Visualizing Indie-Film:
Audience Prediction and the Effect of the Genre Animation

Jongwon Lee
I590 - Data Visualization
Professor Yong-Yeol (yy) Ahn
December 16, 2019

Abstract

Film audience prediction even with the advent of machine learning and the viability of data collection had been extremely difficult. This is especially true during the production process or before the release date. (Dey, 2016) However, indie-film's situation may be different. This is because indie-film's distinctiveness contributes to it, having less kind of responsible variables on the number of audience, in comparison to the blockbuster films. More specifically, this project focuses on indie-films that are imported to South Korea. Another reason that makes audience prediction viable is because it's release date in Korea is usually a while after the initial release from the original country. This is mainly due to the fact that importers can make choice on importing films according to its expected audiences. This project further investigates the regression result of the indie-film by visualizing it. It focuses on how a genre "Animation" can affect the number of audience which is this project's dependant variable and other independent variables as well. It tries various ways of visualizing multivariate regression data and discuss the pros and cons of each method. Ultimately, this project suggests how the decision makers in South Korea (eg. importer and distributors) should conceive each variable. Codes used to analyze and visualize are available in this link.

Introduction

This project explores a dataset that contains the size of audiences of indie-films imported to South Korea and their metadata. It is based on a previous project (Lee, 2016) that found that the review score (tomatometer), which is mainly assessed by western movie critiques, and index that represents movie maniac's opinion rather than the general audiences were not significant for predicting the success of films. Therefore, the distributors should not focus on these data on decision-making. Also, the previous project confirmed that

animation was the most significant genre that positively affects the dependent variable (Lee, 2016). This project will delve more deeply how a film being an animation means in relation to other independent variables and the dependent variable. This project compares simple color combination and alpha to derive an optimal visualization. This project will stick to visualizing all the variables in a plot without dimensionality reduction.

Motivation

Although regression or even deep learning can derive a model that explains multiple variables, numbers sometimes hide valuable insights. By visualizing the model, we not only can convince the reader of an article persuasively but also detect a hidden insight in the numbers.

Related Works

Merwe achieved 52% accuracy predicting movie box office with variables such as actor, director, release month, genre (Merwe, 2013). These variables are traditional and standard approach to this problem. These factors definitely would have significant effect on the box office logically, however, nowadays it is treated as inefficient and indirect variables. Goodman tried to decide which variable among critic score, audience score, US Gross, US Opening, IMDB Rating best defines the success of a film (Goodman, 2013). This suggests simply declaring the number of audience as a measure for success can be problematic. However, this project defines a film's success to gathering enough number of audiences to cover the cost. Zhou took the poster image, extracted features and utilized them to predict the movie revenue. Posters were classified either negative and positive and tried to utilize it for revenue prediction (Zhou, 2017). This idea could be considered in this project as well in the future. However, since the goal of this project is not about researching poster's effect, for now it is not considered.

Urpa suggested results derived from dimension reducing visualizations such as distnet and focusedMDS have real biological insights (Urpa, 2019).

Tools

This project uses Python mainly, sci-kit learn, pandas, numpy, statsmodel, matplotlib, seaborn as libraries.

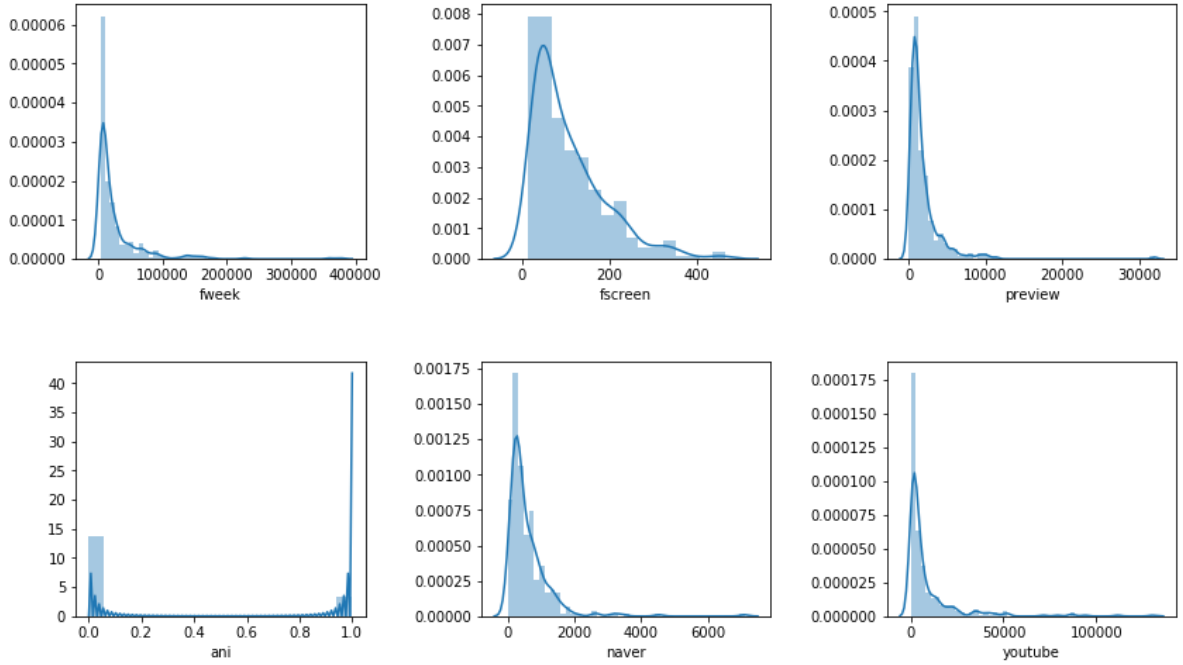
Dataset

In the aforementioned project (Lee, 2016), there were 9 variables considered. Among them, 6 variables (1 independent and 5 dependent) remained and others were dropped. As a result, the independent variables well predict the dependent variables with all P-values 0.

Figure 1

	fweek	fscreen	preview	ani	naver	youtube
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000
mean	27929.993311	107.294314	1888.304348	0.197324	609.103679	9236.173913
std	44286.903827	82.507760	2569.141668	0.398646	684.379388	16884.911619
min	3101.000000	13.000000	14.000000	0.000000	5.000000	35.000000
25%	5985.500000	46.000000	660.000000	0.000000	232.000000	1226.500000
50%	11876.000000	82.000000	1123.000000	0.000000	396.000000	3171.000000
75%	29721.000000	146.500000	2181.000000	0.000000	762.500000	9070.000000
max	377109.000000	465.000000	32040.000000	1.000000	7104.000000	130407.000000

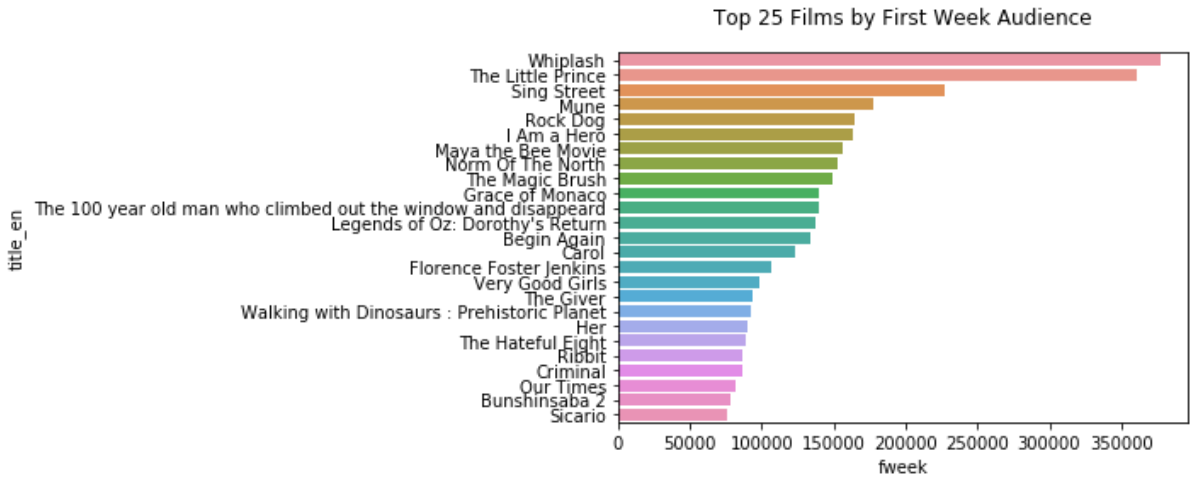
Figure 2



Dataset is collected by getting access to public data or crawling relevant social media. Initially, 962 movies were considered but 300 remained relevant and others were dropped. (Lee, 2016)

Dependent variable is the first week audiences, ‘fweek’. Instead of the total audience, first week audiences were chosen because it helps generalize the problem (Dey, 2016). Figure 2’s fweek shows that its highly skewed. This is because the film's success highly differs by films. This pattern can also be observed in other four independent variables ‘fscreen’, ‘preview’, ‘naver’ and ‘youtube’.

Figure 3



When deciding which variable to regress on for ‘predicting audiences’, an analyst’s insight should be highly valued, rather than putting in any available variables unlike other problems (Assady et al, 2013). Data used in this project were also selected in this manner rather than regressing over all the available data.

Independent variable 1 is the number of screens of the first week, ‘fscreen’. It is decided according to the distributors, somewhat arbitrarily. Opened screens may not be full with the audiences, so it does not conform to ‘fweek’.

Independent variable 2 is the number of audiences of previews, ‘preview’. It is also decided according to the distributors. Movie maniacs (audiences that are more interested in this movies in general) attend these previews and affect general audiences in a positive or negative way. Thus, it may be neutral or slightly above neutral according to this assumption. However, this project concludes larger preview leads to higher first week audience turnout.

Independent variable 3 is a dummy variable that is 1 if the genre is animation, else 0. Among other genre, animation h ad the most significant effect on the audiences, other variables fixed (Lee, 2016); thus chosen as a dummy variable. It is our only categorical data.

Dependent variable and independent variable 1,2,3 are retrieved from Korean film council api.

Independent variable 4 is the ‘expectation index’ from the largest portal in South Korea, Naver, ‘naver’. (https://movie.naver.com/) It can only be updated before the release,

thus should highly be responsible for dependent variable. It is a poll available before the release and can be participated by every individual with one account.

Independent variable 5 is the official Korean YouTube trailer's view, 'youtube'.

OLS Regression Result

R-squared is 0.726 which implies the dataset is valid. Each variable's P-value is 0, which means all independent variables are statistically significant at contributing to the dependent variable.

Figure 4

OLS Regression Results

Dep. Variable:	fweek	**R-squared:**	0.726
Model:	OLS	**Adj. R-squared:**	0.721
Method:	Least Squares	**F-statistic:**	155.2
Date:	Mon, 16 Dec 2019	**Prob (F-statistic):**	3.76e-80
Time:	00:40:52	**Log-Likelihood:**	-3429.1
No. Observations:	299	**AIC:**	6870.
Df Residuals:	293	**BIC:**	6892.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.184e+04	2333.384	-9.358	0.000	-2.64e+04	-1.72e+04
fscreen	250.1716	20.572	12.161	0.000	209.684	290.659
preview	4.8656	0.655	7.428	0.000	3.576	6.155
ani	1.482e+04	3497.896	4.237	0.000	7936.231	2.17e+04
naver	9.7015	2.414	4.019	0.000	4.950	14.453
youtube	0.5308	0.092	5.757	0.000	0.349	0.712

Omnibus:	139.859	**Durbin-Watson:**	1.887
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1440.040
Skew:	1.629	**Prob(JB):**	0.00
Kurtosis:	13.246	**Cond. No.**	5.08e+04

OLS Params

```
model.params
```

const	-21836.251855
fscreen	250.171582
preview	4.865602
ani	14820.417220
naver	9.701502
youtube	0.530833

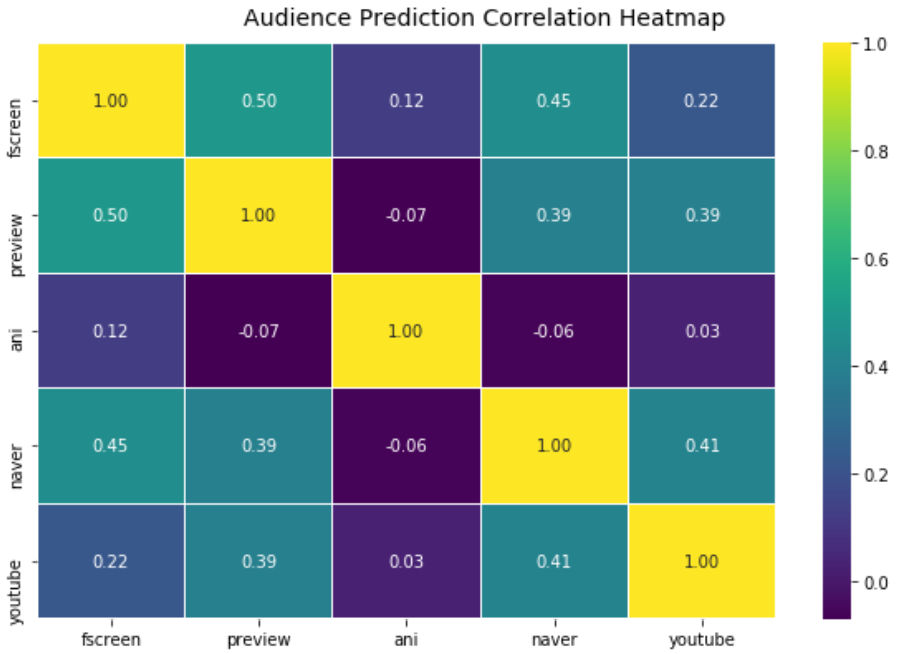
first week audiences
= β0 + β1*(first week screen) + β2*(preview) + β3(animation) + β4(Naver expectation index) + β5(YouTube trailer view)
β0 = -21836.251855, β1 = 250.171582, β2 =4.865602, β3 = 14820.417220, β4 = 9.701502, β5 = 0.530833

1 increase in 'fscreen', first week screen, is associated with 250.17 increase in 'fweek', 'first week audiences'. 1 increase in 'preview', preview audience, is associated with 250.17 increase in 'fweek', 'first week audiences'. 1 increase in 'Naver', Naver expectation index, is associated with 9.70 increase in 'fweek', 'first week audiences'. 1 increase in 'youtube', YouTube trailer view, is associated with 0.53 increase in 'fweek', 'first week audiences'. A film being an 'animation' genre is associated with 14820.41 increase in 'fweek', 'first week audiences'.

Not only other independent variables have meaningful effects, the idea that a film just being an animation increasing the first week audiences by 14820 was intriguing. Thus, this project will put a lot of effort researching further on it by visualizing. Also, this project will look into the relationship between each dependent variables and the situation fixing one variable (eg. fixing the genre to animation) and visualize it. This will reveal what is not observable just by observing OLS result.

OLS Model Heatmap

Figure 5



Heat map is extremely efficient and intuitive method of visualization. Most importantly, all the sets correlation is 0.5 or below, which is acceptable. Moreover, animation has negative relationship between preview and naver. It implies that distributors host less preview for animation and people online shows less interest on animation films on average. However, since we already confirmed that animation films have more audiences, this seems intriguing. It may be dangerous for a distributor if he or she decides to open less screens due to its success at preview or naver index than he or she ideally should. However, since the first week screen and animation is positively correlated, distributors are already making right decisions. This also implies that there is a hidden variable that the distributors consider. They may be opening more screens since it is an animation but their can be another variable which this model lacks.

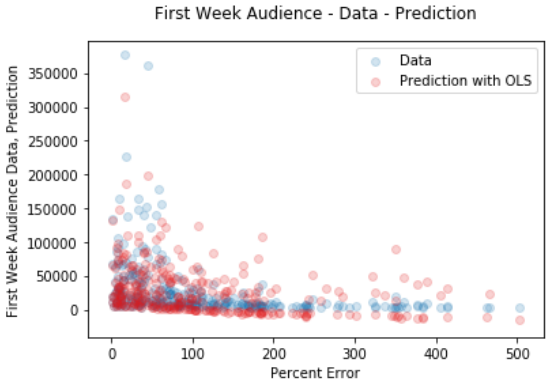
Visualization

This project divides visualization into two parts: 1. color, 2. data analysis.

Visualization - Discussion on Color

1. OLS Result - Colormap Paired, Light Red (Paired 5) / Light Blue (Paired 1) / alpha = 0.2

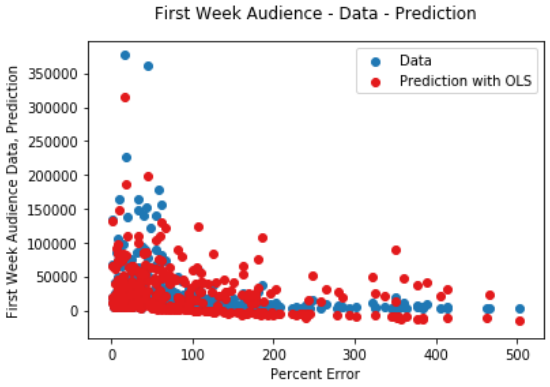
Figure 6



The objective of this visualization is to check if the data points (Prediction and actual Data) genuinely match. Therefore, alpha 0.2, was chosen and since I want the two data to be distinct, chose red and blue. The drawback is it could be hard to observe when printed out, due to its alpha. However, being able to see the overlap should be prioritized.

2. OLS Result - Light Red (Paired 5) / Light Blue (Paired 1) / alpha = 1

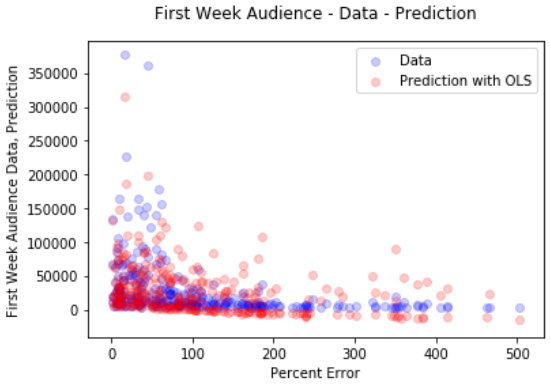
Figure 7



We can still see that data points overlap, however red data points cover a large amount of blue which is not desirable.

3. OLS Result - Non Colormap Red / Blue alpha = 0.2

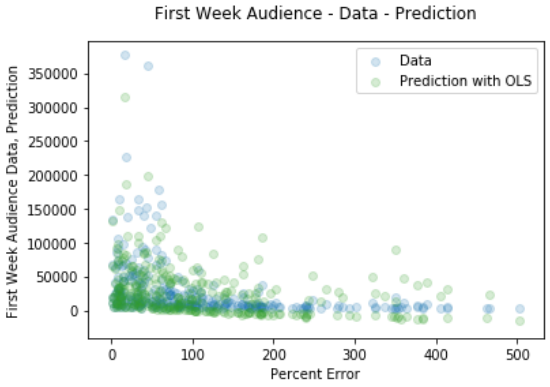
Figure 8



Also, matplotlib's 'Paired' colormap solved the issue with 'red' and 'blue' with 0.2 alpha, which was that it was difficult to distinguish red and blue when printed on paper. Figure 7 is Red and Blue that are not from colormap. When this chart is printed out, red and blue color are difficult to distinguish.

4. OLS Result - Green (Paired 3) / Blue (Paired 1) / alpha = 0.2

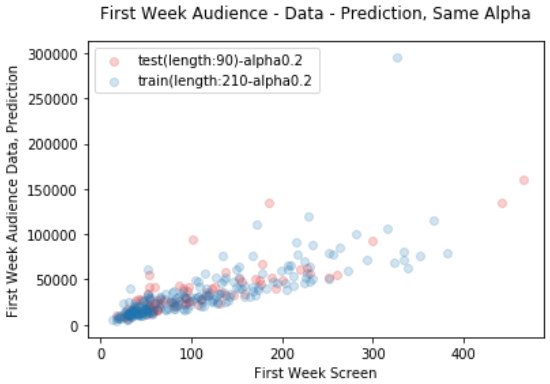
Figure 9



Green, Blue was another color pair that was considered, however, it was not distinct enough as Red and Blue. Thus, this project will use red and blue mostly when the goal is similar to this OLS Prediction / Data comparison. Red and Green was avoided for red-green colorblindness and yellow is hard to see in general.

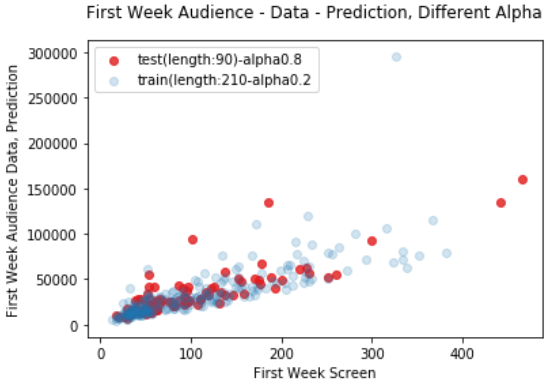
5. OLS Result with 210 Train Dataset - alpha 0.2 and 90 Test Dataset - alpha 0.8

Figure 10



Main model uses all the dataset (length:300) to derive the optimal model. However, we can divide train and test dataset to increase its validity. In this case, since the test dataset has significantly small length than the train dataset, if we use the same alpha, it is difficult to observe the test data.

Figure 11

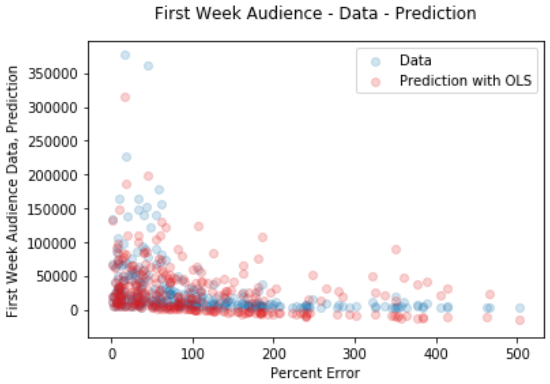


Thus, like Figure11, setting the alpha of the test data to 0.8 and the train data to 0.2 would visualize both dataset and show they overlap well like below.

Visualization - Analysis with 2D pt.1

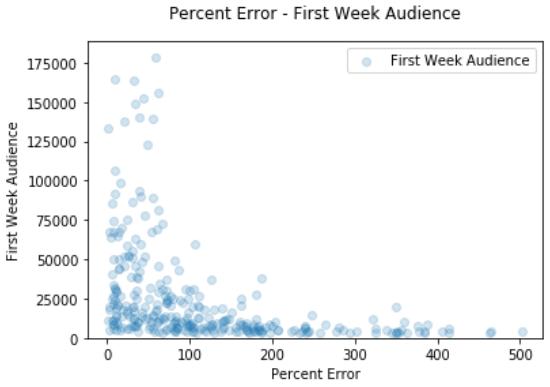
1. OLS Result - Prediction vs Data

Figure 12



This graph shows that Data and Prediction overlap well which visually implies that the OLS regression was successful. However, this also shows that a significant number of points are above 100 Percent Error (Prediction and Data) which is not ideal. To confirm this issue and improve the visualization, only the actual data is plotted again to the Percent Error and ylim was given at the ½ of the maximum y.

Figure 13



Now, we can see that in comparison, not many of the data points are above the 100 percent error. Moreover, we can observe that a successful film tend to have a less percentage error. Therefore, this model's accuracy improves as the expected audience is higher.

2. Pairwise Plots with seaborn

Figure 14

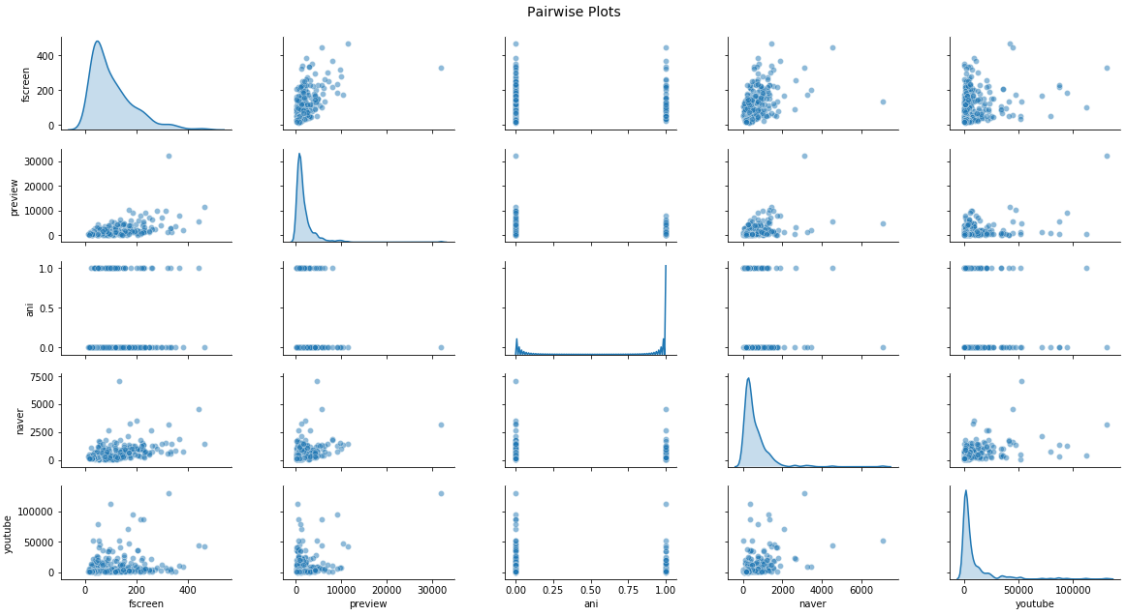
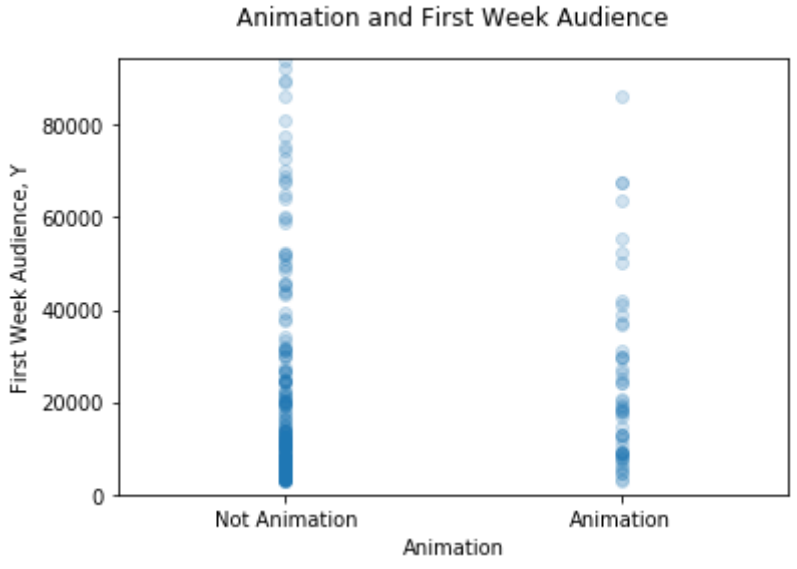


Figure 14 contains more information than heatmap (Figure 5) but unclear due to its size. We can check distribution, however with 300 data points, overlaps of data in a small area makes it difficult to observe meaningful insight. For example, if we see ani-fscreen plot, it is doubtful to state that animation or non-animation film on average have higher fscreen, which was easily seen on an aforementioned heatmap.

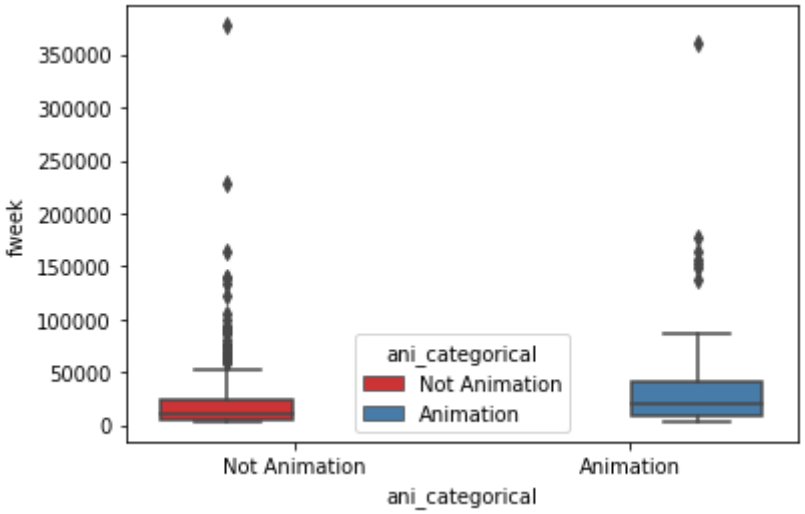
3. Scatter Plot with Dummy Variable / alpha = 0.2, ylim to ¼ of max fweek

Figure 15



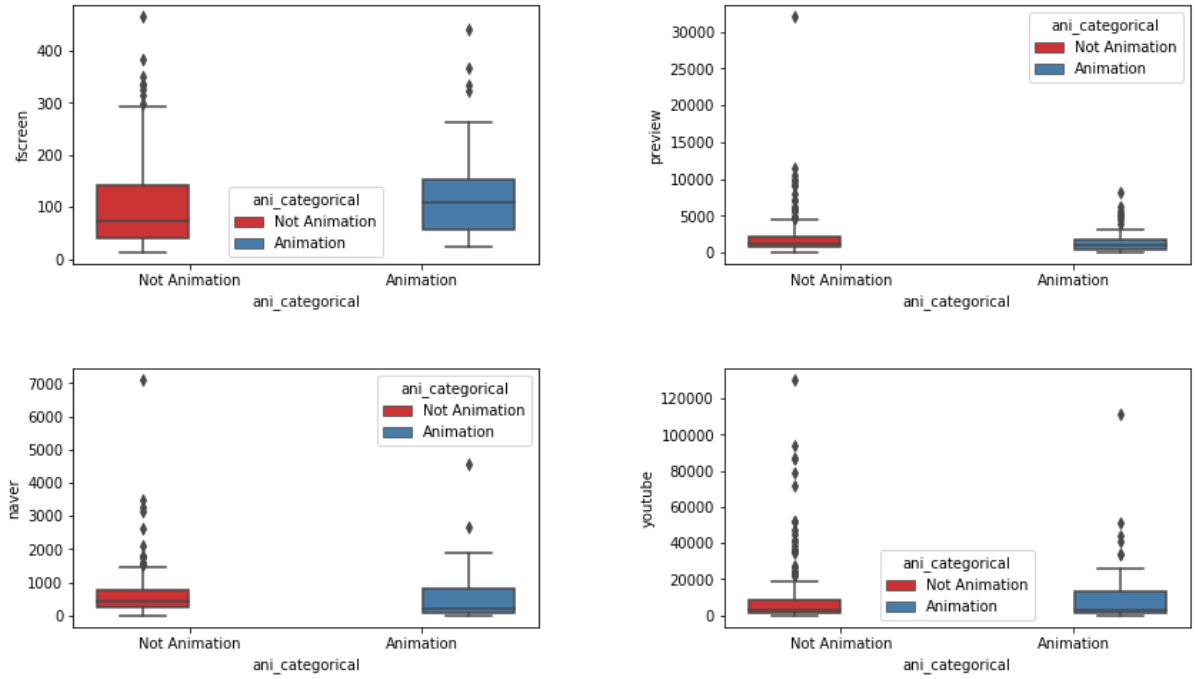
We can improve the above issue with controlling alpha. However, even with controlled alpha, it is not clear that Animation tend to have higher ‘first week audience’ in general.

Figure 16



Boxplot solves this problem that it shows animation tend to have higher ‘first week audience’ than non-animation.

Figure 17



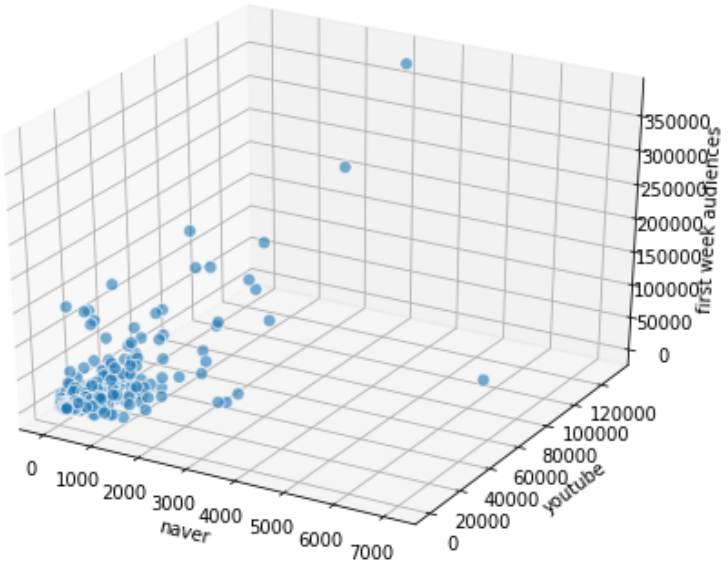
These boxplots are representing dependant variables(‘preview’, ‘naver’, ‘youtube’) other than ‘ani’ showing that except for ‘fscreen’, films tend to have less value when they are ‘animation’. This implies that if a distributor would make a decision with just

‘preview’, ‘naver’, ‘youtube’ data for ‘animation’ genre indie-film, it could be misleading. However, ‘fscreen’ is significantly higher for ‘animation’ than ‘non-animation’. This means distributors are allocating more screens for ‘animation’ because they know ‘animation’ will have more audience even with other factors say it will not be.

Visualization - Analysis with 3D

1. 3 variables in one plot on each axis

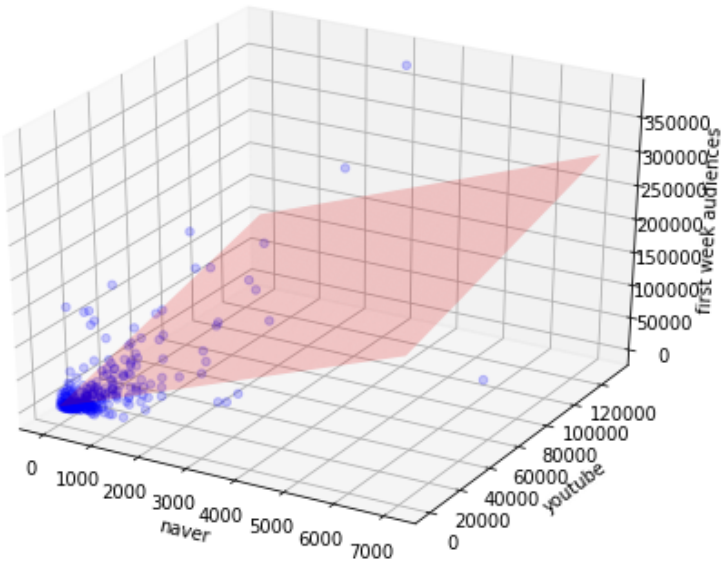
Figure 18



Although this may look cool, this does not derive much information. It is extremely difficult to observe correlation between variables. Most of the data points are highly overlapped.

2. 3 variables in one plot on each axis with a meshgrid

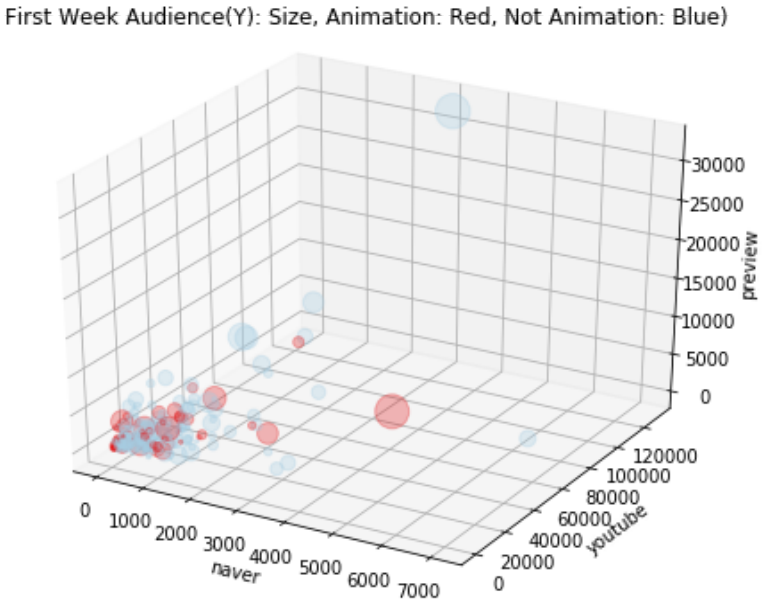
Figure 19



A surface is added that shows all three variables does have correlation. However, this is pretty much all we can say with this visualization.

3. Size and color is added to 3 axis scatter plot

Figure 20



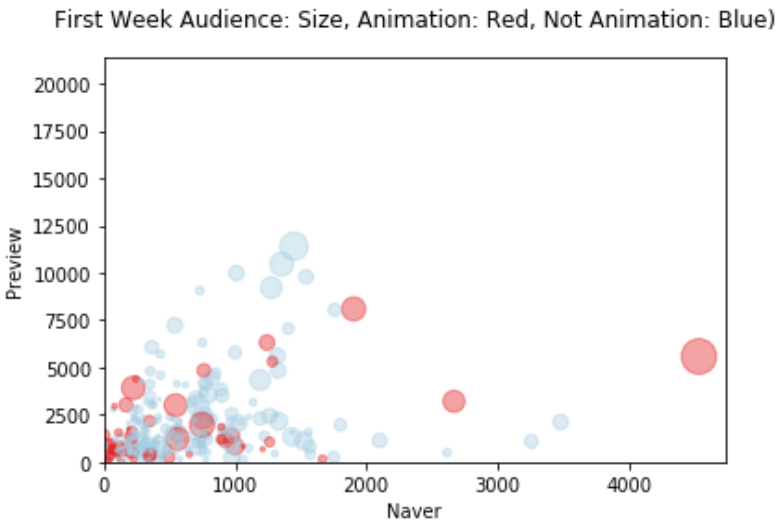
Now this contains some information by adding size and color. We can see that blue dots which are not animation tend to have smaller size on average while red points tend to have larger size. This is because the size contains first week audience, our dependent variable. Also, red dots almost 0 previews on this graph which states

animation films have less previews. However, these were already available with just a heatmap.

Visualization - Analysis with 2D pt.2

1. Naver, Preview, Animation - Color, Audience - Size

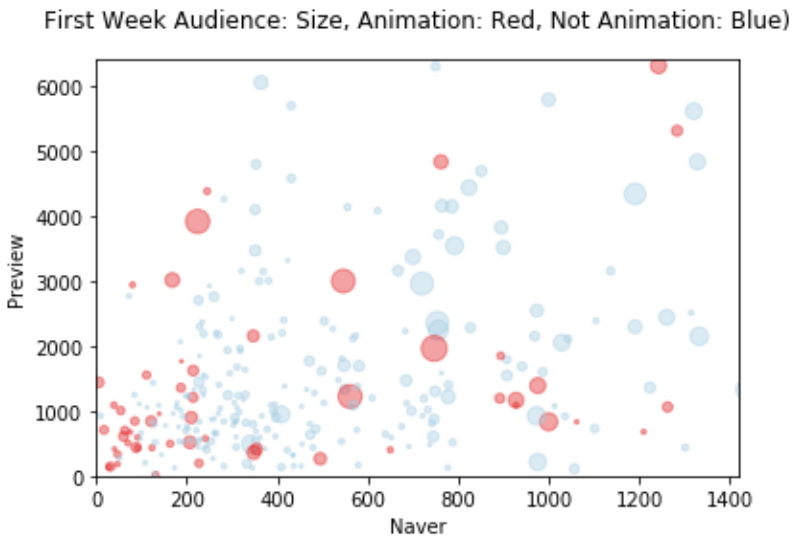
Figure 21



We can just give up one axis and have all other variables still working. This visualization is much better than Figure 19. It does show the correlation between 'Preview' and 'Naver' like heatmap (0.39). Also, it is much easier to confirm that red dots in general are larger than blue dots, which implies that 'animation' has more 'first week audience'. Also, it shows red dots have less 'preview' than blue dots which implies that animations have less 'previews'.

2. Naver, Preview, Animation - Color, Audience - Size : Enlarged

Figure 22



This is the enlarged version of the same figure that emphasizes how blue dots are smaller than red dots in general.

Future Work

The regression analysis would much improve if Twitter and Facebook data are added. Sentiment analysis on a movie platform may also improve the statistics than just a simple data. Also, other variables such as age, gender and release season (spring...winter) could be included in this analysis but could not collect the data. Moreover, economic status of each year or quarter should also be controlled. Lastly, larger dataset would much improve this analysis, however relevant indie-film was limited. If cost of each film is added, cost-benefit analysis could be possible too.

Conclusion

The aim of this project was to explore hidden insight from data from a previously completed regression analysis by visualizing in various ways. This project found that although indie-films that are animation may seem to attract less interest from potential audiences, it ends up having more audiences compared to other genres.

Moreover, this project showed that 2D visualization can much better explain the multivariate dataset. In 2D, other than axis, the size of data points and color was effective for plotting more than two variables. Color was especially useful for plotting a categorical data or a dummy variable.

This project also suggests how alpha and color should be used when plotting regression results. Dataset that has less length should have higher alpha than the others so that it could be observed well when datasets are scattered and overlapped.

Furthermore, this project has shown that the accuracy of an audience prediction model improves as the dependent variable is larger for this dataset.

The model used in this project is far from perfect. It requires more independent variables to fill the logical gap. However, the fact that this project discovered possible variables to solve this issue is meaningful.

Reference

Dey S. Predicting Gross Movie Revenue. (2016)

Lee, J. Indie Film Audience Prediction (Korean). Sungkyunkwan Univ. Korea. (2016)

Assady, M et al. Visual Analytics for the Prediction of Movie Rating and Box Office Performance, Konstanz Univ. Germany. (2013)

Merwe, J. Predicting Movie Box Office Gross. Stanford Univ. (2013)

Goodman E, A Predictor for Movie Success. Stanford Univ. (2013)

Zhou, Y, Predicting movie box-office revenues using deep neural networks. (2016)

Urpa, L.M., Anders, S. Focused multidimensional scaling: interactive visualization for exploration of high-dimensional data. *BMC Bioinformatics* 20, 221 (2019) doi:10.1186/s12859-019-2780-y