The Effect of Lead Actor's Gender and Movie Title on Gross Earnings

Group 4 - Ronald Lee, Lawrence Jiang, Adam Weintraut, Victor Ramirez

April 4, 2022

Contents

1. Introduction

1.1 Motivation

Movies have been in existence since 1888. Since then, watching movies has been one of our favorite pastimes. Movies are a form of blending visual and sound communication to tell a vivid story.

The movie entertainment industry is a high profile multi-billion-dollar global industry. The movie industry has exponentially grown over the past decades. With this growth there has come a windfall of revenue generating profits. As technology evolves and enhances the viewing experience, you can watch movies from various devices. People can enjoy movies in the comfort of their homes or while traveling.

The ability to predict movie revenue can be a very insightful opportunity. With the predicted revenue information, movie makers can intelligently plan their movie budget. Movie budget line items including star salaries, production, and distribution costs can be smartly negotiated and set.

1.2 Research Question

The success or failure of a movie depends on a variety of different factors: star cast, cast gender, budget, and title. With the breadth of data available today, making accurate revenue predictions is extremely difficult. However, we do have many data science tools and methodologies at our disposal to help in attempting to make movie revenue predictions.

Our research question is:

How the Lead actor's gender affect a movie's gross revenue?

In this study we will explore the relationship between the following features, cast gender, budget, runtime, and title. First, to harness the power of the regression testing harness we engineered three linear regression models to predict the movie revenue. Second, we engineer the model using various data features. We iterated over different features including the following extracted features, cast gender, budget, runtime, and title. Third, we collected various types of datasets from different freely open sources. We then joined and sanitized the data for inspections and prediction modeling.

2. Description of the Data and Research Design

We use two datasets for our analysis:

IMDb dataset (url): For this dataset, the data that we are interested in are the movie title, budget, and gross. Budget and Gross are in US dollar.

Movie roles by gender (url): This dataset has the gender info for movie actors. The source has instructions on how to execute R code to parse the original data file, which has data in JSON format, and convert it into CSV format that can be imported to R studio. The values of gender is either a value of 2 for "Male", or 1 for "Female".

With this two datasets, we then perform an inner-join on them by their movie lead actor name to generate a single dataset that has all the movie info plus lead actor's gender info.

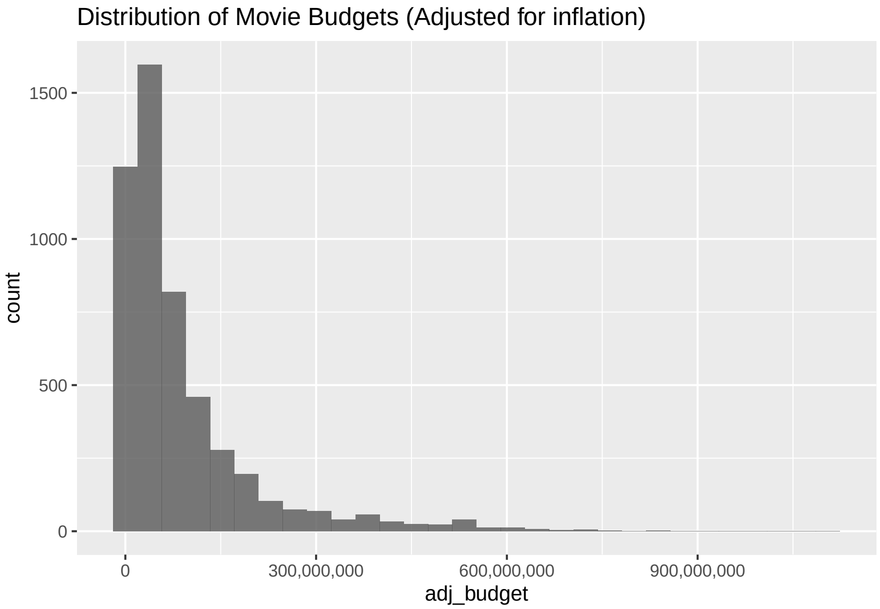
Before we perform our analysis, we also clean up the data by removing all the entries that have empty values for the fields that we are interested (budget, gross, runtime). The final dataset has over 5000 rows of movie data.

In addition, we also generate a "max_title_similarity" score for how similar a movie title to the rest of the movies in the dataset. The score is a value between 0 and 1, where 1 is the highest similarity. We use TFIDF Vectorizer to convert titles to numbers. Then we use Cosine similarity as a measurement of title similarity. It does not account for semantic information.

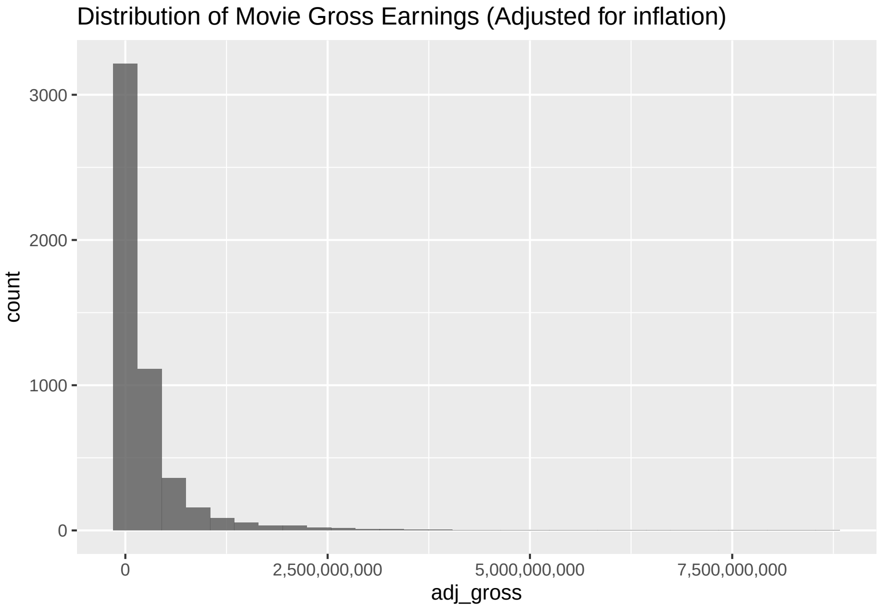
Finally, we uses a R package called 'quantmod' to retrieve CPI data for inflation adjustment for our Gross and Budget data. Then we perform adjustment to the Gross and Budget data in our main dataset. We use 1980 as the base year for our inflation adjustment, then perform inner-join with the main dataset by the years. Additional calculations to adjust Gross and Budget are performed afterwards and added as columns to the main dataset.

2a. Data Characteristics

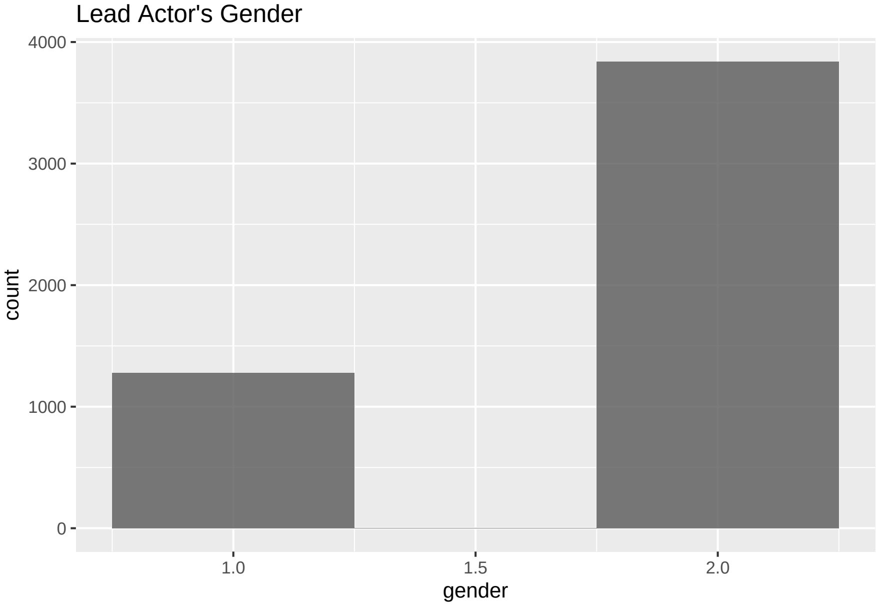
In this section we take a look at the data characteristics for our variables that we are interested and see if there is any need to perform cleanup or transformation before we build our models.



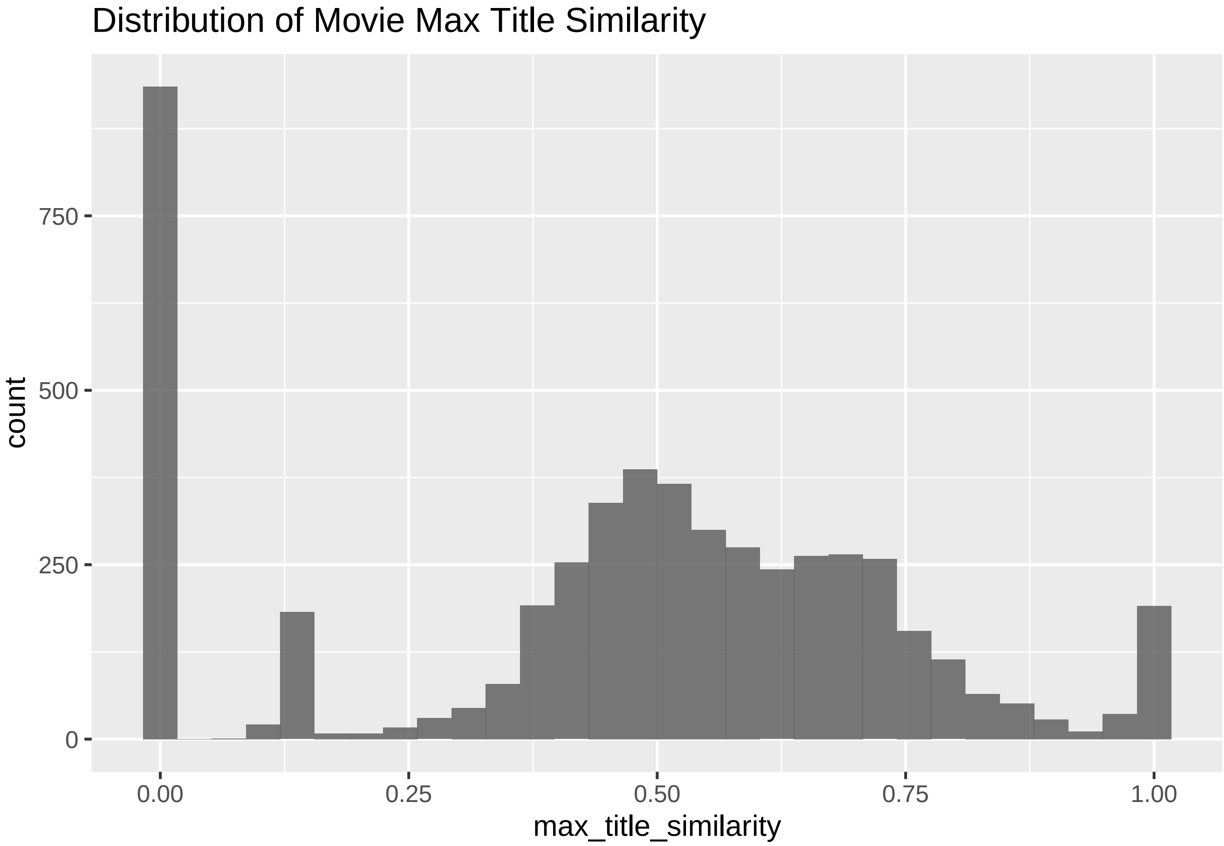
Movie Budget is highly skewed to the left which means most movies have relatively low budget, and there are only a few movies that have really high budget.



Movie Gross Earnings has similar characteristics as the budget, also highly skewed to the low end.



In this histogram that shows the number of Male versus Female Lead Actor. The value “2.0” represent ‘Male’, and you can see that there are much more Male lead actors in our movie dataset.



The Max Title Similarity score has 3 spikes. The biggest one is at the 0 score, which means that most of the movies have title name that are not similar to the others at all. There are another two small spikes at around 0.125 and 1. The spike at 1 indicates that there are movies that are very similar to some other movies in the dataset. The data near the middle of the range has a bell shape similar to a normal distribution.

3. Modeling

3.1 Base Model: Female indicator and Gross Earnings

In order to answer our research question, we build our first model with 'Gross earnings' as our outcome variable and 'Female Indicator' as our feature variable. A female indicator is a binary flag which has the value of 0 or 1. When it is '1', it means that the lead actor of that movie is a female. We generate this indicator by processing the 'gender' data and add a new column in the dataset to store the resulting indicator. We also include the 'Budget' variable since it has strong correlation with Gross, also help setting up the baseline for R-square.
Moreover, since the Gross earning and Budget are highly skewed, we also apply Log transformation to them in our first model.

Log(Gross Adjusted for Inflation) = Female_indicator + Log(Budget Adjusted for Inflation)

3.2 Second Model: Adding Genre Indicators

For the second model, we add explanatory variables that reduce effect caused by the female_indicator variable. We suspect that genre might have correlation with the gender of the lead actor. By introducing genre indicators, they should reduce female_indicator's effect on gross earnings.

Log(Gross Adjusted for Inflation) = Female_indicator + Log(Budget Adjusted for Inflation) + Is_Genre_Action + Is_Genre_Comedy + Is_Genre_Drama + Is_Genre_Adventure + Is_Genre_Biography + Is_Genre_Animation + Is_Genre_Horror

3.3 Third Model: Adding Title Similarity and Movie Duration Indicator

We are also interested to find out if the movie title and movie duration have effects on gross earning. We suspect that title similarity and movie duration might affect the gross earnings. Therefore, we added this "max_title_similarity" score as a control variable, along with another indicator "is_longer_than_2hrs" to our third model to try improve the R-square of the model.

Log(Gross Adjusted for Inflation) = Female_indicator + Log(Budget Adjusted for Inflation) + Is_Genre_Action + Is_Genre_Comedy + Is_Genre_Drama + Is_Genre_Adventure + Is_Genre_Biography + Is_Genre_Animation + Is_Genre_Horror + Max_Title_Similarity + Is_Longer_Than_2hrs

3.3 Robust Standard Errors

Our dataset is not IID, which will be explained in the "Limitations of your Model" section. Therefore, our models are built based on robust standard errors.

4. Results

Table 1:

	Dependent variable:		
	log(adj_gross)		
	(1)	(2)	(3)
female_indicator	0.124*** (0.047)	0.123*** (0.047)	0.147*** (0.047)
log(adj_budget)	1.026*** (0.020)	1.009*** (0.021)	0.969*** (0.022)
is_genre_action		0.415*** (0.078)	0.398*** (0.077)
is_genre_comedy		0.341*** (0.078)	0.375*** (0.078)
is_genre_drama		0.122 (0.090)	0.096 (0.088)
is_genre_adventure		0.271** (0.105)	0.279*** (0.104)
is_genre_biography		0.218** (0.104)	0.152 (0.104)
is_genre_animation		0.804*** (0.103)	0.882*** (0.104)
is_genre_horror		1.014*** (0.132)	0.959*** (0.129)
max_title_similarity			0.676*** (0.069)
is_longer_than_2hrs			0.345*** (0.050)
Constant	−0.064 (0.360)	−0.101 (0.375)	0.197 (0.377)
Observations	5,118	5,118	5,118
R^2	0.493	0.506	0.519
Adjusted R^2	0.493	0.505	0.518
Residual Std. Error	1.404 (df = 5115)	1.387 (df = 5108)	1.369 (df = 5106)
F Statistic	2,485.789*** (df = 2; 5115)	581.469*** (df = 9; 5108)	500.923*** (df = 11; 5106)
Note:			*p<0.1; **p<0.05; ***p<0.01

In the first model, we see that the p-value of the “Female_indictor” is less than 0.01 which means there is significant evidence that female gender has effects on “Log(Adj Gross)”. The coefficient is a positive value, meaning that a movie with a female lead actor actually brings more gross earnings than a movie with a male lead actor. The coefficient value of 0.124 means that it will be 12% more on gross if a movie uses a female lead actor. The adjusted R-squared is 0.493, mainly because the variable ‘Budget’ we added to this model can explain almost half of the model.

After introducing the genre indicators into our second model, most of them have a p-value less that 0.01, except for “is_genre_drama”. Coefficients for both female_indicator and budget have been reduced, but remain significant. Adjusted R-square of the model has also improved to 0.505.

For our third model, the variable “max_title_similarity” and “is_longer_than_2hrs” both have low p-values, showing evidence that they also have effects on the gross earnings. P-values for female indicator remains significant, and its coefficient has increased. Adjusted R-square of the model has also improved to 0.518.

5. Limitations of your Model

5.1 Large-Sample Assumptions

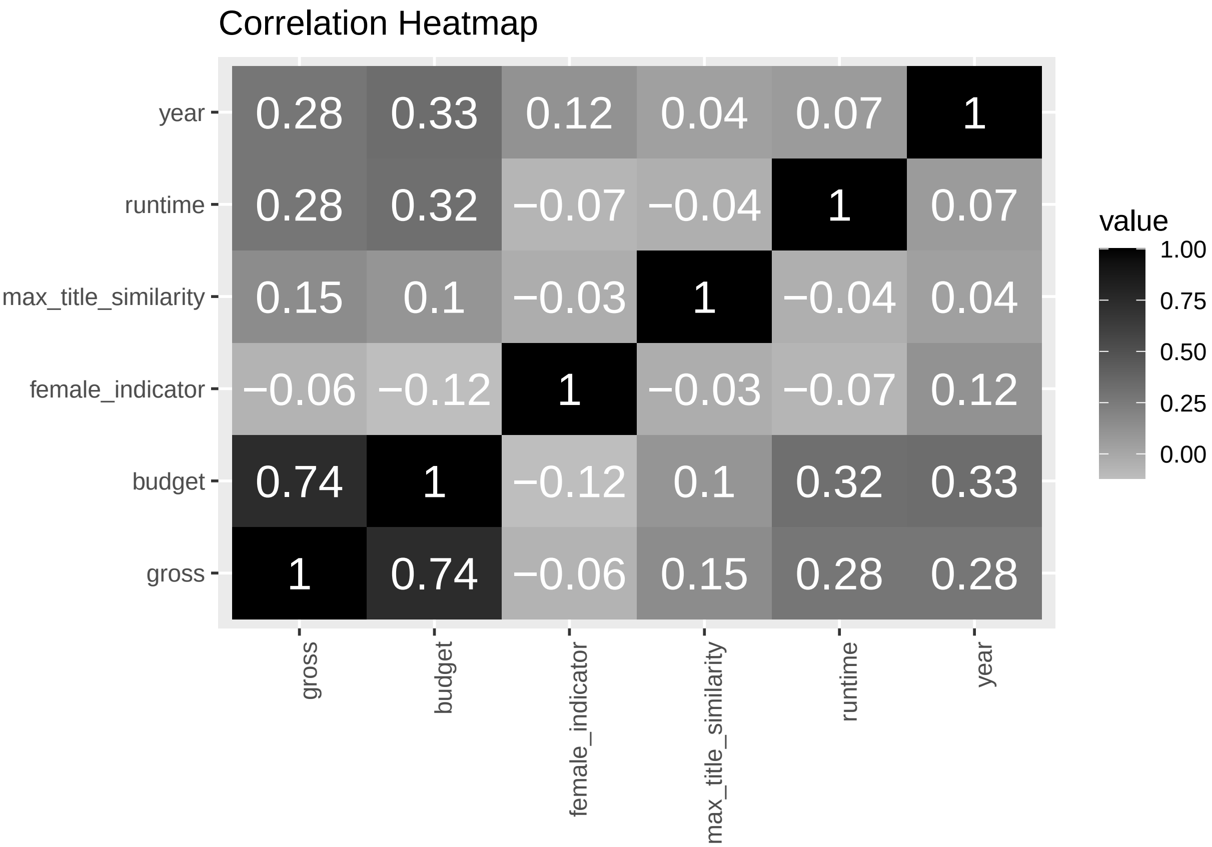
Since our analysis is based on a dataset with more than 5000 entries, we can use the large-sample assumptions.

5.1.1 Independent and Identically Distributed (I.I.D.)

Our dataset doesn’t satisfy the IID assumption. Our dataset has movies between year 1980 and 2020 scraped from IMDb. For movies that are actually in the same series, like ‘Star Wars’ or “Mission Impossible”, their data will not be independent. Since our dataset is also a time series data, there might be dependencies between data in different timeline. Our analysis is based on robust standard errors which can adjust for dependency issues in our dataset.

5.1.2 Unique BLP Exists

We want to make sure that no variables are perfectly collinear with any one of the other variables. From the correlation heatmap below, we can see that perfect collinearity only happens at the diagonal of the matrix which has all 1’s and no where else. Also, while building our model, R didn’t drop any variables considered to cause perfect collinearity with any other variables.


Correlation Heatmap
year
runtime
max_title_similarity
female_indicator
budget
gross
0.28 0.33 0.12 0.04 0.07 1
0.28 0.32 −0.07 −0.04 1 0.07
0.15 0.1 −0.03 1 −0.04 0.04
−0.06 −0.12 1 −0.03 −0.07 0.12
0.74 1 −0.12 0.1 0.32 0.33
1 0.74 −0.06 0.15 0.28 0.28
gross
budget
female_indicator
max_title_similarity
runtime
year
value
1.00
0.75
0.50
0.25
0.00

5.2 Omitted Variables

5.2.1 Intentional Omitted Variables

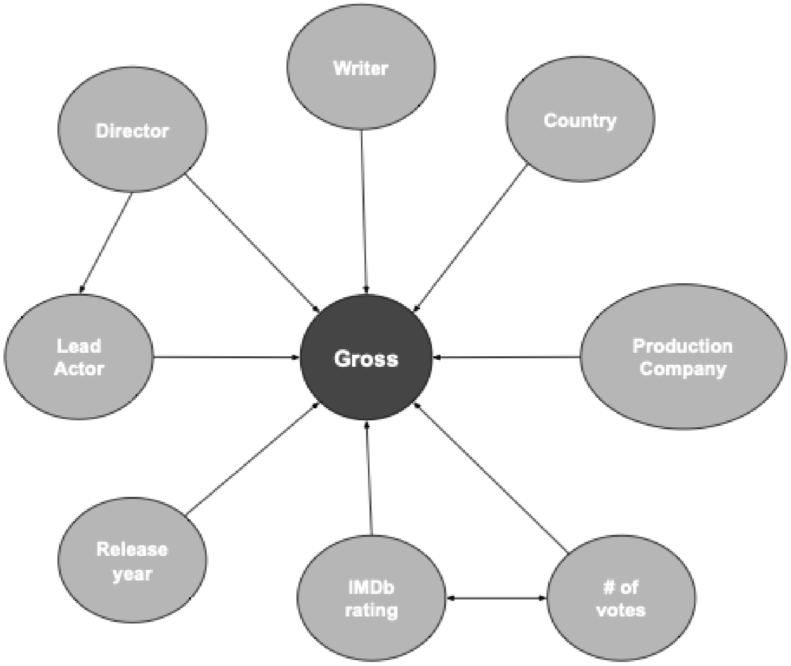


Figure 1: Intentional Omitted Variables

The variables that are available in our dataset but are omitted in building our models are:

- Director
- Writer
- Production Company
- Lead Actor
- Country
- Release year
- IMDb rating
- Number of votes in IMDb

Our EDA shows that these variables have low correlation with Gross earnings, so we simply omit them in our models. For "Release year", since we already adjusted our Gross and Budgets with inflation, so we didn't include year as a variable that could affect the model outcomes.

5.2.2 Unintentional Omitted Variables

There are other variables not available in our dataset that might have causal effects on a movie's gross earnings.

- Film distribution platforms: We don't have data to tell us what platforms (theaters, streaming, etc) the movies were released on. If a movie was released to more platforms, there are more chance to make money from different audience.
- In theater duration: We don't have data to tell us how long the movies have been on the distributon platforms. The longer a movie is on, the more time it can make more money.
- Story of the movie: How good is the story is subjective and hard to quantify. We don't have data on that but it could be a big factor on gross. A good movie could mean that people are willing to pay to watch couple more times, or purchase it and put it into their collection.
- Acting skills of the cast: Acting skills could be subjective and hard to quantify, but could affect how people willing to pay for the movies.
- World population: Population grows over the years, and more population means more people paying for movies.

6. Conclusion

In conclusion, our analysis shows strong evidence that a female lead actor brings more gross revenue to a movie than a male lead actor. In this case, it is about 14% more on average. The "Max Title Similarity" score also seems to have positive effects on the Gross. The score value is between 0 and 1. When it is 1 (high similarity), it will improve gross earnings by a massive 67%. The indicator on whether a movie is longer than 2 hours also shows a significant effect on gross, bringing in 34% more gross if is longer than 2 hours.