross
7.5e+07
5.0e+07
2.5e+07
0.0e+00
−0.5
0.0
0.5
1.0
1.5
War


6e+07
4e+07
2e+07
0e+00
avg_real_gross
−0.5
0.0
0.5
1.0
1.5
Western