Data Exploration : Data Set Overview

The table below lists each of the files available for analysis with a short description of what is found in each one.

FILE NAME	ERD TABLE	DESCRIPTION	FIELDS	
ad-clicks.csv	AdClicks	A line is added to this file when a player clicks on an advertisement in the Flamingo app.	**timestamp**	when the click occured
			txId	unique id for the click (within ad-clicks.log) for the click
			userSessionid	id of user session for user who made click
			teamid	current team id of user who made the click
			userid	user id of user who made the click
			adId	id of the ad licked on
			adCategory	category/type of ad clicked on
buy-clicks.csv	InAppPurchases	A line is added when a player makes an in-app purchase on Flamingo app	**timestamp**	when the click occured
			txId	unique id for the click (within ad-clicks.log)
			userSessionid	id of user session for user who made click
			team	current team id of user who made the purchase
			userid	user id of user who made the click
			buyId	id of the item purchased
			price	price of the item purchased
users.csv	User	File contains a line for each user playing the game.	**timestamp**	when the click occured
			userId	user id of user who made the click
			nick	nickname chosen by the user

			twitter	twitter handle of the user
			dob	date of birth of the user
			country	2-letter country code where the user lives
team.csv	Team	File contains a line for each team terminated in the game.	**teamId**	id of the team
			name	name of the team
			teamCreationTime	timestamp when team was created
			teamEndTime	timestamp when last member of the team
			strength	measure of team strength roughly corresponding to the success of a time
			currentLevel	current level of a team
team-assignments.csv	TeamAssignment	A line is added each time a user joins a team. A user can be in at most a single team at a time.	**timestamp**	when the user joined the team
			team	id of the teamuser
			userId	id of the user
			assignmentId	unique id for this assignment
level-events.csv	LevelEvent	A line is added each time a team starts or finishes a level in the game.	**timestamp**	when the click occured
			eventId	unique id for the event
			teamid	id of the team
			teamLevel	level started or completed
			eventType	type of event (start or end)

<table>
<tr>
<td>user-session.csv</td>
<td>User_Sessions</td>
<td>Each line describes a user session, which denotes when a user starts and stops playing the game.
When a team goes to next game level, the session is ended for each user in the team and a new one is started.</td>
<td>
<table>
<tr><td><b>timestamp</b></td><td>when the click occured</td></tr>
<tr><td><b>userSessionid</b></td><td>unique id for the session</td></tr>
<tr><td><b>userId</b></td><td>current user’s ID</td></tr>
<tr><td><b>teamid</b></td><td>current user’s team</td></tr>
<tr><td><b>assignmentId</b></td><td>team assignment id for the user to the team</td></tr>
<tr><td><b>sessionType</b></td><td>whether the event is the start or end of a session</td></tr>
<tr><td><b>teamLevel</b></td><td>level of team during the session</td></tr>
<tr><td><b>platformType</b></td><td>type of platform of the user during the session</td></tr>
</table>
</td>
</tr>
<tr>
<td>game-clicks.csv</td>
<td>GameClicks</td>
<td>A line is added each time a user performs a click in the game.</td>
<td>
<table>
<tr><td><b>timestamp</b></td><td>when the click occured</td></tr>
<tr><td><b>clickId</b></td><td>unique id for the click</td></tr>
<tr><td><b>userId</b></td><td>click user’s ID</td></tr>
<tr><td><b>userSessionId</b></td><td>id of the session of user when click occurs</td></tr>
<tr><td><b>idHit</b></td><td>if click hits flamingo (val=1) or missed (val=0)</td></tr>
<tr><td><b>teamId</b></td><td>id of the team of the user</td></tr>
<tr><td><b>teamId</b></td><td>id of the team of user</td></tr>
<tr><td><b>teamLevel</b></td><td>level of team during the session</td></tr>
</table>
</td>
</tr>
</table>

Aggregation

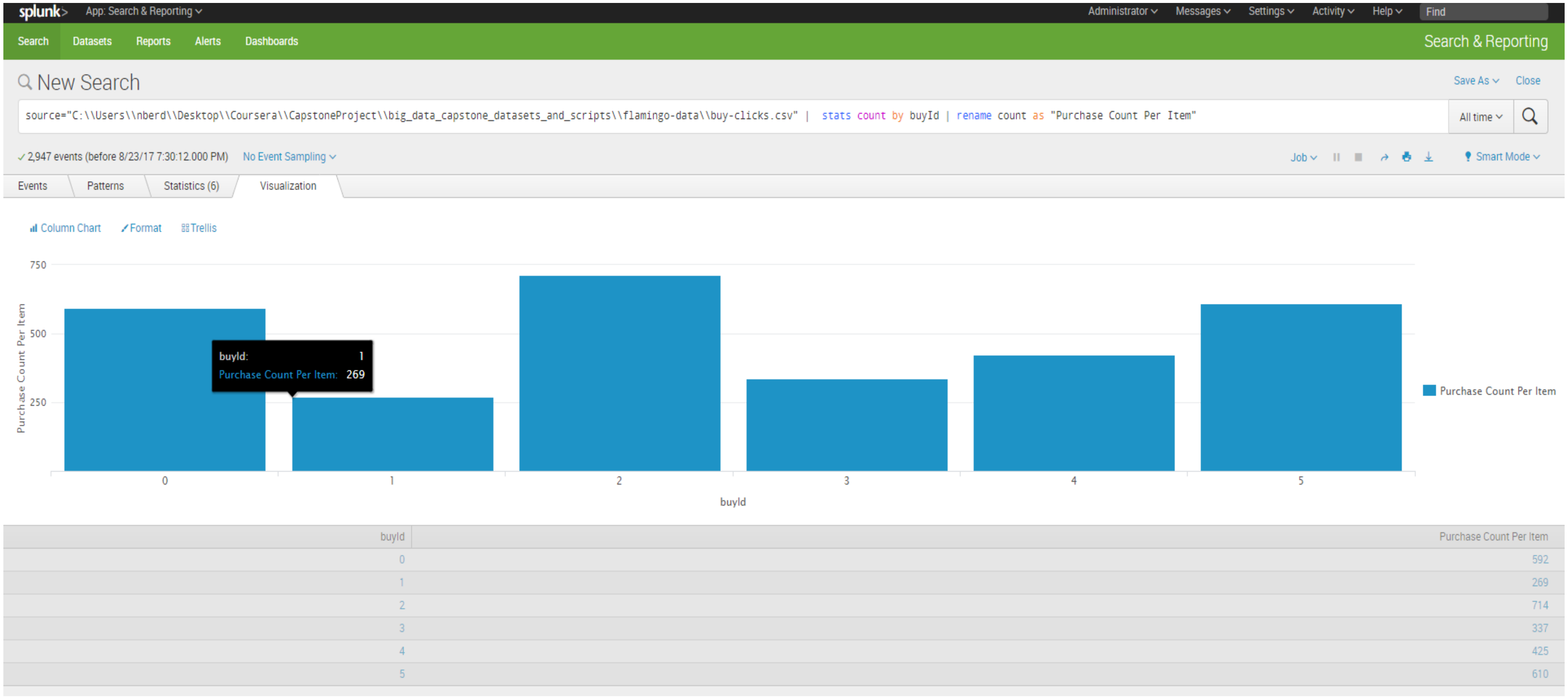
Amount spent buying items	source="buy-clicks.csv" | stats sum(price)	21407.0 This is a simple aggregation where we get all the products that gamers bought through the application and then we apply the aggregation function sum() on the price column to calculate the amount that gamers spent.
Number of unique items available to be purchased	source="buy-clicks.csv" | stats dc(buyId)	6 To find the available unique items we have to calculate the number of different item categories. Each line/record has an id of the item purchased (buyId), thus we have to fetch all items and make a DISTINCT COUNT on the buyId column.

Histograms

1. Purchase Count Per Item

This query is compiled as:
source="C:\\...\\buy-clicks.csv" | stats count by buyId | rename count as "Purchase Count Per Item"

The produced histogram is the following (screenshot from Splunk console):



```
source="C:\\Users\\nberd\\Desktop\\Coursera\\CapstoneProject\\big_data_capstone_datasets_and_scripts\\flamingo-data\\buy-clicks.csv" |  stats count by buyId | rename count as "Purchase Count Per Item"
```

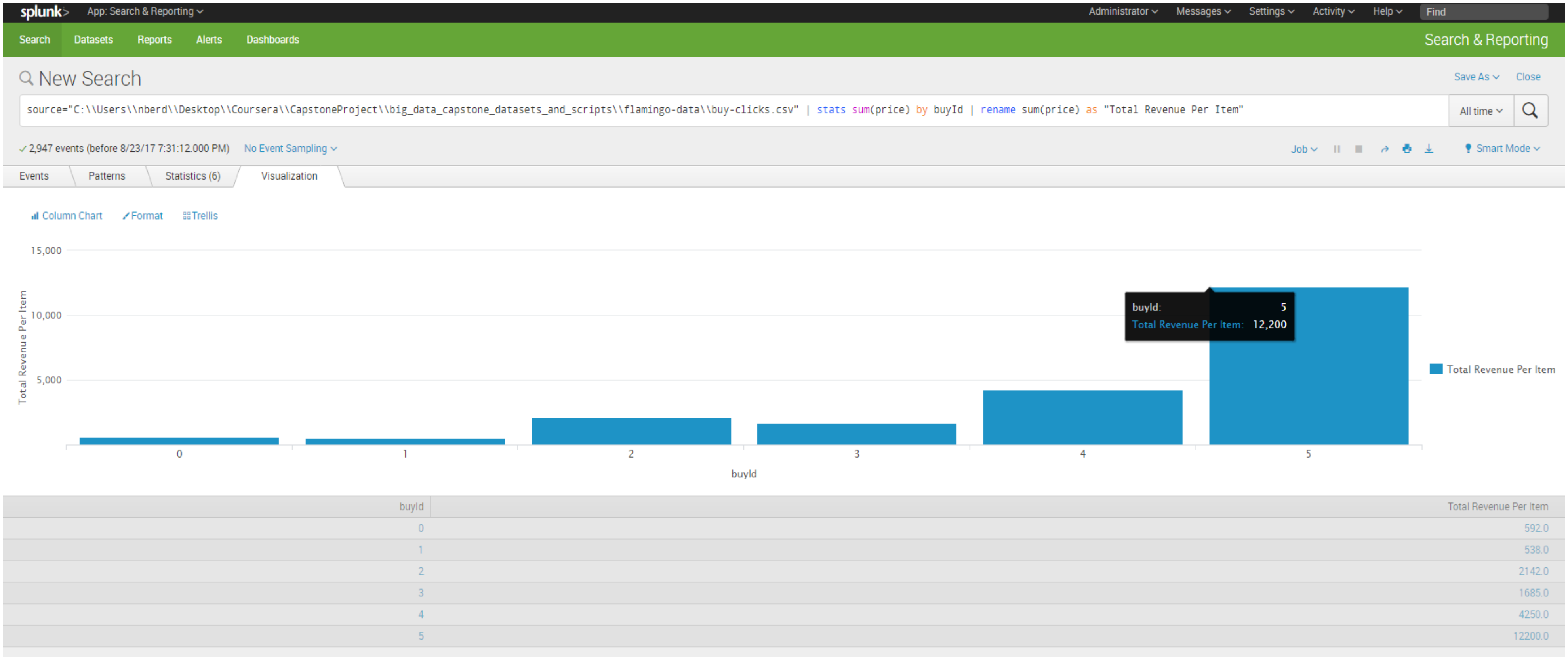
✓ 2,947 events (before 8/23/17 7:30:12.000 PM)   No Event Sampling

buyId	Purchase Count Per Item
0	592
1	269
2	714
3	337
4	425
5	610

2. Total Revenue Per Item

This query is compiled as:
source="C:\\Users…\\buy-clicks.csv" | stats sum(price) by buyId | rename sum(price) as "Total Revenue Per Item"

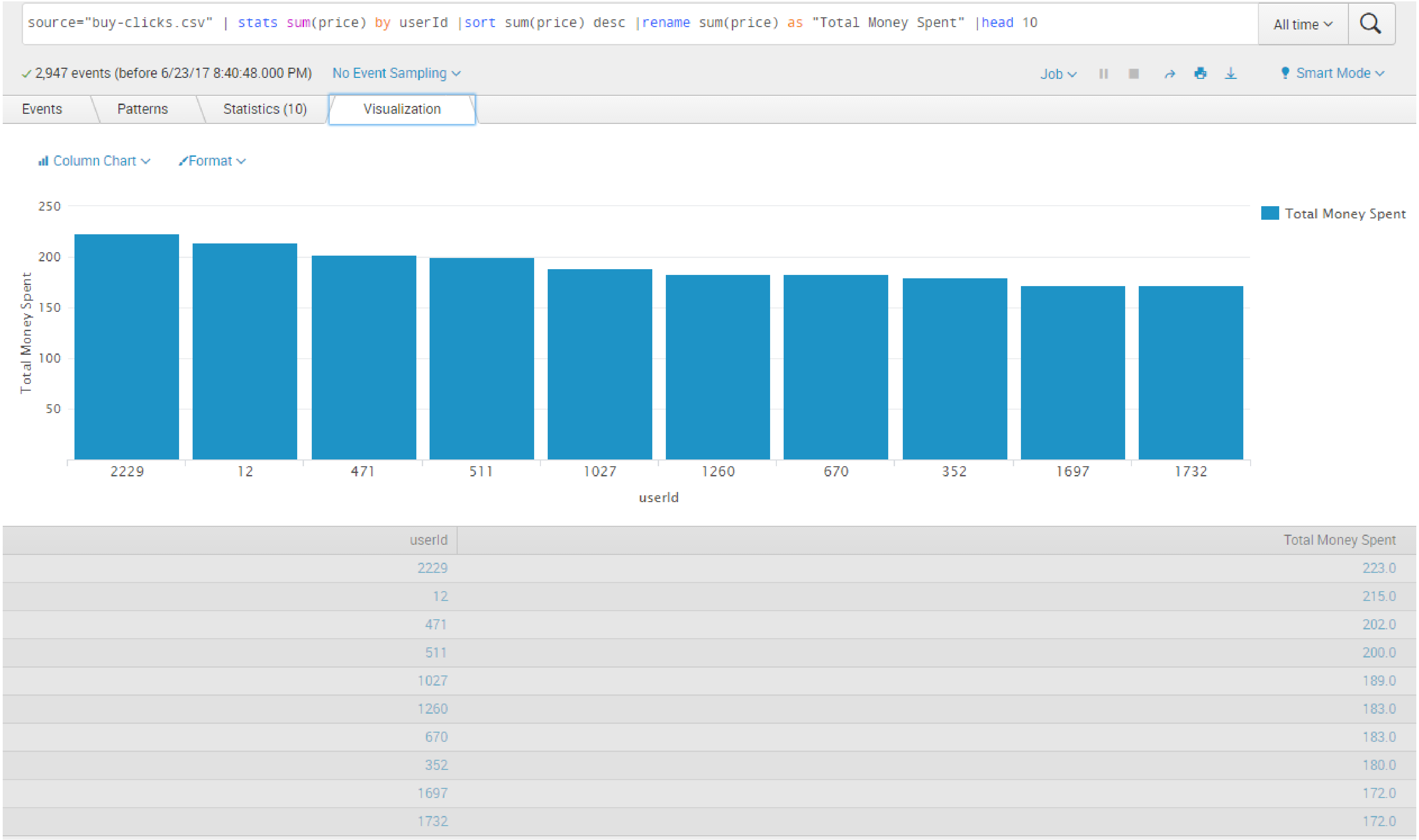
The produced histogram is the following (screenshot from Splunk console):



```
source="C:\\Users\\nberd\\Desktop\\Coursera\\CapstoneProject\\big_data_capstone_datasets_and_scripts\\flamingo-data\\buy-clicks.csv" | stats sum(price) by buyId | rename sum(price) as "Total Revenue Per Item"
```

✓ 2,947 events (before 8/23/17 7:31:12.000 PM) No Event Sampling

buyId	Total Revenue Per Item
0	592.0
1	538.0
2	2142.0
3	1685.0
4	4250.0
5	12200.0



```
source="buy-clicks.csv" | stats sum(price) by userId |sort sum(price) desc |rename sum(price) as "Total Money Spent" |head 10
```

✓ 2,947 events (before 6/23/17 8:40:48.000 PM)

userId	Total Money Spent
2229	223.0
12	215.0
471	202.0
511	200.0
1027	189.0
1260	183.0
670	183.0
352	180.0
1697	172.0
1732	172.0

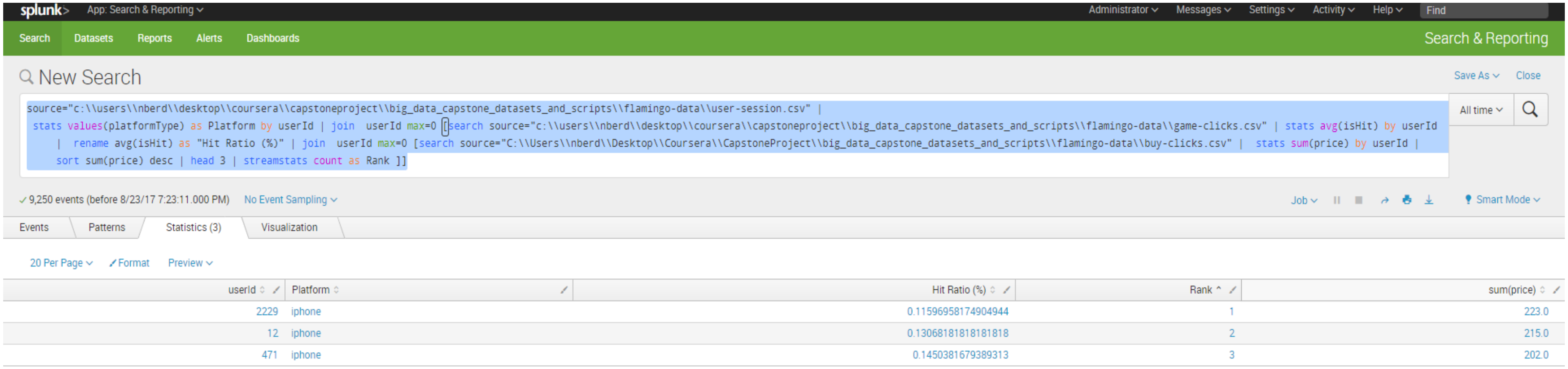
The above query and the histogram indicate the top 10 buyers among the gamers. Each bar indicates the Total money that a gamer spent to for purchases.

Filtering

A histogram showing total amount of money spent by the top ten users (ranked by how much money they spent).

The following table can be produce by writing a more complex query with sub searches. We'll use the JOIN command to achieve it. Find below the query and the screenshot from my Splunk Console:

source="c:\\users\\nberd\\desktop\\coursera\\capstoneproject\\big_data_capstone_datasets_and_scripts\\flamingo-data\\user-session.csv" | stats values(platformType) as Platform by userId | join userId max=0 [search source="c:\\users\\nberd\\desktop\\coursera\\capstoneproject\\big_data_capstone_datasets_and_scripts\\flamingo-data\\game-clicks.csv" | stats avg(isHit) by userId | rename avg(isHit) as "Hit Ratio (%)" | join userId max=0 [search source="C:\\Users\\nberd\\Desktop\\Coursera\\CapstoneProject\\big_data_capstone_datasets_and_scripts\\flamingo-data\\buy-clicks.csv" | stats sum(price) by userId | sort sum(price) desc | head 3 | streamstats count as Rank]]



```
source="c:\\users\\nberd\\desktop\\coursera\\capstoneproject\\big_data_capstone_datasets_and_scripts\\flamingo-data\\user-session.csv" |
 stats values(platformType) as Platform by userId | join  userId max=0 [search source="c:\\users\\nberd\\desktop\\coursera\\capstoneproject\\big_data_capstone_datasets_and_scripts\\flamingo-data\\game-clicks.csv" | stats avg(isHit) by userId
    |  rename avg(isHit) as "Hit Ratio (%)" | join  userId max=0 [search source="C:\\Users\\nberd\\Desktop\\Coursera\\CapstoneProject\\big_data_capstone_datasets_and_scripts\\flamingo-data\\buy-clicks.csv" |  stats sum(price) by userId |
    sort sum(price) desc | head 3 | streamstats count as Rank ]]
```

✓ 9,250 events (before 8/23/17 7:23:11.000 PM) No Event Sampling

userId	Platform	Hit Ratio (%)	Rank	sum(price)
2229	iphone	0.11596958174904944	1	223.0
12	iphone	0.13068181818181818	2	215.0
471	iphone	0.1450381679389313	3	202.0

The nested query find the top 3 userId ranked by how much they spent. Then we pass these ids to find the Platform and calculate the Hit-Ratio percentage

The following table shows the user id, platform, and hit-ratio percentage for the top three buying users:

Rank	User Id	Platform	Hit-Ratio (%)

1	2229	iphone	11.6% (61/526)
2	12	iphone	13.1% (92/704)
3	471	iphone	14.5% (76/524)