ELM CITY STORIES

DATAFEST 2022

BY

KANMANI NATARAJAN

MSBA/CSUEB

DATA CLEANING/WRANGLING

- First 13 columns are used for analysis (row_id, player_id, school, wave, session, date, event_id, event_description, event_category, event_time, event_time_dbl, stack_id, stack_title)
- No missing values in the first 10 columns
- stack_id & stack_title- 1816739 missing records
- **Filtered records using unique player_id and event_description with "Player completes Stack".**
- Duplicate entries were removed using subset ['player_id','event_time_dbl','stack_id']
- Some players played each stack more than once. For those cases, the completion time for the first attempt is taken for analysis.

PRILIMINARY FINDINGS!

How many players finished the last stack(My first paycheck)?
40 players

How many players have completed all the stacks/finished the game?
36 players

What is the average time spent across all the players?
55 hours 52 minutes(More than 2 days!!)

TIME DURATION TO COMPLETE EACH STACK

	player_id	event_time_dbl	stack_id
174	6427001	852	12
1856	6427001	4687	0
2420	6427001	6275	1
3383	6427001	8677	2
3948	6427001	10065	3
4600	6427001	11932	4
6006	6427001	14977	5
7340	6427001	17949	6
8156	6427001	19874	7
9219	6427001	22683	8
11965	6427001	28419	10
12361	6427001	29444	9
14925	6427001	37485	11

Records are sequentially placed. So, the time taken to complete the first stack is the same as event_time_dbl.
For subsequent stacks, the current "event_time_dbl" is subtracted from the previous record.

Question: Have players completed each stack sequentially in the same order?

CONTD.,

```
time_index
```

```
[[852, 3835, 1588, 2402, 1388, 1867, 3045, 2972, 1925, 2809, 5736, 1025, 8041],
 [789, 2057],
 [705, 2561, 1414, 2535, 739, 3855, 461, 3735, 1151, 1304, 2615, 1450, 6901],
 [810],
 [711, 2008, 1721, 4903, 1523, 1540, 4229, 2759, 2460, 1253, 3652, 1178, 3717],
 [702],
 [644, 2656, 3128, 2419, 6146, 1191, 4633],
 [968],
 [576, 2535, 950, 2594, 2046, 197, 2671, 1584, 2617, 917, 2891, 975],
 [570, 3234, 1390, 3256, 609, 1369, 2391, 1882, 1429, 1474, 1740, 893, 5490],
 [508, 4475, 1560, 3817, 1456, 1071, 10725],
 [711, 4607, 1779, 2910],
 [661, 3096, 1980, 5546, 922],
 [775, 4354, 2785, 4448, 1292, 4111, 5226, 3385, 1799, 3520],
 [962, 178119, 12388, 7721, 3991]
```

```
time_index[58:78]
```

```
[[965, 5509, 3034, 11342, 3043, 1792, 8702, 7128, 1429, 1596],
 [1140, 3153, 2249, 5298, 730, 1696, 5239, 875, 2044, 1433, 2335, 923, 3263
 [89824, 1134, 3714, 2269, 4320],
 [91446, 2655],
 [727, 3232, 1899, 2324, 1766, 1011, 3033, 2858, 3136, 7008, 8545, 534],
 [861, 7493, 5093, 10141, 900],
 [511, 3252, 2791, 4204, 757, 1969, 3192, 4201, 2360, 3350, 4775, 1226],
 [633, 264166, 7555, 268166],
 [784, 4527, 1426, 11681, 1586, 258042],
 [554, 5822, 1923],
 [658, 4708, 2374, 3995, 1172, 2048, 254285, 4041, 3375, 6255],
 [815, 5750, 2440, 3277, 3029, 766, 2704, 6553, 2153129],
 [889, 6794, 2039, 3623, 1008, 8526, 519, 1634, 691, 3789],
 [883],
 [642, 3937, 1261, 7622, 1038, 127, 3366, 2976, 2195, 3726, 2360, 615, 2062
 [790, 5109, 1737424, 2122, 2354, 3444, 1032, 2858],
 [840, 2241, 1595, 608974, 1837, 2572, 1171, 2048, 1650, 756, 1596],
 [779, 4229, 2568, 3885, 1056, 1167, 6208, 3789, 4026, 3382],
 [583, 10067, 348983, 844, 6097],
 [717, 3713, 1708, 2643, 1618, 867, 348152, 1046, 2986, 349604, 620]]
```

```
col_index
```

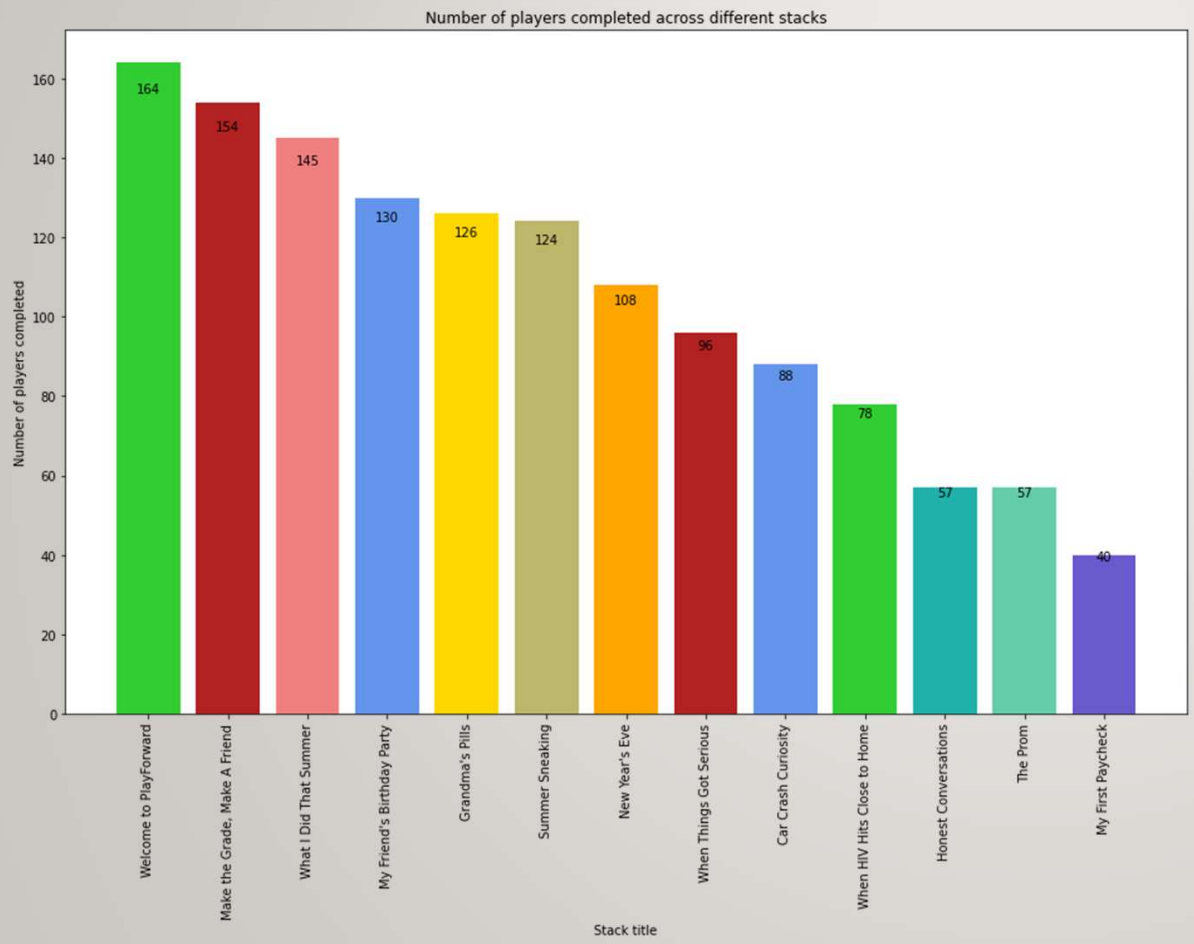
```
[[12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9, 11],
 [12, 0],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9, 11],
 [12],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9, 11],
 [12],
 [12, 0, 1, 2, 4, 3, 5],
 [12],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9, 11],
 [12, 0, 1, 2, 4, 3, 5],
 [12, 0, 1, 2],
 [12, 0, 1, 2, 3],
```

```
col_index[58:78]
```

```
[[12, 0, 1, 4, 2, 3, 5, 6, 7, 8],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9, 11],
 [0, 1, 2, 3, 4],
 [0, 1],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9],
 [12, 0, 1, 2, 3],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9],
 [12, 0, 1, 2],
 [12, 0, 1, 2, 3, 5],
 [12, 0, 1],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8],
 [12, 0, 1, 2, 4, 3, 5, 6, 7],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8],
 [12],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 9, 11],
 [12, 0, 6, 7, 8, 10, 9, 11],
 [12, 0, 1, 4, 5, 6, 7, 8, 10, 9, 11],
 [12, 0, 1, 2, 3, 4, 5, 6, 7, 8],
 [12, 0, 2, 3, 4],
```

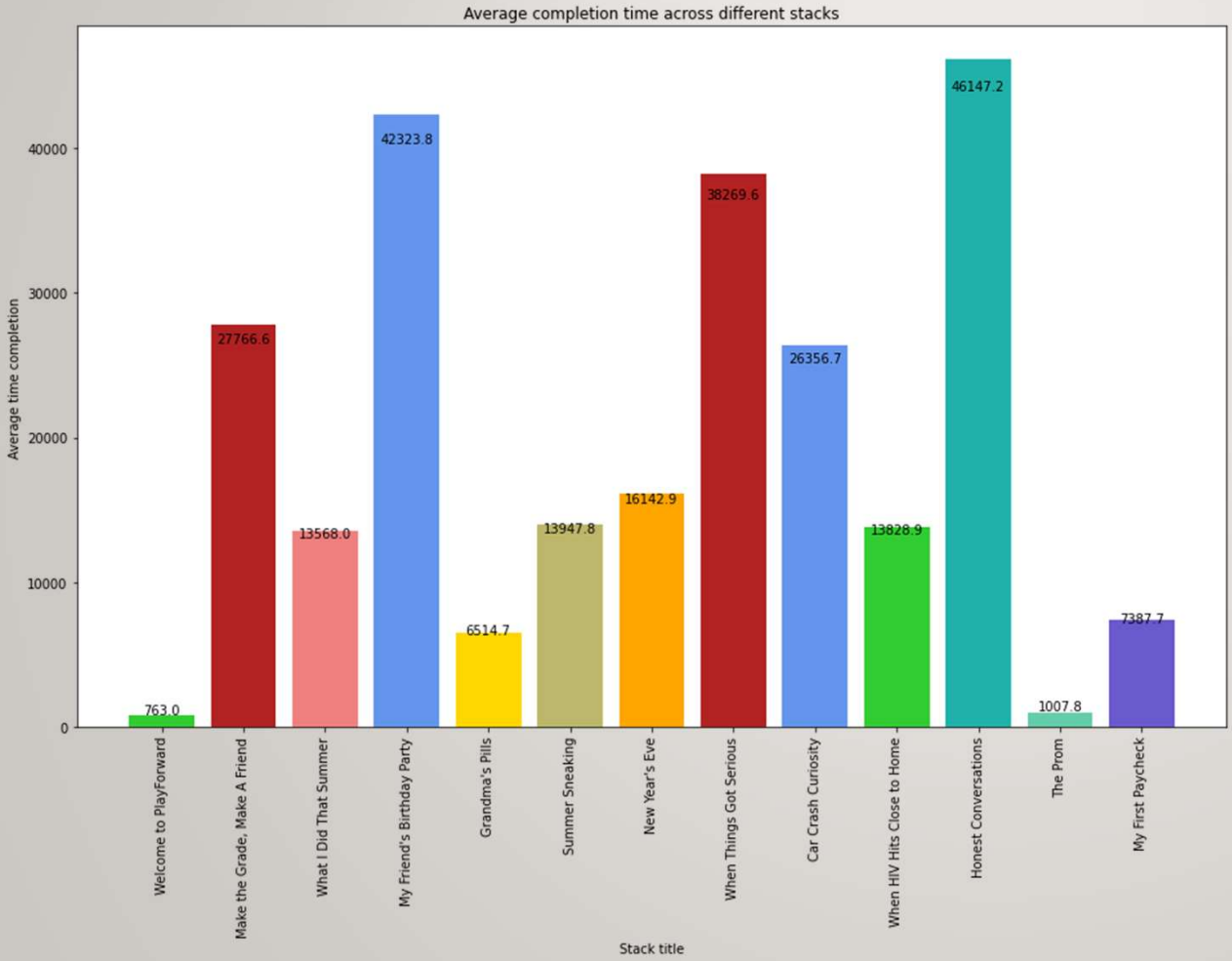
playerid	Welcome to PlayForward	Make the Grade, Make A Friend	What I Did That Summer	My Friend's Birthday Party	Grandma's Pills	Summer Sneaking	New Year's Eve	When Things Got Serious	Car Crash Curiosity	When HIV Hits Close to Home	Honest Conversations	The Prom	My First Paycheck
6486008	965.00000	5509	3034	3043	1792	11342	8702	7128	1429	1596	NC	NC	NC
6486009	1140.00000	3153	2249	5298	730	1696	5239	875	2044	1433	2335	923	3263
6486010	NC	89824	1134	3714	2269	4320	NC	NC	NC	NC	NC	NC	NC
6486011	NC	91446	2655	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
6486012	727	3232	1899	2324	1766	1011	3033	2858	3136	7008	8545	534	NC
6486013	861	7493	5093	10141	900	NC	NC	NC	NC	NC	NC	NC	NC
6486014	511	3252	2791	4204	757	1969	3192	4201	2360	3350	4775	1226	NC
6486015	633	264166	7555	268166	NC	NC	NC	NC	NC	NC	NC	NC	NC
6486016	784	4527	1426	11681	1586	NC	258042	NC	NC	NC	NC	NC	NC
6486017	554	5822	1923	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
6486018	658	4708	2374	3995	1172	2048	254285	4041	3375	6255	NC	NC	NC
6486019	815	5750	2440	3277	766	3029	2704	6553	2153129	NC	NC	NC	NC
6486020	889	6794	2039	3623	1008	8526	519	1634	691	3789	NC	NC	NC
6486021	883	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
6486022	642	3937	1261	7622	1038	127	3366	2976	2195	3726	2360	615	2062
6486023	790	5109	NC	NC	NC	NC	NC	1737424	2122	2354	3444	1032	2858
6486024	840	2241	1595	NC	NC	608974	1837	2572	1171	2048	1650	756	1596
6486025	779	4229	2568	3885	1056	1167	6208	3789	4026	3382	NC	NC	NC
6486026	583	10067	NC	348983	844	6097	NC	NC	NC	NC	NC	NC	NC
6486027	717	3713	1708	2643	1618	867	NC	348152	1046	2986	349604	620	NC

- **Created nested list for each player's time duration and their corresponding stack_id. Time duration is matched with the relevant stack_id**
- **NC means the stack is not completed by the player**



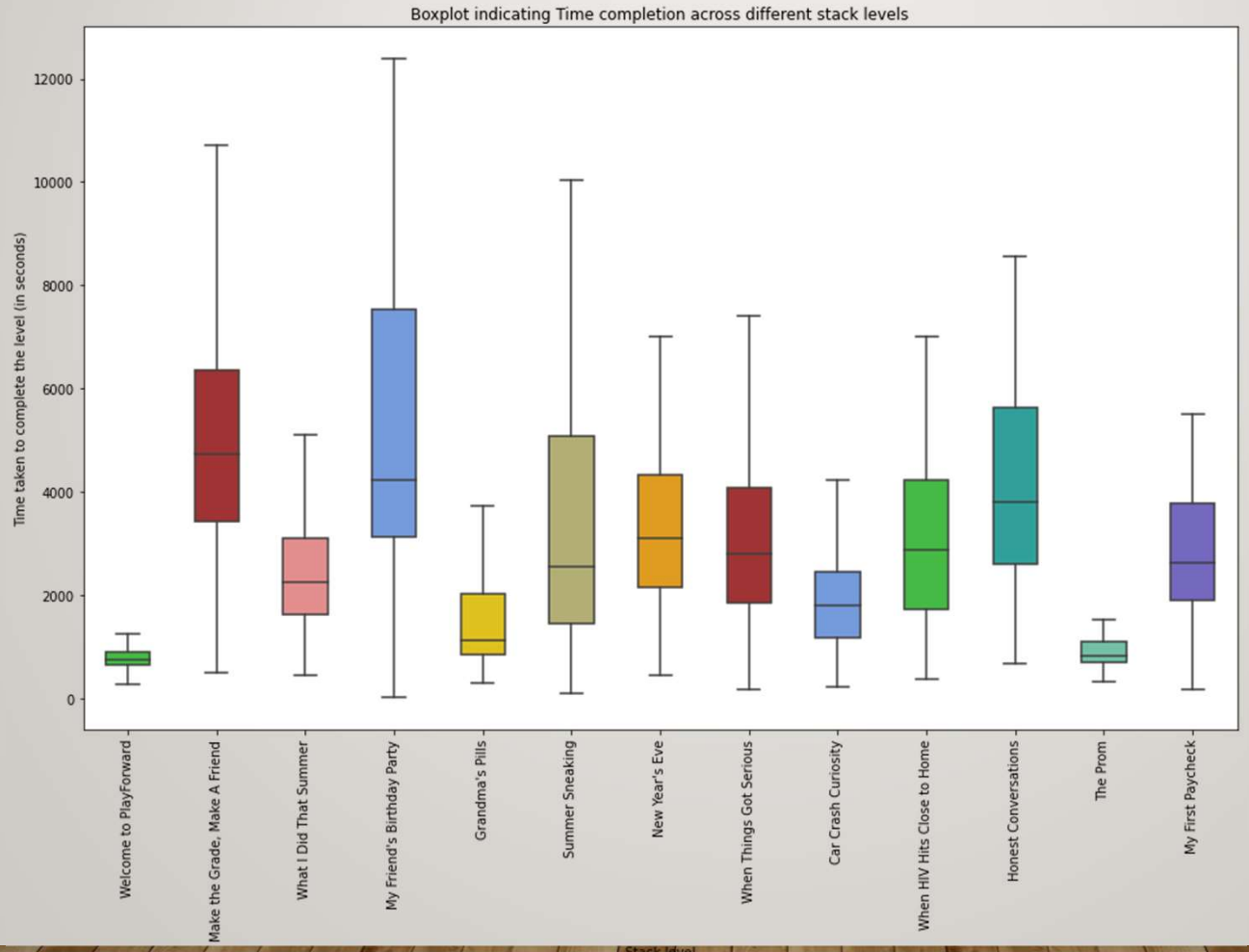
Percentage of players who finished the stack:

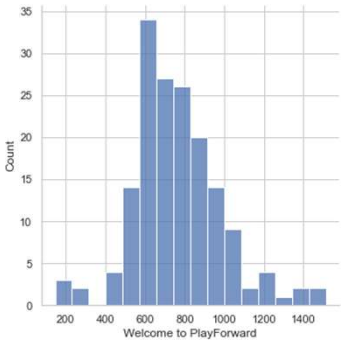
1. Welcome to PlayForward (98.8%)
2. Make the grade, Make a friend (92.8%)
3. What I did that summer (87.3%)
4. My friend's birthday party (78.3%)
5. Grandma's pills (75.9%)
6. Summer sneaking(74.7%)
7. New year's Eve (65%)
8. When things got serious (57.8%)
9. Car crash curiosity (53%)
10. When HIV hits close to home (46.9%)
11. Honest Conversation (34.3%)
12. The Prom (34.3%)
13. My first paycheck (24%)

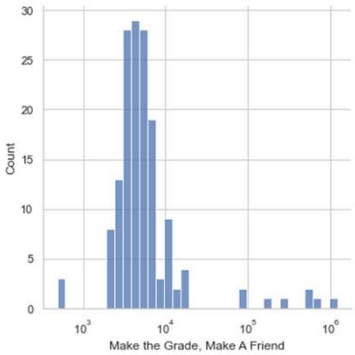


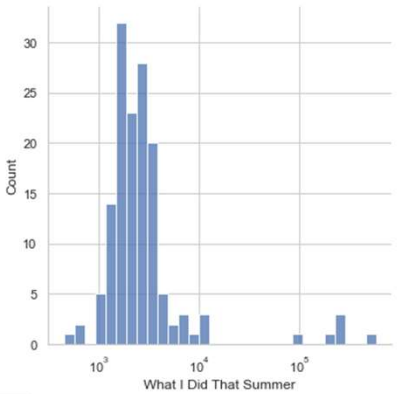
Highest Average time at each stack:

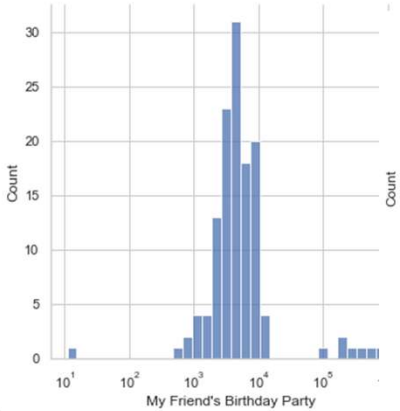
1. Honest Conversation
2. My friend's birthday party
3. When things got serious
4. Make the grade, Make a friend
5. Car crash curiosity
6. New year's Eve
7. Summer sneaking
8. When HIV hits close to home
9. What I did that summer
10. My first paycheck
11. Grandma's pills
12. The Prom
13. Welcome to PlayForward

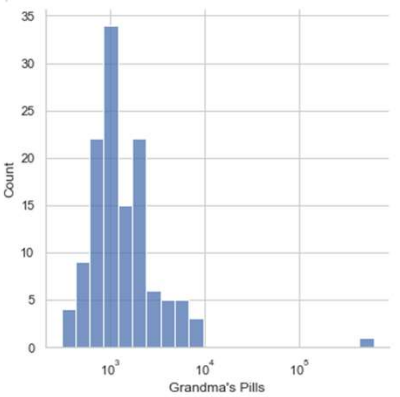

Boxplot indicating Time completion across different stack levels
Time taken to complete the level (in seconds)
0
2000
4000
6000
8000
10000
12000
Welcome to PlayForward
Make the Grade, Make A Friend
What I Did That Summer
My Friend's Birthday Party
Grandma's Pills
Summer Sneaking
New Year's Eve
When Things Got Serious
Car Crash Curiosity
When HIV Hits Close to Home
Honest Conversations
The Prom
My First Paycheck
Stack level

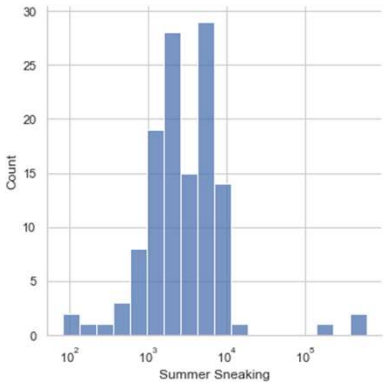

Count
Welcome to PlayForward

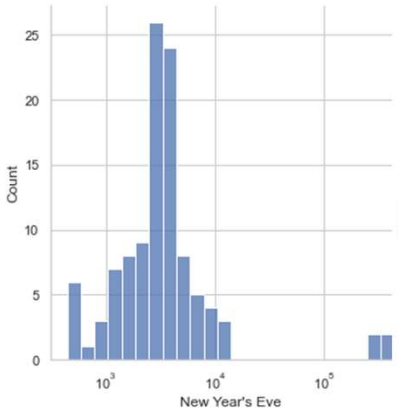

Count
Make the Grade, Make A Friend

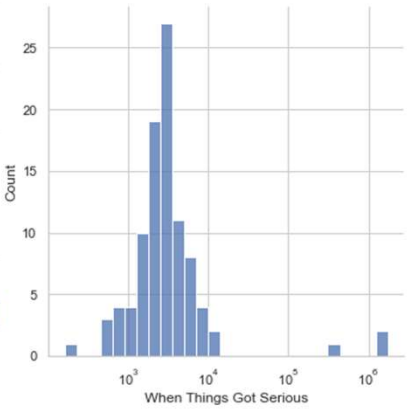

Count
What I Did That Summer

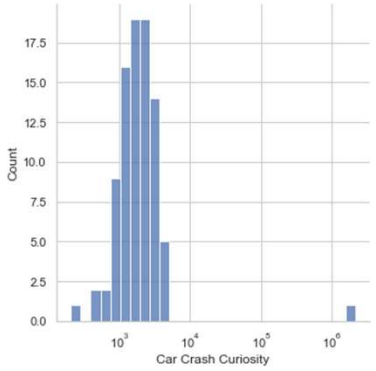

Count
My Friend's Birthday Party

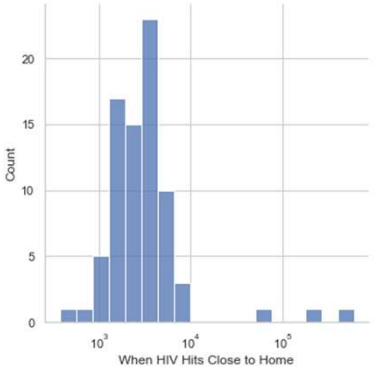

Count
Grandma's Pills

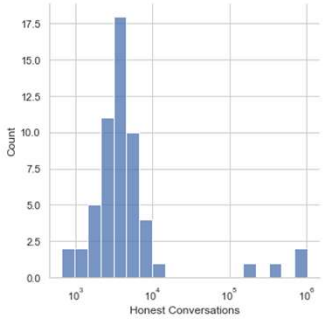

Count
Summer Sneaking

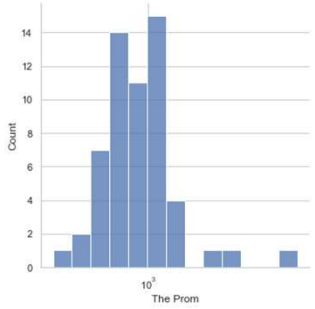

Count
New Year's Eve

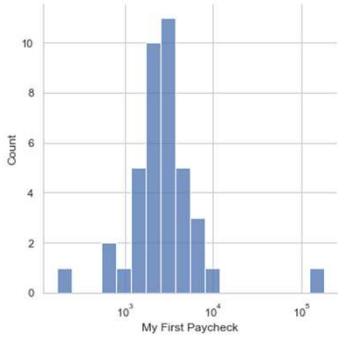

Count
When Things Got Serious


Count
Car Crash Curiosity


Count
When HIV Hits Close to Home


Count
Honest Conversations


Count
The Prom


Count
My First Paycheck

IS THE COMPLETION TIME OF PLAYERS WHO FINISHED DIFFERING ACROSS DIFFERENT STACKS?

Statistical analysis

```
data=pd.melt(df1)
data.head()
```

451]:

	variable	value
0	Welcome to PlayForward	852.00000
1	Welcome to PlayForward	789.00000
2	Welcome to PlayForward	705.00000
3	Welcome to PlayForward	810.00000
4	Welcome to PlayForward	711.00000

```
#one way ANOVA
from statsmodels.formula.api import ols
import statsmodels.api as sm
model=ols('value ~ variable',data=data).fit()
sm.stats.anova_lm(model,typ=1)
# p not less than 0.01. we do not reject null. Hence no population m
```

452]:

	df	sum_sq	mean_sq	F	PR(>F)
variable	12.00000	273369650698.67148	22780804224.88929	1.52705	0.10777
Residual	1354.00000	20199278048286.52734	14918226032.70792	NaN	NaN

Testing procedure: One way ANOVA (To test for differences among at least 3 groups)

Test assumptions:

- Response variable residuals are normally distributed (or approximately normally distributed)
- Variances of populations are equal.
- Responses for a given group are independent and identically distributed normal random variables

Hypothesis:

H0: Performance of players doesn't differ across different stacks

Ha: Performance of players differ across different stacks

Test result: P>0.05, Do not reject null

Answer: The completion time of players doesn't differ across different stacks!

To what extent do the 13 stack levels differ in terms of their effect on players'(who finished the game) completion time?

```
> #Load  File
> top_players<-read.csv("top_players.csv",header=TRUE)
> summary(top_players)
     index                                 treatments      value
 Min.   : 6427001   Car Crash Curiosity          : 36   Min.   :    127
 1st Qu.: 6430003   Grandma's Pills              : 36   1st Qu.:    993
 Median : 6486034   Honest Conversations         : 36   Median :   1915
 Mean   : 8151984   Make the Grade, Make A Friend: 36   Mean   :   5352
 3rd Qu.: 6607003   My First Paycheck            : 36   3rd Qu.:   3153
 Max.   :65670031   My Friend's Birthday Party   : 36   Max.   :1034035
                    (Other)                      :252
> top_players$index<-as.factor(top_players$index)
> #GiveAname <- aov(Response variable ~ treatment+block, data = dataFileName)
> test<-aov(value~treatments+index,data=top_players)
> summary(test)
             Df    Sum Sq   Mean Sq F value Pr(>F)
treatments   12 3.025e+10 2.521e+09   1.034  0.416
index        35 8.377e+10 2.394e+09   0.982  0.501
Residuals   420 1.024e+12 2.438e+09
```

Stack levels don't differ in terms of their effect on top players' completion time!

Experiment: Randomized block design (Introduce block as each player is different)

Assumptions:

- **Independence**. The dependent variable score for each experimental unit is independent of the score for any other unit.
- **Normality**. In the population, dependent variable scores are normally distributed within treatment groups.
- **Equality of variance**. In the population, the variance of dependent variable scores in each treatment group is equal. (Equality of variance is also known as **homogeneity of variance** or **homoscedasticity.)**
- H_0: $\beta_i = 0$ for all i
- H_1: $\beta_i \neq 0$ for some j
- Test result: P>0.05, Do not reject null hypothesis.

RANKING PLAYERS BASED ON THE COMPLETION TIME

playerid	Stack_finish_time	Total_time_spent	Extra_time	Rank_stack_finish	Rank_tot_time_spent
65670031	10863	11100	237	1	1
6430003	10925	24908	13983	2	4
6567003	20876	185411	164535	3	33
6506005	23565	23616	51	4	2
6608004	23850	24263	413	5	3
6427012	25727	26126	399	6	5
6486002	26217	26675	458	7	8
6486005	26271	26449	178	8	6
6486029	26451	26483	32	9	7
6486001	26764	27078	314	10	9
6427019	27750	27810	60	11	10
6626004	28236	28706	470	12	11
6427004	29426	30307	881	13	13
6427026	29714	30187	473	14	12
6608006	30142	30440	298	15	14
6486009	30378	30452	74	16	15
6486030	30523	632242	601719	17	35
6430002	30564	30620	56	18	16
6607002	31068	31130	62	19	17
6486038	31526	32221	695	20	20
6427006	31654	31927	273	21	18
6608002	31704	31983	279	22	19
6486022	31927	32302	375	23	21
6608003	31955	32628	673	24	22
6427033	32712	32741	29	25	23
6546009	33082	33260	178	26	24
6607005	33097	33281	184	27	25
6608005	34872	35124	252	28	26
6506004	34903	39696	4793	29	29
6486003	37249	37318	69	30	27
6427001	37485	37718	233	31	28
6427040	39800	39875	75	32	30
6546013	99490	100006	516	33	31
6506003	124164	124265	101	34	32
6486037	288757	288801	44	35	34
6608001	1061282	1061951	669	36	36

Who is the top player?

Player id: **65670031**

Shortest time spent to finish all the stacks:

~ 3hrs 1 minute (Extra time: Just around 3 minutes)

Who spent extra time to replay the stacks again?

Player id: **6486030, 6567003**

6486030:

8 hrs 28 min to finish all the stacks(shortest).

spent extra 167 hours of time to replay the stacks.

6567003:

5 hrs 47 min to finish all the stacks(shortest)

spent extra 45 hrs 42 min of time to replay the stacks.

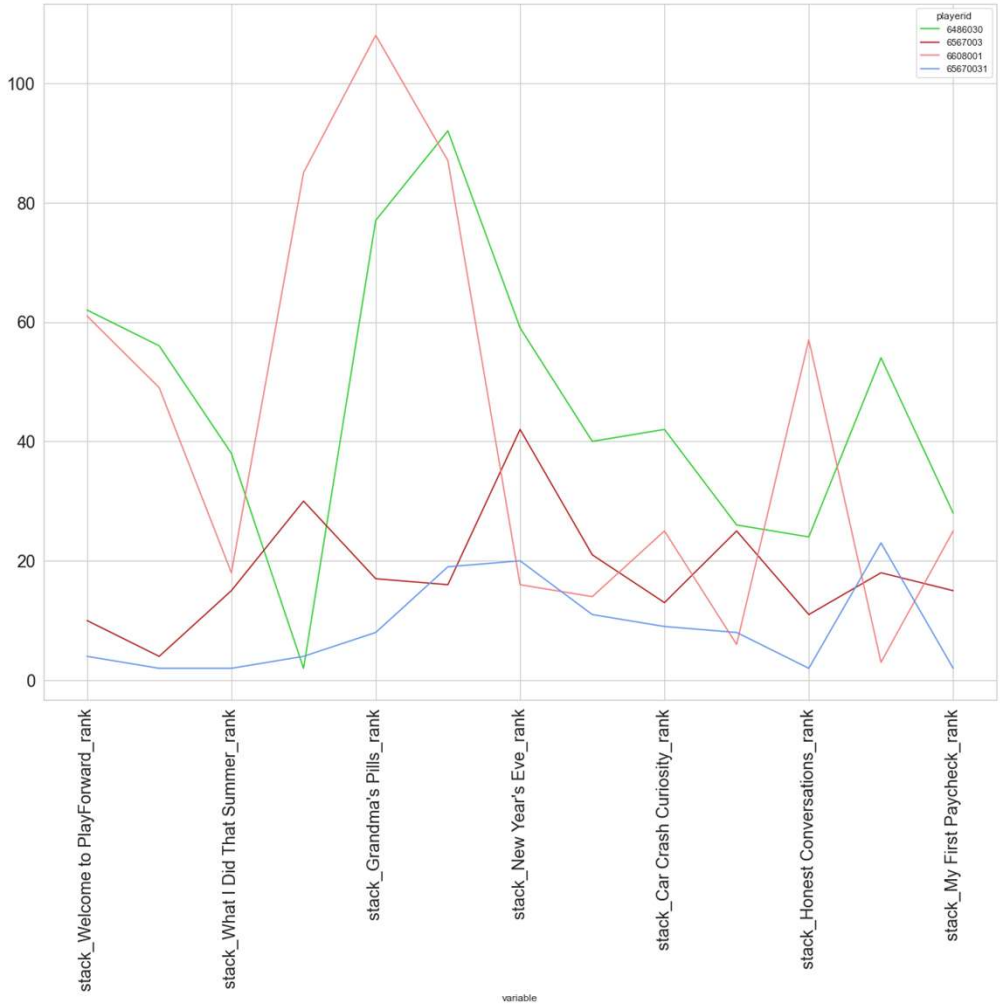
Who spent more time finishing the game once?

6608001

12 days 6 hrs 48 minutes

Maybe they took this game more seriously!!

Line plot indicating player's rank (based on completion time) across different stacks



THANK YOU!