Horror Movies

2023-08-24

Installing the required packaged.

install.packages("tidyverse")  
install.packages("ggplot2")  
install.packages("datapasta")  
install.packages("ggpubr")

## R Markdown

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)  
library(datapasta)  
library(ggpubr)

Importing the datafile from local storage. Data is downloaded from TidyTuesday github repository.

Horror\_movies <- read.csv("horror\_movies.csv")

Get a overview of the structure of our data

str(Horror\_movies)

## 'data.frame': 32540 obs. of 20 variables:  
## $ id : int 760161 760741 882598 756999 772450 1014226 717728 762504 927341 760104 ...  
## $ original\_title : chr "Orphan: First Kill" "Beast" "Smile" "The Black Phone" ...  
## $ title : chr "Orphan: First Kill" "Beast" "Smile" "The Black Phone" ...  
## $ original\_language: chr "en" "en" "en" "en" ...  
## $ overview : chr "After escaping from an Estonian psychiatric facility, Leena Klammer travels to America by impersonating Esther,"| \_\_truncated\_\_ "A recently widowed man and his two teenage daughters travel to a game reserve in South Africa. However, their j"| \_\_truncated\_\_ "After witnessing a bizarre, traumatic incident involving a patient, Dr. Rose Cotter starts experiencing frighte"| \_\_truncated\_\_ "Finney Blake, a shy but clever 13-year-old boy, is abducted by a sadistic killer and trapped in a soundproof ba"| \_\_truncated\_\_ ...  
## $ tagline : chr "There's always been something wrong with Esther." "Fight for family." "Once you see it, it’s too late." "Never talk to strangers." ...  
## $ release\_date : chr "2022-07-27" "2022-08-11" "2022-09-23" "2022-06-22" ...  
## $ poster\_path : chr "/pHkKbIRoCe7zIFvqan9LFSaQAde.jpg" "/xIGr7UHsKf0URWmyyd5qFMAq4d8.jpg" "/hiaeZKzwsk4y4atFhmncO5KRxeT.jpg" "/lr11mCT85T1JanlgjMuhs9nMht4.jpg" ...  
## $ popularity : num 5089 2172 1864 1071 1021 ...  
## $ vote\_count : int 902 584 114 2736 83 1 125 1684 73 1035 ...  
## $ vote\_average : num 6.9 7.1 6.8 7.9 7 1 5.8 7 6.5 6.8 ...  
## $ budget : int 0 0 17000000 18800000 0 0 20000000 68000000 0 10000000 ...  
## $ revenue : int 9572765 56000000 45000000 161000000 0 0 2892594 170800000 0 14257609 ...  
## $ runtime : int 99 93 115 103 0 0 88 130 90 106 ...  
## $ status : chr "Released" "Released" "Released" "Released" ...  
## $ adult : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ backdrop\_path : chr "/5GA3vV1aWWHTSDO5eno8V5zDo8r.jpg" "/2k9tBql5GYH328Krj66tDT9LtFZ.jpg" "/mVNPfpydornVe4H4UCIk7WevWjf.jpg" "/AfvIjhDu9p64jKcmohS4hsPG95Q.jpg" ...  
## $ genre\_names : chr "Horror, Thriller" "Adventure, Drama, Horror" "Horror, Mystery, Thriller" "Horror, Thriller" ...  
## $ collection : int 760193 NA NA NA NA NA 94899 NA NA 950289 ...  
## $ collection\_name : chr "Orphan Collection" NA NA NA ...

We select the variables which is of interest.

sd\_horror\_movies <- Horror\_movies %>% select("id","original\_title","original\_language","popularity","vote\_average","budget","revenue")

We look at our subset data, and we can see that the languages are annotated with abbreviations.

str(sd\_horror\_movies)

## 'data.frame': 32540 obs. of 7 variables:  
## $ id : int 760161 760741 882598 756999 772450 1014226 717728 762504 927341 760104 ...  
## $ original\_title : chr "Orphan: First Kill" "Beast" "Smile" "The Black Phone" ...  
## $ original\_language: chr "en" "en" "en" "en" ...  
## $ popularity : num 5089 2172 1864 1071 1021 ...  
## $ vote\_average : num 6.9 7.1 6.8 7.9 7 1 5.8 7 6.5 6.8 ...  
## $ budget : int 0 0 17000000 18800000 0 0 20000000 68000000 0 10000000 ...  
## $ revenue : int 9572765 56000000 45000000 161000000 0 0 2892594 170800000 0 14257609 ...

To get the actual names of the language from the abbreviations we copy a table from this website: <https://www.science.co.il/language/Codes.php> and saved it into a new R-script. We converted it into a table.

lang <- data.table::fread("lang", data.table = FALSE, fill = TRUE)

We view our language data:

head(lang)

## Afar aa aar afar  
## 1 Abkhazian ab abk abkhaze  
## 2 Avestan ae ave avestique  
## 3 Afrikaans af afr afrikaans  
## 4 Akan ak aka akan  
## 5 Amharic am amh amharique  
## 6 Aragonese an arg aragonais

Then we join our horror movie data with our language data, but only adding the Afar (language) column:

joined\_sd\_horror\_movies <- left\_join(sd\_horror\_movies, lang, by=c("original\_language"="aa")) %>% select(-"aar", -"afar")

Relocating Afar next original language column and renaming Afar to language:

joined\_sd\_horror\_movies <- joined\_sd\_horror\_movies %>% relocate(Afar,.after="original\_language") %>% rename(Language=Afar)

We arrange the data to find the top three language of the movies:

joined\_sd\_horror\_movies %>% group\_by(Language) %>% summarise(Number=n()) %>% arrange(desc(Number))

## # A tibble: 95 × 2  
## Language Number  
## <chr> <int>  
## 1 English 21923  
## 2 Spanish; Castilian 1661  
## 3 Japanese 1639  
## 4 Portuguese 676  
## 5 German 631  
## 6 Indonesian 604  
## 7 French 591  
## 8 Italian 575  
## 9 <NA> 523  
## 10 Chinese 496  
## # ℹ 85 more rows

Making a new dataset with only movies with a budget above 500.000(unit?), vote average above 1 and only of the top three language of movies:

sd\_horror\_movies\_w500000 <- joined\_sd\_horror\_movies %>% filter(budget >= 500000, vote\_average > 1, Language=="English"|Language=="Spanish; Castilian"|Language=="Japanese")

We examine whether a larger budget gives the movies a larger revenue by plotting scatter plot of budget vs. revenue for the three top language of movies(English, Japanese, Spanish) and linear regression:

av\_vote\_revenue\_d <- ggplot(sd\_horror\_movies\_w500000, mapping=aes(x=budget, y=revenue, color=Language)) + geom\_point() + scale\_x\_log10() + facet\_grid(cols=vars(Language)) + geom\_smooth(method="lm", formula=y~log(x), color="black", level=FALSE) + xlab("Budget (Log scale)") + ylab("Revenue") + ggtitle("Budget vs. Revenue") + theme\_bw()   
av\_vote\_revenue\_d

A graph of different colored dots

Description automatically generated with medium confidence

We look into the linear regression for each the top three movie language estimating the R-squared and p-value to test the model fit.

lr1 <- sd\_horror\_movies\_w500000 %>% group\_by(Language) %>% do(model=lm(formula=revenue~log(budget), .))  
lr1

## # A tibble: 3 × 2  
## # Rowwise:   
## Language model   
## <chr> <list>  
## 1 English <lm>   
## 2 Japanese <lm>   
## 3 Spanish; Castilian <lm>

summary(lr1[[2]][[1]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -309387952 14753716.8 -20.97017 2.541957e-86  
## log(budget) 21949135 964560.4 22.75558 2.356515e-99

summary(lr1[[2]][[1]])$r.squared

## [1] 0.2471895

summary(lr1[[2]][[2]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -121550253 34109251 -3.563557 0.0010800986  
## log(budget) 8800481 2282173 3.856185 0.0004722738

summary(lr1[[2]][[2]])$r.squared

## [1] 0.2981775

summary(lr1[[2]][[3]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -63932490 46824830 -1.365354 0.1822932  
## log(budget) 4785220 3189193 1.500449 0.1439503

summary(lr1[[2]][[3]])$r.squared

## [1] 0.06980626

Putting the R-squared and p-value for each group into a dataframe:

dat\_text <- data.frame(  
 label = c("R = 0.247, p=0.001", "R = 0.298, p=0.001", "R = 0.070, p=0.144"),  
 Language = c("English", "Japanese", "Spanish; Castilian")  
)

Adding the R-squared and p-value for each group into the figure:

av\_vote\_revenue\_d\_wn <- ggplot(sd\_horror\_movies\_w500000, mapping=aes(x=budget, y=revenue, color=Language)) + geom\_point() + scale\_x\_log10() + facet\_grid(cols=vars(Language)) + geom\_smooth(method="lm", formula=y~log(x), color="black", level=FALSE) + xlab("Budget (Log scale)") + ylab("Revenue") + ggtitle("Budget vs. Revenue") + theme\_bw() + geom\_text(  
 data = dat\_text,  
 mapping = aes(x = 1e+05, y = 6.3e+08, label = label),  
 hjust = -0.1,  
 vjust = -1  
)  
av\_vote\_revenue\_d\_wn

A graph of numbers and a budget

Description automatically generated with medium confidence

We have budget of the movies on X axis and its revenue on the Y axis. We looked into the top three language movies. Budget of movies in all three language found to have a very low positive correlation with its revenue. However, only English and Japanese movies showed a significant correlation.

We examine whether a larger budget gives the movies a higher vote average by plotting scatter plot of budget vs. vote average for the three top language of movies(English, Japanese, Spanish) and linear regression:

av\_vote\_budget\_d <- ggplot(sd\_horror\_movies\_w500000, mapping=aes(x=budget, y=vote\_average, color=Language)) + geom\_point() + scale\_x\_log10() + facet\_grid(cols=vars(Language)) + geom\_smooth(method="lm", color="black", level=FALSE) + xlab("Budget (Log scale)") + ylab("Vote average") + ggtitle("Budget vs. vote average") + theme\_bw()   
  
av\_vote\_budget\_d

## `geom\_smooth()` using formula = 'y ~ x'

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

We look into the linear regression for each the top three movie language estimating the R-squared and p-value to test the model fit.

lr2 <- sd\_horror\_movies\_w500000 %>% group\_by(Language) %>% do(model=lm(formula=vote\_average~log(budget), .))  
lr2

## # A tibble: 3 × 2  
## # Rowwise:   
## Language model   
## <chr> <list>  
## 1 English <lm>   
## 2 Japanese <lm>   
## 3 Spanish; Castilian <lm>

summary(lr2[[2]][[1]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.6812006 0.31564577 2.158117 3.106924e-02  
## log(budget) 0.3120145 0.02063612 15.119828 2.467747e-48

summary(lr2[[2]][[1]])$r.squared

## [1] 0.1266106

summary(lr2[[2]][[2]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.6849932 1.8952101 1.416726 0.1654029  
## log(budget) 0.2387745 0.1268042 1.883017 0.0680305

summary(lr2[[2]][[2]])$r.squared

## [1] 0.09198813

summary(lr2[[2]][[3]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.4234114 5.1962819 0.08148352 0.9355986  
## log(budget) 0.3659685 0.3539137 1.03406153 0.3093716

summary(lr2[[2]][[3]])$r.squared

## [1] 0.03441609

Putting the R-squared and p-value for each group into a dataframe:

dat\_text2 <- data.frame(  
 label = c("R = 0.127, p=0.001", "R = 0.092, p=0.068", "R = 0.034, p=0.309"),  
 Language = c("English", "Japanese", "Spanish; Castilian")  
)

Adding the R-squared and p-value for each group into the figure:

av\_vote\_budget\_d\_wn <- ggplot(sd\_horror\_movies\_w500000, mapping=aes(x=budget, y=vote\_average, color=Language)) + geom\_point() + scale\_x\_log10() + facet\_grid(cols=vars(Language)) + geom\_smooth(method="lm", color="black", level=FALSE) + xlab("Budget (Log scale)") + ylab("Vote average") + ggtitle("Budget vs. vote average") + theme\_bw() + geom\_text(  
 data = dat\_text2,  
 mapping = aes(x = 1e+05, y = 9, label = label),  
 hjust = -0.1,  
 vjust = -1  
)  
  
av\_vote\_budget\_d\_wn

## `geom\_smooth()` using formula = 'y ~ x'

A screen shot of a graph

Description automatically generated

We have budget of the movies on X axis and its vote average on the Y axis. We looked into the top three language movies. Budget of movies in all three language found to have a very low positive correlation with its vote average. However, only English movies showed a significant correlation.

We examine whether larger budget movies has higher popularity by plotting scatter plot of budget vs. popularity for the three top language of movies(English, Japanese, Spanish) and linear regression:

popu\_budget\_d <- ggplot(sd\_horror\_movies\_w500000, mapping=aes(x=budget, y=popularity, color=Language)) + geom\_point() + scale\_x\_log10() + facet\_grid(cols=vars(Language)) + geom\_smooth(method="lm", color="black", level=FALSE) + xlab("Budget (Log scale)") + ylab("Popularity") + ggtitle("Budget vs. popularity") + theme\_bw()  
popu\_budget\_d

## `geom\_smooth()` using formula = 'y ~ x'

A graph of different colored dots

Description automatically generated with medium confidence

We look into the linear regression for each the top three movie language estimating the R-squared and p-value to test the model fit.

lr3 <- sd\_horror\_movies\_w500000 %>% group\_by(Language) %>% do(model=lm(formula=popularity~log(budget), .))  
lr3

## # A tibble: 3 × 2  
## # Rowwise:   
## Language model   
## <chr> <list>  
## 1 English <lm>   
## 2 Japanese <lm>   
## 3 Spanish; Castilian <lm>

summary(lr3[[2]][[1]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -149.90024 19.524485 -7.677552 2.826570e-14  
## log(budget) 11.37172 1.276461 8.908788 1.389469e-18

summary(lr3[[2]][[1]])$r.squared

## [1] 0.04791603

summary(lr3[[2]][[2]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -5.059776 21.792346 -0.2321813 0.8177494  
## log(budget) 1.092239 1.458077 0.7490957 0.4588048

summary(lr3[[2]][[2]])$r.squared

## [1] 0.0157797

summary(lr3[[2]][[3]])$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -106.637828 58.644851 -1.818366 0.07900574  
## log(budget) 8.529535 3.994243 2.135457 0.04100470

summary(lr3[[2]][[3]])$r.squared

## [1] 0.1319489

Putting the R-squared and p-value for each group into a dataframe:

dat\_text3 <- data.frame(  
 label = c("R = 0.048, p=0.001", "R = 0.016, p=0.459", "R = 0.132, p=0.041"),  
 Language = c("English", "Japanese", "Spanish; Castilian")  
)

Adding the R-squared and p-value for each group into the figure:

popu\_budget\_d\_wn <- ggplot(sd\_horror\_movies\_w500000, mapping=aes(x=budget, y=popularity, color=Language)) + geom\_point() + scale\_x\_log10() + facet\_grid(cols=vars(Language)) + geom\_smooth(method="lm", color="black", level=FALSE) + xlab("Budget (Log scale)") + ylab("Popularity") + ggtitle("Budget vs. popularity") + theme\_bw() + geom\_text(  
 data = dat\_text3,  
 mapping = aes(x = 1e+05, y = 1550, label = label),  
 hjust = -0.1,  
 vjust = -1  
)  
  
popu\_budget\_d\_wn

## `geom\_smooth()` using formula = 'y ~ x'

A graph with different colored dots

Description automatically generated with medium confidence

We have budget of the movies on X axis and its popularity on the Y axis. We looked into the top three language movies. Budget of movies in all three language were found to have a very low positive correlation with its popularity. However, only English movies showed a significant correlation.