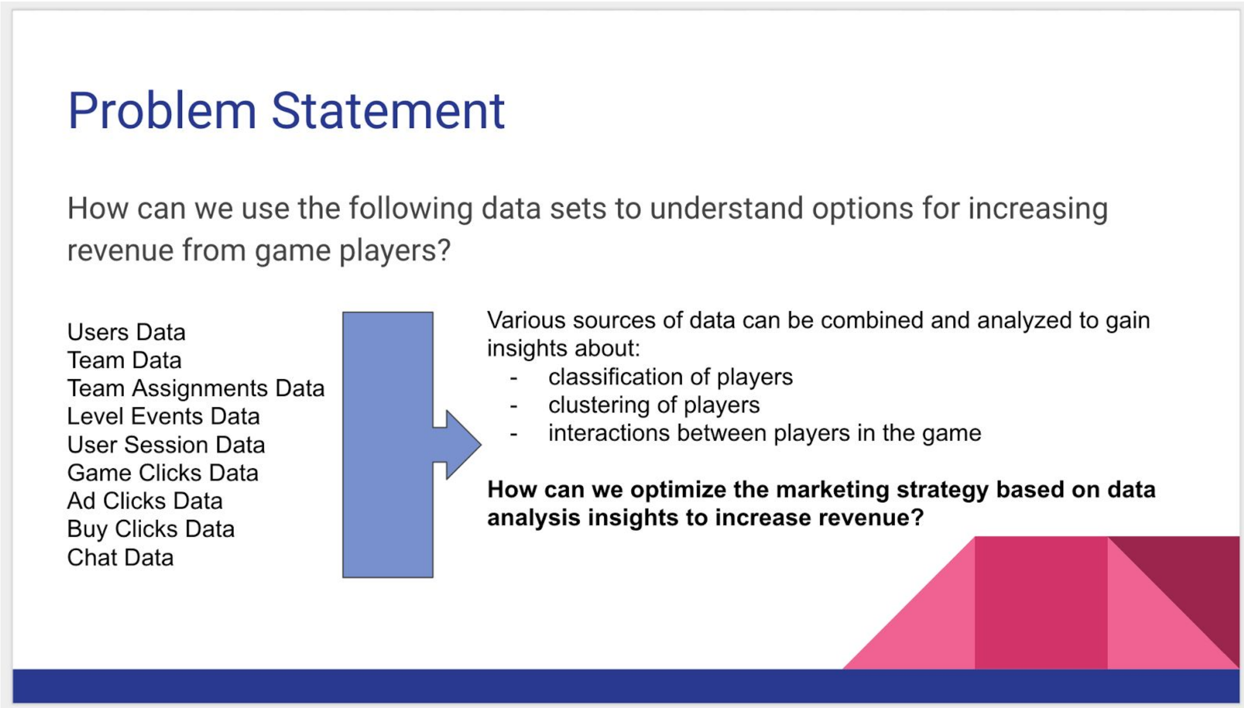
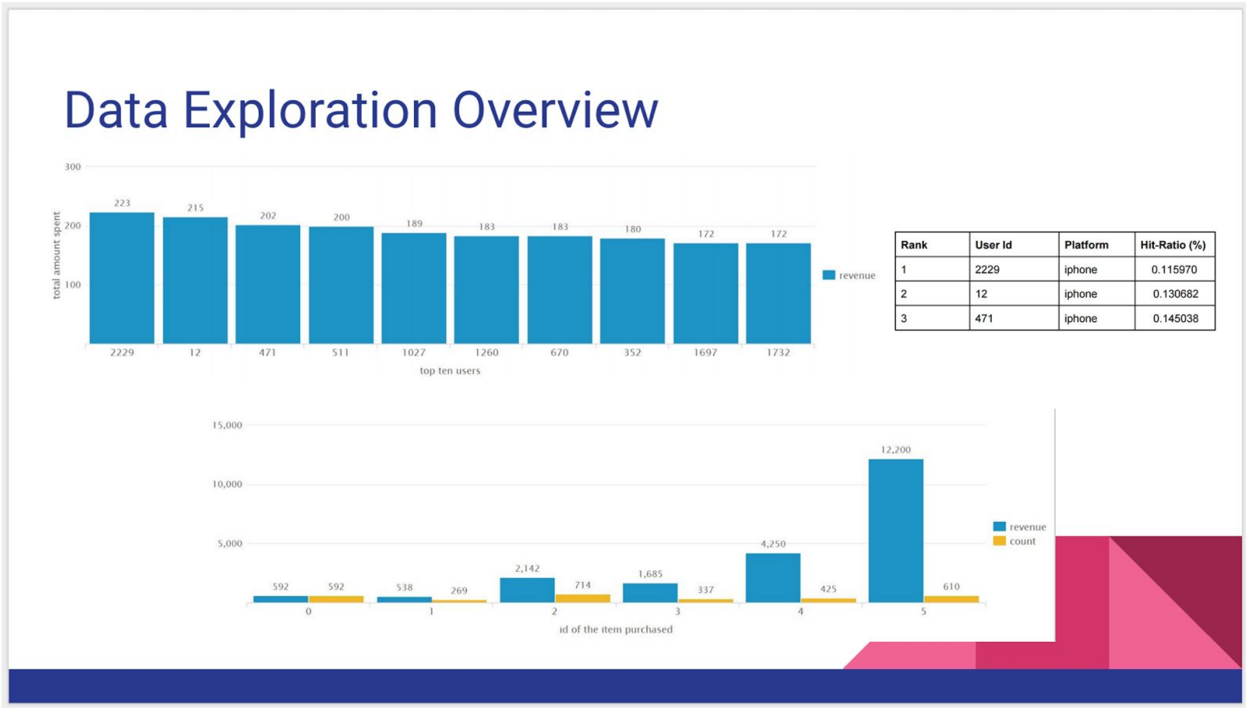
How can we increase revenue from Catch the Pink Flamingo?



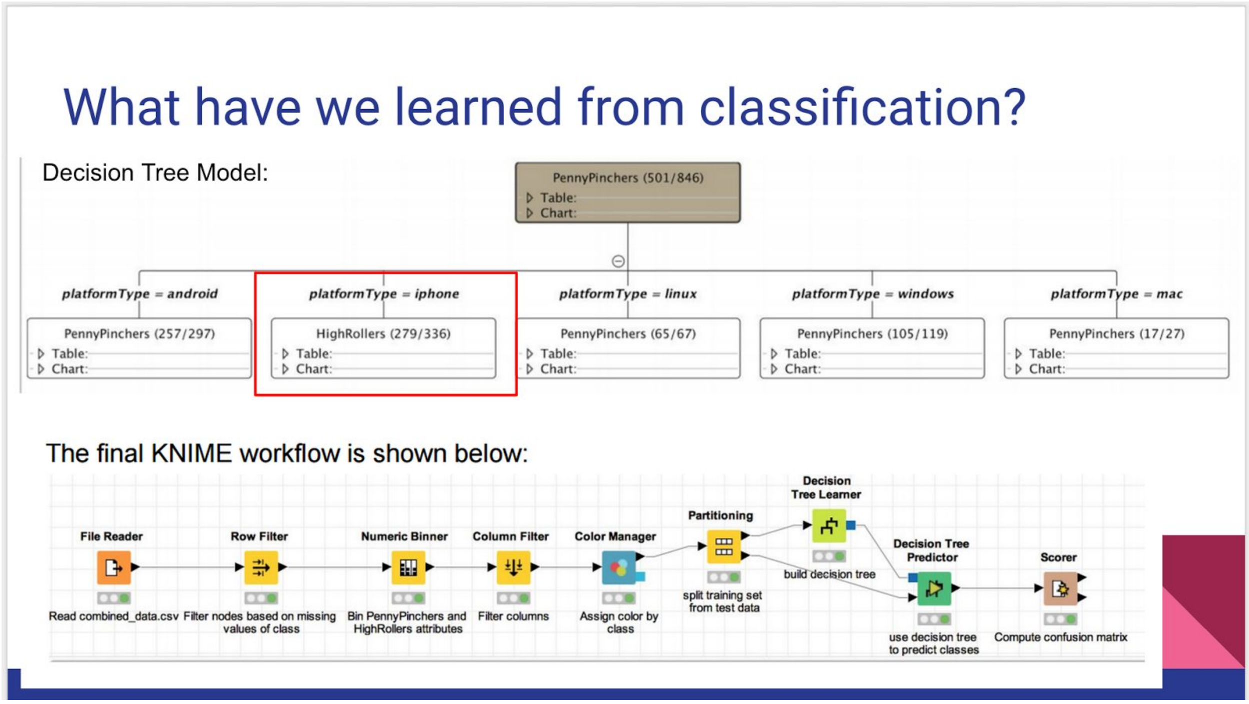
How can we use the following data sets to understand options for increasing revenue from game players?

Users data describes data about each user playing the Flamingo app game including when the user started playing the game, their twitter handle, date of birth and country where they live.
Team data describes data about each team playing the game including when the team was created, a measure of strength of the team and the current level of the team.
Team assignment data describes team assignments of players in the game and time when a user joins a team.
Level events data describes team levels in the game, when a team starts or completes a level.
User session data let us know when a user starts or stops playing the game and what platform she used.
Game clicks data describes whether the player clicked a flamingo or not.
Ad Clicks data lets us know what player clicked the ad in the game and ad category.
Buy Clicks data lets us know what player make a purchase in the app and item's price.
In addition, Chat data lets us know about chatting patterns among players and teams in the game.

The various sources of data can be combined and analyzed to gain insights about classification of users, clustering of players and users interactions to predict their buying behaviour and adjust marketing strategy, the number and price of ads shown and to increase revenue. So quality of data matters. Collecting user data, in game click data, purchasing data and in-game chat data will enable Eglence to identify new revenue opportunity with thorough data analysis.



Data exploration can reveal patterns in the data. For example, we can identify the top ten spenders in the game and plot revenue as a histogram to convey our findings (top left image on the slide). In addition, we can filter the data to find that the top 3 buyers use Iphone platform which can help us to target advertisement and higher price ads to them (top right image).
We can analyze the revenue in the game from user's purchases. There 6 different items for purchase and the total revenue is $21,407. We plot a histogram of revenue and number of purchases for each 6 items (bottom image). Grouping the items by count and revenue we see that the item 2 was purchased the most with 714 purchases but only generated revenue of $2,142, while item 5 revenue was $12,200 with 610 purchases.



We used KNIME to identify big spenders. Players who spend $5 or less on a purchase are classified “PennyPinchers” and players who spend more than $5 are classified “HighRollers”. The training dataset was used to build a decision tree model shown on the slide. The trained model is then applied on a test dataset. The overall accuracy of the model is 88.5% which is very good. Our model, however, has 27 false negative predictions and 38 false positive predictions out of 565.

From classification, we learned that the main attribute used to make classification decision is platformType. If the player uses Iphone, then the player is predicted as “HighRollers”, while for the platform types android, linux, windows and mac, the user is predicted to be “PennyPinchers” class.

What have we learned from clustering?

K-means clustering with Spark MLlib:

Cluster #	Cluster Center (revenue, average hit rate, total clicks per user)
1	array([30.20, 0.11365, 293.292])
2	array([51.78, 0.115574, 732.62])
3	array([33.92, 0.11272, 1407.02])
4	array([39.40, 0.11211, 2737.48])

Users are clustered into 4 distinct clusters based on which we can utilize different marketing strategy and increase revenue.

Assume you have 2 minutes to present what you perceive to be the most important or remarkable points from your clustering analysis.

For K-means clustering analysis, we chose the following 3 attributes: revenue, averageHitRate, and totalClicks. We chose to cluster players into 4 clusters.

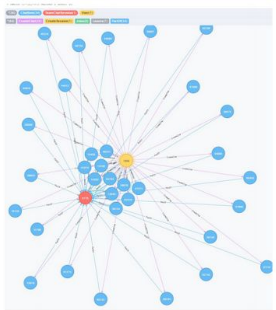
These clusters can be differentiated from each other as follows:
Cluster 1 is different from the others in that it has the lowest average revenue and the lowest average total click count per user. It clusters users who spend the least amount of money in the game and click the least while playing the game.
Cluster 2 is different from the others in that it has the highest average revenue and the highest average hit rate per user. It clusters users who generate the most revenue and are the most accurate players in the game.
Cluster 3 is different from the others in that it does not have any extreme values of the cluster centers. All the attributes revenue, average hit rate or total clicks per user has the average values for the center. It has lower total clicks than Cluster 4 but higher total clicks than Cluster 1 and 2. Similarly, its center revenue value is lower than Cluster 2 and Cluster 4 but higher than Cluster 1. Cluster 4 is different from the others in that it has the highest total clicks per user. It has the lowest average hit rate. Also, it has the second highest average revenue after Cluster 2. It clusters users who click the most in the game and also are the least accurate players.

Based on this analysis, we can increase ad prices shown to Cluster 2 and increase number of ads shown to Cluster 4 to increase the revenue.

From our chat graph analysis, what further exploration should we undertake?

Chattiest Users

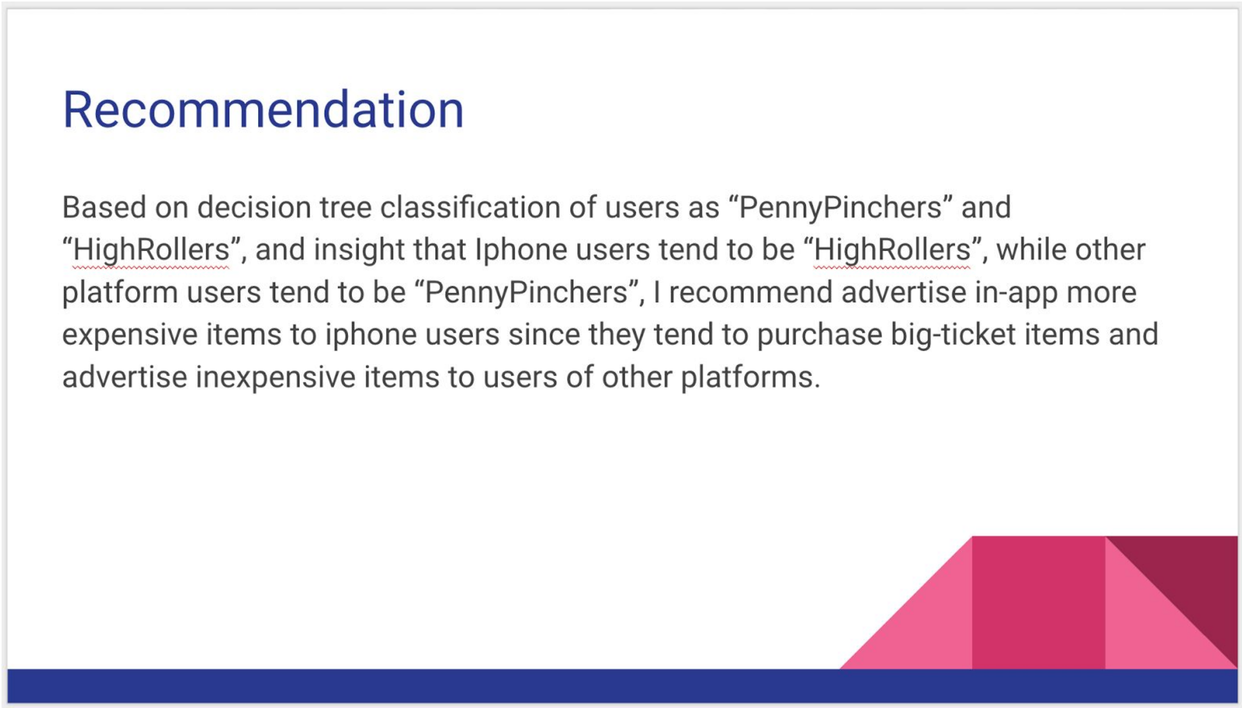
Users	Number of Chats
394	115
2067	111
1087 or 209	109



user	clusteringcoeff
461	1
668	1
516	0.9523809523809523
209	0.9523809523809523
394	0.9166666666666666
999	0.8194444444444444
554	0.8095238095238095
2067	0.7678571428571429
1627	0.7678571428571429
1087	0.7666666666666667

The graph model for chat data contains data about chatting patterns among users and teams in the game. With Neo4j, we can find the longest conversation chain and identify users who participated in it. We can identify the chattiest users and the chattiest teams.

We analyzed how active groups of users are for the top 10 chattiest users to find dense neighborhood of players who interact among the neighborhood. We compute a clustering coefficient shown in the table on the right. Based on this analysis, user 461 and 668 have a clustering coefficient of 1 meaning every neighbor interacts with every other neighbor in this group.



I recommend to advertise in-app more expensive items to iphone users since they tend to purchase big-ticket items and advertise inexpensive items to users of other platforms. The decision tree model accuracy was high and the classification result is easy to interpret as opposed to K-means modeling when we have to chose the number of clusters (was 4 a good number of clusters and we could have chosen other attributes for clustering) or chat data which is harder to interpret and needs a real-time analysis.