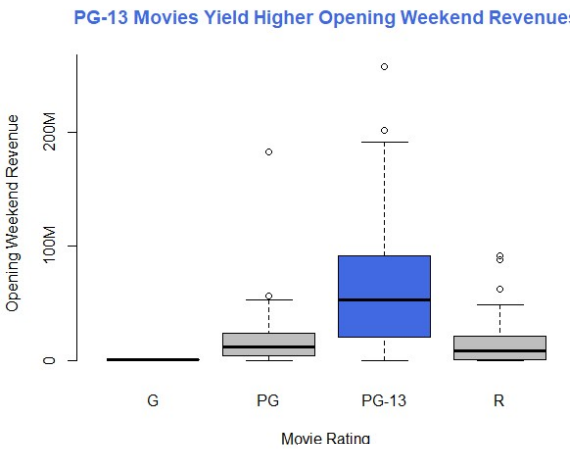
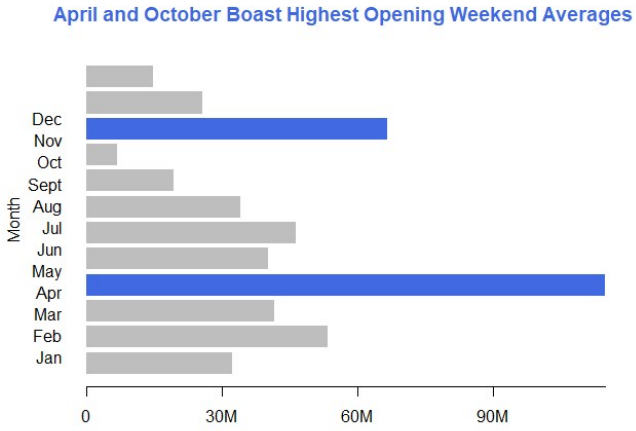
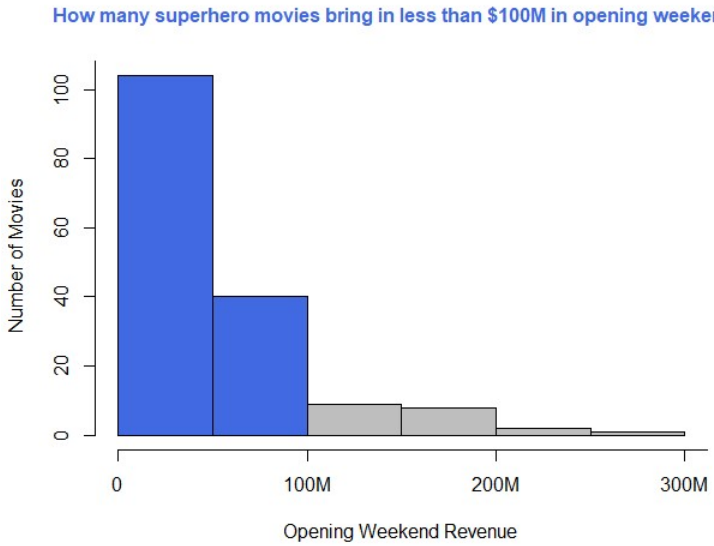
Hypothesis

This analysis uses superhero movie data from the Internet Movie Database API cleaned by Professor Spence for use by our class. The original set covered 35 variables and 860 observations. However, this project focuses on opening weekend revenue as the variable of interest. Figures 1-3 show insights on the opening weekend data with 6 outliers removed:



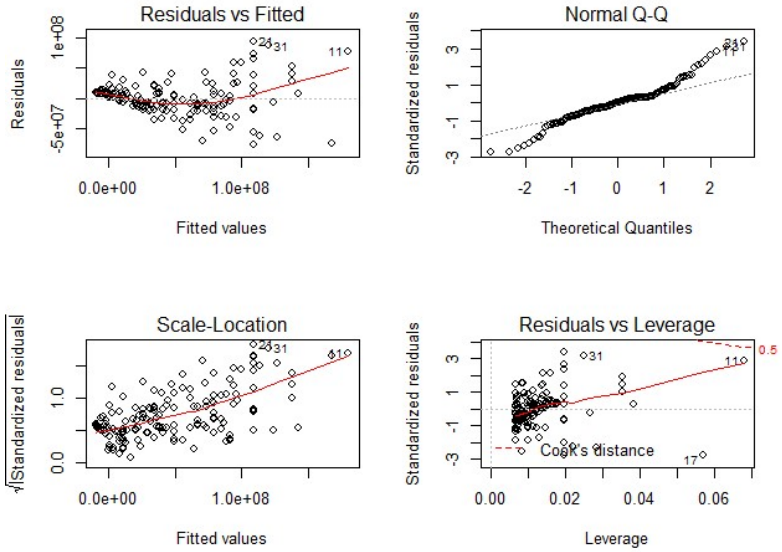




The null hypothesis states that no variables have a significant effect on opening weekend. I decided to test the alternative hypothesis that movie budget and month of movie release would have a significant relationship.

I applied linear regression modeling in R Studio. While the month variable was stored as integers, it does not make mathematical sense to calculate them as such. I used the *as.factor* method to turn each month into a dummy variable. I used the *lm* method to create a linear model where opening weekend is predicted by budget and month. One might notice in the summary output that month 1 is not represented; this is because *as.factor* sets January as a baseline so a predictive model with all other months set to 0 would represent January. Surprisingly, some months were significant to the model while others were not. However, using pieces of a factor and dummy variables would not make sense, so the choice must be made whether to include the dummy variables or not. This model including the release month did, in fact, increase the amount of variation explained (74.55%) over a simpler model only using budget (70.56%). This alternative hypothesis is promising, yet there is still room to improve this model.

The following plots check the normality assumptions for the second model predicting opening weekend using budget and release month after 6 outliers are removed. The Q-Q plot in particular suggests a non-linear model might work better.



Peer review and improvements

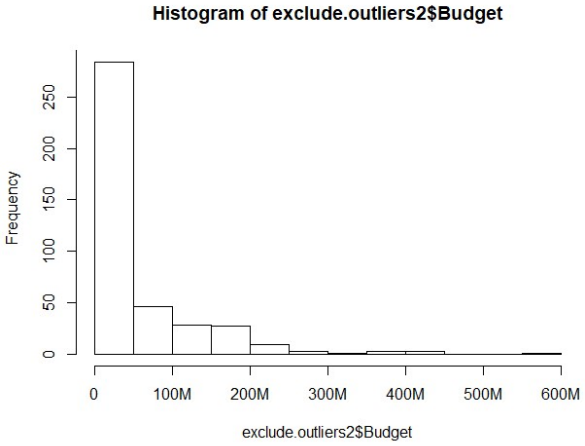
My first iteration had several issues:

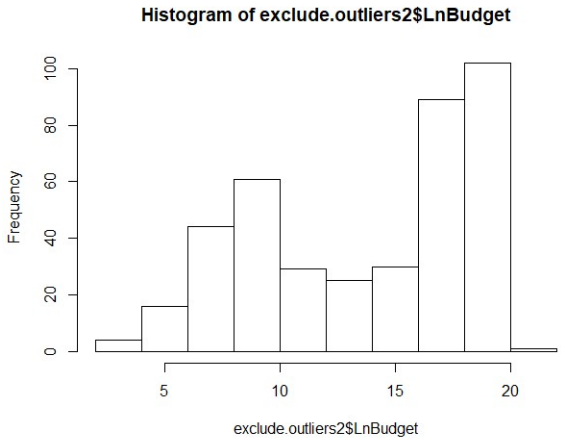
- I did not address outliers.
- When I used *abline* to plot my regression model, I received a warning that R only took 2 of 13 variables into account.
- I assumed all the data imported was in US currency, which was not the case.
- The normality assumptions are not met.

After receiving my first round of feedback and investigating others' projects, I was able to see where Professor Spence had pointed out the movie Devilman had a budget of five billion because the data import was in Japanese Yen. I started removing outliers three at time. The first three I removed had incorrect data imports: Krrish 3, Eliza on the Ice, and Devilman which improved the adjusted R-squared of my model and allowed my plot to write correctly. I continued with a second and third rounds removing three outliers each. The improvement to adjusted R-squared was only 0.001, so I kept the model removing only six outliers instead of nine.

Removing the outliers helped with plotting visualizations and increasing accuracy of the model, but they did not fix the problem with unmet normality assumptions. Reflecting on feedback from graded assignments and the discussion boards, I found the next step for improvement might be adding a logarithmic or exponential component to the model. I tried both, using the *log()* function (which applies natural log in R) in models 3 and 4 on Budget and Opening Weekend, and *I()* on Budget for interpretation in a 5th model.

The histograms below show the difference in the budget variable after being transformed with natural log. The distribution improves, but the interpretability is lost.





Transforming the variables resulted in the following adjusted R-squared values:

- 75.84% of variation explained by Model 3 (Logarithmic Budget only)
- 76.80% of variation explained by Model 4 (Logarithmic Budget and release month)
- 57.33% of variation explained by Model 5 (Budget-squared and Budget-cubed)

Moving forward with model 4, I tried removing 2 more rounds of 3 outliers each, so the adjusted R-squared becomes 0.8658. Now it explains 86.58% of the variation in the sampled data. After removing a total of 12 outliers and doing a log transform, the model has greatly improved. Here might be a good place to stop manipulating the data to avoid overfitting.

However, the purpose of this analysis is to find the best model to explain the data. A higher adjusted R-squared value may explain more of the variation, but transforming the variables using natural log complicates the model in terms of interpretability (Mike Marin https://www.youtube.com/watch?v=tOzwEv0PoZk). For this reason, this analysis will keep the first linear model using budget and release month to be its final predictive model for opening weekend.

As you can see, implementing feedback had a positive effect. With so many ways to present and manipulate the data, examining feedback was very important to the process of incrementally improving the model. Good data analysis relies on trying new coding methods, focusing on data visualization clarity and best practices, and interpreting math calculations and underlying implications of your data set accurately; it's very easy to focus on one thing and forget the others without getting input from colleagues.

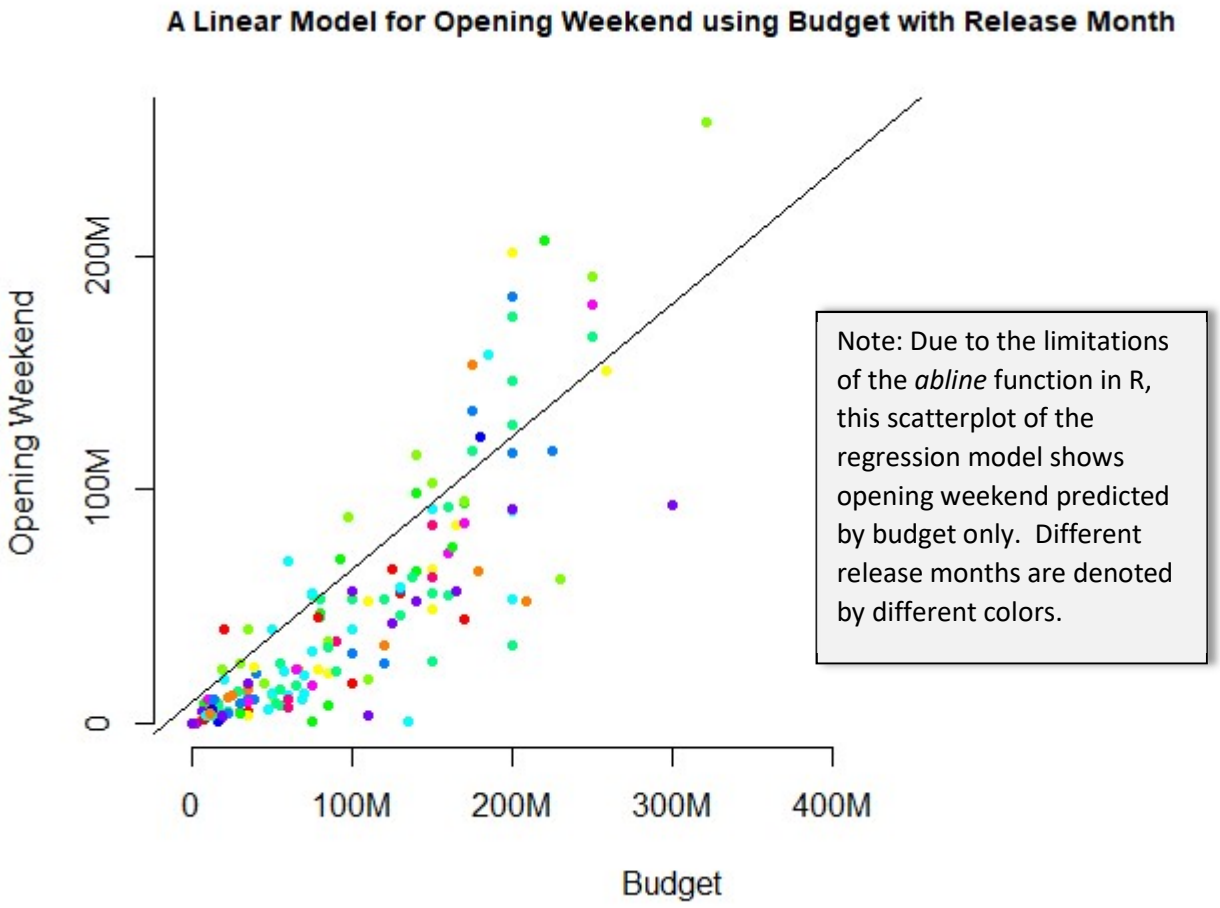
Summary of Results

The summary of model_data suggests the following formula can be used to predict opening weekend revenue, where we insert 1 for the chosen month and 0 for all others:

Predicted opening weekend = 8933000 + 0.5710*Budget + January*1 - 6476000*February - 17240000*March + 10330000*April - 29770000*May - 24650000*June - 15790000*July -16310000*August - 16090000*September - 26780000*October -33630000*November - 24390000*December

```
> summary(model_data)

Call:
lm(formula = exclude.outliers2$Opening.Weekend ~ exclude.outliers2$Budget +
    DV_month)

Residuals:
      Min        1Q    Median        3Q       Max
-65661718 -12481969   -345887  12836126  84197839

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.933e+06  8.884e+06   1.006  0.31635
exclude.outliers2$Budget  5.710e-01  3.331e-02  17.144  < 2e-16 ***
DV_month2                -6.476e+06  1.190e+07  -0.544  0.58707
DV_month3                -1.724e+07  1.189e+07  -1.450  0.14927
DV_month4                 1.033e+07  1.118e+07   0.924  0.35709
DV_month5                -2.997e+07  1.129e+07  -2.655  0.00884 **
DV_month6                -2.465e+07  9.880e+06  -2.495  0.01372 *
DV_month7                -1.579e+07  1.037e+07  -1.522  0.13013
DV_month8                -1.631e+07  1.078e+07  -1.512  0.13265
DV_month9                -1.609e+07  1.728e+07  -0.931  0.35321
DV_month10               -2.678e+07  1.218e+07  -2.198  0.02953 *
DV_month11               -3.363e+07  1.306e+07  -2.575  0.01103 *
DV_month12               -2.439e+07  1.219e+07  -2.002  0.04722 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25840000 on 143 degrees of freedom
  (699 observations deleted due to missingness)
Multiple R-squared:  0.7652,    Adjusted R-squared:  0.7455
F-statistic: 38.84 on 12 and 143 DF,  p-value: < 2.2e-16
```

The adjusted R-squared explains 74.55% of the variability for this sample of data, which is a reasonable amount that avoids overfitting and offers clearer interpretations than the logarithmic version. Notice at the beginning of analysis 699 records were removed, as these observations had missing data (opening weekend, budget, or release month). We can assume that the model accuracy is actually lower because a large amount of the population data was not studied.



For a more robust analysis, we can examine if certain directors, producers, or companies have a greater success rate in order to find examples to follow. For future improvements to this analysis, I suggest using other variables from the superhero movie data to explore whether DC, Marvel, or other branding has a significant effect on the model. I suspect that exploring variables like year of release or country of origin will have an effect on the outcome as well. As always, adding more quality data to the dataset will increase its accuracy.

In practical terms, this analysis is not a golden ticket guaranteeing a movie's success just by implementing the correct budget and release month. But it does provide an idea of what kind of revenue to expect on opening weekend depending on what month the movie is released and how big the budget is.

Appendix: source code, sample data, and outliers

Source code:

```
#read in csv
data1 <- read.csv(file.choose("C:\\Users\\chris\\Documents\\Fall_2019\\DA460\\edited_movie_data.csv"))
#print first row and column names
data1[1,]

#remove outliers
exclude.outliers <- subset(data1, Title != "Krrish 3" & Title !="Eliza on the Ice" & Title != "Devilman")
data1[3,]
data1[29,]
data1[36,]
exclude.outliers2 <- subset(exclude.outliers, Title != "Avengers: Endgame" & Title != "Deadpool" & Title != "The Lone Ranger")
exclude.outliers[11,]
exclude.outliers[21,]
exclude.outliers[31,]

#exclude.outliers3 <- subset(exclude.outliers2, Title !="Glass" & Title !="Watchmen" & Title != "Batman Begins")
#running the model with this 3rd set of outliers did not affect R-square enough to justify dropping them
#so final model will use exclude.outliers2
#after accepting model4, delete more outliers
exclude.outliers2[290,]
exclude.outliers2[66,]
exclude.outliers2[123,]
exclude.outliers4 <- subset(exclude.outliers2, Title != "Special" & Title != "Wanted" & Title != "Teenage Mutant Ninja Turtles: Out of the Shadows")
#try another round of outliers
exclude.outliers4[16,]
exclude.outliers4[219,]
exclude.outliers4[278,]
exclude.outliers5 <- subset(exclude.outliers4, Title != "Deadpool 2" & Title != "The Toxic Avenger" & Title != "Super Capers: The Origins of Ed and the Missing Bullion")
#create dummy variable for release month
DV_month <- as.factor(exclude.outliers2$Month)
#run first model predicting opening weekend with budget and release month
model_data <- lm(exclude.outliers2$Opening.Weekend ~ exclude.outliers2$Budget + DV_month)
model2_data <-lm(exclude.outliers2$Opening.Weekend ~ exclude.outliers2$Budget)
#print summary statistics
summary(model_data)
summary(model2_data)
#plot residuals
par(mfrow=c(1,1))
plot(model_data)

#try a logarithmic model
#model4 has highest adj R squared
exclude.outliers5$LnOpening.Weekend <- log(exclude.outliers5$Opening.Weekend)
exclude.outliers5$LnBudget <- log(exclude.outliers5$Budget)
model3_data <- lm(exclude.outliers2$LnOpening.Weekend ~ exclude.outliers2$LnBudget)
model4_data <- lm(exclude.outliers5$LnOpening.Weekend ~ exclude.outliers5$LnBudget +DV_month)
```

```
summary(model3_data)
summary(model4_data)
plot(model3_data)
plot(model4_data)

#try adding exponential variables
#reduces adj R squared
model5_data <- lm(exclude.outliers2$LnOpening.Weekend ~ exclude.outliers2$Budget + I(exclude.outliers2$Budget^2)
+ I(exclude.outliers2$Budget^3))
summary(model5_data)
plot(model5_data)

#First Visualization: Regression
#color code months on scatter plot, add regression line
#need to add month legend and adjust y-axis for scale
month = cut(exclude.outliers5$Month, breaks = 12)
cols = rainbow(12)[as.numeric(month)]
plot(exclude.outliers2$Opening.Weekend ~ exclude.outliers2$Budget,
    main = "A Linear Model for Opening Weekend using Budget with Release Month",
    cex.main = 0.9,
    col=cols,
    axes=F,
    las =1, pch = 20,
    xlab = "Budget",
    ylab = "Opening Weekend")
axis(1, at=c(0,100000000,200000000,300000000,400000000), labels = c(0,"100M","200M","300M","400M"))
axis(2, at=c(0,100000000,200000000,300000000,400000000), labels = c(0,"100M","200M","300M","400M"))
abline(model_data)
#legend(1, 1, legend, col=cols, title = "Month")

#Second Visualization: Barplot
#compare average amount for opening weekend by month
#would like to improve the y-axis labeling and reorder data
plot.new()
barplot.data <-aggregate(Opening.Weekend ~ Month, data = exclude.outliers2, mean, na.rm = TRUE)
as.data.frame(barplot.data)

barplot(barplot.data$Opening.Weekend,
      las=1, border=NA, xlab="$", ylab = "Month", horiz = TRUE,
      main = "April and October Boast Highest Opening Weekend Averages", col.main= "royalblue", axes = F,
      col = c("gray","gray","gray","royalblue","gray","gray","gray","gray","gray","royalblue","gray","gray"))
axis(2, at=c(1,2,3,4,5,6,7,8,9,10,11,12),tick = 0, labels = c("Jan", "Feb", "Mar", "Apr", "May","Jun","Jul","Aug", "Sept",
"Oct","Nov","Dec"), las = 1)
axis(1, at=c(0,30000000,60000000,90000000,120000000), labels = c("0","30M","60M","90M","120M"))

#Third Visualization: Histogram
#histogram of opening weekend

hist(exclude.outliers2$Opening.Weekend,
    xlab = "Opening Weekend Revenue", ylab = "Number of Movies",
    main = "How many superhero movies bring in less than $100M in opening weekend?",
    cex.main = 1, col.main = "royalblue", axes = F,
```

```
    col = c("royalblue", "royalblue", "gray","gray","gray","gray","gray","gray"))
axis(1, at=c(0,100000000,200000000,300000000,400000000), labels = c("0","100M","200M","300M","400M"))
axis(2, at=c(0, 20, 40, 60, 80, 100, 120))

#Fourth Visualization: Boxplot
#compare Opening Weekend by MPAArating buckets

boxplot(Opening.Weekend ~ MPAA, data = exclude.outliers2, subset = (MPAA =="G"| MPAA == "PG"| MPAA == "PG-13"| MPAA == "R"),
      drop = TRUE, axes=F,
      xlab = "Movie Rating", ylab = "Opening Weekend Revenue",
      main = "PG-13 Movies Yield Higher Opening Weekend Revenues", col.main= "royalblue",
      col = c("gray","gray","royalblue","gray"))
axis(1, at=c(1,2,3,4), labels = c("G", "PG", "PG-13", "R"), tick = 0)
axis(2, at=c(0,100000000,200000000,300000000,400000000), labels = c("0", "100M", "200M","300M","400M"))

#Fifth Visualization: Histogram of Budget and BudgetLn
hist(exclude.outliers2$Budget)
hist(exclude.outliers2$LnBudget)
```

Sample data:

ï..ID	Title	Runtime	Rating	Votes	MPAA	ReleaseDate	Year	Month	Day	Budget	Opening.Weekend	Cumulative.Gross	IsDc	IsMarvel
4154756	Avengers: Infinity War	149	8.5	710195	PG-13	4/23/2018 0:00	2018	4	23	321000000	257698183	2048709917	NA	1
848228	The Avengers	143	8.0	1200653	TV-14::(LV)	4/11/2012 0:00	2012	4	11	220000000	207438708	1519557910	NA	1
1825683	Black Panther	134	7.3	542201	PG-13	1/29/2018 0:00	2018	1	29	200000000	202003951	1347071259	NA	1
2395427	Avengers: Age of Ultron	141	7.3	682974	PG-13	4/13/2015 0:00	2015	4	13	250000000	191271109	1405413868	NA	1
3606756	Incredibles 2	118	7.7	215615	PG	6/5/2018 0:00	2018	6	5	200000000	182687905	1242770554	NA	NA
3498820	Captain America: Civil War	147	7.8	602061	PG-13	4/12/2016 0:00	2016	4	12	250000000	179139142	1153304495	NA	1
1300854	Iron Man 3	130	7.2	706627	PG-13	4/12/2013 0:00	2013	4	12	200000000	174144585	1215439994	NA	1
2975590	Batman v Superman: Dawn of Justice	151	6.5	585742	PG-13	3/12/2016 0:00	2016	3	12	250000000	166007347	873634919	1	NA
468569	The Dark Knight	152	9.0	2102542	TV-14::(LV)	7/14/2008 0:00	2008	7	14	185000000	158411483	1004558444	1	NA
4154664	Captain Marvel	123	7.0	361286	PG-13	2/27/2019 0:00	2019	2	27	175000000	153433423	1128274794	NA	1
413300	Spider-Man 3	139	6.2	468924	TV-PG	4/3/2007 0:00	2007	4	3	258000000	151116516	890871626	NA	1
3896198	Guardians of the Galaxy Vol. 2	136	7.6	504328	PG-13	4/10/2017 0:00	2017	4	10	200000000	146510104	863756051	NA	1

Note: This snip includes the variables of interest for this analysis, not all 35 variables available in this dataset.

Outliers:

Krrish 3
Eliza on the Ice
Devilman
Avengers: Endgame
Deadpool
The Lone Ranger