Analysis of IMDB Movies Over Time

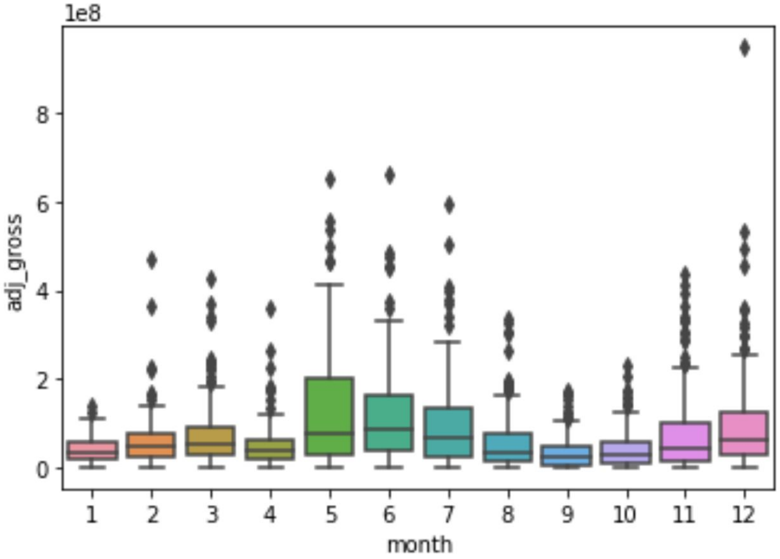
STATS 131 FInal Project

Valerie Chen
Valeria Lopez
Cristina Sanchez
Jericho Villareal

Background

- data set derived from Kaggle titled **Movie Industry**: *Three decades of movies*
- IMDb provides a variety of information such as runtime, user rating, year released, US certificate rating,
- as well as box office reportings of films
- analyzing data on popular movies pulled from the IMDb website using their advanced search tool
- popularity determined by the number of visits to the movies' page from when the movie was listed on IMDb to the day the data was scraped n 2017
- original data set contains 6820 observations and 15 variables; final data set contains 2131 observations and 12 variables
- data set filtered using the movie's popularity and year it was released, top 220 movies collected for each year spanning three decades starting from 1986 to 2016
- data subsetted to only USA movies due to US gross
- variables in original data set are “budget”, “company”, “country”, “director”, “genre”, “gross”, “name”, “rating”, “released”, “runtime”, “score”, “votes”, “star”, “writer”, “year”
- variables in final data set are “country”, “gross”, “name”, “rating”, “released”, “runtime”, and “year”

Exploratory Data Analysis

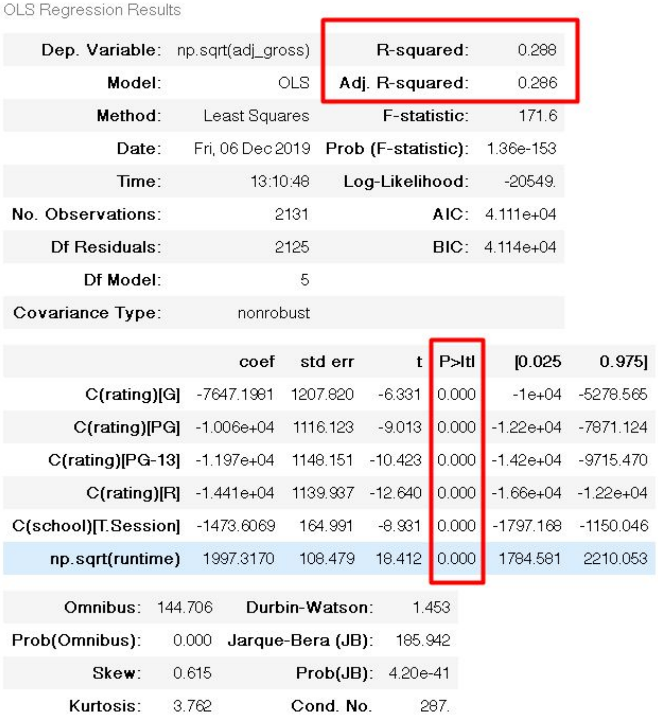


- There is a **sine-like curve** trend in the means of adjusted gross across the 12 months.
- This trend corresponds with the academic calendar year; school is typically in session during the months of January, February, March, April, September, October, and November, and breaks are typically during the months of *May, June, July, and August* for summer and *December* for winter.
- This may indicate that adjusted gross is influenced in part by the amount of free time people may have during breaks versus in-session periods of school.

Linear Regression

“Can a movie’s gross be explained by certain factors?”

A movie’s rating, runtime, and release period can explain ~ 29% of the variability in its gross sales.

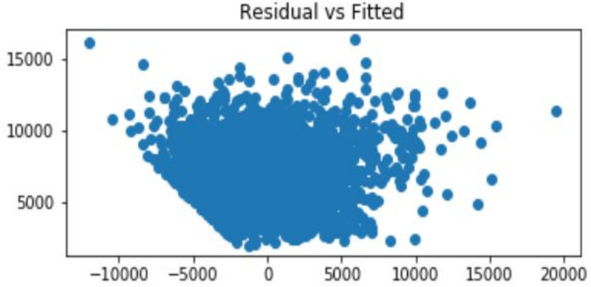


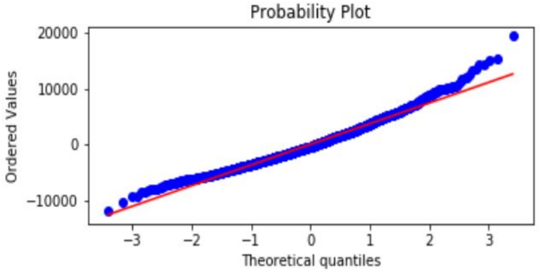
```
OLS Regression Results

    Dep. Variable:  np.sqrt(adj_gross)        R-squared:      0.288
            Model:                 OLS   Adj. R-squared:      0.286
           Method:       Least Squares      F-statistic:      171.6
             Date:    Fri, 06 Dec 2019 Prob (F-statistic):  1.36e-153
             Time:            13:10:48   Log-Likelihood:    -20549.
No. Observations:                2131              AIC:  4.111e+04
     Df Residuals:                2125              BIC:  4.114e+04
         Df Model:                   5
  Covariance Type:           nonrobust

                         coef   std err        t  P>|t|     [0.025      0.975]
      C(rating)[G]  -7647.1981  1207.820   -6.331  0.000     -1e+04   -5278.565
     C(rating)[PG]  -1.006e+04  1116.123   -9.013  0.000  -1.22e+04   -7871.124
  C(rating)[PG-13]  -1.197e+04  1148.151  -10.423  0.000  -1.42e+04   -9715.470
      C(rating)[R]  -1.441e+04  1139.937  -12.640  0.000  -1.66e+04   -1.22e+04
C(school)[T.Session] -1473.6069  164.991   -8.931  0.000  -1797.168   -1150.046
  np.sqrt(runtime)   1997.3170   108.479   18.412  0.000   1784.581    2210.053

      Omnibus:  144.706    Durbin-Watson:    1.453
Prob(Omnibus):    0.000 Jarque-Bera (JB):  185.942
         Skew:    0.615         Prob(JB): 4.20e-41
     Kurtosis:    3.762        Cond. No.      287.
```





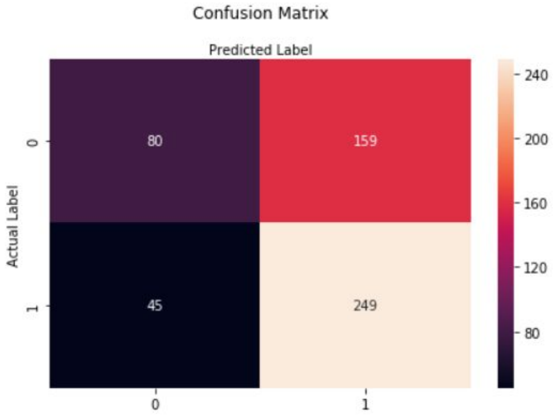
Binary Classification

“What are the odds that a movie will fall into group 1 or 2?”

Classified as **“Break”** or **“Session”**

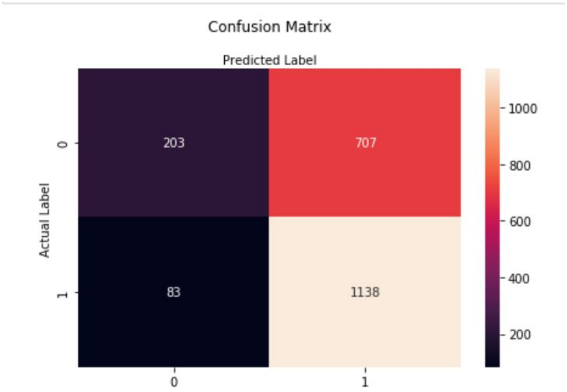
LR: 62% accuracy
K-Means: 63% accuracy

Logistic Regression



```
Accuracy: 0.6172607879924953
Precision: 0.6102941176470589
Recall: 0.8469387755102041
```

K-Means Clustering (K = 2)



```
Accuracy: 0.6292820272172689
Precision: 0.6168021680216802
Recall: 0.9320229320229321
```

Results and Conclusions

- After performing our exploratory data analysis, we were able to see some interesting patterns and trends.
- The most interesting finding was the significance between adjusted gross and months in the year.
- We observed a similar pattern between adjusted gross and month released to that of the academic calendar. There was a higher adjusted gross during the months of summer and winter break versus months of when school was in session.
- Through data modeling, we found that the variables "school" (derived from months), "rating", and "runtime" were statistically significant with respect to adjusted gross. Our findings suggest that movies released during vacation gross more than those released during Session.
- These conclusions were able to support our hypothesis that the (adjusted) gross of movies is influenced by the academic calendar and follows a periodic pattern that is relatively time dependent.