PROBLEM STATEMENT

With more than 2.5 billion video gamers from all over the world, more gamers are switching towards mobile gaming compared to traditional video game consoles which are slowly moving out of phase.

In this project of mine, I'll analyze sales data from 8k video games, identify most correlated to hits (games which sell over 1M units) and implement a prediction model to show which games released in 2016 can still become hit and which can't.

Also, I'll create a predictive model based on video gaming market revenue based on the data and identify what are the most important features that affects the revenue of video games.

Additionally, through analysis, I would like to identify the console among PS, XBOX, Nintendo, Play Station is the console that will possibly generate the most sales and effectively maximizing profits and minimizing the cost for further game development.

Lastly, I'll be building a dashboard/report in PowerBI for better Data Visualization to sales director which will be helpful to him for taking better decisions.

OUTCOMES:

1. Which games (released in 2016) can still become hit and which can't.
2. Predictions for sales volume.
3. PowerBI dashboard.

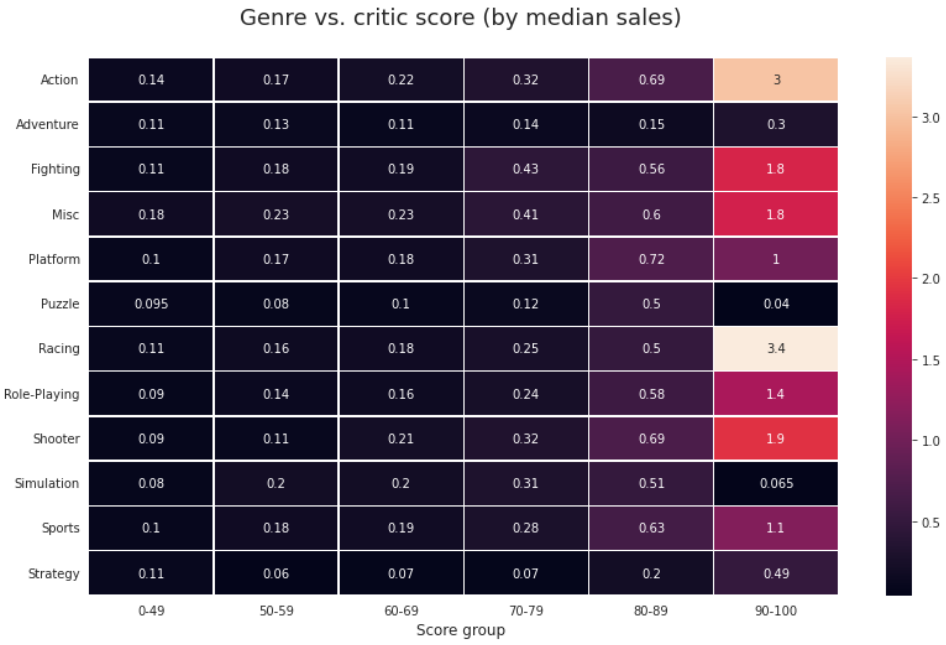
PREDICTING VIDEO GAMES HITS WITH MACHINE LEARNING

Data Exploration and Analysis

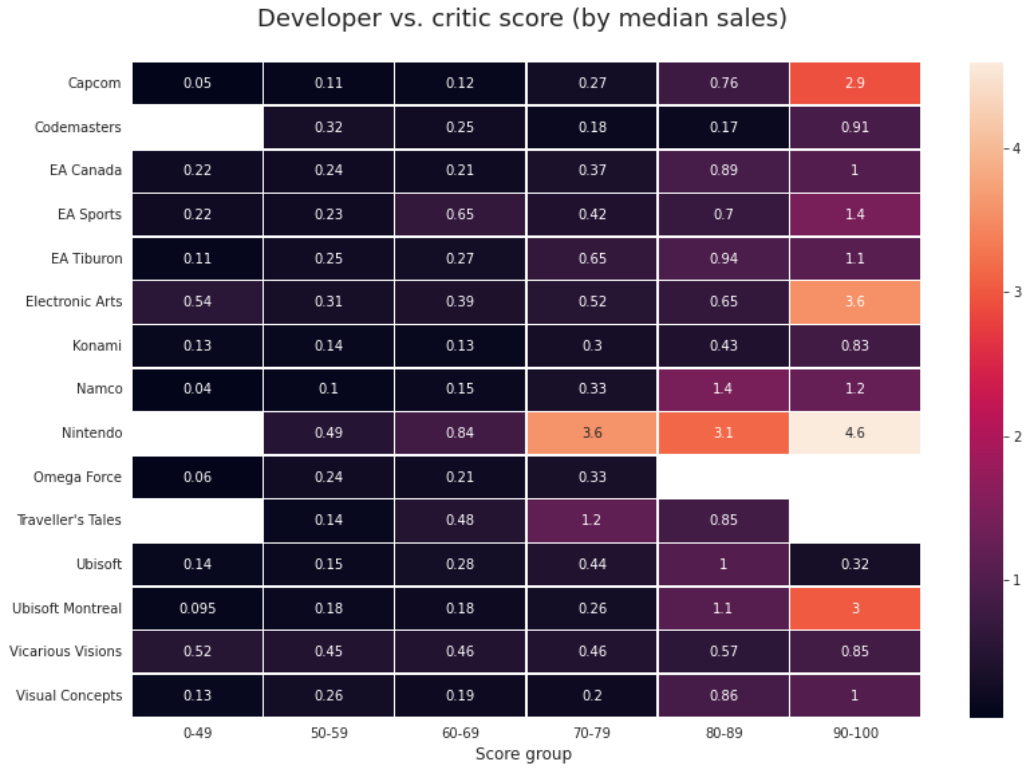
Median sales (in millions of units) vs. critic scores

The following four heatmaps show how game sales vary according to critic scores, which are split into six scoring groups. Additionally, each heatmap segments the data further by one of the following features: genre, developer, publisher, and platform (in order of appearance).

Under each heatmap, we identify the categories where games sell best. This is done for okay, good, and great games, as defined by games with scores in the 70s, 80s, and 90s, respectively.

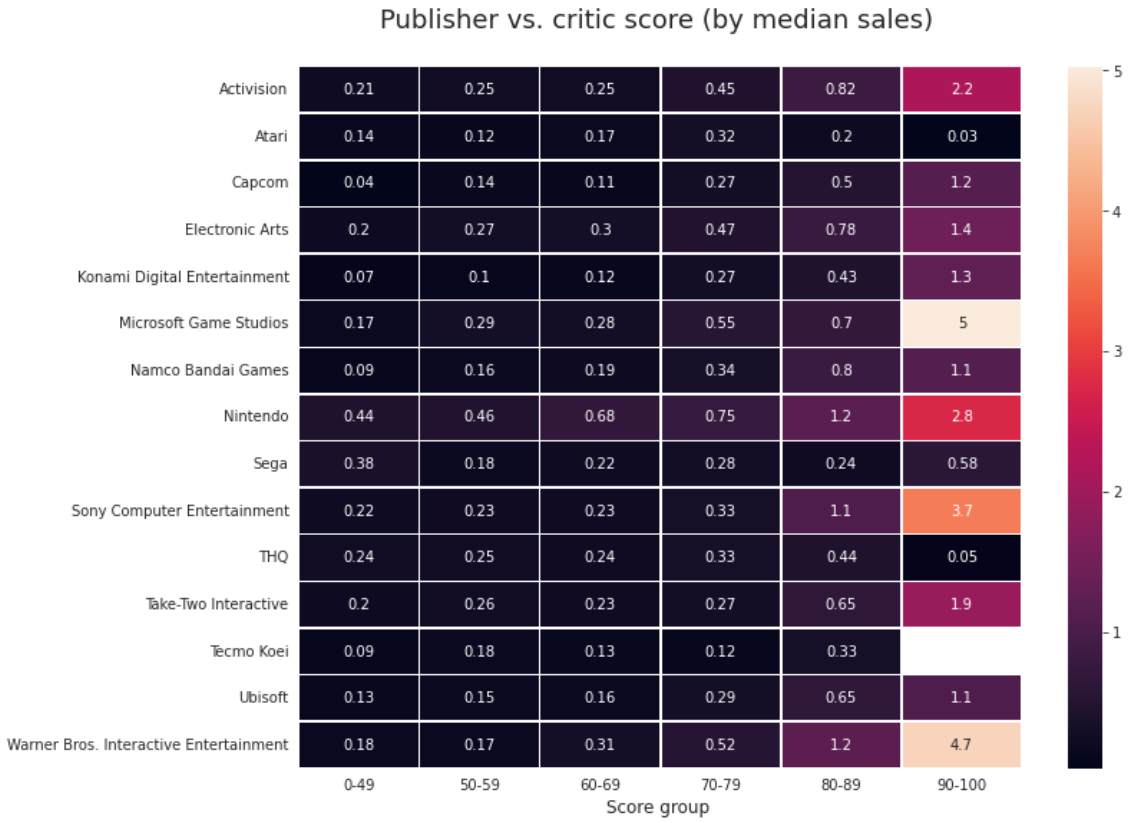


- **Genres where great games sell best:** Racing, Action
- **Genres where good games sell best:** Platform, Shooter/Action
- **Genres where okay games sell best:** Fighting, Misc

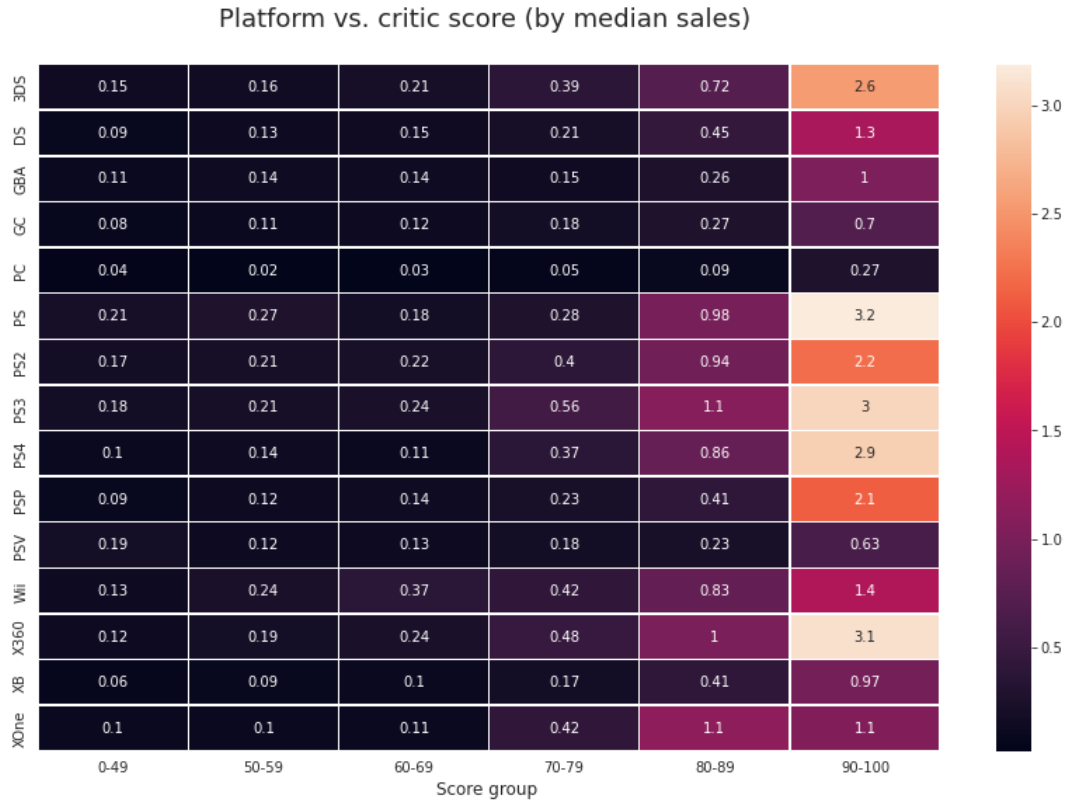


- **Developers whose great games sell best:** Nintendo, Electronic Arts
- **Developers whose good games sell best:** Nintendo, Namco
- **Developers whose okay games sell best:** Nintendo, Traveler's Tales

Interpretations: In the **great** scores column (last), Nintendo has the highest median sales (in millions of units) per game, at 4.6M. Interestingly, Nintendo also has the highest median sales per game in both the **good** and **okay** scoring columns.



- **Publishers who sell great games best:** Microsoft Game Studios**,** Warner Bros. Interactive Entertainment
- **Publishers who sell good games best:** Nintendo/ Warner Bros. Interactive Entertainment
- **Publishers who sell okay games best:** Nintendo, Microsoft Game Studios

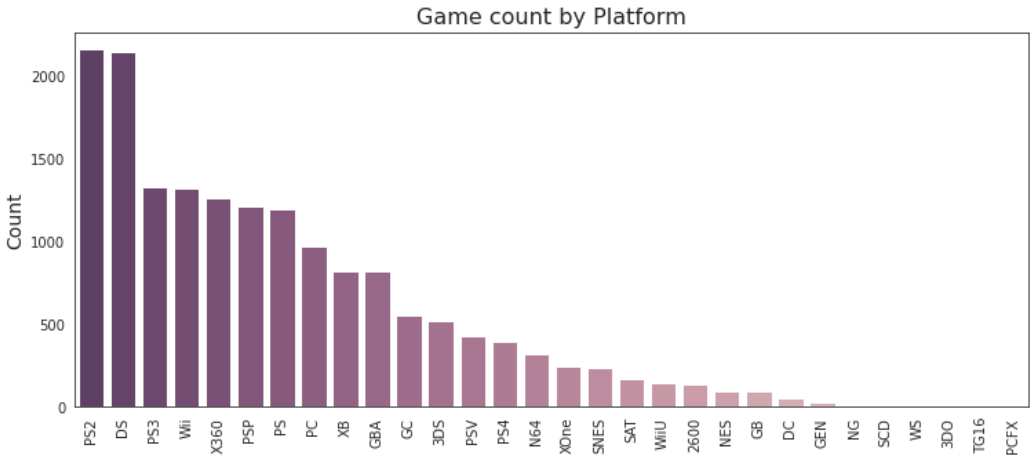


- **Platforms where great games sell best:** PS, X360
- **Platforms where good games sell best:** PS3, XOne
- **Platforms where okay games sell best:** PS3, X360

It's interesting how sensitive game sales in the whole PlayStation line seem to be to high critic scores, especially when sales in the mid-score ranges look relatively on par with other consoles.

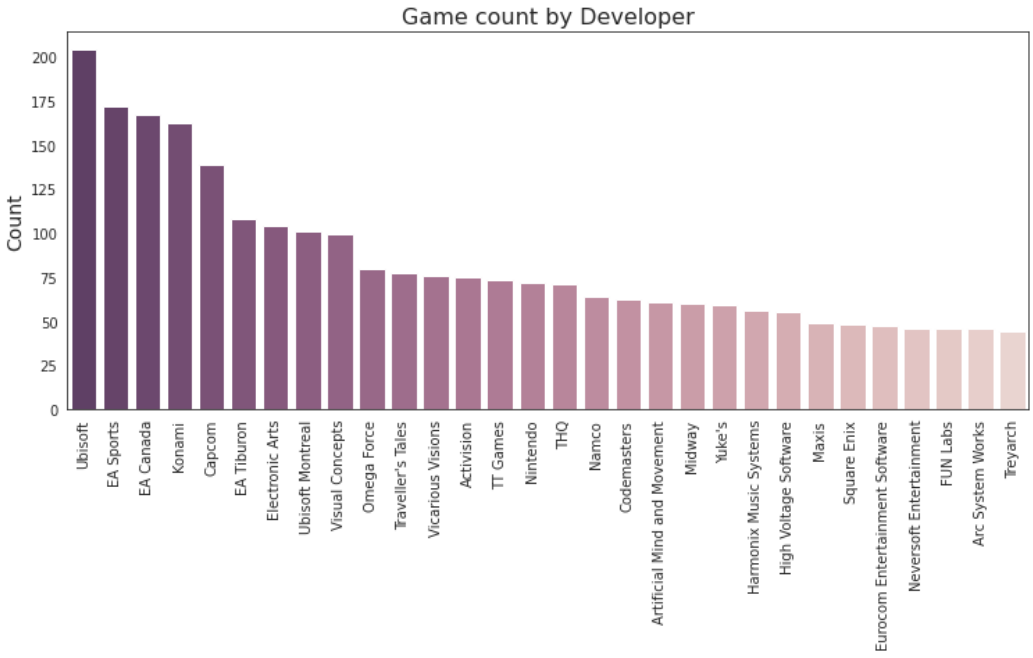
TOP VALUES IN THE DATASET - (BY PLATFORM, DEVELOPER, PUBLISHER AND GENRE)

Platforms with most games in dataset:



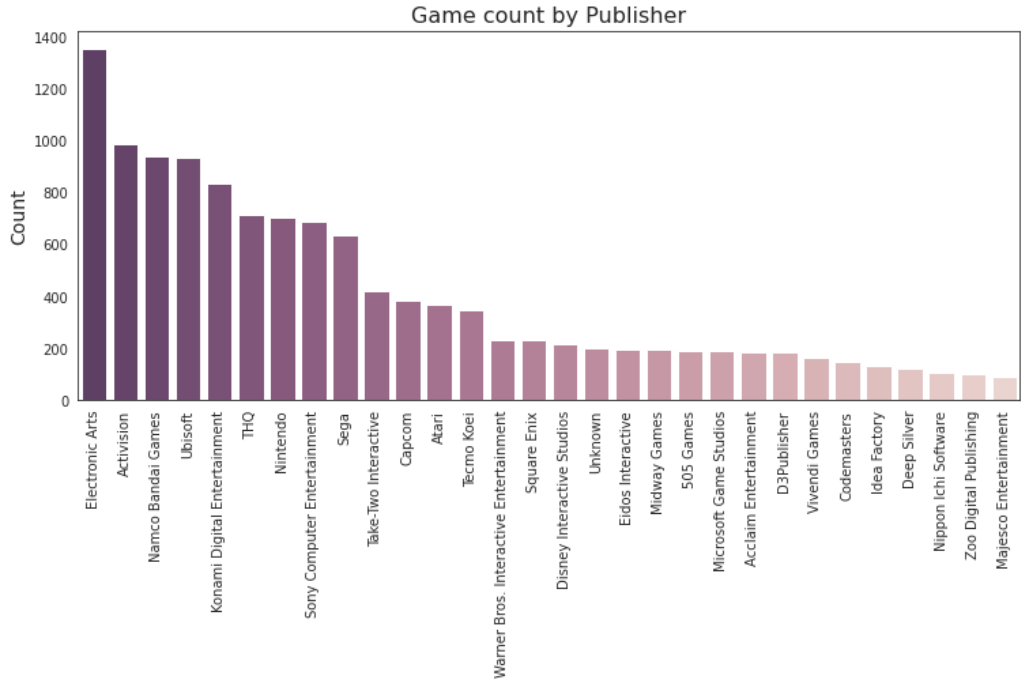
1. PS2
2. DS
3. PS3

Developers with most games in dataset:



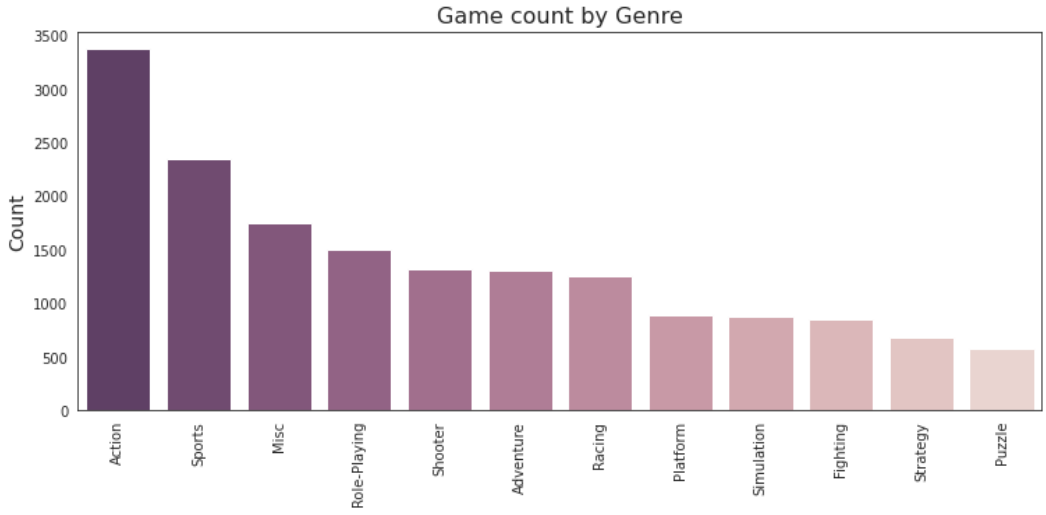
1. Ubisoft
2. EA Sports
3. EA Canada

Publishers with most games in dataset:



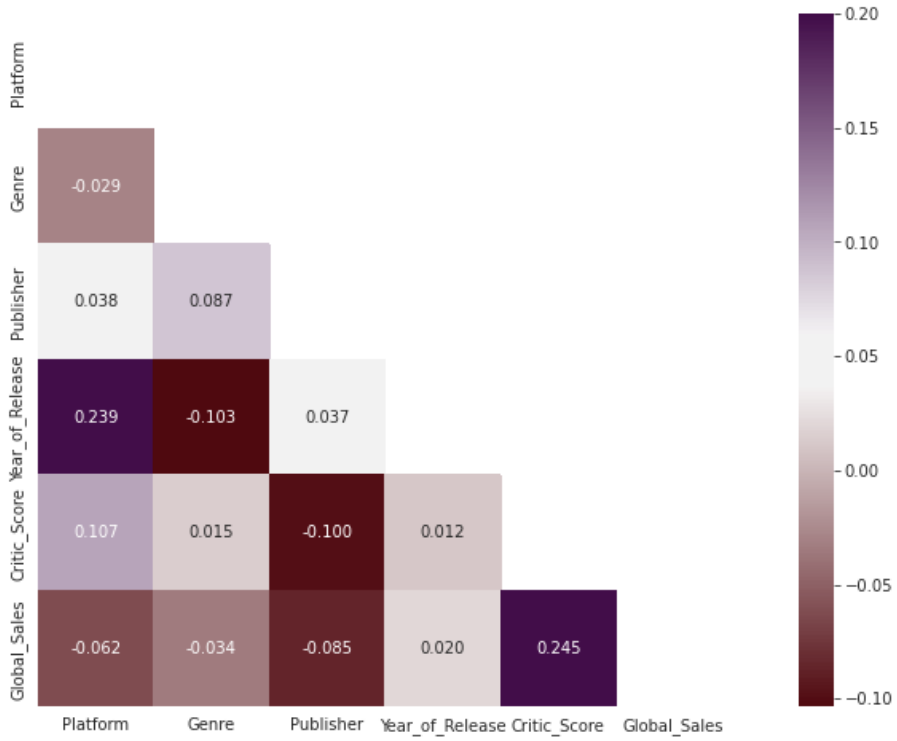
1. Electronic Arts
2. Activision
3. Namco Bandai Games

Genres with most games in dataset



1. Action
2. Sports
3. Misc

DATASET CORRELATIONS – (for numeric and categorical variables)

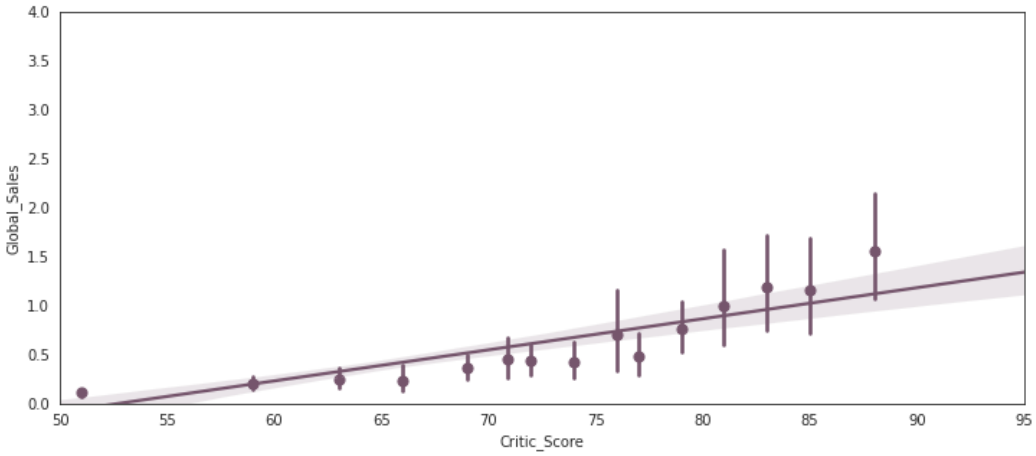


Strongest correlations:

- **Critic score-to-global sales:** We'll take a closer look at this in the next two sections.
- **Year of release-to-platform:** This makes sense since new platforms are released periodically.

Note: Categorical columns (platform, genre, publisher) were converted to numeric in order of game count, as seen in previous section. The slightly negative correlations the have to global sales can be interpreted as "the higher the ID number, the smaller the [platform, genre, publisher], and thus the slightly lower the sales figure".

Critic score vs global sales – (for all years in the dataset)

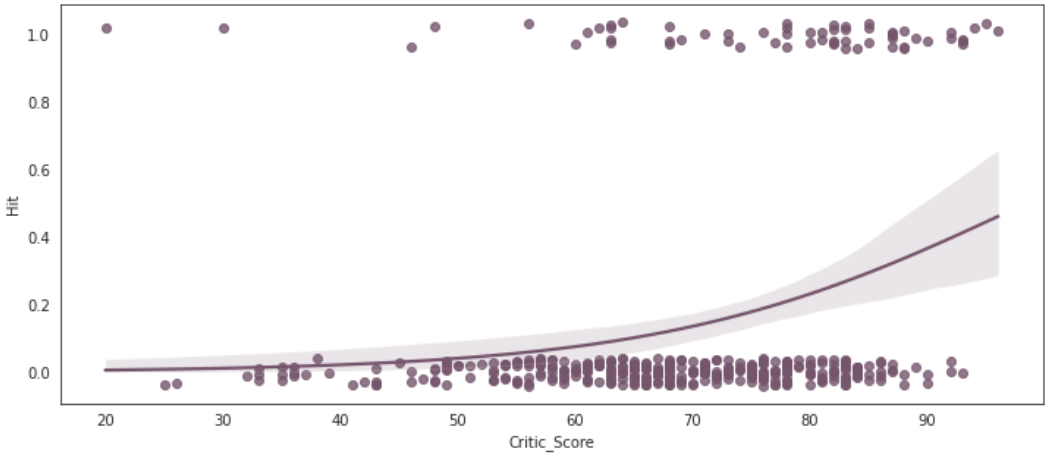


It's interesting how the slope gets steeper in the 80's. it seems once a video game gets high critic score, every additional point has a higher impact. For example, in this 2014-16 subset, **an 8-point increase in critic score seems to have positive effect on sales of about 250k when starting from a score of 65, but around 1M when starting from77.**

DEFINING HITS AS THOSE WITH SALES ABOVE 1 MILLION UNITS

This will be our target in our prediction model, where we'll predict if a game will be a hit or not. The target is binary: 1 if Hit, else 0.

Here's the relationship between critic scores and VG hits using a 5% sample:



As expected, it seems **hits are mostly found near critic scores,** while non-hits can vary in scores but begin to lose presence in the high score ranges (as interpreted by the steepening regression curve near the 70's).

PREDICTION MODEL

For predicting the likelihood of a given game to reach sales of 1 million units or higher, referred to as “hit” games. Classification approach is applied to separate hits from non-hits.

Generating features and train/test splitting

```
[ ] from sklearn.feature_selection import SelectFromModel
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import classification_report, f1_score, accuracy_score, confusion_matrix
    from sklearn import svm
```

```
[ ] df2[:5]
```

	Platform	Genre	Publisher	Year_of_Release	Critic_Score	Hit
0	Wii	Sports	Nintendo	2006.0	76.0	1
1	Wii	Racing	Nintendo	2008.0	82.0	1
2	Wii	Sports	Nintendo	2009.0	80.0	1
3	DS	Platform	Nintendo	2006.0	89.0	1
4	Wii	Misc	Nintendo	2006.0	58.0	1

```
[ ] from pandas import get_dummies
    df_copy = pd.get_dummies(df2)
```

```
[ ] df_copy[:5]
```

	Year_of_Release	Critic_Score	Hit	Platform_3DS	Platform_DC	Platform_DS	Platform_GBA	Platform_GC	Platform_PC
0	2006.0	76.0	1	0	0	0	0	0	0
1	2008.0	82.0	1	0	0	0	0	0	0
2	2009.0	80.0	1	0	0	0	0	0	0
3	2006.0	89.0	1	0	0	1	0	0	0
4	2006.0	58.0	1	0	0	0	0	0	0

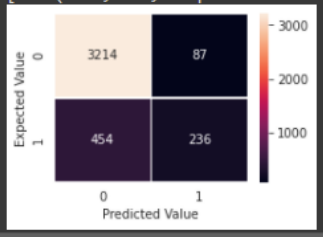
5 rows × 334 columns

```
[ ] df3 = df_copy
    y = df3['Hit'].values
    df3 = df3.drop(['Hit'],axis=1)
    X = df3.values
```

```
[ ] Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, test_size=0.50, random_state=2)
```

Testing prediction accuracy with RFC and LR

```
+ Code  + Text                                                    Connect

[ ] radm = RandomForestClassifier(random_state=2).fit(Xtrain, ytrain)
    y_val_1 = radm.predict_proba(Xtest)
    print("Validation accuracy: ", sum(pd.DataFrame(y_val_1).idxmax(axis=1).values
                                       == ytest)/len(ytest))

    Validation accuracy:  0.864946128789777

[ ] log_reg = LogisticRegression().fit(Xtrain, ytrain)
    y_val_2 = log_reg.predict_proba(Xtest)
    print("Validation accuracy: ", sum(pd.DataFrame(y_val_2).idxmax(axis=1).values
                                       == ytest)/len(ytest))

    Validation accuracy:  0.8644450012528189
    /usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_logistic.py:818: ConvergenceWarning: lbfgs failed to con
    STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

    Increase the number of iterations (max_iter) or scale the data as shown in:
        https://scikit-learn.org/stable/modules/preprocessing.html
    Please also refer to the documentation for alternative solver options:
        https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
      extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,

[ ] all_predictions = log_reg.predict(Xtest)
    print(classification_report(ytest, all_predictions))

                  precision    recall  f1-score   support

               0       0.88      0.97      0.92      3301
               1       0.73      0.34      0.47       690

        accuracy                           0.86      3991
       macro avg       0.80      0.66      0.69      3991
    weighted avg       0.85      0.86      0.84      3991

[ ] fig, ax = plt.subplots(figsize=(3.5,2.5))
    sns.heatmap(confusion_matrix(ytest, all_predictions), annot=True, linewidths=.5, ax=ax, fmt="d").set(xlabel='Predict

    [Text(10.5, 0.5, 'Expected Value'), Text(0.5, 1.5, 'Predicted Value')]
```



Ranking feature performance

```
[ ]  indices = np.argsort(radm.feature_importances_)[::-1]

     # Print the feature ranking
     print('Feature ranking (top 10):')

     for f in range(10):
         print('%d. feature %d %s (%f)' % (f+1 , indices[f], df3.columns[indices[f]],
                                           radm.feature_importances_[indices[f]]))
```

```
Feature ranking (top 10):
1. feature 1 Critic_Score (0.326461)
2. feature 0 Year_of_Release (0.166946)
3. feature 216 Publisher_Nintendo (0.027233)
4. feature 19 Genre_Action (0.021519)
5. feature 99 Publisher_Electronic Arts (0.019508)
6. feature 27 Genre_Shooter (0.019204)
7. feature 42 Publisher_Activision (0.017003)
8. feature 29 Genre_Sports (0.016701)
9. feature 9 Platform_PS2 (0.014028)
10. feature 10 Platform_PS3 (0.014002)
```

WHICH 2016 VIDEO GAMES CAN STILL BECOME HITS?

Video games with highest probability of becoming hits:

	Name	Platform	Hit_Probability
0	Titanfall 2	PS4	0.763205
1	Dishonored 2	PS4	0.758210
2	Dishonored 2	XOne	0.729189
3	Titanfall 2	XOne	0.699592
4	Skylanders Imaginators	PS4	0.661389
5	Kirby: Planet Robobot	3DS	0.630537
6	Plants vs. Zombies: Garden Warfare 2	PS4	0.617735
7	Fast Racing Neo	WiiU	0.606393
8	Lego Star Wars: The Force Awakens	PS4	0.591593
9	Quantum Break	XOne	0.586057

Video games with lowest probability of becoming hits:

	Name	Platform	Hit_Probability
0	Psycho-Pass: Mandatory Happiness	PSV	0.003695
1	RollerCoaster Tycoon World	PC	0.003723
2	Dino Dini's Kick Off Revival	PS4	0.004151
3	Bus Simulator 16	PC	0.004442
4	Battle Worlds: Kronos	PC	0.005096
5	The Technomancer	PC	0.005913
6	Sherlock Holmes: The Devil's Daughter	PC	0.005939
7	Dead or Alive Xtreme 3: Fortune	PS4	0.006217
8	Homefront: The Revolution	PC	0.006291
9	Agatha Christie: The ABC Murders	PC	0.006305