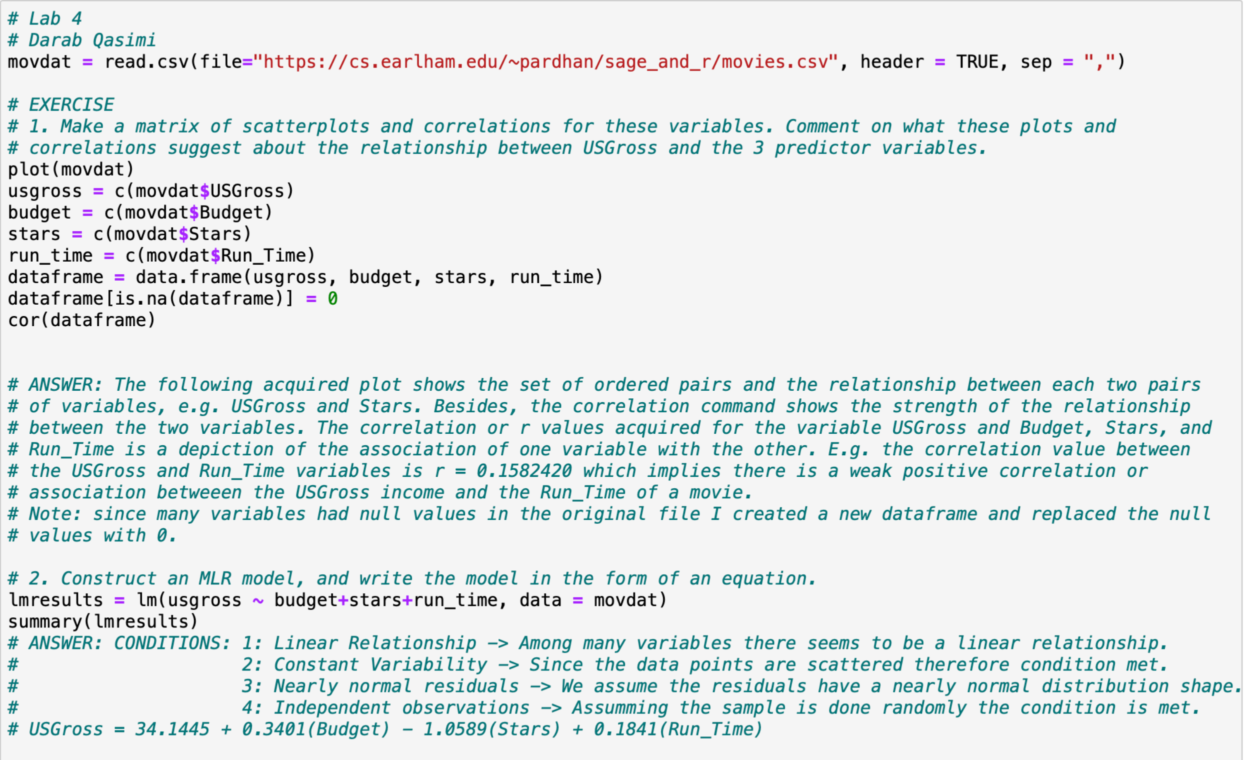
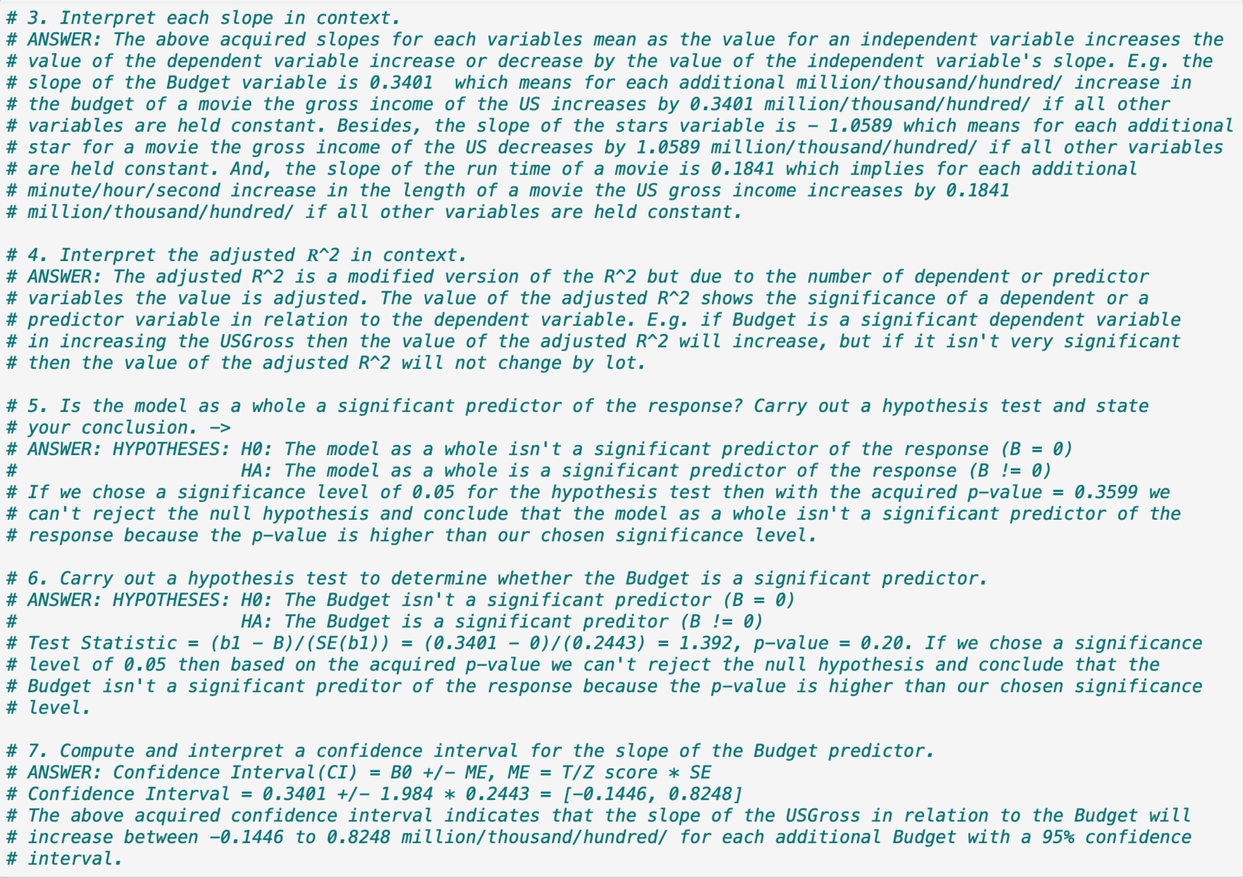
LAB 4

Darab Qasimi

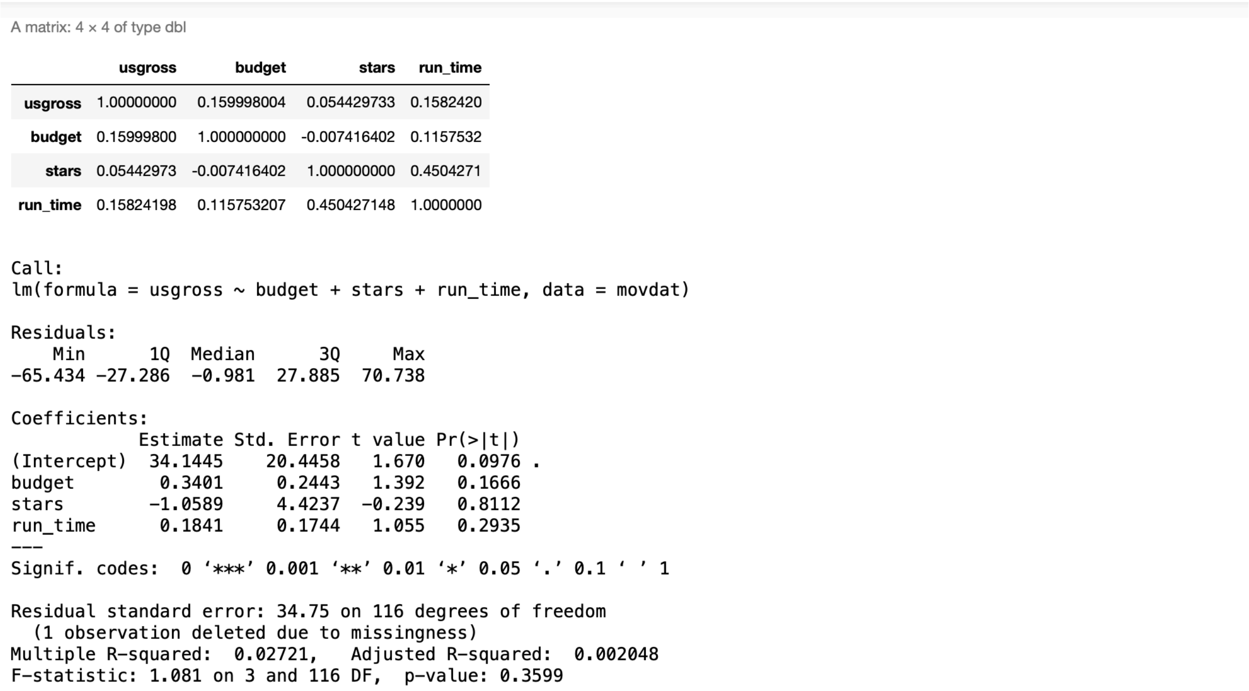


```
# Lab 4
# Darab Qasimi
movdat = read.csv(file="https://cs.earlham.edu/~pardhan/sage_and_r/movies.csv", header = TRUE, sep = ",")

# EXERCISE
# 1. Make a matrix of scatterplots and correlations for these variables. Comment on what these plots and
# correlations suggest about the relationship between USGross and the 3 predictor variables.
plot(movdat)
usgross = c(movdat$USGross)
budget = c(movdat$Budget)
stars = c(movdat$Stars)
run_time = c(movdat$Run_Time)
dataframe = data.frame(usgross, budget, stars, run_time)
dataframe[is.na(dataframe)] = 0
cor(dataframe)


# ANSWER: The following acquired plot shows the set of ordered pairs and the relationship between each two pairs
# of variables, e.g. USGross and Stars. Besides, the correlation command shows the strength of the relationship
# between the two variables. The correlation or r values acquired for the variable USGross and Budget, Stars, and
# Run_Time is a depiction of the association of one variable with the other. E.g. the correlation value between
# the USGross and Run_Time variables is r = 0.1582420 which implies there is a weak positive correlation or
# association betweeen the USGross income and the Run_Time of a movie.
# Note: since many variables had null values in the original file I created a new dataframe and replaced the null
# values with 0.

# 2. Construct an MLR model, and write the model in the form of an equation.
lmresults = lm(usgross ~ budget+stars+run_time, data = movdat)
summary(lmresults)
# ANSWER: CONDITIONS: 1: Linear Relationship -> Among many variables there seems to be a linear relationship.
#                     2: Constant Variability -> Since the data points are scattered therefore condition met.
#                     3: Nearly normal residuals -> We assume the residuals have a nearly normal distribution shape.
#                     4: Independent observations -> Assumming the sample is done randomly the condition is met.
# USGross = 34.1445 + 0.3401(Budget) - 1.0589(Stars) + 0.1841(Run_Time)
```



```
# 3. Interpret each slope in context.
# ANSWER: The above acquired slopes for each variables mean as the value for an independent variable increases the
# value of the dependent variable increase or decrease by the value of the independent variable's slope. E.g. the
# slope of the Budget variable is 0.3401  which means for each additional million/thousand/hundred/ increase in
# the budget of a movie the gross income of the US increases by 0.3401 million/thousand/hundred/ if all other
# variables are held constant. Besides, the slope of the stars variable is - 1.0589 which means for each additional
# star for a movie the gross income of the US decreases by 1.0589 million/thousand/hundred/ if all other variables
# are held constant. And, the slope of the run time of a movie is 0.1841 which implies for each additional
# minute/hour/second increase in the length of a movie the US gross income increases by 0.1841
# million/thousand/hundred/ if all other variables are held constant.

# 4. Interpret the adjusted R^2 in context.
# ANSWER: The adjusted R^2 is a modified version of the R^2 but due to the number of dependent or predictor
# variables the value is adjusted. The value of the adjusted R^2 shows the significance of a dependent or a
# predictor variable in relation to the dependent variable. E.g. if Budget is a significant dependent variable
# in increasing the USGross then the value of the adjusted R^2 will increase, but if it isn't very significant
# then the value of the adjusted R^2 will not change by lot.

# 5. Is the model as a whole a significant predictor of the response? Carry out a hypothesis test and state
# your conclusion. ->
# ANSWER: HYPOTHESES: H0: The model as a whole isn't a significant predictor of the response (B = 0)
#                     HA: The model as a whole is a significant predictor of the response (B != 0)
# If we chose a significance level of 0.05 for the hypothesis test then with the acquired p-value = 0.3599 we
# can't reject the null hypothesis and conclude that the model as a whole isn't a significant predictor of the
# response because the p-value is higher than our chosen significance level.

# 6. Carry out a hypothesis test to determine whether the Budget is a significant predictor.
# ANSWER: HYPOTHESES: H0: The Budget isn't a significant predictor (B = 0)
#                     HA: The Budget is a significant preditor (B != 0)
# Test Statistic = (b1 - B)/(SE(b1)) = (0.3401 - 0)/(0.2443) = 1.392, p-value = 0.20. If we chose a significance
# level of 0.05 then based on the acquired p-value we can't reject the null hypothesis and conclude that the
# Budget isn't a significant preditor of the response because the p-value is higher than our chosen significance
# level.

# 7. Compute and interpret a confidence interval for the slope of the Budget predictor.
# ANSWER: Confidence Interval(CI) = B0 +/- ME, ME = T/Z score * SE
# Confidence Interval = 0.3401 +/- 1.984 * 0.2443 = [-0.1446, 0.8248]
# The above acquired confidence interval indicates that the slope of the USGross in relation to the Budget will
# increase between -0.1446 to 0.8248 million/thousand/hundred/ for each additional Budget with a 95% confidence
# interval.
```

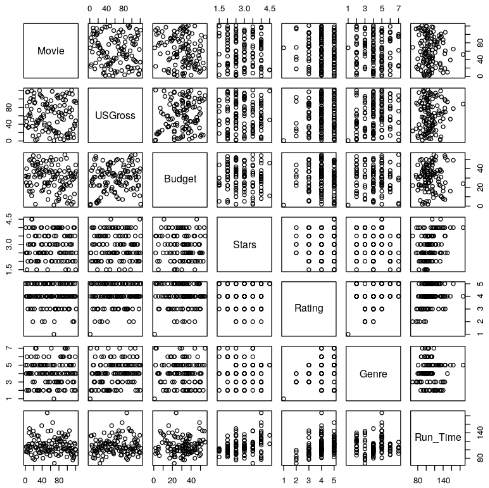
SOFTWARE OUTPUT



A matrix: 4 × 4 of type dbl

	usgross	budget	stars	run_time
usgross	1.00000000	0.159998004	0.054429733	0.1582420
budget	0.15999800	1.000000000	-0.007416402	0.1157532
stars	0.05442973	-0.007416402	1.000000000	0.4504271
run_time	0.15824198	0.115753207	0.450427148	1.0000000

```
Call:
lm(formula = usgross ~ budget + stars + run_time, data = movdat)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.434 -27.286  -0.981  27.885  70.738 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  34.1445    20.4458   1.670   0.0976 .
budget        0.3401     0.2443   1.392   0.1666  
stars        -1.0589     4.4237  -0.239   0.8112  
run_time      0.1841     0.1744   1.055   0.2935  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.75 on 116 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.02721,	Adjusted R-squared:  0.002048 
F-statistic: 1.081 on 3 and 116 DF,  p-value: 0.3599
```


Movie
USGross
Budget
Stars
Rating
Genre
Run_Time