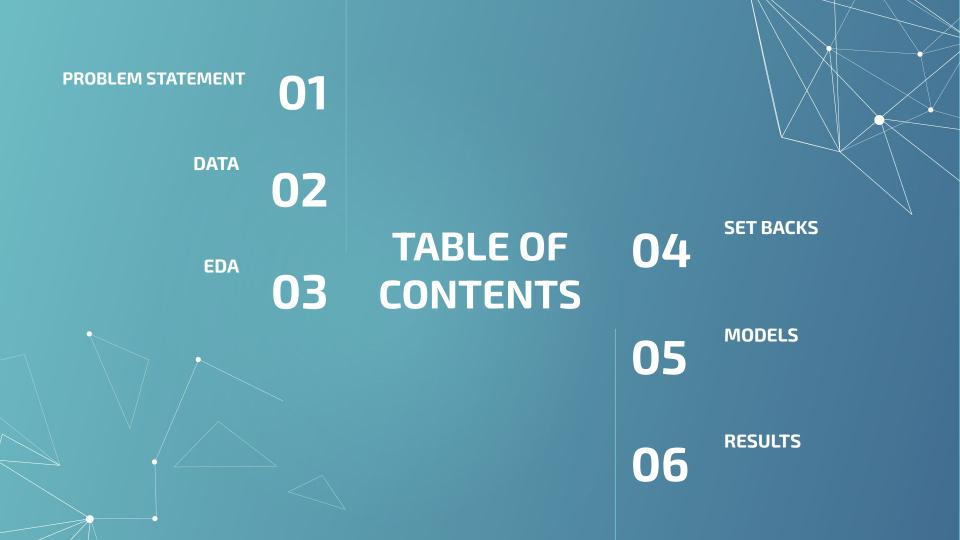
# PREDICTING STEAM GAME PLAY TIME

Dominika Jones, General Assembly





#### Introduction

Steam platform is currently the largest PC game online distributor in the world and has accumulated a vast amount of user and game data.

One metric is particularly important to the developers who plan the game's support cycle - expected play time.



#### **Problem Statement**

Customer engagement is a measurement of a user's response to the product. Average playtime metric is a major factor influencing the player's buying decision. By predicting the average total playtime through various regression models, I hope to improve user experience and customer engagement.

# O2 DATA



### **DATA ACQUISITION**

#### **Steam Web API**

- Public access to player and game data.

#### **Steam Spy API**

- 3rd party site that aggregates statistics on Steam games.-

#### **Data Issues Along the Way**

- Data gaps due to users setting accounts to private.
- Distribution of playtimes is very skewed.
- Difficult to choose a good cost function.
- Daily issues with steamspy API

#### Hi Sergey

My significant other (CCed) is attempting to use your steamspy API to get some initial steam data for a data science class. She was trying to reproduce the code written here but is having issues. She is trying to use the request all type to retrieve a bunch of data but now is getting empty jsons. It was not defined in the API document but does an empty json entry mean that we attempted to poll the site too quickly. It's extra strange because appdetail requests always return data so the behavior is not consistent. If we do start to get empty APIs how long do we need to abstain from pulling data from the site before we can retrieve data gain?

#### **STEAMTIME.INFO**



# 840 hrs ->





# **Feature Processing**

#### **One-hot encoded**

CATEGORY GENRE

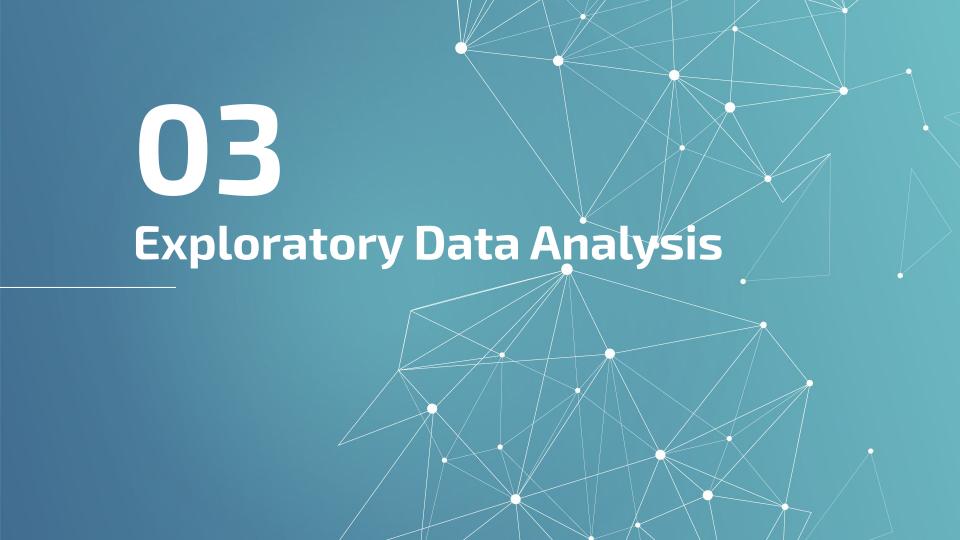
**DEVELOPERS** 

# Raw numerical input

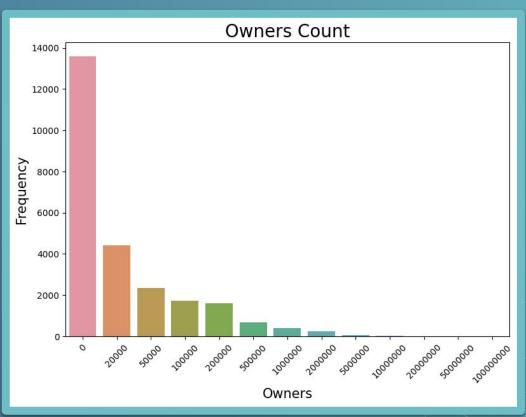
PRICE

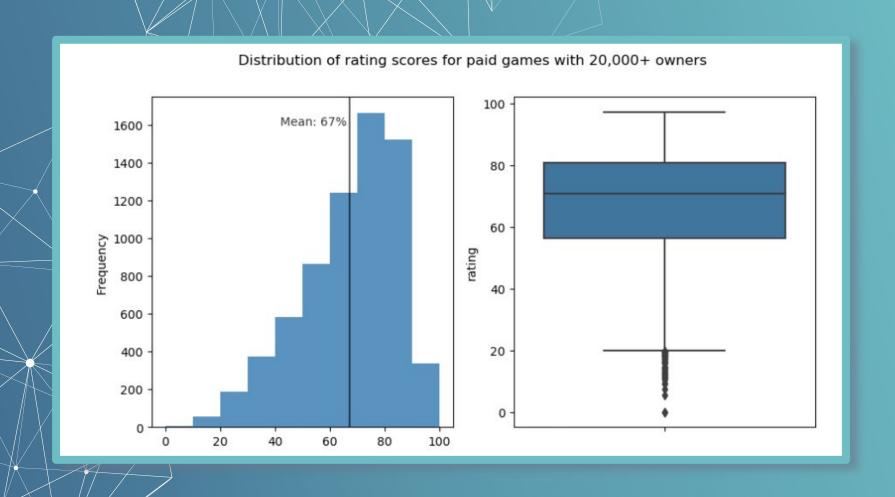
**RATINGS** 

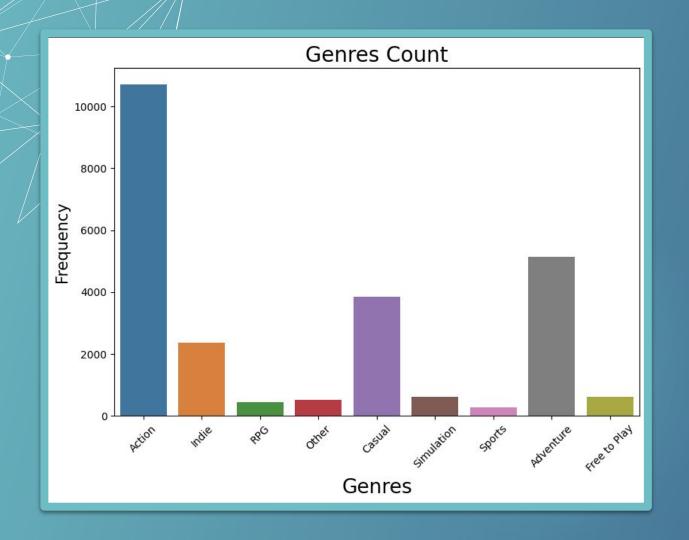
**OWNERS** 



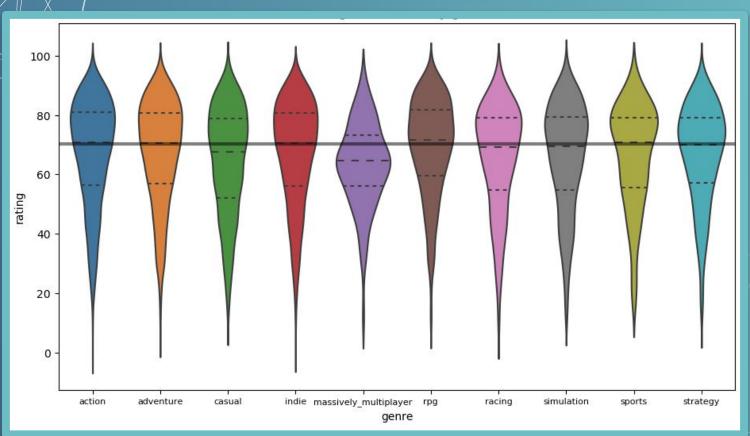
# **OWNERS AND RATINGS**







# Rating Distribution by Genre



average_forever			
average_neeks	average_forever -		
total_ratings -	owners -	0.2	
Description	average_2weeks -	0.19	
negative	total_ratings -	0.18	
Dositive	ccu -	0.18	
price - 0.11 tagories_fixed_Single_Multi - 0.849 required_age - 0.841 catagories_fixed_Other - 0.04 genres_Other - 0.035 rating_ratio - 0.039 gerres_Simulation - 0.029 gerres_Simulation - 0.024 catagories_fixed_Multi - 0.024 genres_Fixed_Multi - 0.023 genres_Fixed_Multi - 0.023 genres_Fixed_Multi - 0.015 tforms_windows_mac_inux - 0.015 tforms_windows_mac_inux - 0.0066 platforms_minder_inux - 0.0067 genres_Indie - 0.00077 english - 0.00078 genres_Action - 0.001 platforms_mac - 0.0012 discount - 0.0012 discount - 0.0024 genres_Free to Play - 0.0038 platforms_windows_linux - 0.0054 free - 0.013 platforms_windows_linux - 0.0054 free - 0.013 platforms_windows_linux - 0.0054 release_year - 0.036	negative -		
agories_fixed_Single/Multi	positive -	0.18	
required_age - 0.041  catagories_fixed_Other - 0.035  rating_ratio - 0.029  genres_Simulation - 0.024  catagories_fixed_Multi - 0.023  genres_RPG - 0.021  genres_RPG - 0.015  ttorms_windows.mac_liniux - 0.0066  platforms_mac_liniux - 0.0067  genres_finitive - 0.00077  genres_finitive - 0.00078  genres_fiction - 0.001  platforms_mac - 0.001  genres_fiction - 0.001  ge	price -		
catagories fixed Other - 0.04	gories_fixed_Single/Multi -	0.049	
genres_Other	required_age -	0.041	
rating ratio genres_Simulation	catagories_fixed_Other -		
Genres_Simulation	genres_Other -		
Catagones_Rived_Multi-   0.073	rating_ratio -		
genres_Sports	genres_Simulation -		
genres_Sports - 0.015  tforms_windows;mac-linux - 0.0066  platforms_mac,linux - 0.0067  genres_Indie - 0.00077  english - 0.00078  genres_Action - 0.0012  platforms_mac - 0.0012  discount - 0.0012  discount - 0.0034  genres_Free to Play - 0.0038  platforms_windows;linux - 0.0044  free - 0.013  platforms_windows - 0.014  genres_Adventure - 0.018  genres_Casual - 0.021  release_year - 0.036	catagories_fixed_Multi -		
tforms_windows_mac_linux - 0.0066  platforms_windows_mac_linux - 0.00067  genres_Indie - 0.00077  english - 0.00078  genres_Action - 0.001  platforms_mac - 0.0012  discount - 0.0012  genres_Free to Play - 0.0038  platforms_windows_linux - 0.0054  free - 0.013  platforms_windows - 0.014  genres_Casual - 0.021  release_year - 0.030	genres_RPG -		
platforms_windows.mac - platforms_mac_linux -	genres_Sports -		
platforms_mac,linux	orms_windows;mac;linux -		
Qenres_Indie	platforms_windows;mac -	0.0066	
genres_Indie     -0.00077       english -     -0.001       genres_Action -     -0.001       platforms_mac -     -0.0012       discount -     -0.0024       genres_Free to Play -     -0.0038       platforms_windows,linux -     -0.0054       free -     -0.013       platforms_windows -     -0.014       genres_Adventure -     -0.018       genres_Casual -     -0.021       release_year -     -0.036	platforms_mac;linux -	-0.00067	
english - 0.00078 genres_Action		-0.00077	
genres_Action	11.410.50.344.000	-0.00078	
platforms_mac -		-0.001	
discount -	A A A A A A A A A A A A A A A A A A A	0.0012	
genres_Free to Play	100000000000000000000000000000000000000	0.0024	
platforms_windows;linux -		-0.0038	
free			
platforms_windows -			
genres_Adventure -         -0.018           genres_Casual -         -0.021           release_year -         -0.036	1.000		
genres_Casual - 40,021 release_year - 40,036	(1,0,1,0) (1,0,0) (1,0,0)		
release_year			
	/		
abelia.			
catagories fixed_Single			

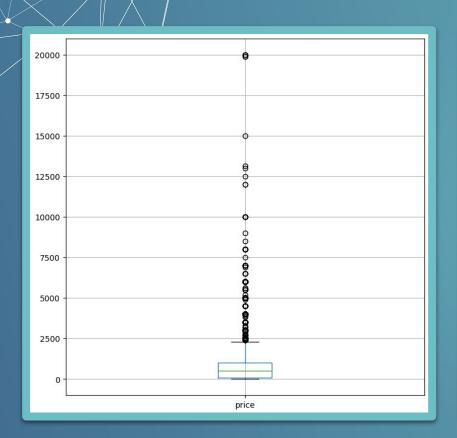
#### **Correlations**

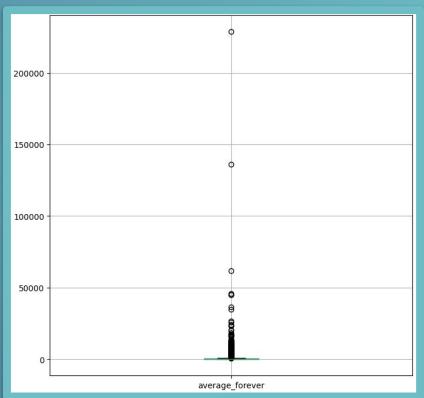
Price, Owners and Ratings have a positive correlation on average playtime (average\_forever)

Release year surprisingly has a negative effect or little correlation on the average playtime.

# Price Distribution

# **Playtime Distribution**







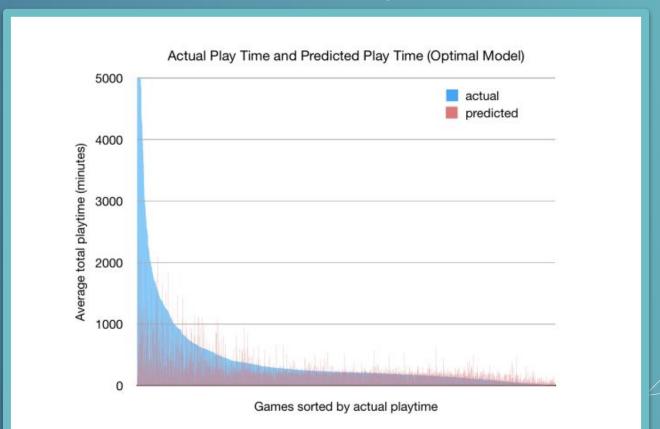
#### THE MODELS

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Tree
- GridSearchCV

### **MODEL RESULTS**

	MSE Train/Test	RMSE Train/Test	R2	ACC. Score
Linear Regression	2.5/3.3	1.5/1.8	0.1	7%
Random Forest Regression	0.26/2.05	0.5/1.4	0.9	33%
Gradient Boosting	2.5/2.6	1.6/1.7	0.14	14%

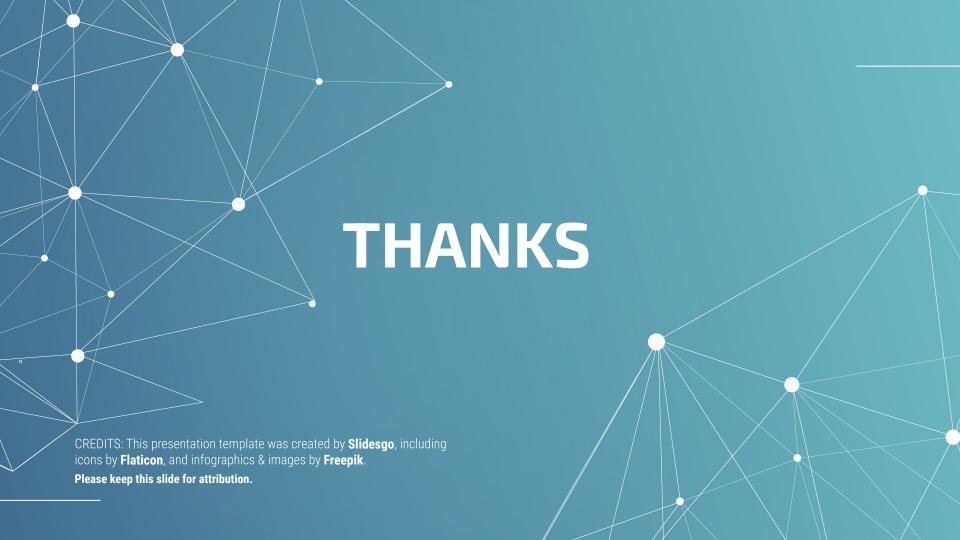
#### **Predicted Play Time**





#### **CONCLUSION AND FUTURE WORK**

- Positive influence features:
  - Positive Ratings, Number of Owners, Price
- Negative influence features:
  - Total Ratings, Category, Genre
- > The model does not capture the skewed distribution of playtime.
  - Underestimates the higher end
  - Overestimates the lower end
- More informative features
- More flexible model (tuning gradient boost tree)



#### **RESOURCES**

- http://store.steampowered.com/
- https://developer.valvesoftware.com/wiki/Steam\_Web\_API
- https://steamspy.com/about
- https://steamtime.info/

