Which “Genre” of Entertainment Would You Prefer?

- **Group members**
  Kexin Fang(kfang5), Zhaochen Gao(zgao22), Tianying Zhou(tzhou26)

- **Research Interest:**

  -As we got more and more ways to entertain ourselves, we became curious about one question: is there a particular kind of entertainment that people prefer? In order to solve this problem, we picked two popular approaches: videogames and movies. They seemed different at the first sight, but they share a common character: genre. For some genres, they could be applied to both videogames and movies; and for some genres, we could roughly categorize them as the same, such as “Sports” in video games and “Action” in movies.

  -By using merge and sort techniques, we could eliminate useless variables, generate relationship between genres and sales or feedback from the market of video games and movies.

  -There could be missing data, but since they are popular topics, more information could be easily accessed by using the Internet.

- **Background Information:**

  These data sets are all retrieved from the internet. They are uploaded by different individual users on Kaggle.com. The first file is the result of crawl on http://www.ign.com/games/reviews.

  The second file is a scrape of http://www.vgchartz.com/. The third file is a scrape on IMDB website by using a Python library called "scrapy". The code for “scrapy” is posted on GitHub: https://github.com/sundeepblue/movie_rating_prediction/blob/master/movie/spiders/movie_budget_spider.py.

  -File 1 20-years-of-games
  https://www.kaggle.com/egrinstein/20-years-of-games
  -File 2 videogamessales
  https://www.kaggle.com/gregorut/videogamesales
  -File 3 imdb-5000-movie
  https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

- **Methods section:**
  Original Data Files

For all three raw data files that we use, they all have huge amount of observations and variables. We would select only the variables that we need.

"Ign.csv"

The raw data contains 18625 observations and 10 variables.Four variables that are not very related are dropped and left 18625 observations and 6 variables. Those 6 variables provide information about editors_choice, genre, release_year, score, score_phrase and title. It creates a dataset called "project_game".

"Vgsales.csv"

The raw data contains 16598 observations and 11 variables.Four variables that are not very related are dropped and left 16598 observations and 7 variables. Those 7 variables provide information about Genre, Name, Genre, EU_Sales, JP_Sales, NA_Sales and Global_Sales.It creates a dataset called"project_vsales".

"Metadata.csv"

The raw data contains 5043 observations and 28 variables. 16 variables that are not very related are dropped and left 5043 observations and 12 variables. Those 12 variables provide information about actor_1_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes, cast_total_facebook_likes, country, director_facebook_likes, genres, imdb_score, movie_facebook_likes, movie_title, num_critic_for_reviews, title_year. It creates a dataset called"project_movie".

Data Validation

File1: "ign.csv"-"project_game"-"game1"

The "proc freq" technique is used twice to make sure different genres(70 in all) are roughly categorized into 7 different genres. It also helped us to find 36 missing values that we would fill out later. The "proc univariate" technique is also used to make sure there are no extreme values.

File2: "vgsales"-"project_vsales"-"game2"

The "proc freq" technique is used twice to make sure different genres(7 in all) are roughly categorized into 7 different genres. Its genre type is also the our "standard" list. The "proc univariate" technique is also used to make sure there are no extreme values.

File3: "metadata"-"project_movie"-"movie"

The "proc freq" technique is used twice to make sure different genres(more than 200) are roughly categorized into 7 different genres.The "proc univariate" technique is also used to make sure there are no extreme values.

Data Clean

In both “project_game” and “project_movie” datasets, we found missing values after “proc freq”. In “project_game” dataset, there are 36 missing values in the genre variable, so we used the internet to find out each missing genre and fill out our “project_game” dataset by looking at the output. In “project_movie” datasets, the missing values are all belong to those dropped variable. Therefore, we ignored those missing values.

Additional Steps.

To process and review datasets, we self-studied to input ‘csv.’ profiles into SAS and then started to clean the data. After cleaning three datasets, we first merged ‘game_analysis’ and ‘score_movie’ by genres and used the IF-THEN loop to get the average of scores. Therefore we were able to compare the average scores for both movies and video games of the same genre. Then we used the similar way to merge the new combination of datasets with the third dataset ‘game2_analysis’ to add the sales of video games for each type. We sorted the final dataset to observe the relationship between the genre of entertainment and its popularity.

Variables Analyzed

While dealing with the original data sets, we picked variables that we think are most relevant: genre and score in “project_game”, genre and sales in “project_vsales” and genre and score in “project_movie”. For our final comparison, the three major variables are the average sales/score of each genre in each datasets. By analyzing those numeric values, we could conclude which genre is the best-seller and achieve the highest score.

- **Results Section:**

  Charts and tables pertaining to validation and cleaning

1. Descriptor Portion of Datasets

Data Set Name	WORK.PROJECT_GAME	**Observations**	18625
Member Type	DATA	**Variables**	7
Engine	V9	**Indexes**	0

Created	12/14/2016 16:46:02	**Observation Length**	104
Last Modified	12/14/2016 16:46:02	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	15
First Data Page	1
Max Obs per Page	1258
Obs in First Data Page	1232
Number of Data Set Repairs	0
Filename	/saswork/SAS_work6B0E00007661_odaws02-prod-us/SAS_work4E0D00007661_odaws02-prod-us/project_game.sas7bdat
Release Created	9.0401M3

Host Created	Linux
Inode Number	14417961
Access Permission	rw-r--r--
Owner Name	zgao220
File Size	2MB
File Size (bytes)	2097152

Alphabetic List of Variables and Attributes					
#	**Variable**	**Type**	**Len**	**Format**	**Informat**
1	,	Num	8	BEST12.	BEST32.
6	editors_choice	Char	1	$1.	$1.
5	genre	Char	19	$19.	$19.
7	release_year	Num	8	BEST12.	BEST32.
4	score	Num	8	BEST12.	BEST32.
2	score_phrase	Char	8	$8.	$8.
3	title	Char	52	$52.	$52.

We also used PROC CONTENTS procedure for the other two datasets to get a general idea about those three raw datasets in the first step.

2. Frequency table for dataset project_game

genre	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**

Action	3797	20.43	3797	20.43
Action, Adventure	765	4.12	4562	24.54
Action, Compilation	89	0.48	4651	25.02
Action, Editor	1	0.01	4652	25.03
Action, Platformer	3	0.02	4655	25.04
Action, Puzzle	1	0.01	4656	25.05
Action, RPG	330	1.78	4986	26.82
Action, Simulation	32	0.17	5018	26.99
Action, Strategy	1	0.01	5019	27.00
Adult, Card	2	0.01	5021	27.01
Adventure	1175	6.32	6196	33.33
Adventure, Adult	1	0.01	6197	33.34
Adventure, Adventur	5	0.03	6202	33.36
Adventure, Compilat	11	0.06	6213	33.42
Adventure, Episodic	4	0.02	6217	33.44
Adventure, Platform	1	0.01	6218	33.45
Adventure, RPG	3	0.02	6221	33.47
Baseball	1	0.01	6222	33.47
Battle	32	0.17	6254	33.64
Board	116	0.62	6370	34.27
Board, Compilation	7	0.04	6377	34.31

Card	108	0.58	6485	34.89
Card, Battle	54	0.29	6539	35.18
Card, Compilation	3	0.02	6542	35.19
Card, RPG	9	0.05	6551	35.24
Casino	31	0.17	6582	35.41
Compilation	54	0.29	6636	35.70
Compilation, Compil	1	0.01	6637	35.70
Compilation, RPG	2	0.01	6639	35.71
Educational	20	0.11	6659	35.82
Educational, Action	11	0.06	6670	35.88
Educational, Advent	3	0.02	6673	35.90
Educational, Card	1	0.01	6674	35.90
Educational, Produc	5	0.03	6679	35.93
Educational, Puzzle	25	0.13	6704	36.06
Educational, Simula	2	0.01	6706	36.08
Educational, Trivia	2	0.01	6708	36.09
Fighting	547	2.94	7255	39.03
Fighting, Action	77	0.41	7332	39.44
Fighting, Adventure	5	0.03	7337	39.47
Fighting, Compilati	13	0.07	7350	39.54

Fighting, RPG	2	0.01	7352	39.55
Fighting, Simulatio	3	0.02	7355	39.57
Flight	24	0.13	7379	39.70
Flight, Action	125	0.67	7504	40.37
Flight, Racing	3	0.02	7507	40.38
Flight, Simulation	37	0.20	7544	40.58
Hardware	2	0.01	7546	40.59
Hunting	112	0.60	7658	41.20
Hunting, Action	2	0.01	7660	41.21
Hunting, Simulation	1	0.01	7661	41.21
Music	371	2.00	8032	43.21
Music, Action	39	0.21	8071	43.42
Music, Adventure	1	0.01	8072	43.42
Music, Compilation	4	0.02	8076	43.45
Music, Editor	6	0.03	8082	43.48
Music, RPG	1	0.01	8083	43.48
Other	20	0.11	8103	43.59
Other, Action	1	0.01	8104	43.60
Other, Adventure	1	0.01	8105	43.60
Party	141	0.76	8246	44.36
Pinball	77	0.41	8323	44.77

Pinball, Compilatio	1	0.01	8324	44.78
Platformer	823	4.43	9147	49.21
Platformer, Action	11	0.06	9158	49.27
Platformer, Adventu	8	0.04	9166	49.31
Productivity	39	0.21	9205	49.52
Productivity, Actio	2	0.01	9207	49.53
Puzzle	776	4.17	9983	53.70
Puzzle, Action	200	1.08	10183	54.78
Puzzle, Adventure	47	0.25	10230	55.03
Puzzle, Compilation	9	0.05	10239	55.08
Puzzle, Platformer	1	0.01	10240	55.09
Puzzle, RPG	1	0.01	10241	55.09
Puzzle, Word Game	6	0.03	10247	55.12
RPG	980	5.27	11227	60.40
RPG, Action	1	0.01	11228	60.40
RPG, Compilation	4	0.02	11232	60.42
RPG, Editor	2	0.01	11234	60.43
RPG, Simulation	8	0.04	11242	60.48
Racing	1228	6.61	12470	67.08
Racing, Action	210	1.13	12680	68.21

Racing, Compilation	2	0.01	12682	68.22
Racing, Editor	3	0.02	12685	68.24
Racing, Shooter	2	0.01	12687	68.25
Racing, Simulation	25	0.13	12712	68.38
Shooter	1610	8.66	14322	77.05
Shooter, Adventure	1	0.01	14323	77.05
Shooter, First-Pers	4	0.02	14327	77.07
Shooter, Platformer	3	0.02	14330	77.09
Shooter, RPG	22	0.12	14352	77.21
Simulation	567	3.05	14919	80.26
Simulation, Adventu	1	0.01	14920	80.26
Sports	1916	10.31	16836	90.57
Sports, Action	196	1.05	17032	91.62
Sports, Baseball	3	0.02	17035	91.64
Sports, Compilation	14	0.08	17049	91.72
Sports, Editor	1	0.01	17050	91.72
Sports, Fighting	1	0.01	17051	91.73
Sports, Golf	1	0.01	17052	91.73
Sports, Other	1	0.01	17053	91.74
Sports, Party	1	0.01	17054	91.74
Sports, Racing	5	0.03	17059	91.77

Sports, Simulation	44	0.24	17103	92.01
Strategy	1071	5.76	18174	97.77
Strategy, Compilati	1	0.01	18175	97.77
Strategy, RPG	77	0.41	18252	98.19
Strategy, Simulatio	1	0.01	18253	98.19
Trivia	119	0.64	18372	98.83
Virtual Pet	82	0.44	18454	99.27
Wrestling	134	0.72	18588	99.99
Wrestling, Simulati	1	0.01	18589	100.00
Frequency Missing = 36				

The FREQ procedure shows there are 36 missing values for variable genre.

3. Check Extreme Values of Three Datasets

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.5	5243	10	18434
0.7	891	10	18435
0.8	12514	10	18512
1.0	16410	10	18624
1.0	16373	10	18625

In order to check extreme values in three datasets, we did PROC UNIVARIATE focusing on the numeric values in each dataset: score, sales and score. All three datasets show no abnormal data.

4. Calculated Average of Genre’s Score for Dataset game1

genre	VG_Avg
Action	6.889
Adventure	6.879
Misc	6.980
Puzzle	7.097
Role-Playing	7.266
Simulation	6.802
Strategy	7.169

We use IF-THEN loop to calculate the average score for seven genres based on game1 dataset. We found out genre “Role-Playing” has the highest score. Genre “Strategy” is in second place and genre “Puzzle” is in third place.

5. Calculated Average of Genre’s Score for Dataset movie

genre	Movie_Ave
Action	6.240
Adventure	6.525
Misc	6.753
Puzzle	5.872
Role-Playing	7.133

Simulation	7.011
Strategy	6.907

We use IF-THEN loop to calculate the average score for seven genres based on movie dataset. We found out genre “Role-Playing” has the highest average score. Genre “Simulation” is in second place and genre “Strategy” is in third Place.

6. Calculated Average of Genre’s sales for Dataset game2

Genre	Vg_sales_Ave
Action	0.52810
Adventure	0.18588
Misc	0.46576
Puzzle	0.42088
Role-Playing	0.62323
Simulation	0.45236
Strategy	0.25585

We use IF-THEN loop to calculate the average sales for seven genres based on game2 dataset. We found out genre “Role-Playing” has the highest average sales. Genre “Action” is in second place and genre “Misc” is in third Place.

7. Overall Analysis

Obs	genre	Movie_Ave	VG_Avg	Vg_sales_Ave
1	Role-Playing	7.133	7.266	0.62323
2	Simulation	7.011	6.802	0.45236

3	Strategy	6.907	7.169	0.25585
4	Misc	6.753	6.980	0.46576
5	Adventure	6.525	6.879	0.18588
6	Action	6.240	6.889	0.52810
7	Puzzle	5.872	7.097	0.42088

In general, we can found out genre “Role-Playing” has the highest rank in all of our datasets. Therefore, we can conclude that Genre “Role-Playing” is most favorite genre for most of people. This result may be owing to that this kind of games encourage players to become a character—often one who is very different from a player’s real-life persona. They could help player increased empathy toward people with different lifestyles or appearances and make player develop critical thinking when facing ongoing challenges.