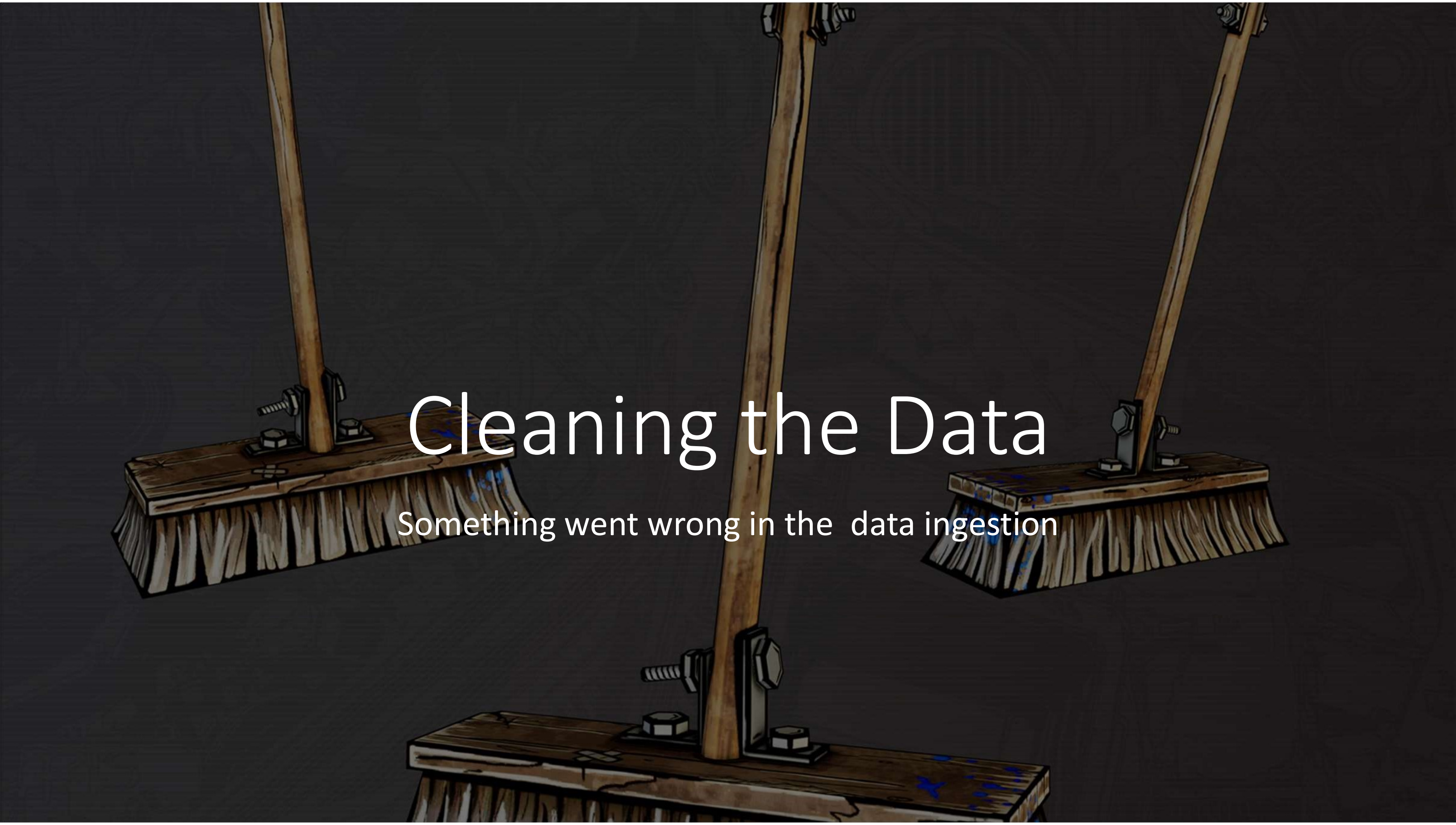
Cleaning the Data and Analyzing It

Alvaro Hu

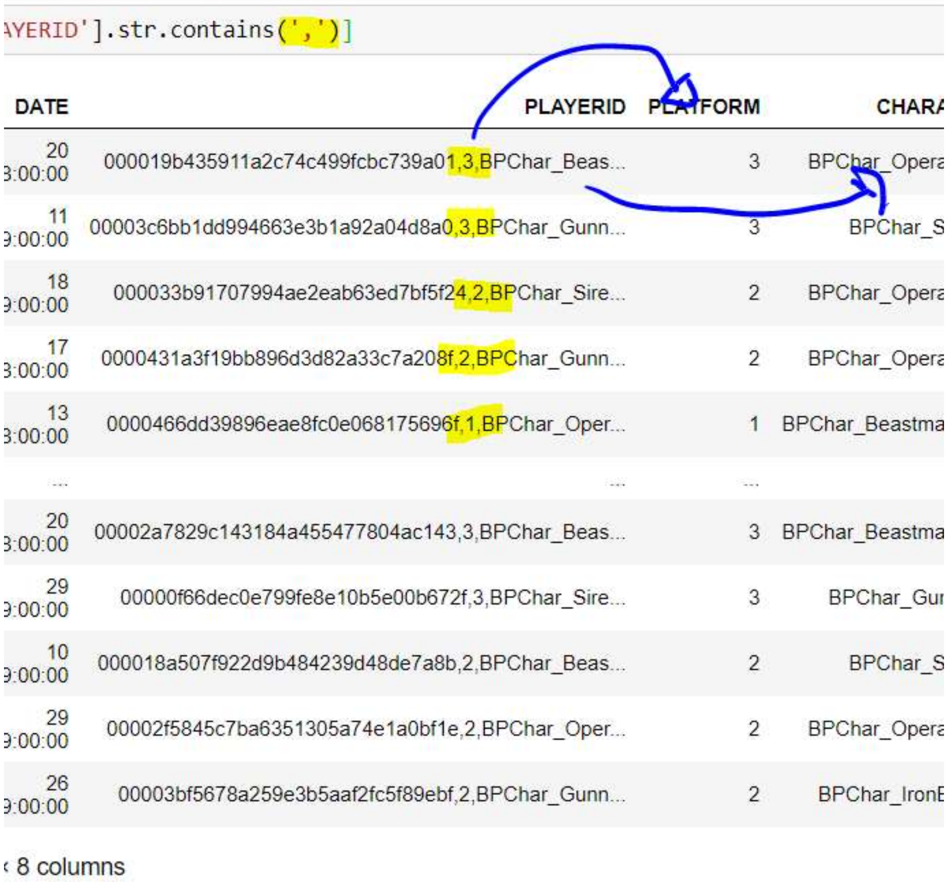


Cleaning the Data

Something went wrong in the data ingestion

The CSV

- 1,700,000 rows; 500,000 corrupted
    - Almost 1/3… can’t drop these
- What’s wrong with them?
    - Data in many columns shifted several cells over
- Coincidence?
    - There were also ~500,000 rows in the “clean” data with *extra* rows of data within a single cell!
    - Because of the similarity in numbers, I came to the conclusion that the 500,000 corrupted rows of data were caused because of these rows
- Rest of the clean data was ingested into **Python**

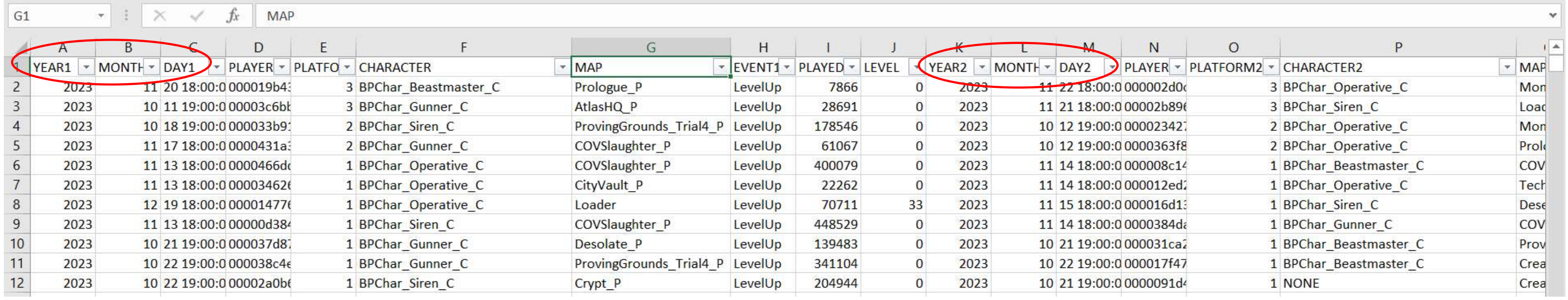


```
YERID'].str.contains(',')]
```

DATE	PLAYERID	PLATFORM	CHARA
20 3:00:00	000019b435911a2c74c499fcbc739a01,3,BPChar_Beas...	3	BPChar_Opera
11 9:00:00	00003c6bb1dd994663e3b1a92a04d8a0,3,BPChar_Gunn...	3	BPChar_S
18 9:00:00	000033b91707994ae2eab63ed7bf5f24,2,BPChar_Sire...	2	BPChar_Opera
17 3:00:00	0000431a3f19bb896d3d82a33c7a208f,2,BPChar_Gunn...	2	BPChar_Opera
13 3:00:00	0000466dd39896eae8fc0e068175696f,1,BPChar_Oper...	1	BPChar_Beastma
...	...	...	
20 3:00:00	00002a7829c143184a455477804ac143,3,BPChar_Beas...	3	BPChar_Beastma
29 9:00:00	00000f66dec0e799fe8e10b5e00b672f,3,BPChar_Sire...	3	BPChar_Gur
10 9:00:00	000018a507f922d9b484239d48de7a8b,2,BPChar_Beas...	2	BPChar_S
29 9:00:00	00002f5845c7ba6351305a74e1a0bf1e,2,BPChar_Oper...	2	BPChar_Opera
26 9:00:00	00003bf5678a259e3b5aaf2fc5f89ebf,2,BPChar_Gunn...	2	BPChar_IronE

8 columns

Cleaning in Excel



G1 MAP

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	YEAR1	MONTH	DAY1	PLAYER	PLATFO	CHARACTER	MAP	EVENT1	PLAYED	LEVEL	YEAR2	MONTH	DAY2	PLAYER	PLATFORM2	CHARACTER2	MAP
2	2023	11	20 18:00:0	000019b43	3	BPChar_Beastmaster_C	Prologue_P	LevelUp	7866	0	2023	11	22 18:00:0	000002d0c	3	BPChar_Operative_C	Mon
3	2023	10	11 19:00:0	00003c6bb	3	BPChar_Gunner_C	AtlasHQ_P	LevelUp	28691	0	2023	11	21 18:00:0	00002b896	3	BPChar_Siren_C	Load
4	2023	10	18 19:00:0	000033b91	2	BPChar_Siren_C	ProvingGrounds_Trial4_P	LevelUp	178546	0	2023	10	12 19:00:0	000023427	2	BPChar_Operative_C	Mon
5	2023	11	17 18:00:0	0000431a3	2	BPChar_Gunner_C	COVSlaughter_P	LevelUp	61067	0	2023	10	12 19:00:0	0000363f8	2	BPChar_Operative_C	Prol
6	2023	11	13 18:00:0	0000466dc	1	BPChar_Operative_C	COVSlaughter_P	LevelUp	400079	0	2023	11	14 18:00:0	000008c14	1	BPChar_Beastmaster_C	COV
7	2023	11	13 18:00:0	000034626	1	BPChar_Operative_C	CityVault_P	LevelUp	22262	0	2023	11	14 18:00:0	000012ed2	1	BPChar_Operative_C	Tech
8	2023	12	19 18:00:0	000014776	1	BPChar_Operative_C	Loader	LevelUp	70711	33	2023	11	15 18:00:0	000016d13	1	BPChar_Siren_C	Dese
9	2023	11	13 18:00:0	00000d384	1	BPChar_Siren_C	COVSlaughter_P	LevelUp	448529	0	2023	11	14 18:00:0	0000384da	1	BPChar_Gunner_C	COV
10	2023	10	21 19:00:0	000037d87	1	BPChar_Gunner_C	Desolate_P	LevelUp	139483	0	2023	10	21 19:00:0	000031ca2	1	BPChar_Beastmaster_C	Prov
11	2023	10	22 19:00:0	000038c4e	1	BPChar_Gunner_C	ProvingGrounds_Trial4_P	LevelUp	341104	0	2023	10	22 19:00:0	000017f47	1	BPChar_Beastmaster_C	Crea
12	2023	10	22 19:00:0	00002a0b6	1	BPChar_Siren_C	Crypt_P	LevelUp	204944	0	2023	10	21 19:00:0	0000091d4	1	NONE	Crea

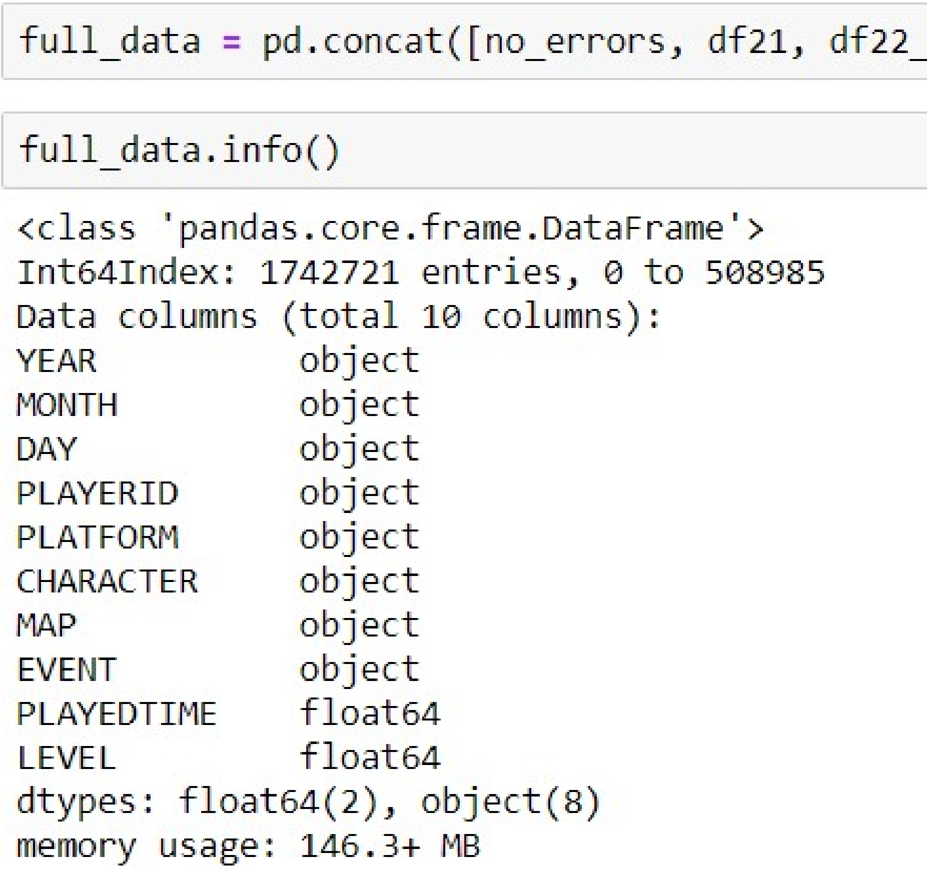
- Brought corrupted rows into excel to work in real time with them
- Split by commas and made a table with two tables' worth of columns
- Brought the data into python and split them into two tables, renaming for consistency

```
df21 = df2[['YEAR1','MONTH1','DAY1','PLAYERID1','PLATFORM1','CHARACTER','MAP','EVENT1','PLAYEDTIME','LEVEL']]
```

```
df21.rename({'YEAR1':'YEAR', 'MONTH1':'MONTH','DAY1':'DAY','PLAYERID1':'PLAYERID','PLATFORM1':'PLATFORM','CHARACTER1':'CHARACTER
            'MAP1':'MAP','EVENT1':'EVENT','PLAYEDTIME1':'PLAYEDTIME','LEVEL1':'LEVEL'}, axis = 1, inplace = True)
```

Bringing it together

- Used Python to bring all of the data together
- There were still some issues with the data. Some outliers and corrupted data, but only <1%
    - Examples: Data remaining in wrong row
        - Used RegEx and Type constraints to find them
    - Dropped them
- Result: 1.7 mil clean rows of data



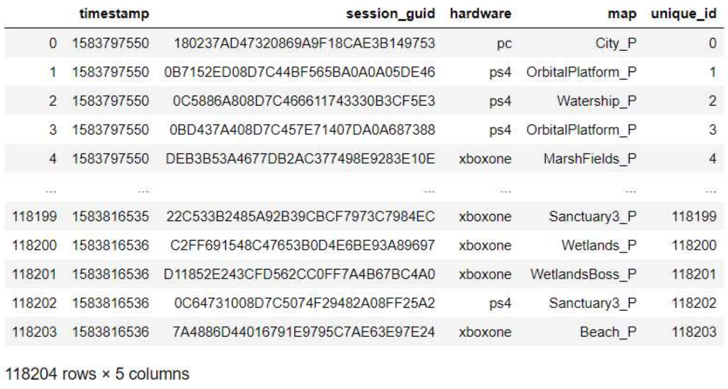
```
full_data = pd.concat([no_errors, df21, df22_
```

```
full_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1742721 entries, 0 to 508985
Data columns (total 10 columns):
YEAR          object
MONTH         object
DAY           object
PLAYERID      object
PLATFORM      object
CHARACTER     object
MAP           object
EVENT         object
PLAYEDTIME    float64
LEVEL         float64
dtypes: float64(2), object(8)
memory usage: 146.3+ MB
```

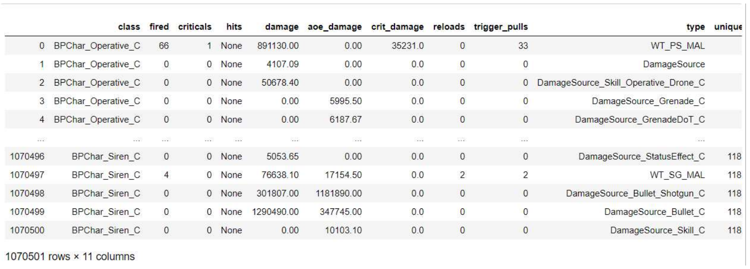
The JSON file

- The JSON was a clean dataset, but needed to be worked on for Python and Jupyter Notebook use
- Created two tables: one for ‘context’ and one for ‘weapons’
  - Assigned Unique Identifier for each JSON line
- 5 Hours worth of Data
- ~100,000 in context and 1,000,000 in weapons



	timestamp	session_guid	hardware	map	unique_id
0	1583797550	180237AD47320869A9F18CAE3B149753	pc	City_P	0
1	1583797550	0B7152ED08D7C44BF565BA0A0A05DE46	ps4	OrbitalPlatform_P	1
2	1583797550	0C5886A808D7C466611743330B3CF5E3	ps4	Watership_P	2
3	1583797550	0BD437A408D7C457E71407DA0A687388	ps4	OrbitalPlatform_P	3
4	1583797550	DEB3B53A4677DB2AC377498E9283E10E	xboxone	MarshFields_P	4
...	...	...	...	...	...
118199	1583816535	22C533B2485A92B39CBCF7973C7984EC	xboxone	Sanctuary3_P	118199
118200	1583816536	C2FF691548C47653B0D4E6BE93A89697	xboxone	Wetlands_P	118200
118201	1583816536	D11852E243CFD562CC0FF7A4B67BC4A0	xboxone	WetlandsBoss_P	118201
118202	1583816536	0C64731008D7C5074F29482A08FF25A2	ps4	Sanctuary3_P	118202
118203	1583816536	7A4886D44016791E9795C7AE63E97E24	xboxone	Beach_P	118203

118204 rows × 5 columns



	class	fired	criticals	hits	damage	aoe_damage	crit_damage	reloads	trigger_pulls	type	unique
0	BPChar_Operative_C	66	1	None	891130.00	0.00	35231.0	0	33	WT_PS_MAL	
1	BPChar_Operative_C	0	0	None	4107.09	0.00	0.0	0	0	DamageSource	
2	BPChar_Operative_C	0	0	None	50678.40	0.00	0.0	0	0	DamageSource_Skill_Operative_Drone_C	
3	BPChar_Operative_C	0	0	None	0.00	5995.50	0.0	0	0	DamageSource_Grenade_C	
4	BPChar_Operative_C	0	0	None	0.00	6187.67	0.0	0	0	DamageSource_GrenadeDoT_C	
...	...	...	...	...	...	...	...	...	...	...	
1070496	BPChar_Siren_C	0	0	None	5053.65	0.00	0.0	0	0	DamageSource_StatusEffect_C	118
1070497	BPChar_Siren_C	4	0	None	76638.10	17154.50	0.0	2	2	WT_SG_MAL	118
1070498	BPChar_Siren_C	0	0	None	301807.00	1181890.00	0.0	0	0	DamageSource_Bullet_Shotgun_C	118
1070499	BPChar_Siren_C	0	0	None	1290490.00	347745.00	0.0	0	0	DamageSource_Bullet_C	118
1070500	BPChar_Siren_C	0	0	None	0.00	10103.10	0.0	0	0	DamageSource_Skill_C	118

1070501 rows × 11 columns

The Final Data Tables

full_data -> from CSV

Contains each player session, character they used, map, playtime, etc. Essentially usage data

1.7 million rows

Added columns: SESSION_TIME, DAY_NUM

- Will help in analysis

Context -> from JSON

More usage metrics, but these can be joined to the “weapons” table.

Subset of 5 hours of data

~100,000 rows

Added “date” column from Timestamp

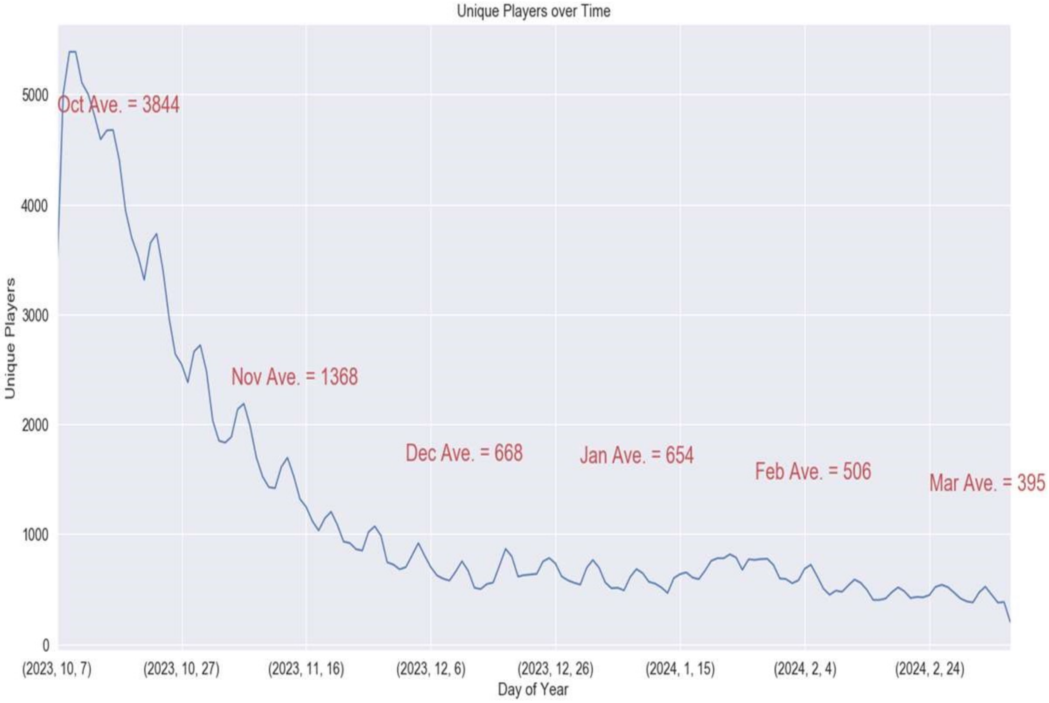
Weapons -> from JSON

Contains the character, type of weapon, amount and types of damage, etc.

~700,000 rows

Joins to “context” table in one-to-many relationship

Exploring the Data


Unique Players over Time

Oct Ave. = 3844

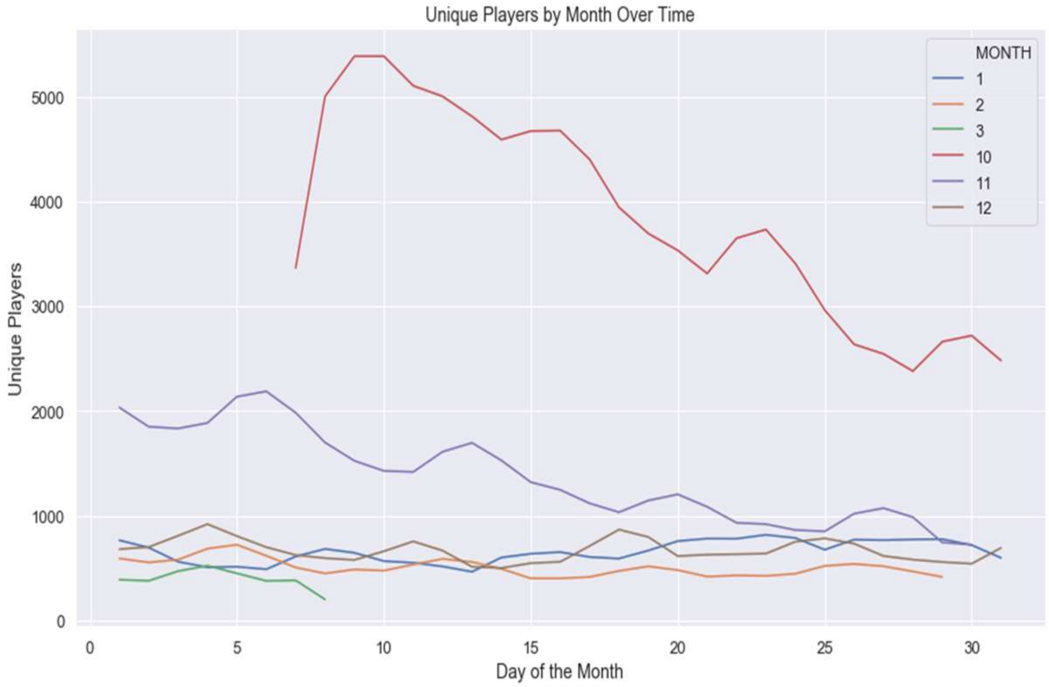
Nov Ave. = 1368

Dec Ave. = 668

Jan Ave. = 654

Feb Ave. = 506

Mar Ave. = 395

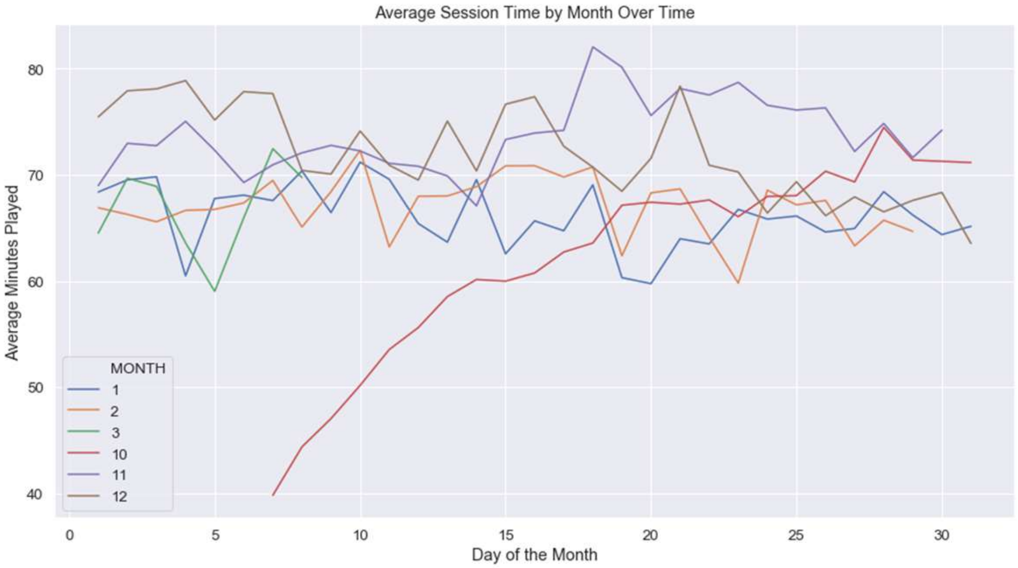

Unique Players by Month Over Time

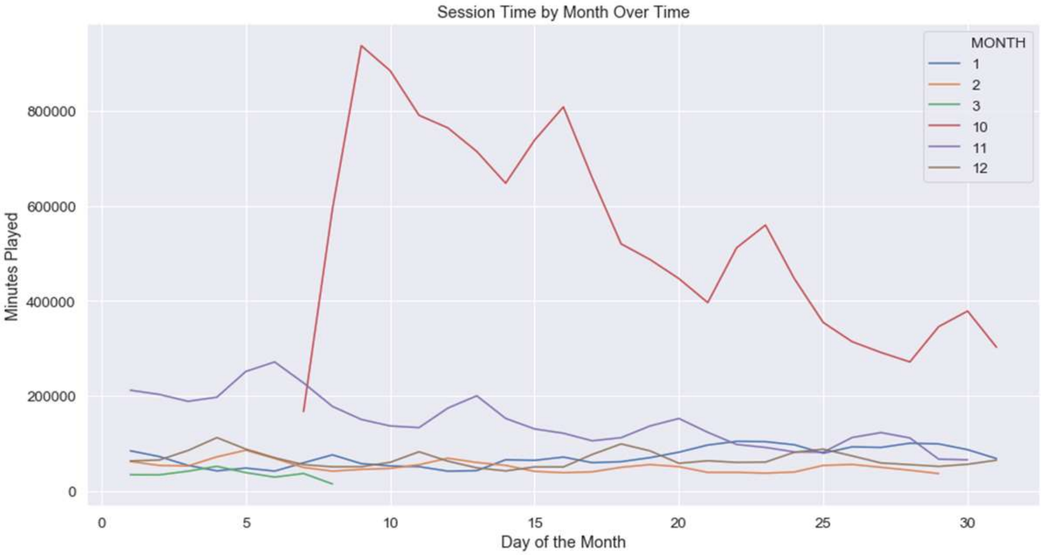
Unique Players per Month

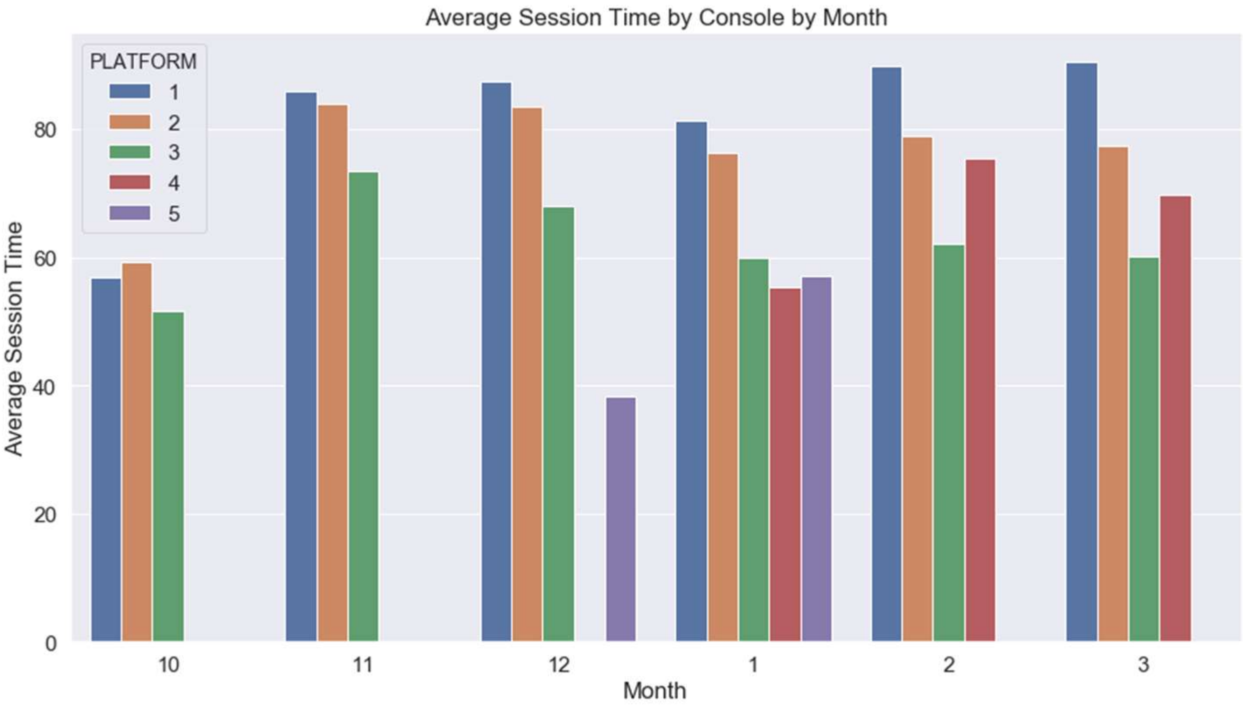
- Peak single-day Unique Users just after launch at 5,000+
- Has decreased to ~400 per day, 5 months after release
- Peaks on the weekends

The number of Total minutes played per day also increases on the weekends

People’s session times remain fairly constant over the months.

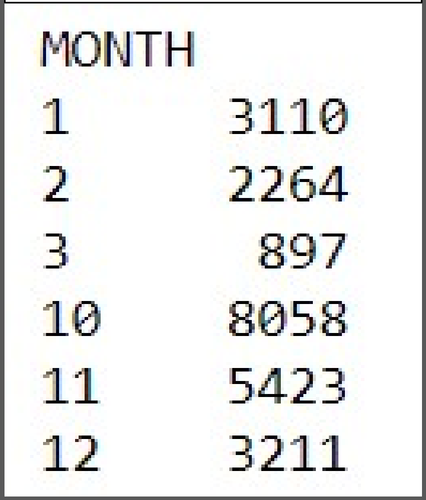




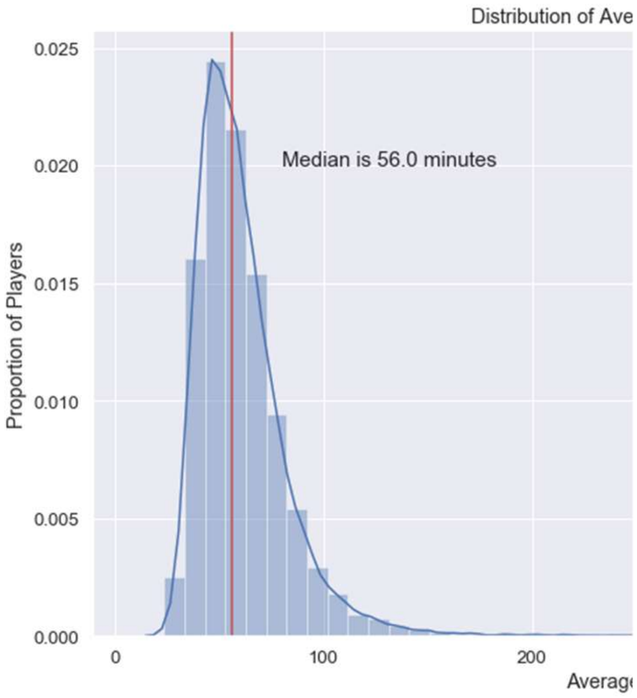


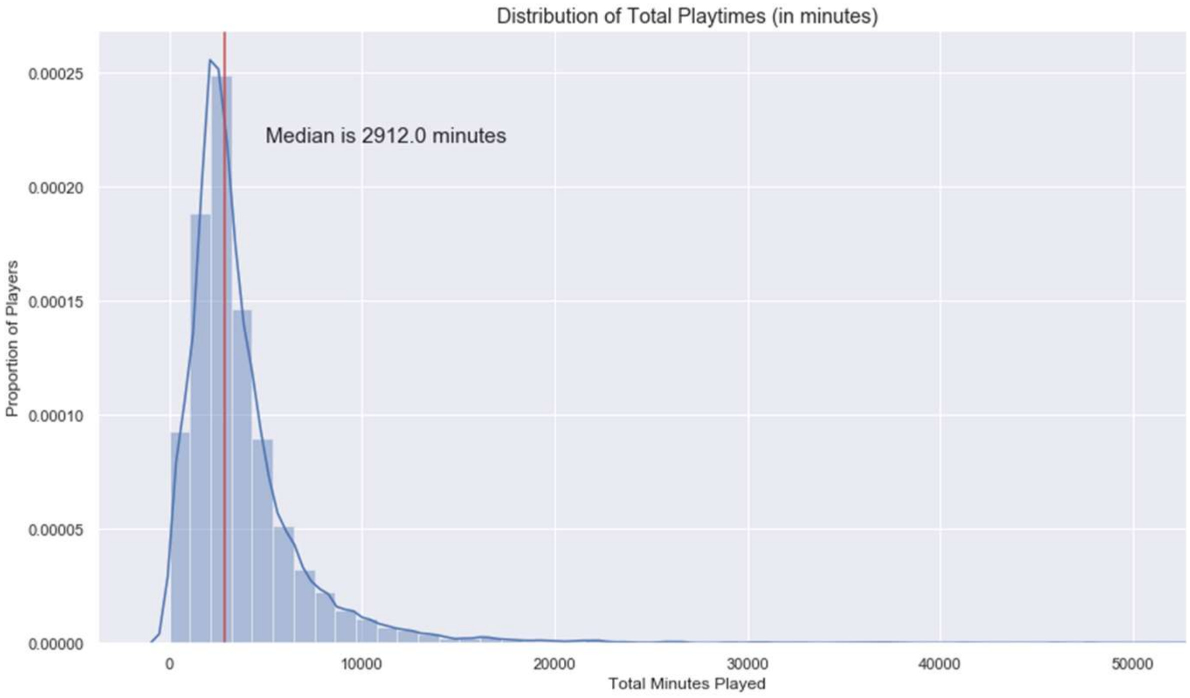
People played it for shorter amount of time in October, but there were more players

Number of Players



```
MONTH
1        3110
2        2264
3         897
10       8058
11       5423
12       3211
```





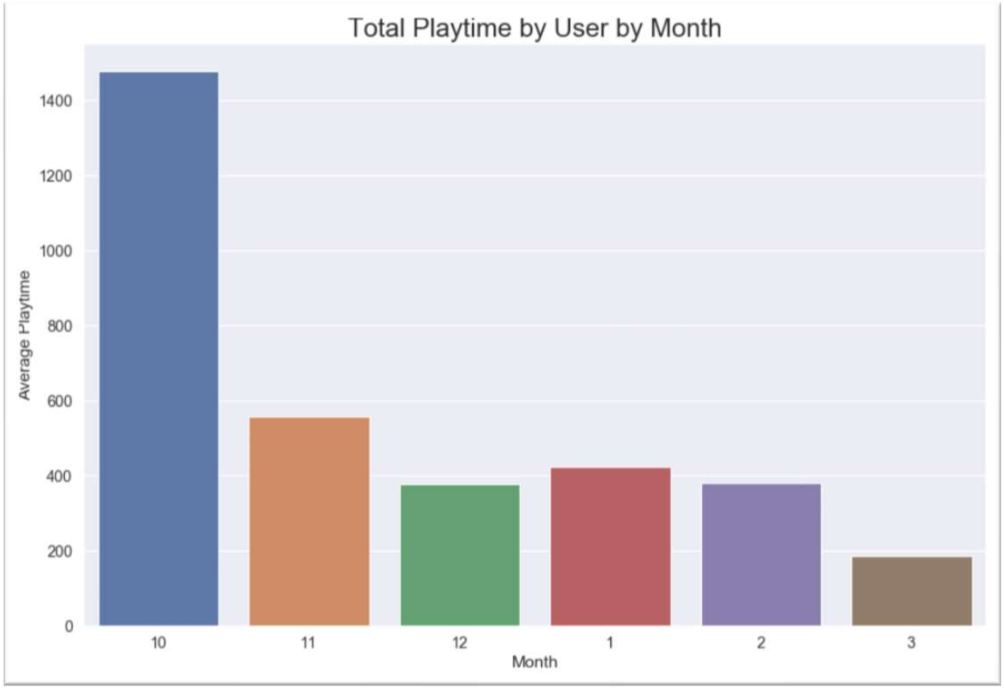
How long are people playing?

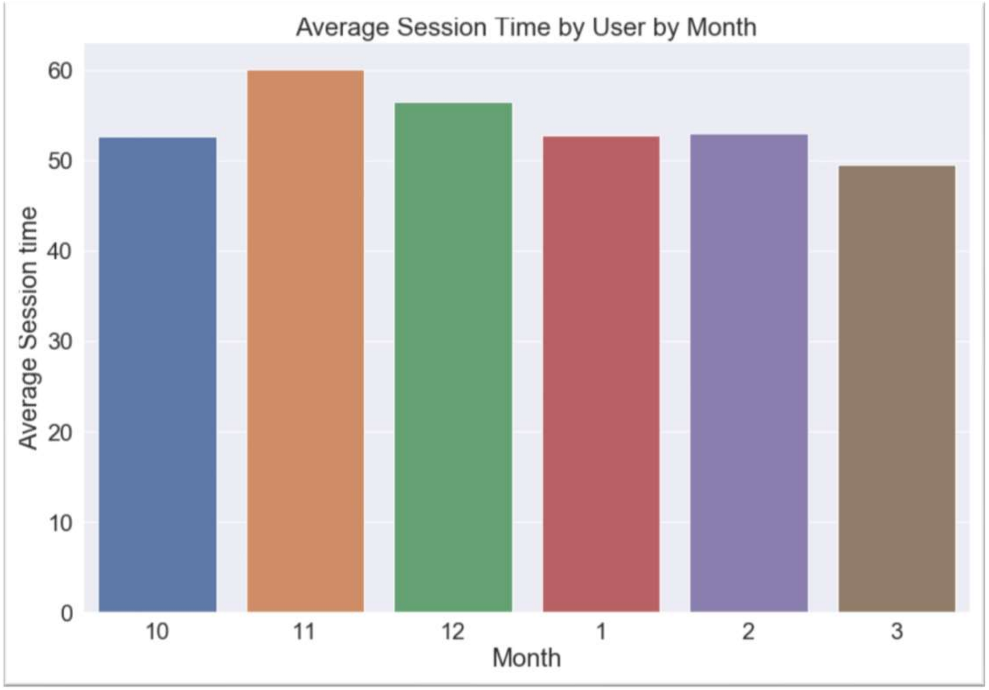
- Players are likely to spend almost 3,000 minutes (50 hours) total in the world of borderlands
- Most playing sessions are roughly an hour, but many extend to 2 or 3
  - Does it have the “Just one more” component?

Playtime By Month

Total Playtime heavily decreased by November, but increased again in January. Winter break?

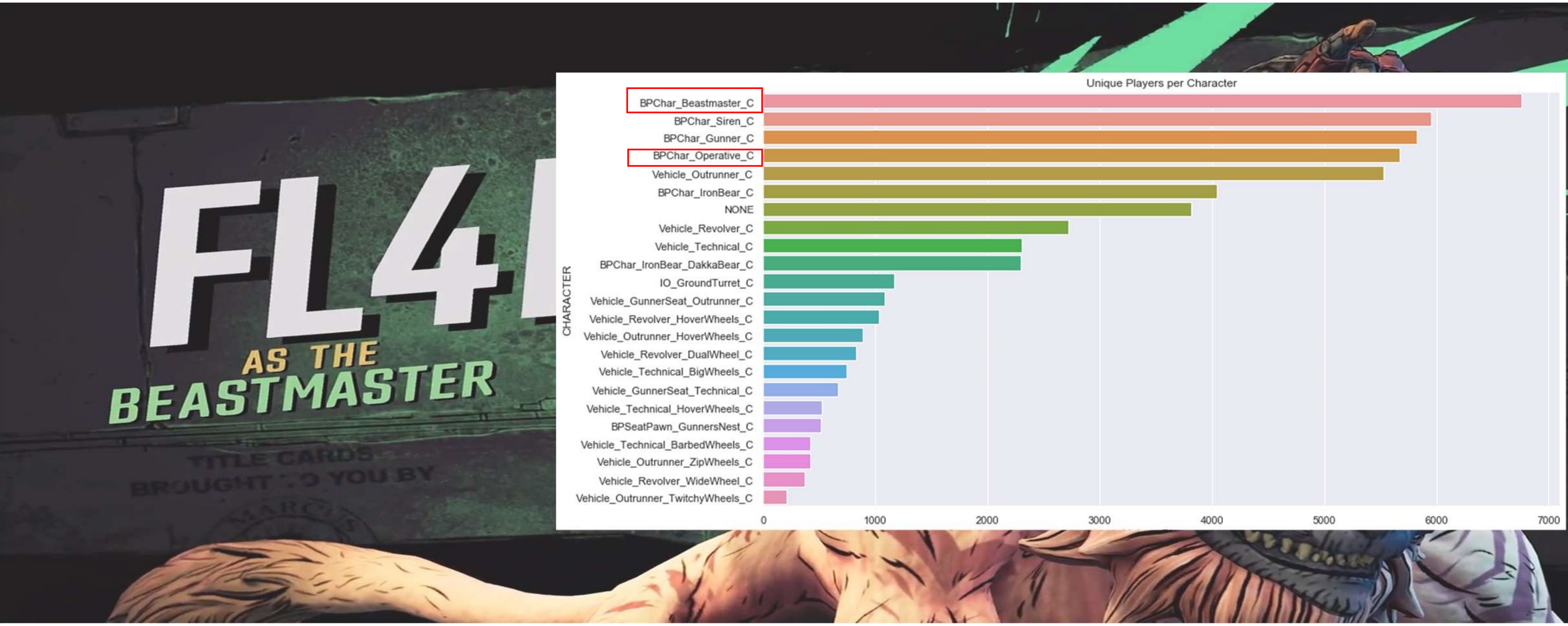
There is not a large different in average session over time. So even if less people are playing, they are still spending just as long playing





Summary on Playtimes

- October (release month) had the most unique players by far
  - Player count peaks on weekends.
- Though less players now, still play for the same time (~1 hour)
  - For some reason, average playtime was low in October
    - Possible reasons:
  - People played more on average in November and December. Thanksgiving and Holiday vacation?
- Total amount of time people play is ~3,000 minutes or about 50 hours of gameplay
  - This makes sense, as Borderlands is an open-world RPG



People LOVE Beast Master (FL4K)

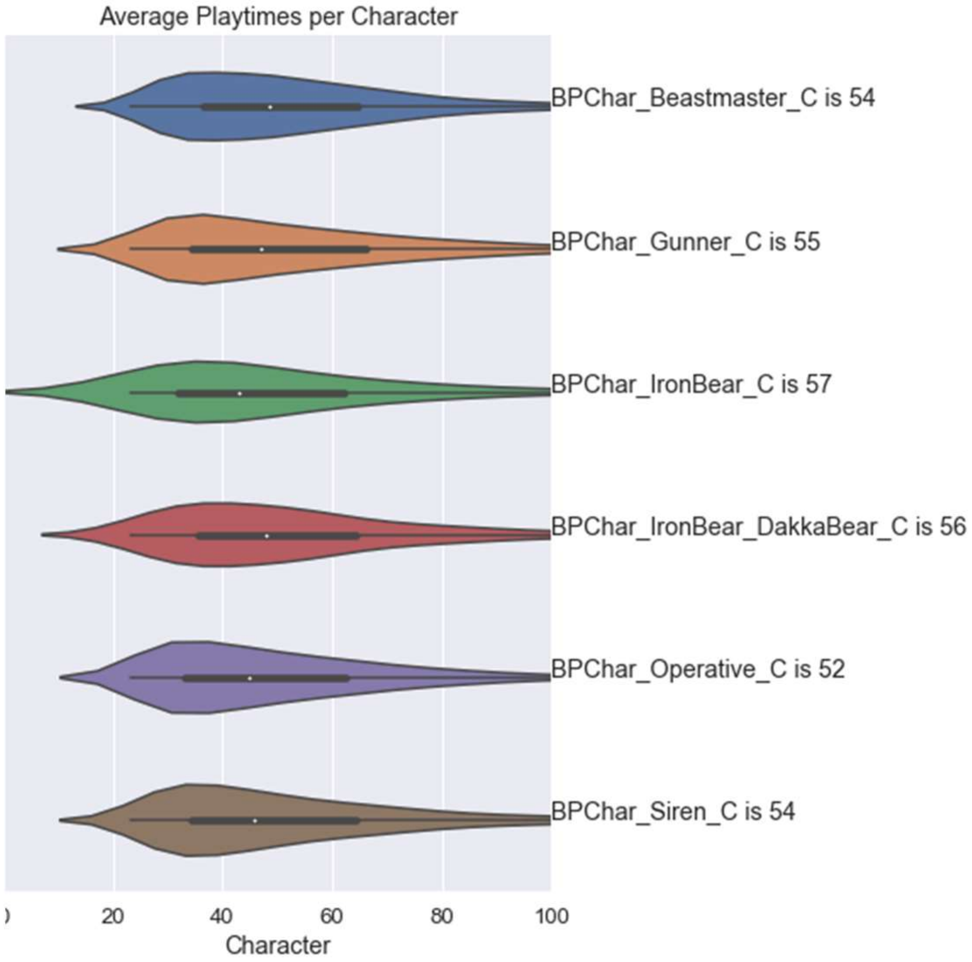
Almost 1,000 more Players than the next most

Operative comes last of 4 vault hunters

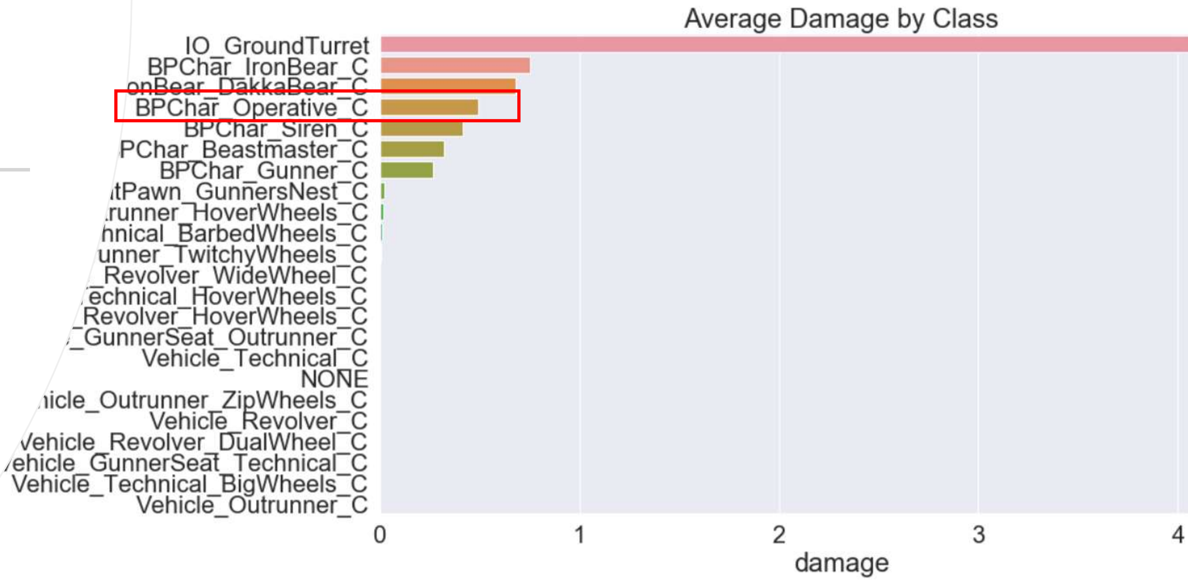
Which Vault Hunter is played for the longest

The Operative also is played with the lowest number of average minutes.

Why is this? Is there something wrong with the character?



Operative (Zane) does the Most Damage



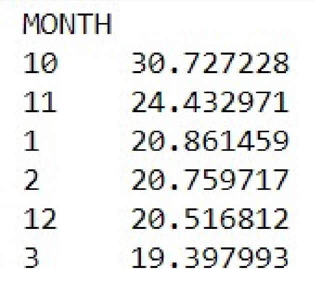
Summary on Characters

- Beastmaster has the highest number of Unique Players
    - Operative has the lowest
- Gunner is played the longest on average, while the Operative is played the shortest
    - Smaller differences. Can do statistical significance testing to follow up
- The Operative does the *most* damage on average. The Gunner does the least

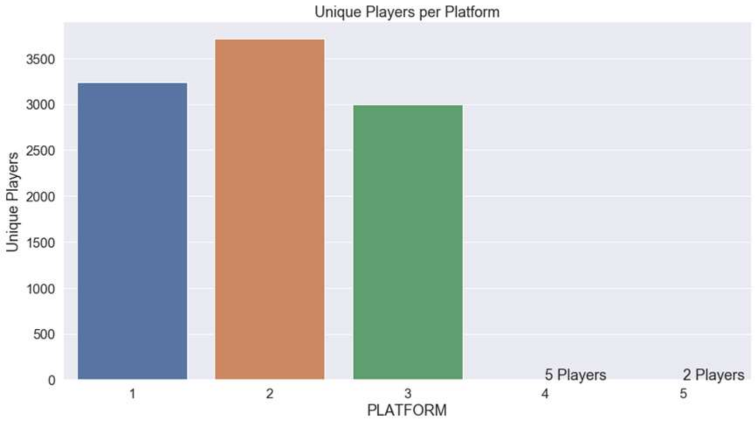
What are people using to play Borderlands 5?

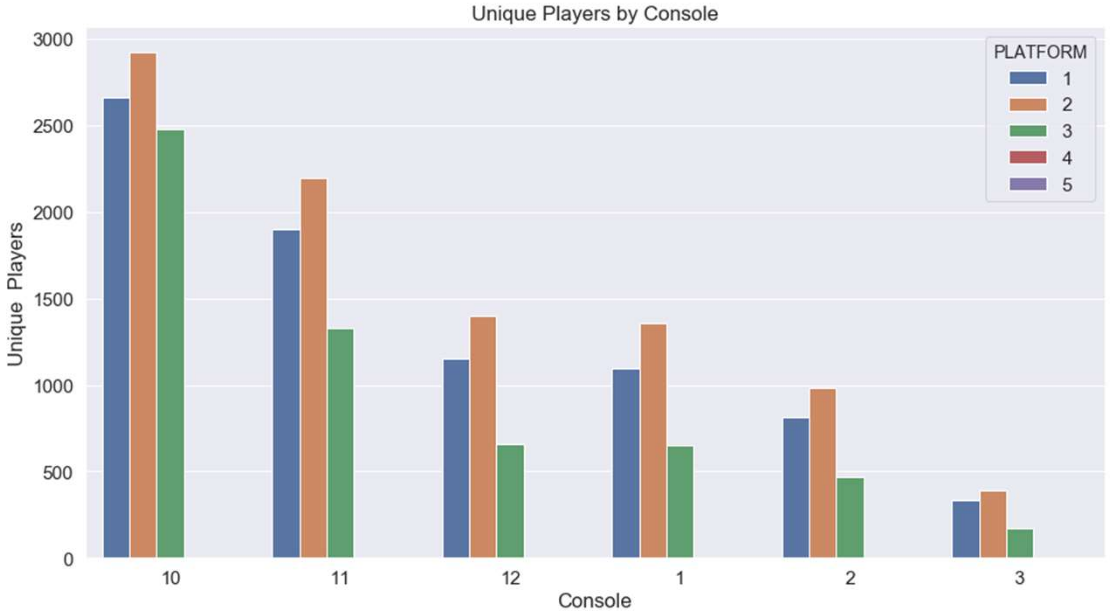
3 Platforms above the rest

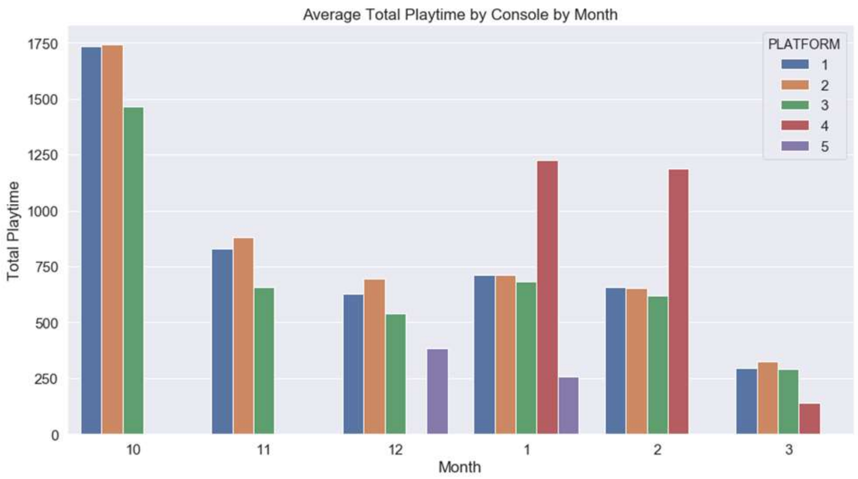
- Platform 3 starts off very strong, but quickly declines in proportion of users
    - 30% to %20 by January

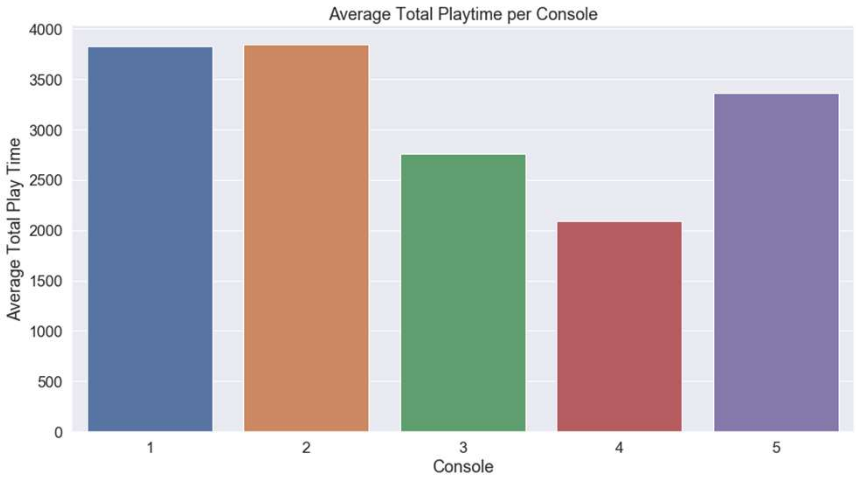


```
MONTH
10      30.727228
11      24.432971
1       20.861459
2       20.759717
12      20.516812
3       19.397993
```









Platform User Behaviors

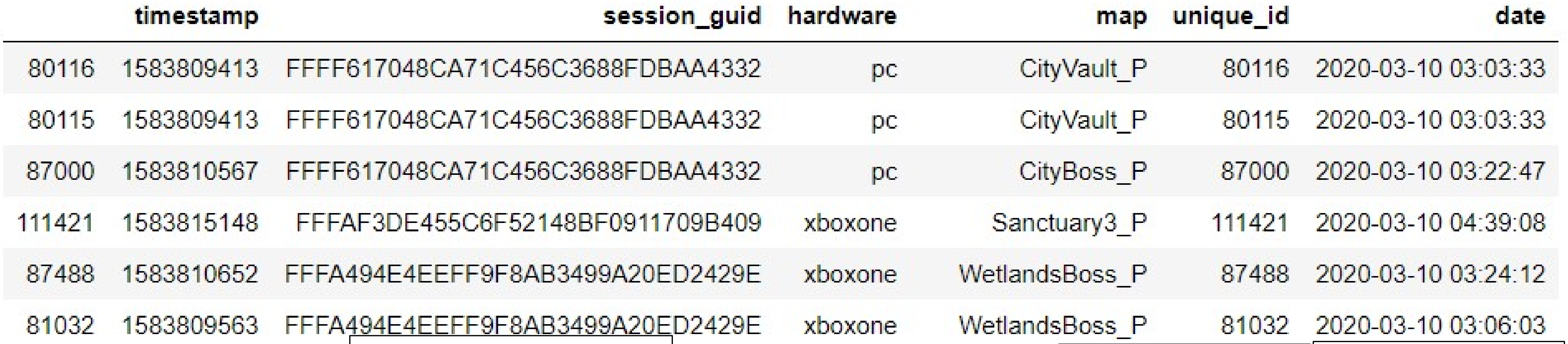
- The top two consoles (at this point I'm suspecting PS4 and Xbox), are played with similar habits
- Console 4 (I'm suspecting the Stadia) increases in January because it was released then
    - Quickly decreases. Not working well?

Summary on Platform Habits

- Platform 4 (Stadia?) didn’t release until January
  - Playtimes dropped fast though. Maybe poor experience?
- Platform 3 (PC?) started strong in platform-share but decreased in proportion
  - 30% of share to 20%
- Platforms 1 and 2 (PS4 and Xbox?) have the strongest playtimes and are roughly equal throughout

Let’s look at a subset of data

Looking at a subset of 5 hours of data



	timestamp	session_guid	hardware	map	unique_id	date
80116	1583809413	FFFF617048CA71C456C3688FDBAA4332	pc	CityVault_P	80116	2020-03-10 03:03:33
80115	1583809413	FFFF617048CA71C456C3688FDBAA4332	pc	CityVault_P	80115	2020-03-10 03:03:33
87000	1583810567	FFFF617048CA71C456C3688FDBAA4332	pc	CityBoss_P	87000	2020-03-10 03:22:47
111421	1583815148	FFFAF3DE455C6F52148BF0911709B409	xboxone	Sanctuary3_P	111421	2020-03-10 04:39:08
87488	1583810652	FFFA494E4EEFF9F8AB3499A20ED2429E	xboxone	WetlandsBoss_P	87488	2020-03-10 03:24:12
81032	1583809563	FFFA494E4EEFF9F8AB3499A20ED2429E	xboxone	WetlandsBoss_P	81032	2020-03-10 03:06:03

Session ID: Can appear more than once

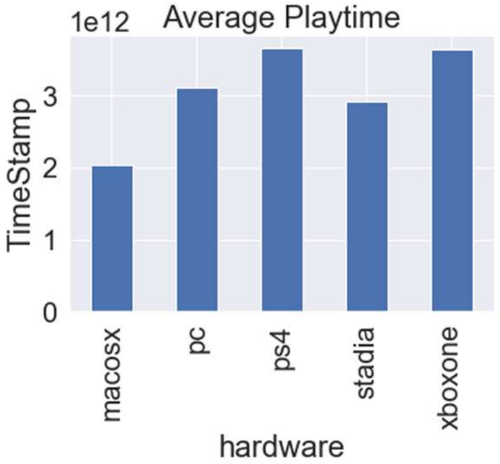
Created to join weapons table

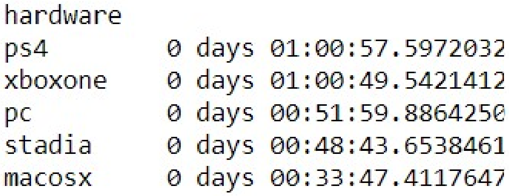
Created from timestamp: 5 hours of data

Checking the specifics

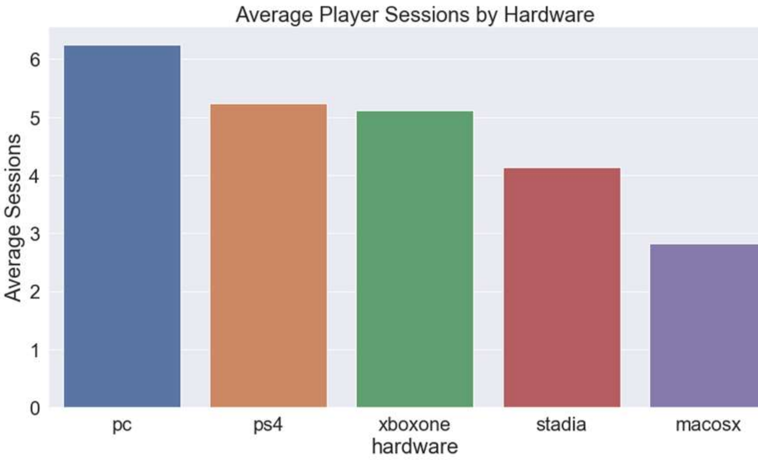
Less PC Players, but they’re… better?

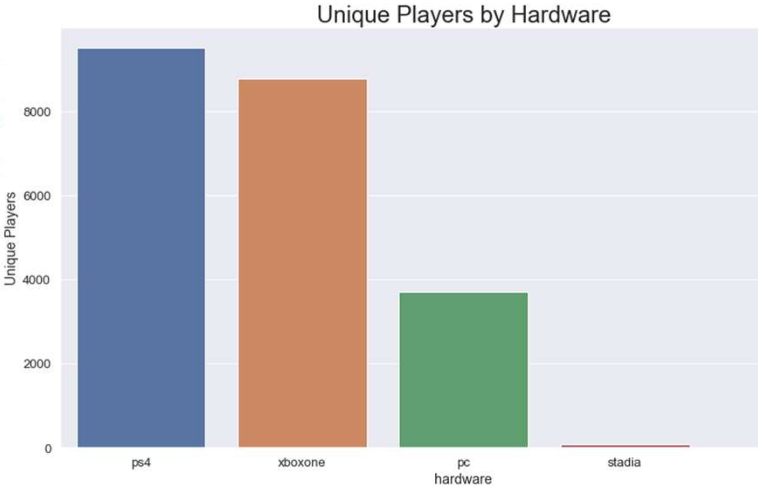
- There are only half the amount of PC players than PS4 and Xbox One players
- They “get more done” though, and in less time
- More experienced? Faster load times?





```
hardware
ps4          0 days 01:00:57.5972032
xboxone      0 days 01:00:49.5421412
pc           0 days 00:51:59.8864250
stadia       0 days 00:48:43.6538461
macosx       0 days 00:33:47.4117647
```





Final Recommendations