

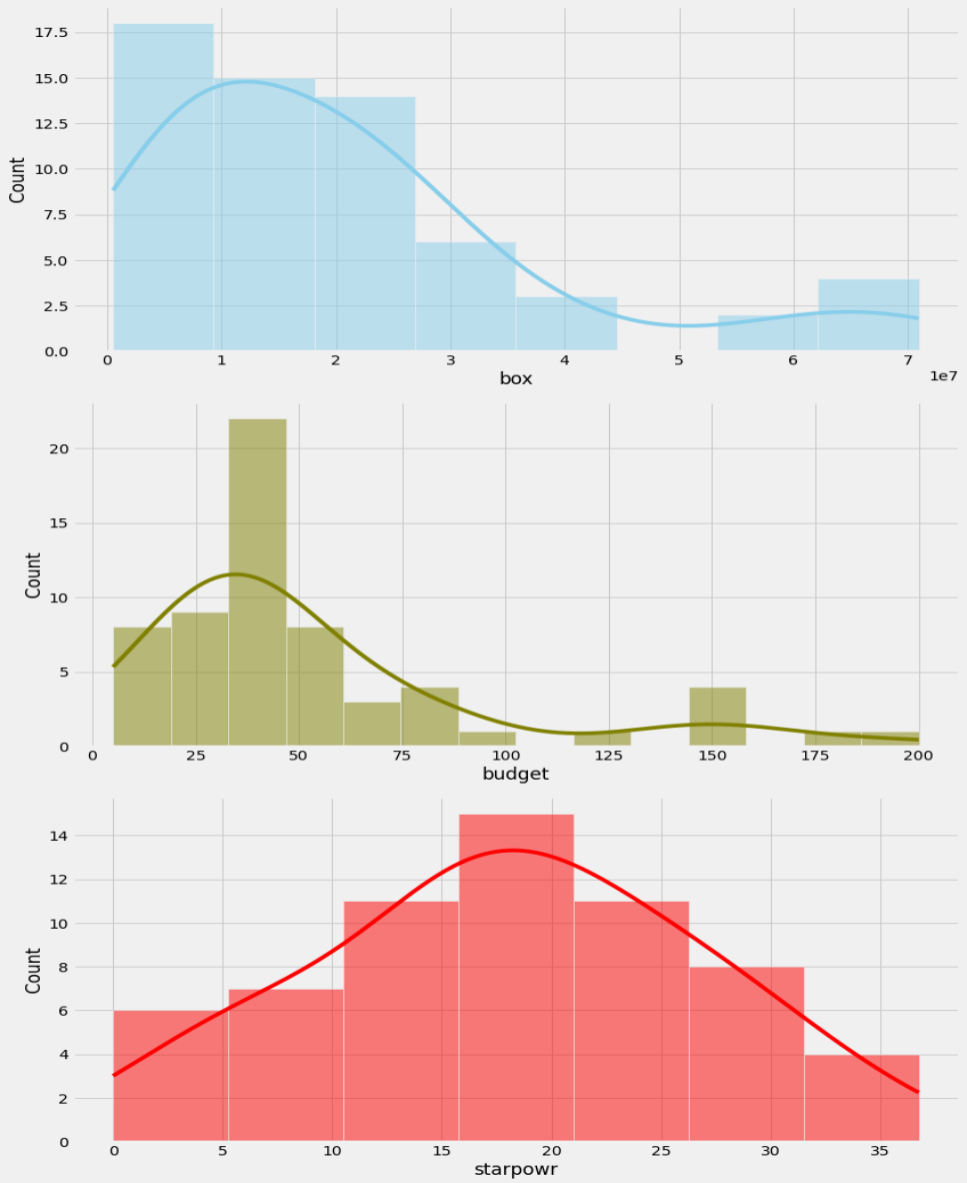


# Internet Buzz

# Box-Office Revenue Analysis


# Table of Contents

# 1. Histograms of the continuous variables (box, budget, starpwr)




Observations 1.1

- Based on the histograms of the three continuous variables(box, budget, and starpowr) shown above, we could see the ranges are quite different.
- Also, box and budget are right skewed while star power is closer to a normal distribution.
- Thus, box and budget need transformations. Next, we apply log-transformations to box and budget.

## 2. Linear regression (1) of box office revenues on the "traditional" variables (except buzz variables)

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.342 |
| Model: | OLS | Adj. R-squared: | 0.214 |
| Method: | Least Squares | F-statistic: | 2.656 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 0.0109 |
| Time: | 09:58:12 | Log-Likelihood: | -70.832 |
| No. Observations: | 62 | AIC: | 163.7 |
| Df Residuals: | 51 | BIC: | 187.1 |
| Df Model: | 10 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| G | 0.2524 | 0.692 | 0.365 | 0.717 | -1.137 | 1.642 |
| PG | 0.3297 | 0.372 | 0.885 | 0.380 | -0.418 | 1.077 |
| PG13 | 0.0718 | 0.272 | 0.264 | 0.793 | -0.474 | 0.618 |
| starpowr | 0.0065 | 0.016 | 0.403 | 0.689 | -0.026 | 0.039 |
| sequel | 0.6437 | 0.331 | 1.942 | 0.058 | -0.022 | 1.309 |
| action | -0.3068 | 0.344 | -0.892 | 0.377 | -0.997 | 0.384 |
| comedy | -0.0385 | 0.321 | -0.120 | 0.905 | -0.682 | 0.605 |
| animated | -0.8203 | 0.539 | -1.523 | 0.134 | -1.902 | 0.261 |
| horror | 1.0264 | 0.440 | 2.332 | 0.024 | 0.143 | 1.910 |
| log_budget | 0.7091 | 0.208 | 3.407 | 0.001 | 0.291 | 1.127 |
| const | 13.5768 | 0.688 | 19.727 | 0.000 | 12.195 | 14.959 |

| | | | |
|---|---|---|---|
| Omnibus: | 7.632 | Durbin-Watson: | 2.080 |
| Prob(Omnibus): | 0.022 | Jarque-Bera (JB): | 7.280 |
| Skew: | -0.626 | Prob(JB): | 0.0262 |
| Kurtosis: | 4.119 | Cond. No. | 153. |

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

Observations 2.1

- According to the results of linear regression above, the R-squared is 0.342 and the adjusted R-squared is 0.214.
- The variables budget, sequel, and horror are statistically significant at 0.10 confidence level based on the t-statistics and associated probabilities (P>|t|).

## 3. Linear Regression (2) including only significant "traditional" variables except the "buzz" variables.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.291 |
| Model: | OLS | Adj. R-squared: | 0.254 |
| Method: | Least Squares | F-statistic: | 7.929 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 0.000162 |
| Time: | 09:58:12 | Log-Likelihood: | -73.175 |
| No. Observations: | 62 | AIC: | 154.4 |
| Df Residuals: | 58 | BIC: | 162.9 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| log_budget | 0.6568 | 0.159 | 4.124 | 0.000 | 0.338 | 0.976 |
| sequel | 0.4994 | 0.298 | 1.676 | 0.099 | -0.097 | 1.096 |
| horror | 0.9908 | 0.385 | 2.574 | 0.013 | 0.220 | 1.761 |
| const | 13.8718 | 0.611 | 22.700 | 0.000 | 12.649 | 15.095 |

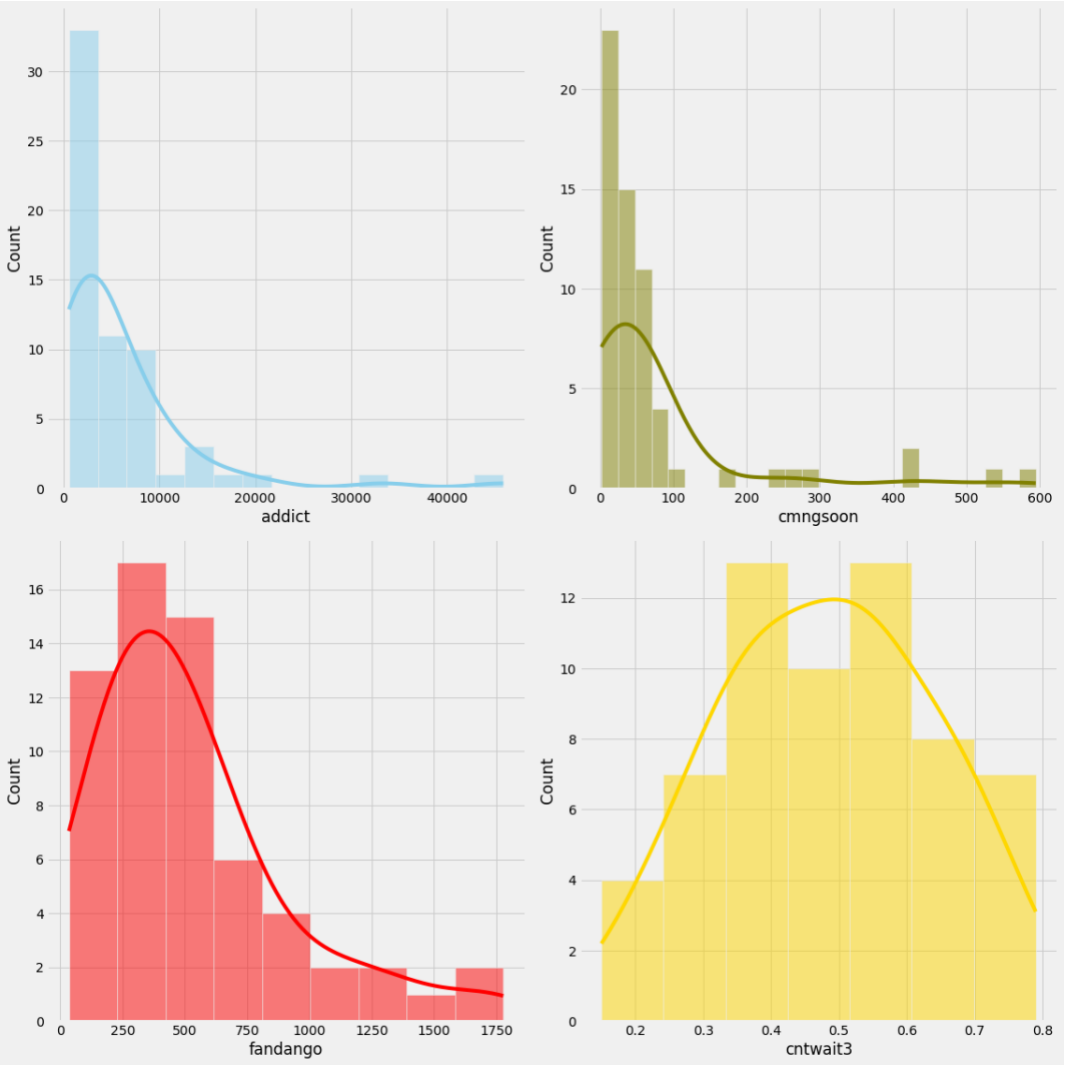
| | | | |
|---|---|---|---|
| Omnibus: | 17.577 | Durbin-Watson: | 2.065 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 28.607 |
| Skew: | -0.969 | Prob(JB): | 6.14e-07 |
| Kurtosis: | 5.704 | Cond. No. | 25.1 |

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

Observations 3.1

- The R-squared is 0.291 and the adjusted R-squared is 0.254
- The adjusted R-squared has increased and R-squared as decreased for the model using only significant variables as compared to model using all tradition variables except buzz variables.
- The increase in adjusted R-squared signifies that the model has improved when only significant variables were used.
- Also, all the variables are still significant at the 0.10 level.

## 4. Histograms of the four “buzz” variables




Observations 4.1

- Based on the histograms of the four "buzz" variables above, we could see the distributions of addict, cmngsoon, and fandango are right skewed.
- cantwait3 is approximately normally distributed.
- Thus, we need to do log transformation of these three variables.

# 5. Running a linear regression of box office revenues on all the independent variables

Linear Regression 3 – including all the independent variables

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.624 |
| Model: | OLS | Adj. R-squared: | 0.512 |
| Method: | Least Squares | F-statistic: | 5.576 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 3.77e-06 |
| Time: | 09:58:13 | Log-Likelihood: | -53.492 |
| No. Observations: | 62 | AIC: | 137.0 |
| Df Residuals: | 47 | BIC: | 168.9 |
| Df Model: | 14 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| G | 0.6375 | 0.600 | 1.063 | 0.293 | -0.569 | 1.844 |
| PG | 0.6054 | 0.316 | 1.913 | 0.062 | -0.031 | 1.242 |
| PG13 | 0.2171 | 0.220 | 0.986 | 0.329 | -0.226 | 0.660 |
| starpowr | 0.0012 | 0.013 | 0.089 | 0.930 | -0.026 | 0.028 |
| sequel | 0.4277 | 0.305 | 1.402 | 0.167 | -0.186 | 1.041 |
| action | -0.8419 | 0.301 | -2.801 | 0.007 | -1.447 | -0.237 |
| comedy | -0.0720 | 0.255 | -0.282 | 0.779 | -0.586 | 0.442 |
| animated | -0.8965 | 0.437 | -2.050 | 0.046 | -1.776 | -0.017 |
| horror | 0.3233 | 0.370 | 0.874 | 0.386 | -0.421 | 1.067 |
| cntwait3 | 2.5943 | 0.927 | 2.798 | 0.007 | 0.729 | 4.459 |
| log_budget | 0.2344 | 0.187 | 1.256 | 0.215 | -0.141 | 0.610 |
| log_addict | 0.2946 | 0.135 | 2.175 | 0.035 | 0.022 | 0.567 |
| log_cmngsoon | 0.0588 | 0.134 | 0.439 | 0.663 | -0.211 | 0.328 |
| log_fandango | 0.0274 | 0.117 | 0.235 | 0.815 | -0.207 | 0.262 |
| const | 11.4958 | 0.993 | 11.581 | 0.000 | 9.499 | 13.493 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.099 | Durbin-Watson: | 2.087 |
| Prob(Omnibus): | 0.577 | Jarque-Bera (JB): | 0.502 |
| Skew: | -0.150 | Prob(JB): | 0.778 |
| Kurtosis: | 3.322 | Cond. No. | 304. |

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

Observations 5.1

- R-squared is 0.624 and adjusted R-squared is 0.512.
- PG, action, animated, cntwait3 and log_addict is statistically significant at 0.10 confidence level based on the t-statistics and associated probabilities (P>|t|).

## 6. Running another linear regression using only the variables that were significant.

Linear Regression 4 – including all significant independent variables

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.558 |
| Model: | OLS | Adj. R-squared: | 0.519 |
| Method: | Least Squares | F-statistic: | 14.15 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 6.06e-09 |
| Time: | 09:58:13 | Log-Likelihood: | -58.512 |
| No. Observations: | 62 | AIC: | 129.0 |
| Df Residuals: | 56 | BIC: | 141.8 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| PG | 0.3464 | 0.228 | 1.519 | 0.134 | -0.110 | 0.803 |
| action | -0.6530 | 0.225 | -2.896 | 0.005 | -1.105 | -0.201 |
| animated | -0.5455 | 0.322 | -1.693 | 0.096 | -1.191 | 0.100 |
| cntwait3 | 3.7234 | 0.702 | 5.303 | 0.000 | 2.317 | 5.130 |
| log_addict | 0.2810 | 0.106 | 2.662 | 0.010 | 0.069 | 0.492 |
| const | 12.5048 | 0.751 | 16.644 | 0.000 | 11.000 | 14.010 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.995 | Durbin-Watson: | 1.992 |
| Prob(Omnibus): | 0.369 | Jarque-Bera (JB): | 1.257 |
| Skew: | -0.156 | Prob(JB): | 0.533 |
| Kurtosis: | 3.624 | Cond. No. | 80.9 |

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

Observations 6.1

- R-squared is 0.558, decreased from linear regression 3, it's obvious because we have reduced the number of independent variables.
- Adjusted R-squared is 0.519, increased from linear regression 3, that means that model has improved when only significant variables are used instead of all the variables.
- Except for PG, all the other variables are still significant at the 0.10 level.

# 7. Comparing the models developed so far.

We have run four linear regression models until now.

Observations 7.1

- Linear regression 1 and 2 have relatively low R-squared and adjusted R-squared values, which indicates the independent variables only explain 20% - 30% proportion of variation for the dependent variable -- box office revenues.

- Linear regression 3 and 4 have relatively higher R-squared and adjusted R-squared values. The models 3 and 4 are an improvement over the models 1 and 2.

- Between Model 3 and 4, Model 4 with 51.9% adjusted R-squared is the better model.

- Both R-squared and adjusted R-squared are utilized to measure how much percent of the change in the dependent variables are justified by the independent variables. However, adjusted R-squared, a modified version of R-squared, adds precision and reliability by considering the impact of additional independent variables that tend to skew the results of R-squared measurements. Therefore, we can say that Model 4 is the overall model yet.

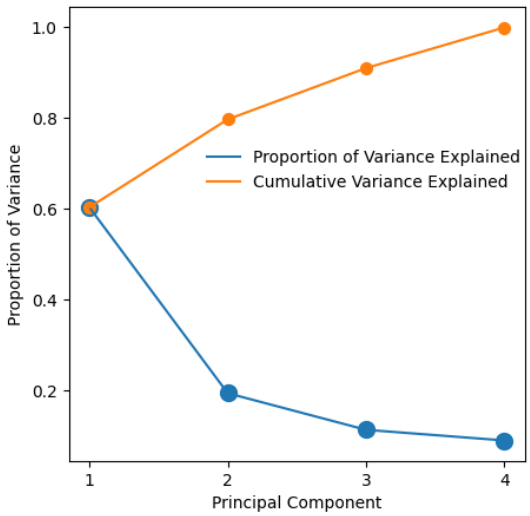
# 8. Applying Principal Component Analysis to just the 4 "buzz" variables

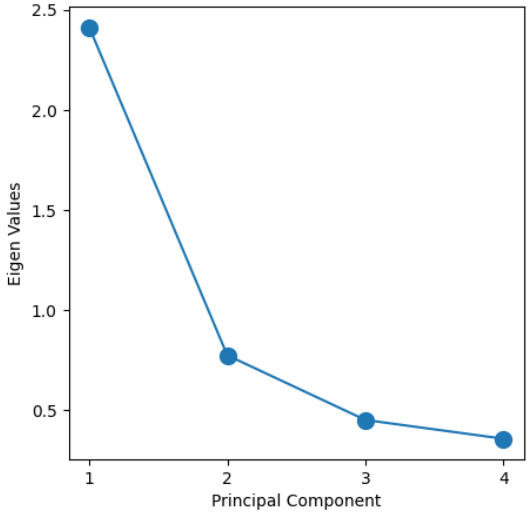
- After applying Principal Component Analysis to just the 4 "buzz" variables, we could get the eigen values associated with each component as follows: [2.41420026, 0.77519959, 0.45214886, 0.3584513].

- The fraction of the variance is explained by each component is as follows: [0.60355006, 0.1937999, 0.11303721, 0.08961282].

- The cumulative sum of the explained variance: [0.60355006, 0.79734996, 0.91038718, 1].

Plot of explained variance and Scree Plot




Observations 8.1

- Based on Kaiser's Rule and "explained variance" thresholds
    1. We select 1 principal component to explain 60% variance.
    2. We select 2 principal components to explain 70/80% variance.
    3. We select 3 principal components to explain 90% variance.

# 9. Running a linear regression using all the "traditional" independent variables and all 4 principal components

Linear Regression 5 – including all "traditional" independent variables except "buzz" variables and the four principal components generated based on four "buzz" variables

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.624 |
| Model: | OLS | Adj. R-squared: | 0.512 |
| Method: | Least Squares | F-statistic: | 5.576 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 3.77e-06 |
| Time: | 09:58:13 | Log-Likelihood: | -53.492 |
| No. Observations: | 62 | AIC: | 137.0 |
| Df Residuals: | 47 | BIC: | 168.9 |
| Df Model: | 14 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| G | 0.6375 | 0.600 | 1.063 | 0.293 | -0.569 | 1.844 |
| PG | 0.6054 | 0.316 | 1.913 | 0.062 | -0.031 | 1.242 |
| PG13 | 0.2171 | 0.220 | 0.986 | 0.329 | -0.226 | 0.660 |
| starpowr | 0.0012 | 0.013 | 0.089 | 0.930 | -0.026 | 0.028 |
| sequel | 0.4277 | 0.305 | 1.402 | 0.167 | -0.186 | 1.041 |
| action | -0.8419 | 0.301 | -2.801 | 0.007 | -1.447 | -0.237 |
| comedy | -0.0720 | 0.255 | -0.282 | 0.779 | -0.586 | 0.442 |
| animated | -0.8965 | 0.437 | -2.050 | 0.046 | -1.776 | -0.017 |
| horror | 0.3233 | 0.370 | 0.874 | 0.386 | -0.421 | 1.067 |
| log_budget | 0.2344 | 0.187 | 1.256 | 0.215 | -0.141 | 0.610 |
| PC1 | 0.4183 | 0.079 | 5.279 | 0.000 | 0.259 | 0.578 |
| PC2 | 0.1424 | 0.106 | 1.341 | 0.186 | -0.071 | 0.356 |
| PC3 | 0.0091 | 0.159 | 0.057 | 0.955 | -0.311 | 0.329 |
| PC4 | 0.2474 | 0.180 | 1.376 | 0.175 | -0.114 | 0.609 |
| const | 15.5293 | 0.675 | 22.991 | 0.000 | 14.170 | 16.888 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.099 | Durbin-Watson: | 2.087 |
| Prob(Omnibus): | 0.577 | Jarque-Bera (JB): | 0.502 |
| Skew: | -0.150 | Prob(JB): | 0.778 |
| Kurtosis: | 3.322 | Cond. No. | 179. |

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

Observations 9.1

- Based on the results of linear regression 5, the R-squared is 0.624 and the adjusted R-squared is 0.512.
- The variables PG, action, and animated are statistically significant at the 0.10 confidence level. As for the principal components, PC1 is significant.
- According to R-squared and adjusted R-squared, linear regression 5 is same as linear regression 3, which indicates that these four principal components have the same explanatory power as the four "buzz" variables.

# 10. Now running regressions using the number of principal components based on Kaiser's Rule and "explained variance."

linear regression 6 -- based on Kaiser's Rule and "explained variance" thresholds of 60% -- including PC1.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.589 |
| Model: | OLS | Adj. R-squared: | 0.498 |
| Method: | Least Squares | F-statistic: | 6.510 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 1.39e-06 |
| Time: | 09:58:13 | Log-Likelihood: | -56.278 |
| No. Observations: | 62 | AIC: | 136.6 |
| Df Residuals: | 50 | BIC: | 162.1 |
| Df Model: | 11 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| G | 0.3844 | 0.553 | 0.695 | 0.490 | -0.727 | 1.495 |
| PG | 0.5336 | 0.300 | 1.780 | 0.081 | -0.069 | 1.136 |
| PG13 | 0.2150 | 0.219 | 0.983 | 0.331 | -0.225 | 0.655 |
| starpowr | 0.0043 | 0.013 | 0.337 | 0.738 | -0.021 | 0.030 |
| sequel | 0.2751 | 0.273 | 1.007 | 0.319 | -0.274 | 0.824 |
| action | -0.8693 | 0.293 | -2.964 | 0.005 | -1.458 | -0.280 |
| comedy | -0.0162 | 0.256 | -0.063 | 0.950 | -0.531 | 0.498 |
| animated | -0.8332 | 0.430 | -1.937 | 0.058 | -1.697 | 0.031 |
| horror | 0.3746 | 0.371 | 1.009 | 0.318 | -0.371 | 1.120 |
| log_budget | 0.2609 | 0.185 | 1.408 | 0.165 | -0.111 | 0.633 |
| PC1 | 0.4291 | 0.078 | 5.473 | 0.000 | 0.272 | 0.587 |
| const | 15.4002 | 0.643 | 23.960 | 0.000 | 14.109 | 16.691 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.622 | Durbin-Watson: | 2.131 |
| Prob(Omnibus): | 0.444 | Jarque-Bera (JB): | 0.994 |
| Skew: | -0.282 | Prob(JB): | 0.608 |
| Kurtosis: | 3.260 | Cond. No. | 171. |

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

## Observations 10.1.1

- R-squared is 0.589 and adjusted R-squared is 0.498
- Adjusted R-squared has gone down when considering only 1 principal component, which was expected.

Linear Regression 7 -- based on Kaiser's Rule and "explained variance" thresholds of 70% -- including PC1, PC2

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.609 |
| Model: | OLS | Adj. R-squared: | 0.513 |
| Method: | Least Squares | F-statistic: | 6.357 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 1.27e-06 |
| Time: | 09:58:13 | Log-Likelihood: | -54.729 |
| No. Observations: | 62 | AIC: | 135.5 |
| Df Residuals: | 49 | BIC: | 163.1 |
| Df Model: | 12 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| G | 0.4933 | 0.549 | 0.898 | 0.374 | -0.611 | 1.597 |
| PG | 0.5807 | 0.297 | 1.956 | 0.056 | -0.016 | 1.177 |
| PG13 | 0.2511 | 0.217 | 1.158 | 0.252 | -0.185 | 0.687 |
| starpowr | 0.0063 | 0.013 | 0.495 | 0.623 | -0.019 | 0.032 |
| sequel | 0.3543 | 0.274 | 1.295 | 0.202 | -0.196 | 0.904 |
| action | -0.9138 | 0.290 | -3.147 | 0.003 | -1.497 | -0.330 |
| comedy | -0.0224 | 0.252 | -0.089 | 0.929 | -0.530 | 0.485 |
| animated | -0.8304 | 0.424 | -1.959 | 0.056 | -1.682 | 0.021 |
| horror | 0.3254 | 0.367 | 0.887 | 0.379 | -0.412 | 1.063 |
| log_budget | 0.2830 | 0.183 | 1.546 | 0.129 | -0.085 | 0.651 |
| PC1 | 0.4219 | 0.077 | 5.454 | 0.000 | 0.266 | 0.577 |
| PC2 | 0.1637 | 0.103 | 1.585 | 0.119 | -0.044 | 0.371 |
| const | 15.2565 | 0.640 | 23.849 | 0.000 | 13.971 | 16.542 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.798 | Durbin-Watson: | 2.070 |
| Prob(Omnibus): | 0.407 | Jarque-Bera (JB): | 1.232 |
| Skew: | -0.334 | Prob(JB): | 0.540 |
| Kurtosis: | 3.173 | Cond. No. | 172. |

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

Observations 10.1.2

- R-squared is 0.609 and adjusted R-squared is 0.513
- Adjusted R-squared has improved when considering 2 principal components PC1 and PC2

Linear regression 8 -- based on Kaiser's Rule and "explained variance" thresholds of 80% or 90% -- including PC1, PC2, PC3

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.609 |
| Model: | OLS | Adj. R-squared: | 0.503 |
| Method: | Least Squares | F-statistic: | 5.752 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 3.32e-06 |
| Time: | 10:31:16 | Log-Likelihood: | -54.715 |
| No. Observations: | 62 | AIC: | 137.4 |
| Df Residuals: | 48 | BIC: | 167.2 |
| Df Model: | 13 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| G | 0.4635 | 0.592 | 0.783 | 0.437 | -0.726 | 1.653 |
| PG | 0.5654 | 0.318 | 1.778 | 0.082 | -0.074 | 1.205 |
| PG13 | 0.2467 | 0.221 | 1.115 | 0.270 | -0.198 | 0.691 |
| starpowr | 0.0059 | 0.013 | 0.448 | 0.656 | -0.021 | 0.032 |
| sequel | 0.3372 | 0.301 | 1.121 | 0.268 | -0.267 | 0.942 |
| action | -0.9212 | 0.298 | -3.094 | 0.003 | -1.520 | -0.323 |
| comedy | -0.0214 | 0.255 | -0.084 | 0.933 | -0.534 | 0.491 |
| animated | -0.8174 | 0.437 | -1.868 | 0.068 | -1.697 | 0.062 |
| horror | 0.3317 | 0.373 | 0.889 | 0.379 | -0.419 | 1.082 |
| log_budget | 0.2831 | 0.185 | 1.531 | 0.132 | -0.089 | 0.655 |
| PC1 | 0.4243 | 0.080 | 5.313 | 0.000 | 0.264 | 0.585 |
| PC2 | 0.1607 | 0.106 | 1.512 | 0.137 | -0.053 | 0.375 |
| PC3 | 0.0232 | 0.160 | 0.145 | 0.885 | -0.299 | 0.345 |
| const | 15.2721 | 0.655 | 23.314 | 0.000 | 13.955 | 16.589 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.768 | Durbin-Watson: | 2.068 |
| Prob(Omnibus): | 0.413 | Jarque-Bera (JB): | 1.219 |
| Skew: | -0.334 | Prob(JB): | 0.544 |
| Kurtosis: | 3.158 | Cond. No. | 173. |

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

## Observations 10.1.3

- R-squared is 0.609 and adjusted R-squared is 0.503
- Adjusted R-squared has decreased when considering 3 principal components PC1 and PC2 and PC3.
- Based on the results of all linear regression models involving the principal components, linear regression 7 which contains the first two principal components -- PC1 and PC2 based on Kaiser's Rule and "explained variance" threshold of 70% has the highest adjusted R-squared compared to the others.
- We found this a little surprising that PC1 PC2 and PC3 together explains maximum variance but the model 8 has not improved over model 7.

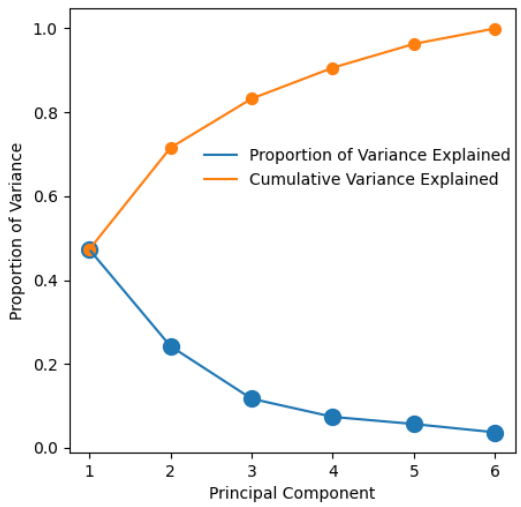
## 11. Now applying Principal Component Analysis to the 4 "buzz" variables and the other continuous variables (budget and starpowr).

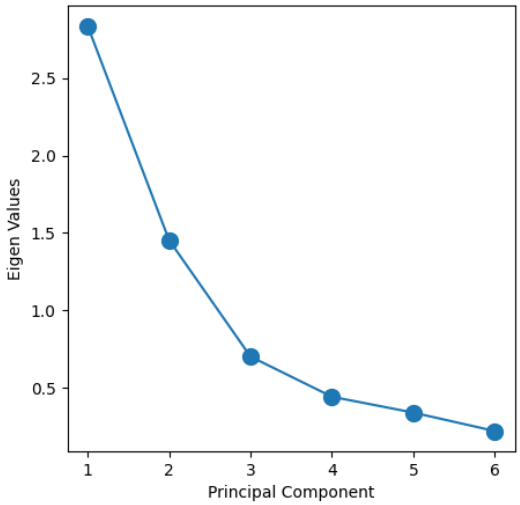
- After applying Principal Component Analysis on the four "buzz" and two continuous variables, we get the eigen values associated with each component as follows: [2.82694346, 1.4072405, 0.71089336, 0.4760487, 0.36108278, 0.21779121].

- The fraction of the variance is explained by each component is as follows: [0.47115724, 0.23454008, 0.11848223, 0.07934145, 0.06018046, 0.03629853].

- Cumulative sum of the explained variance [0.47303897 0.71544342 0.83249711 0.90632927 0.96307878 1.]

Plot of explained variance and Scree Plot




Observations 11.1

- Based on Kaiser's Rule and "explained variance" thresholds
    1. We select 2 principal components to explain 60/70% variance.
    2. We select 3 principal components to explain 80% variance.
    3. We select 4 principal components to explain 90% variance.

## 12. Running regressions using then number of principal components based on Kaiser's Rule and "explained variance" thresholds of 60%, 70%, 80% and 90%.

Linear regression 9 -- based on Kaiser's Rule and "explained variance" thresholds of 60% or 70% -- including PC1, PC2

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.590 |
| Model: | OLS | Adj. R-squared: | 0.509 |
| Method: | Least Squares | F-statistic: | 7.327 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 4.70e-07 |
| Time: | 01:18:45 | Log-Likelihood: | -56.220 |
| No. Observations: | 62 | AIC: | 134.4 |
| Df Residuals: | 51 | BIC: | 157.8 |
| Df Model: | 10 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| PC1 | -0.4596 | 0.061 | -7.586 | 0.000 | -0.581 | -0.338 |
| PC2 | 0.1107 | 0.094 | 1.175 | 0.245 | -0.078 | 0.300 |
| G | 0.3972 | 0.544 | 0.731 | 0.468 | -0.694 | 1.488 |
| PG | 0.5528 | 0.298 | 1.852 | 0.070 | -0.046 | 1.152 |
| PG13 | 0.2317 | 0.219 | 1.058 | 0.295 | -0.208 | 0.671 |
| sequel | 0.2697 | 0.268 | 1.008 | 0.318 | -0.267 | 0.807 |
| action | -0.8831 | 0.291 | -3.031 | 0.004 | -1.468 | -0.298 |
| comedy | -0.0196 | 0.253 | -0.077 | 0.939 | -0.528 | 0.488 |
| animated | -0.8129 | 0.424 | -1.917 | 0.061 | -1.664 | 0.038 |
| horror | 0.3219 | 0.363 | 0.887 | 0.379 | -0.407 | 1.051 |
| const | 16.4428 | 0.215 | 76.618 | 0.000 | 16.012 | 16.874 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.552 | Durbin-Watson: | 2.136 |
| Prob(Omnibus): | 0.460 | Jarque-Bera (JB): | 0.952 |
| Skew: | -0.280 | Prob(JB): | 0.621 |
| Kurtosis: | 3.236 | Cond. No. | 12.7 |

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

### Observations 12.1.1

- R-squared is 0.590 and adjusted R-squared is 0.509

Linear regression 10 -- based on Kaiser's Rule and "explained variance" thresholds of 80% -- including PC1, PC2, PC3

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.613 |
| Model: | OLS | Adj. R-squared: | 0.528 |
| Method: | Least Squares | F-statistic: | 7.198 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 3.64e-07 |
| Time: | 01:21:38 | Log-Likelihood: | -54.407 |
| No. Observations: | 62 | AIC: | 132.8 |
| Df Residuals: | 50 | BIC: | 158.3 |
| Df Model: | 11 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| PC1 | -0.4542 | 0.060 | -7.632 | 0.000 | -0.574 | -0.335 |
| PC2 | 0.1038 | 0.092 | 1.123 | 0.267 | -0.082 | 0.290 |
| PC3 | 0.1768 | 0.102 | 1.735 | 0.089 | -0.028 | 0.381 |
| G | 0.5136 | 0.537 | 0.956 | 0.344 | -0.566 | 1.593 |
| PG | 0.5639 | 0.293 | 1.926 | 0.060 | -0.024 | 1.152 |
| PG13 | 0.2335 | 0.215 | 1.087 | 0.282 | -0.198 | 0.665 |
| sequel | 0.3638 | 0.268 | 1.358 | 0.181 | -0.174 | 0.902 |
| action | -0.8945 | 0.286 | -3.130 | 0.003 | -1.469 | -0.320 |
| comedy | -0.0340 | 0.248 | -0.137 | 0.892 | -0.533 | 0.465 |
| animated | -0.8266 | 0.416 | -1.988 | 0.052 | -1.662 | 0.009 |
| horror | 0.3377 | 0.356 | 0.948 | 0.348 | -0.378 | 1.053 |
| const | 16.4289 | 0.211 | 77.991 | 0.000 | 16.006 | 16.852 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.083 | Durbin-Watson: | 2.027 |
| Prob(Omnibus): | 0.582 | Jarque-Bera (JB): | 0.693 |
| Skew: | -0.255 | Prob(JB): | 0.707 |
| Kurtosis: | 3.087 | Cond. No. | 12.8 |

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

## Observations 12.1.2

- R-squared is 0.613 and adjusted R-squared is 0.528
- Adjusted R-squared has improved when considering 3 principal components PC1, PC2 and PC3

Linear regression 11 -- based on Kaiser's Rule and "explained variance" thresholds of 90% -- including PC1, PC2, PC3, PC4

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log_box | R-squared: | 0.614 |
| Model: | OLS | Adj. R-squared: | 0.519 |
| Method: | Least Squares | F-statistic: | 6.491 |
| Date: | Thu, 08 Sep 2022 | Prob (F-statistic): | 9.69e-07 |
| Time: | 01:22:44 | Log-Likelihood: | -54.333 |
| No. Observations: | 62 | AIC: | 134.7 |
| Df Residuals: | 49 | BIC: | 162.3 |
| Df Model: | 12 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| PC1 | -0.4480 | 0.063 | -7.141 | 0.000 | -0.574 | -0.322 |
| PC2 | 0.1023 | 0.093 | 1.095 | 0.279 | -0.085 | 0.290 |
| PC3 | 0.1814 | 0.104 | 1.750 | 0.086 | -0.027 | 0.390 |
| PC4 | 0.0562 | 0.165 | 0.340 | 0.735 | -0.276 | 0.388 |
| G | 0.5957 | 0.593 | 1.004 | 0.320 | -0.597 | 1.788 |
| PG | 0.6003 | 0.314 | 1.911 | 0.062 | -0.031 | 1.231 |
| PG13 | 0.2405 | 0.218 | 1.105 | 0.275 | -0.197 | 0.678 |
| sequel | 0.4099 | 0.302 | 1.356 | 0.181 | -0.198 | 1.018 |
| action | -0.8706 | 0.297 | -2.933 | 0.005 | -1.467 | -0.274 |
| comedy | -0.0407 | 0.251 | -0.162 | 0.872 | -0.546 | 0.464 |
| animated | -0.8602 | 0.431 | -1.996 | 0.052 | -1.726 | 0.006 |
| horror | 0.3227 | 0.362 | 0.891 | 0.377 | -0.405 | 1.050 |
| const | 16.4091 | 0.220 | 74.466 | 0.000 | 15.966 | 16.852 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.997 | Durbin-Watson: | 2.028 |
| Prob(Omnibus): | 0.607 | Jarque-Bera (JB): | 0.596 |
| Skew: | -0.234 | Prob(JB): | 0.742 |
| Kurtosis: | 3.108 | Cond. No. | 14.1 |

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

## Observations 12.1.3

- R-squared is 0.614 and adjusted R-squared is 0.519
- Adjusted R-squared has decreased when considering 4 principal components PC1, PC2,PC3 and PC4
- Based on the results of linear regression 9, 10, and 11, we could see the adjusted R-squared is best for Model 10 with 52.8% variance being explained in the target variable box office revenue.

## 13. Conclusion

- By comparing the linear regressions without "buzz" variables, including linear regression 1 and 2, to the linear regressions with "buzz" variables, including linear regression 3 and 4, we could detect that the "buzz" variables help build a better model since adding them increase the R-squared and adjusted R-squared approximately by 0.2 - 0.3.

- By comparing the linear regressions without "buzz" variables, including linear regression 1 and 2, to the linear regressions with principal components based on "buzz" variables, including linear regression 5, 6, 7 and 8, we could detect that PCA also helps build a better model and adding them also could increase the R-squared and adjusted R-squared approximately by 0.2 - 0.3.

- In other words, "buzz" variables and PCA helped us building superior models.

## 14. Key takeaway and surprises

Surprises along the way:

- When we did PCA to reduce dimensions in question 12, we had 6 variables, the first 4 principal components explain the mor than 90% of variance in the data.

- Now when we started making models using First 2 components and then gradually added more principal components to the model, the expectation was that as we add more components, model's prediction ability would improve. But what we observed that from model 9 to Model 10 adjusted R squared did improve but then dropped for model 11. It was surprising because first principal components explain 80% of the variance as compared to first 4 components explain more than 90% of the variance.

- What we have learned is that, we just can't go only by the thumb rule and select the principal components based on the cumulative variance explained by them. By using the exhaustive approach and building multiple models we were able to narrow down to a better model.

Managerial takeaways:

- The data from addict, cmngsoon, fandango, cntwait3 provide additional predictive information to predict the box office revenues, so these "buzz" variables are important variables in predicting box office revenues. We can also look into ways to creating the buzz of the movie on these websites could help in increasing box office revenues.

- Since we did a PCA, we don't have the exact recipe for combination of 4 buzz variable, starpowr and movie budget. But we can definitely say that these variables play a role (pun intended 😊) in predicting the box office revenue

- PG, action, animated and comedy are significant variables from the best model.

- There is a scope for building a better model here. We need more data points as well as we need to capture more features to better predict the box office revenue.

- The examples of more features that can be captured to improve model can be movie distribution (region by region release) data, movie marketing data for all regions. Besides genre we can also look analyze scripts of the movies. We have also not considered features like director power, technician power, original sound score etc.