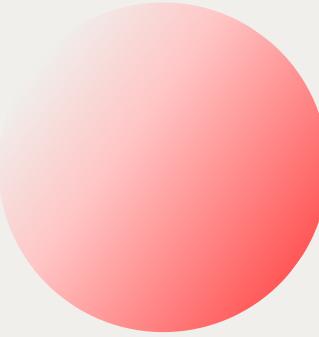


DOAN VU MINH THANH

# Bài test ROUND 1 - CADS/DC5 FPT 2023



doanvuminhthanh2404@gmail.com

source code: GitHub

# Table of Content

- Câu hỏi 1: tập log mới với format là dạng row, column. Mỗi row tương ứng dòng log, mỗi column tương ứng trường thuộc tính. Mỗi column cách nhau dấu tab.
- Câu hỏi 2: Kết hợp file user\_info.txt và tập data đã parse, phân tích hành vi đặc điểm sử dụng dịch vụ của những user này.
- Câu hỏi 3: Dựa vào kết quả phân tích, hãy đề ra giải pháp Dự đoán user có khả năng hủy sử dụng dịch vụ. Hãy hiện thực giải pháp đó (nếu có thể).

# Câu hỏi 1

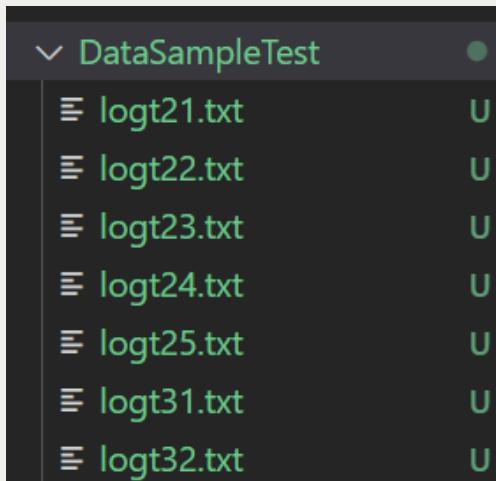
Tạo 1 tập log mới với format là dạng row, column. Mỗi row tương ứng dòng log, mỗi column tương ứng trường thuộc tính. Mỗi column cách nhau dấu tab.



# Phân tích đề bài

## Hướng xử lý

Có tổng cộng 6 file log dưới dạng .txt



đọc từng file .txt  
trong folder

dùng thư viện pandas  
để chuyển đổi sang  
dạng DataFrame

Mỗi record được lưu dưới dạng JSON hoặc dictionary, bao gồm các cặp key-value, trong đó key được biểu diễn bằng Unicode (u") và value là chuỗi. Các key đại diện cho các thuộc tính như 'ItemId', 'Firmware', 'DefaultGetway', vv.

```
{u'ItemId': u'100052388', u'RealTimePlaying': u'570.3', u'Firmware': u'2.4.14', u'DefaultGetway': u'192.168.1.1', u'BoxTime': u'2018-01-01T00:00:00', u'AppName': u'IPTV', u'Ip': u'192.168.1.25', u'ItemId': u'100052388', u'Firmware': u'2.4.14', u'DefaultGetway': u'192.168.1.1', u'BoxTime': u'2018-01-01T00:00:00', u'AppName': u'IPTV', u'Ip': u'192.168.1.25', u'ItemId': u'100052388', u'Firmware': u'2.4.14', u'DefaultGetway': u'192.168.1.1', u'BoxTime': u'2018-01-01T00:00:00', u'AppName': u'IPTV', u'Ip': u'192.168.1.25', u'ItemId': u'100052388', u'Firmware': u'2.4.14', u'DefaultGetway': u'192.168.1.1', u'BoxTime': u'2018-01-01T00:00:00', u'AppName': u'IPTV', u'Ip': u'192.168.1.25', u'ItemId': u'100052388', u'Firmware': u'2.4.14', u'DefaultGetway': u'192.168.1.1', u'BoxTime': u'2018-01-01T00:00:00', u'AppName': u'IPTV', u'Ip': u'192.168.1.25'}
```

lưu DataFrame dưới dạng .txt,  
các cột cách nhau bởi dấu cách

ghép các DataFrame từ các  
file và chọn những thuộc tính  
được yêu cầu

## DOAN VU MINH THANH

```
dataframes = []

# Iterate through each file in the folder
for file_name in file_list:
    if file_name.endswith('.txt'):
        file_path = os.path.join(folder_path, file_name)
        print(file_path)

        # Read the file
        with open(file_path, 'r') as file:
            data = file.readlines()

        # Process each line in the file
        cleaned_data = [ast.literal_eval(entry.replace('u\'', '\')) for entry in data]

        # Convert to DataFrame
        df = pd.DataFrame(cleaned_data)
        dataframes.append(df)
```

- Trở vào thư mục chứa các file log
- Duyệt từng file log, bỏ kí tự ‘u’ ở đầu các thuộc tính và chuyển đổi từng file log thành dạng list → DataFrame
- tổng hợp DataFrame của các file log vào 1 list dataframe

## DOAN VU MINH THANH

ghép các DataFrame của từng file thành 1 DataFrame

```
# Combine all DataFrames into a single one
combined_df = pd.concat(dataframes, ignore_index=True)
```

	ItemId	RealTimePlaying	Firmware	DefaultGetway	SubMenuId	Folder	\
0	100052388	570.3	2.4.14	192.168.1.1	11	20	
1	NaN	NaN	2.4.14	192.168.1.1	NaN	NaN	
2	100052388	NaN	2.4.14	192.168.1.1	11	20	
3	NaN	NaN	2.4.14	192.168.1.1	NaN	NaN	
4	100052388	NaN	2.4.14	192.168.1.1	11	20	
...	...	...	...	...	...	...	
914055	175	1412.898	2.4.18	192.168.0.1	NaN	NaN	
914056	1	NaN	2.4.18	192.168.0.1	1	NaN	
914057	1	NaN	2.4.18	192.168.0.1	1	NaN	
914058	NaN	NaN	2.4.18	192.168.0.1	NaN	NaN	
914059	100054410	NaN	2.4.18	192.168.0.1	3	2	
			ItemName	AppName	ElapsedTimePlaying	\	
0	Trường Học Moorim (20 Tập)	VOD			3811		
1		NaN	IPTV		NaN		
2	Trường Học Moorim (20 Tập)	VOD			NaN		
3		NaN	IPTV		NaN		
...							
914058	NaN	NaN	NaN	NaN	NaN	NaN	
914059	NaN	NaN	NaN	NaN	NaN	NaN	
[914060 rows x 43 columns]							

#	Column	Non-Null Count	Dtype
0	ItemId	760938	non-null object
1	RealTimePlaying	295537	non-null object
2	Firmware	914060	non-null object
3	DefaultGetway	914060	non-null object
4	SubMenuId	202716	non-null object
5	Folder	120220	non-null object
6	ItemName	757649	non-null object
7	AppName	914060	non-null object
8	ElapsedTimePlaying	41481	non-null object
9	Screen	792552	non-null object
10	SecondaryDNS	914060	non-null object
11	LogId	914060	non-null object
12	Mac	914060	non-null object
13	LocalType	914060	non-null object
14	SubnetMask	914060	non-null object
15	ip_wan	914060	non-null object
16	CustomerID	914060	non-null object
17	Url	219709	non-null object
18	ListOnFolder	120220	non-null object
19	Contract	914060	non-null object
20	Directors	120220	non-null object

21	SessionMainMen	914032	non-null	object
22	PrimaryDNS	914060	non-null	object
23	ChapterID	161484	non-null	object
24	Ip	914060	non-null	object
25	BoxTime	914060	non-null	object
26	PublishCountry	120220	non-null	object
27	Session	914060	non-null	object
28	SessionSubMen	213429	non-null	object
29	AppId	914060	non-null	object
30	Duration	380805	non-null	float64
31	Event	914060	non-null	object
32	DateStamp	78739	non-null	object
33	isLandingPage	571834	non-null	object
34	Key	246711	non-null	object
35	Multicast	488942	non-null	object
36	Title	692	non-null	object
37	IsPersonal	60648	non-null	object
38	IDRelated	26041	non-null	object
39	keyword	13448	non-null	object
40	Original	1319	non-null	object
41	Hit	1010	non-null	object
42	Path	396	non-null	object

DataFrame tổng hợp có 914060 dòng và 43 thuộc tính

## DOAN VU MINH THANH

DataFrame mới có 7 thuộc tính được chọn lọc. Loại trừ các dòng bị trùng lắp và thay thế các giá trị bị thiếu bằng khoảng trắng

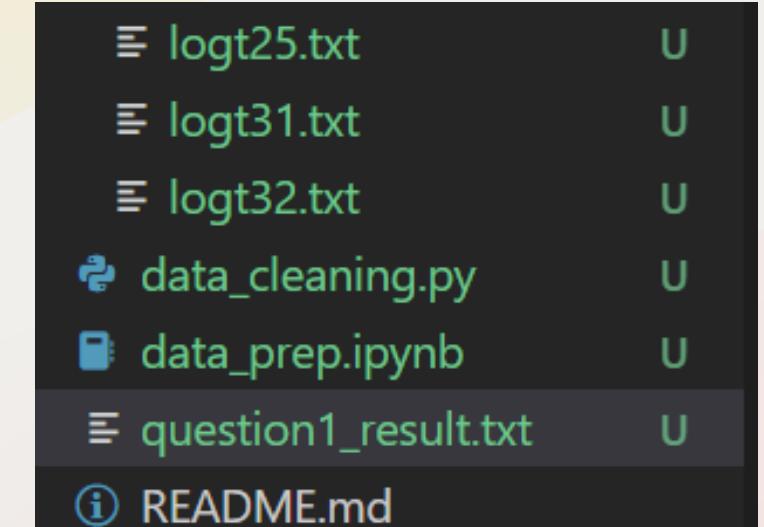
```
selected_df = combined_df[['Mac', 'SessionMainMen', 'AppName', 'LogId', 'Event', 'ItemId', 'RealTimePlaying']]  
selected_df = selected_df.fillna('')  
selected_df = selected_df.drop_duplicates()  
selected_df = selected_df.reset_index(drop=True)  
display(selected_df)
```

	Mac	SessionMainMen	AppName	LogId	Event	ItemId	RealTimePlaying
0	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:12:35:13:437	VOD	52	StopVOD	100052388	570.3
1	B046FCAC0DC1	B046FCAC0DC1:2016:02:11:01:01:56:838	IPTV	40	EnterIPTV		
2	B046FCAC0DC1	B046FCAC0DC1:2016:02:11:01:02:29:258	VOD	55	NextVOD	100052388	
3	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:04:44:59:143	IPTV	18	ChangeModule		
4	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:12:35:13:437	VOD	54	PlayVOD	100052388	
...	...	...	...	...	...	...	...
770928	B046FCB626E5	B046FCB626E5:2016:03:14:12:22:22:382	IPTV	42	StopChannel	148	8.779
770929	B046FCB626E5	B046FCB626E5:2016:03:14:12:18:04:612	VOD	57	EnterFolderVOD	4	
770930	B046FCB626E5	B046FCB626E5:2016:03:14:12:15:19:629	IPTV	451	ExitChannelList	1	6.82
770931	B046FCB626E5	B046FCB626E5:2016:03:14:12:45:57:625	IPTV	42	StopChannel	1	24.418
770932	B046FCB626E5	B046FCB626E5:2016:03:14:17:34:06:842	IPTV	42	StopChannel	1	9.986
770933 rows × 7 columns							

sau khi xử lý, DataFrame còn 770933 quan sát

lưu DataFrame dưới dạng .txt, các cột cách nhau bởi dấu tab '/t'

```
selected_df.to_csv('question1_result.txt', sep='\t', index=False)
```



kết quả lưu file log:

```
data_prep.ipynb U ●   question1_result.txt U X  
question1_result.txt  
1 Mac SessionMainMen AppName LogId Event ItemId RealTimePlaying  
2 B046FCAC0DC1 B046FCAC0DC1:2016:02:12:12:35:13:437 VOD 52 StopVOD 100052388 570.3  
3 B046FCAC0DC1 B046FCAC0DC1:2016:02:11:01:01:56:838 IPTV 40 EnterIPTV  
4 B046FCAC0DC1 B046FCAC0DC1:2016:02:11:01:02:29:258 VOD 55 NextVOD 100052388  
5 B046FCAC0DC1 B046FCAC0DC1:2016:02:12:04:44:59:143 IPTV 18 ChangeModule  
6 B046FCAC0DC1 B046FCAC0DC1:2016:02:12:12:35:13:437 VOD 54 PlayVOD 100052388  
7 B046FCAC0DC1 B046FCAC0DC1:2016:02:12:04:44:59:143 IPTV 40 EnterIPTV  
8 B046FCAC0DC1 B046FCAC0DC1:2016:02:12:12:35:13:437 VOD 55 NextVOD 100052388  
9 B046FCAC0DC1 B046FCAC0DC1:2016:02:12:12:35:13:437 VOD 52 StopVOD 100052388 3384.6  
10 B046FCAC0DC1 B046FCAC0DC1:2016:02:13:17:25:40:373 IPTV 40 EnterIPTV
```

# Câu hỏi 2

Kết hợp file user\_info.txt và tập data đã parse, phân tích hành vi đặc điểm sử dụng dịch vụ của những user này.



# Nhận xét dữ liệu đầu vào

# Hướng xử lý

1. Về file user\_info.txt:

- 2 thuộc tính: userID (MAC) và số ngày user sử dụng dịch vụ (# of days)
- MAC chưa Mac trong file log (FBOX<Mac>)

FBOXB046FCB79E0B

2. Về file log từ câu 1:

- SessionMainMen: đang có format <MAC> <thời gian>
- Event: 1 số sự kiện có kèm ItemId (<sự kiện> <Item ID>)
- 1 lượng lớn dữ liệu bị thiếu ở cột RealTimePlaying

B046FCAC0DC1:2016:02:11:01:01:56:838

StopVOD

RealTimePlaying 289015 non-null

770933 entries

- định dạng lại cột SessionMainMen, Event
- xử lý missing values
- đọc file user\_info.txt, xử lý MAC
- Pivot table / groupby
- Visualization - nhận xét

# Data cleaning

💡 Click here to ask Blackbox to help you code faster  
display(selected\_df[selected\_df['SessionMainMen'].isnull() == True])

	Mac	SessionMainMen	AppName	LogId	Event	ItemId	RealTimePlaying
99808	B046FCB58B44	NaN	IPTV	41	StartChannel	11	NaN
99865	B046FCB58B44	NaN	IPTV	40	EnterIPTV	NaN	NaN
350118	B046FCB467E9	NaN	IPTV	42	StopChannel	181	2256.61
350120	B046FCB467E9	NaN	IPTV	42	StopChannel	95	281.477
350171	B046FCB467E9	NaN	IPTV	42	StopChannel	161	7.001
350180	B046FCB467E9	NaN	IPTV	41	StartChannel	55	NaN
350228	B046FCB467E9	NaN	IPTV	42	StopChannel	55	19.32
350241	B046FCB467E9	NaN	IPTV	45	ShowChannelList	161	NaN
350242	B046FCB467E9	NaN	IPTV	42	StopChannel	55	1047.052
350243	B046FCB467E9	NaN	IPTV	41	StartChannel	95	NaN
350244	B046FCB467E9	NaN	IPTV	42	StopChannel	55	6.062
350245	B046FCB467E9	NaN	IPTV	45	ShowChannelList	95	NaN
350302	B046FCB467E9	NaN	IPTV	40	EnterIPTV	undefined	NaN
350303	B046FCB467E9	NaN	IPTV	41	StartChannel	161	NaN
350305	B046FCB467E9	NaN	IPTV	41	StartChannel	181	NaN
350307	B046FCB467E9	NaN	IPTV	42	StopChannel	55	1091.276
350309	B046FCB467E9	NaN	IPTV	42	StopChannel	161	4.162
350350	B046FCB467E9	NaN	IPTV	42	StopChannel	181	355.325
498495	B046FCA984F2	NaN	IPTV	40	EnterIPTV	NaN	NaN
499134	B046FCA984F2	NaN	IPTV	41	StartChannel	1	NaN
738388	B046FCA86D3A	NaN	IPTV	41	StartChannel	3	NaN
738482	B046FCA86D3A	NaN	IPTV	40	EnterIPTV	NaN	NaN

Cột SessionMainMen có 22 / 770933 quan sát bị thiếu dữ liệu, vì số lượng quan sát thiếu dữ liệu trên cột này nhỏ nên sẽ loại bỏ các quan sát này

# Data cleaning

trên cột SessionMainMen, loại bỏ userID được đổi chiếu từ cột Mac

```
#remove Mac in sessionmainmenu  
selected_df['SessionMainMen'] = selected_df.apply(lambda row: row['SessionMainMen'].replace(str(row['Mac'])+':'), ''), axis=1)
```

tạo 2 cột dữ liệu mới là Date (dd-mm-yyyy) và Time(hh:mm:ss)

```
#format the date time  
selected_df['Date'] = pd.to_datetime(selected_df['SessionMainMen'], format='%Y:%m:%d:%H:%M:%S:%f').dt.strftime('%d-%m-%Y')  
selected_df['Time'] = pd.to_datetime(selected_df['SessionMainMen'], format='%Y:%m:%d:%H:%M:%S:%f').dt.strftime('%H:%M:%S')
```

SessionMainMen
B046FCAC0DC1:2016:02:12:12:35:13:437
B046FCAC0DC1:2016:02:11:01:01:56:838

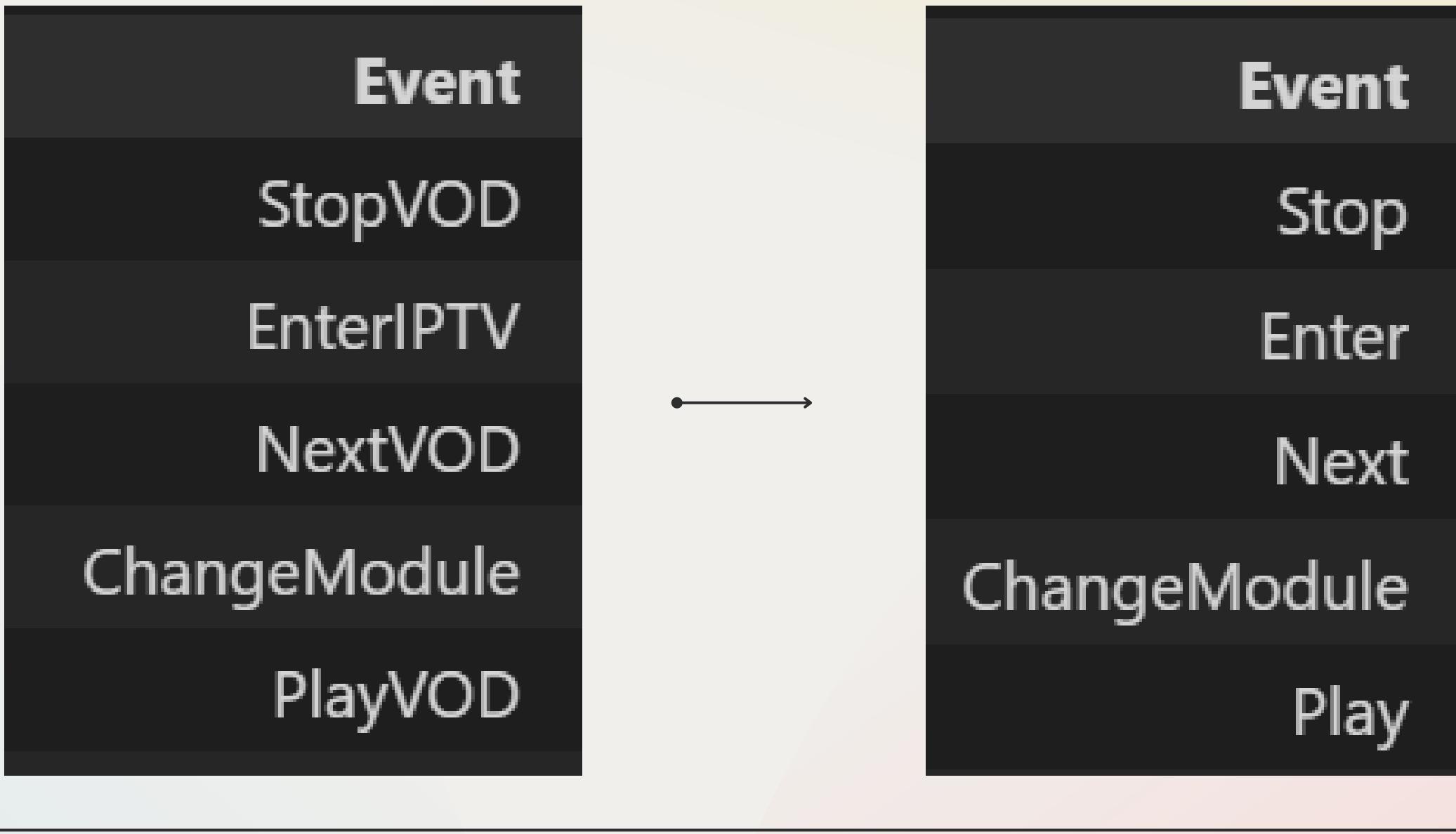
sau xử lý

Date	Time
12-02-2016	12:35:13
11-02-2016	01:01:56

# Data cleaning

trên cột Event, loại bỏ tên app được đổi chiếu ở cột AppName

```
selected_df['Event'] = selected_df.apply(lambda row: row['Event'].replace(row['AppName'], ''), axis=1)
```



# Data cleaning

vì số lượng missing values trên 2 cột ItemID và RealTimePlaying lớn (hơn 100000), nên không thể loại bỏ các quan sát có missing values. Đối với ItemID, dữ liệu rỗng được thay thế bằng từ 'undefined', còn ở RealTimePlaying sẽ thay thế bằng giá trị 0 (không xem).

```
selected_df['ItemId'] = selected_df['ItemId'].fillna('undefined')
```

```
selected_df['RealTimePlaying'] = selected_df['RealTimePlaying'].fillna(0)
```

ItemId	RealTimePlaying
100052388	570.3
NaN	NaN
100052388	NaN
NaN	NaN
100052388	NaN

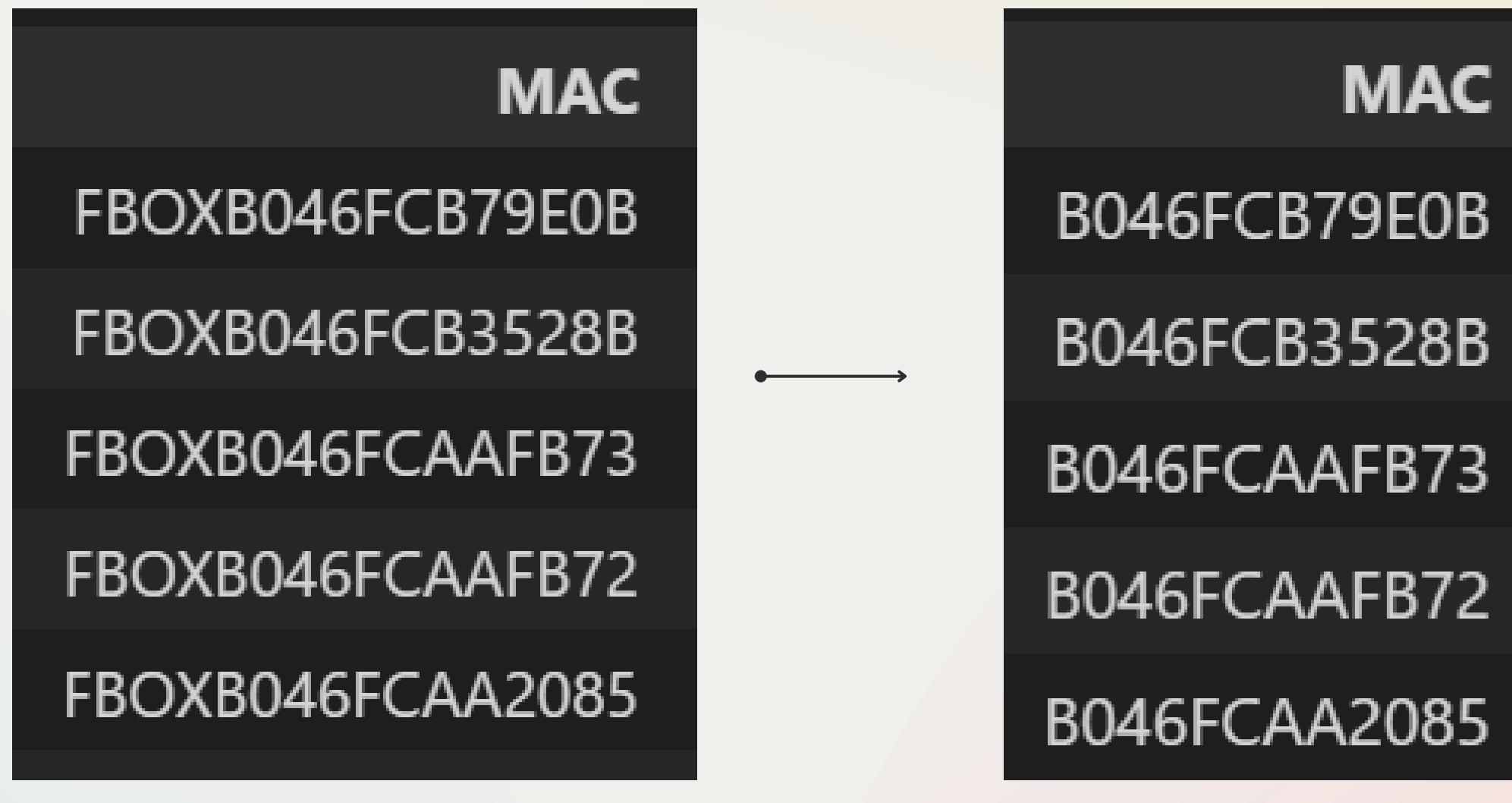


ItemId	RealTimePlaying
100052388	570.3
undefined	0
100052388	0
undefined	0
100052388	0

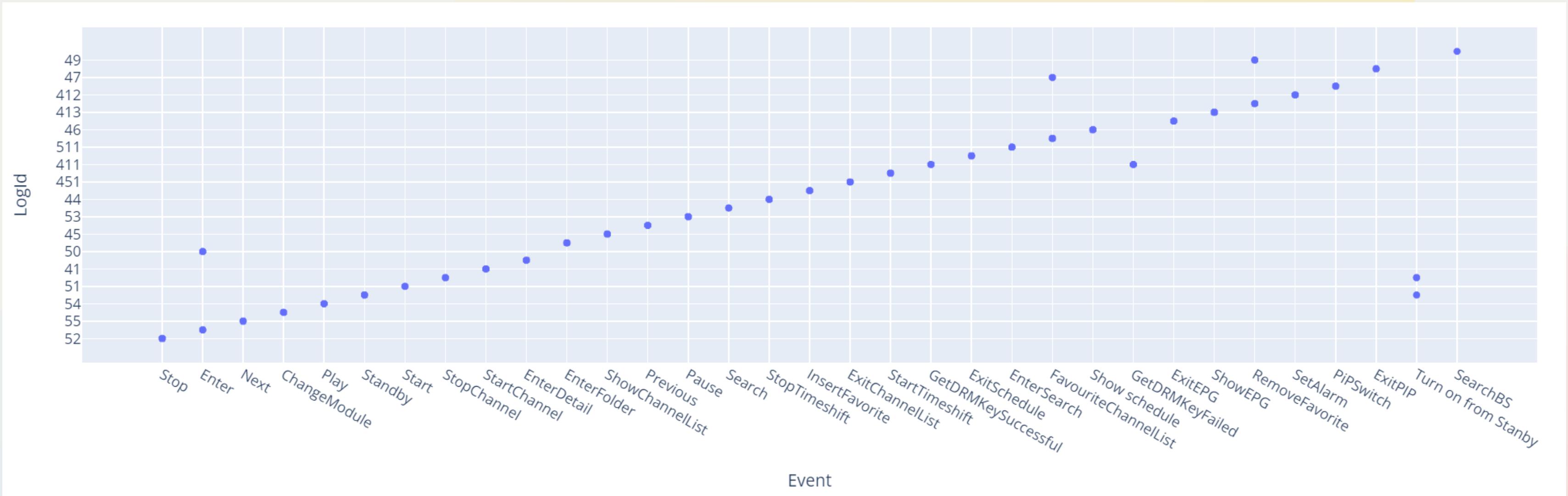
# Data cleaning

sau khi đọc file user\_info.txt và chuyển đổi thành DataFrame, cột MAC được định dạng lại bằng cách loại bỏ từ 'FBOX' ở đầu chuỗi

```
user_info['MAC'] = user_info['MAC'].str.replace(r'FBOX', '', regex=True)
```

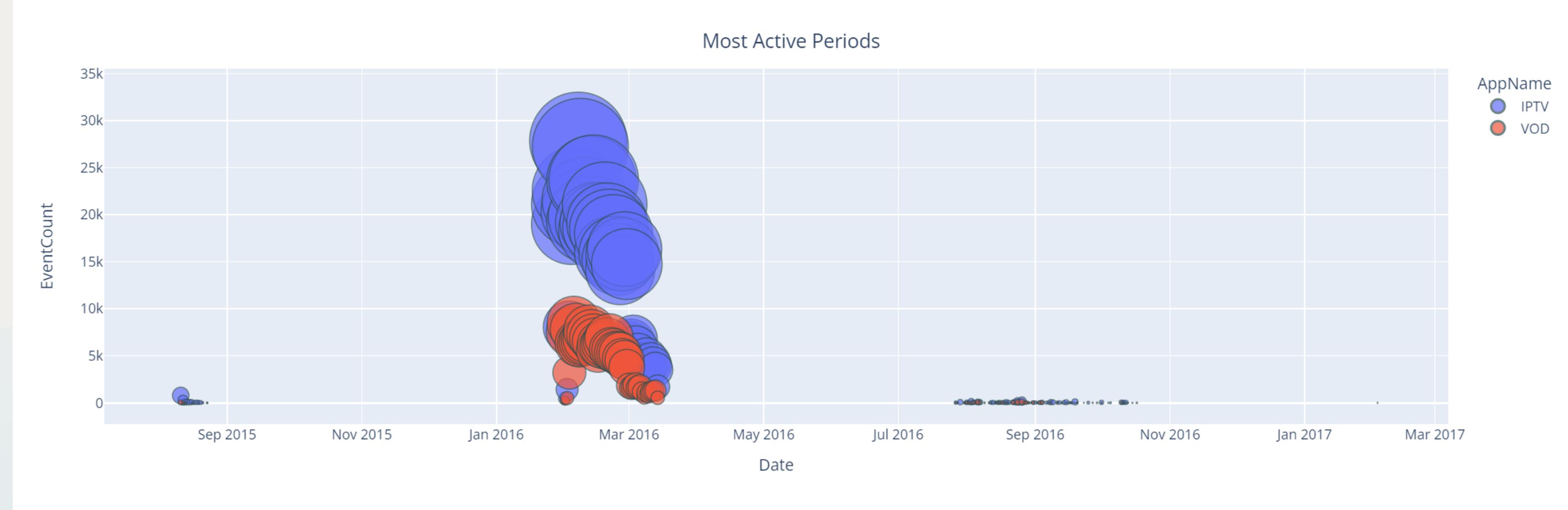


## DOAN VỤ MINH THANH



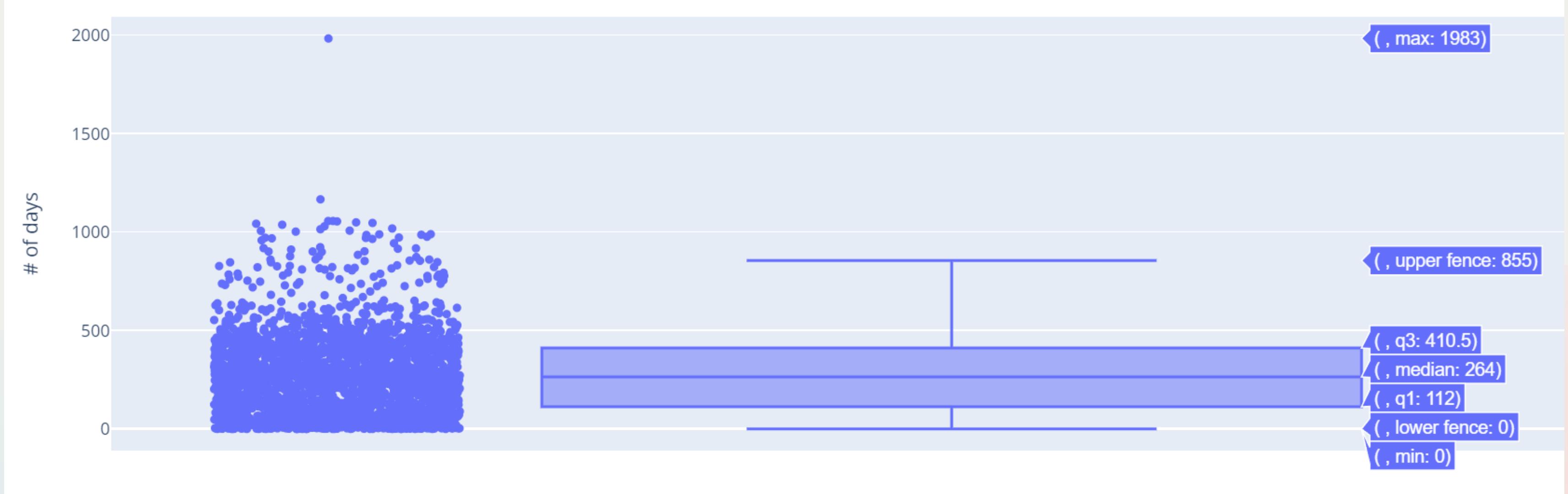
Kiểm tra tính tương quan của 2 biến Event và LogId cho thấy 2 biến có độ tương đồng cao, hầu như mỗi sự kiện sẽ có 1 logID riêng. Vì thế, ta sẽ giữ lại biến Event để tiếp tục phân tích vì biến Event có ý nghĩa hơn biến LogId.

## DOAN VU MINH THANH



- Dữ liệu được thu thập từ tháng 8/2015 đến tháng 3/2017.
- **Cao điểm** sử dụng dịch vụ kéo dài từ cuối tháng **1/2016** đến cuối tháng **3/2016**, cụ thể app **IPTV** có lượt sử dụng, **tương tác cao hơn** so với app **VOD** (**IPTV: 10k-33k lượt tương tác, VOD: 500-9k lượt tương tác**)
- Ngoài ra, còn 2 khoảng thời gian khác (tháng 8/2015 và từ cuối tháng 7 - đầu tháng 10/2016) ghi nhận có sự giao dịch dịch vụ nhưng với tần số thấp, không sôi động.

## DOAN VU MINH THANH



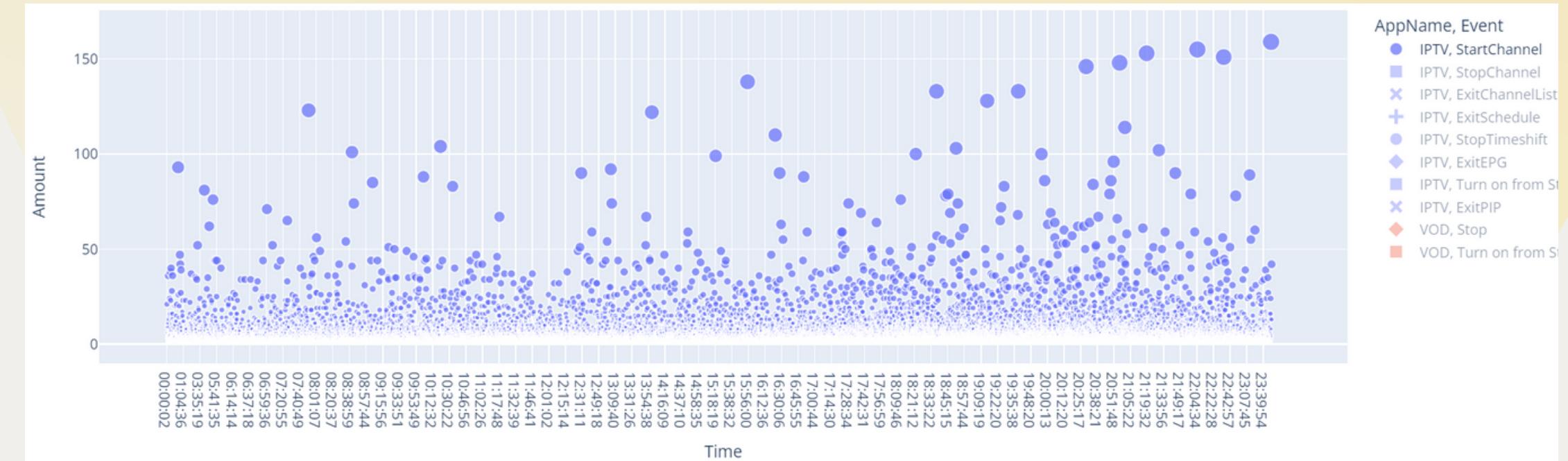
- Biểu đồ trên biểu thị độ phân tán của thời gian khách hàng sử dụng dịch vụ.
- Khách hàng thường sử dụng gói dịch vụ **ngắn hạn**, từ 112 ngày đến 410 ngày, tương đương từ **3 đến 12 tháng**. Ngoài ra 1 phần nhỏ khách hàng sử dụng dịch vụ lâu hơn, kéo dài đến hơn 2 năm, và thời gian khách hàng gắn bó với dịch vụ lâu nhất lên tới 5 năm.

## DOAN VỤ MINH THANH

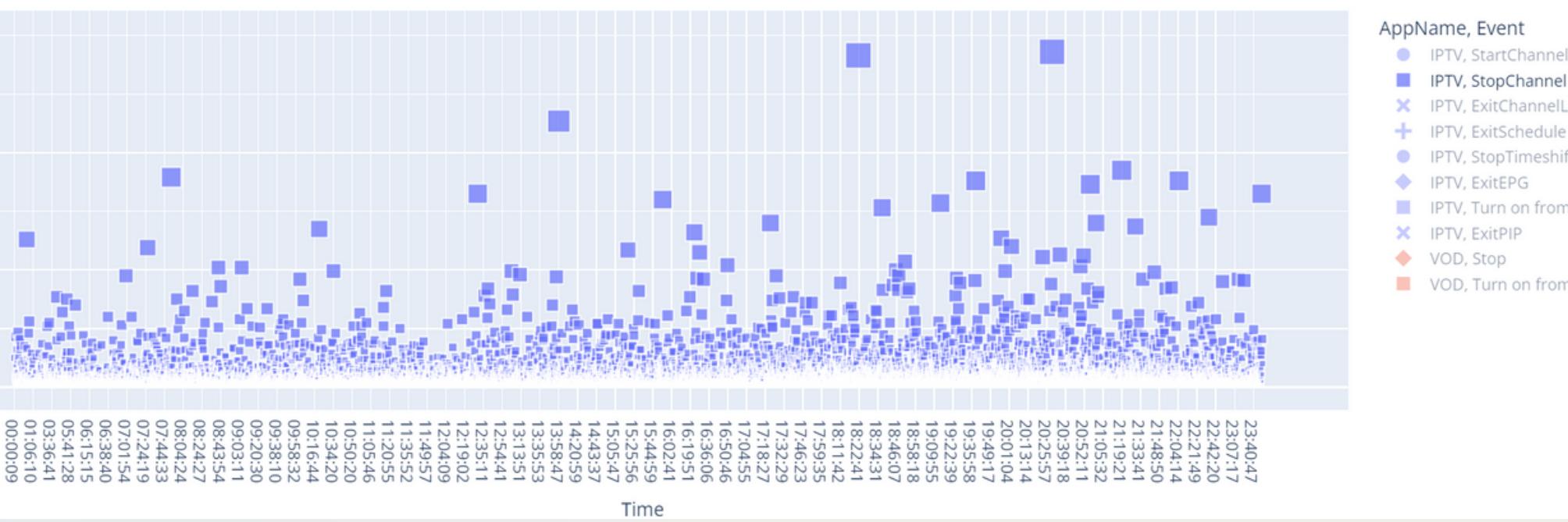
Event	
ChangeModule	0.000000
Enter	0.000000
EnterDetail	0.000000
EnterFolder	0.000000
EnterSearch	0.000000
ExitChannelList	34.321316
ExitEPG	107.539640
ExitPIP	417.997182
ExitSchedule	17.342581
FavouriteChannelList	0.000000
GetDRMKeyFailed	0.000000
GetDRMKeySuccessful	0.000000
InsertFavorite	0.000000
Next	0.000000
Pause	0.000000
PiPSwitch	0.000000
Play	0.000000
Previous	0.000000
RemoveFavorite	0.000000
Search	0.000000
SearchBS	0.000000
SetAlarm	0.000000
Show schedule	0.000000
ShowChannelList	0.000000
ShowEPG	0.000000
Standby	0.000000
Start	0.000000
StartChannel	0.067342
StartTimeshift	0.000000
Stop	1254.898539
StopChannel	1257.232046
StopTimeshift	3100.114908
Turn on from Stanby	20.892000

Khi thống kê thời gian thực trung bình sử dụng dịch vụ của các sự kiện, có thể thấy chỉ có các sự kiện Exit, Stop, Turn on from Standby và StartChannel có ghi nhận thời gian. Ta sẽ chú trọng vào các sự kiện này.

## DOAN VU MINH THANH



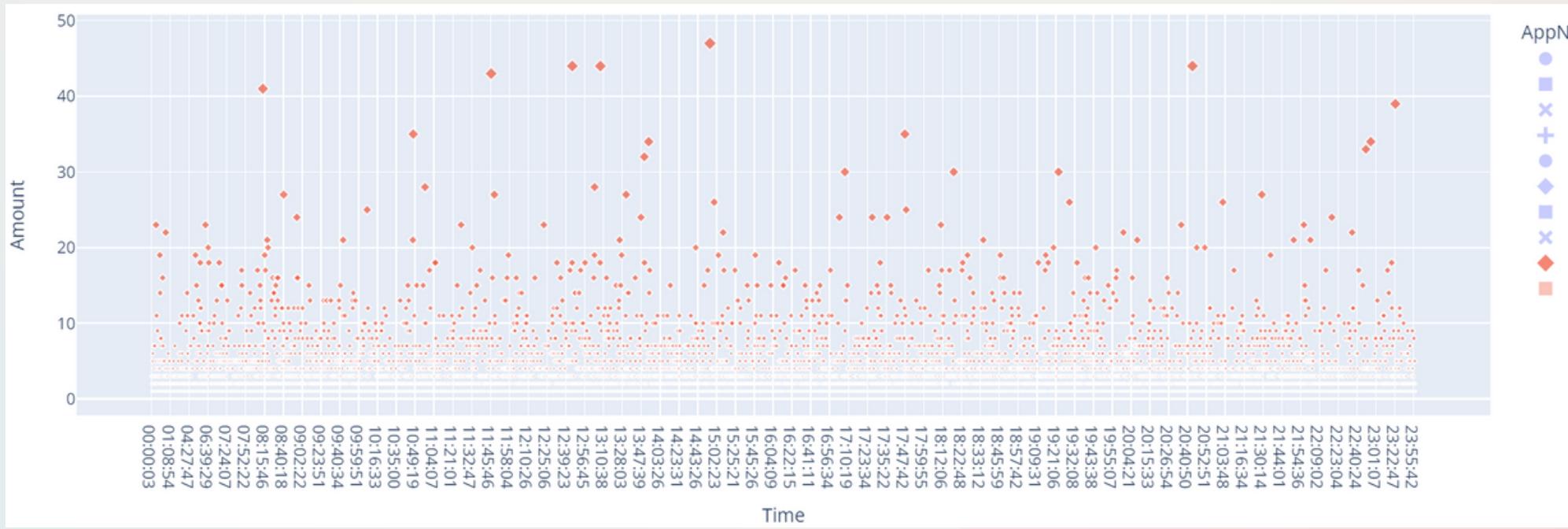
các hoạt động tương tác dịch vụ chủ yếu là Start Channel, StopChannel và Stop. Tuy nhiên, app IPTV không có sự kiện STOP, trong khi app VOD có, và ta giả định rằng STOP là hoạt động thoát khỏi app, ngưng kết nối; như vậy, app VOD không được ưa chuông bằng app IPTV, góp phần cao đến việc hủy hợp đồng hệ thống.



**AppName, Event**

- IPTV, StartChannel
- IPTV, StopChannel
- IPTV, ExitChannelList
- IPTV, ExitSchedule
- IPTV, StopTimeshift
- IPTV, ExitEPG
- IPTV, Turn on from Schedule
- IPTV, ExitPIP
- VOD, Stop
- VOD, Turn on from Schedule

khách hàng sử dụng và tương tác với dịch vụ cả ngày.



**AppName, Event**

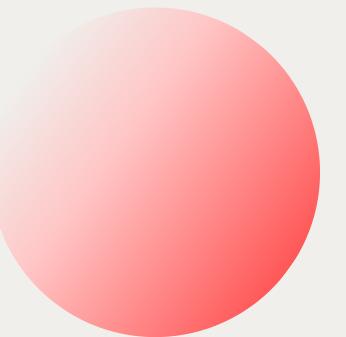
- IPTV, StartChannel
- IPTV, StopChannel
- IPTV, ExitChannelList
- IPTV, ExitSchedule
- IPTV, StopTimeshift
- IPTV, ExitEPG
- IPTV, Turn on from Schedule
- IPTV, ExitPIP
- VOD, Stop
- VOD, Turn on from Schedule

	<b>AppName</b>	<b>Average_Time</b>
0	IPTV	754.427594
1	VOD	1236.485936

Khách hàng dành nhiều thời gian dùng app VOD hơn app IPTV

DOAN VU MINH THANH

# Thanks for your attention



source code: GitHub



0346365024



doanvuminhthanh2404@gmail.com