

Video Game Analysis Data Report

By: Robert Fusting

In this Analysis of two Kaggle datasets based around “gamer” behavior we will discuss the findings of 4 different statistical models which predict playtime and sales based on genre and publisher.

Model 1: Predicting playtime based on genre.

	Df	Sum Sq	Mean Sq	F Value	P(>F)
Genre	11	3499	318.1	87.94	<2e-16 ***
Residuals	11387	41189	3.6		

Looking at the summary output from our model we can see that we have a p value of essentially 0 which means there is strong evidence that the model is explaining playtime through genre. *Note this model is the log of playtime since it led to a better model due to the spread of playtime.

Now we can examine the means of the individual genres, to assess which genres are contributing to higher playtime. The below numbers are on the log scale so to better interpret the playtime we can exponentiate them.

Grand mean

2.068083

Action	Adventure	Fighting	Misc	Platform	Puzzle	Racing	Role-Playing
2.043	1.305	1.608	2.813	0.8474	0.625	1.574	1.968

Shooter	Simulation	Sports	Strategy
1.914	2.262	4.13	2.328

Table of Means (exponentiated)

Action	7.714
Adventure	3.688
Fighting	4.993
Misc.	16.660

Platform	2.334
Puzzle	1.868

Racing	4.826
Role-Playing	7.156
Shooter	6.780
Simulation	9.602
Sports	62.178
Strategy	10.257

From the above table we can note that the top 5 most played genres are: Sports, Misc, Strategy, Simulation, and Action. We can also note that the 3 least played are: Puzzle, Adventure, and Platform.

Model 2: Predicting playtime based on publisher.

	Df	Sum Sq	Mean Sq	F value	P(>F)
Publisher	23	4066	176.80	50.14	<2e-16 ***
Residuals	10184	35912	3.53		

Looking at the summary output from our model we can see that we have a p value of essentially 0 which means there is strong evidence that the model is explaining playtime through publisher. *Note this model is the log of playtime since it led to a better model due to the spread of playtime.

Now we can examine the means of the individual publishers, to assess which publishers are contributing to higher playtime. The below numbers are on the log scale so to better interpret the playtime we can exponentiate them.

Grand mean

2.157271

505 Games	Atari	Bethesda Softworks	Capcom	Codemasters	Deep Silver
1.979	1.693	2.746	1.383	1.591	1.953

Eidos Interactive	Electronic Arts	Iceberg Interactive	Kalypso Media
1.939	1.648	1.287	2.101

Sega	Sony Online Entertainment	Square Enix	Take-Two Interactive
3.931	1.726	1.742	2.55
Telltale Games	THQ	Ubisoft	Unknown
2.31	1.302	2.173	1.775
			Vivendi Games
			0.6785
Warner Bros. Interactive Entertainment			
2.097			

Table of Means (exponentiated)

505 Games	7.236
Atari	5.436
Bethesda Softworks	15.580
Capcom	3.987
Codemasters	4.909
Deep Silver	7.050
Eidos Interactive	6.952
Electronic Arts	5.197
Iceberg Interactive	3.622
Kalypso Media	8.174
Sega	50.958
Sony Online Entertainment	5.618
Square Enix	5.709
Take-Two Interactive	12.807
Telltale Games	10.074
THQ	3.677
Ubisoft	8.785
Unknown	5.783

Vivendi Games	1.971
Warner Bros. Interactive Entertainment	8.142

From the above table we can note that the top 5 most played publishers are: Sega, Bethesda Softworks, Take-Two Interactive, Telltale Games, and Ubisoft. We can also note that the 3 least played are: Vivendi Games, Iceberg Interactive, and THQ.

Model 3: Predicting sales based on Genre.

```

                Df Sum Sq Mean Sq F value    P(>F)
Genre           11  112.1   10.187    4.791 3.33e-07 ***
Residuals      945 2009.5    2.126

```

Looking at the summary output from our model we can see that we have a p value of essentially 0 which means there is strong evidence that the model is explaining sales through genre. *Note this model is the log of sales since it led to a better model due to the spread of sales.

Now we can examine the means of the individual genres, to assess which genres are contributing to higher sales. The below numbers are on the log scale so to better interpret the sales we can exponentiate them. It's also worth noting that all of these are negative because sales are in millions. Once we exponentiate them they will become positive decimals less than 1, because many games don't sell in the millions.

Grand mean

-2.692014

```

Action  Adventure Fighting  Misc Platform  Puzzle  Racing  Role-Playing
-2.577  -3.08   -3.526      -3.06   -3.274   -3.686   -3.267      -2.292

Shooter  Simulation  Sports  Strategy
      -2.39    -2.404      -2.755   -2.859

```

Table of Means (exponentiated)

Action	.076
Adventure	.046
Fighting	.0294
Misc.	.047

Platform	.038
Puzzle	.025

Racing	.038
Role-Playing	.101
Shooter	.092
Simulation	.090
Sports	.064
Strategy	.057

From the above table we can note that the top 5 bestselling genres are: Role-Playing, Shooter, Simulation, Sports, and Strategy. We can also note that the 3 least selling are: Puzzle, Platform, and Racing.

Model 4: Predicting sales based on Publisher.

	Df	Sum Sq	Mean Sq	F value	P(>F)
Publisher	15	170.4	11.357	5.217	7.18e-10 ***
Residuals	615	1338.8	2.177		

Looking at the summary output from our model we can see that we have a p value of essentially 0 which means there is strong evidence that the model is explaining sales through publisher. *Note this model is the log of sales since it led to a better model due to the spread of sales.

Now we can examine the means of the individual publishers, to assess which publishers are contributing to higher sales. The below numbers are on the log scale so to better interpret the sales we can exponentiate them. It's also worth noting that all of these are negative because sales are in millions. Once we exponentiate them they will become positive decimals less than 1, because many games don't sell in the millions.

Activision	Atari	Codemasters	Deep Silver	Eidos Interactive	Electronic Arts
-1.95	-3.348	-3.44	-2.696	-2.813	-2.013
Focus Home Interactive	Microsoft	Game Studios	Paradox Interactive	Sega	
-3.181		-2.65	-3.582	-2.1	
Square Enix					
-2.478					

Take-Two Interactive	THQ	Ubisoft	Unknown	Vivendi Games
-2.103	-2.753	-2.534	-3.264	-3.013

Table of Means (exponentiated)

Activision	.142
Atari	.035
Codemasters	.032
Deep Silver	.067
Eidos Interactive	.060
Electronic Arts	.134
Focus Home Interactive	.042
Microsoft Game Studios	.071
Paradox Interactive	.028
Sega	.122
Square Enix	.084
Take-Two Interactive	.122
THQ	.064
Ubisoft	.079
Unknown	.038
Vivendi Games	.049

From the above table we can note that the top 5 bestselling publishers are: Activision, Electronic Arts, Sega, Take-Two Interactive, and Square Enix. We can also note that the 3 least selling are: Paradox Interactive, Codemasters, and Atari.

Problems in the Models.

Before we conclude which genres and publishers should be prioritized by Microsoft, we must consider the problems in the models. The largest problem in both playtime models is due to the nature of the playtime dataset. Since the playtime dataset has many counts of the same game, but with different users,

popular games are dominating their respective genre or publisher. This is a violation of creating a model since the datapoints are not individual from each other. Another problem in the playtime models is the videogame Football Manager has some high playtimes. Football manager is both a sports game and published by Sega, and therefore both have very high means. The largest problem with both sales models is significance. Both sales models have overall small p-values, however, when we run the TukeyHSD test (see r files for more on this) we can see that the p-values between many of the genres and publishers on the individual scale has no significant difference.

Discussions of Findings.

Despite the problems in both models the genres that appeared in the top 5 of sales and playtime were simulation, sports, and strategy. The publishers that appeared in the top 5 sales and playtime were Sega and Take-Two Interactive. Given these genres and publishers performed well in both models, we can recommend that Microsoft prioritize these genres and publishers in their game pass.

We can also observe that puzzle and platform genres performed badly in both sales and playtime models. We would advise that Microsoft avoid spending large amounts of money trying to purchase games that fall under these genres.