Modeling For Business Analytics BUAN 6383

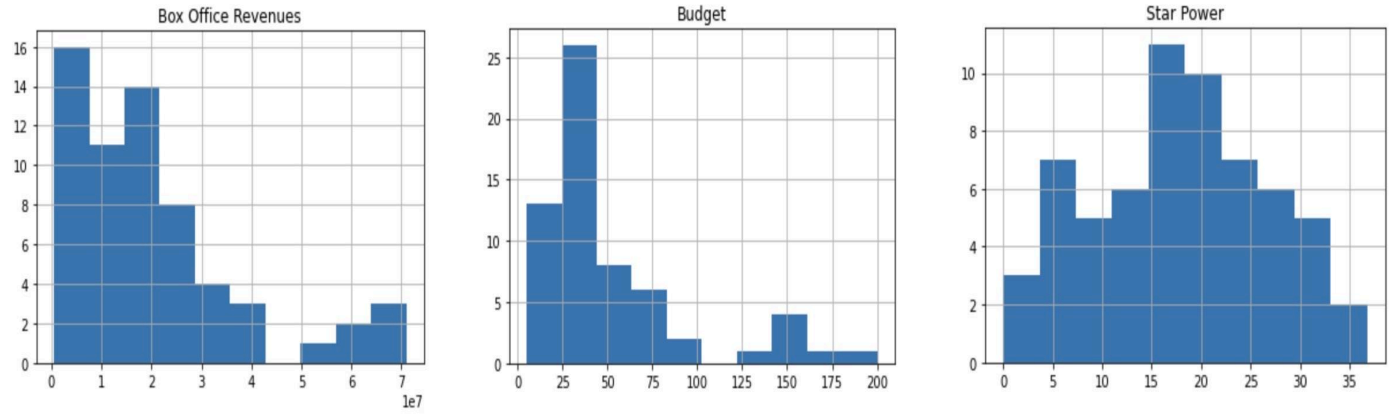
Mini-Project:
Principal Component Analysis for Exploring the ability of Buzz-variables in Predicting Box-Office Collections

Compilation of Results

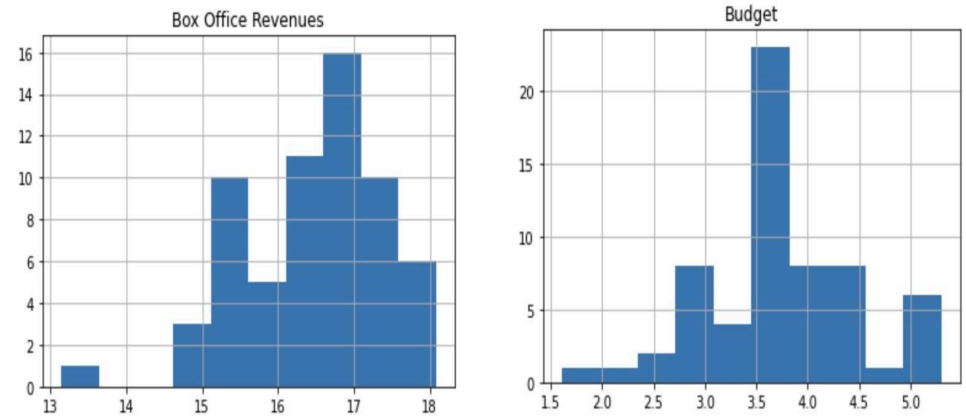
Data Dictionary:

box	domestic opening weekend box office revenues ($)
G, PG, PG13	binary variables indicating MPAA rating code (of all 3 are 0, the movie is rated R)
budget	production budget (in millions of $)
starpowr	star power rating based on the Forbes 2009 Star Currency list (range: 0 to 10)
sequel	binary (1 → sequel; 0 → if not)
action, comedy, animated, horror	binary variables indicating movie genre (of all 4 are 0, the movie genre was drama)
addict	number of trailer views at traileraddict.com
cmngsoon	number of message board comments at comingsoon.net
fandango	sum of "can't wait" and "don't care" votes at fandango.com
cntwait3	percent of "can't wait" votes at fandango.com

The histograms of the continuous variables, before running a Log Transformation are as follows:

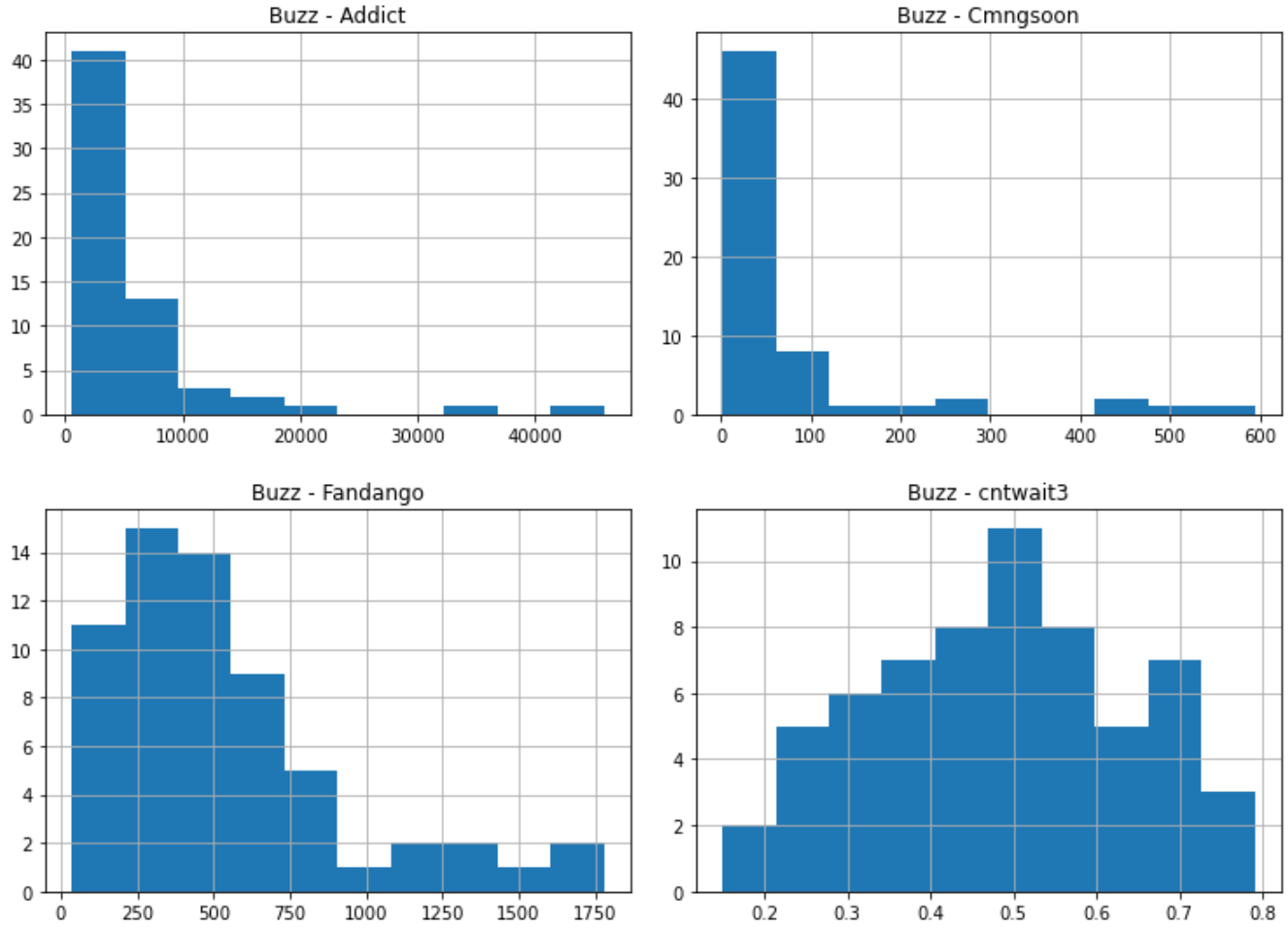


From the above 3 histograms of box, budget and starpower we can see that box and budget are right skewed. So we apply log transformation to the box and budget variables to normalize it. Starpower, however,already has a normal distribution.

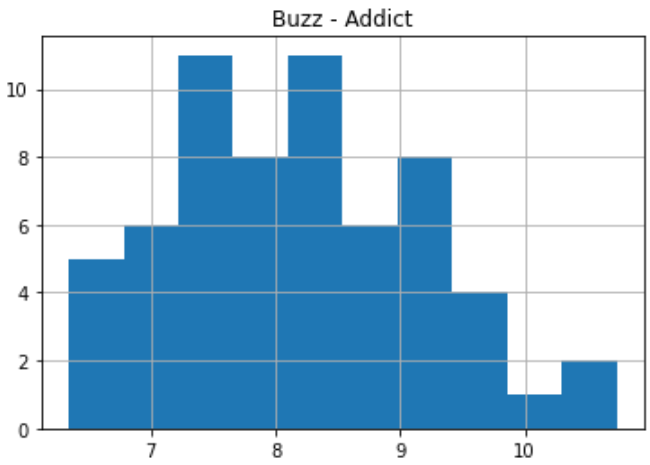


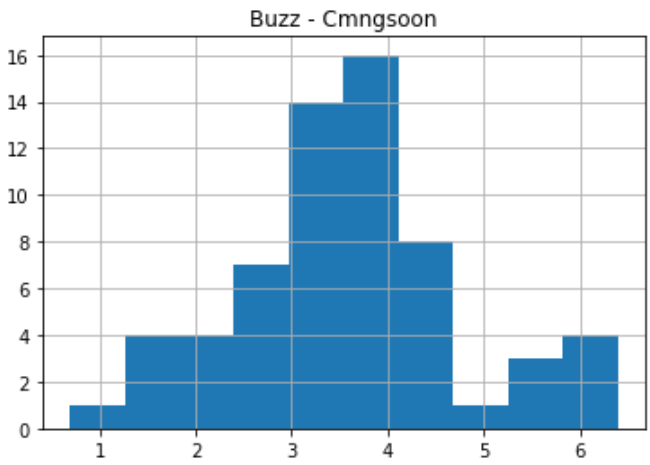
After running a log transformation on skewed variables., we can see that the variables are normalized.

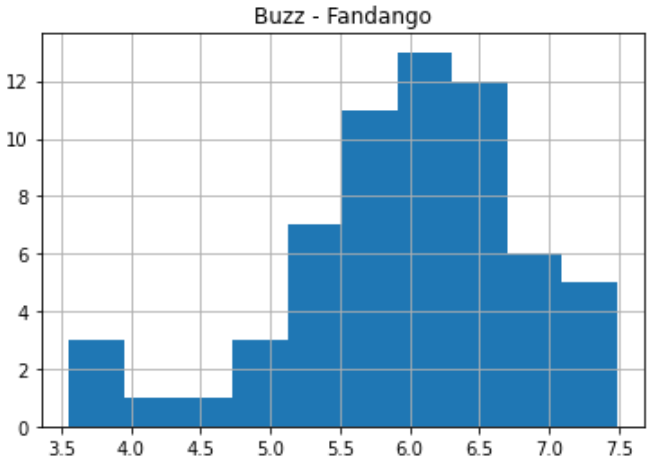
The histograms of the buzz variables, before running a Log Transformation are as follows:



From the above 4 histograms of addict, cmngsoon, fandango and cntwait3 variables we can see that addict, cmngsoon and fandango variables appear right skewed while cntwait3 variable appears normal in distribution. So we apply log transformation to the addict, cmngsoon and fandango variables to normalize them.







We developed 4 models initially:

1. based on all traditional variables
2. Based on only significant traditional variables
3. Based on all independent variables including buzz variables
4. Based on only significant independent variables including buzz variables which are significant

Model	**Variables**	**Significant Variables (P value < 0.1)**	R^2	**Adjusted** R^2
1	Based on all traditional variables: G, PG, PG13, budget(log), starpowr, sequel, action, comedy, animated, horror Note: R and Drama are implicit variables	budget (log) - 0.001 sequel - 0.058 horror - 0.024	0.342	0.214

2	Based on only significant traditional variables from Model 1: budget(log) , sequel, horror	budget (log) - 0.000 sequel - 0.099 horror - 0.013 (All 3 still significant)	0.291	0.254
3	Based on all independent variables including buzz variables: G, PG, PG13, budget(log), starpowr, sequel, action, comedy, animated, horror, addict(log), cmngsoon(log), fandango(log), cntwait Note: R and Drama are implicit variables	PG - 0.062 action - 0.007 animated - 0.046 addict(log) - 0.035 cntwait - 0.007 (budget, sequel and horror do not appear significant here)	0.624	0.512
4	Based on only significant variables from Model 3: PG, action, animated, addict(log), cntwait	action - 0.005 animated - 0.096 addict(log) - 0.010 cntwait - 0.000 (PG does not appear significant here)	0.558	0.519

When we compare the four models, the last model **(Model 4)** gives us the **best 'adjusted-R^2 -value (0.519**) and therefore appears to be the best model for predicting box-office revenues of the given dataset

Next, we develop Principal Components using only the traditional variables:

Principal Component	**EigenValue**	**Explained Variance Ratio**	**Cumulative Sum of Explained Variance (%)**
PC1	2.41420026	0.60355006 (60.35%)	60.35 %
PC2	0.77519959	0.1937999 (19.37%)	79.73 %
PC3	0.45214886	0.11303721 (11.30%)	91.03 %
PC4	0.3584513	0.08961282 (8.9%)	100 %

Selection Criteria	Principal Components selected
Kaiser's Rule (Eigenvalue > 1)	Only PC1 (Eigenvalue = 2.41420026)
Explained variance threshold → 60%	Only PC1
Explained variance threshold → 70%	PC1 and PC2
Explained variance threshold → 80%	PC1 and PC2 (rounding the cumulative variance of 79.73% to 80%)
Explained variance threshold → 90%	PC1 , PC2 and PC3

Comparing the new models developed using Principal Components

Model	Variables	Significant Variables (P value < 0.1)	R^2	Adjusted R^2
5	Based on all traditional variables and all 4 principal components: G, PG, PG13, budget(log), starpowr, sequel, action, comedy, animated, horror, PC1, PC2, PC3, PC4	PG - 0.062 action - 0.007 animated - 0.046 PC1 - 0.000	0.624	0.512

PC1 appears to be a significant principal component
The model (Model 5) presents $\mathbf{R^2}$ and adjusted $\mathbf{R^2}$ values identical to Model 3 as shown under.

3	Based on all independent variables including buzz variables: G, PG, PG13, budget(log), starpowr, sequel, action, comedy, animated, horror, addict(log), cmngsoon(log), fandango(log), cntwait Note: R and Drama are implicit variables	PG - 0.062 action - 0.007 animated - 0.046 addict (log) - 0.035 cntwait - 0.007 (budget, sequel and horror do not appear significant here)	0.624	0.512

While PG, action and animated continue to remain significant variables in Model 5 as in Model 3, the significant variables addict and cntwait are replaced by PC1, meaning PC1 reflects the significance of addict and cntwait variables by itself.
However, Model 4 with **adjusted-$\mathbf{R^2}$-value of 0.519** continue to remain the best model so far

Next comparing the regressions using the number of principal components based on Kaiser's Rule and "explained variance" thresholds of 60%, 70%, 80% and 90%

Model	**Criteria**	**Variables**	**Significant Variables at 0.10 level of Significance** **P-Value < 0.10**	R^2	**Adj.** R^2
6	1. Selecting Principal Components based on kaiser rule i.e., eigenvalues > 1 also 2. explained variance threshold of 60%.	G, PG, PG13, budget, starpowr, sequel, action, comedy, animated, horror, PC1	PG Action Animated PC1	0.589	0.498
7	Selecting Principal Components based explained variance threshold of 70%	G, PG, PG13, budget, starpowr, sequel, action, comedy, animated, horror, PC1, PC2	PG, action, animated, PC1	0.609	0.513
8	Selecting Principal Components based explained variance threshold of 80%	G, PG, PG13, budget, starpowr, sequel, action, comedy, animated, horror, PC1, PC2, PC3	PG, action, animated, PC1	0.609	0.503
9	All traditional Variables and all Principal Components	G, PG, PG13, budget, starpowr, sequel, action, comedy, animated, horror, PC1, PC2, PC3, PC4	PG, action, animated, PC1	0.624	0.512

Comparing all 4 models (including the one which has all 4 Principal Components) we can say that **Model 7** which uses 2 Principal components based on **explained variance threshold of 70%** is the model that can be recommended.

This model has the best **adjusted-R^2 -value of 0.513** among the models 6 - 9 and hence appears to be the best model that can predict the box office revenues of the given data set

Next, applying Principal Component Analysis to the 4 “buzz” variables and the other continuous variables (budget and starpowr).

Principal Component	EigenValue	Explained Variance Ratio	Cumulative Sum of Explained Variance Ratio
PC1	2.83823382	0.47303897	0.47303897
PC2	1.45442671	0.24240445	0.71544342
PC3	0.70232212	0.11705369	0.83249711
PC4	0.44299297	0.07383216	0.90632927
PC5	0.34049709	0.05674951	0.96307878
PC6	0.2215273	0.03692122	1.0

Selecting Principal Components

Selection Criteria	Principal Components selected
Kaiser’s Rule (Eigenvalue > 1)	PC1, PC2
Explained variance threshold → 60%	PC1, PC2
Explained variance threshold → 70%	PC1, PC2
Explained variance threshold → 80%	PC1, PC2, PC3
Explained variance threshold → 90%	PC1, PC2, PC3, PC4

Model	Criteria	Variables	Significant Variables at 0.10 level of Significance P-Value < 0.10	R^2	Adj. R^2
10	1. Selecting Principal Components based on kaiser rule i.e., eigenvalues > 1 also	G, PG, PG13, sequel, action, comedy, animated, horror, PC1, PC2	PG - 0.069 action - 0.004 animated - 0.060 PC1 - 0.053	0.594	0.495

	2. explained variance threshold of 60% as well as 70%				
11	Selecting Principal Components based explained variance threshold of 80%	G, PG, PG13, sequel, action, comedy, animated, horror, PC1, PC2, PC3	PG - 0.079 action - 0.003 animated - 0.057 PC1 - 0.009 PC3 - 0.075	0.620	0.518
12	Selecting Principal Components based explained variance threshold of 90%	G, PG, PG13, sequel, action, comedy, animated, horror, PC1, PC2, PC3, PC4	PG - 0.062 action - 0.007 animated - 0.046 PC1 - 0.008 PC3 - 0.059	0.624	0.512

Comparing all 3 models (model 10 - model 12) we can say that **Model 11** which uses 2 Principal components **PC1 and PC3** based on **explained variance threshold of 80%** is the model that can be recommended.
This model has the best **adjusted-R^2 -value of 0.518** among the models 10 - 12 and hence appears to be the best model among these that can predict the box office revenues of the given data set.

Comparing the results of Model 1 which excluded buzz variables (modeled only using traditional variables) and Model 3 which included buzz variables, we find that adjusted R^2 values doubled from Model 1 to Model 3. This further improved in Model 4 where we used only significant variables from Model 3

Model	**Variables**	**Significant Variables (P value < 0.1)**	R^2	**Adjusted** R^2
1	Based on all traditional variables: G, PG, PG13, budget(log), starpowr, sequel, action, comedy, animated, horror Note: R and Drama are implicit variables	budget (log) - 0.001 sequel - 0.058 horror - 0.024	0.342	0.214
3	Based on all independent variables including buzz variables: G, PG, PG13, budget(log), starpowr, sequel, action, comedy, animated, horror, addict(log), cmngsoon(log), fandango(log), cntwait Note: R and Drama are implicit variables	PG - 0.062 action - 0.007 animated - 0.046 addict(log) - 0.035 cntwait - 0.007 (budget, sequel and horror do not appear significant here)	0.624	0.512
4	Based on only significant variables from Model 3: PG, action, animated, addict(log), cntwait	action - 0.005 animated - 0.096 addict(log) - 0.010 cntwait - 0.000 (PG does not appear significant here)	0.558	0.519

As a result, we can conclude that the inclusion of buzz variables has led to the marked enhancement of our models by improving the explained variance by over 30% after adjusting for number of parameters (adjusted R2)

Now to understand the effect of PCA on predicting the box office revenues of the data set, we can compare the best PCA model ie Model 11, with the best linear regression model without PCA i.e., Model 4

Model	Variables	Significant Variables (P value < 0.1)	R^2	Adjusted R^2
11	G, PG, PG13, sequel, action, comedy, animated, horror, PC1, PC2, PC3	PG - 0.079 action - 0.003 animated - 0.057 PC1 - 0.009 PC3 - 0.075	0.620	0.518

The 2 models are almost identical in their adjusted R^2 values with PCA Model (Model 11) trailing Linear Regression Model by 0.001 (0.518 vs 0.519). **One could conclude that PCA has given us an equally good model as a linear regression model without PCA.**

The techniques employed in these analyses showcased the influence of PCA on regression models. We were intrigued to see the impact of varying thresholds for Principal Component Selection. Our findings indicated that the optimal model was not always the one with the low thresholds or the one that included the most number of PCs (high thresholds), but rather was found to be somewhere in between. Therefore, it was crucial to experiment with different thresholds to arrive at the optimal models.

Also, We initially thought that budget and starpower would be very important factors for Box Office revenue, however these variables were not significant in all models except for Model 1 in which budget was significant. This was surprising as well.

Managerial takeaways:

Buzz Variables are one of the key factors in the correlation between internet activity and box office revenues. It shows the importance of generating buzz and attention online. Companies should aim to create online conversations and engage with their audience to increase their visibility and ultimately, their performance.