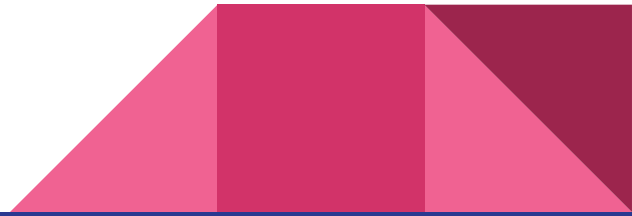
Game Streamer Analysis

Sanjoy Kumar
Data Scientist & Analyst

Outline

1. Observing and Preprocessing
2. Finding Things
3. Summary
4. Next Steps



Observation and Preprocessing

Feature include Country, Gender, Game, Total Follower, Hours, PaidStarPerWatchedHour and the other 121 Facial Estimation columns

58 games: '8 Ball Pool', 'Age of Empires', 'Apex Legends', 'Arena of Valor', "Assassin's Creed Odyssey", 'Audition Online'.....

4 countries: Indonesia(ID), Vietnam(VN), Philippines(PH)

2 genders: Female, Male

Observation and Preprocessing

There are some games name with different name or typo in the csv, such as:

```
'Garena Liên Quân Mobile'
'Liên Quân Mobile'
'Arena of Valor'
```

```
'Audition'
'Audition Online'
```

```
'Agge of Empires'
'Age of Empires'
```

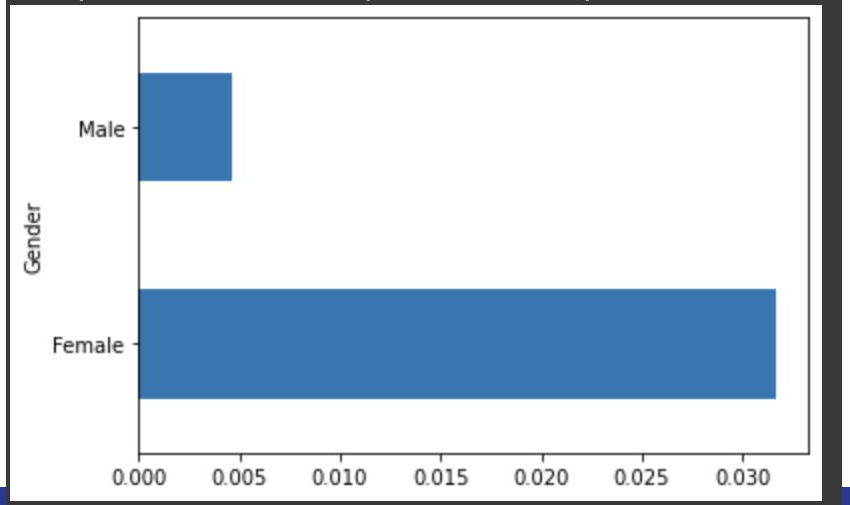
```
'Call of Duty: Mobile VN'
'Call of Duty: Mobile'
```

Finding Things - Comparison with features

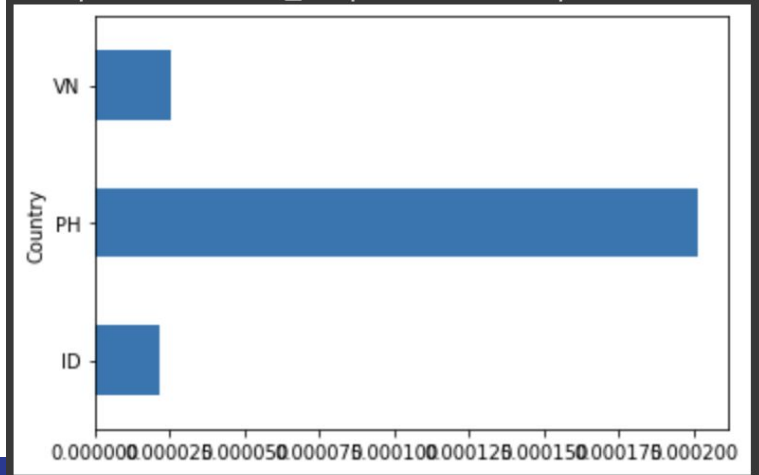
As shown in the graph below, (1) Female get more stars in general

(2) PH get more stars among 3 countries

```
Gender
Female    0.031696
Male      0.004646
Name: PaidStarPerWatchedHour, dtype: float64
<matplotlib.axes._subplots.AxesSubplot at 0x7fc04
```




```
Country
ID    0.000022
PH    0.000202
VN    0.000025
Name: PaidStarPerWatchedHour, dtype: float64
<matplotlib.axes._subplots.AxesSubplot at 0x7fc
```



Finding Things - Correlation in general

To find which feature may higly affect the result of “PaidStarPerWatchedHour”, I do correlation matrix to check the relationship between these features.

After sorting by the abolute value of correlation as shown in the graph below, it’s hard to select features because the value of correlation are too small.

Personal_Values_Facet_Cont_Hedonism	0.142095
Role_Philanthropist	-0.140508
Personal_Values_Facet_Cont_ConformityRules	-0.140314
Character_Facet_Cont_AchievementStriving	-0.134136
Role_Manager	-0.130683
Character_Cont_Conscientiousness	-0.126654
Role_Marshal	-0.116596

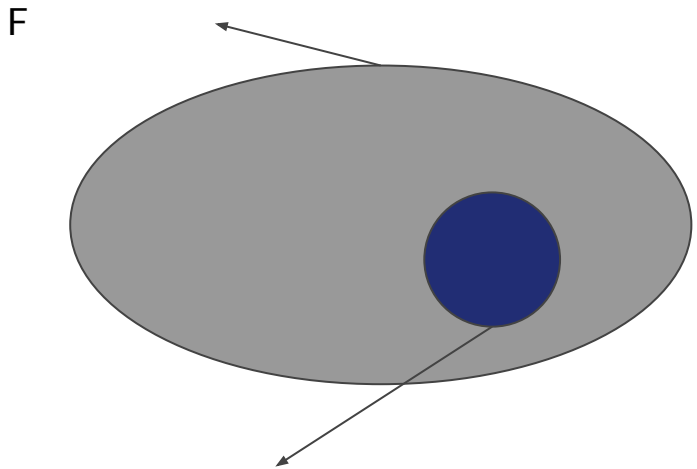
Finding Things - Correlation in specific class

The reason why the correlation value in general is small is that characteristic feature may vary from different data. Therefore, we have to divide data into different classes to observe the relationship of these features.

For example, I select data grouped in fixed value (VN, PUBG, Male). As shown in the graph below, the correlation is higher than in general.

Character_Cont_Neuroticism	-0.365415
Self_Esteem_Cont_SEDiscrepancyOriginality	0.303852
Temperament_Stable	0.284451
Personal_Values_Facet_Cont_Achievement	0.263908
Personal_Values_Facet_Cont_SelfDirectionAction	0.262849
Temperament_Phlegmatic	0.242283

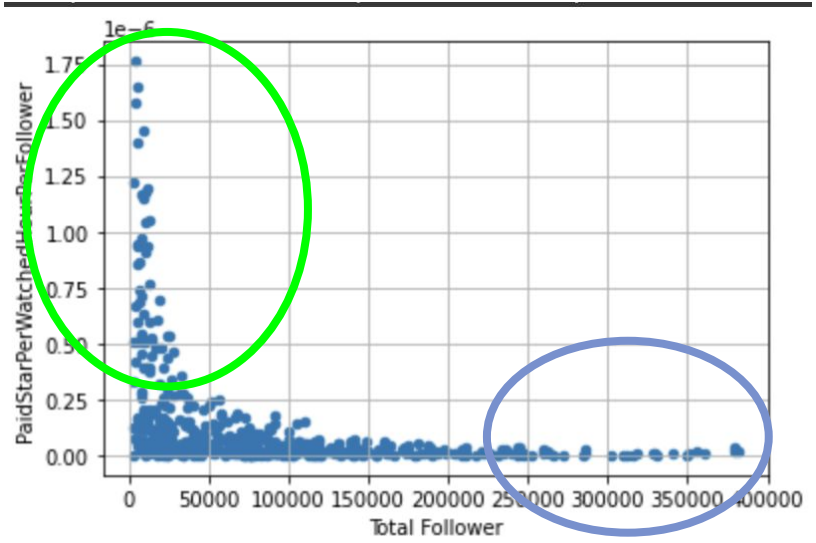
Finding Things



In the gaming industry, there's a saying that "90% of the revenue is generated by 10% of the players." As shown in the left, followers contains those who donate stars, and the range of the blue circle may decide whether this phenomenom ocurrs in game streaming.

Although We can't check whether this phenomenom occurs in the game streaming as well if only given this dataset, it still implies that the ratio of followers who donate stars is a significant feature to classify game streamers.

Finding Things



To group those gamers, I add another columns called “PaidStarPerWatchedHourPerFollower”.

As shown in the graph, streamers in the red circle means that they have many followers but the donated stars don’t grow accordingly by the the amount of followers. Streamers in the green circle means that they although their followers are not as many as the red, they have high ratio between the donated followers and the free followers. It indicate that the green circle have more potentials.

Finding Things

Role_Philanthropist	-0.282005
Role_Manager	-0.254636
Character_Cont_Conscientiousness	-0.245829
Character_Facet_Cont_AchievementStriving	-0.245191
Role_Marshal	-0.237801
Personal_Values_Facet_Cont_Hedonism	0.236380
Character_Cont_Openness	-0.230269

To find out what matters in those potential streamers, we can take a look at the correlation matrix.

As shown in the graph, I sort the value by abolution. In the word cloud, tt indicate that those feature may be more highlt-related with the start they get paid.

Summary

1. Data preprocessing is conducted to drop invalid values, merge equivalent values, and exclude outlier by IQR
2. Female streamers and streamers in Philippines have extraordinary advantage of getting stars.
3. After clustering streamers into different classes, the correlation value will be more obvious.
4. It's interesting that streamers who get more PaidStarPerWatchedHourPerFollower tend to be less philanthropic, less conscientiousness. The results seems to be be different from what we think a good people shold be like, and it indicate that maybe people prefer to donate stars to these features.

Next Step

1. Streamers should be grouped into more classification (), and add some tags of these classification to the data as new features(column).

2. Data should be divided into several subset to analysis. For example, streamers playing hardcore games like PUBG may be quite different from other streamers playing Audition Online. On the other hand, streamers playing Leauge of Legend may share same feature with those playing Arena of Valor.

3. To fulfill these steps, I can build machine learning medels (tree-based or neural network) to utilize some clustering algorithm like K-means instead of using correlation matrix and other simple statistical methods to analyze data.

Thanks for sharing knowledge :)