Exploratory Data Analysis for Machine Learning

IBM Machine Learning -
Project 1
PRASON SOOD
JUNE 2021

About the data

- The data originally came from the Board Game Geek database, including 90,000+ board games, their description, and ratings.
- This data set was collected by R for Data Science (R4DS) - Online Learning Community and posted on their GitHub in March 2019. The .csv file can be found in Tidy Tuesday repository.

Data exploration plan

This analysis is the initial step in an attempt to build a baseline model to predict game average ratings based on their characteristics.

1. Data Overview
2. Data Cleaning and Feature Engineering: Categorical Data
3. Data Cleaning and Feature Engineering: Numeric Data
4. Hypothesis Testing

Data overview

- The train set has 8,425 rows and 22 columns
- There are missing data only in most of the categorical variables

game_id	0
year_published	0
average_rating	0
playing_time	0
name	0
min_playtime	0
users_rated	0
min_age	0
max_playtime	0
max_players	0
description	0
min_players	0
image	1
thumbnail	1
publisher	2
category	79
designer	94
mechanic	751
artist	2238
family	2255
expansion	6236
compilation	8103

Categorical data

1. Data Cleaning:

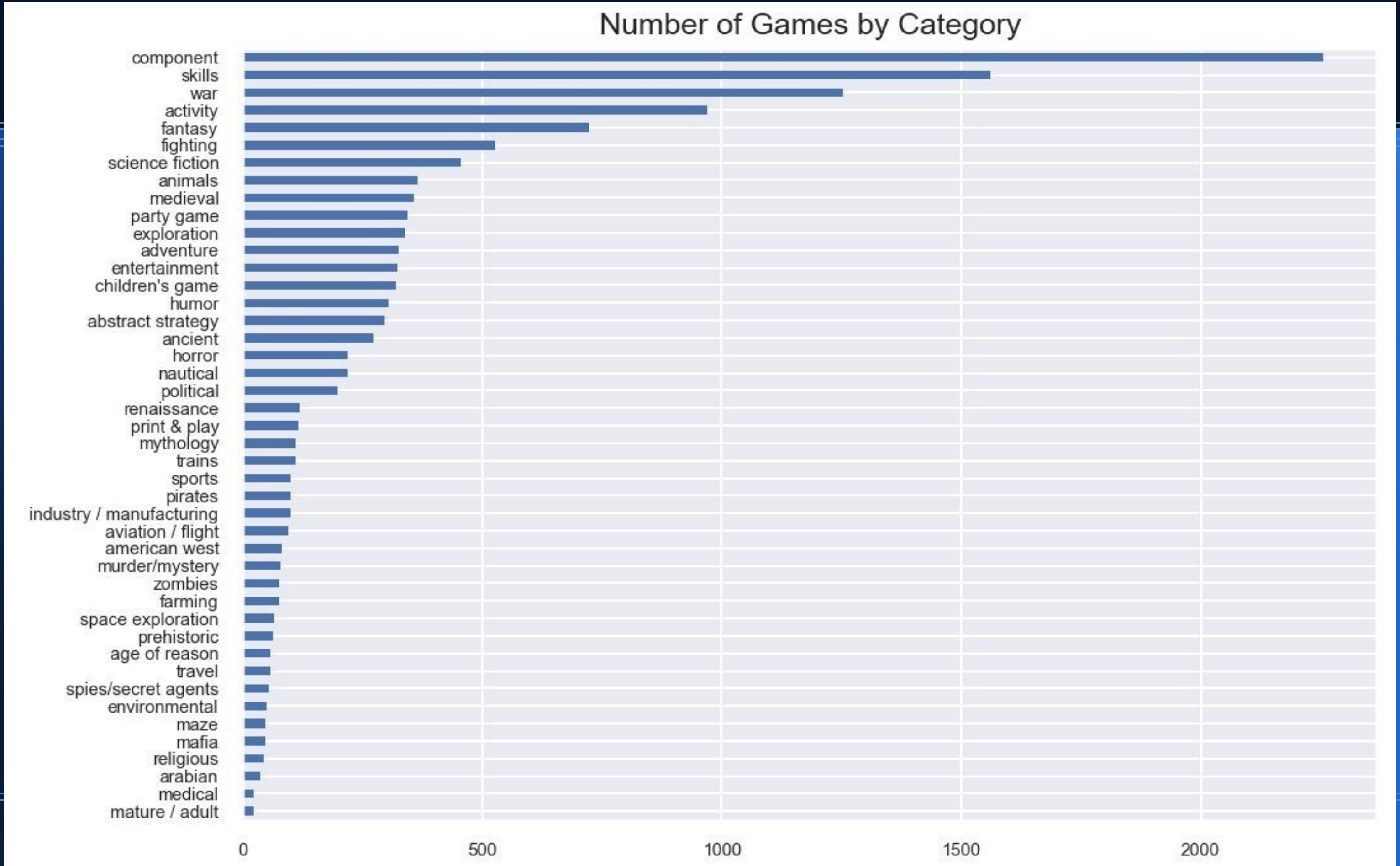
- Remove features that are not useful to discriminate the target: *description, image, name, thumbnail, family, expansion,* and *compilation*
- Also remove *game_id*

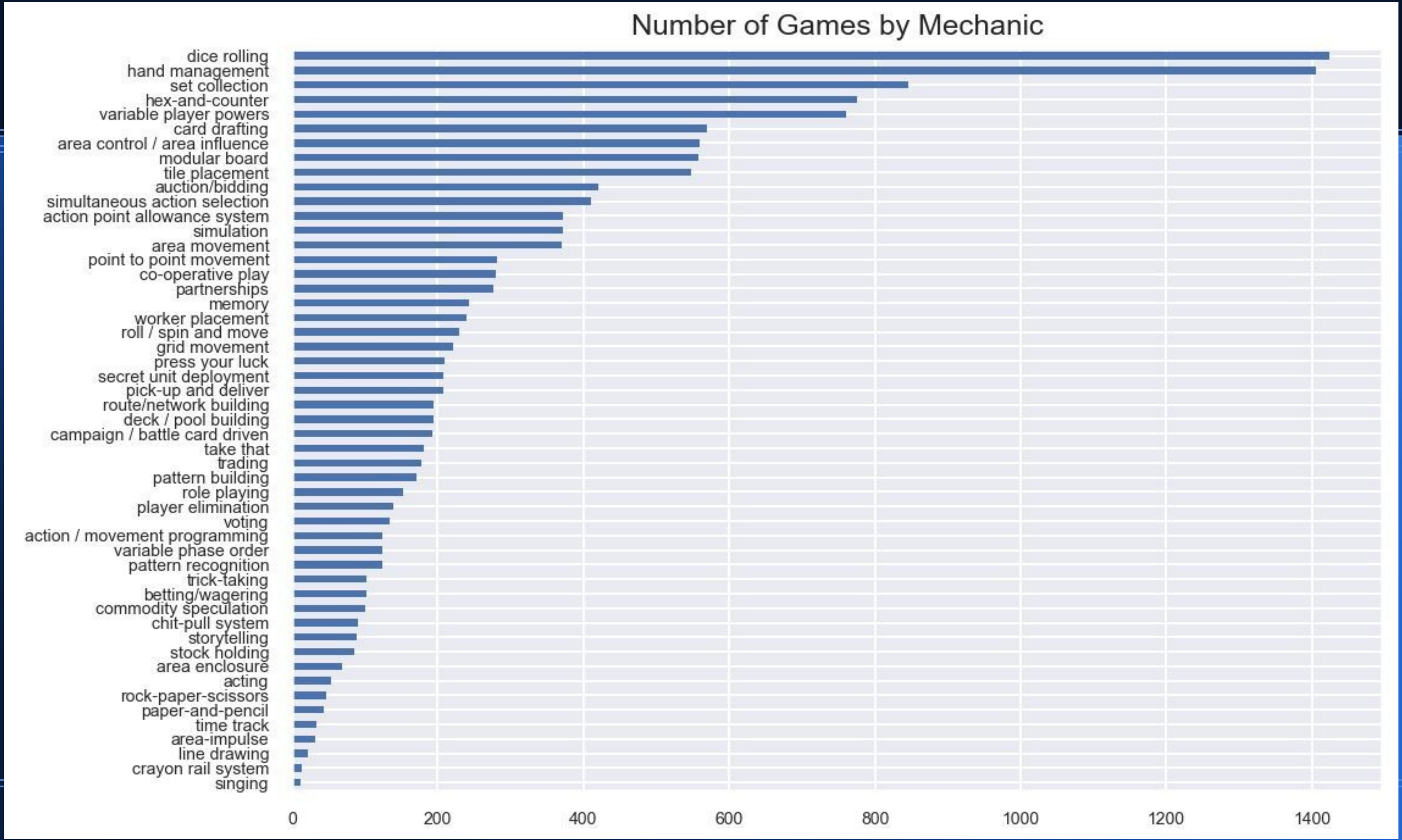
	count	unique	top	freq
description	8425	8423	How could that have happened? Black Stories ar...	2
image	8424	8422	//cf.geekdo-images.com/images/pic2262580.png	2
name	8425	8314	Robin Hood	5
thumbnail	8424	8422	//cf.geekdo-images.com/images/pic2410035_t.png	2
artist	6187	3881	Franz Vohwinkel	141
category	8346	3310	Wargame,World War II	364
compilation	322	269	Traveller: The Classic Games, Games 1-6+	6
designer	8331	3978	(Uncredited)	442
expansion	2189	2106	Règlement de l'An XXX,Regulations of the Year ...	7
family	6170	3321	Crowdfunding: Kickstarter	312
mechanic	7674	2708	Hex-and-Counter	406
publisher	8423	4538	GMT Games	140

Categorical data

Categories derived from category aggregates

- Get a set of all unique values in each variable
- Create new columns based on these values


Number of Games by Category
component
skills
war
activity
fantasy
fighting
science fiction
animals
medieval
party game
exploration
adventure
entertainment
children's game
humor
abstract strategy
ancient
horror
nautical
political
renaissance
print & play
mythology
trains
sports
pirates
industry / manufacturing
aviation / flight
american west
murder/mystery
zombies
farming
space exploration
prehistoric
age of reason
travel
spies/secret agents
environmental
maze
mafia
religious
arabian
medical
mature / adult
0
500
1000
1500
2000


Number of Games by Mechanic
dice rolling
hand management
set collection
hex-and-counter
variable player powers
card drafting
area control / area influence
modular board
tile placement
auction/bidding
simultaneous action selection
action point allowance system
simulation
area movement
point to point movement
co-operative play
partnerships
memory
worker placement
roll / spin and move
grid movement
press your luck
secret unit deployment
pick-up and deliver
route/network building
deck / pool building
campaign / battle card driven
take that
trading
pattern building
role playing
player elimination
voting
action / movement programming
variable phase order
pattern recognition
trick-taking
betting/wagering
commodity speculation
chit-pull system
storytelling
stock holding
area enclosure
acting
rock-paper-scissors
paper-and-pencil
time track
area-impulse
line drawing
crayon rail system
singing
0
200
400
600
800
1000
1200
1400

Numeric data

Data description

	max_players	max_playtime	min_age	min_players	min_playtime	playing_time	year_published	average_rating	users_rated	num_artist	num_category	num_designer	num_mechanic	num_publisher
count	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000	5608.000000
mean	5.010521	105.758559	9.955599	2.059379	91.313302	105.758559	2004.717725	6.546314	1166.660663	2.203994	2.651926	1.411733	2.600927	2.824893
std	7.543777	866.538797	3.301289	0.674542	848.267125	866.538797	11.284651	0.775103	3548.581155	7.690679	1.300462	0.802652	1.501255	3.683774
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1951.000000	2.339400	50.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	4.000000	30.000000	8.000000	2.000000	30.000000	30.000000	2001.000000	6.051200	100.000000	1.000000	2.000000	1.000000	1.000000	1.000000
50%	4.000000	45.000000	10.000000	2.000000	45.000000	45.000000	2009.000000	6.548855	237.000000	1.000000	2.000000	1.000000	2.000000	2.000000
75%	6.000000	90.000000	12.000000	2.000000	90.000000	90.000000	2013.000000	7.065962	755.250000	2.000000	3.000000	2.000000	3.000000	3.000000
max	362.000000	60000.000000	25.000000	9.000000	60000.000000	60000.000000	2016.000000	9.003920	67655.000000	510.000000	14.000000	11.000000	12.000000	92.000000

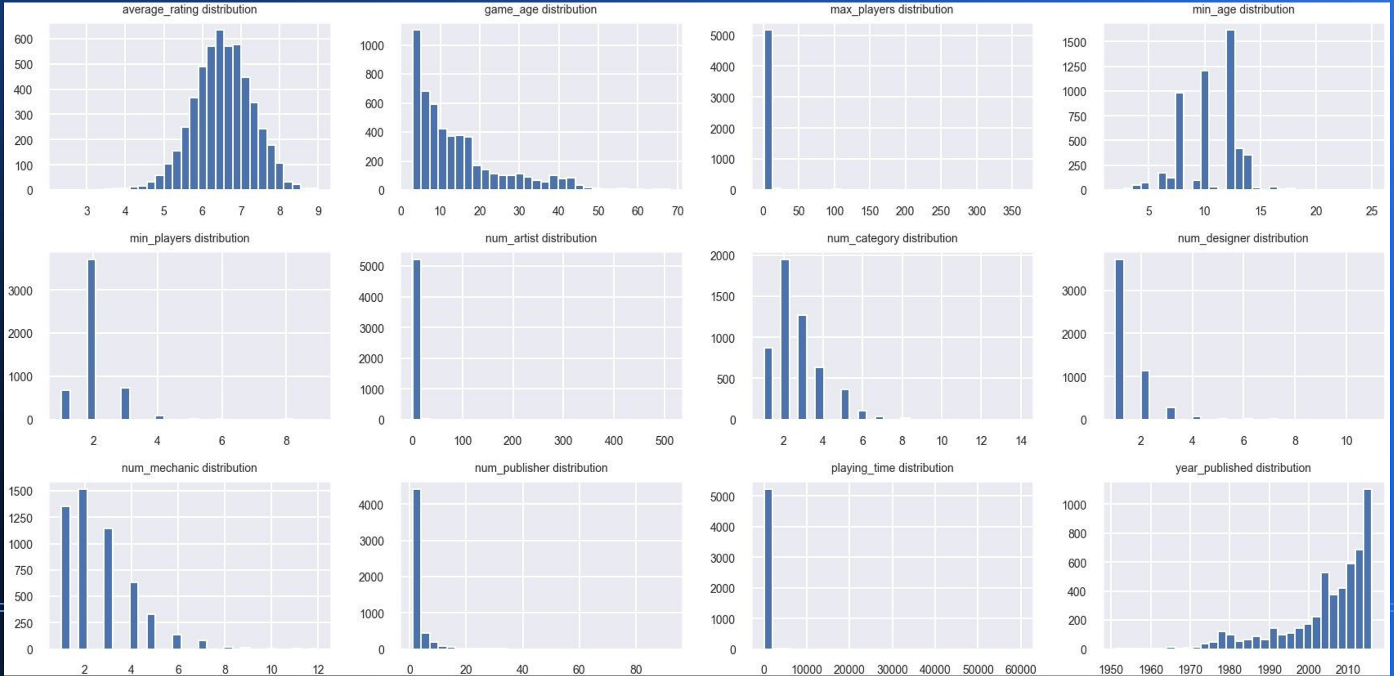
Numeric data

1. Data cleaning

- Derive *game_age* from *year_published*
- Remove *max_playtime, min_playtime,* and *users_rated*
- Select data that have non zero values

	max_players	min_age	min_players	playing_time	year_published	average_rating	num_artist	num_category	num_designer	num_mechanic	num_publisher	game_age
count	5240.000000	5240.000000	5240.000000	5240.000000	5240.000000	5240.000000	5240.000000	5240.000000	5240.000000	5240.000000	5240.000000	5240.000000
mean	5.102290	10.471756	2.070611	101.502481	2004.554008	6.525332	2.213550	2.663740	1.406870	2.624046	2.909542	14.445992
std	7.753493	2.441990	0.666585	858.286053	11.397775	0.765409	7.936809	1.316168	0.794444	1.512648	3.783639	11.397775
min	1.000000	2.000000	1.000000	1.000000	1951.000000	2.339400	1.000000	1.000000	1.000000	1.000000	1.000000	3.000000
25%	4.000000	8.000000	2.000000	30.000000	2000.000000	6.036300	1.000000	2.000000	1.000000	1.000000	1.000000	6.000000
50%	4.000000	10.000000	2.000000	45.000000	2009.000000	6.525335	1.000000	2.000000	1.000000	2.000000	2.000000	10.000000
75%	6.000000	12.000000	2.000000	90.000000	2013.000000	7.032408	2.000000	3.000000	2.000000	3.000000	3.000000	19.000000
max	362.000000	25.000000	9.000000	60000.000000	2016.000000	9.003920	510.000000	14.000000	11.000000	12.000000	92.000000	68.000000

Numeric data



Numeric data

- The target (*average_rating*) has a normal distribution
- Most features are right skewed
- Severe outliers

Numeric data

2. Feature engineering

Log transformation for skewed variables

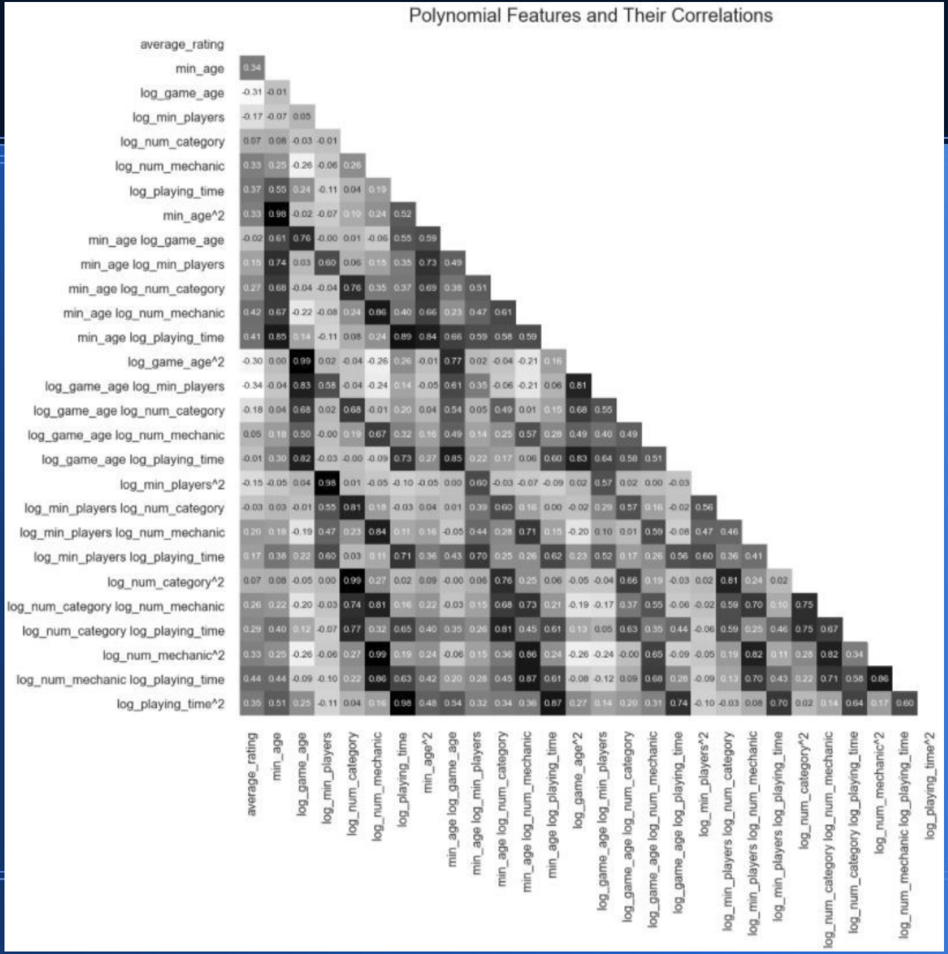
- Apply log transformation and check for skewness again.
- The result shows that log transformation does not work well for n*um_artist, num_designer, num_publisher,* and *year_published*

Next page present a pairplot of numeric features that have nearly normal distribution.

Numeric data

Adding polynomial and interaction terms

- This plot shows that polynomial and interaction terms do not have significantly higher correlations with the target comparing to the original features



Numeric data

Binning numeric data that cannot be scaled by log transformation

- These are *num_artist, num_designer, num_publisher,* and *year_published*
- Apply dummy transformation to these bins
- New columns from these bins: *group_artist_three_or_more, group_designer_three_or_more, group_max_players_five_or_six, group_max_players_seven_or_more, group_publisher_four_or_more, group_year_published_between_2001_and_2009, group_year_published_between_2010_and_2013,* and *group_year_published_between_2014_and_2016*

Hypothesis testing

- Main purpose: check if there are differences in average ratings between one group and others
- Due to different variances between two groups, Welch's t-test is used
- Perform multiple tests across all categories, mechanics, and groups (derived from numeric data)
- Sample of hypotheses:
  - H_0: War games and other games have similar ratings on average
  - H_a: There is a difference in average ratings between war games and other games

Hypothesis testing

- Result tables are shown on the next page. These values are sorted by p-values with colored bars (green for positive values and red for negative ones)
- For those that have $p\text{-value} < 0.05$ and $|t\text{-value}| > 1.96$, we reject the null hypotheses
- The sign of t-value suggests the direction of the test. A positive sign means that the group of interest has higher average ratings than others. On the contrary, a negative sign means that the group of interest has lower average ratings than others.

category_name	t-value	p-value
children's game	-15.841916	0.000000
war	13.893726	0.000000
component	-10.584794	0.000000
humor	-9.138182	0.000000
party game	-7.005245	0.000000
animals	-6.487482	0.000000
trains	4.741813	0.000006
renaissance	4.690531	0.000007
activity	4.241476	0.000024
space exploration	4.478895	0.000032
fighting	3.980679	0.000077
industry / manufacturing	4.088935	0.000090
age of reason	3.980063	0.000221
ancient	3.669080	0.000289
abstract strategy	-3.629330	0.000328
medieval	3.532035	0.000461
fantasy	3.358155	0.000818
farming	3.341165	0.001299
science fiction	2.937326	0.003463
nautical	2.900014	0.004100

maze	-3.009117	0.004285
pirates	-2.329630	0.021775
political	2.278630	0.023735
mythology	2.144308	0.034218
spies/secret agents	2.004349	0.050047
entertainment	-1.937038	0.053610
religious	1.763866	0.084909
print & play	1.709978	0.090178
aviation / flight	1.708230	0.091375
skills	-1.677003	0.093654
exploration	1.409116	0.159621
environmental	1.298456	0.200263
adventure	0.972847	0.331317
mature / adult	-0.791506	0.437057
arabian	-0.638325	0.527225
murder/mystery	0.600392	0.550026
horror	-0.465960	0.641699
sports	0.423544	0.672805
medical	0.424722	0.675144
travel	0.345949	0.730699
prehistoric	-0.332909	0.740358
american west	0.231495	0.817530
mafia	-0.115298	0.908744
zombies	-0.002223	0.998233

Hypothesis testing

We can conclude by these tables thatt on average:

- People generally like war games
- People do not like children's games and component games.
- People like games that use area control / area influence, worker placement, simulation, variable player powers, and deck / pool building

group_name	t-value	p-value
year_published_between_2014_and_2016	21.686049	0.000000
artist_three_or_more	10.349241	0.000000
year_published_between_2001_and_2009	-9.785744	0.000000
max_players_five_or_six	-9.068495	0.000000
publisher_four_or_more	6.625750	0.000000
year_published_between_2010_and_2013	5.153826	0.000000
designer_three_or_more	2.842000	0.004688
max_players_seven_or_more	-2.561497	0.010632

Hypothesis testing

- Since these features might have effects on each other, there need to be more analyses before jumping to a conclusion. For example, perhaps area control mechanic is mostly used in war games, or children’s games are mostly played by rolling and spinning. War games might be more complex and need more artists to complete.

Conclusion

WE can conclude that LINEAR REGRESSION might not be the best but workable algorthms for this dataset
jupyter Notebook for this analysis can be found here:
https://github.com/prason3106/IBM-PROJECT/blob/main/Project-1.ipynb