

PC Video Game Sales

By Richard Mei

A dark blue diagonal gradient bar that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the page.

Context

- Why specifically PC Platform?
- Goal: Model global sales of PC Games
- Some characteristics to consider:
 - Genres
 - Rating
 - Publisher
 - Series



SPOILERS:

GENRE

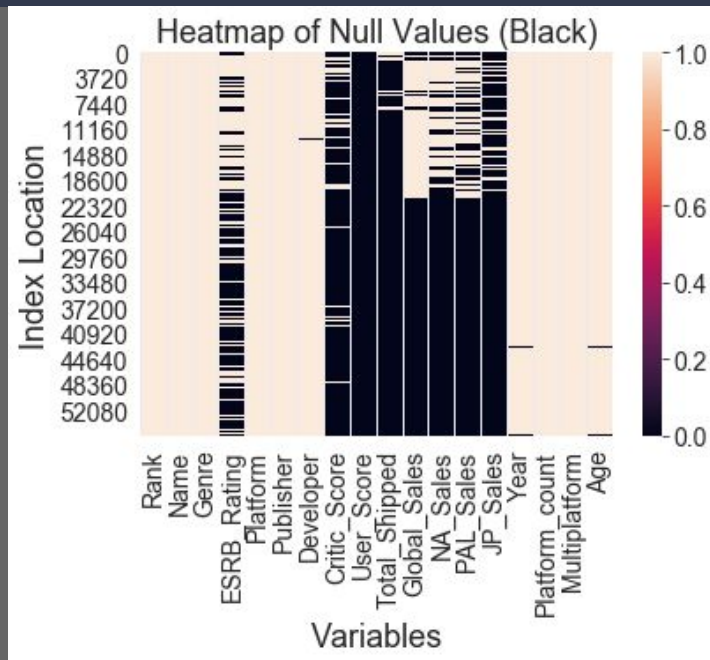
PUBLISHERS

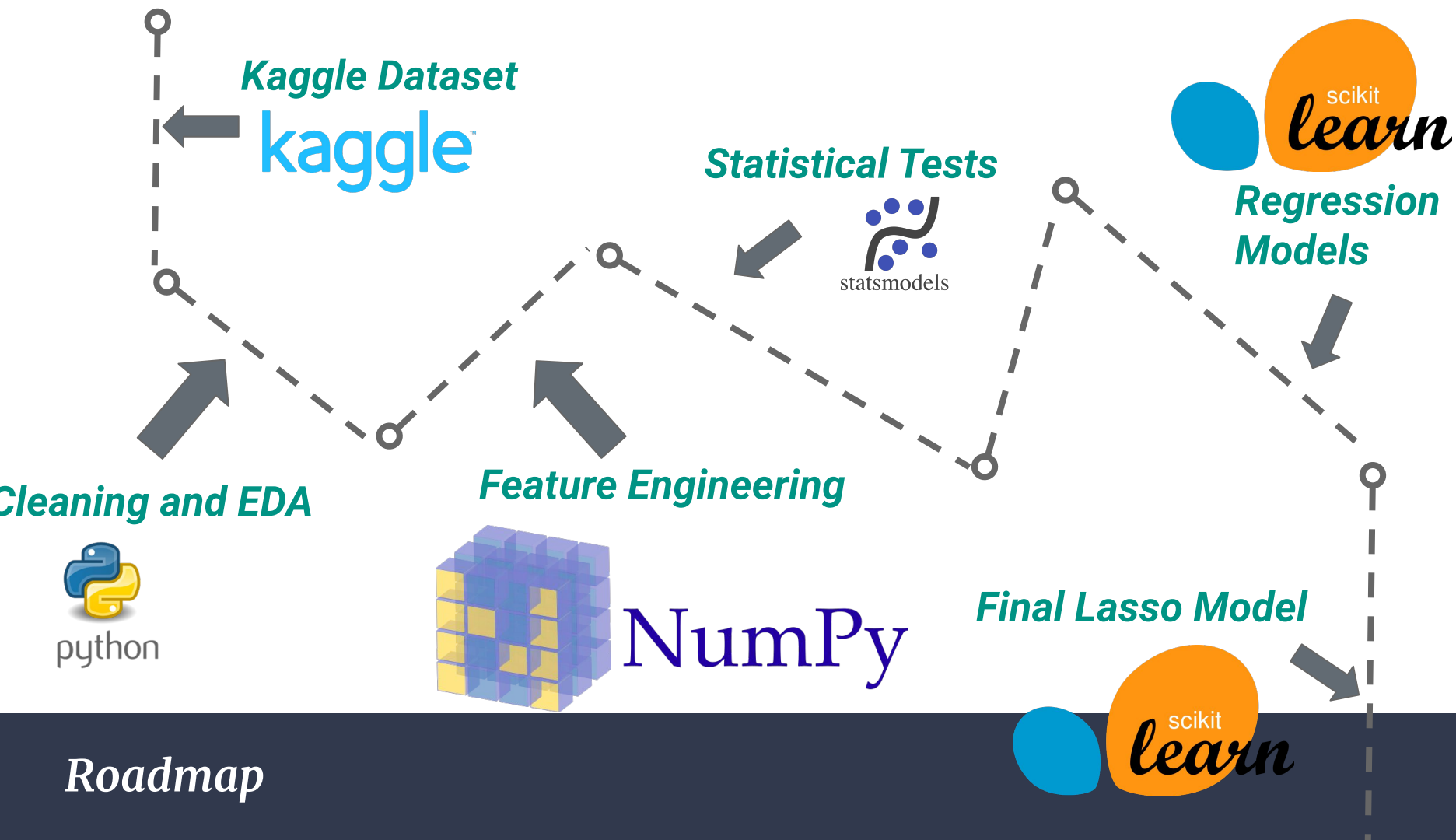
AGE

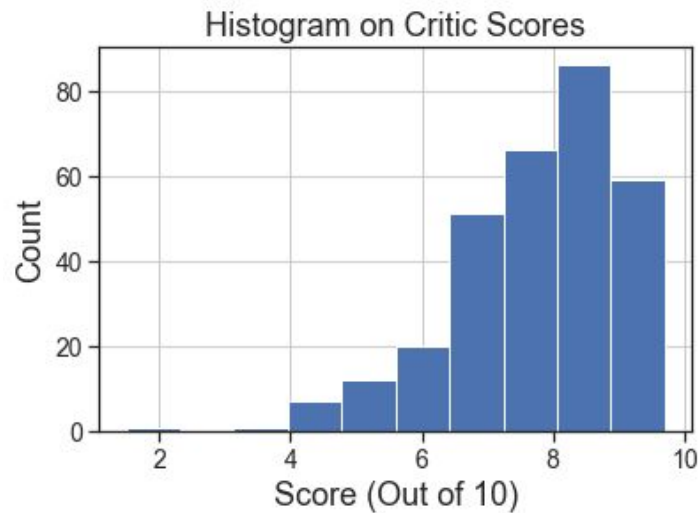
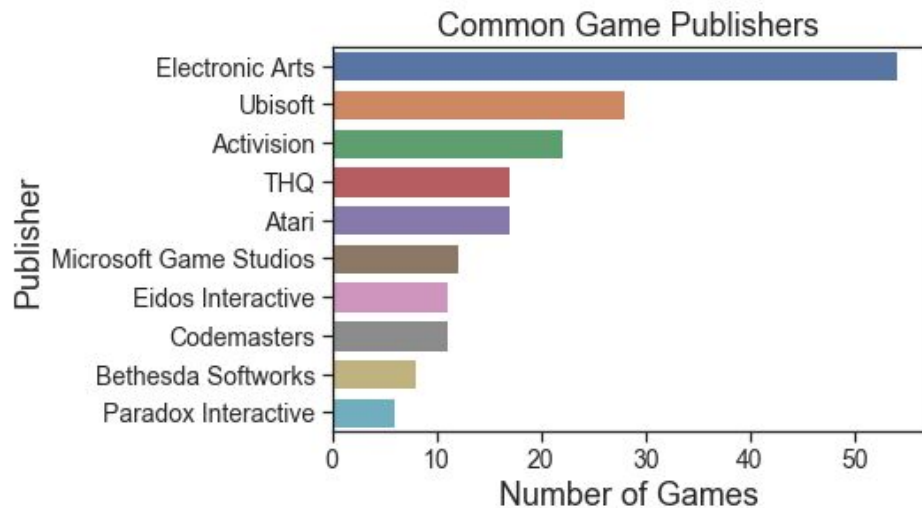
CRITIC- SCORE

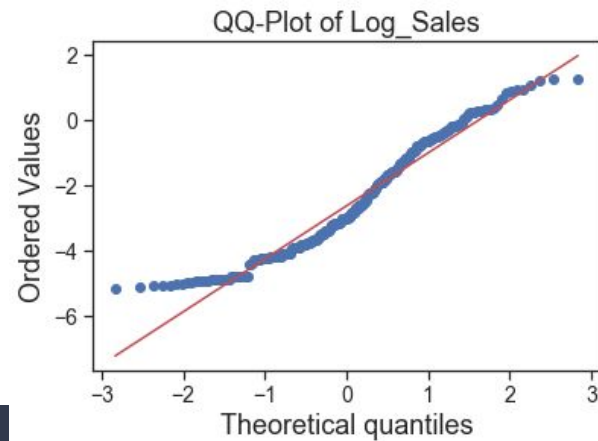
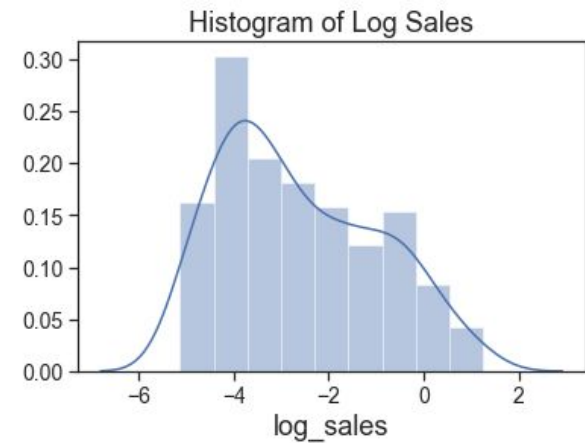
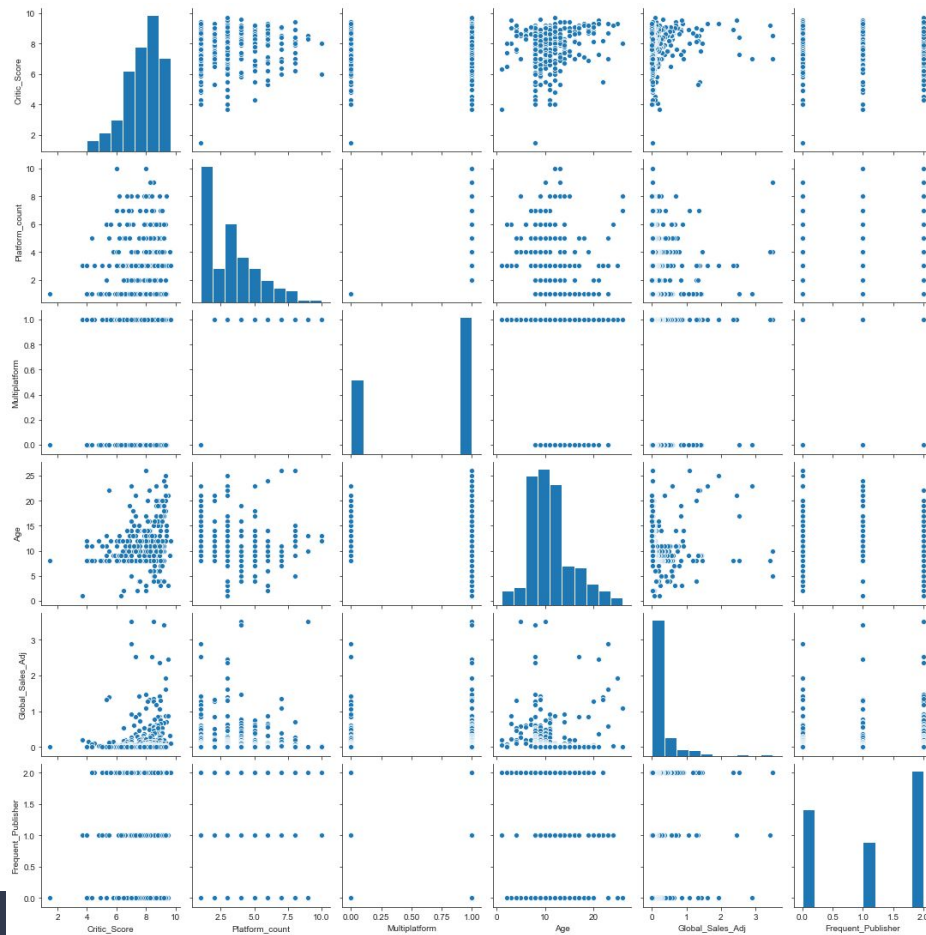
Starting Data from Kaggle

- **Started with:**
 - 55,792 entries x 23 variables
- **Ended with:**
 - 303 entries x 6 variables
 - **Variables:**
 - Name, Genre, ESRB Rating, Publisher, Critic Score, Global Sales
 - **Engineered:**
 - Multiplatform, Age, "*Frequent Publisher*",
 - Adjusted Global Sales
 - Dummies



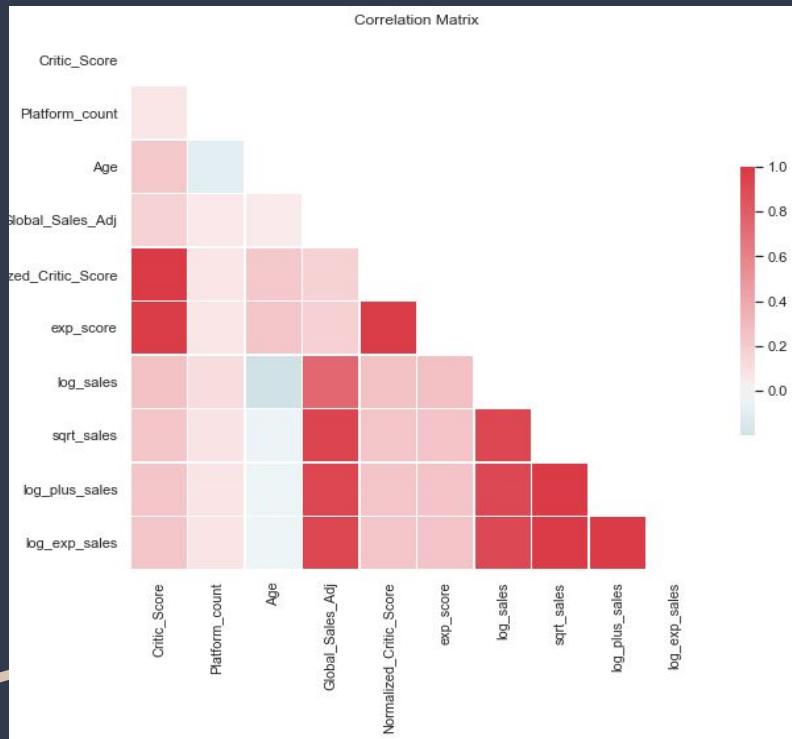






Data Graphics Continued

Statistical Tests



- The only highly correlated data (in red) are transformations of. Independent variables are safe to model with.

- ANOVA independent vars vs target

- Genre (*p-value: 0.0120*)
- Multiplatform (*p-value: 0.0018*)
- ESRB Ratings (*p-value: 0.04811*)

- Results

- Keep Genre and Multiplatform variables
- Drop ESRB Ratings

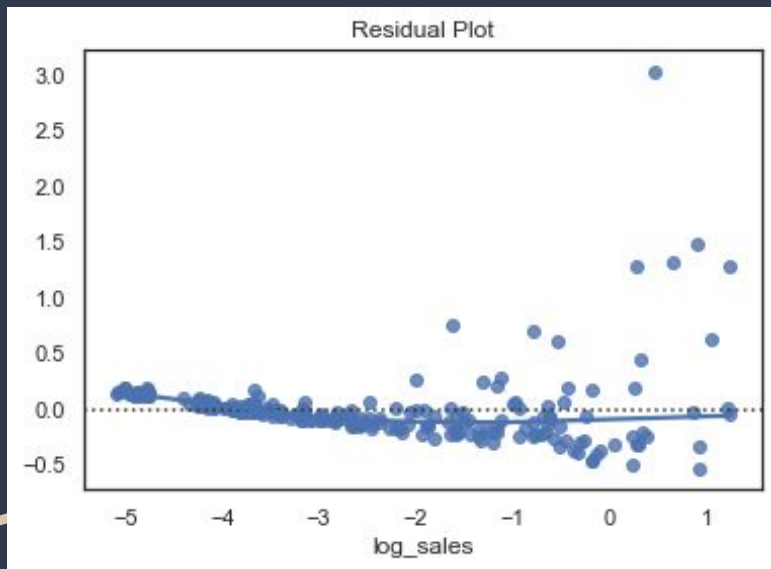
Regression Models

- Evaluation metric used was *Root Mean Square Error* and modeled using Lasso
- Best alpha level is at 0.4, which then selects 9 features
- Very High RMSE when average is 0.3

Lasso models at different alpha levels

	alpha	Number of Features	Training RMSE	Test RMSE
0	0.150	1	3.132060	3.235864
1	0.100	3	3.130799	3.236111
2	0.075	6	3.129578	3.235153
3	0.050	9	3.128189	3.232238
4	0.040	9	3.128419	3.231295
5	0.030	13	3.128884	3.230848
6	0.025	16	3.128747	3.230394
7	0.010	30	3.126165	3.238358

Regression Model Continued



- Selected Features:

	0
Age_13	-0.646847
Age_12	-0.137748
Genre_Action	-0.052290
Age_9	0.097800
Multiplatform_1	0.330200
Frequent_Publisher_2	0.364410
Genre_Simulation	0.417475
Age_8	0.748525
exp_score	1.157879

- Interpretation:
 - Age: references newly released
 - Genre: references compares again high frequency genre, Shooter.
 - Multiplatform games result in higher log sales
 - Frequent Publisher; references low frequency
 - As exp_score increases, log sales increases

Conclusion

- Lasso model has high error because of being unable to predict higher total sales values
- Age potentially can be an indicated that time is more important for total sales, which makes sense
- People trying to get into the PC Gaming business could use this model when building a game to decide to:
 - Make sure it is/becomes multiplatform
 - Make it a simulation game
 - Seek out popular publishers

Features	0
Age_13	-0.646847
Age_12	-0.137748
Genre_Action	-0.052290
Age_9	0.097800
Multiplatform_1	0.330200
Frequent_Publisher_2	0.364410
Genre_Simulation	0.417475
Age_8	0.748525
exp_score	1.157879