QRM II Graded Assignment (5), Period 1 2025

Material by Sjoerd van Alten and Klervie Toczé

Assignment Group 8, Catherina Mikhail, Mayte Leegwater, Merel Vonk, Yasmina Andaou

03-10-2025

0 Introduction

This assignment is to be completed in groups of 3-4. Further, all students in your group need to be assigned to the same R tutorial group (Friday's tutorial). You can sign yourself up for a group on Canvas. Please do so **before the start of your first R tutorial on Friday September 5th.** You can use the Discussion Board in Canvas if you do not have a group yet or if your group is incomplete.

The assignment has 5 parts, and each part corresponds to the course material of that week (with the exclusion of week 6, for which there is no R programming material).

You are supposed to hand in these assignments on Canvas at the following dates:

- **Deadline 1** *Thursday September 25th, at 23:59pm*: you are supposed to hand in weeks 1, 2, and 3 of this assignment. This will determine 18% of your overall course grade
- **Deadline 2** *Thursday October 9th, at 23:59pm*: you are supposed to hand in weeks 4, and 5 of this assignment. This will determine 12% of your overall course grade

The R tutorials (each Friday) will consist of two halves. During the first half, you will discuss the tutorial exercises. These can be downloaded separately from Canvas. During the second half, you can work on this graded assignment within your own group. The purpose is that you find out how to work with R for doing statistical analyses by yourself. The tutorial exercises are meant to teach you basic commands to get you started, but to answer the problem sets in this assignment, you might need to research your own solutions, and use functions and commands not described in the tutorial exercises. Learning how to solve your own research problems is integral part of learning R. When you and your group get stuck on how to approach an exercise, the hierarchy in finding your way is as follows:

- use the concepts from the tutorial exercises;
- use the cheat sheets available on Canvas;
- use Google, YouTube, StackOverflow, or another website;
- ask the teacher.

The use of generative AI is **not** permitted and may result in a grade of 0. See the AI protocol in the course manual for details.

To answer the assignment, you can simply fill out this R markdown document. There are designated places which you can fill with R code. There are also designated spaces for you to answer each question. Often, the structure of an answer will be as follows. First, you type the R code in the designated box. This will show how you analyzed the data to get the answer to the question. Below the box for the R code, you will then summarize your answer to the question, i.e. what are the conclusions that you draw from the data analysis?

When handing in, you are supposed to submit this .Rmd file, and a knitted version of this document. You can knit this document to pdf, word, or html. Knitting to pdf requires you to have a .tex distribution installed on your computer. Knitting to Word requires you to have Word installed.

The exercises are designed such that you should be able to finish the majority of them during the tutorial each week. If you are not able to finish them fully during that time, you are expected to work on it in your own time using the computers on campus or your own device. It is best to meet as a group in-person when working together. If you want to work remotely, github is a good platform to guarantee smooth collaboration. Alternatively, you can email this .Rmd file back and forth to one another as a group, but this is not recommended as it is more cumbersome.

We encourage you to keep your code blocks, printing statements, and final answers, as short as possible. In any case, there is a page limit of 6 pages per week, which encompasses the total length of this document which consists of the questions, your coding lines, and your answers. When your answers to questions of the respective week exceed this page limit, they will not be graded, resulting in zero points.

Each week consists of 1, 2, or 3 subquestions. The total amount of points you can earn per week is 20 points.

1 Week 1

1. Find the dataset "movies2.tsv" on Canvas. Describe your data set: How many observations does it have. How many variables are there? How many subjects? What consists of a subject? [**4 points**]

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.5.1
```

```
movies2 = read_tsv("movies2.tsv")
```

```
## Rows: 606 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr   (8): keywords, original_language, title, genre, first_actor, first_act...
## dbl  (10): index, budget, popularity, revenue, runtime, vote_average, vote_c...
## 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.
```

```
#Here I am computing the number of observations, variables and subjects
ncol(movies2)
```

```
## [1] 19
```

```
nrow(movies2)
```

```
## [1] 606
```

Your Answer:

This dataset has 606 observations. There are 19 variables and 606 subjects. A subject in a dataset is the unit of analysis that each row represents, in this case for example each row is gives all the information about 1 movie with variables such as the revenue, runtime, average voting and the other 16 variables.

2. Which of the following types of variables are present in your data set? (i) nominal; (ii) ordinal; (iii); interval; (iv) ratio. If present, name one example of such a variable present in your data set. **[4 points]**

```
#Here I am looking at the head of the dataset to determine whether the variables are
↪  nominal, ordinal, interval or ratio
head(movies2)
```

```
## # A tibble: 6 x 19
##   index  budget keywords original_language title popularity release_date revenue
##   <dbl>   <dbl> <chr>    <chr>             <chr>      <dbl> <date>         <dbl>
## 1  3793   4  e6 <NA>     en                The ~      0.409 1999-07-16    0
## 2  3853   3.5e6 thrille~ en                2:13       1.27  2009-04-25    0
## 3  2476   0     poison ~ en                Whit~      9.42  2002-10-11    0
## 4   491   1.3e8 gladiat~ en                Pomp~     50.6   2014-02-18    1.18e8
## 5   540   7.5e7 rap mus~ en                Holl~     10.6   2003-06-09    5.11e7
## 6  3238   8  e6 califor~ en                Litt~     14.8   2006-07-26    1.01e8
## # i 11 more variables: runtime <dbl>, vote_average <dbl>, vote_count <dbl>,
## #   genre <chr>, release_year <dbl>, release_month <dbl>, release_day <dbl>,
## #   first_actor <chr>, first_actor_gender <chr>, director_first_name <chr>,
## #   director_gender <chr>
```

Your Answer:

The variable Keywords is nominal. The variable Vote Average is ordinal. The variable Budget is ratio. The variable release_year is interval.

3. A movie studio wants to know which types of movies give maximal profit. Perform the following steps to provide the movie studio with an analysis which corresponds to their request:

a. Create the variable profits as the revenue of a movie minus its budget. Report its mean, median, maximum, and minimum. **[2 points]**
b. Which movie has the highest profits in your data set and how much are these profits. Which movie has the lowest and how much are its profits? If multiple movies have the exact same highest or lowest profits, give only one example. **[2 points]**
c. Create a boxplot of the variable profits. Make sure it has an appropriate title, and appropriate titles and labels for the x- and y-axis. Give Q1, Q2, Q3, and Q4. What does this tell you about the nature of making money in the movies industry? **[2 points]**
d. Add a new variable to your data set the log of profits. When creating this variable, what happens to movies for which profits is zero or negative? What then happens when you calculate the mean of log of profits? **[2 points]**
e. For movies that have a profit of zero or less, replace log of profits with "NA". What is now the mean of log of profits? Create a boxplot for log of profits, again with an appropriate title, x- and y-axis labels. How does it compare to the boxplot you made under c.)? **[2 points]**
f. Create a scatterplot of with the runtime of movies on the x-axis and the average vote of movies on the y-axis. What do you conclude from the scatterplot? Are movies with a longer runtime considered worse or better by the audience, or does the audience not have a preference? Why do you think this is the case? **[2 points]**

For each step, you should provide first all the code you used to answer the question and then formulate an answer using full sentences.

Step a

```r
#Here I am making a new variable called profits
library(tidyr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4
## -- 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
```

```r
movies2 = movies2 %>%
  mutate(profits = revenue -budget)
#Here I am computing the minimum, maximum and mean values of the variable profits
summary(movies2$profits)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -111007242   -1498570          0   55265374   59412351 1316249360
```

Your Answer:

The mean of profits is 55265374 dollars. The median of profits is 0. The minimum of profits is -111007242. The maximum of profits is 1316249360.

Step b

```r
#Here I am making a seperate dataset with 1 variable and 1 observation to see which movie
↪  has the highest profits and how much it is.
max_profit_movie = movies2[which.max(movies2$profits), "title"]
print(max_profit_movie)
```

```
## # A tibble: 1 x 1
##   title
##   <chr>
## 1 Furious 7
```

```r
max_profit = max(movies2$profits)

#Here I am doing the same thing as above but then for the lowest profits
min_profit_movie = movies2[which.min(movies2$profits), "title"]
print(min_profit_movie)
```

```
## # A tibble: 1 x 1
##   title
##   <chr>
## 1 Mars Needs Moms
```

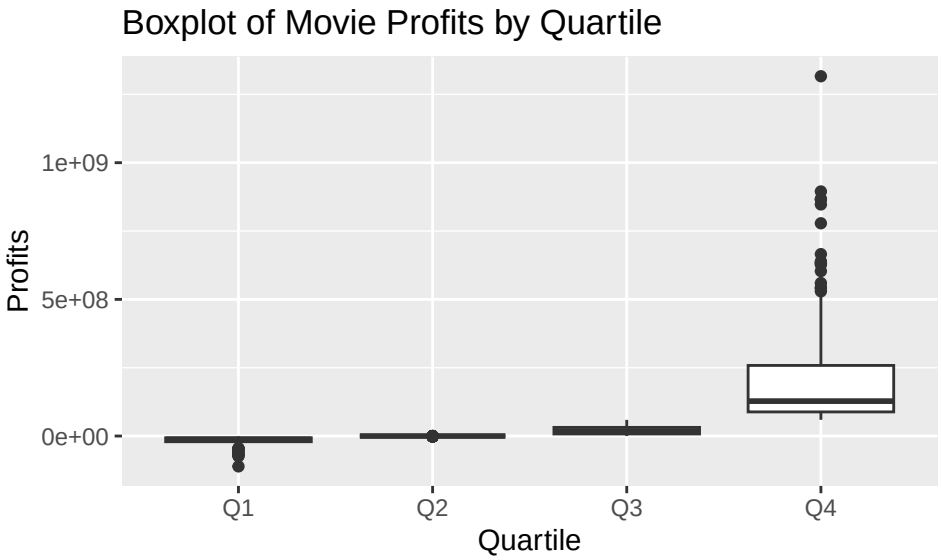
```r
min_profit = min(movies2$profits)
```

Your Answer:

The movie with the highest profit of 1316249360 dollars is Furious 7. The movie with the lowest profit is Mars Needs Moms and has a negative profit of minus 111007242 dollars.

Step c

```r
#Here I am dividing the profits into quartiles as I have to plot a boxplot
movies2$quartile = cut(movies2$profits, breaks = quantile(movies2$profits, probs = c(0,
↪  0.25,0.5, 0.75, 1)), labels = c("Q1", "Q2", "Q3", "Q4"), include.lowest = T)

#Here I am plotting a boxplot for the profits with quartiles
library(ggplot2)

ggplot(movies2, aes(x = quartile, y = profits)) +
  geom_boxplot() +
  labs(title = "Boxplot of Movie Profits by Quartile", x = "Quartile", y = "Profits")
```



Your Answer:

The boxplot shows that most movies make very little profit, with many barely breaking even in Q2 or even losing money in Q1. Only a small share of films in the top quartile in Q4 generate significant profits, and a handful of blockbuster hits drive the industry's earnings. Overall, the movie industry is high-risk, high-reward, relying heavily on rare mega-hits to stay profitable.

Step d

```r
#Here I make a new variable with the log transformed profits
movies2 = movies2 %>%
  mutate(log_profits = log(profits))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `log_profits = log(profits)`.
```

```
## Caused by warning in `log()`:
## ! NaNs produced
```

```
#Here I compute the mean but we were not sure to remove the NAs or not so in the first
↪  one we did not remove it and in the second we did to see if we would get different
↪  answers
mean(movies2$log_profits)
```

```
## [1] NaN
```

```
mean(movies2$log_profits, na.rm = T)
```

```
## [1] -Inf
```

Your Answer:

When you do not remove the NAs, while computing the mean of the log transformed profits you get an NaN, so a not a number error. However, if you do remove the NAs while computing the mean of the log transformed profits you get a minus Inf.

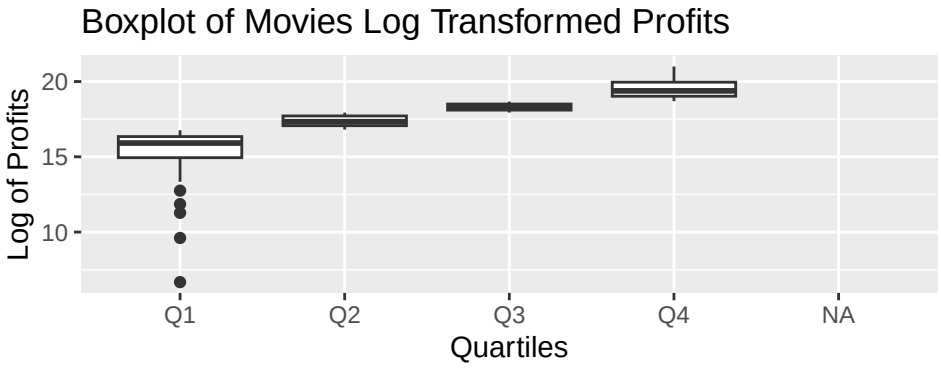
Step e

```
#Here I put NA instead of the profits that are negative or equal to 0, then I used the
↪  same code as before to calculate the logs and then I calculated the mean of the logs
↪  of the profits again.
movies2$profits = ifelse(movies2$profits <= 0 , NA, movies2$profits)
movies2 = movies2 %>%
  mutate(log_profits = log(profits))
mean(movies2$log_profits, na.rm = T)
```

```
## [1] 17.62456
```

```
#As I have to make a boxplot, I first divided the logs into quartiles
movies2$quartiles = cut(movies2$log_profits, breaks = quantile(movies2$log_profits, probs
↪  = c(0, 0.25,0.5, 0.75, 1), na.rm = T), labels = c("Q1", "Q2", "Q3", "Q4"),
↪  include.lowest = T)

# Here I am plotting the boxplot with the log transformed profits with quartiles
ggplot(movies2, aes(x = quartiles, y = log_profits)) +
  geom_boxplot() +
  labs(title = "Boxplot of Movies Log Transformed Profits", x = "Quartiles", y = "Log of
  ↪  Profits")
```

```
## Warning: Removed 310 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



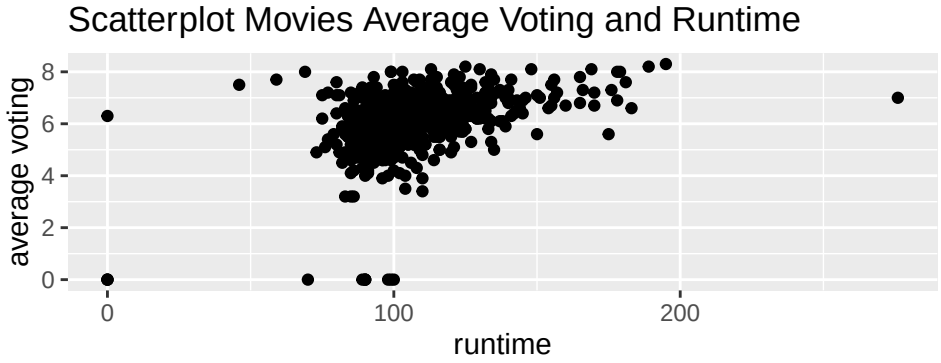
Your Answer:

The mean of the log of the profits is now 17.62456.

The first boxplot of the profits before being log-transformed shows that movie profits are extremely skewed, with most films making very little and a few blockbusters dominating the scale. This highlights the "winner-takes-most" nature of the industry but makes it hard to see differences among the majority of movies. The second boxplot, so the log-transformed profits, reduces the effect of extreme outliers, spreading the data more evenly across quartiles. This makes the relative differences between groups clearer and shows a smoother upward trend in profits from Q1 to Q4. Together, they reveal both the extreme inequality of raw profits and the more balanced progression when viewed on a log scale.

Step f

```
#Here I plot a scatterplot with runtime on x-axis and on the y-axis the vote average

ggplot(movies2, aes(x = runtime, y = vote_average)) +
  geom_point() +
  labs(title = "Scatterplot Movies Average Voting and Runtime", x = "runtime", y =
  ↪  "average voting")
```



Your Answer:

The scatterplot shows a weak positive relationship between runtime and ratings: longer movies tend to score slightly higher, though the effect is small. Most films cluster around 90–120 minutes with average ratings, while very long films (150+ minutes) often maintain strong scores around 7–8. This may reflect that only higher-quality projects get extended runtimes, supported by bigger budgets and stronger storytelling. Short films under 60 minutes generally perform poorly. Overall, runtime has some influence, but we think that audience satisfaction depends far more on quality than length.

2 Week 2

1 Is your dataset movies2.tsv the full population, or is it a sample of a larger population? If the latter, how would you describe the full population? [**4 points**]

Your Answer:

Our dataset is probably a sample of a larger population as we only see 606 subjects. I would describe the full population as all the movies existing starting from 1990 until 2016.

2

a. For which actor in your data set do you observe the most movies? [**2 points**]
b. What is the average revenue of the movie in which this actor plays and does the revenue lie above or below the revenue of an average movie according to your data set? [**2 points**]

c. How trustworthy do you consider your conclusion to answer 2b? Use the term "law of large numbers" in your explanation. [**2 points**]

step a

```
#looking for the actor with the most movies
library(dplyr)
actor = movies2 %>%
  count(first_actor, sort = T) %>%
  head(1)
print(actor)
```

```
## # A tibble: 1 x 2
##   first_actor      n
##   <chr>        <int>
## 1 Bruce Willis     5
```

Your Answer:

Actor Bruce Willis played in 5 movies, which is the most.

step b

```
#computing the average revenue of Bruce
average_rev_Bruce = movies2 %>%
  filter(first_actor == "Bruce Willis") %>%
  summarise(av_rev_Bruce = mean(revenue))
print(average_rev_Bruce)
```

```
## # A tibble: 1 x 1
##   av_rev_Bruce
##          <dbl>
## 1   175160583.
```

```
#computing the average revenue general
average_rev_generally = movies2 %>%
  summarise(rev_gen = mean(revenue))
print(average_rev_generally)
```

```
## # A tibble: 1 x 1
##     rev_gen
##       <dbl>
## 1 84953602.
```

Your Answer:

The average revenue of the movies in which Bruce Willis played is 175.160.583 and the average revenue of all movies is 84.953.602. So the average revenue of Bruce Willis lies above the general average revenue.

step c

```
#computing the varience of the subset Bruce of the revenue and the general variance of
↪  the full dataset
library(tidyverse)

var_bruce = movies2 %>%
  filter(first_actor == "Bruce Willis") %>%
  summarise(var_bruce = var(revenue))
print(var_bruce)
```

```
## # A tibble: 1 x 1
##   var_bruce
##       <dbl>
## 1   7.83e16
```

```
var_gen = movies2 %>%
  summarise(var_gen = var(revenue))
print(var_gen)
```

```
## # A tibble: 1 x 1
##   var_gen
##     <dbl>
## 1 2.84e16
```

Your Answer: Based on the variances we conclude that our answer is not trustworthy because the variance of the revenue of Bruce was far greater than the variance of the full data set, which is based on a higher n. This means that the expected value of the sample of Bruce Willis will deviate from his full population, which would consists of all the movies he played in, as our average is only based on 5 movies of Bruce. So the higher the n, the lower the variance and the more the sample mean approaches the expected value, which is also known as the law of large numbers. This is not the case with Bruce Willis.

3 For this question, you will assume that your data set is the full population.

a. Recode profits such that it is expressed in millions. What is the variance of the variable profits (in millions) in your data set? **[2 points]**
b. Create a new data set, called movies_sample. Make sure that it is a random sample of your data set of 25 movies. What is the variance of profits in this random sample? How does it compare to the variance of profits in 2a? **[2 points]**
c. In a for loop, create 100 different samples of 25 movies, as in b, and estimate the variance within each sample. Save the variance of each sample in a vector called sample_vars. So the first position of the vector would have the variance of the first sample, the second position the variance of the second sample, etc. Print the start of this vector. **[2 points]**

d. Summarize and make a histogram of sample_vars. What is the mean, standard deviation and shape of its distribution? **[2 points]**
e. In your opinion, is a sample of 25 movies sufficient to get a reliable estimate of the population variance of profits, using the sample variance? Explain? **[2 points]**

step a

```
#recoding profits into millions and calculating the variance of the mean
movies2$profits = movies2$profits/1000000
library(tidyverse)

var_profits2 = var(movies2$profits, na.rm = T)
print(var_profits2)
```

```
## [1] 32211.42
```

Your Answer:

The variance of profits is 3.2211422×10^4 million.

step b

```
#creation of a sample of 25 movies, new data set and computing the variance of this
↪  sample
set.seed(123)
movies_sample = movies2[sample(nrow(movies2), 25),]
var_sample_profits = var(movies_sample$profits, na.rm =T)
print(var_sample_profits)
```

```
## [1] 20649.6
```

Your Answer:

The variance of the profits of the sample is 2.06496×10^4 million and is higher than the variance of the full population of 3.2211422×10^4 million.

step c

```
#creating a for loop to make different samples with vectors computing the variance
set.seed(123)
sample_vars = numeric(100)
for (i in 1:100) {
  samples_movies = movies2[sample(nrow(movies2), 25),]
  sample_vars[i] = var(samples_movies$profits, na.rm = T)
}
head(sample_vars)
```

```
## [1] 20649.600  8090.688  1767.455 12320.060 34236.920 43705.870
```

Your Answer:

the variances of the start of the vector with the sample variances are: 2.06496×10^4, 8090.6884611, 1767.4546965, 1.232006×10^4, 3.423692×10^4, 4.370587×10^4

step d

```r
#computing the mean, standard deviation and plotting the histogram of the sample
↪  variances
summary(sample_vars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1047    7776   21823   30140   43775  170696
```

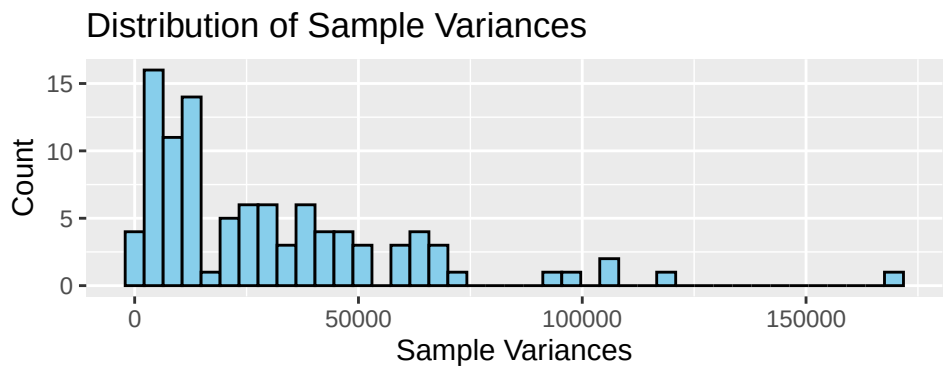
```r
mean(sample_vars)
```

```
## [1] 30139.83
```

```r
sd(sample_vars)
```

```
## [1] 29945.74
```

```r
sample_vars = data.frame(sample_vars)
library(ggplot2)
ggplot(sample_vars, aes(x = sample_vars)) +
  geom_histogram(binwidth = diff(range(sample_vars))/40, colour = "black", fill =
  ↪  "skyblue") +
  labs(title = "Distribution of Sample Variances", x = "Sample Variances", y = "Count")
```



Your Answer:

The mean of the sample variances is 30139.83 million and the standard deviation is 29945.74 million.The shape of the distribution is more right skewed.

step e

Your Answer:

Given the fact that the full population consists of 606 movies, I think personally that a sample size of 25 is too small to get a reliable estimate of the population variance of profits, using the sample variance and I also think that because there are big differences in profits across movies that a sample may be not representative for the whole population.

Your answer here

3 Week 3

For the next part of the assignment, assume that the movies in your data frame are a random sample of a larger population of movies.

1

a. Create a new data set that only includes movies that are of the genre "Thriller". For these thriller movies, give a 99 percent confidence interval for the variable *runtime*. Interpret the result. **[2 points]**
b. Now, assume that the variance of *runtime* amongst thriller movies in your data is exactly the same as the variance of *runtime* in the population. Under this assumption, give a 99 percent confidence interval for the variable *runtime* among thriller movies. Interpret the result. Is you confidence interval wider or less wide than the one you found under question 1a? Why is that the case? **[2 points]**

step a

```
#WRITE YOUR CODE HERE
thriller = movies2 %>%
  filter(genre == "Thriller")

N<-length(thriller$runtime)
mean_runtime <- mean(thriller$runtime)
sd_runtime   <- sd(thriller$runtime)
t_a2 <- qt(0.995, df = N - 1)


LOW <- mean_runtime - t_a2 * (sd_runtime/ sqrt(N))
UP  <- mean_runtime + t_a2 * (sd_runtime/ sqrt(N))


CI<-paste0("(",round(LOW,4),",",round(UP,4),")")
print(CI)
```

```
## [1] "(98.6087,106.0333)"
```

Your Answer:

There is a 99 percent certainty that the value of runtime is between the 98.6087 and 106.0333.

step b

```
#WRITE YOUR CODE HERE
N<-length(thriller$runtime)
mean_runtime <- mean(thriller$runtime)
sd_runtime   <- sd(thriller$runtime)
z_a2 <- qnorm(0.995)


LOW <- mean_runtime - z_a2 * (sd_runtime/ sqrt(N))
UP  <- mean_runtime + z_a2 * (sd_runtime/ sqrt(N))


CI<-paste0("(",round(LOW,4),",",round(UP,4),")")
print(CI)
```

```
## [1] "(98.6971,105.9448)"
```

Your Answer:

There is a 99 percent certainty that the value of runtime is between 98.6971 and 105.9448. The interval is less wide than the interval of question 1a. This is because in 1a we don't know the standard deviation, while in question b the sd is certain. We have to take the uncertainty of the sd into account in the interval of question 1a, and thats why the interval of 1a is wider than 1b.

2

a. Using an appropriate five-step procedure, set up a test for the null hypothesis that the variance of runtime equals 500. Clearly state your null hypothesis, alternative hypothesis your test statistic, your critical value, and your conclusion. [**2 points**]
b. For the validity of your test in 2a, what assumption about the distribution of revenue needs to hold? Make an appropriate plot to test this assumption. What do you conclude? [**2 points**]

step a

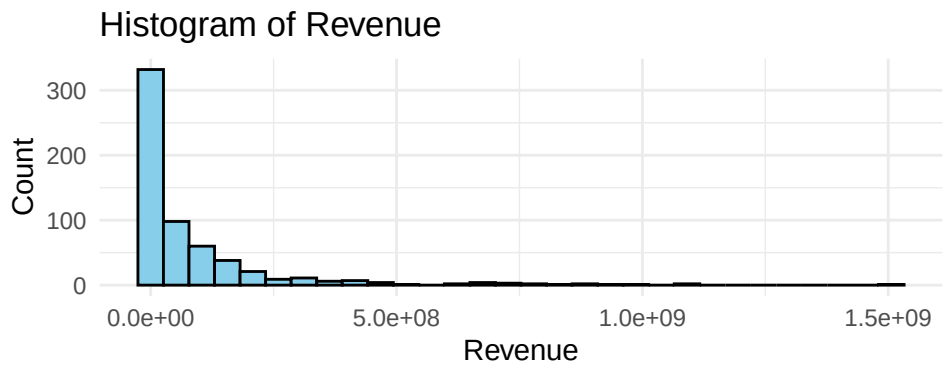
```
#WRITE YOUR CODE HERE
s2 <- var(movies2$runtime)

sigma_sq <- 500

chi2_stat <- (N - 1) * s2 / sigma_sq

alpha <- 0.05
chi2_L <- qchisq(alpha/2, df = N-1)
chi2_U <- qchisq(1 - alpha/2, df = N-1)
```

Your Answer:

H0 –> sigma^2 = 500 H1 –> sigma^2 IS NOT 500. The test statistic is 81.16 if you take a alpha of 0.05 then you will get critical values 57 and 106.63. The test statistic falls in the interval of the critical values so the hypothesis should not be rejected.

step b

```
#WRITE YOUR CODE HERE
ggplot(movies2, aes(x = revenue)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Histogram of Revenue", x = "Revenue", y = "Count") +
  theme_minimal()
```



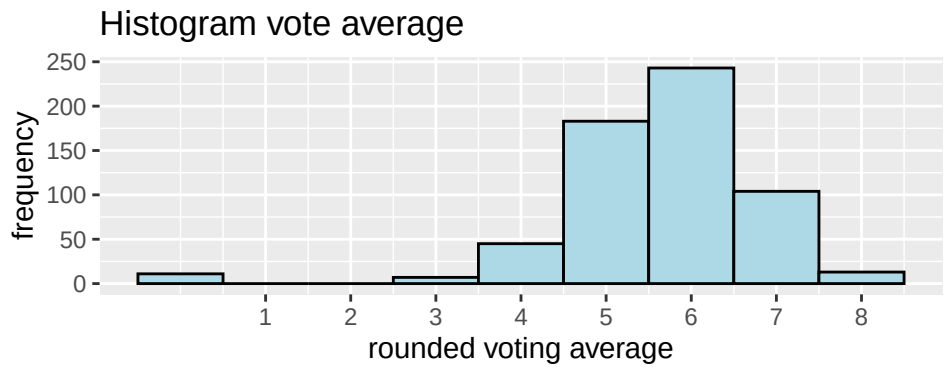
Your Answer: For the validity of the test the values should be normally distributed, but as can be seen in the graph, this is not the case. So the conclusions made in the previous excercise aren't valid.

3. There is an argument going on in the movie studio. *Bob* claims that they should make higher-quality movies, as this will bring in more profits. *Chantal* disagrees. She tells Bob that mediocre movies bring in the most profits. You are asked to advise on who is right.

a. Create a new variable called vote_average_rounded. Make sure this variable is the same as vote_average, but without any decimals (i.e., a 6.3 becomes a 6, a 8.7 an 8, etc.). Display a histogram of vote_average_rounded. [**2 points**]
b. Create a scatter plot with vote_average_rounded on the x axis and the mean of profits within each category of vote_average_rounded on the y-axis. Make sure it has an appropriate title, and appropriate titles and labels for the x- and y-axis. At which rating of movies are profits the highest? [**3 points**]
c. Recreate the scatter plot with year on the x axis and mean_profits on the y-axis, but now add bars around each point, indicating the 95% confidence interval. [**3 points**]
d. Write an advice to settle the argument between Bob and Chantal. [**4 points**]

step a

```
#WRITE YOUR CODE HERE
movies2 <- movies2 %>%
  mutate(vote_average_rounded = floor(vote_average))

# Histogram met ggplot2
ggplot(movies2, aes(x = vote_average_rounded)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  labs(title = "Histogram vote average",
       x = "rounded voting average",
       y = "frequency") +
  scale_x_continuous(breaks = seq(1, 10, 1))
```



Your Answer:

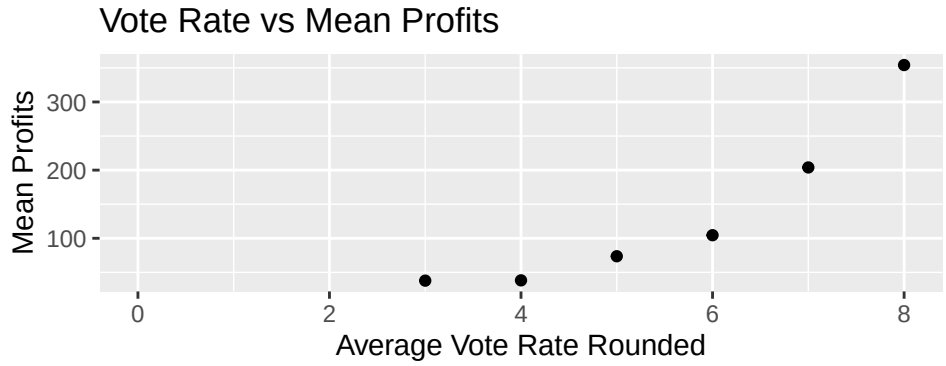
The floor function is used to round the numbers and ggplot + geom histogram to plot the graph/

step b

```
#calculting mean profits per category and plotting scatterplot

mean_profits_plot = movies2 %>%
  group_by(vote_average_rounded) %>%
```

```
  summarise(mean_profit = mean(profits, na.rm = T))

ggplot(mean_profits_plot, aes(x = vote_average_rounded, y = mean_profit)) +
  geom_point()+
  labs(title = "Vote Rate vs Mean Profits", x = "Average Vote Rate Rounded", y = "Mean
  ↪  Profits")
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```

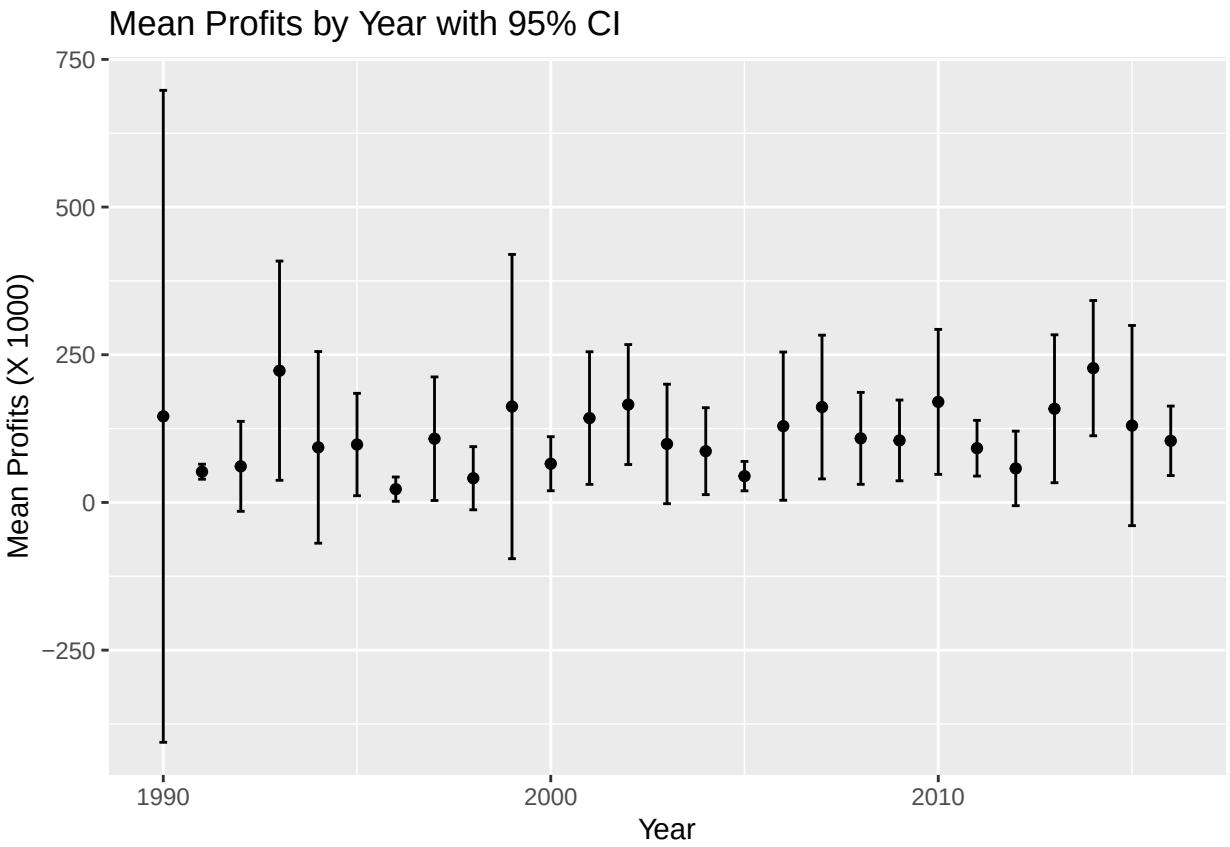


Your Answer:

The profits are the highest if the movie has a rating of 8.

step c

```
#new sub data set with yearly mean profits and 95% confidence interval.
yearly_profits = movies2 %>%
  group_by(release_year) %>%
  summarise(mean_profit = mean(profits, na.rm = T),
            n = sum(!is.na(profits)),
            se = sd(profits, na.rm = T)/ sqrt(n),
            lower_ci = mean_profit - qt(0.975, df = n - 1) * se,
            upper_ci = mean_profit + qt(0.975, df = n-1) * se)

ggplot(yearly_profits, aes(x  = release_year, y = mean_profit)) +
  geom_point()+
  geom_errorbar(aes(ymin = lower_ci, ymax = upper_ci), width = 0.2) +
  labs(title = "Mean Profits by Year with 95% CI", x = "Year", y = "Mean Profits (X
  ↪  1000)")
```



Your Answer:

Here we plotted the yearly average profits with a 95% confidence interval

step d

Your Answer:

When we put the results together, the rating group at the top in plot 3b looks most promising, because it has the highest average profit on our data. At the same time, the year plot in 3c shows that the results can change over time and that our certainty is different for each year. A good plan is to aim for projects that are likely to fall in the higher rating group; while also checking the recent years with narrow bars when setting expectations. We have to be careful with years that have wide bars, because those averages are less stable. This means focusing on the higher rating group, but also keeping an eye on changes between years. Also do not rely on one single year with low certainty

4 Week 4

1. There is another argument going on in the movie studio. *Bob* claims that production budgets are getting out of hand, and that the studio should focus on making cheaper movies. *Chantal* disagrees. She tells Bob that "Every dollar we spend on movie production is more than offset by the increase in movie profits''.

a. Set up a regression model to test Chantal's claim, and estimate it. That is, estimate:

$$\text{Profits}_i = \beta_0 + \beta_1 \text{Budget}_i + \varepsilon_i.$$

Print a summary of your estimated model. **[2 points]**

b. What is the estimated value of β_1 and how do you interpet it? **[2 points]**
c. Test for the null hypothesis that $\beta_1 \geq 0$. Report the p-value and state your conclusion. **[2 points]**
d. Next, estimate the model

$$\text{Log Profits}_i = \beta_0 + \beta_1 \text{Log Budget}_i + \varepsilon_i.$$

When creating the variables Log Profits and Log Budget, make sure that movies with a Revenue or Budget of zero are assigned the value "NA". Print a summary of your estimated model **[2 points]**
e. What is the estimated value of β_1 and how do you interpet it? **[2 points]**
f. Which model has better fit? The level-level model or the log-log model? Explain. **[2 points]**
g. Who do you think is correct? Bob or Chantal? What would you advise the movie studio to do? **[2 points]**

step a

```
#WRITE YOUR CODE HERE
model = lm(profits ~ budget, data = movies2)
summary(model)
```

```
##
## Call:
## lm(formula = profits ~ budget, data = movies2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -390.64  -52.94  -18.52   21.82  864.30
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.926e+01  1.146e+01   1.681   0.0938 .
## budget      2.277e-06  1.748e-07  13.025   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.2 on 294 degrees of freedom
##   (310 observations deleted due to missingness)
## Multiple R-squared:  0.3659, Adjusted R-squared:  0.3637
## F-statistic: 169.6 on 1 and 294 DF,  p-value: < 2.2e-16
```

Your Answer:

Here we set up a regression model and printed the summary of the model

step b

```
#look at summary of model above
```

Your Answer:

Looking at the summary of the model above, we see that the estimated value of beta 1 is 2.277e-06. This means that an increase in the budget is associated with an increase in profits by 2.277e-06. So the coefficient is very small and implies that there could be a weak relationship between budget and profits.

step c

```
#look at summary of model above
```

Your Answer:

H0: beta 1 is 0 and H1: beta 1 is not equal to 0. The test statistic is beta 1 and we reject for small values. Looking at the summary of the model above, we see under the column variable Pr(>|t|) that the p-value is <2e-16 which is extremely small and smaller than any significance rate. Therefore we reject the null hypothesis that beta 1 is equal to 0.

step d

```
#We already created the new variable log profits before so here we start first with
↪  creating the new variable log budget and make a new regression model

movies2 = movies2%>%
  mutate(budget = ifelse(budget == 0, NA_real_, budget))

movies2 = movies2 %>%
  mutate(log_budget = log(budget))

movies2 = movies2 %>%
  mutate(log_budget = ifelse(is.infinite(log_budget), NA_real_, log_budget))

model_new = lm(log_profits ~ log_budget, data = movies2)
summary(model_new)
```

```
##
## Call:
## lm(formula = log_profits ~ log_budget, data = movies2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6650 -0.5956  0.1941  0.7592  3.0653
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.11309    0.95830   6.379 7.48e-10 ***
## log_budget   0.68289    0.05581  12.235  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.22 on 276 degrees of freedom
##   (328 observations deleted due to missingness)
## Multiple R-squared:  0.3517, Adjusted R-squared:  0.3493
## F-statistic: 149.7 on 1 and 276 DF,  p-value: < 2.2e-16
```

Your Answer:

Here we have a new regression model of the logarithm of budgets and profits

step e

```
#Look at summary model_new above
```

Your Answer: The new estimated value of beta 1 is 0.68289. We can interpret this now as an elasticity. So in this case a 1% increase in budget is associated with a 0.68289% increase in the profit.

step f

```
#Look at summary model and summary model_new above
```

Your Answer:

When looking at which model has a better fit, we compare the R^2 values. In the level level model the value of the R^2 is 0.3659 and in the log log model the R^2 is 0.3517. This implies that the level level model is relatively a better fit as the R^2 of the level level model is greater than the R^2 of the log log model.

step g

```
#Concluding by looking at the summary of model and model_new above
```

Your Answer:

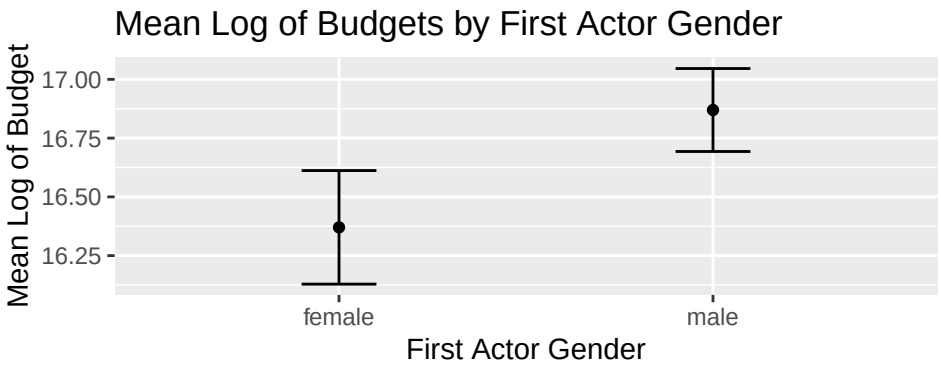
I think that Bob is right as 1% increase in the budget is associated to an increase of 0.68% in profits.This implies that the budget is getting out of hand just as Bob stated and that Chantal's argument that spending more brings more than proportional profits is not backed by our regression model. In conclusion, the movie studio should also be producing more cheaper movies.

2

a. Make a plot with a 95% confidence interval with the mean log of budget on the y-axis, and whether the first actor of the movie is male or female on the x-axis. What do you conclude? **[2 points]**
b. Estimate the following simple OLS model: $log(budget)_i = \beta_0 + \beta_1(FirstActorMale)_i + \varepsilon_i$. Is the estimated coefficient for β_1 significantly different from zero? How do you interpret its estimate, and how does this relate to your conclusion in 2a? **[2 points]**
c. Now, have a close look at your data frame. Can you find any instances of male first actors who are wrongly labeled as being female, or vice versa? What would such mislabelling mean for the coefficient you estimated under 2b? **[2 points]**

step a

```
#WRITE YOUR CODE HERE
plot_log = movies2%>%
  filter(!is.na(first_actor_gender))%>%
  group_by(first_actor_gender) %>%
  summarise(mean_log_budget = mean(log_budget, na.rm = T),
            se = sd(log_budget, na.rm = T)/sqrt(n()),
            low_ci = mean_log_budget - 1.96 * se,
            up_ci = mean_log_budget + 1.96 * se)

ggplot(plot_log, aes(x = first_actor_gender, y = mean_log_budget)) +
  geom_point()+
  geom_errorbar(aes(ymin = low_ci, ymax = up_ci), width = 0.2)+
  labs(title = "Mean Log of Budgets by First Actor Gender", y = "Mean Log of Budget", x =
  ↪  "First Actor Gender")
```



Your Answer: Movies with male first actors have higher average budgets compared to those with female leads based on the mean of the log of budgets. This suggests a budget disparity associated with the gender of the actor.

step b

```
#WRITE YOUR CODE HERE
ols_model = lm(log_budget ~ first_actor_gender, data = movies2)
summary(ols_model)
```

```
##
## Call:
## lm(formula = log_budget ~ first_actor_gender, data = movies2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8695  -0.5691   0.3472   1.0353   2.4989
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            16.3703     0.1667  98.214   <2e-16 ***
## first_actor_gendermale   0.4992     0.1932   2.584   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.787 on 448 degrees of freedom
##   (156 observations deleted due to missingness)
## Multiple R-squared:  0.01469,    Adjusted R-squared:  0.01249
## F-statistic: 6.678 on 1 and 448 DF,  p-value: 0.01007
```

Your Answer:

The value of beta 1 of 0.4992 is significantly greater than 0. This means that the log of budget is 0.4992 higher for male leads versus female leads on average. So male actors are linked to higher budgets and this perfectly confirms our conclusion of the plot where we saw the same thing: male actors are associated with higher budgets compared to female leads

step c

```
#WRITE YOUR CODE HERE
```

Your Answer:

Yes, we have found female actors mislabeled as male like Brit Marling who is a female and was labeled as male and also Jackie Chan who is mistakenly labeled as female. This implies that the coefficient is undervalued for male as some female actors are also taken into account of which the budget is usually lower than the male.

5 Week 5

a. Create a plot of the mean profits by month of release. Do you see any indication that month of release matters to the profits of the movie? [**2 points**]

b. Estimate an OLS model which has as dependent variable the log of profits of a movie, and as independent variable the log of budget, a dummy for whether the movie was released in english or not, and a linear term for the month of release. Show a summary of the resulting model and interpret each coefficient. [**4 points**]

c. Test for the hypothesis that the coefficient that belongs to month of release is zero. [**2 points**]

d. Based on your plot in a.) do you consider the choice that month of release enters the model linearly under b.) reasonable? Estimate a specification that allows for a more flexible curve. In this new specification, test for the null hypothesis that month of release does not impact profits. This might require testing multiple terms at once. [**4 points**]

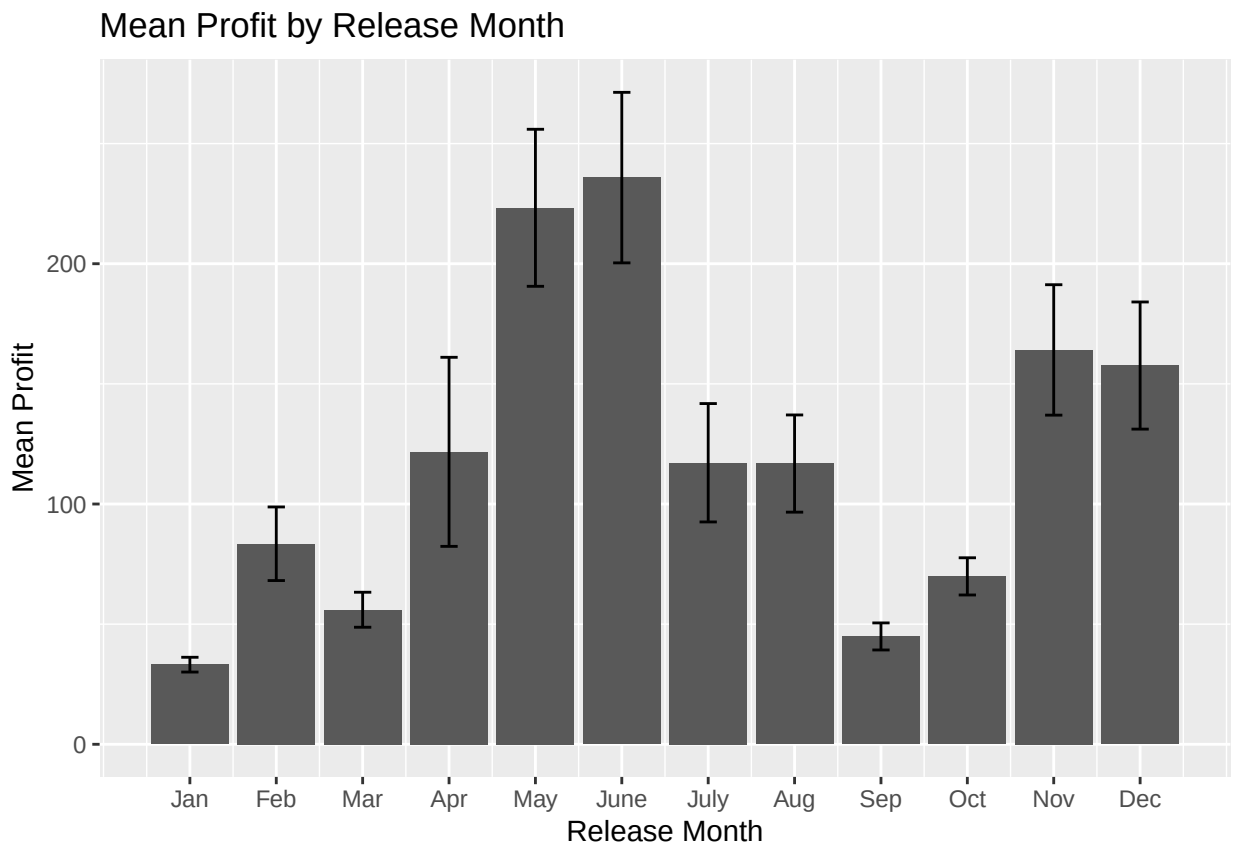
e. One executive at the studio wants to time the release of the movie to a specific month of the year such that they can maximize revenue. Based on your model under d.), What would you advise the movie studio regarding the timing of the release of the movie? [**2 points**]

The movie studio that you work at is releasing a new movie in 2026. It will be an English-spoken Thriller movie with a budget of 40,000,0000.

f. Estimate a model that is able to predict the revenue of this movie. Give its predicted revenue and include a 99% prediction interval. [**6 points**]

step a

```
#WRITE YOUR CODE HERE
monthly_profits = movies2 %>%
  group_by(release_month) %>%
  summarise(mean_profits = mean(profits, na.rm = T),
            se_profit = sd(profits, na.rm = T) / sqrt(n()))

ggplot(monthly_profits, aes(x = release_month, y = mean_profits)) +
  geom_col()+
  geom_errorbar(aes(ymin = mean_profits - se_profit, ymax = mean_profits + se_profit),
  ↪  width = 0.2) +
  scale_x_continuous(breaks = 1:12, labels = c("Jan", "Feb", "Mar", "Apr", "May", "June",
  ↪  "July", "Aug", "Sep", "Oct", "Nov", "Dec")) +
  labs(title = "Mean Profit by Release Month", y = "Mean Profit", x = "Release Month")
```



Your Answer:

We see that there is significant variety between months, so we can indicate that the release month does matter to profits with peaks in May, June, November and December.

step b

```
#creating dummy variable english
movies2$English_release = ifelse(movies2$original_language == "en", 1, 0)

#OLS model
model_week_5 = lm(log_profits ~ log_budget + English_release + release_month, data =
↪  movies2)
summary(model_week_5)
```

```
##
## Call:
## lm(formula = log_profits ~ log_budget + English_release + release_month,
##     data = movies2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8694 -0.6303  0.1712  0.7701  3.1939
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.90179    0.95983   6.149 2.75e-09 ***
```

```
## log_budget       0.65976    0.05773  11.428  < 2e-16 ***
## English_release  0.36321    0.34519   1.052    0.294
## release_month    0.03811    0.02102   1.813    0.071 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.215 on 274 degrees of freedom
##   (328 observations deleted due to missingness)
## Multiple R-squared:  0.3618, Adjusted R-squared:  0.3548
## F-statistic: 51.78 on 3 and 274 DF,  p-value: < 2.2e-16
```

Your Answer: The first coefficient of log_budget is 0.65976, which implies that an increase in the budget of 1% is associated with an 0.6597 percent increase in the profits, holding all the other variables constant. The second coefficient of English_release implies an association of an increase of 0.36321 percent of profits if the original language of the movie is English, compared to non-English movies and holding all the other variables constant. The third coefficient of the linear term of the release months implies that each additional month number is associated with an increase of 0.03811, holding all the other variables constant

step c

```
#WRITE YOUR CODE HERE
```

Your Answer:

The null hypothesis is that beta 3, which is the coefficient of the release months, is 0 and the alternative hypotheis is that beta 3 is not equal to 0 and therefore significant. If we look in the summary of the OLS model, we observe a p-value for release months of 0.071. At common significance level of alfa equals 0.05, a p-value of 0.071 is greater than the significance level and therefore we fail to reject the null hypothesis. This means that there is not enough evidence that the coefficient of the release month is significant, holding the other variables constant.

step d

```
#Converting release month as a factor
movies2$release_month = as.factor(movies2$release_month)

#new OLS model more flexible

flexible_model = lm(log_profits ~ log_budget + English_release + release_month, data =
↪  movies2)
summary(flexible_model)
```

```
##
## Call:
## lm(formula = log_profits ~ log_budget + English_release + release_month,
##     data = movies2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9313 -0.5571  0.1426  0.7887  2.8722
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.34049    1.01900   6.222 1.91e-09 ***
```

```
## log_budget       0.62601    0.06186  10.120  < 2e-16 ***
## English_release  0.35895    0.34537   1.039  0.29960
## release_month2   0.48413    0.40152   1.206  0.22899
## release_month3  -0.16927    0.36567  -0.463  0.64381
## release_month4   0.37131    0.35626   1.042  0.29825
## release_month5   0.93623    0.41063   2.280  0.02341 *
## release_month6   0.87211    0.33466   2.606  0.00968 **
## release_month7   0.04033    0.33693   0.120  0.90481
## release_month8   0.25800    0.35775   0.721  0.47144
## release_month9   0.33850    0.34413   0.984  0.32619
## release_month10  0.55732    0.36112   1.543  0.12396
## release_month11  0.51879    0.33882   1.531  0.12692
## release_month12  0.67560    0.32970   2.049  0.04144 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.201 on 264 degrees of freedom
##   (328 observations deleted due to missingness)
## Multiple R-squared:  0.3985, Adjusted R-squared:  0.3689
## F-statistic: 13.46 on 13 and 264 DF,  p-value: < 2.2e-16
```

```
#Testing null hypothesis, f-test for multiple terms
lm(log_profits ~ log_budget + English_release, data = movies2)
```

```
##
## Call:
## lm(formula = log_profits ~ log_budget + English_release, data = movies2)
##
## Coefficients:
##     (Intercept)       log_budget  English_release
##          6.0406           0.6672           0.3583
```

```
anova(lm(log_profits ~ log_budget + English_release, data = movies2), flexible_model)
```

```
## Analysis of Variance Table
##
## Model 1: log_profits ~ log_budget + English_release
## Model 2: log_profits ~ log_budget + English_release + release_month
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    275 409.16
## 2    264 381.06 11    28.101 1.7699 0.0593 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Your Answer:

Looking at the plot under question a, the pattern was not clearly linear as there was more variety. So a linear specification did not fit well. After testing whether all the coefficients of the release months are equal to 0, we observe a p-value of 0.0593, which is higher than the common significance level of 0.05, so again we fail to reject the null hypothesis as we do not have enough evidence to fail the null hypothesis. So the coefficients of the release months are not significant, holding the other variables constant.

step e

```
#WRITE YOUR CODE HERE
```

Your Answer:

Based on the flexible model and the test hypothesis, we came to the conclusion that we do not have enough evidence that, while holding the other variables constant, the month of release had significant impact on profits. So maybe the executive should not look at a particular month to maximize revenue but rather look at marketing strategies and competition in the market. So just other factors to maximize revenue than release month.

step f

```
#prediction of profits new movie
model_week_5 = lm(log_profits ~ log_budget + English_release + release_month, data =
↪  movies2)
new_data = data.frame(log_budget = log(40e6), English_release = 1, release_month =
↪  factor(6, level = levels(movies2$release_month)))
predicted_log_profit = predict(model_week_5, newdata = new_data)
predicted_profit = exp(predicted_log_profit)
print(predicted_profit)
```

```
##         1
## 111495824
```

```
#prediction interval
pred_interval = predict(model_week_5, newdata = new_data, interval = "prediction", level
↪  = 0.99)
lower_bound = exp(pred_interval[1, "lwr"])
upper_bound = exp(pred_interval[1, "upr"])

cat("Predicted Profit:", predicted_profit, "\n")
```

```
## Predicted Profit: 111495824
```

```
cat("99% Prediction Interval:(", lower_bound, ",", upper_bound,") \n")
```

```
## 99% Prediction Interval:( 4695840 , 2647304477 )
```

Your Answer: The predicted profit is 68.502.641 dollars and the 99% prediction interval is (2.913.186 , 1.610.817.918)