

---

title: "Pixar Movies Analysis"  
author: "Sofia Martinez"  
date: '2022-03-17'  
output: html\_document

---

## The Business Task

Pixar Animation Studios wants to focus on how they can generate more income in the next 10 years with their future films. Will increasing a budget lead to more money returned? Does focusing on the rotten tomato score increase the income? The job is to look at Pixar's dataset and analyze any trends to determine what Pixar should focus on to generate more revenue.

The cleaning and manipulation has been done on Google Sheets, so its best to check the change log on what was changed. To see the change long, click on the link. [https://docs.google.com/document/d/189LYblb\\_zOnhuLS13PpFrIXBEQ1xGcNSrzPuHac\\_o1k/edit?usp=sharing](https://docs.google.com/document/d/189LYblb_zOnhuLS13PpFrIXBEQ1xGcNSrzPuHac_o1k/edit?usp=sharing)

## Loading in the data

First, we need to install & load in the tidyverse package in order to use its functions.

```
install.packages("tidyverse")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)

library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.1 --
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.6      v dplyr   1.0.8
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Next, we need to create a dataframe so we can use the Pixar dataset. The Pixar data is from by Sandip Devre on Kaggle, available on an Open Database License.

```
pixar_movies <- read.csv("pixar_data.csv")
```

## Manipulating the data

In order to be able to make some comparisons, we need to manipulate the data first. To find some comparisons, finding the max budget, max income, max money return, max rotten tomatoes score, and max north america % on income. We don't need to find the minimum income and minimum money return because it easily can be identified to the movie Onward, which released in theaters a few days before the COVID-19 Pandemic lock down.

```
max_budget <- max(pixar_movies$budget_million)
max_income <- max(pixar_movies$worldwide_million)
max_money_return <- max(pixar_movies$worldwide_million)
max_northamerica <- max(pixar_movies$north_america_.on_income)
max_rotten <- max(pixar_movies$rotten_tomatoes_percent)
max_budget
```

```
## [1] 2000
```

```
max_income
```

```
## [1] 12428
```

```
max_money_return
```

```
## [1] 12428
```

```
max_northamerica
```

```
## [1] 0.53
```

After manipulating the data, we can see some comparisons. We see that the max budget of 2000 (million) belongs to many movies, but it belongs to *The Incredibles*, who also had the max income and the max money return (which is the worldwide income subtracted by the budget). However, the maximum North American income percent was 53% belonging to the movie *Cars*. Is there a relationship? To see this clearly, we need to graph. We take the North American income percent because Pixar's headquarters and animation studios is in North America.

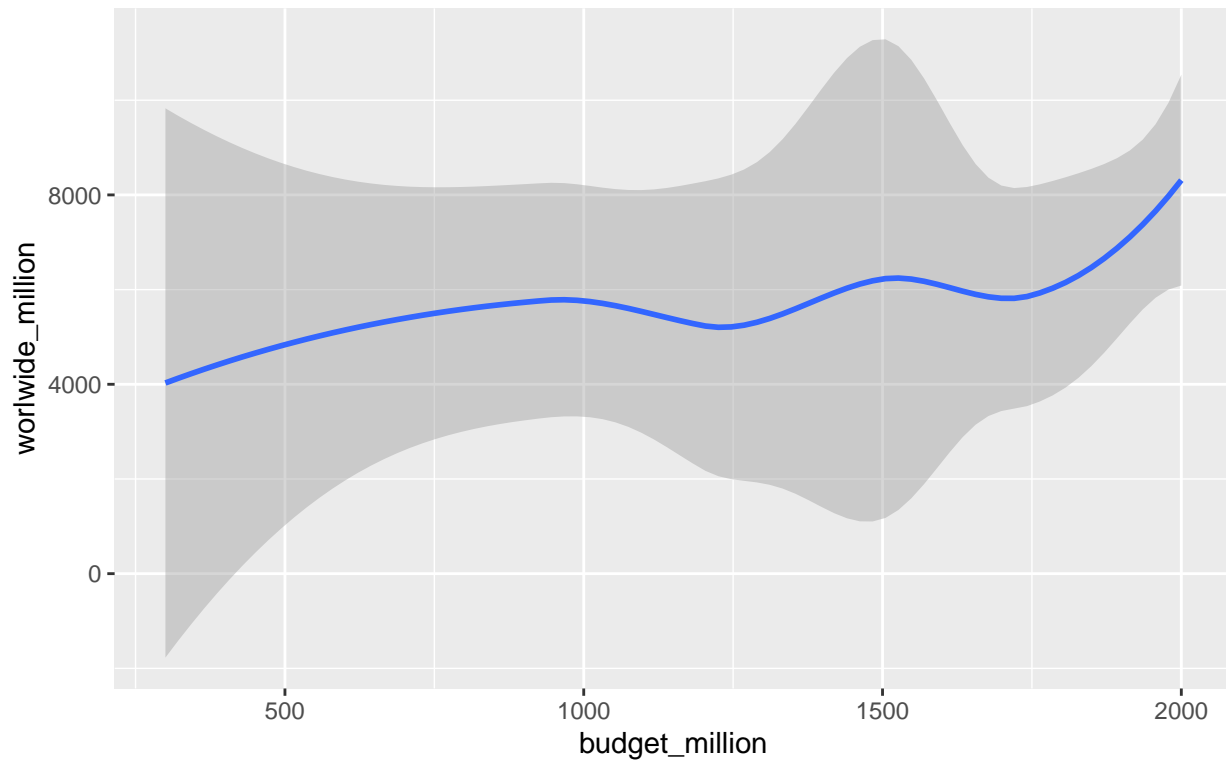
## Graphing Relationships

First, we should graph the relationship between the budget and the income worldwide. We should save this to a value so its easier to access. Because there are only 50 movies, its best to use the `geom_smooth` to better detect a relationship.

```
budget_x_worldwide <- ggplot(data= pixar_movies) + geom_smooth(mapping = aes(x= budget_million, y = wor  
budget_x_worldwide
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

## Budget and the Worldwide income for Pixar Movies



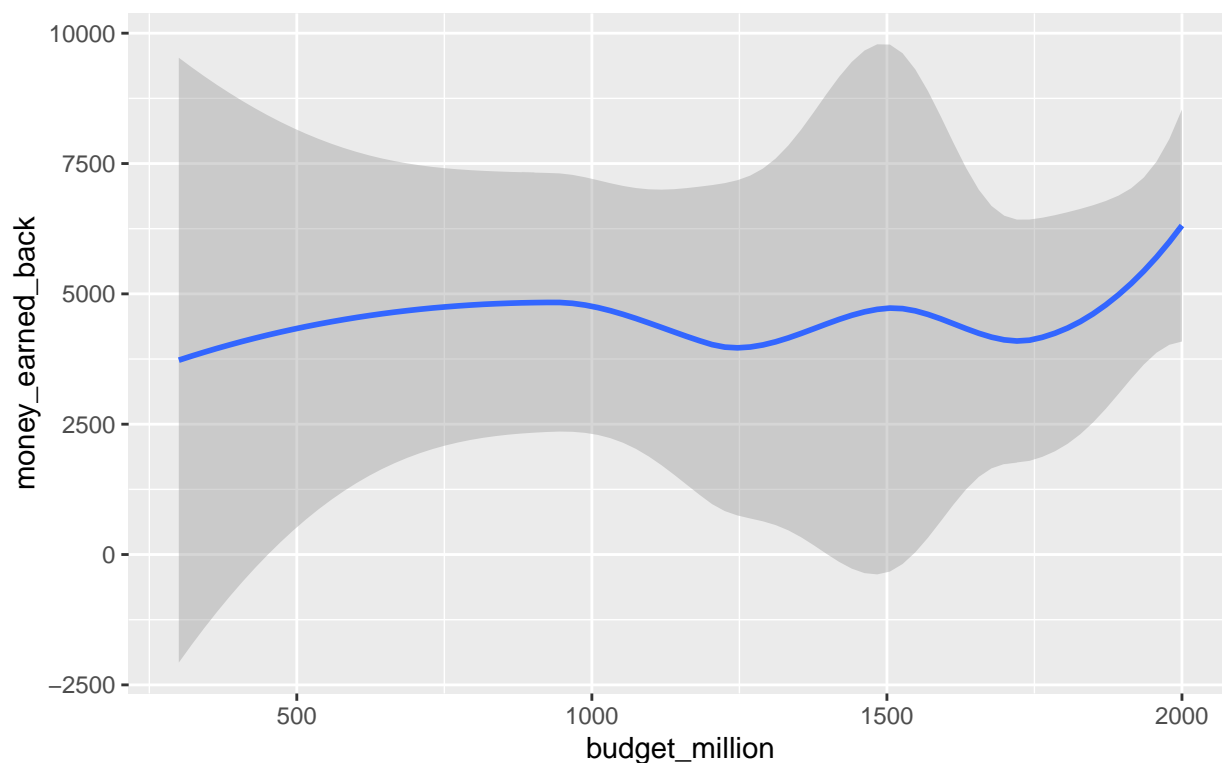
Data provided by Sandip Devre on Kaggle

As we can see, there is a moderate positive relationship between the budget and the worldwide income. However, it is not strong, so we should look at other relationships. Lets now look at the relationship between the budget and the total money earned(or the money they got back).

```
budget_x_moneyback <- ggplot(data= pixar_movies) + geom_smooth(mapping = aes(x= budget_million, y = moneyback))
budget_x_moneyback
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

## Budget and the Money earned back for Pixar Movies



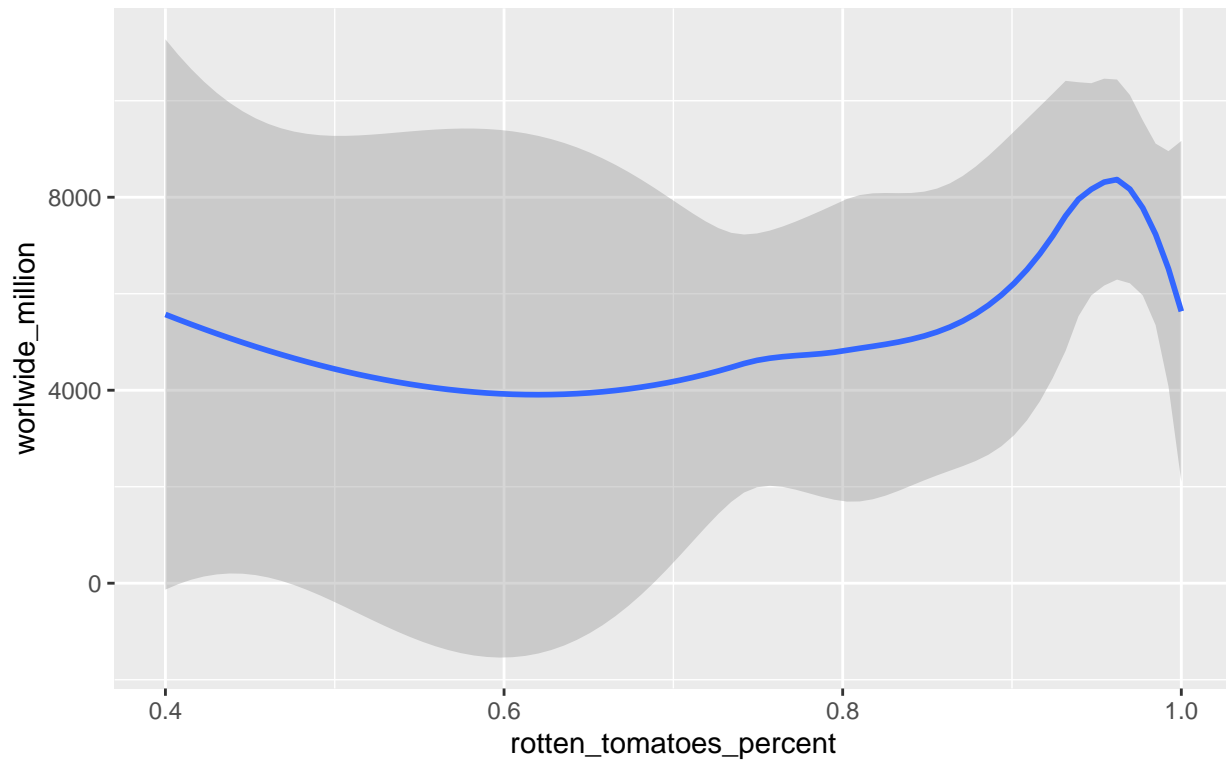
Data provided by Sandip Devre on Kaggle

The relationship is similar to the budget and worldwide income. Lets now look at a relationship that doesn't include the budget, which is the Rotten Tomatoes percent score and the worldwide income.

```
rotten_tomatoes_x_worldwide <- ggplot(data= pixar_movies) + geom_smooth(mapping = aes(x= rotten_tomatoes_x_worldwide))
rotten_tomatoes_x_worldwide
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

## Rotten Tomatoes % Score and the Worldwide Income

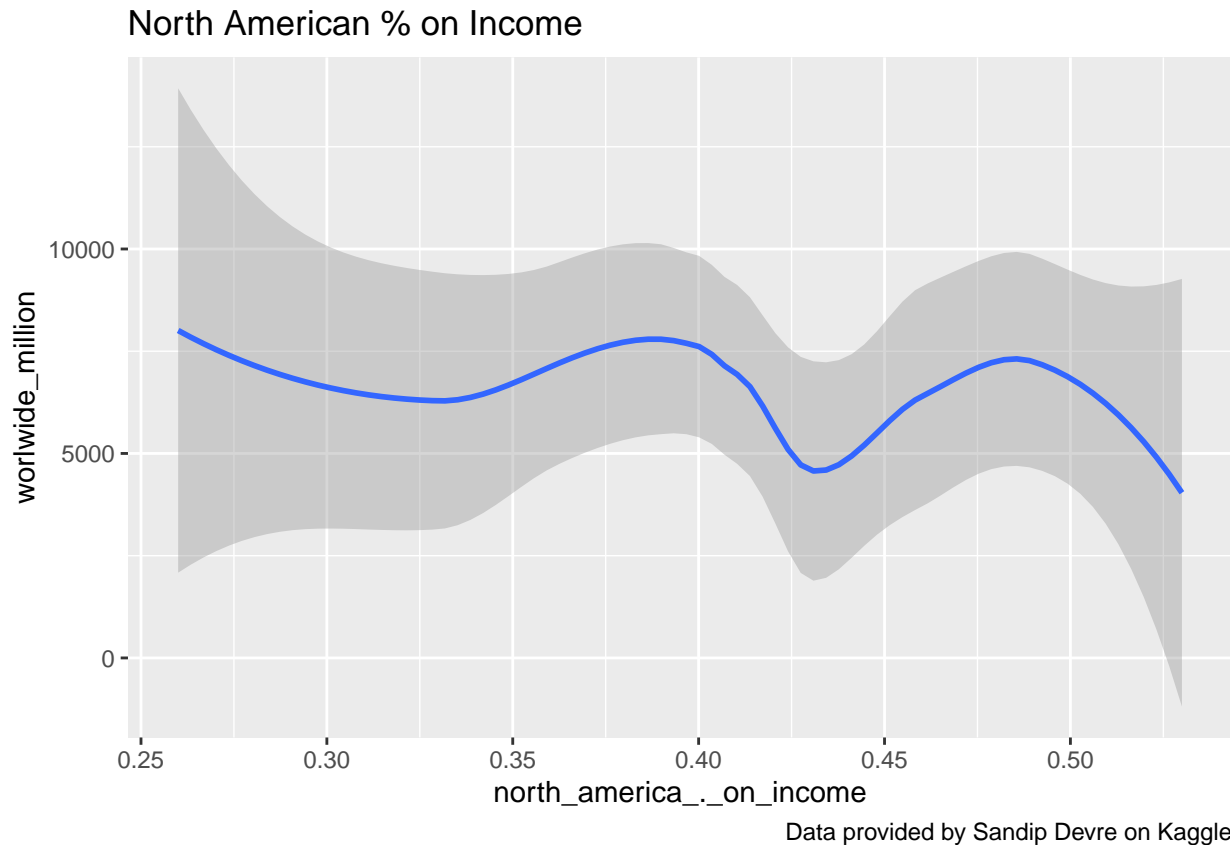


Data provided by Sandip Devre on Kaggle

When looking at the graph, I don't really see a steady relationship between the two variables, so its safe to not make any conclusions based on this relationship. Finally, I want to look at the north american % on worldwide income to the worldwide income, since Pixar makes its movies in North America. Lets check it out.

```
north_america_.on_income_x_worldwide<- ggplot(data= pixar_movies) + geom_smooth(mapping = aes(x= north_america_.on_income_x_worldwide
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



With this graph, we can see a steady negative relationship between the percent on North American income and the worldwide income. It's safe to make a estimate that as the percent on North America to worldwide income increased, the worldwide income decreased.

### Conclusion on Analysis

After looking at some graphs, some conclusions can be made. A higher budget can lead to a higher income/money earned back, and a less % of north american income on worldwide income and lead to a greater worldwide income. With this information, Pixar can focus on having a higher budget to earn more money, or focus on advertising to other places worldwide to increase the % on worldwide income for other continents rather than North America.