R Notebook

Import Necessary packages

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.3

## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.2.3
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(caTools)
library(pander) #Table format
```

```
## Warning: package 'pander' was built under R version 4.2.3
```

The dataset can be located from this github repo

```
# Import data
Merged_data <- read_csv('Cleaned_data.csv')
```

```
## New names:
## Rows: 1337 Columns: 15
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (5): primary_title, genres, original_language, original_title, studio dbl (9):
## ...1, runtime_minutes, averagerating, popularity, vote_average, pr... date (1):
## release_date
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
head(Merged_data)
```

```
## # A tibble: 6 x 15
##    ...1 primary_title     runtime_minutes genres averagerating original_language
##   <dbl> <chr>                       <dbl> <chr>          <dbl> <chr>
## 1     0 On the Road                   124 Adven~           6.1 en
## 2     1 On the Road                    89 Drama            6   en
## 3     2 On the Road                   121 Drama            5.7 en
## 4     3 The Secret Life ~             114 Adven~           7.3 en
## 5     4 A Walk Among the~             114 Actio~           6.5 en
## 6     5 Jurassic World                124 Actio~           7   en
## # i 9 more variables: original_title <chr>, popularity <dbl>,
## #   vote_average <dbl>, release_date <date>, production_budget <dbl>,
## #   worldwide_gross <dbl>, studio <chr>, domestic_gross <dbl>,
## #   foreign_gross <dbl>
```

Data Cleaning/ Preparation

```r
# Check for duplicates
sum(duplicated(Merged_data))
```

```
## [1] 0
```

```r
# Check missing values
pander::pander(colSums(is.na(Merged_data)))
```

Table 1: Table continues below

...1	primary_title	runtime_minutes	genres	averagerating
0	0	0	0	0

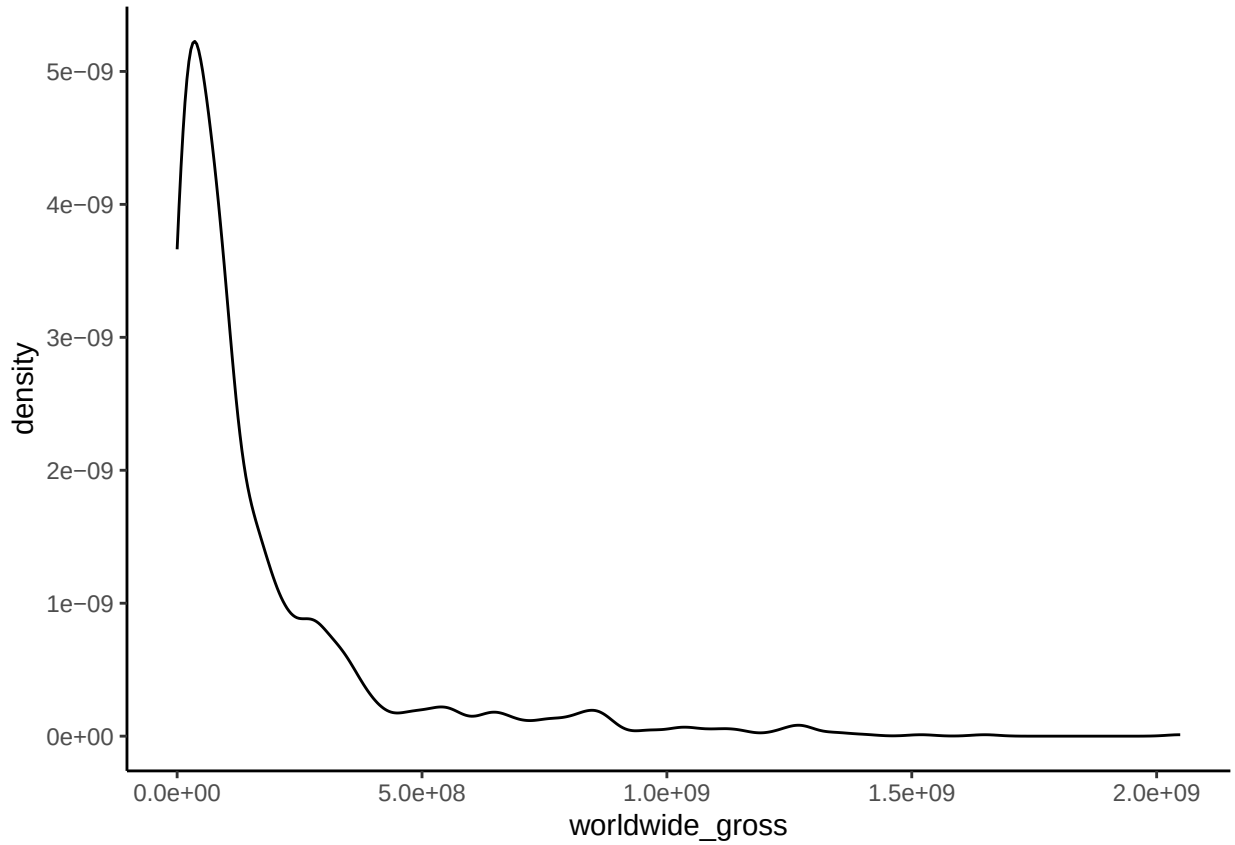
Table 2: Table continues below

original_language	original_title	popularity	vote_average	release_date
0	0	0	0	0

production_budget	worldwide_gross	studio	domestic_gross	foreign_gross
0	0	0	0	0

For this regression model we are interested in understanding the relationship between the worldwide gross for movie production against various variables in our models. The Dependent variable is **Worldwide_gross**

```r
# Invetsigate the properties of dependent variable
ggplot(data = Merged_data,aes(x=worldwide_gross))+geom_density()+theme_classic()
```




```
# There is alot of skewness in the data (positive skewness)

# Test for normality in the data
ks.test(Merged_data$worldwide_gross,"pnorm")
```

```
## Warning in ks.test.default(Merged_data$worldwide_gross, "pnorm"): ties should
## not be present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Merged_data$worldwide_gross
## D = 0.99701, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# Normality test violated a glm model that uses the gamma distribution would be effective to try and mo
```

```
# remove rows where the gross returns is zero
Merged_data %>% filter(!worldwide_gross <= 0) -> Merged_data
```

GLM models

```r
#
Merged_data %>% select(runtime_minutes,averagerating,popularity,production_budget,domestic_gross,foreig

# Split the data for Train and test data
Y = Model_data$worldwide_gross
sample_sl =sample.split(Y,SplitRatio = 3/4)

# Create Train and test data
train_data =subset(Model_data,sample_sl==TRUE)

test_data =subset(Model_data,sample_sl==FALSE)

# Fit GLM gamma model
Fit_model=glm(worldwide_gross~.,family = Gamma(link = "log"),data = train_data)


# Summary of the model
pander::pander(summary.glm(Fit_model))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.57	0.1411	124.5	0
runtime_minutes	0.00032	0.001046	0.3058	0.7598
averagerating	-0.03366	0.01983	-1.697	0.08998
popularity	0.01278	0.002931	4.361	1.432e-05
production_budget	3.655e-09	5.496e-10	6.65	4.829e-11
domestic_gross	6.458e-09	3.893e-10	16.59	9.611e-55
foreign_gross	2.299e-09	2.489e-10	9.237	1.5e-19

(Dispersion parameter for Gamma family taken to be 0.3905132)

Null deviance:	1802.4 on 998 degrees of freedom
Residual deviance:	707.1 on 992 degrees of freedom

Interpreting the glm model

This generalized linear model (GLM) with a Gamma distribution and log link function is used to model the relationship between `worldwide_gross` and several predictors related to movie characteristics. Here's an interpretation of the results:

Residuals

The deviance residuals summary provides a quick look at the spread of residuals:

- **Min, 1Q, Median, 3Q, Max** values suggest that residuals are moderately centered around zero, indicating an acceptable fit. However, some residuals have relatively large negative and positive values, implying a few observations deviate notably from the model's predictions.

Coefficients

Each predictor in the model has an associated estimated coefficient, its standard error, and p-value:

- **(Intercept)**: The intercept is significant, with a large positive estimate (17.78), representing the baseline log of worldwide gross when all other predictors are at zero.
- **runtime_minutes**: This variable has a negative coefficient (-0.0007785), but it is not statistically significant ($p = 0.472$), meaning `runtime_minutes` does not appear to have a substantial impact on worldwide gross in this model.
- **averagerating**: This has a significant negative effect (-0.0446, $p = 0.027$), suggesting that, holding other factors constant, higher ratings are weakly associated with lower worldwide gross, though the effect is relatively small.
- **popularity**: This variable has a positive and statistically significant coefficient (0.009753, $p = 0.0005$), indicating a positive association with worldwide gross—higher popularity scores are associated with higher gross earnings.
- **production_budget**: A positive and significant effect (3.83e-09, $p < 0.001$), implying that larger production budgets lead to higher worldwide gross, though the impact is small per unit increase in budget.
- **domestic_gross** and **foreign_gross**: Both are highly significant ($p < 0.001$) and positively associated with `worldwide_gross`, with each having an estimated impact proportional to their gross values, underscoring their strong, direct contributions to worldwide gross.

Model Fit

- **Dispersion parameter**: The estimated dispersion parameter for the Gamma family is 0.399962, suggesting moderate variability around the fitted values.
- **Deviance**: The residual deviance (745.68) is much lower than the null deviance (1894.53), indicating that the model significantly reduces deviance (improves fit) compared to a null model without predictors.
- **AIC**: The Akaike Information Criterion (AIC) of 38770 can be used to compare this model with other potential models, where a lower AIC generally indicates a better fit.

Conclusion

This model suggests that `popularity`, `production_budget`, `domestic_gross`, and `foreign_gross` are the primary predictors of `worldwide_gross`, with statistically significant positive associations. Although `averagerating` shows a significant but small negative impact, the impact of `runtime_minutes` does not appear statistically relevant in this context.