TidyTuesday - Analyzing Movie Profits

2022-08-12

Horror Movies and Profit

https://github.com/rfordatascience/tidytuesday/tree/master/data/2018/2018-10-23

Horror seems to be a great category for movies in terms of profit, a phenomenon described by FiveThirtyEight. Horror movies usually turn out to be great investments, let's see how this is happening with data.

Data Source: TidyTuesday

A weekly social data project in R, borne out of the R4DS Online Learning Community and the R for Data Science textbook, an emphasis was placed on understanding how to summarize and arrange data to make meaningful charts with ggplot2, tidyr, dplyr, and other tools in the tidyverse ecosystem.

Custom Magic

An utility to help me pretty print graphs.

```
`&.gg` <- function(e1, e2) e2(e1)
# see https://stackoverflow.com/a/53599958/19042045
pprint <- function(p) {
    # pass by value somehow...
    p$labels$x <- p$labels$x %>% gsub("_", " ", .) %>% str_to_title()
    p$labels$y <- p$labels$y %>% gsub("_", " ", .) %>% str_to_title()
    p$labels$colour <- p$labels$colour %>% gsub("_", " ", .) %>% str_to_title()
    return(p)
}
```

Cleansing & Preprocessing

```
# horror_movies %>%
#  View()

movie_profit <- readr::read_csv("https://github.com/rfordatascience/tidytuesday/raw/master/data/2018/20

library(lubridate)
movie_profit <- movie_profit %>%
  mutate(release_date = mdy(release_date))

movie_profit <- movie_profit %>%
  mutate(profit = worldwide_gross - production_budget,
         markup = profit / production_budget)

movie_profit <- movie_profit %>%
  rename(idx = ...1)

movie_profit %>%
  colnames()
```

```
##  [1] "idx"               "release_date"      "movie"
##  [4] "production_budget" "domestic_gross"    "worldwide_gross"
##  [7] "distributor"       "mpaa_rating"       "genre"
## [10] "profit"            "markup"
```

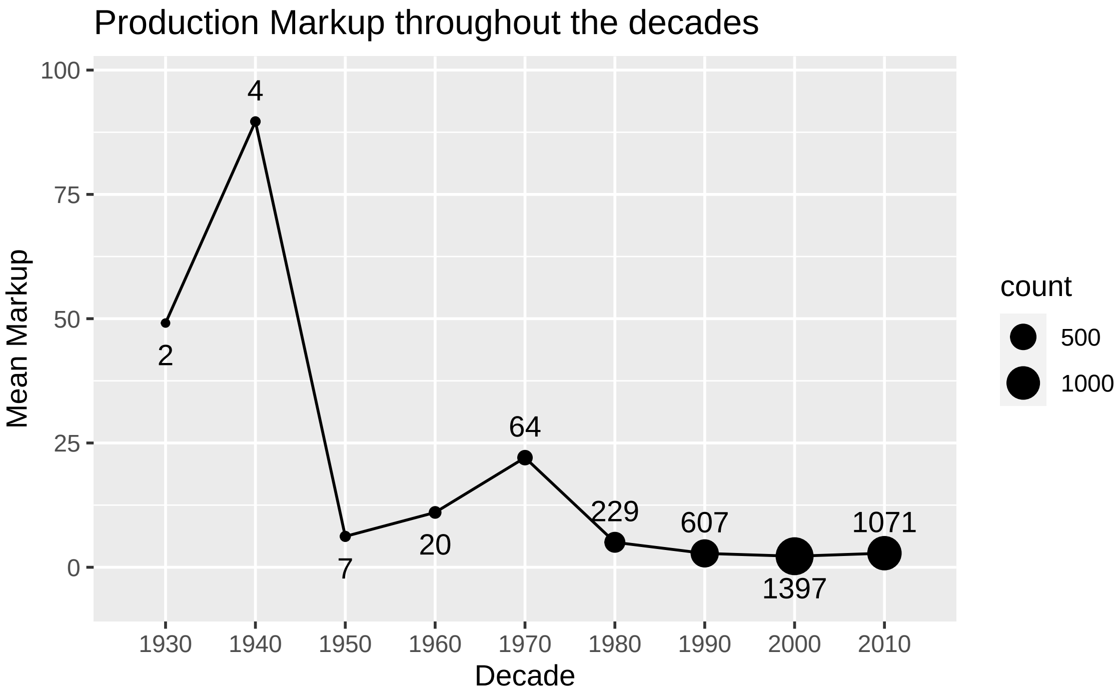
Quality Control

```
library(scales)

movie_profit %>%
  mutate(year=release_date %>% year) %>%
  group_by(genre) %>%
  summarise(min(year), max(year))
```

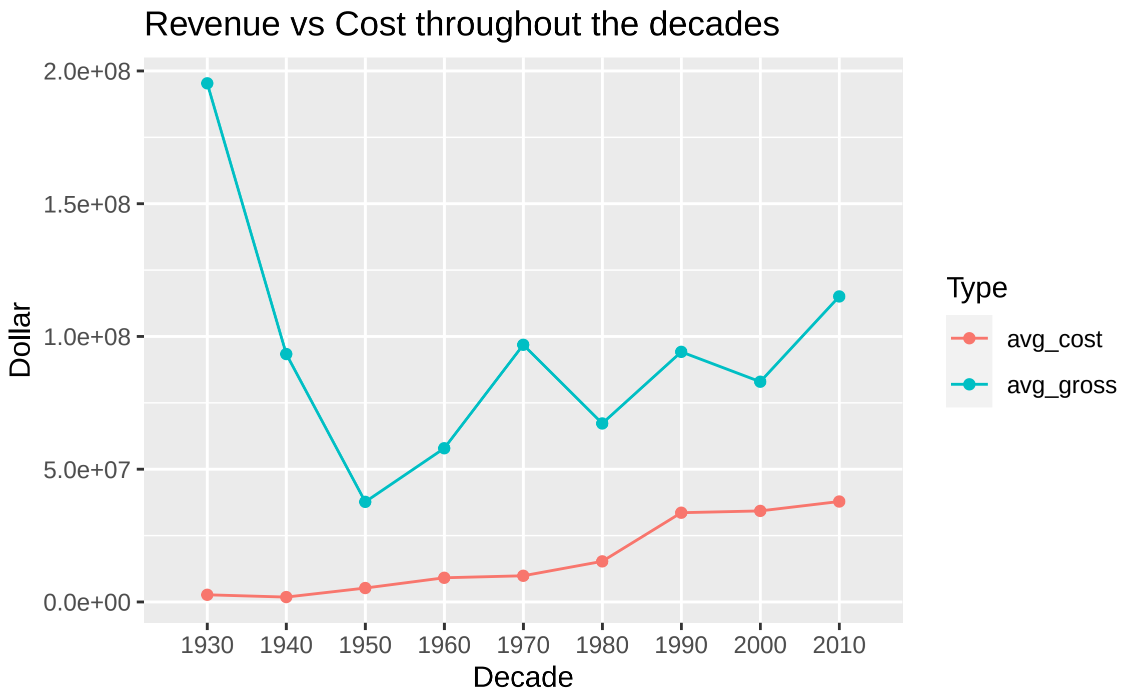
```
## # A tibble: 5 x 3
##   genre     ‘min(year)‘ ‘max(year)‘
##   <chr>           <dbl>       <dbl>
## 1 Action           1960        2018
## 2 Adventure        1940        2019
## 3 Comedy           1936        2018
## 4 Drama            1939        2018
## 5 Horror           1973        2018
```

The data is slightly imbalanced because the Horror category only appeared after 1973, while Comedy dated back as far as 1936.

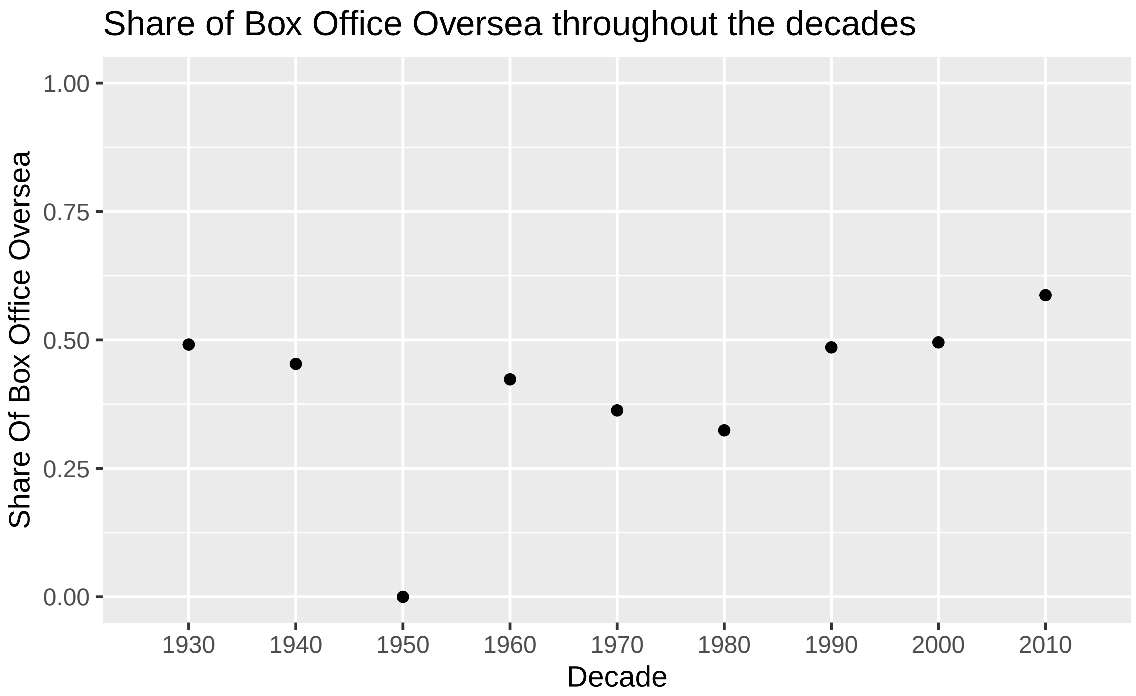
```
# automatic vjsut for count
auto_vjust <- function(data) {
  vjust <- sign(diff(data));
  vjust <- c(vjust, -1*tail(vjust, 1));
  vjust <- ifelse(vjust == 1, 2, -1);
  return(vjust)
}

movie_profit %>%
  mutate(year=release_date %>% year) %>%
  mutate(decade = year%/%10*10) %>%
  group_by(decade) %>%
  summarise(mean_markup=mean(markup), count=n()) %>%
  ggplot(aes(decade, mean_markup)) +
  geom_point(aes(size=count)) +
  geom_line() +
  geom_text(aes(label = count, vjust=auto_vjust(mean_markup))) +
  scale_y_continuous(expand = expansion(mult = c(0.15, 0.15))) +
  scale_x_discrete(limits=seq(1930, 2010, 10), expand = expansion(mult = c(0.1, 0.1))) +
  ggtitle("Production Markup throughout the decades") & pprint
```



```
movie_profit %>%
  mutate(year=release_date %>% year) %>%
  mutate(decade = year%/%10*10) %>%
  group_by(decade) %>%
  summarise(avg_cost=mean(production_budget), avg_gross=mean(worldwide_gross)) %>%
  gather(avg_cost, avg_gross, key = "type", value = "dollar") %>%
  ggplot(aes(decade, dollar)) +
  geom_point(aes(color=type)) +
  geom_line(aes(color=type)) +
  scale_x_discrete(limits=seq(1930, 2010, 10), expand = expansion(mult = c(0.1, 0.1))) +
  ggtitle("Revenue vs Cost throughout the decades") & pprint
```



```
movie_profit %>%
  mutate(year=release_date %>% year) %>%
  mutate(decade = year%/%10*10) %>%
  group_by(decade) %>%
  summarise(total_domestic=sum(domestic_gross), total_worldwide=sum(worldwide_gross)) %>%
  ggplot(aes(decade, 1-total_domestic/total_worldwide)) +
  geom_point() +
  ylim(0.0, 1.0) +
  scale_x_discrete(limits=seq(1930, 2010, 10), expand = expansion(mult = c(0.1, 0.1))) +
  ylab("Share of Box Office Oversea") +
  ggtitle("Share of Box Office Oversea throughout the decades") & pprint
```

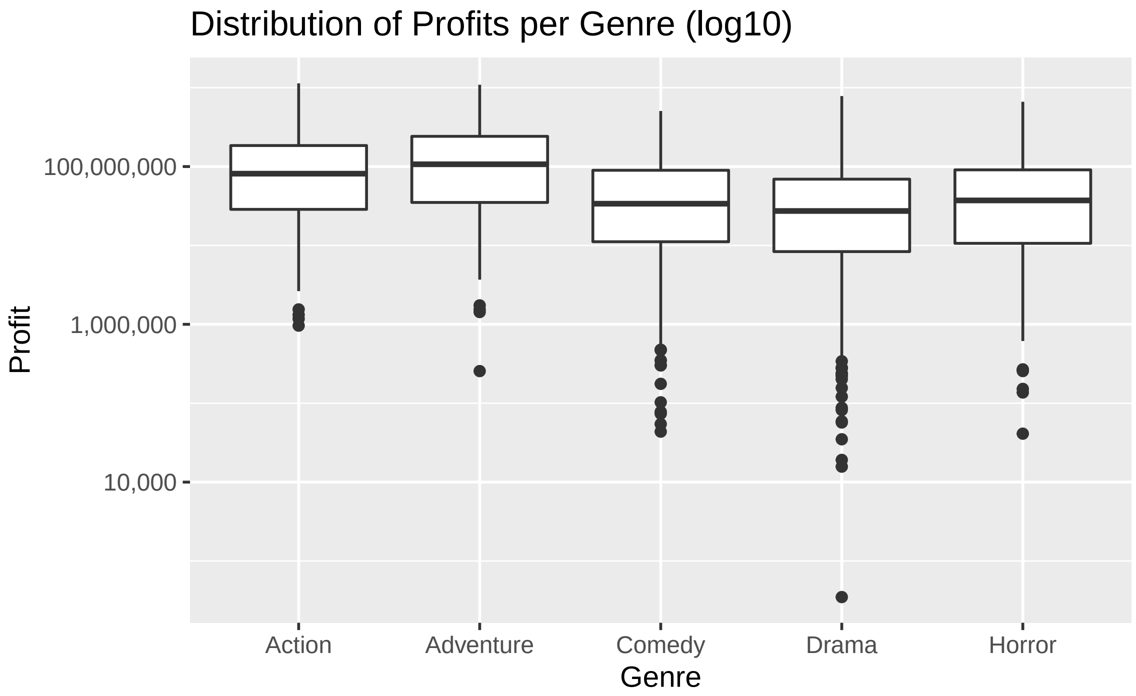


As a measure to avoid outliers and irrelevant observations, we will only include movies dated after 1980.

```
movie_profit <- movie_profit %>%
  filter(year(release_date) >= 1980)
```

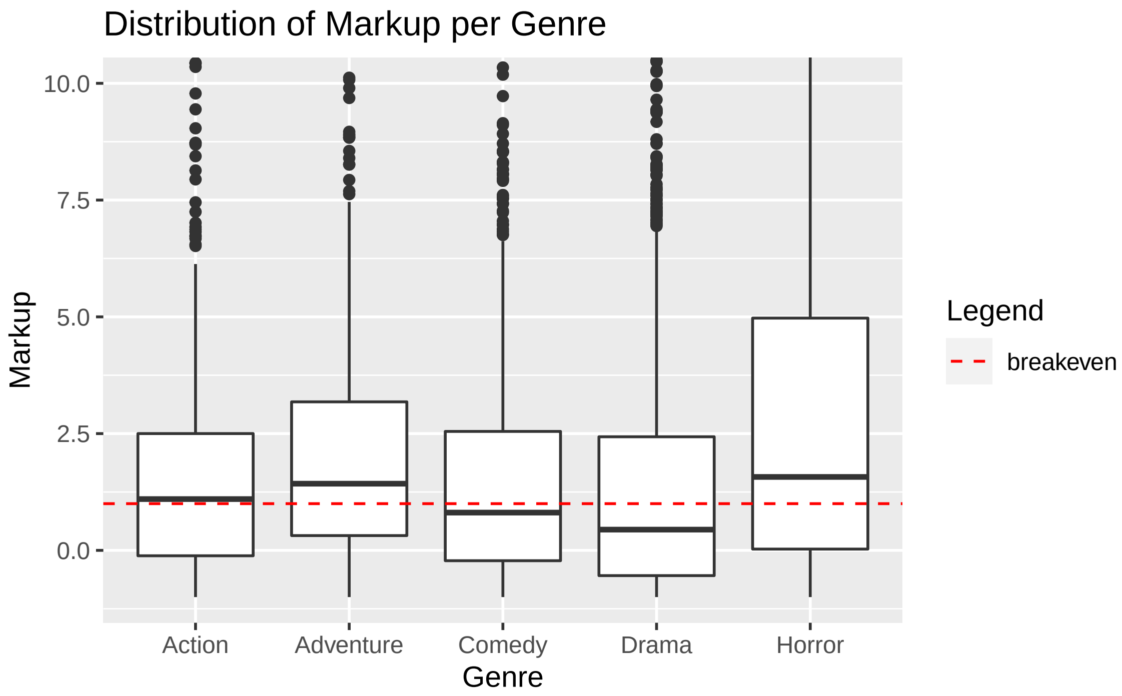
Overview

```
# movie_profit %>% View()

movie_profit %>%
  ggplot(aes(genre, profit)) +
  geom_boxplot() +
  scale_y_log10(label = comma) +
  ggtitle("Distribution of Profits per Genre (log10)") & pprint
```



When considering only profitable films, all categories amounts to similar amount of profits.

```
movie_profit %>%
  ggplot(aes(genre, markup)) +
  geom_boxplot() +
  coord_cartesian(ylim = c(-1,10)) +
  geom_hline(aes(yintercept = 1.0, linetype = "breakeven"), color = "red") +
  scale_linetype_manual(name = "Legend", values = "dashed",
                        guide = guide_legend(override.aes = list(color = c("red")))) +
  ggtitle("Distribution of Markup per Genre") & pprint
```

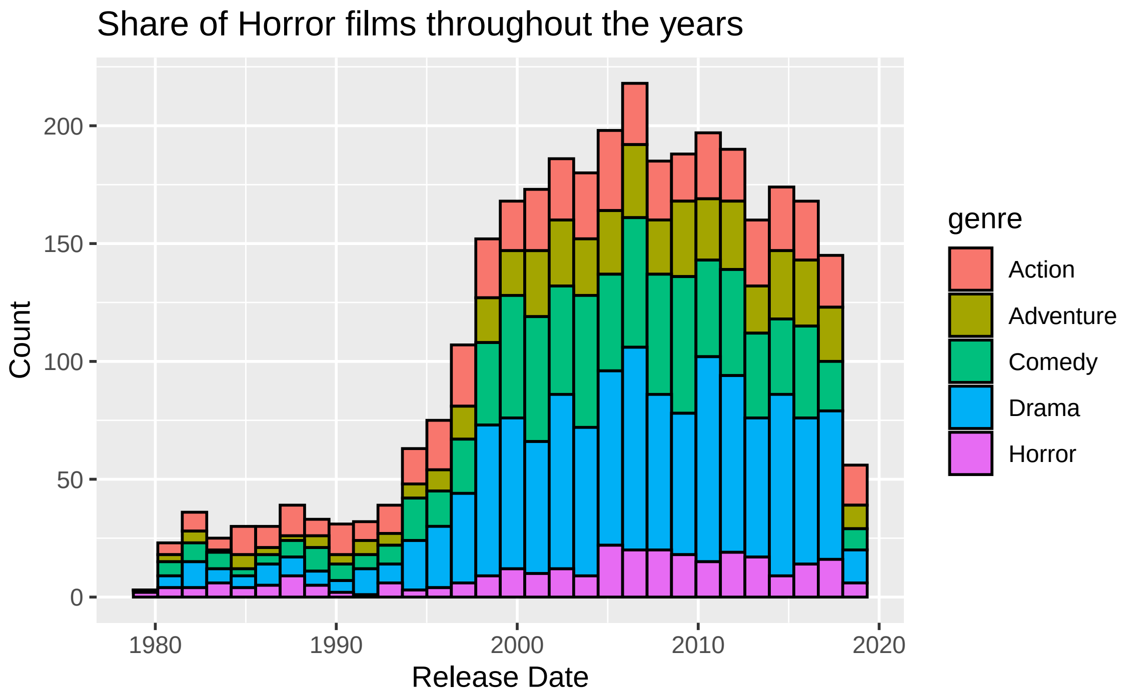


```
movie_profit %>%
  group_by(genre) %>%
  summarise(median(markup), mean(markup))
```

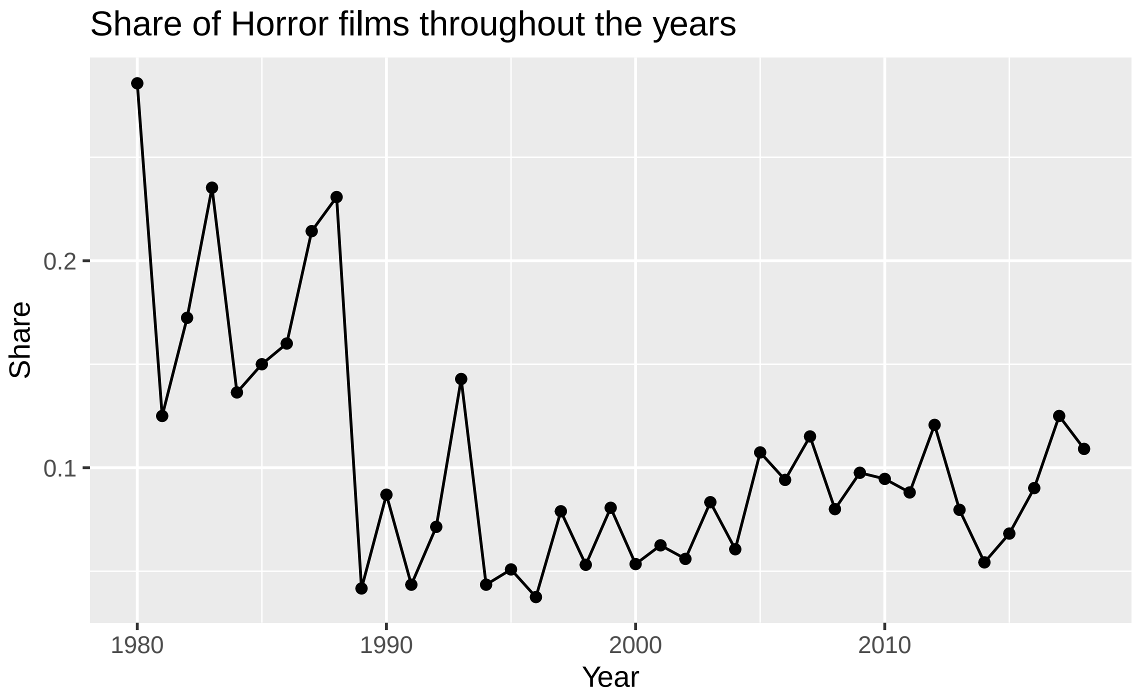
```
## # A tibble: 5 x 3
##   genre     `median(markup)` `mean(markup)`
##   <chr>                <dbl>          <dbl>
## 1 Action               1.10            1.75
## 2 Adventure            1.43            2.32
## 3 Comedy               0.809           2.33
## 4 Drama                0.445           2.06
## 5 Horror               1.57            8.99
```

As confirmed with visualization, the horror category is the most profitable, with a median markup of 1.62 (62% return on budget), and a whopping 10x average return. This is probably due to some outliers on the extreme.

```
movie_profit %>%
  ggplot(aes(release_date, fill = genre)) +
  geom_histogram(color = "black") +
  ggtitle("Share of Horror films throughout the years") & pprint
```



```
movie_profit %>%
  mutate(year=release_date %>% year) %>%
  group_by(year, genre) %>%
  summarise(count=n()) %>%
  mutate(share=count/sum(count)) %>%
  filter(genre=="Horror") %>%
  ggplot(aes(year, share)) +
  geom_point() +
  geom_line() +
  ggtitle("Share of Horror films throughout the years") & pprint
```

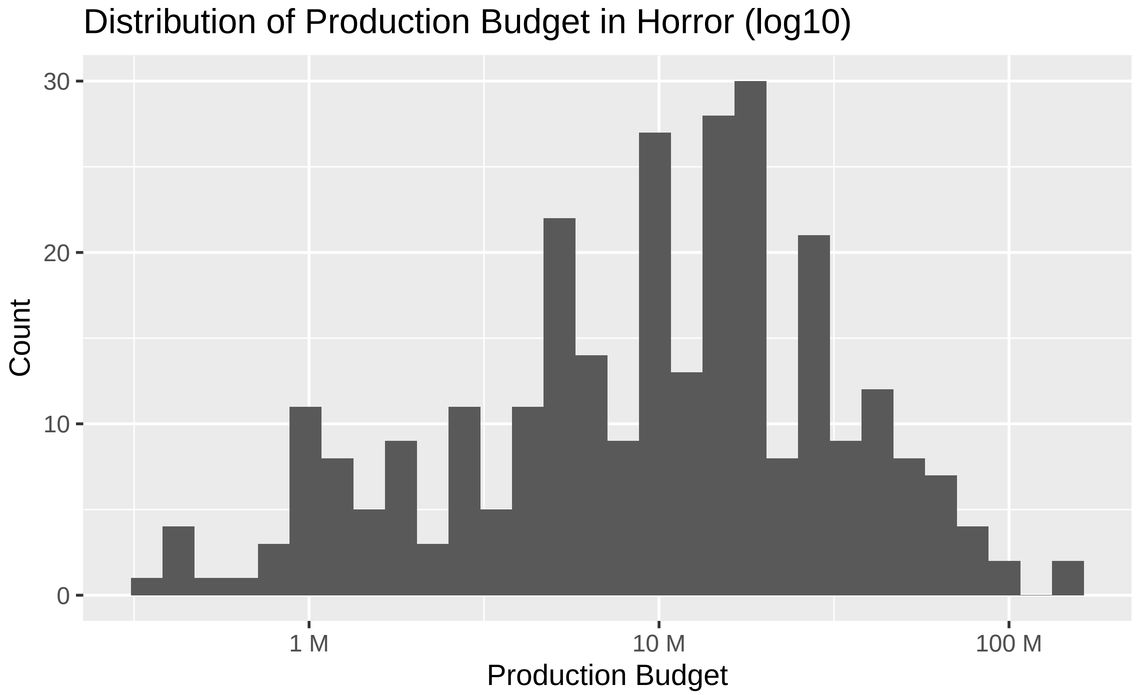


The horror category is the least popular genre in the dataset, so seemingly production companies are not capitalizing as much off horrors. We have fewer observations but nevertheless, we got ~300 data points.

```
horror_movies <- movie_profit %>%
  filter(genre == "Horror") %>%
  select(-genre)
```

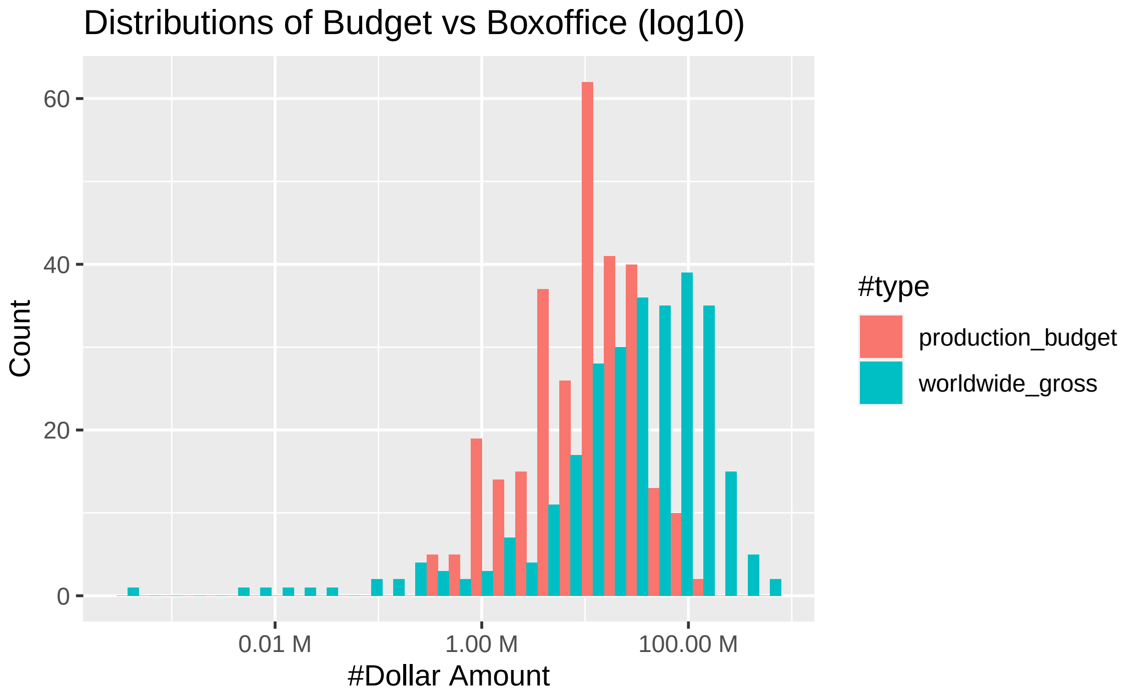
Analysis

```
dollar_format <- unit_format(unit="M", scale = 1e-6)

horror_movies %>%
  ggplot(aes(production_budget)) +
  geom_histogram(position = "dodge") +
  scale_x_log10(labels = dollar_format) +
  ggtitle("Distribution of Production Budget in Horror (log10)") & pprint
```

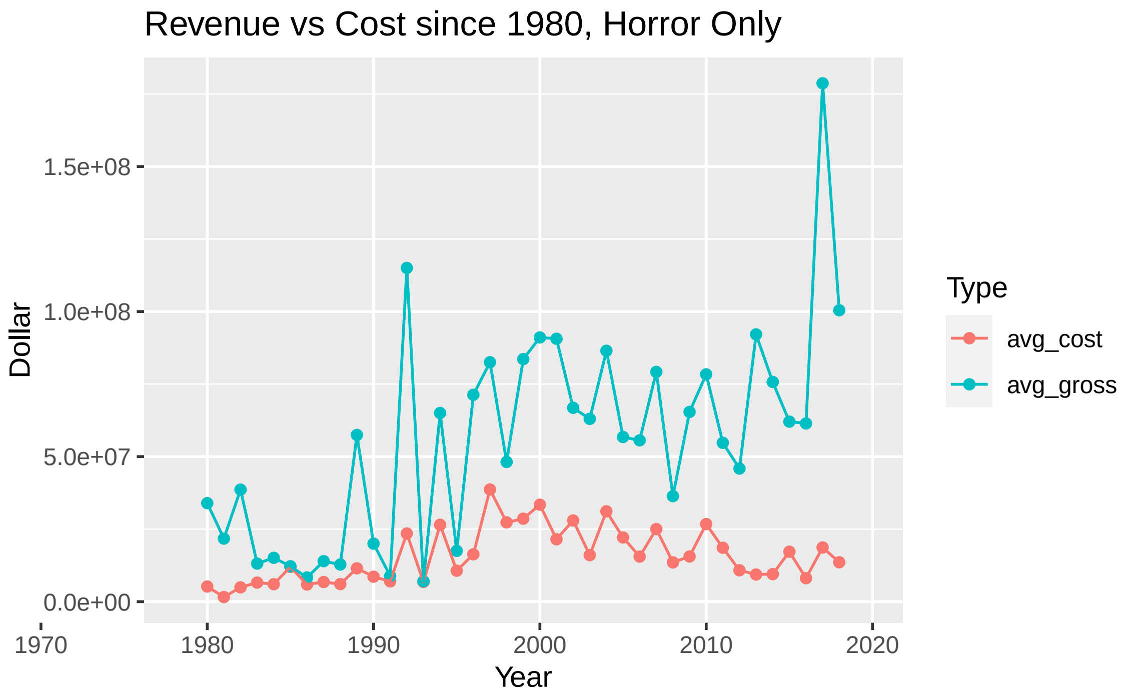


The production budget resembles the log-normal distribution to a certain degree.

```
horror_movies %>%
  gather(production_budget, worldwide_gross, key = "#type", value = "#dollar_amount") %>%
  ggplot(aes(`#dollar_amount`, fill = `#type`)) +
  geom_histogram(position = "dodge") +
  scale_x_log10(labels = dollar_format) +
  ggtitle("Distributions of Budget vs Boxoffice (log10)") & pprint
```

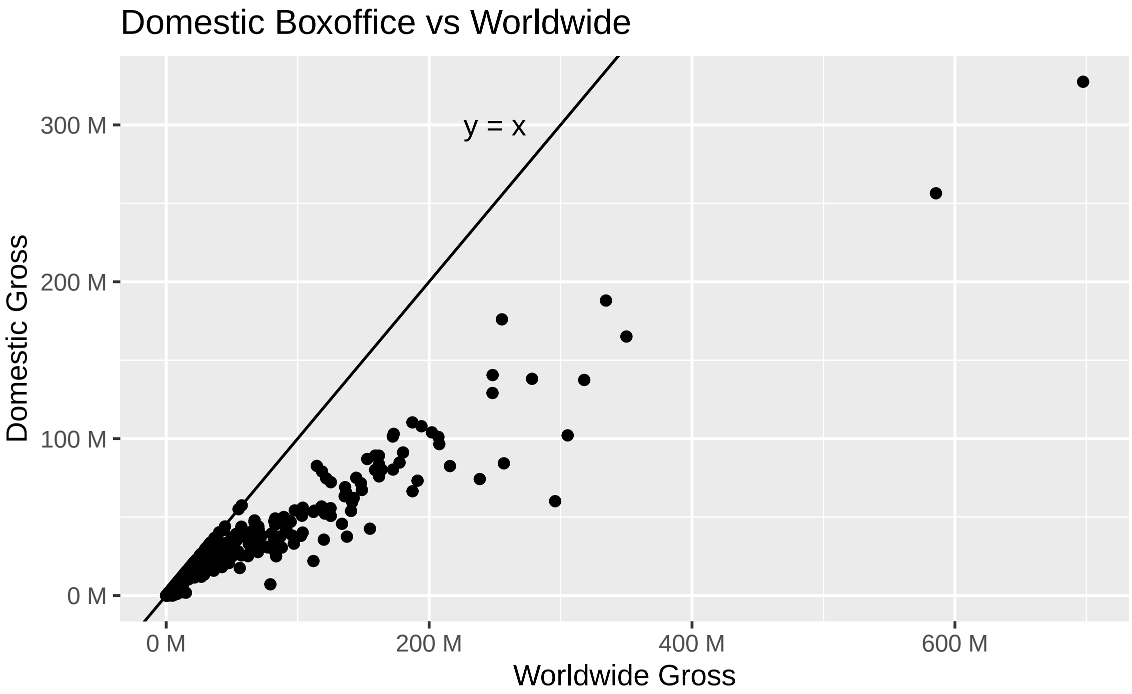


```
horror_movies %>%
  mutate(year=release_date %>% year) %>%
  group_by(year) %>%
  summarise(avg_cost=mean(production_budget), avg_gross=mean(worldwide_gross)) %>%
  gather(avg_cost, avg_gross, key = "type", value = "dollar") %>%
  ggplot(aes(year, dollar)) +
  geom_point(aes(color=type)) +
  geom_line(aes(color=type)) +
  scale_x_discrete(limits=seq(1930, 2020, 10), expand = expansion(mult = c(0.1, 0.1))) +
  ggtitle("Revenue vs Cost since 1980, Horror Only") & pprint
```



Comparing production budget and worldwide gross amount, we see that the revenues are usually higher than the costs.

```
horror_movies %>%
  ggplot(aes(worldwide_gross, domestic_gross)) +
  geom_point() +
  scale_x_continuous(labels = dollar_format) +
  scale_y_continuous(labels = dollar_format) +
  geom_abline(intercept = 0, slope = 1) +
  annotate(
    "text",
    x = 250e6,
    y = 300e6,
    label = "y = x"
  ) + ggtitle("Domestic Boxoffice vs Worldwide") & pprint
```



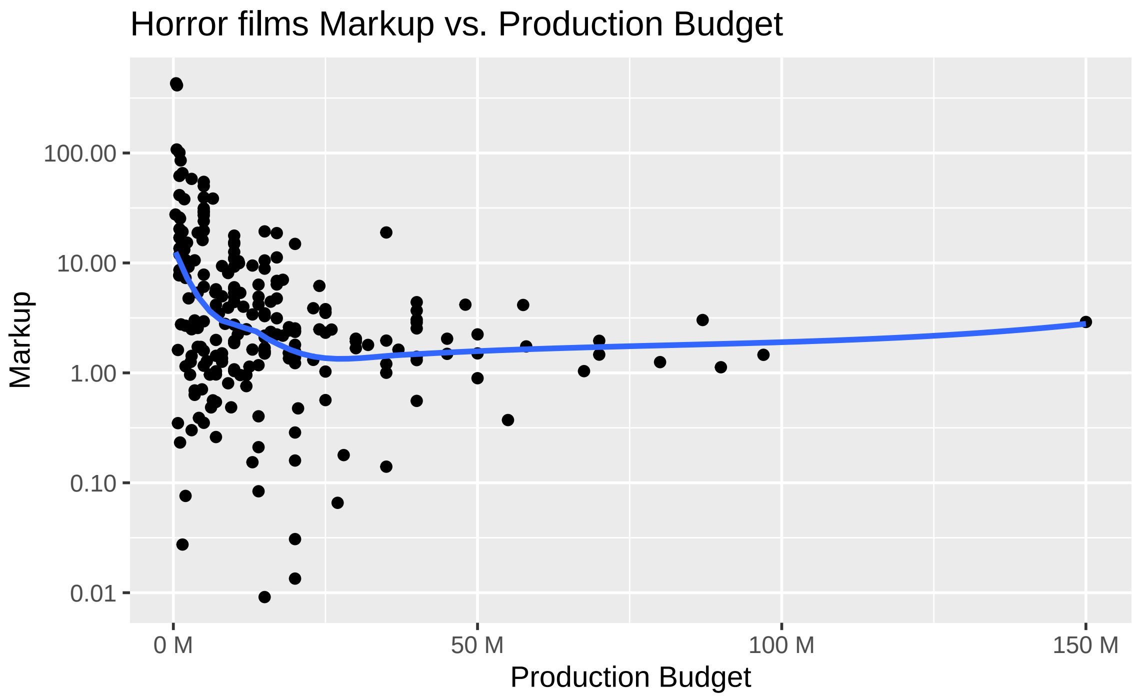
As a sanity check, worldwide gross is always greater than domestic gross.

Profitability and Size Effect

In all likelihood, the bigger the budget, the more stable the profitability should be.

Let's evaluate this hypothesis with visualization.

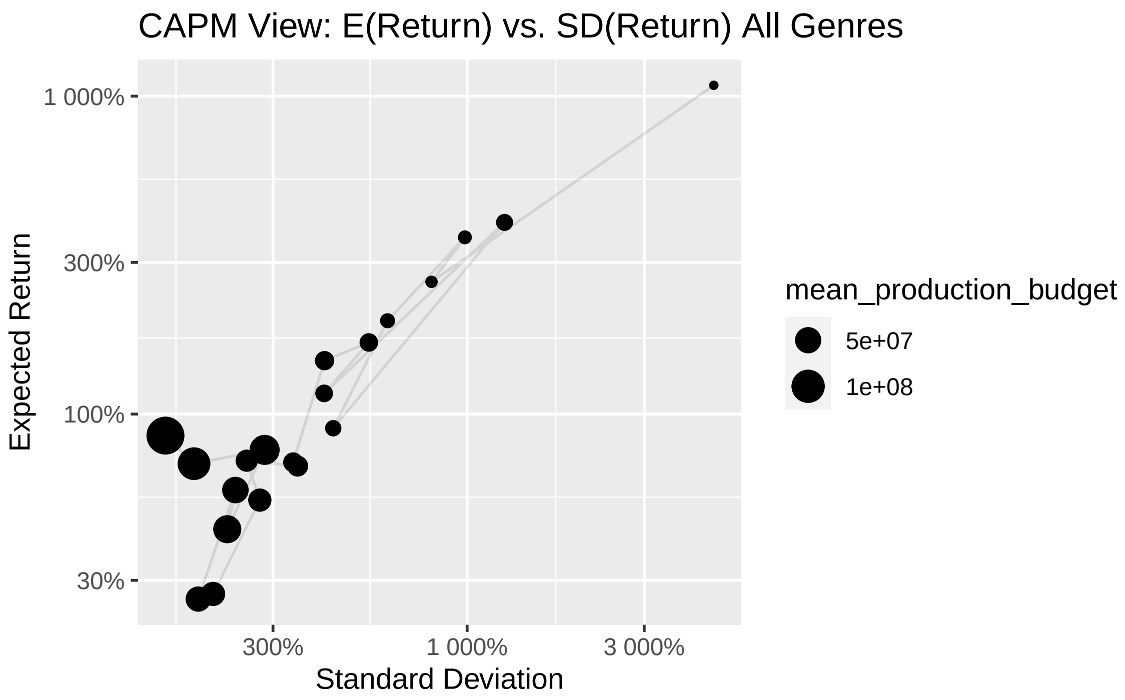
```
horror_movies %>%
  ggplot(aes(production_budget, markup)) +
  geom_point() +
  scale_x_continuous(labels = dollar_format) +
  scale_y_log10(labels = comma) +
  geom_smooth(se = FALSE) +
  ggtitle("Horror films Markup vs. Production Budget") & pprint
```



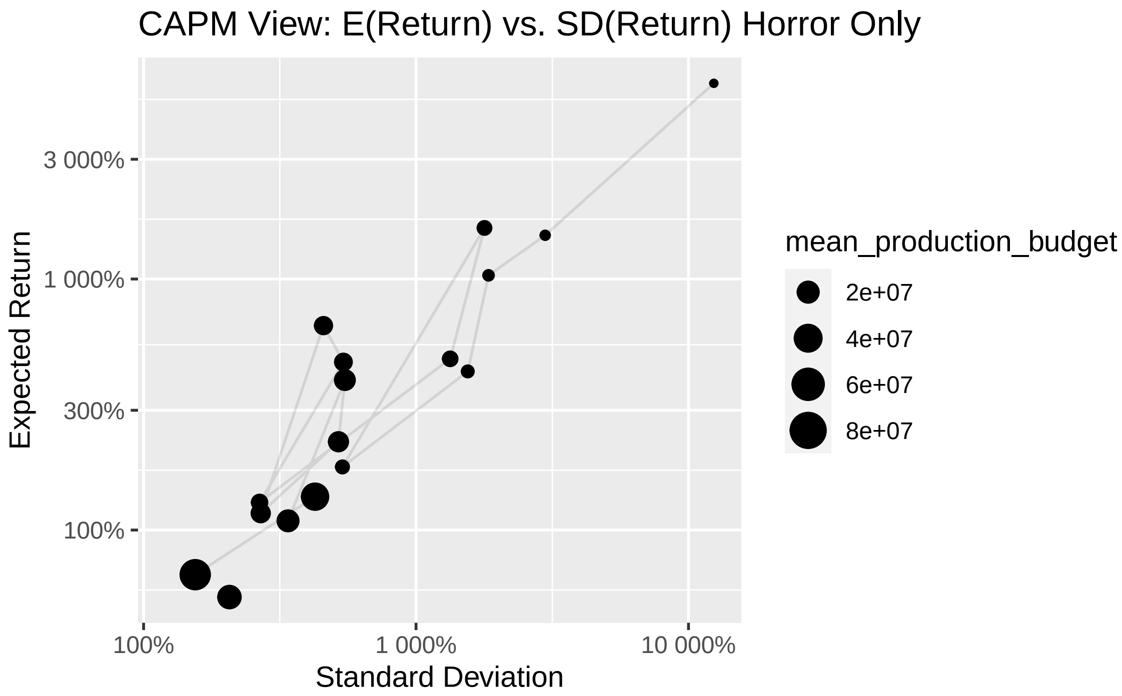
Small budget horror films have a high expected markup, but are also less stable in terms of profitability.

Seemingly, more production factors (labor, capital) aggregates to a more stable outcome under law of large number.

```
movie_profit %>%
  mutate(rank=percent_rank(production_budget),
         tier=rank%/%0.05*0.05) %>%
  group_by(tier) %>%
  summarise(expected_return=mean(markup)-1, standard_deviation=sd(markup), mean_production_budget=mean(
  ggplot(aes(standard_deviation, expected_return)) +
  geom_point(aes(size=mean_production_budget)) +
  scale_x_log10(labels = scales::percent) +
  scale_y_log10(labels = scales::percent) +
  geom_path(alpha=0.1) +
  ggtitle("CAPM View: E(Return) vs. SD(Return) All Genres") & pprint
```



```
horror_movies %>%
  mutate(rank=percent_rank(production_budget),
         tier=rank%/%0.05*0.05) %>%
  group_by(tier) %>%
  summarise(expected_return=mean(markup)-1, standard_deviation=sd(markup), mean_production_budget=mean(
  ggplot(aes(standard_deviation, expected_return)) +
  geom_point(aes(size=mean_production_budget)) +
  scale_x_log10(labels = scales::percent) +
  scale_y_log10(labels = scales::percent) +
  geom_path(alpha=0.1) +
  ggtitle("CAPM View: E(Return) vs. SD(Return) Horror Only") & pprint
```

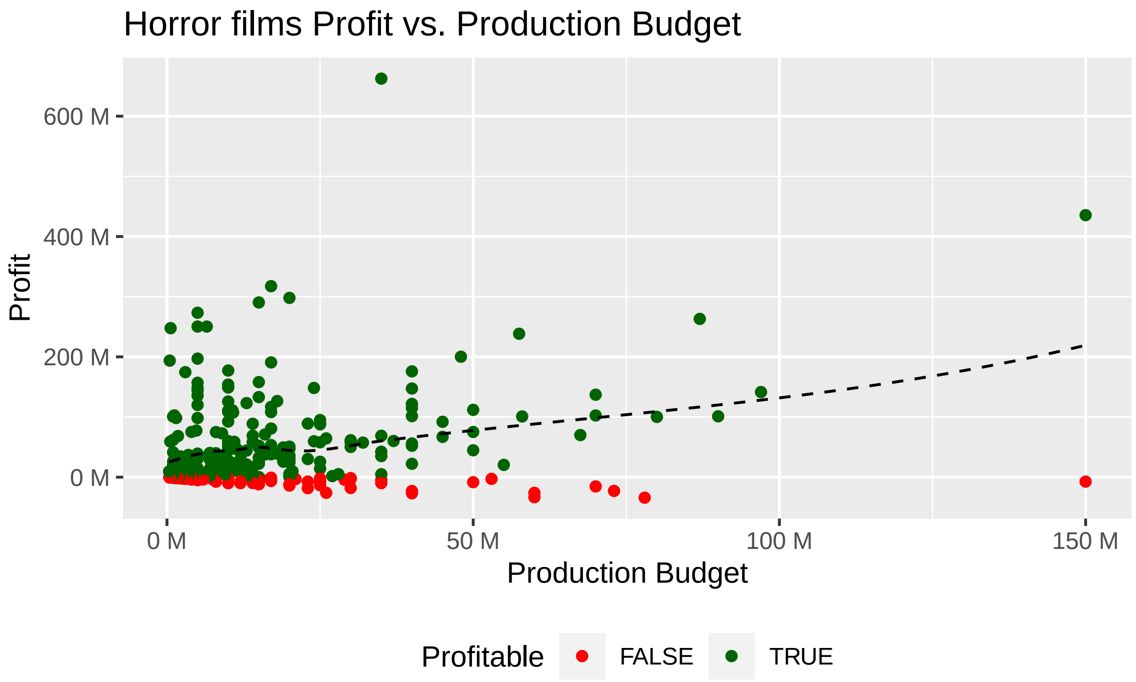


In another representation, we can use the CAPM dimensions to visualize the relationship between risk and return.

```
horror_movies %>%
  summarise(median_budget = median(production_budget),
            mean_markup_below_median = mean(markup[production_budget<=median(production_budget)]),
            mean_markup_above_median = mean(markup[production_budget>median(production_budget)])) %>%
  print.data.frame()
```

```
##   median_budget mean_markup_below_median mean_markup_above_median
## 1      10800000                 15.76432                  2.17084
```

```
horror_movies %>%
  mutate(profitable = profit > 0) %>%
  ggplot(aes(production_budget, profit)) +
  geom_point(aes(color = profitable)) +
  geom_smooth(se=FALSE, size=0.5, linetype="dashed", color="black") +
  scale_x_continuous(labels = dollar_format) +
  scale_y_continuous(labels = dollar_format) +
  theme(legend.position = "bottom") +
  scale_colour_manual(
    values = c("FALSE" = "red", "TRUE" = "darkgreen")
  ) + ggtitle("Horror films Profit vs. Production Budget") & pprint
```



```
sel <- horror_movies %>%
  mutate(rank_budget = percent_rank(production_budget)) %>%
  filter(profit == max(profit)) %>%
  select(movie, profit, production_budget, rank_budget)

# raw
sel %>% print()
```

```
## # A tibble: 1 x 4
##   movie     profit production_budget rank_budget
##   <chr>      <dbl>             <dbl>       <dbl>
## 1 It     662459228          35000000       0.854
```

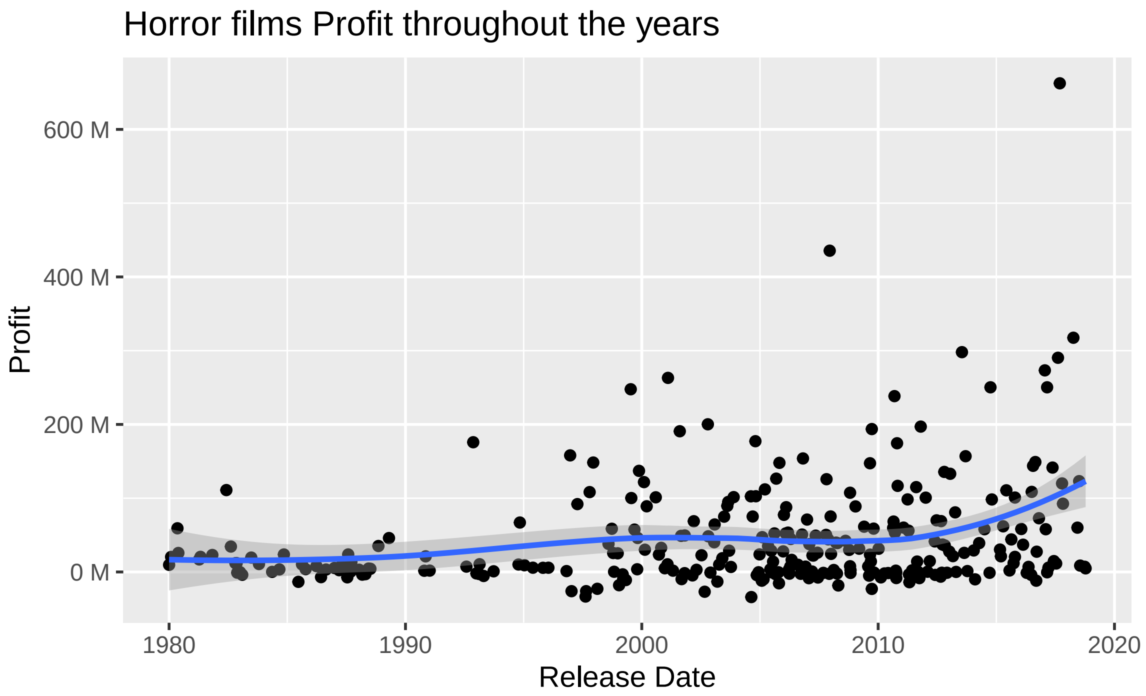
```
# adding big marks
# sel %>% formatC(format="d", big.mark=",") %>% print()
```

Higher production budget is associated with higher profit, but small-to-medium budget films can also gross very high.

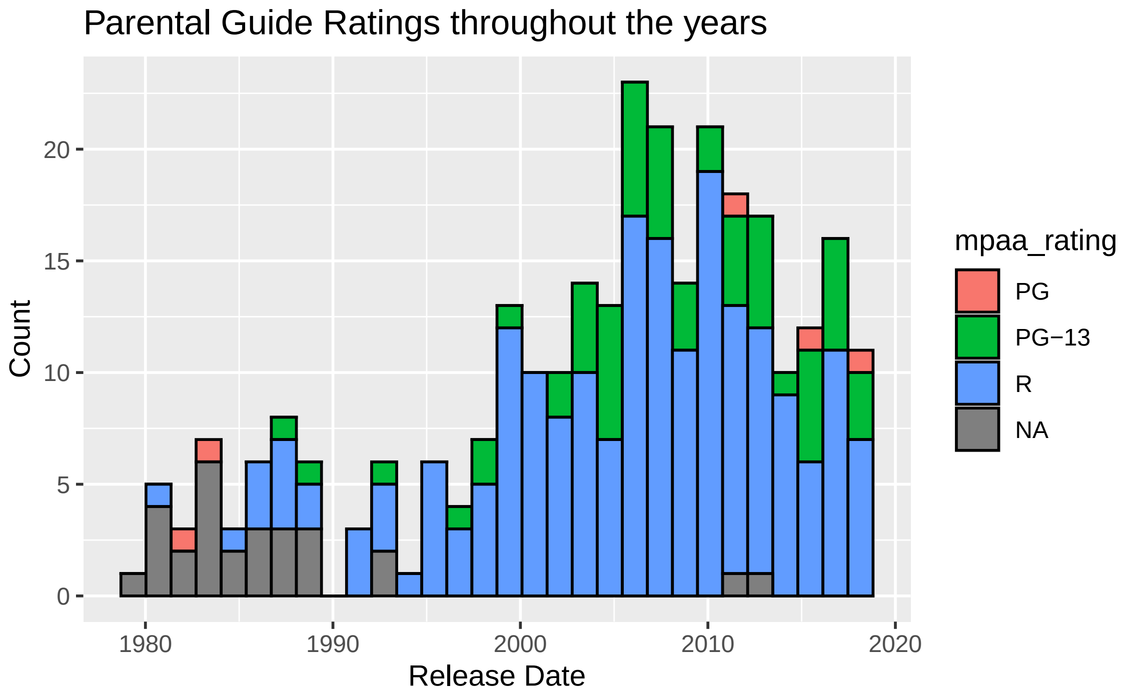
Horror movies seldom incur huge loss, even the worse performing films can earn back most of the budget. This supports the notion that horror productions are a great investment.

The highest grossing movie was “It” at 662 million dollars, but only spent 35 million in budget.

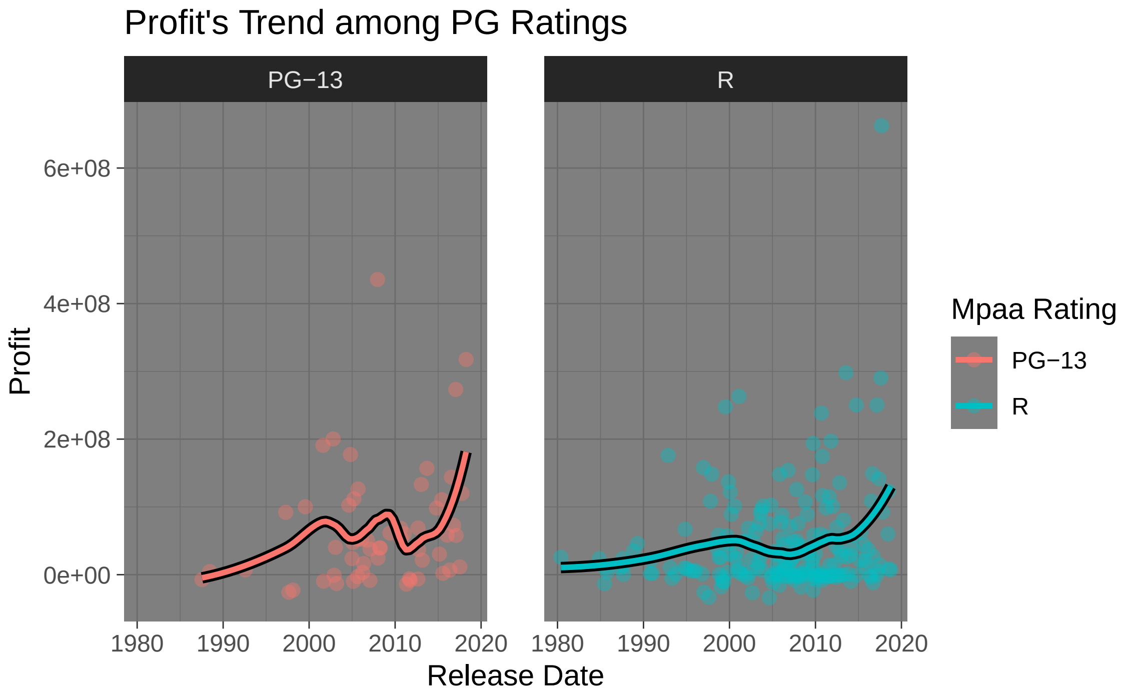
```
horror_movies %>%
  ggplot(aes(release_date, profit)) +
  geom_point() +
  geom_smooth() +
  scale_y_continuous(labels = dollar_format) +
  ggtitle("Horror films Profit throughout the years") & pprint
```



```
horror_movies %>%
  ggplot(aes(release_date, fill = mpaa_rating)) +
  geom_histogram(color = "black") +
  ggtitle("Parental Guide Ratings throughout the years") & pprint
```



```
horror_movies %>%
  filter(mpaa_rating %in% c("PG-13", "R")) %>%
  ggplot(aes(release_date, profit, color=mpaa_rating, fill=mpaa_rating)) +
  guides(fill="none") +
  geom_point(size=2, alpha=0.4) +
  geom_smooth(se=FALSE, color="black", size = 2, span=0.5) +
  geom_smooth(se=FALSE, span=0.5) +
  theme_dark() +
  ggtitle("Profit's Trend among PG Ratings") +
  facet_wrap(~mpaa_rating) +
  theme(panel.spacing = unit(1.5, "lines")) & pprint
```



As can be seen clearly, the trend of producing more R-rated horror movies started around mid 1990s.

We are seeing more and more outliers in terms of profit, and a very mild trend of improving profits.

We don't see a clear difference in profit between `R` and `PG-13` films, except occasional outperformance of `PG-13`s.