Final Report: Analysis of Top One Thousand Movies

Allie Ayrapetyan, Ahmad Azizi, Shehzad Shahbuddin
W200 MIDS Final Project
Summer-2021

Introduction:

Movies are the modern version of storytelling where people can get immersed in characters and sometimes idolize them. Movies, in some cases, could be a glimpse to reality, in other cases to fantasy. They have become an inseparable part of our society. Not only are movies an integral part of our culture but they have also become a lucrative business empire worth billions of dollars.

With multiple movies reaching over 2 billion dollars in sales, it has intrigued us as to what has made such movies financially successful. Since this question is too broad to answer, we have decided to narrow down our scope to gain valuable insight. We will explore multiple variables including Worldwide, international, and Domestic Box Offices, movie release date, genre, studio, and director for the top 1000 movies worldwide in terms of box office sales. We did not take inflation into account for our figures. However, to run a better statistical analysis, we took inflation into account for our analysis and conclusion.

Research Questions:

For this study, we have narrowed down our analysis and, therefore, intend to explore the following questions.

1. How does genre influence how great a movie does?
2. Does the number of directors a movie has impact its success?
3. Does the number of production studios a movie has impact its success?
4. Does the year of movie release determine its success?
5. Does the number of main stars in a movie impact the box office?

Research Method:

I. Data Sources:

https://www.the-numbers.com

We came across this website dataset while searching for movie box office information. The following details where we gathered specific information for our research.

- *Top 1000 List*
  https://www.the-numbers.com/box-office-records/worldwide/all-movies/cumulative/all-time
  this was the primary data source to identify the top 1000 grossing box office movies of all time. This data included the Movie Name, Rank, and Year of release, and Box Office numbers broken out by Worldwide, Domestic and International. With 100 movies listed per page, we gathered data on the first 10 pages. Each Movie Name was hyperlinked to a page that contained additional information about the movie. Avatar is the highest-ranked movie on the list and will be

used as an example for the links provided to highlight the additional information that was collected.

- *Genre & Production Companies*
  https://www.the-numbers.com/movie/Avatar#tab=summary
  The Summary tab includes information about the genre of the film as well as the production companies involved in the making of the movie.

- *Actor & Director Information*
  https://www.the-numbers.com/movie/Avatar#tab=cast-and-crew
  The Cast and Crew tab includes information about the leading actors in the movie and the directors responsible for making the movie.

- *Inflation Ranking*
  https://www.the-numbers.com/box-office-records/domestic/all-movies/cumulative/all-time-inflation-adjusted
  The original top 1000 list did not take into account the effects of inflation. To take this into account, we found the inflation-adjusted numbers for the movies in our list.

Below, we explain how the data from the website was pulled for our analysis.

II. Web Scraping and Data Cleaning.

In a Python Jupyter Notebook, we utilized the Beautiful Soup (bs4) library to scrape and parse the HTML of the various links used in our study. Here were the steps taken:

1. Use a for loop to go through the first ten pages of the top box office capturing the HTML using bs4
2. The table on each page contained the information for each movie on a distinct row between HTML codes <tr>...</tr> which was used as the parser.
3. Each link captured in the prior pull was then traversed and used bs4 to capture the desired information from the Summary and Cast & Crew tabs. This was done using a loop to go through all 1000 movie URLs. The HTML captured was reviewed to identify the necessary parsers for each of the desired data elements.
4. The inflation page was scraped in the same fashion as the top 1000 list.

See Appendix I to see sample HTML captured using bs4.

Once all the data was scrapped, it was placed in a data frame and joined with all the information collected from the various pages traversed. Once the data frame was built, the following steps were taken to clean the data.

1. All blank fields were filled with NaN
2. All Box Office numbers were formatted into floats removing the commas and dollar signs so Python could interact with the data as a number rather than a string.

3. The list of genres, production companies, actors, and directors were traversed to pull the top in each category.
4. The number of actors, directors, and production companies was also counted as another data element. Actors and directors were limited to a maximum of 3 items in the list for this study.

Check Appendix II for sample data.

Data Exploration:

In order to yield optimal results, we had to first look at the data we were working with and conduct appropriate data cleaning measures. Below, we will explain how specific data sets were optimized and what statistical analyses were performed.

Since the movie box office value had large numbers, we converted the variable *to per million* for easier visualization and interpretation.

We first analyzed the box office numbers. ***Figure 1*** shows a correlational heat map for our box office variables and we can see that domestic and international box offices have a .86 and .96 correlation to worldwide box office respectively. This makes sense because the worldwide box office is the sum of two. Because of this, we made the decision to exclude worldwide box office from the analysis. Additionally, we can see that domestic and international box offices have a correlation of .7 with each other (*Note: exploration on domestic and international box office showed some outliers).*

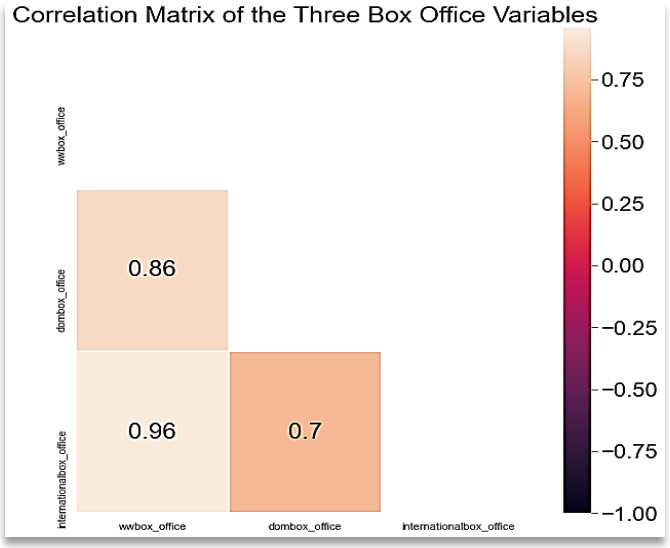


Figure 1. *Shows correlations between the three numbers using a heat map.*

Next, we calculated some statistical values for international and domestic box offices. From ***table 1,*** we can see that the mean and median for the two variables are different. Therefore, it is clear that our data is not normally distributed. Additionally, we can see both variables have unusually high max values, which could indicate outliers.

	domestic_per_million	int_per_million
count	981.000000	998.000000
mean	154.654536	246.955331
std	108.014621	202.551207
min	0.010000	28.700000
25%	90.710000	121.362500
50%	127.710000	182.550000
75%	184.030000	294.122500
max	936.660000	2085.390000

Table1 *shows summary statistics for domestic and international Box offices.*

To study the outliers, we used box plots for our two independent variables—domestic and international box office. As shown in ***figure 2,*** for domestic box office, anything any number over 300 million seems to be an outlier. For the international box office, any number above 550 million indicates an outlier.

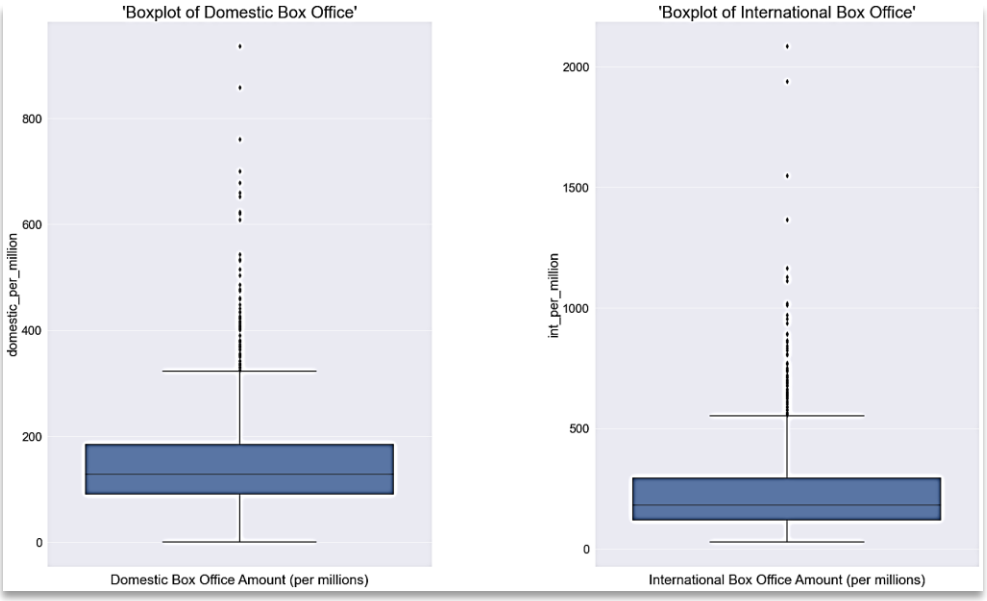


Figure 2. *Shows boxplots of domestic and international box offices. We can observe outliers for both.*

We wanted to analyze the data without the outliers, as with the outliers are models for the anlaysis were not good fits. To do so, we created two separate data frames, one for each variable with the outliers removed. The new data frame for domestic box office numbers went down from 1000 to 893 movies and the international box office numbers went from 1000 to 875. Since both samples are large enough, we decided that it should be fine to remove the outliers. ***Figure 3*** shows the new box plots for domestic and international box offices.

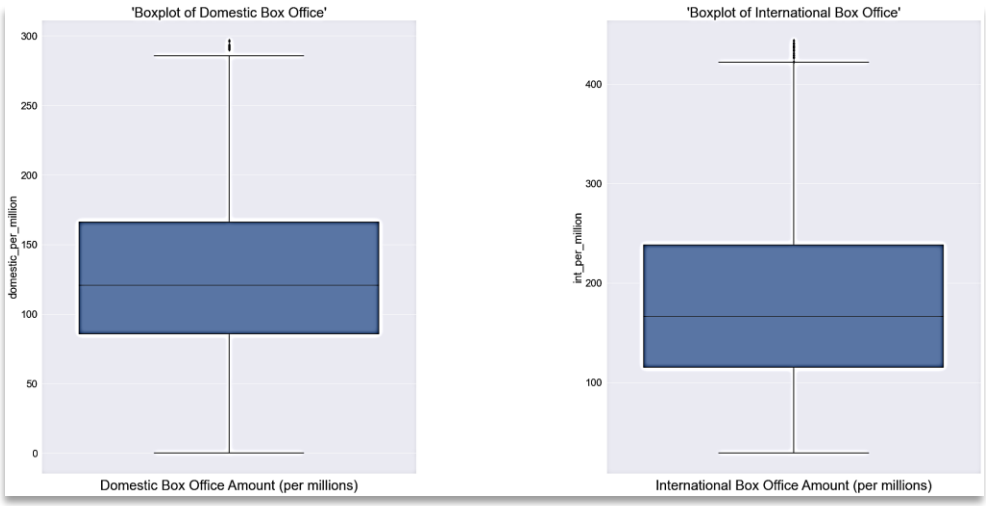


Figure 3. *Shows boxplots of domestic and international box offices with outliers removed.*

Next, we wanted to study the distribution of movies across the years from our database. In other words, how many movies were released each per year. ***Figure 4*** shows the distribution of movies across the years. It appears that most of the movies in our data set are from 2000 onwards.

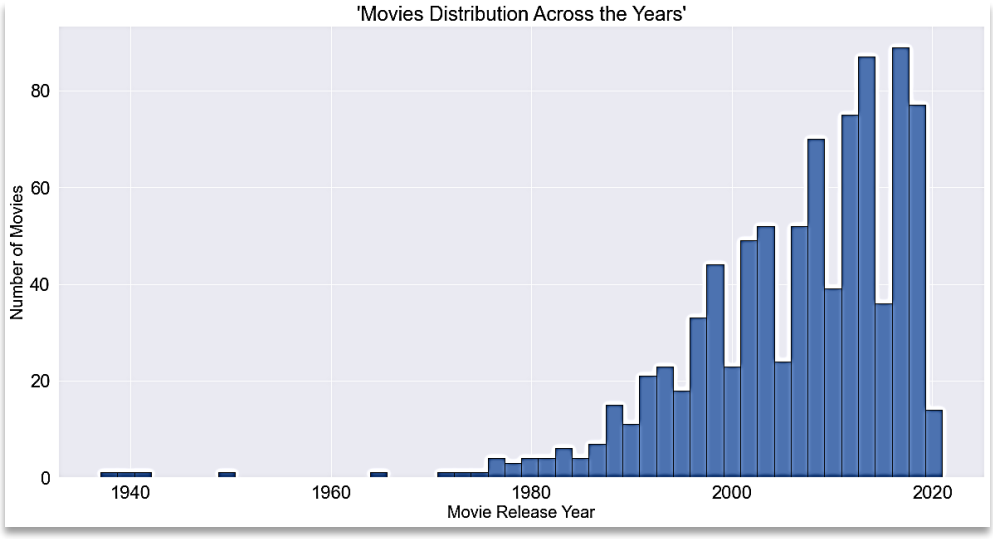


Figure 4 *shows movie distribution across years. The drop in the number of movies in 2020 is the cut-out point of our data.*

We also analyzed the distribution of domestic and international box offices across different genres. ***Figure 5*** shows this in boxplot format and ***Table 2a*** and ***Table 2b*** shows summary statistics for domestic and international box offices for each genre type, respectively. We can observe that the highest box office gross per genre type is different for international and domestic categories. For domestic gross, musicals have the highest average yet only 19 of the top 1000 movies are of this genre. Additionally, we can see that most movies in the top 1000 domestically fall under adventure. For international gross, action movies have the highest average. And just like domestic box office gross, most movies in the top

1000 internationally fall under adventure. It's also worth noting that there is a lot of variability in both the domestic and international box office in movies, as seen in the high standard deviation values.

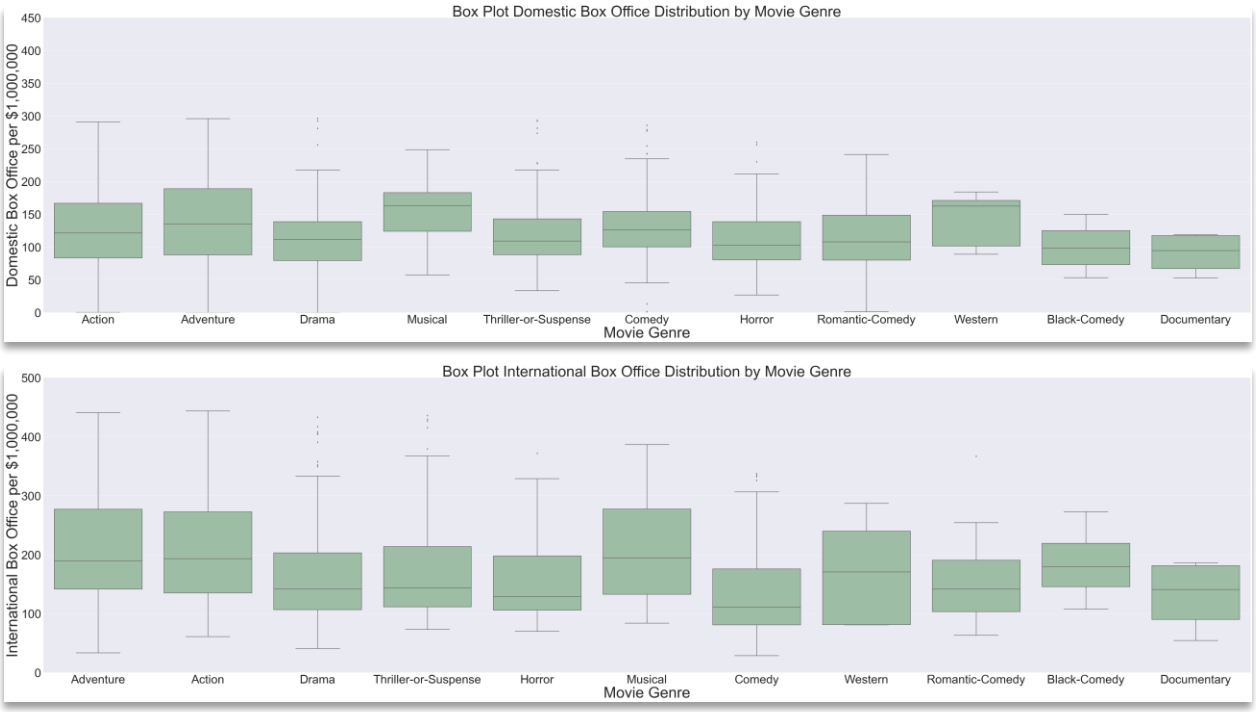


Figure 5 shows the distribution of domestic and international box office for movies across genres.

Summary Stats of Domestic Box Office (per millions) for each Movie Genre Type

genre	count	mean	std	min	25%	50%	75%	max
Musical	19.0	155.673684	47.813171	57.39	124.3150	163.210	183.3700	248.76
Western	5.0	141.838000	43.233706	89.30	101.6300	162.810	171.2400	184.21
Adventure	266.0	141.405150	68.708380	0.02	88.2375	135.330	189.2025	296.13
Comedy	116.0	128.796552	55.448234	0.67	100.4875	126.550	154.5075	285.76
Action	218.0	125.418073	61.386866	0.36	83.6375	121.630	167.0475	291.05
Thriller-or-Suspense	72.0	124.444028	57.616543	33.70	88.4000	109.440	143.1250	293.51
Horror	40.0	117.244500	54.790879	26.84	80.4550	102.985	138.7400	260.00
Drama	110.0	115.548182	58.402147	0.01	79.4675	111.550	138.7075	296.62
Romantic-Comedy	38.0	108.444737	53.544889	1.29	80.3000	108.150	148.6900	241.44
Black-Comedy	4.0	100.037500	42.357749	53.37	73.1400	98.340	125.2375	150.10
Documentary	4.0	90.307500	33.131530	52.80	67.2675	94.660	117.7000	119.11

Table 2a. *Shows summary statistics for domestic box office per genre.*

Summary Stats of International Box Office (per millions) for each Movie Genre Type

genre	count	mean	std	min	25%	50%	75%	max
Action	209.0	214.097608	96.616591	61.10	135.0000	193.140	272.8000	443.71
Adventure	249.0	213.950120	96.068295	33.60	141.7700	189.540	277.2900	441.37
Musical	16.0	213.520000	95.681240	83.95	132.8100	194.750	277.5725	386.85
Black-Comedy	4.0	185.132500	69.938833	107.76	145.8000	179.900	219.2325	272.97
Thriller-or-Suspense	75.0	175.889200	91.671724	73.37	111.5200	143.440	213.8350	436.01
Western	5.0	172.030000	92.665275	81.03	81.3800	170.700	240.0000	287.04
Drama	113.0	167.884956	88.233771	40.99	106.7900	141.600	203.3000	433.24
Horror	41.0	153.738293	70.964816	70.00	105.9800	129.400	197.8700	372.25
Romantic-Comedy	39.0	151.910256	63.418980	63.57	103.4550	141.910	190.7450	366.96
Comedy	119.0	131.432605	70.834485	28.70	81.0900	111.200	176.0750	337.35
Documentary	4.0	130.672500	63.627152	54.46	90.1375	141.015	181.5500	186.20

Table 2b. Shows summary statistics for international box office per genre.

The next variables that we looked at were production studio name, directors, and main stars. Upon studying our data, we found out that there were over 200 unique fields for our overall data among these variables. *Table 3* and *Table 4, in appendix V* show summary statistics for the Production studio and director for each movie, respectively (*Note to preserve space, we are only showing the first 10 rows for each table).* Moreover, each movie could have more than one director, production studio, or main star. That being said, we decided to create a few new quantitative variable to help us with the process: number of directors (number of directors each movie had), number of the studio (number of studios each movie had), and number of stars (number of main stars each movie had). *Table 5* shows summary statistics of our three new variables. From this table, we can see that for a movie from our dataset the average number of production studios is about two, the number of stars is about two, and the number of directors is one. One interesting point stands out; one movie has 20 production companies in charge. This value is uniquely different from the rest of the table.

	number_of_production_companies	number_of_stars	number_of_directors
count	1000.000000	1000.000000	1000.000000
mean	2.592000	1.576000	0.825000
std	1.641411	1.213135	0.551978
min	1.000000	0.000000	0.000000
25%	1.000000	0.000000	1.000000
50%	2.000000	2.000000	1.000000
75%	3.000000	3.000000	1.000000
max	20.000000	3.000000	3.000000

Table 5. Shows the summary statistics for the number of production companies, number of main stars, and number of directors for each movie

We analyzed the number of production studios through a box plot to gain some insight. As shown in ***figure 6,*** there seems to be some outliers, we can leave them in our data. More so, we can tell from ***table 6*** that the number of production studios for a movie is strongly correlated with the number of stars. There may be some multicollinearity involved there so we considered using only one of those variables.

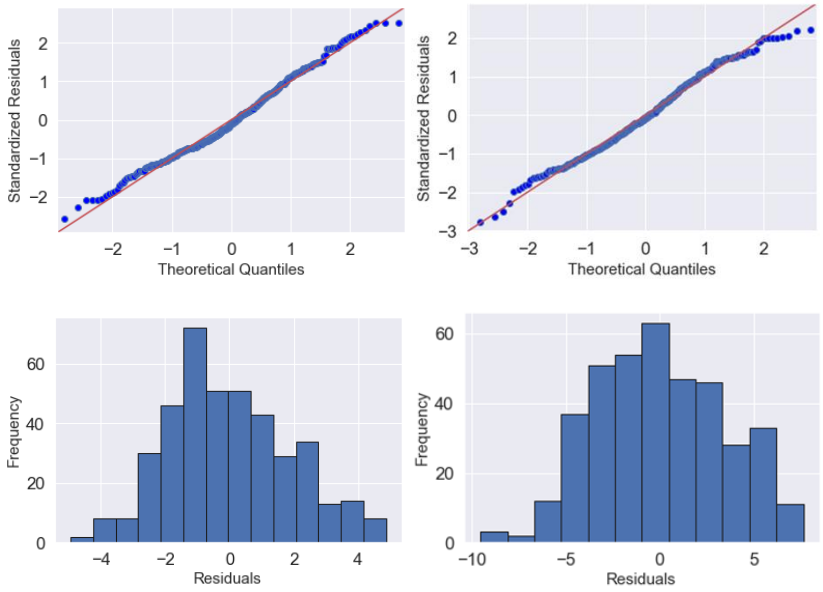
Figure 6. *Shows a boxplot of the number of production studios with its outliers.*

	number_of_production_companies	number_of_stars	number_of_directors
number_of_production_companies	1.000000	0.232251	0.186274
number_of_stars	0.232251	1.000000	0.658939
number_of_directors	0.186274	0.658939	1.000000

Table 6. *Shows the correlation between number of production studios, number of main stars, and number of directors.*

Analysis and Conclusion:

Is genre a factor of the domestic and international box office? To help us answer this question, we performed a *1 way ANOVA* on genre against both box office variables. Transformed domestic and international variables with square root, as neither were normally distributed and residuals were not constant. We, then, ran ANOVA with the transformed variables. ***Figure 7*** shows Residuals of Genre ~ Domestic Box Office and Residuals of Genre ~ International Box Office-normally distributed. Both residuals are now normally distribute because of the bell curve histogram and the variances are constant, Levene's test shows p-value more than .05. ***Figure 1*** in appendix VI show the model diagnostics pre-transfomrations, which you can see the assumptions are not met.



Levene's Test for Equality of Variances

	Parameter	Value
0	Test statistics (W)	1.1921
1	Degrees of freedom (Df)	9.0000
2	p value	0.2983

Levene's Test for Equality of Variances

	Parameter	Value
0	Test statistics (W)	1.0600
1	Degrees of freedom (Df)	9.0000
2	p value	0.3917

Figure 7. *(Left) Model diagnsostics for standard residuals and constant variance for Genre ~ Domestic Box Office. (Right) Model diagnsostics for standard residuals and constant variance for Genre ~ International Box Office*

We found that there is a significant difference domestic and international box office growth among genre types, as the p-values for both are less than 0.05 as shown in ***table 7.***

Anova for Domestic Box Office

	df	sum_sq	mean_sq	F	PR(>F)
C(genre)	10.0	216.773007	21.677301	2.264387	0.01293
Residual	881.0	8433.940460	9.573145	NaN	NaN

Anova for International Box Office

	df	sum_sq	mean_sq	F	PR(>F)
C(genre)	10.0	1233.379227	123.337923	12.446515	2.252398e-20
Residual	863.0	8551.841815	9.909434	NaN	NaN

Table 7. *Shows anova results for domestic and international box office against movie genre.*

Moreover, we conducted a *Tukey Post Hoc Tes*t to determine which genre types are exactly significantly different with each other in the domestic box office. As you can see in ***table 8,*** there is a significant impact on domestic box office between Drama and Adventure, Adventure and Thriller/Suspense.

group1	group2	Diff	Lower	Upper	q-value	p-value
Adventure	Drama	1.109012	0.041539	2.176485	4.674768	0.034408
Adventure	Thriller-or-Suspense	1.225970	0.094113	2.357827	4.873815	0.021925

Table 8. *Shows the domestic Box Office Post Hoc Analysis.*

Results are more interesting when looking at the international box office. There are more genre differences that impact international growth, demonstrated in ***table 9.***

group1	group2	Diff	Lower	Upper	q-value	p-value
Adventure	Drama	2.086899	0.185733	3.988064	4.940903	0.018838
Adventure	Horror	3.078623	0.173181	5.984065	4.769458	0.027953
Adventure	Comedy	4.627777	2.756654	6.498899	11.132560	0.001000
Adventure	Romantic-Comedy	3.714875	0.525807	6.903943	5.243306	0.008999
Action	Drama	2.428910	0.414752	4.443068	5.428036	0.005591
Action	Horror	3.420634	0.440031	6.401237	5.165677	0.010931
Action	Comedy	4.969788	2.983963	6.955613	11.264754	0.001000
Action	Romantic-Comedy	4.056886	0.799194	7.314578	5.605413	0.003480
Drama	Comedy	2.540878	0.257636	4.824120	5.009064	0.016024
Thriller-or-Suspense	Comedy	3.083850	0.695588	5.472112	5.812139	0.001963
Musical	Comedy	4.050217	0.383460	7.716973	4.971884	0.017510

Table 9. *Shows the international Box Office Post Hoc Analysis.*

Next, to answer the following questions, we conducted an OLS linear regression.

- Does movie release year a factor of the domestic and international box office?
- Does the number of studios a movie has impact its box office (international and domestic)?
- Does the number of main stars a movie has impact its box office (international and domestic)?
- Does the number of directors a movie has impact its box office (international and domestic)?
- Does inflation play a role in movie box office?

Since the units for the year and inflation gross differ greatly from the number of stars, directors, and studios, we need to standardize all the variables. Otherwise, our effect size would be overly reported and our model will not be accurate. And so, we applied min_max standardization on all our variables to help resolve this issue. Additionally, there are some missing values from inflation grosses, so our sample size drops a little bit when running the models. However, the sizes are still large enough to proceed with the analysis.

We also ran correlation tests (***Tables 10*** and ***11*** below) against our independent and dependent variables, there seem to not be any issue with muliticollinearity. More so, there is some linear releationship between the independent and dependent variables.

	domestic_per_million	year	number_of_production_companies	number_of_stars	number_of_directors	infl_gross_per_million
domestic_per_million	1.000000	-0.110175	-0.091958	0.206882	0.157368	0.604967
year	-0.110175	1.000000	0.399817	0.380423	0.341786	-0.503827
number_of_production_companies	-0.091958	0.399817	1.000000	0.257788	0.199439	-0.231589
number_of_stars	0.206882	0.380423	0.257788	1.000000	0.659553	0.048216
number_of_directors	0.157368	0.341786	0.199439	0.659553	1.000000	0.080352
infl_gross_per_million	0.604967	-0.503827	-0.231589	0.048216	0.080352	1.000000

Table 10. Correlation matrix for domestic box office with independent variables.

	int_per_million	year	number_of_production_companies	number_of_stars	number_of_directors	infl_gross_per_million
int_per_million	1.000000	0.224452	0.088625	0.226857	0.271726	0.264631
year	0.224452	1.000000	0.393679	0.335207	0.284110	-0.463482
number_of_production_companies	0.088625	0.393679	1.000000	0.264124	0.226224	-0.188438
number_of_stars	0.226857	0.335207	0.264124	1.000000	0.681920	0.076446
number_of_directors	0.271726	0.284110	0.226224	0.681920	1.000000	0.109958
infl_gross_per_million	0.264631	-0.463482	-0.188438	0.076446	0.109958	1.000000

Table 11. Correlation matrix for international box office with independent variables.

Model 2: domestic_per_million ~ number_of_production_companies + number_of_directors + infl_gross_per_million

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     domestic_per_million   R-squared:                       0.535
Model:                              OLS   Adj. R-squared:                  0.533
Method:                   Least Squares   F-statistic:                     231.4
Date:                  Sat, 31 Jul 2021   Prob (F-statistic):          7.08e-100
Time:                          19:38:26   Log-Likelihood:                 426.61
No. Observations:                   607   AIC:                            -845.2
Df Residuals:                       603   BIC:                            -827.6
Df Model:                             3
Covariance Type:              nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        0.1311      0.017      7.588      0.000       0.097       0.165
number_of_production_companies   0.4338      0.071      6.098      0.000       0.294       0.574
number_of_directors              0.3118      0.026     11.953      0.000       0.261       0.363
infl_gross_per_million           0.9889      0.046     21.504      0.000       0.899       1.079
==============================================================================
Omnibus:                        8.567   Durbin-Watson:                   1.204
Prob(Omnibus):                  0.014   Jarque-Bera (JB):               12.597
Skew:                           0.078   Prob(JB):                      0.00184
Kurtosis:                       3.688   Cond. No.                         16.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

To predict domestic box office and see what factors have an impact on domestic box office, we tried 3 different models. Looking at our outputs, considering R squared values along with other model diagnostics, the best model seemed to be model 2 (Model 1 and 3 can be found in Appendix III). Even though model 1 has the higher R squared, its diagnostic checks are not the greatest. Especially when looking at the Omnibus value. Our final model has a 55% of its variability explained by the predictors; the number of production studios of a movie, the number of directors for a movie, and the gross box office adjusted due to inflation. Because the coefficients for all 3 variables are positive, we can conclude that all 3 variables do have a positive effect on the domestic box office. Specifically, the more the number of production companies and directors behind a movie, the better it will do. Additionally, controlling for inflation, recent movies do better. That being said, the number of main stars a movie has doesn't have a significant impact on the domestic box office.

Model 1: int_per_million ~ year + number_of_production_companies + number_of_stars + number_of_directors+ infl_gross_per_million

OLS Regression Results

Dep. Variable:	int_per_million	R-squared:	0.318
Model:	OLS	Adj. R-squared:	0.312
Method:	Least Squares	F-statistic:	53.85
Date:	Sat, 31 Jul 2021	Prob (F-statistic):	6.95e-46
Time:	19:38:26	Log-Likelihood:	137.88
No. Observations:	583	AIC:	-263.8
Df Residuals:	577	BIC:	-237.5
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-23.9217	2.345	-10.199	0.000	-28.528	-19.315
year	24.3261	2.367	10.277	0.000	19.677	28.975
number_of_production_companies	-0.1202	0.124	-0.971	0.332	-0.363	0.123
number_of_stars	-0.0063	0.030	-0.210	0.834	-0.065	0.052
number_of_directors	0.1725	0.064	2.704	0.007	0.047	0.298
infl_gross_per_million	1.2886	0.113	11.397	0.000	1.067	1.511

Omnibus:	8.853	Durbin-Watson:	1.092
Prob(Omnibus):	0.012	Jarque-Bera (JB):	7.819
Skew:	0.220	Prob(JB):	0.0200
Kurtosis:	2.642	Cond. No.	639.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

To consider impacts on international box office, we also looked at 3 different models. Looking at our outputs, considering R squared values along with other model diagnostics, the best model seemed to be model 1. (Model 2 and 3 can be found in Appendix IV). Even though R squared is not that high, the model diagnostic statistics are the lowest, which is more important. This model has 32% of its variability explained by the predictors-movie release year, number of production studios of a movie, number of main stars, number of directors for a movie, and the gross box office adjusted due to inflation. What's interesting for the international box office is the significant predictors are different from those predicting domestic box office. For starters, the movie release year is significant here, where it wasn't in the domestic model. We see from the positive coefficient that newer movies do better internationally. Moreover, since inflation gross is also a significant variable, we can say that this is true even when we control for inflation. Another difference from the domestic box office model is that number of main stars do actually predict a movie's international success. Last but not least, similarly for domestic gross, the number of directors a movie has is also a significant predictor for the international box office- the more directors, the better it will do.

Our two final models plus ANOVA showed:

- number of directors **does** have an impact on both domestic and international box office
- number of production studios **does** impact domestic box office but **not** international
- number of main stars in a movie **does** not impact international or domestic box offices.
- movie release year **does** impact international box office but **not** domestic
- Movie gross due to inflation **does** impact both international and domestic box office
- Movie genre **does** make an impact on both international and domestic box office

Appendix I

Primary Data

```
<tr>
<td class="data">1</td>
<td class="data"><a href="/box-office-records/worldwide/all-movies/cumulative/released-in-2009">2009</a></td>
<td><b><a href="/movie/Avatar#tab=summary">Avatar</a></b></td>
<td align="right">$2,845,899,541</td>
<td align="right">$760,507,625</td>
<td align="right">$2,085,391,916</td>
</tr>
<tr>
<td class="data">2</td>
<td class="data"><a href="/box-office-records/worldwide/all-movies/cumulative/released-in-2019">2019</a></td>
<td><b><a href="/movie/Avengers-Endgame-(2019)#tab=summary">Avengers: Endgame</a></b></td>
<td align="right">$2,797,800,564</td>
<td align="right">$858,373,000</td>
<td align="right">$1,939,427,564</td>
</tr>
<tr>
<td class="data">3</td>
```

Genre & Production Company

```
<tr><td><b>Source:</b></td><td><a href="/market/source/Original-Screenplay">Original Screenplay</a></td></tr>
<tr><td><b>Genre:</b></td><td><a href="/market/genre/Action">Action</a></td></tr>
<tr><td><b>Production Method:</b></td><td><a href="/market/production-method/Animation-and-Live-Action">Animatio
e Action</a></td></tr>
<tr><td><b>Creative Type:</b></td><td><a href="/market/creative-type/Science-Fiction">Science Fiction</a></td></
<tr><td><b>Production Companies:</b></td>
<td><a href="/movies/production-company/Dune-Entertainment">Dune Entertainment</a>, <a href="/movies/production-
ny/20th-Century-Fox">20th Century Fox</a>, <a href="/movies/production-company/Ingenious-Film-Partners">Ingenious
m Partners</a></td></tr>
<tr><td><b>Production Countries:</b></td>
<td><a href="/United-States/movies">United States</a></td></tr>
<tr><td><b>Languages:</b></td>
<td><a href="/language/English/movies">English</a>, <a href="/language/Navi/movies">Na'vi</a></td></tr>
</table>
```

Actor (itemprop = “actor”)

```
<tr>
<td align="right" itemprop="actor" itemscope="" itemtype="https://schema.org/Person" width="49%"><b><a href="/per
60740401-Mark-Hamill" itemprop="url"><span itemprop="name">Mark Hamill</span></a></b></td>
<td>  </td>
<td align="left" width="49%">Luke Skywalker</td>
</tr>
<tr>
<td align="right" itemprop="actor" itemscope="" itemtype="https://schema.org/Person" width="49%"><b><a href="/per
600401-Harrison-Ford" itemprop="url"><span itemprop="name">Harrison Ford</span></a></b></td>
<td>  </td>
<td align="left" width="49%">Han Solo</td>
</tr>
<tr>
<td align="right" itemprop="actor" itemscope="" itemtype="https://schema.org/Person" width="49%"><b><a href="/per
48440401-Carrie-Fisher" itemprop="url"><span itemprop="name">Carrie Fisher</span></a></b></td>
<td>  </td>
<td align="left" width="49%">Princess Leia Organa</td>
</tr>
```

Director

```
<table align="center" cellspacing="0">
<tr>
<td align="right" itemprop="director" itemscope="" itemtype="https://schema.org/Person" width="49%"><b><a href="/pers
on/88340401-George-Lucas" itemprop="url"><span itemprop="name">George Lucas</span></a></b></td>
<td>  </td>
<td align="" width="49%">Director</td>
</tr><tr>
<td align="right" width="49%"><b><a href="/person/203280401-Gary-Kurtz" rel="nofollow">Gary Kurtz</a></b></td>
<td>  </td>
<td align="" width="49%">Producer</td>
</tr><tr>
<td align="right" width="49%"><b><a href="/person/88340401-George-Lucas">George Lucas</a></b></td>
<td>  </td>
<td align="" width="49%">Screenwriter</td>
</tr><tr>
```

Appendix II

rank	year	movie	wwbox_office	dombox_office	internationalbox_office	url	genre	production_companies	inflation_rank
1	2009	Avatar	2.845900e+09	760507625.0	2.085392e+09	/movie/Avatar	Action	[Dune-Entertainment, 20th-Century-Fox, Ingenio...	5
2	2019	Avengers: Endgame	2.797801e+09	858373000.0	1.939428e+09	/movie/Avengers-Endgame-(2019)	Action	[Marvel-Studios]	7
3	1997	Titanic	2.207987e+09	659363944.0	1.548623e+09	/movie/Titanic-(1997)	Drama	[20th-Century-Fox, Paramount-Pictures, Lightst...	2
4	2015	Star Wars Ep. VII: The Force Awakens	2.064616e+09	936662225.0	1.127954e+09	/movie/Star-Wars-Ep-VII-The-Force-Awakens	Adventure	[Lucasfilm, Bad-Robot]	4
5	2018	Avengers: Infinity War	2.044541e+09	678815482.0	1.365725e+09	/movie/Avengers-Infinity-War	Action	[Marvel-Studios]	18

infl_gross_box_office	actors	directors	main	number_of_stars	director_main	number_of_directors	production_companies_main	number_of_production_companies
9.041786e+08	[Sam Worthington, Zoe Saldana]	[James Cameron]	Sam Worthington	2	James Cameron	1	Dune-Entertainment	3
8.583730e+08	[Robert Downey, Jr., Chris Evans, Mark Ruffalo]	[Joe Russo, Anthony Russo]	Robert Downey, Jr.	3	Joe Russo	2	Marvel-Studios	1
1.247411e+09	[Leonardo DiCaprio, Kate Winslet]	[James Cameron]	Leonardo DiCaprio	2	James Cameron	1	20th-Century-Fox	3
1.012399e+09	[Adam Driver, Daisy Ridley, John Boyega]	[J.J. Abrams]	Adam Driver	3	J.J. Abrams	1	Lucasfilm	2
6.825411e+08	[Robert Downey, Jr., Chris Hemsworth, Mark Ruf...	[Joe Russo, Anthony Russo]	Robert Downey, Jr.	3	Joe Russo	2	Marvel-Studios	1

Appendix III

Model 1: domestic_per_million ~ year + number_of_production_companies + number_of_stars + number_of_directors +infl_gross_per_million

OLS Regression Results

Dep. Variable:	domestic_per_million	R-squared:	0.887
Model:	OLS	Adj. R-squared:	0.886
Method:	Least Squares	F-statistic:	944.0
Date:	Mon, 02 Aug 2021	Prob (F-statistic):	8.15e-282
Time:	20:41:28	Log-Likelihood:	855.98
No. Observations:	607	AIC:	-1700.
Df Residuals:	601	BIC:	-1674.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-31.2826	0.740	-42.249	0.000	-32.737	-29.828
year	31.6475	0.746	42.426	0.000	30.183	33.112
number_of_production_companies	-0.0984	0.037	-2.640	0.009	-0.172	-0.025
number_of_stars	-0.0082	0.009	-0.910	0.363	-0.026	0.009
number_of_directors	0.0275	0.019	1.447	0.148	-0.010	0.065
infl_gross_per_million	1.6303	0.027	59.671	0.000	1.577	1.684

Omnibus:	149.403	Durbin-Watson:	1.657
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1123.913
Skew:	-0.871	Prob(JB):	8.82e-245
Kurtosis:	9.435	Cond. No.	674.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Model 3: domestic_per_million ~ year + number_of_production_companies + number_of_directors

OLS Regression Results

Dep. Variable:	domestic_per_million	R-squared:	0.060
Model:	OLS	Adj. R-squared:	0.057
Method:	Least Squares	F-statistic:	18.82
Date:	Sat, 31 Jul 2021	Prob (F-statistic):	7.73e-12
Time:	19:38:26	Log-Likelihood:	159.13
No. Observations:	892	AIC:	-310.3
Df Residuals:	888	BIC:	-291.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.6084	1.455	4.543	0.000	3.753	9.463
year	-6.2645	1.473	-4.254	0.000	-9.155	-3.374
number_of_production_companies	-0.1862	0.090	-2.077	0.038	-0.362	-0.010
number_of_directors	0.2563	0.039	6.512	0.000	0.179	0.334

Omnibus:	11.961	Durbin-Watson:	1.233
Prob(Omnibus):	0.003	Jarque-Bera (JB):	12.090
Skew:	0.282	Prob(JB):	0.00237
Kurtosis:	3.080	Cond. No.	439.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Appendix IV

Model 2: int_per_million ~ number_of_production_companies + number_of_stars + number_of_directors+ infl_gross_per_million

OLS Regression Results

Dep. Variable:	int_per_million	R-squared:	0.193
Model:	OLS	Adj. R-squared:	0.188
Method:	Least Squares	F-statistic:	34.64
Date:	Sat, 31 Jul 2021	Prob (F-statistic):	6.05e-26
Time:	19:38:26	Log-Likelihood:	88.881
No. Observations:	583	AIC:	-167.8
Df Residuals:	578	BIC:	-145.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1800	0.027	6.790	0.000	0.128	0.232
number_of_production_companies	0.3393	0.125	2.705	0.007	0.093	0.586
number_of_stars	0.0508	0.032	1.592	0.112	-0.012	0.114
number_of_directors	0.2968	0.068	4.359	0.000	0.163	0.431
infl_gross_per_million	0.6854	0.105	6.526	0.000	0.479	0.892

Omnibus:	23.465	Durbin-Watson:	0.783
Prob(Omnibus):	0.000	Jarque-Bera (JB):	25.580
Skew:	0.506	Prob(JB):	2.79e-06
Kurtosis:	2.833	Cond. No.	17.9

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Model 3: int_per_million ~ year + number_of_production_companies + number_of_stars + number_of_directors

OLS Regression Results

Dep. Variable:	int_per_million	R-squared:	0.099
Model:	OLS	Adj. R-squared:	0.095
Method:	Least Squares	F-statistic:	23.89
Date:	Sat, 31 Jul 2021	Prob (F-statistic):	9.12e-19
Time:	19:38:26	Log-Likelihood:	160.80
No. Observations:	874	AIC:	-311.6
Df Residuals:	869	BIC:	-287.7
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.3420	1.461	-4.340	0.000	-9.210	-3.474
year	6.7517	1.479	4.566	0.000	3.850	9.654
number_of_production_companies	-0.0867	0.091	-0.951	0.342	-0.266	0.092
number_of_stars	0.0215	0.024	0.902	0.367	-0.025	0.068
number_of_directors	0.2382	0.051	4.625	0.000	0.137	0.339

Omnibus:	60.733	Durbin-Watson:	0.560
Prob(Omnibus):	0.000	Jarque-Bera (JB):	72.592
Skew:	0.706	Prob(JB):	1.73e-16
Kurtosis:	3.018	Cond. No.	466.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Appendix V

Summary Stats of Domestic Box Office (per millions) for each Movie Studio

production_companies_main	count	mean	std	min	25%	50%	75%	max
Vertigo-Entertainment	1.0	257.78	NaN	257.78	257.78	257.78	257.780	257.78
Temple-Hill-Entertainment	3.0	255.46	54.570526	192.77	237.03	281.29	286.805	292.32
Guber-Peters	1.0	251.19	NaN	251.19	251.19	251.19	251.190	251.19
Spyglass-Entertainment	1.0	242.70	NaN	242.70	242.70	242.70	242.700	242.70
Bad-Hat-Harry-Productions	1.0	233.92	NaN	233.92	233.92	233.92	233.920	233.92
...	...	...	...	...	...	...	...	...
Far-East-Films	1.0	0.29	NaN	0.29	0.29	0.29	0.290	0.29
Bona-Film-Group	1.0	0.22	NaN	0.22	0.22	0.22	0.220	0.22
Edko-Films	1.0	0.03	NaN	0.03	0.03	0.03	0.030	0.03
Village-Roadshow-Asia	1.0	0.02	NaN	0.02	0.02	0.02	0.020	0.02
Dirty-Monkey-Films-Group	1.0	0.01	NaN	0.01	0.01	0.01	0.010	0.01

245 rows × 8 columns

Table 3. *Shows summary statistics for unique Production Studios*

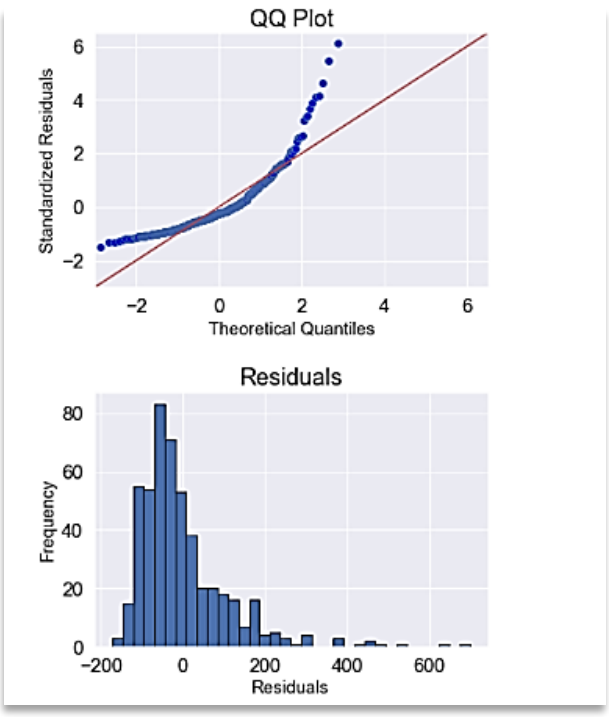
Summary Stats of Domestic Box Office (per millions) by Director

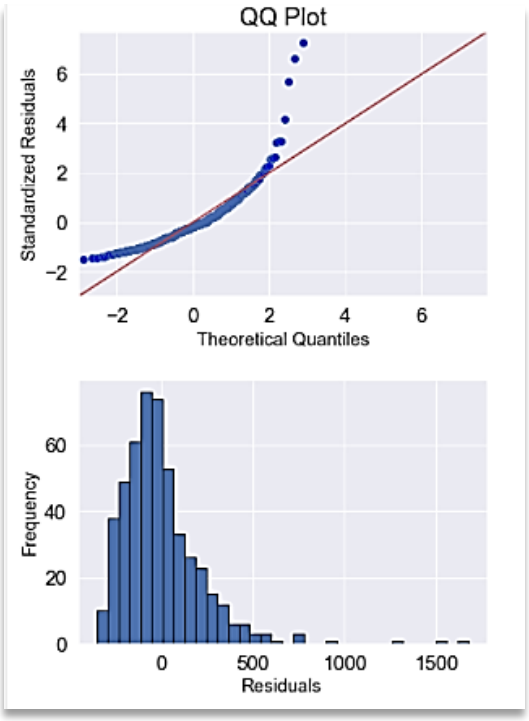
director_main	count	mean	std	min	25%	50%	75%	max
Chris Weitz	1.0	296.620	NaN	296.62	296.6200	296.620	296.6200	296.62
Irvin Kershner	1.0	291.740	NaN	291.74	291.7400	291.740	291.7400	291.74
Pete Docter	2.0	291.575	2.015254	290.15	290.8625	291.575	292.2875	293.00
Bill Condon	2.0	286.805	7.799388	281.29	284.0475	286.805	289.5625	292.32
Andrew Adamson	2.0	279.685	17.005918	267.66	273.6725	279.685	285.6975	291.71
...	...	...	...	...	...	...	...	...
Raman Hui	2.0	0.370	0.480833	0.03	0.2000	0.370	0.5400	0.71
Stanley Tong	1.0	0.360	NaN	0.36	0.3600	0.360	0.3600	0.36
Tony Chan	1.0	0.290	NaN	0.29	0.2900	0.290	0.2900	0.29
Teng Cheng	1.0	0.210	NaN	0.21	0.2100	0.210	0.2100	0.21
Muye Wen	1.0	0.010	NaN	0.01	0.0100	0.010	0.0100	0.01

380 rows × 8 columns

Table 4. *Shows summary statistics for unique Directors.*

Appendix VI





Levene's Test for Constant Variance

	Parameter	Value
0	Test statistics (W)	2.7038
1	Degrees of freedom (Df)	9.0000
2	p value	0.0044

Levene's Test for Constant Variance

	Parameter	Value
0	Test statistics (W)	4.959
1	Degrees of freedom (Df)	9.000
2	p value	0.000

Figure : Model diagnostics for Genre~Domestic Box Office (Left) and Genre ~ International Box Office (Right) pre transformations.