EXPLORATORY DATA ANALYSIS

Analysis Report

Ravi Teja Buddabathuni
Date:10/28/2016

Bowling Green State University

Executive Summary

This report summarizes the data analysis results associated with the popularity of the movie in Facebook based on the genres. The purpose of this report is to document the analysis results of the genre wise popularity of the movie based on its facebook likes using EDA methods to identify the underlying pattern in the dataset.

Firstly, data set is taken from the IMDB website form this link. Later the data set is being cleaned by using MS Excel, Here the priority is taken for the movies released in 2016 and 2015 in USA in English language, and then the below rows were considered:

- Director_Name
- Movie_Name
- Genres
- IMDB_Score
- Facebook_Likes
- Critic_Reviews
- User_Votes
- User_Reviews
- Year

The development of the sampling protocol, including both the initial recommended design and final implemented sampling strategy are discussed in Section 2. The initial Stratified Random sampling design was developed using a Neyman allocation scheme. After presenting this design to the client, a refined GIS analysis was performed and more accurate available sampling areas for each school were calculated. These calculations were used to revise the second-stage random sampling scheme. Additionally, two extra properties were added to the sampling design (one nursery located within 2 Km of the factory and one previously overlooked park) and 12 additional sampling locations were selected along the factory perimeter. After these refinements, the final sampling plan contained 361 sampling locations from 69 distinct non-factory properties (and the factory perimeter). The basic univariate statistics that summarize the contamination data associated with the analyzed metals (for all 360 topsoil samples) are given in Section 3. A total of seven metal concentration measurements were made on each topsoil sample; the metals analyzed in this study include Arsenic (As), Cadmium (Cd), Chromium (Cr), Copper (Cu), Nickel (Ni), Lead (Pb), and Zinc (Zn). The univariate statistics summarize both the raw and natural log transformed metal data, where the transformed data is defined as Y = ln(X+1). The histograms and quantile plots of each log transformed metal data appear to be approximately symmetric (but in some cases also moderately heavy-tailed). Section 4 presents the analysis of the sampling depth effect, based on the 43 sites were topsoil samples were acquired from two sampling depths.. Paired t-tests and sign-

INTRODUCTION:

Meet the Data

Data: US Movies publicity in Facebook from the years 2016 and 2015 from IMDB website

Source: https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

One of the major social networking sites used in the United States for movie publicity is Facebook.

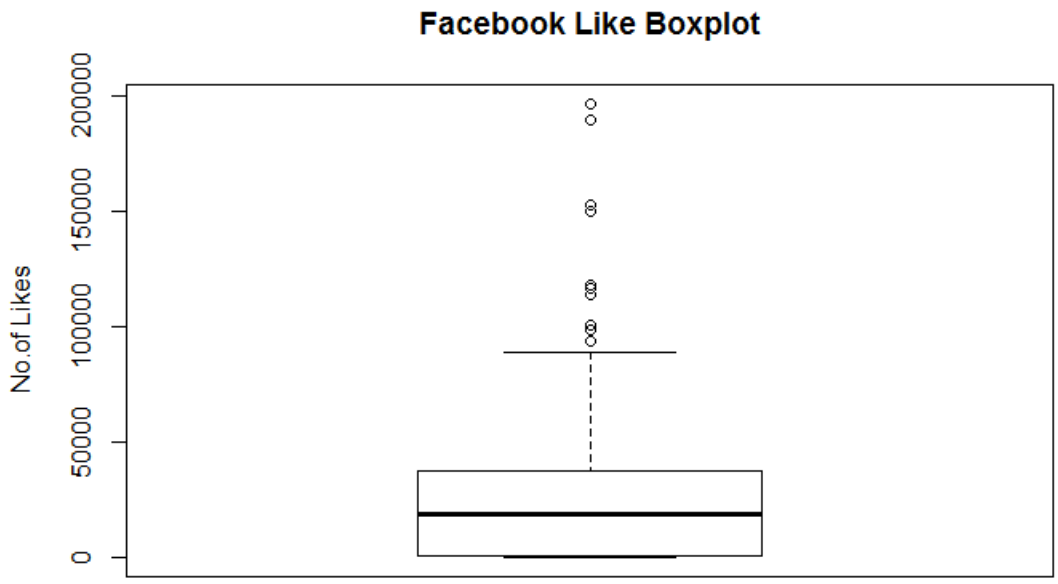
This platform is able to reach many number of audience in an affordable way.

The table below shows the sample data from the data set. Here the main focus is on Genres and Facebook_likes.

Director_Name	Movie_Name	Genres	IMDB_Score	Facebook_Likes	Critic_Reviews	User_Votes	User_Reviews	Year
Joss Whedon	Avengers: Age of Ultron	Action	7.5	118000	635	462669	1117	2015
Zack Snyder	Batman v Superman: Dawn of Justice	Action	6.9	197000	673	371639	3018	2016
Anthony Russo	Captain America: Civil War	Action	8.2	72000	516	272670	1022	2016
Colin Trevorrow	Jurassic World	Action	7	150000	644	418214	1290	2015
James Wan	Furious 7	Action	7.2	94000	424	278232	657	2015
Peter Sohn	The Good Dinosaur	Advent ure	6.8	20000	298	62836	345	2015
Justin Lin	Star Trek Beyond	Action	7.5	30000	322	53607	432	2016
Lana Wachowski	Jupiter Ascending	Action	5.4	44000	384	139593	720	2015
David Yates	The Legend of Tarzan	Action	6.6	29000	248	42372	239	2016

Here the dataset contains 177 rows, so instead of considering the stemplot for identifying shape, spread, skewness, it would be better to prefer boxplot as we have more than 50 values.

Boxplot for Facebook Likes:



Here from the above boxplot we can say that the median is almost equally distributed from the upper fourth and lower fourth

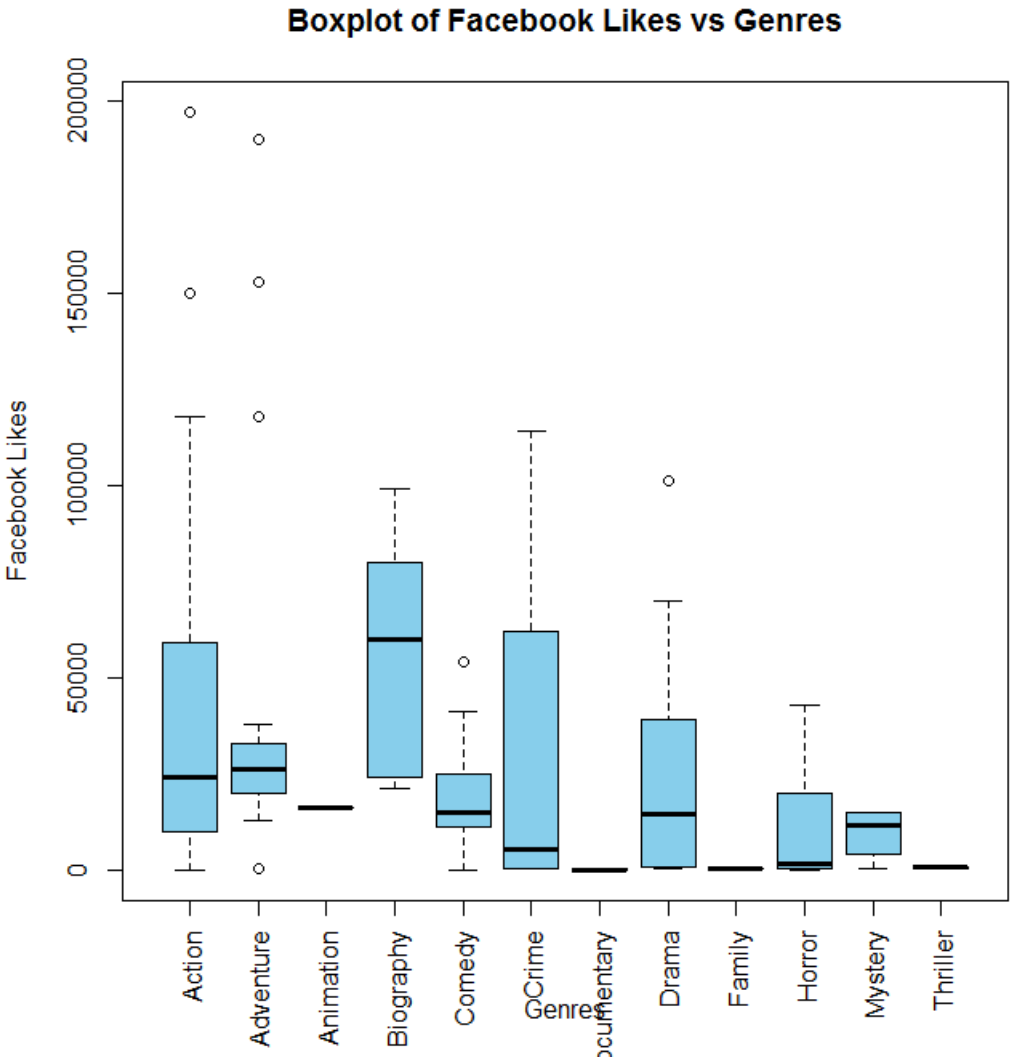
Median distance from Upper fourth = FU – Median = 38000-19000 = 19000

Median distance from Lower fourth = Median – FL = 19000-936 = 18064

Comparing the whisker length, the length of upper whisker is comparatively high to the length of the lower whisker.

Here we have 11 mild outliers and 4 extreme outliers from the whole dataset of Facebook_Likes.

Boxplot to compare Genre batches:



Genre	N	Min	FL	Median	FU	Max
Action	66	34	10000	24000	59000	197000
Adventure	26	215	20000	26000	33000	190000
Animation	1	16000	16000	16000	16000	16000
Biography	6	21000	24000	60000	80000	99000
Comedy	25	14	11000	15000	25000	54000
Crime	4	158	231.5	5152.5	62000	114000
Documentary	2	5	5	63	121	121
Drama	24	84	890	14500	39000	101000
Family	14	9	124.5	240	437	634
Horror	3	52	378	1468	20000	43000
Mystery	4	312	4156	11500	15000	15000
Thriller	2	398	398	563	728	728

Genre	Spread	Step	Inner	Inner	Outer	Outer	Outliers

			Lower Fence	Upper Fence	Lower Fence	Upper Fence	
Action	49000	73500	-63500	132500	-137000	206000	197000, 150000
Adventure	13000	19500	500	52500	-19000	72000	118000, 190000, 153000, 240, 380, 244, 215
Animation	0	0	16000	16000	16000	16000	NA
Biography	56000	84000	-60000	164000	-144000	248000	54000
Comedy	14000	21000	-10000	46000	-31000	67000	NA
Crime	61768.5	92652.75	-92421.25	154652.75	-185074	247305.5	NA
Documentary	116	174	-169	295	-343	469	NA
Drama	38110	57165	-56275	96165	-113440	153330	101000
Family	312.5	468.75	-344.25	905.75	-813	1374.5	NA
Horror	19622	29433	-29055	49433	-58488	78866	NA
Mystery	10844	16266	-12110	31266	-28376	47532	NA
Thriller	330	495	-97	1223	-592	1718	NA

From the boxplots mentioned above,

We can see that the spread of Action, Biography, Crime, and Drama is relatively same with some deviations.

And few of the Genres like Animation, Biography, Crime, Documentary, Horror, Mystery and Thriller have only single digit Observations which is less than 5.

Here we notice a 7 outliers in Adventure movies and 2 in Action, and 1 each in Biography and Drama respectively.

Comparing the dataset by use of Medians:

As the Median comparison is easy to make conclusions for batches having similar spread, so considering only Action, Adventure, Comedy and Drama.

Genre	Spread	Median
Action	49000	24000
Adventure	13000	26000
Comedy	14000	15000
Drama	38110	14500

Here if Drama genre can get 10890 more likes then it can get same publicity as Action genre movies.

And if comedy genre can get 24110 more likes then it can get same publicity as Drama genre movies.

And if Adventure can get 1000 more likes then the publicity of that will be same as Comedy genre.

Although the compared batches data is right skewed and having few outliers, but by comparing the median facebook likes of each genre and considering Action, Adventure, Comedy and Drama as they have some good amount of observations, we can conclude that, American citizens are more likely to watch Action and Adventure genre movies than Comedy and Drama genre movies.

Comparison by Spread VS Level Plot:

Genre	Spread	Median	Log Median	Log Spread
Action	49000	24000	4.38021	4.6902
Adventure	13000	26000	4.414973	4.113943
Animation	0	16000	4.20412	"-Inf"
Biography	56000	60000	4.778151	4.748188
Comedy	14000	15000	4.176091	4.146128
Crime	61768.5	5152.5	3.712018	4.790767
Documentary	116	63	1.799341	2.064458
Drama	38110	14500	4.161368	4.581039
Family	312.5	240	2.380211	2.49485
Horror	19622	1468	3.166726	4.292743
Mystery	10844	11500	4.060698	4.03519
Thriller	330	563	2.750508	2.518514

	Max spread	Min spread	max/min
raw data	0.131762	0.012164	10.83209
re-expressed data	61768.5	116	532.4871