Steam User Behavior

SIENNA ZHU
SOPHIE ZHU
CHLOE CHEN
BRAYDEN MOORE

Agenda

PROJECT OVERVIEW

Industry Overview & Objectives

DATA COLLECTION & PROCESSING

Steam API Calls & Data Amalgamation

DATA VISUALIZATION

Game Specs & User Spending Habits

MODELING

Multiple Regression & Decision Tree Classifier

CONCLUSION & INSIGHTS

Business Insights & Further Improvements

Industry Overview

REVENUE

- $135 Bn in revenue globally in 2019
- $60.59Bn revenue in US

INDUSTRY

- Larger than Film and Music industries combined
- CAGR ~11.2%
- Tech Innovation & Subscription Plans

PLATFORMS

- Largest digital distribution platform for PC gaming
- 34,000+ games available
- over 8,000 releases in 2018
- 21M+ concurrent users

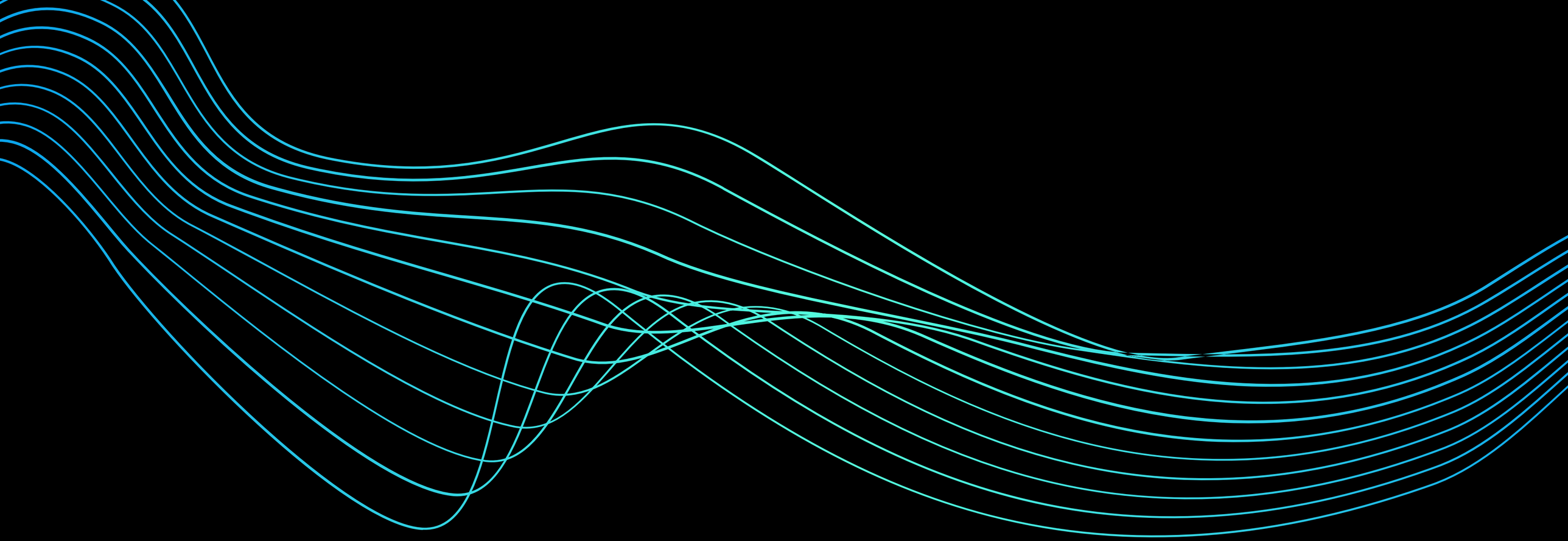
Our Goals

How can we help game developers capture a larger audience?

Understand Steam userbase & user behavior

Use this knowledge to build predictive models

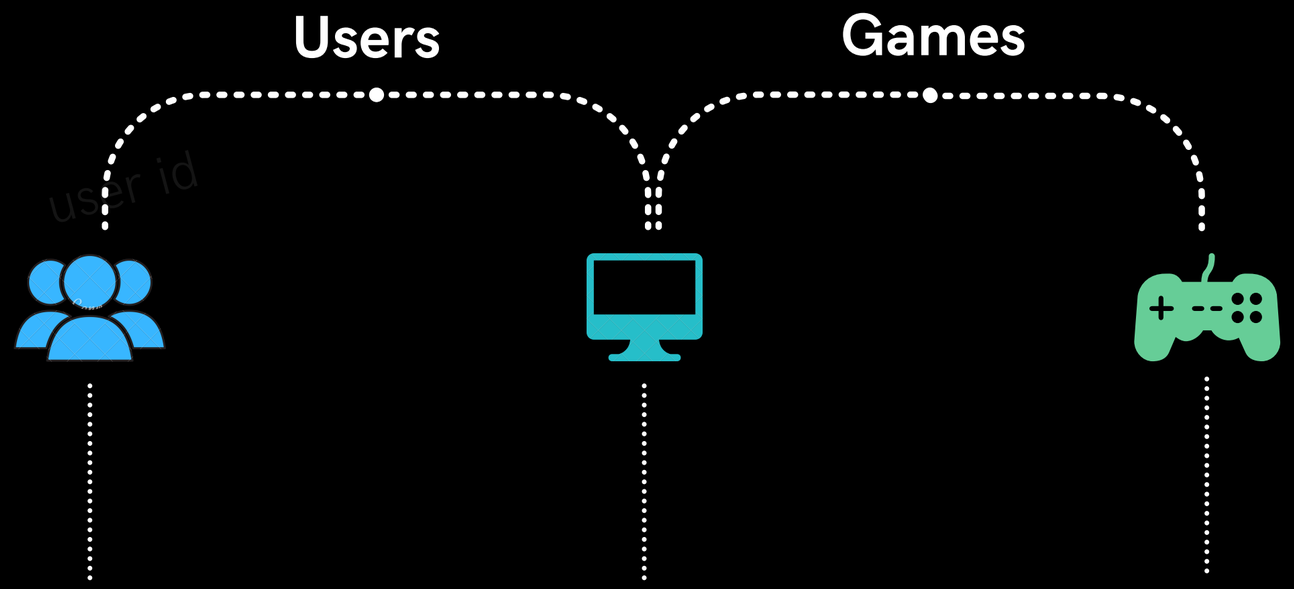
Infer the optimal price of a game for a user



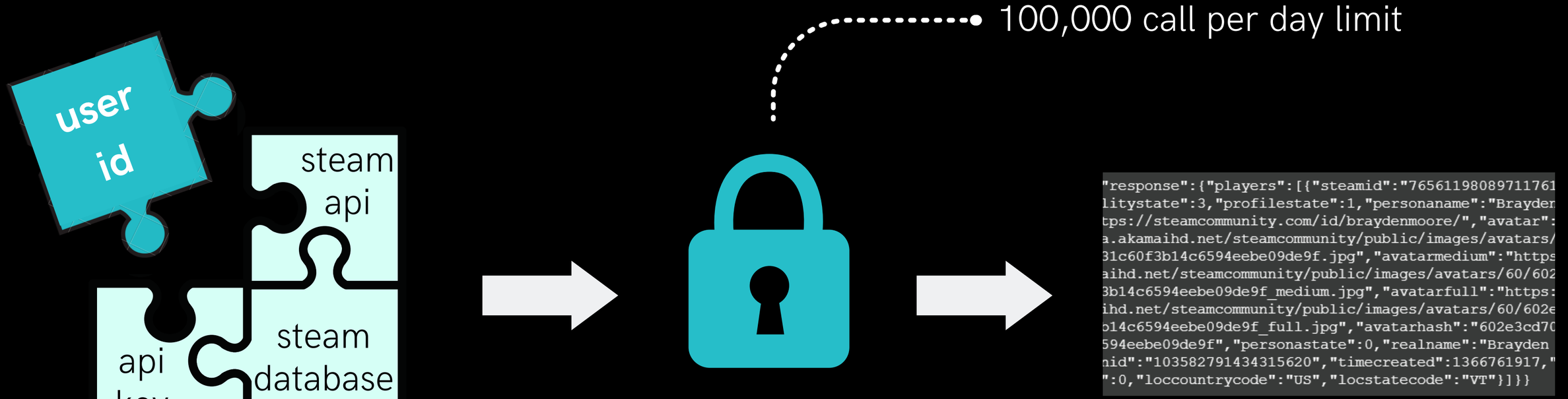
Part II

Data Collection & Processing

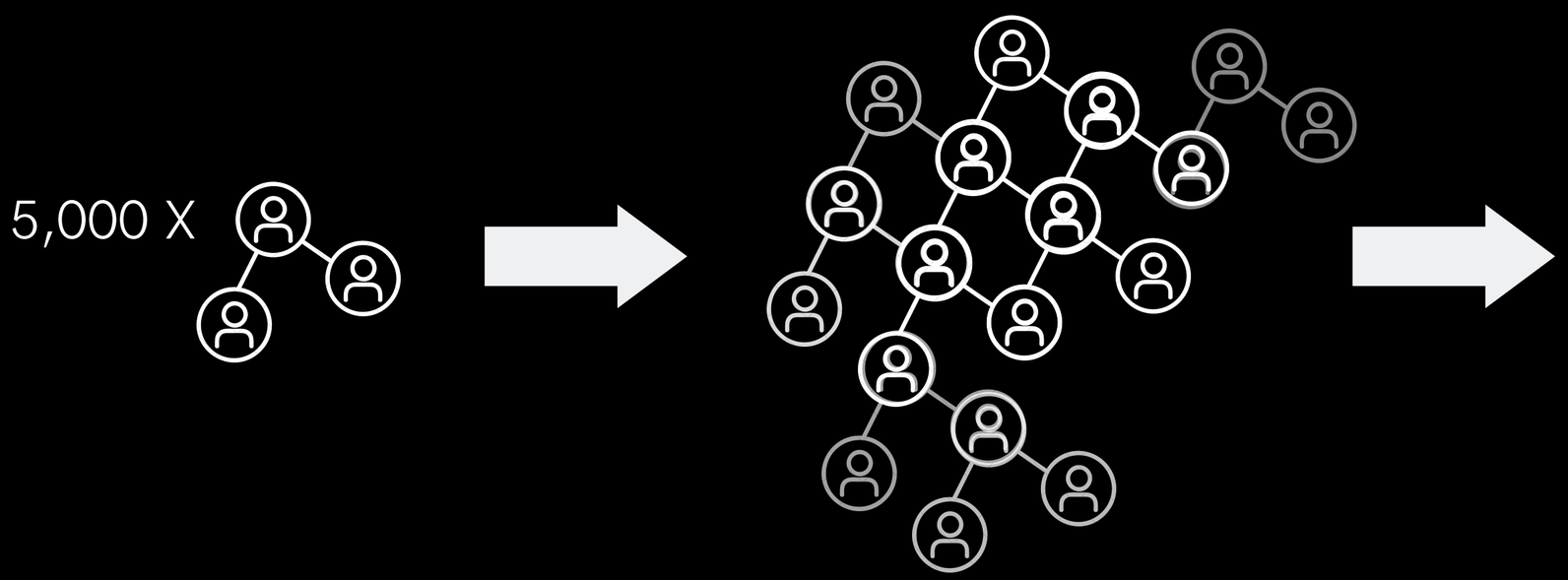
What data do we need?



User Data


user id
steam api
api key
steam database
100,000 call per day limit
"response":{"players":[{"steamid":"76561198089711761
litystate":3,"profilestate":1,"personaname":"Brayden
tps://steamcommunity.com/id/braydenmoore/","avatar":
a.akamaihd.net/steamcommunity/public/images/avatars/
31c60f3b14c6594eebe09de9f.jpg","avatarmedium":"https
aihd.net/steamcommunity/public/images/avatars/60/602
3b14c6594eebe09de9f_medium.jpg","avatarfull":"https:
ihd.net/steamcommunity/public/images/avatars/60/602e
b14c6594eebe09de9f_full.jpg","avatarhash":"602e3cd70
594eebe09de9f","personastate":0,"realname":"Brayden
nid":"103582791434315620","timecreated":1366761917,"
":0,"loccountrycode":"US","locstatecode":"VT"}]}}

Generating User IDs



	steamid	relationship	friend_since
0	76561197967007112	friend	1250885733
1	76561197972326806	friend	1306781161
2	76561197982550012	friend	1258884042
3	76561197986688687	friend	1258566191
4	76561197989642450	friend	1250885738
...	...	...	...
1	76561198353450678	friend	1584570841
2	76561199035760827	friend	1584611130
3	76561199035771249	friend	1584570731
0	76561198023547951	friend	1598371922
1	76561198026722662	friend	1597347255

2099751 rows × 3 columns

Users' Personal Data

	steamid	relationship	friend_since
0	76561197967007112	friend	1250885733
1	76561197972326806	friend	1306781161
2	76561197982550012	friend	1258884042
3	76561197986688687	friend	1258566191
4	76561197989642450	friend	1250885738
...	...	...	...
1	76561198353450678	friend	1584570841
2	76561199035760827	friend	1584611130
3	76561199035771249	friend	1584570731
0	76561198023547951	friend	1598371922
1	76561198026722662	friend	1597347255

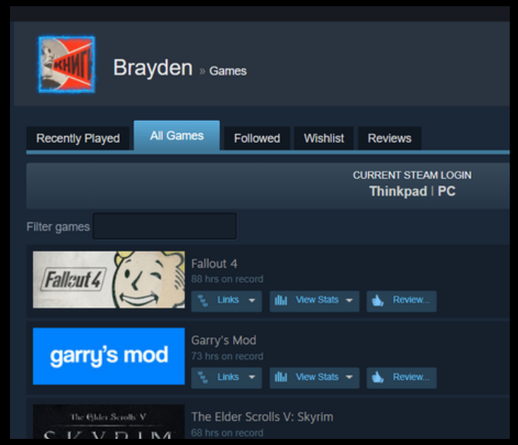
2099751 rows × 3 columns

```python
# function that takes a list of IDs and our API key, runs the query for
def getPlayerSummaries(idList, apiKey):
    df = pd.DataFrame()
    for i in idList:
        url = f'http://api.steampowered.com/ISteamUser/GetPlayerSumm
        try:
            r = requests.get(url)
            result = json.loads(r.text)
            output = pd.DataFrame.from_dict(result['response']['players'])
            df = df.append(output)
        except:
            pass
    return df
```

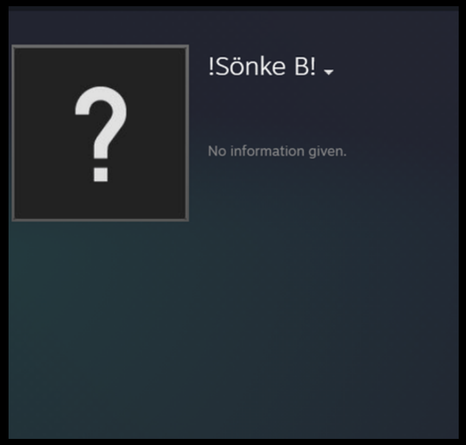
Users' Game Data

Title App ID

Forever 2 Weeks



Game list public



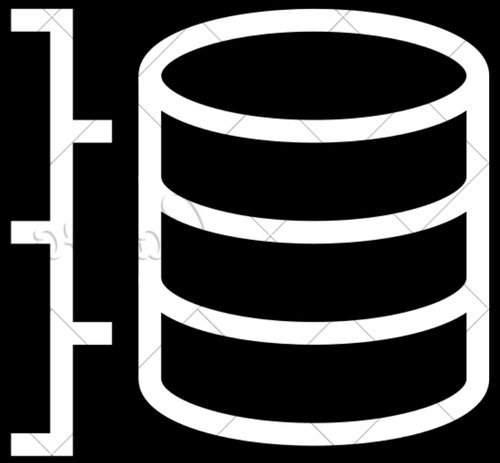
Game list private

Aggregation

Steam Game Data

Users' Game Data

Users' Personal Data



Steam ID

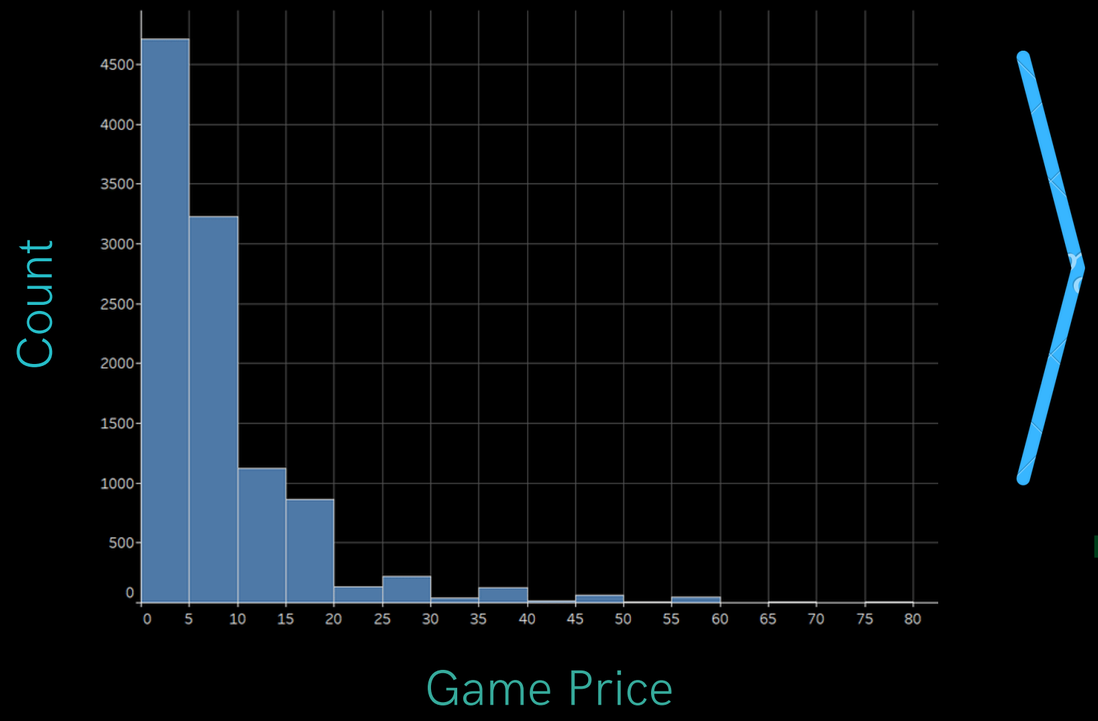
- Country
- Creation Date
- Games Owned
- Playtime Forever
- Playtime 2 Weeks
- % of Library Composed of Each Genre
- **Average Spent per Game**

Part III

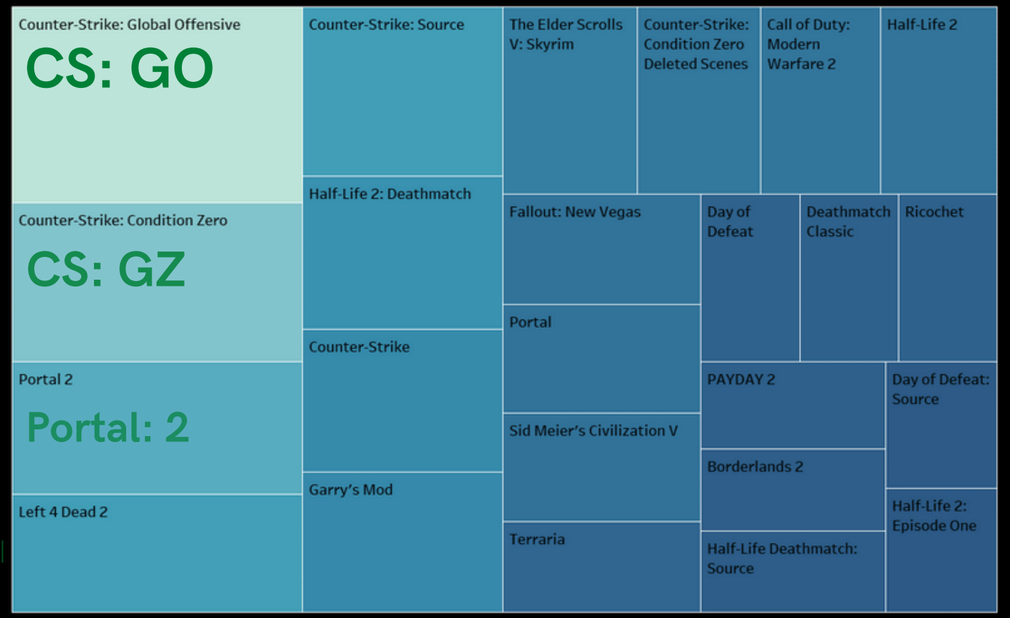
Visualizing Our Dataset

I. Game Data

Game Prices



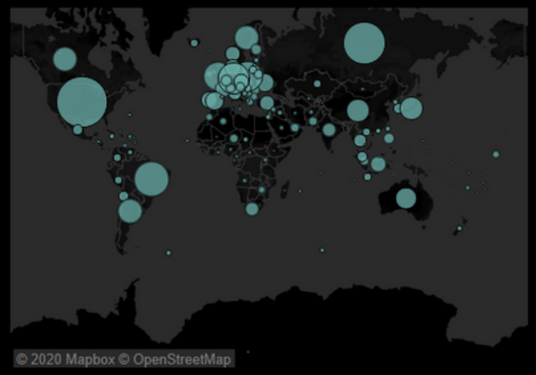
TOP 25 Paid Games



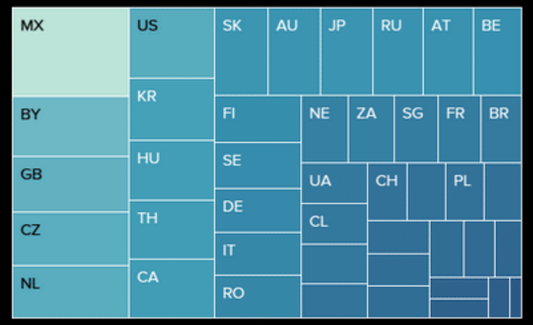
II. User Data

USA has by far the most users

Number of Users



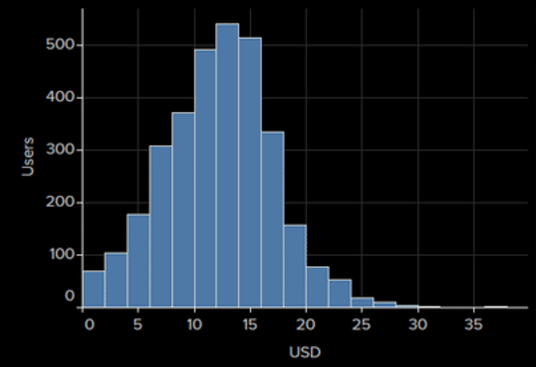
Games Owned Per Player



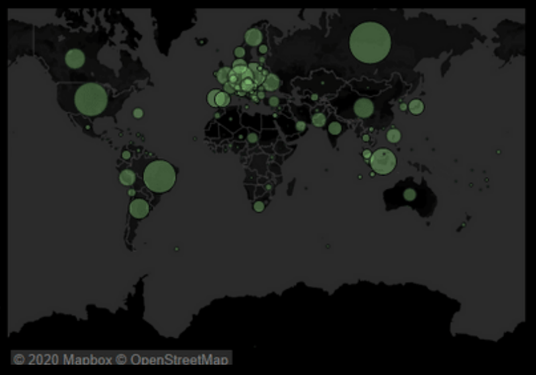
Players in Mexico tend to own the most games

Spending per game is nearly normal, mean of $12.22

Average Spent Per Game

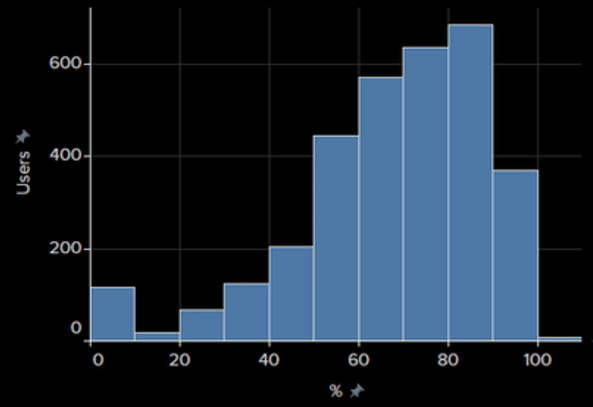


Total Playtime

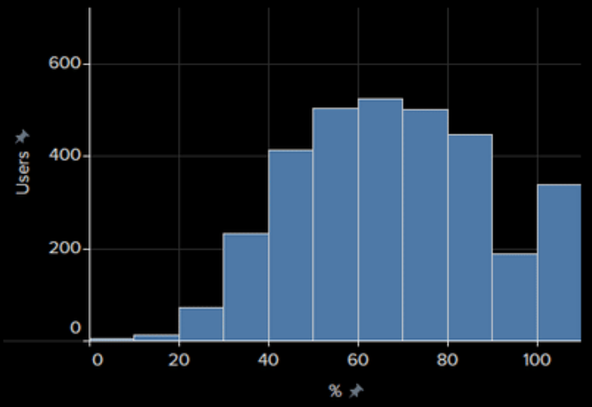


Russia, Brazil, USA have most total playtime

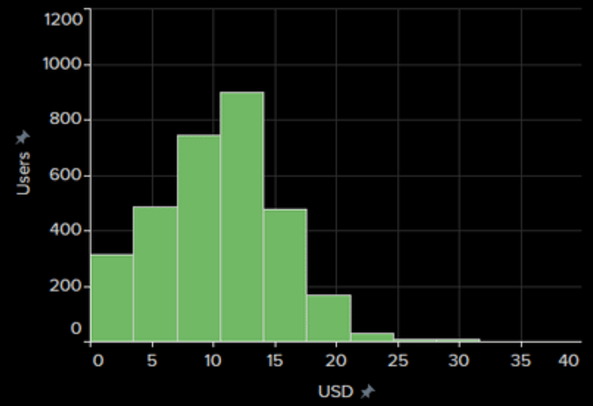
% of Library that is **Single Player**



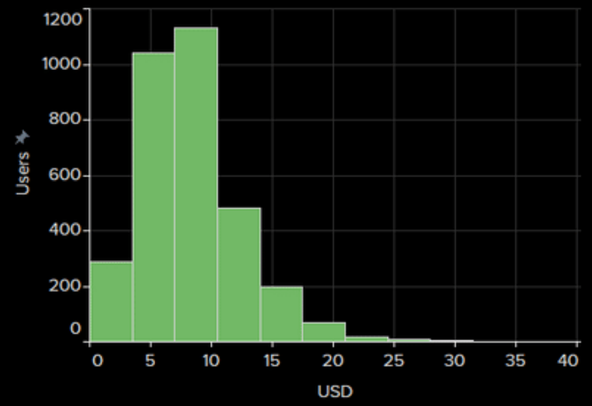
% of Library that is **Multiplayer**



Average Spent on **Single Player** Games



Average Spent on **Multiplayer** Games



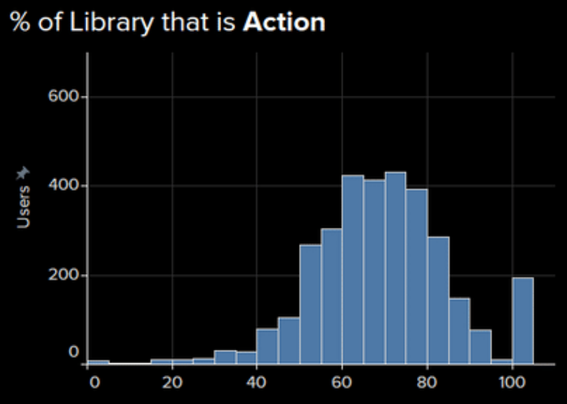
$$\frac{\text{Single Player Games Owned}}{\text{All Games Owned}} = \%\ \text{Single Player}$$

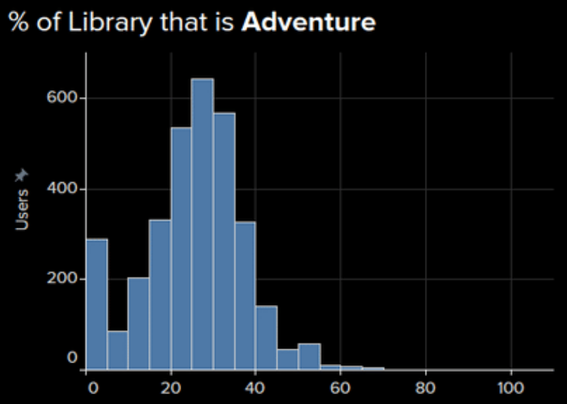
$$\frac{\text{Total Spent on MP Games}}{\text{Multiplayer Games Owned}} = \text{Avg Spent on Multiplayer Games}$$

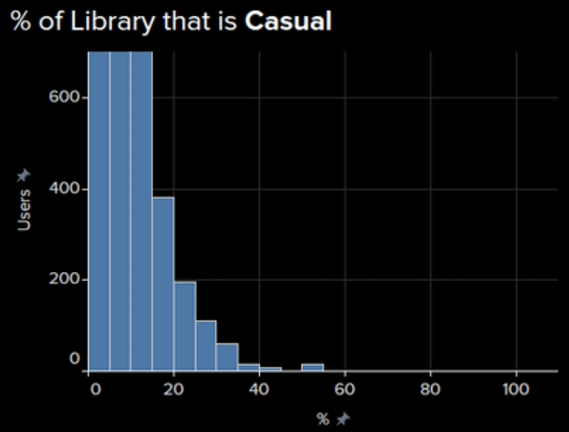
PART III: VISUALIZING OUR DATASET

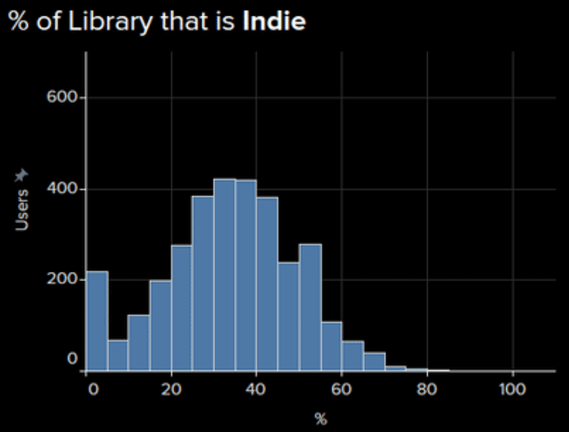
ACTION ADVENTURE CASUAL INDIE

% of Library Composition

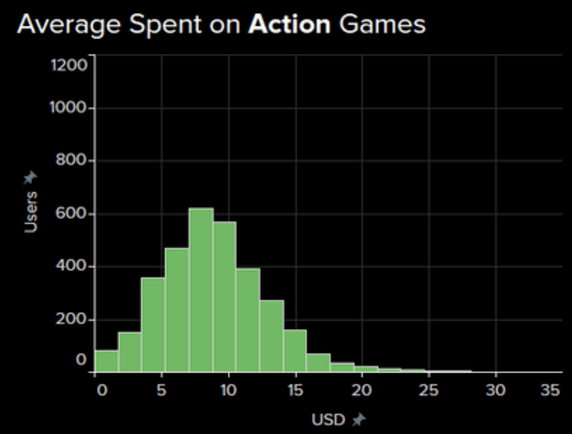


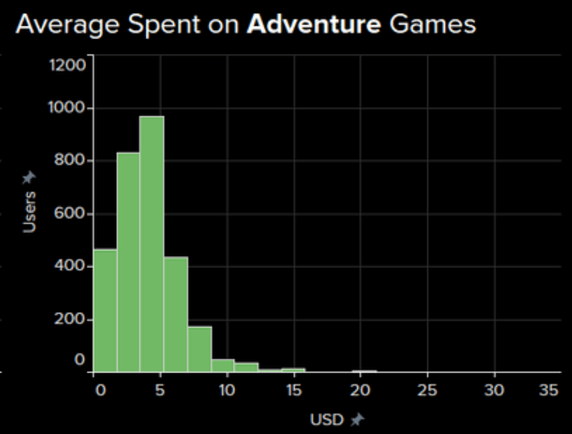


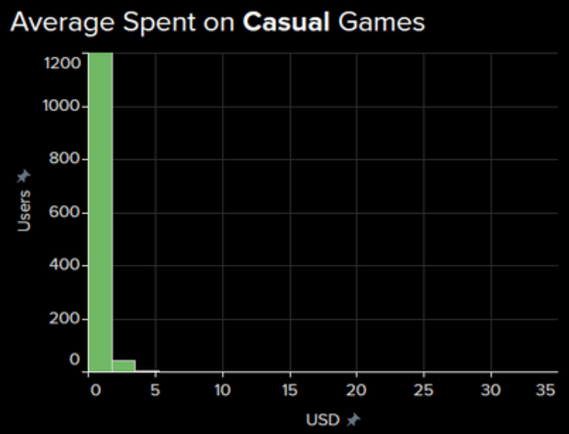


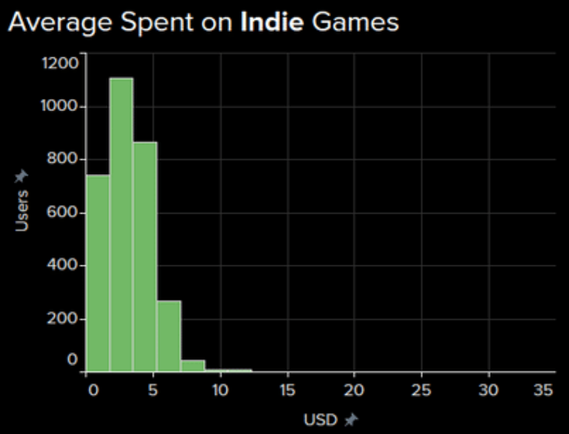


$ Average Spent









Part IV

Modeling our data

> Based on a player's characteristics and game preferences, can we predict how much they are willing to spend on a game?

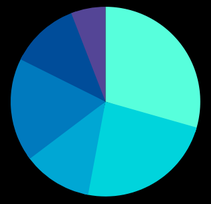
I. Multiple Regression

Best Subset: 15 Predictors

Country

Creation Date

% 11 Genres
% Single
% Multi

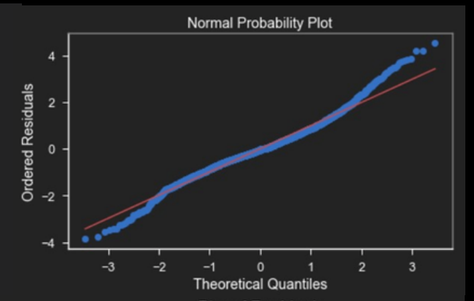


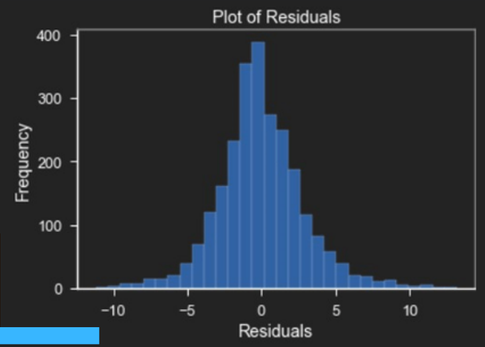
Avg Spent per Game

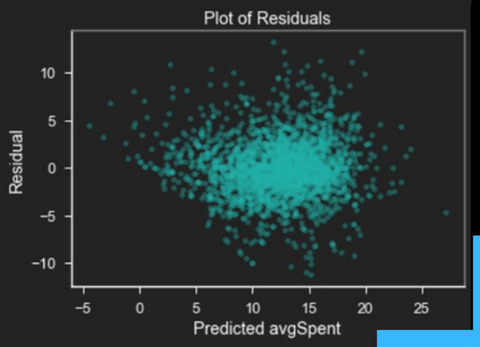
R-squared 62.9%

MSE $3.02

Residual Plots and Q-Q Plot







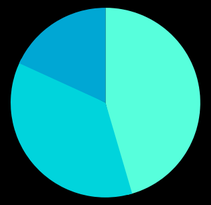
II. Multiple Regression: Simplest Model

Best Subset: 5 Predictors

Country

Creation Date

% Casual
% Single
% Multi

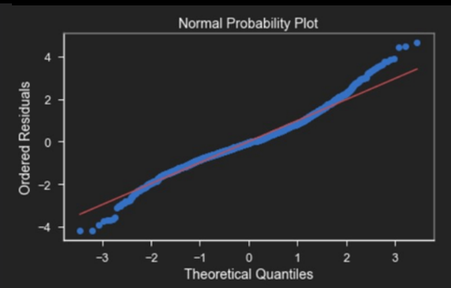


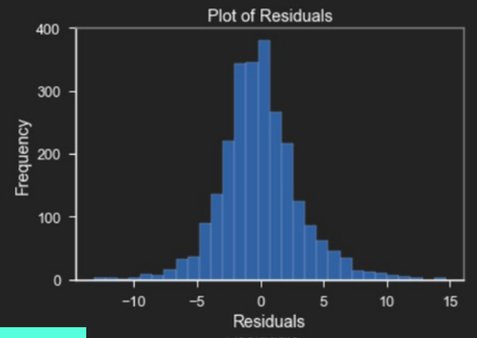
Avg Spent per Game

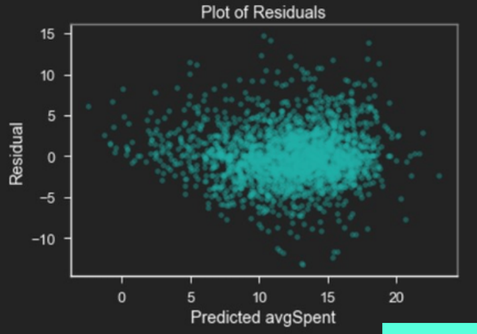
R-squared 56.7%

MSE $3.26

Residual Plots and Q-Q Plot

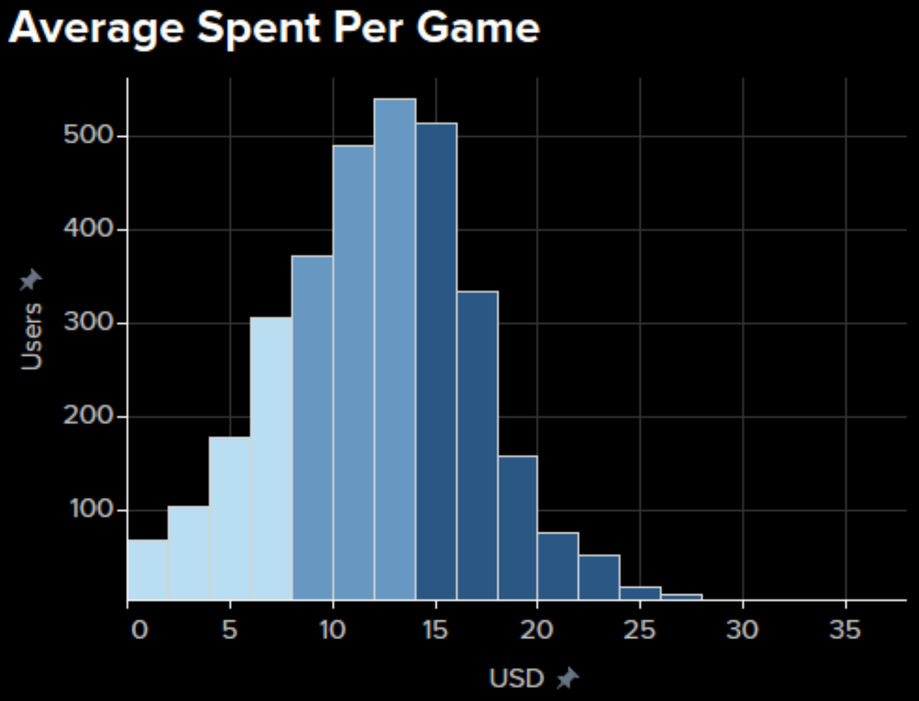






III. Decision Tree Classification

Predicting whether a user has a low, medium, or high optimal game price.



Low: $0 - 8
Medium: $8 - 14
High: $14 - 50

Control: SteamID

Predicted

Actual	Low	Med	High
Low	0%	100%	0%
Med	0%	100%	0%
High	0%	100%	0%

Accuracy Score
42.8%

Creation Date

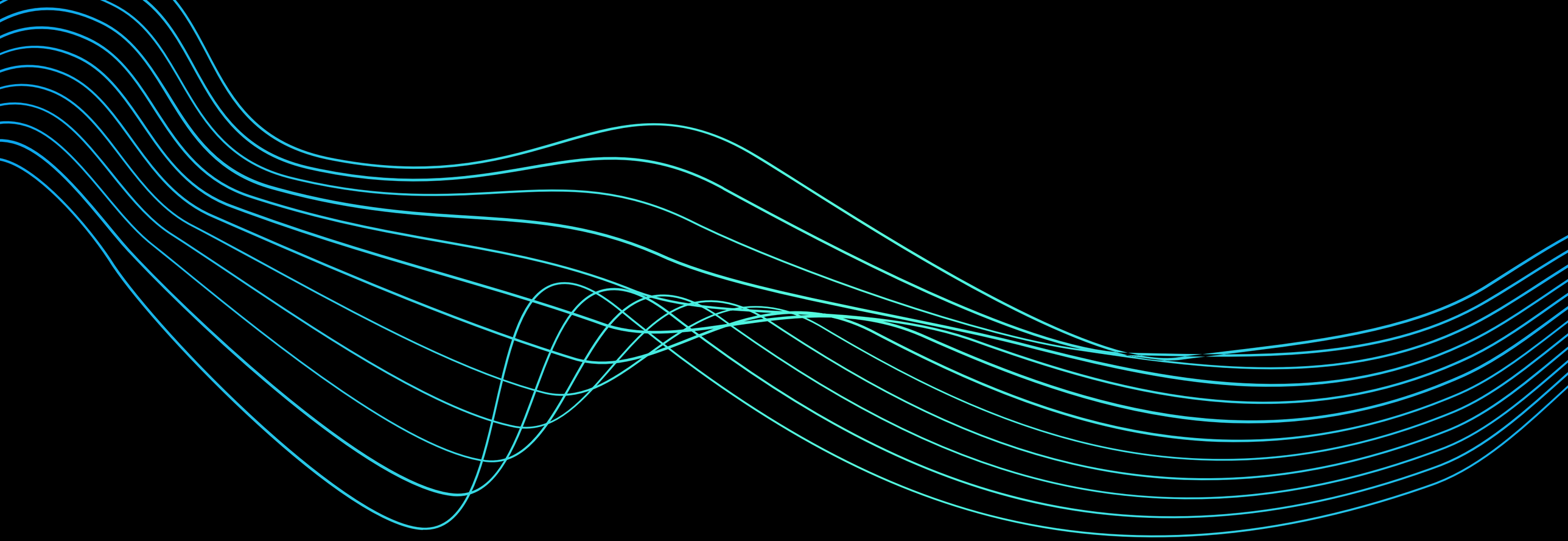
% 11 Genres
% Single/Multi

Avg Spent per Game

Predicted

Actual	Low	Med	High
Low	59%	39%	2%
Med	14%	71%	14%
High	5%	31%	64%

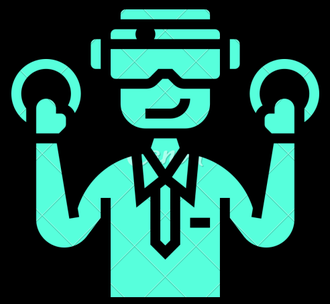
Accuracy Score
66.2%



Part V

Conclusions & Insights

Key Takeaways



Industry Wise

People are willing to spend more on Single Player games and Multi Player games (highest T-value) compared to Indie and Casual games (lowest T-value)

User Wise

People who has been a long time Steam user are generally willing to spend more than new users

Model Recommendation

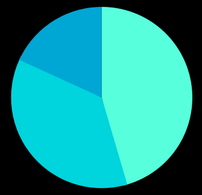
Multiple Regression: Simplest Model

Best Subset: 5 Predictors

Country

Creation Date

% Casual
% Single
% Multi



Avg Spent per Game

R-squared 56.7%

MSE $3.26

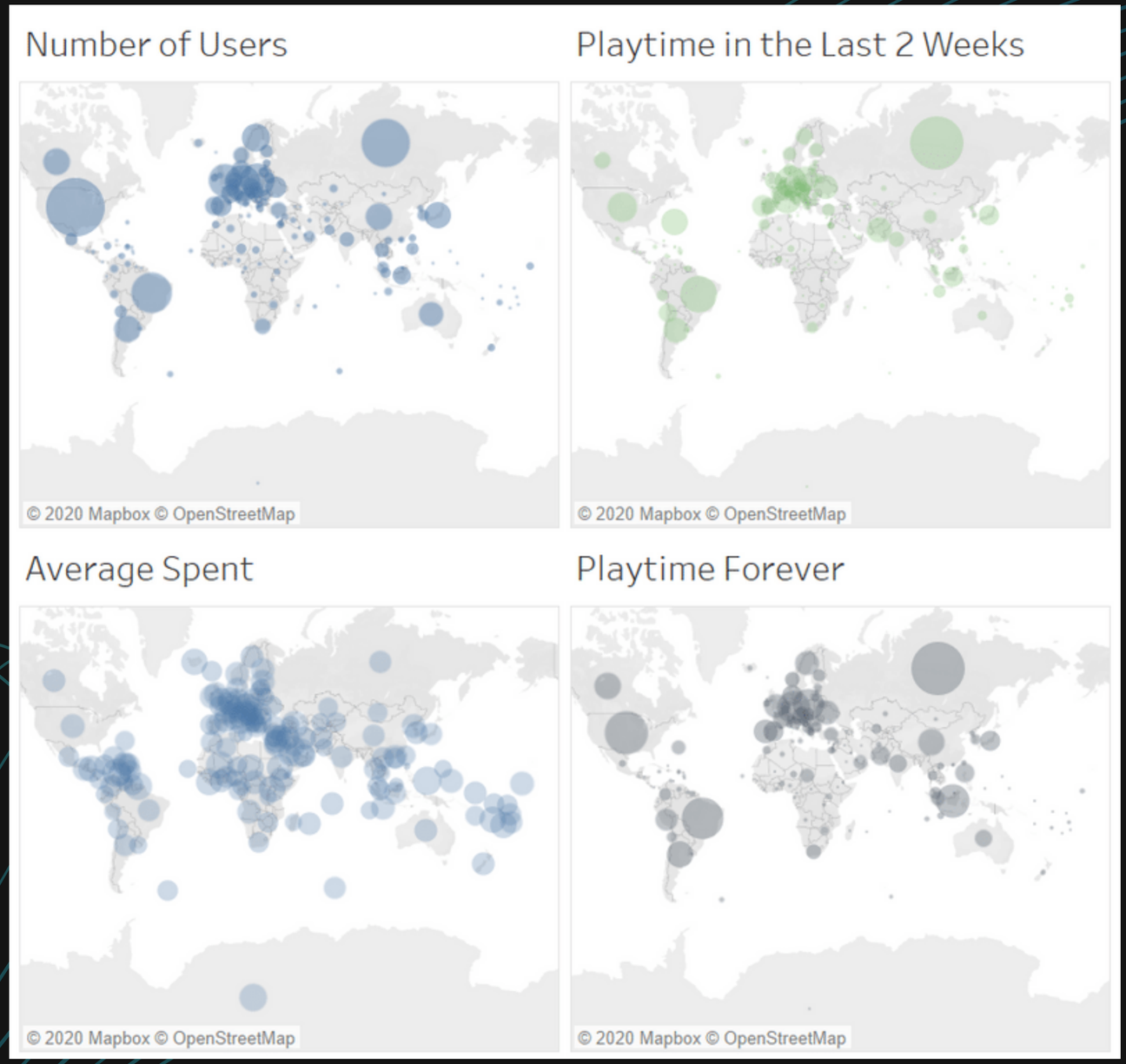
- Higher Precision
- Ease in Interpretation
- Data Accessibility

PART V: CONCLUSIONS AND NEXT STEPS


Q&A

APPENDIX

#1: Visualization by Countries



APPENDIX

#2: 15 Predictor Multiple Regression Coefficients

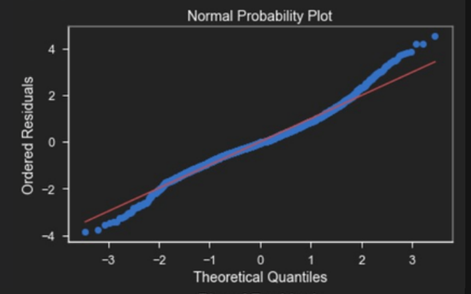
Coefficients

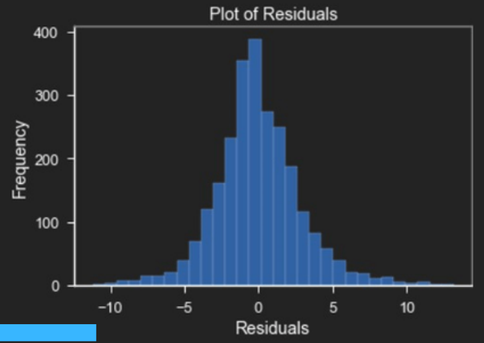
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.873	0.615	-1.42	0.156	
playtimeForever	0.000000	0.000000	0.90	0.366	1.77
playtime2weeks	0.000003	0.000004	0.84	0.399	1.69
indieProp	-0.08003	0.00650	-12.32	0.000	2.63
actionProp	-0.01776	0.00735	-2.42	0.016	3.16
advenProp	0.09557	0.00800	11.94	0.000	2.25
casualProp	-0.1615	0.0107	-15.15	0.000	2.14
stratProp	-0.04174	0.00714	-5.85	0.000	1.78
rpgProp	0.01252	0.00903	1.39	0.166	2.05
simProp	0.06962	0.00903	7.71	0.000	2.00
racingProp	0.0895	0.0188	4.76	0.000	2.25
sportsProp	-0.0918	0.0174	-5.27	0.000	2.50
mmoProp	-0.0623	0.0102	-6.12	0.000	3.07
singProp	0.15802	0.00698	22.63	0.000	5.26
multProp	0.09339	0.00674	13.86	0.000	4.75
coopProp	0.00093	0.00688	0.13	0.893	2.44
inappProp	-0.05474	0.00862	-6.35	0.000	6.02
vrProp	-0.0450	0.0197	-2.29	0.022	1.41
timeCreated	0.000000	0.000000	11.87	0.000	4.46
country					

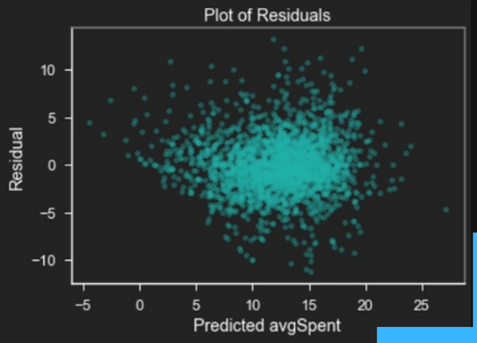
R-squared 62.9%

MSE $3.02

Residual Plots and Q-Q Plot







APPENDIX

#3: 5 Predictor Simplest Multiple Regression

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2.238	0.570	-3.93	0.000	
casualProp	-0.25679	0.00816	-31.48	0.000	1.15
singProp	0.19170	0.00477	40.17	0.000	2.26
multProp	0.05268	0.00433	12.15	0.000	1.80
timeCreated	0.000000	0.000000	14.06	0.000	1.64
country					

R-squared 56.7%

MSE $3.26

Residual Plots and Q-Q Plot