Predicting Video Game Sales

STAT 204 Fall 2019

Meltem Ozcan
Gulzina Kuttubekova

The Data

“Video Game Sales with Ratings” dataset available on Kaggle:
https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/download

16719 observations, 16 variables - Metacritic & VGChartz

Categorical Variables:

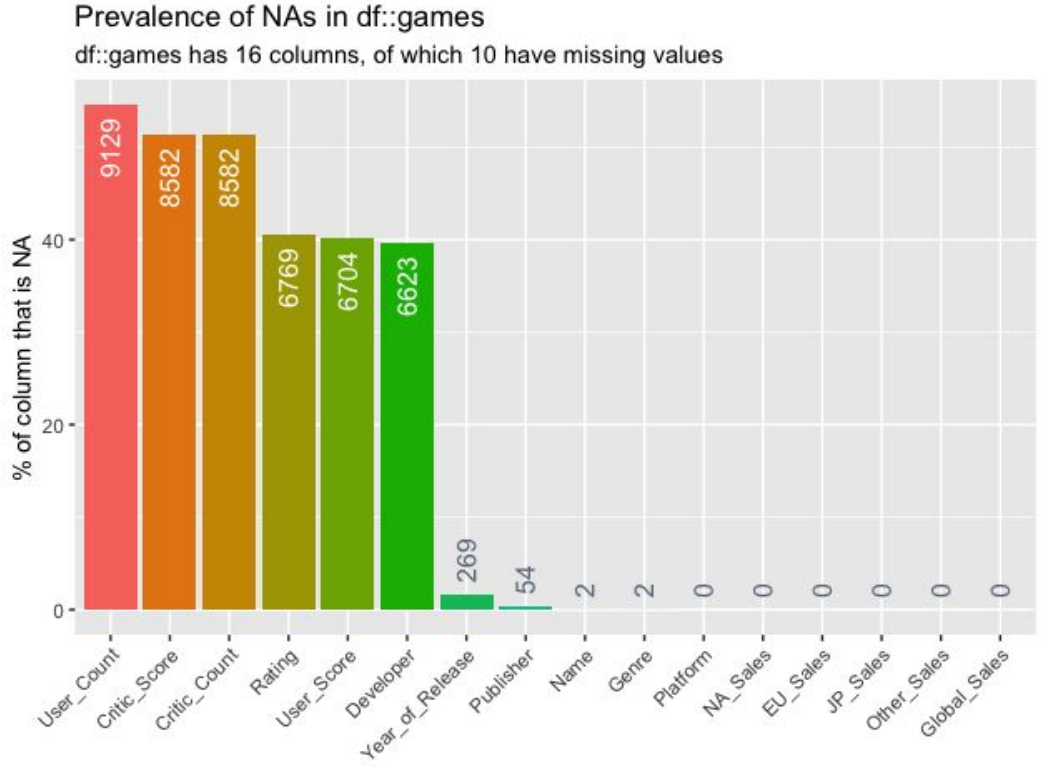
Name, Platform, Genre, Publisher, Developer, Rating

Numerical Variables:

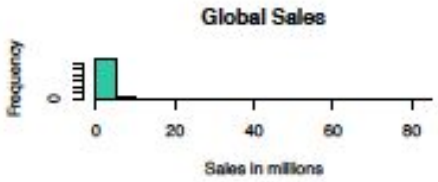
Year, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count

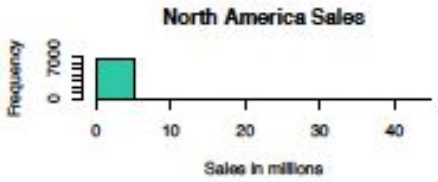
Data Cleaning & Manipulation

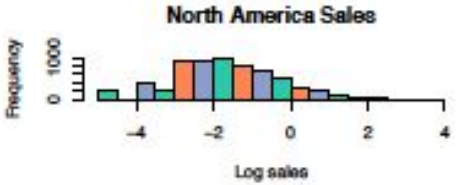
- Treat blanks and missing value as NA
- From 16719 to **6825** complete cases

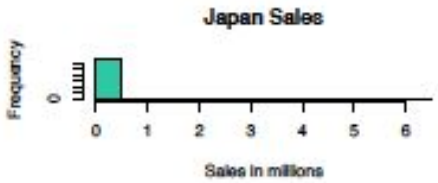


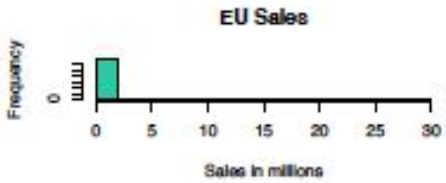
Transformation of the sales variables

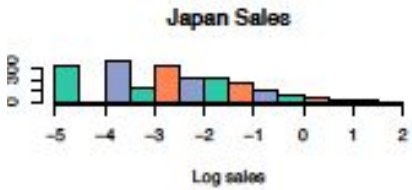


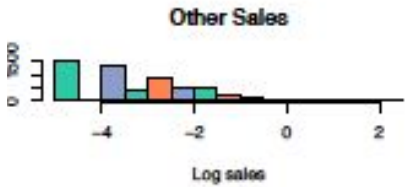












New variables

Platform_Generation: 5th, 6th, 7th, 8th

Family_of_Systems: Nintendo_Wii, Nintendo_DS, SONY_PS, Microsoft_XBox

Platform_Company: Nintendo, SONY, Microsoft and SEGA

Developer_Company: Nintendo, Microsoft, SONY, SEGA, Ubisoft, Activision Blizzard, Tencent, Namco, EA, Square Enix, Other

Developer_Country: USA, Japan, France, China, Other

Years_Since_Release: 0, 1, 2, … , 31

Decade: 80s, 90s, 00s, 10s

The Problems

1) What properties of games best predict global sales?

2) What properties of games best predict sales for each region?

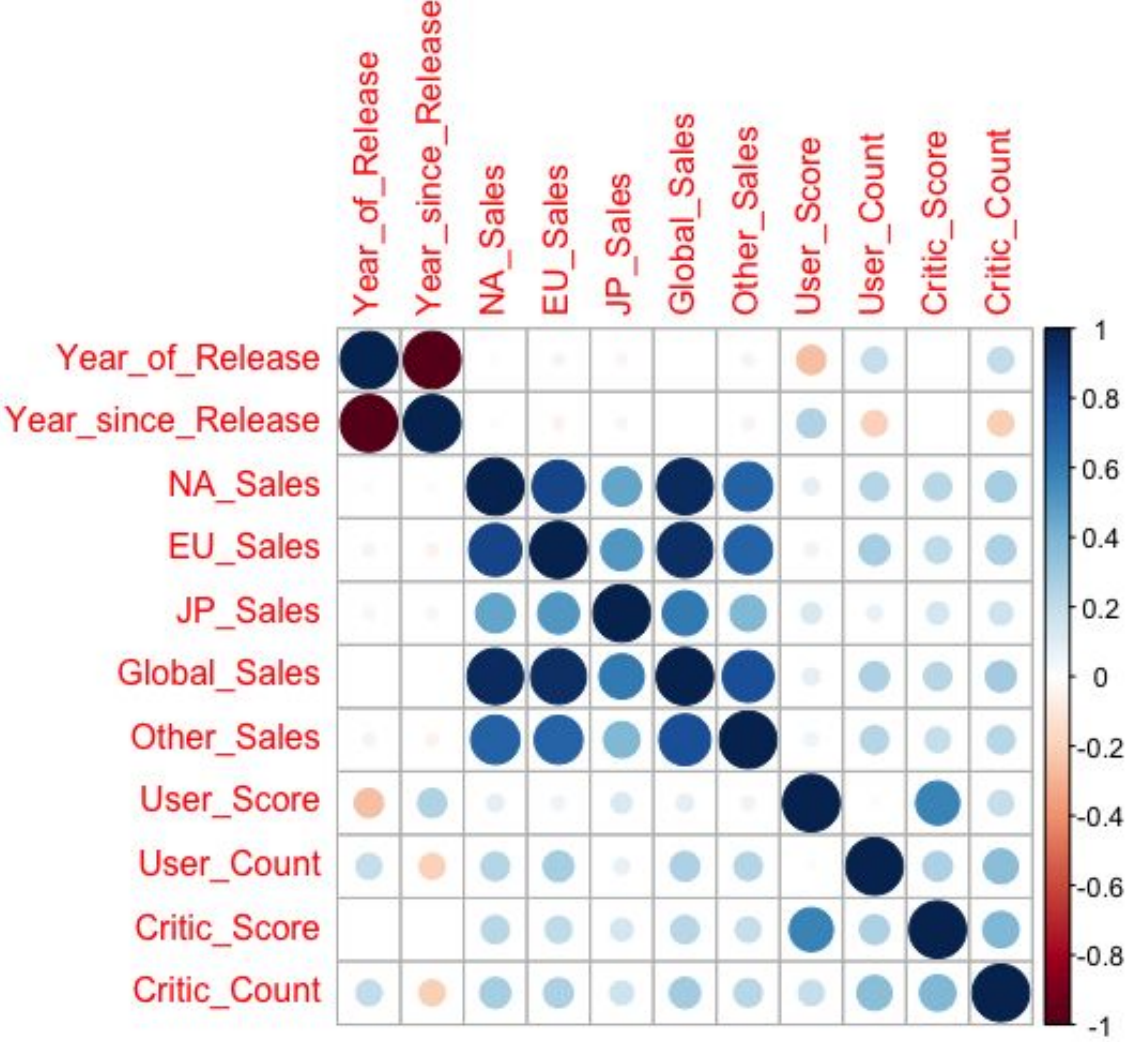
Some additional hypotheses regarding interactions

- **Genre**.
  - ex. Popularity of shooter games in North America or EU vs. in Japan
- **Developer** & **publisher company**.
  - ex. Games produced by SEGA vs. Microsoft in Japan vs. North America
- **Critic scores** & **user scores** on sales in Japan vs. in North America
- **Decade** & **genre** on global sales

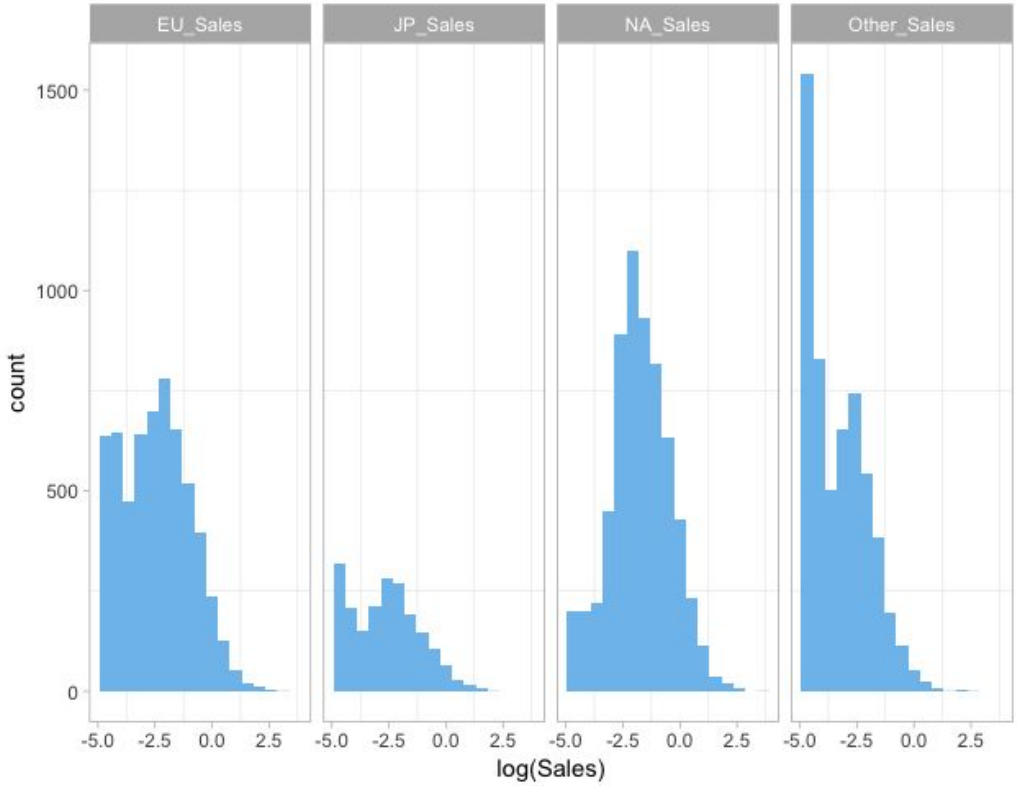
Tools & Models

- Multiple Linear Regression (MLR) to predict sales
- Take subsets of the data to train and test models
- Model comparison and selection - ex. AIC, BIC
- Comparison of groups - t-tests, ANOVA
- Post-hoc analyses - pairwise comparisons with Holm’s correction, Tukey’s HSD

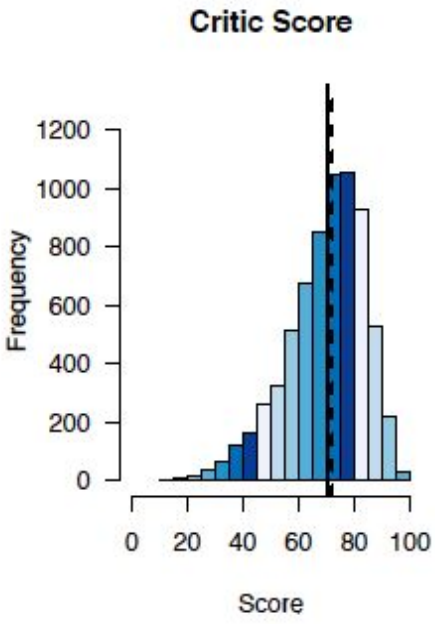
EDA:

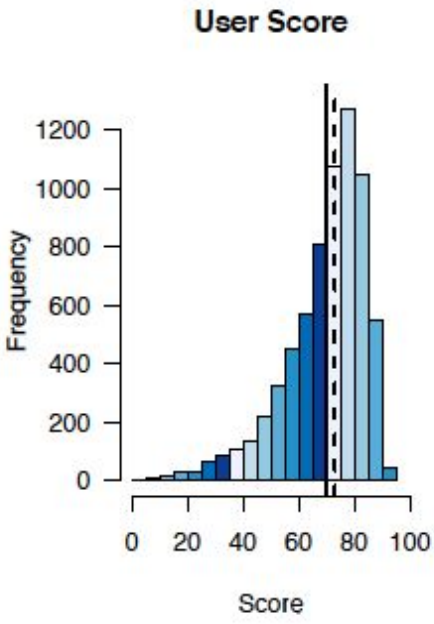

Year_of_Release
Year_since_Release
NA_Sales
EU_Sales
JP_Sales
Global_Sales
Other_Sales
User_Score
User_Count
Critic_Score
Critic_Count
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

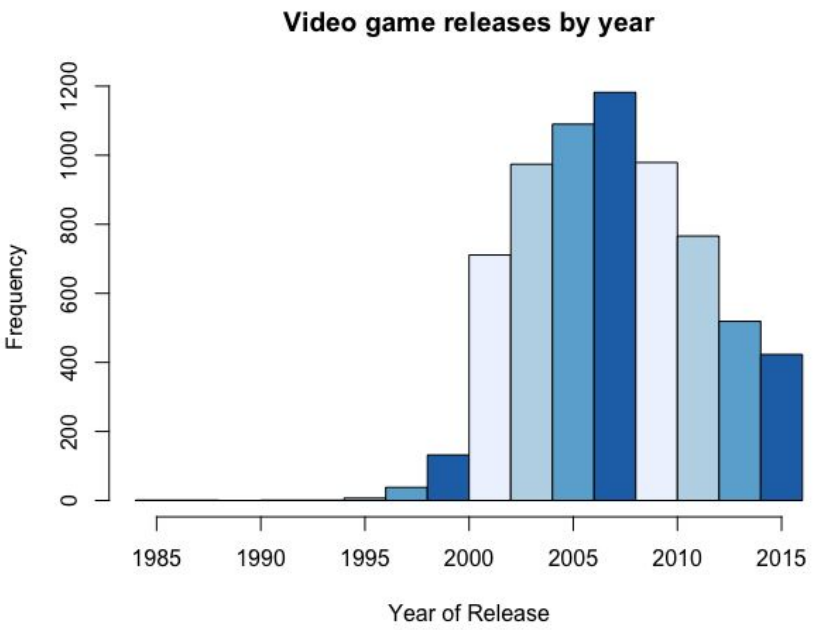
EDA: Sales by region



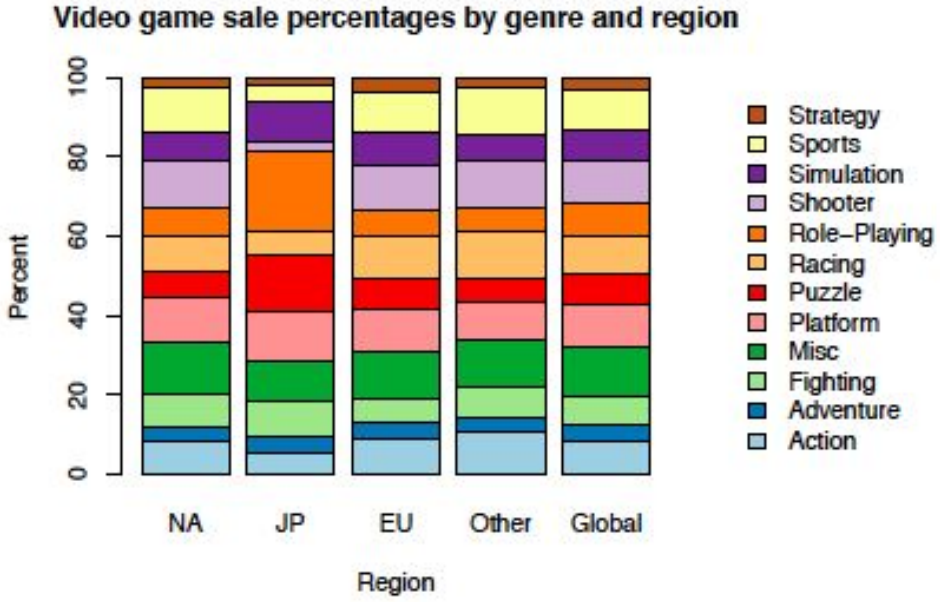
EDA:

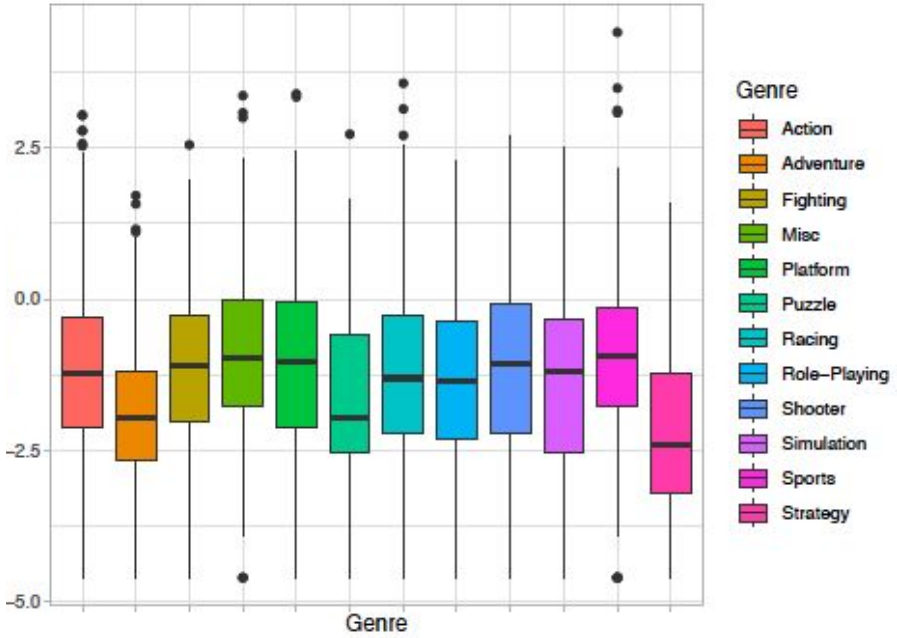

Critic Score
Frequency
0
200
400
600
800
1000
1200
0
20
40
60
80
100
Score


User Score
Frequency
0
200
400
600
800
1000
1200
0
20
40
60
80
100
Score


Video game releases by year
Frequency
0
200
400
600
800
1000
1200
1985
1990
1995
2000
2005
2010
2015
Year of Release

EDA: Genre


Video game sale percentages by genre and region
Percent
0
20
40
60
80
100
NA
JP
EU
Other
Global
Region
Strategy
Sports
Simulation
Shooter
Role–Playing
Racing
Puzzle
Platform
Misc
Fighting
Adventure
Action


2.5
0.0
−2.5
−5.0
Genre
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role–Playing
Shooter
Simulation
Sports
Strategy

Potential Challenges

- Violations of basic assumptions:
  - Non-normality of response variables
  - Heteroskedasticity
- Caused by:
  - Multicollinearity between some numeric variables
  - Dependence between categorical variables
  - Outliers and influential points
- Removal of NA’s
  - Non-generalizable model and inference
- Too many levels in categorical variables

References

https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/download

https://www.vgchartz.com/analysis/platform_totals/

https://en.wikipedia.org/wiki/Video_game_console