**Data Science Practicum Report**

**Preprocessing**

During this round of analysis, we tracked only one record per customer. We did this by taking the most recent customer record. This record confirms either if the customer churned or not. Most of the columns have a cumulative aspect so the last record is mostly a summary of all the customer transactions. We acknowledge that we may have lost some information in this form but we tried to counter this by creating a new column ‘New\_Duration’ which represents the period the customer has maintained an active subscription.

We then went ahead and removed columns we felt would not affect churn or had over 80% of missing data or were deemed to be confounding variables.

Below are the columns we removed:

* SegmentName,
* 'Emails\_Recency,
* status
* Duration
* EventsAttended\_LastDateHits\_LastDate
* CollegeProspects\_LastDate
* created\_atupdated\_atMonth
* Unnamed: 0'
* MSG\_CHURN
* Made\_Team?
* New\_DurationID
* subscription\_Ended
* AthleteSubAthleteSubscription'

We then preprocessed more by creating dummy variables for categorical variables.

Final table shape: **16117 rows 75 columns**

**Model Building**

We implemented three models machine-learning models recorded and visualized Precision and Recall Values.

Below are the three models we chose:

* Decision Trees
* Logistic Regression
* Suppport Vector Machines (SVM)

When it came to building models, we used 3 months worth of data (Jan – March 2014) and used it to predict April 2014 churn.

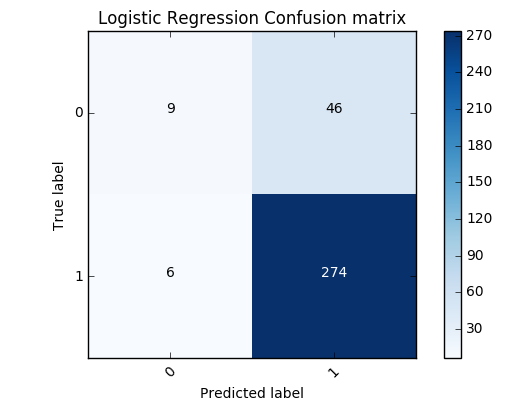
Training data: **827 rows 75 columns**

Test data: **: 335 rows 75 columns**

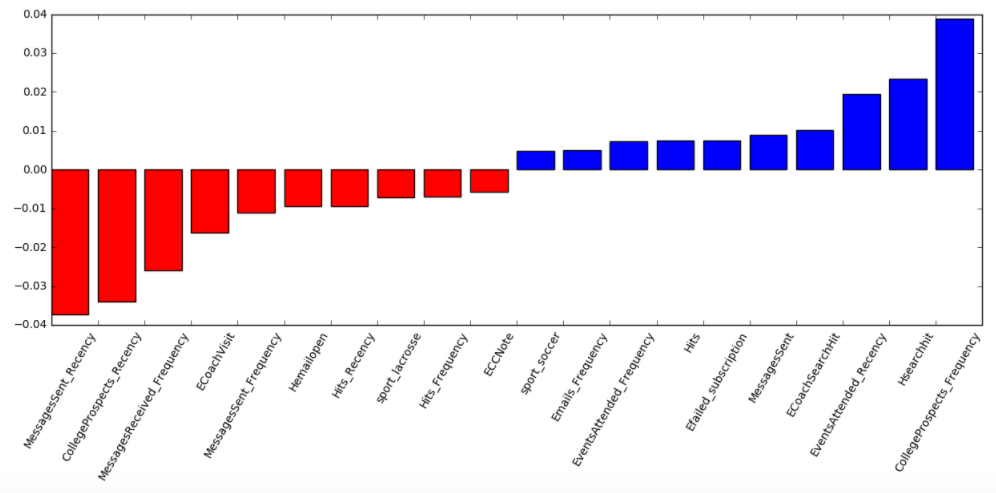
Summary of Precision Recall Values

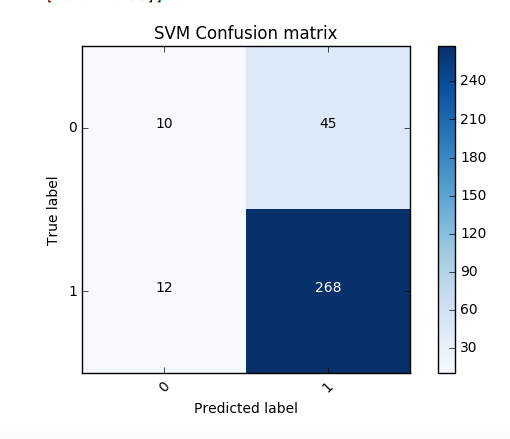
|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** |
| Decision Tree | 0.85 | 0.82 | 0.85 |
| Logistic Regression | 0.86 | 0.97 | 0.91 |
| SVM | 0.86 | 0.95 | 0.90 |

**Logistic Regression Visualization**

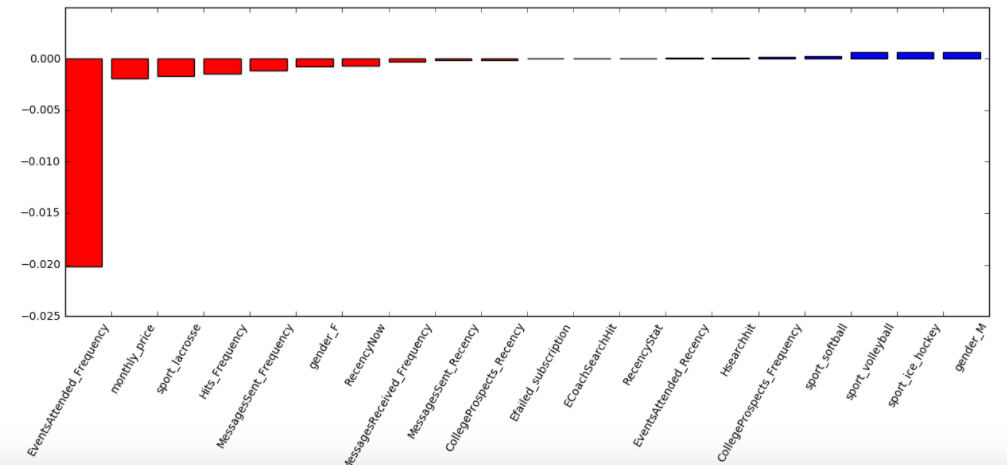


**Important Features According to Logistic Regression**

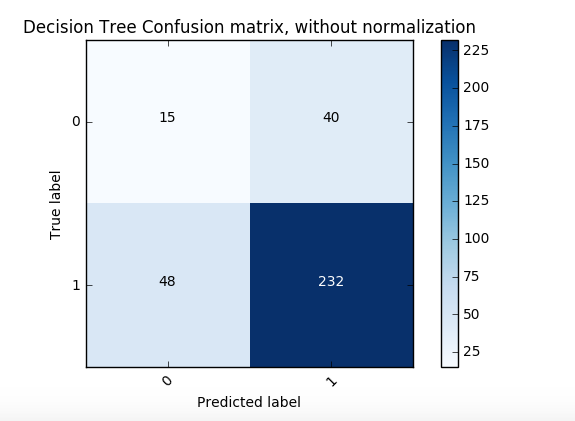


**Support Vector Machine Visualization** 

**Important Features According to SVM**



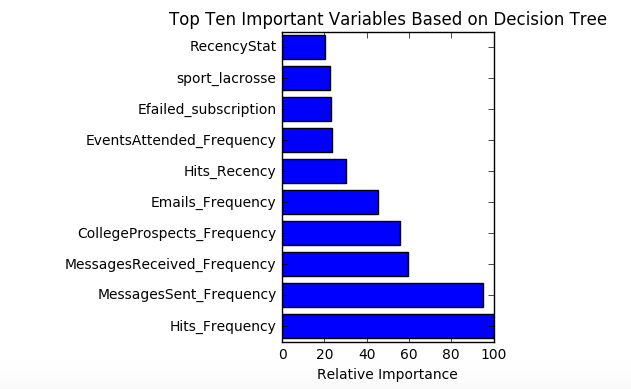
**Decision Tree Visualization**



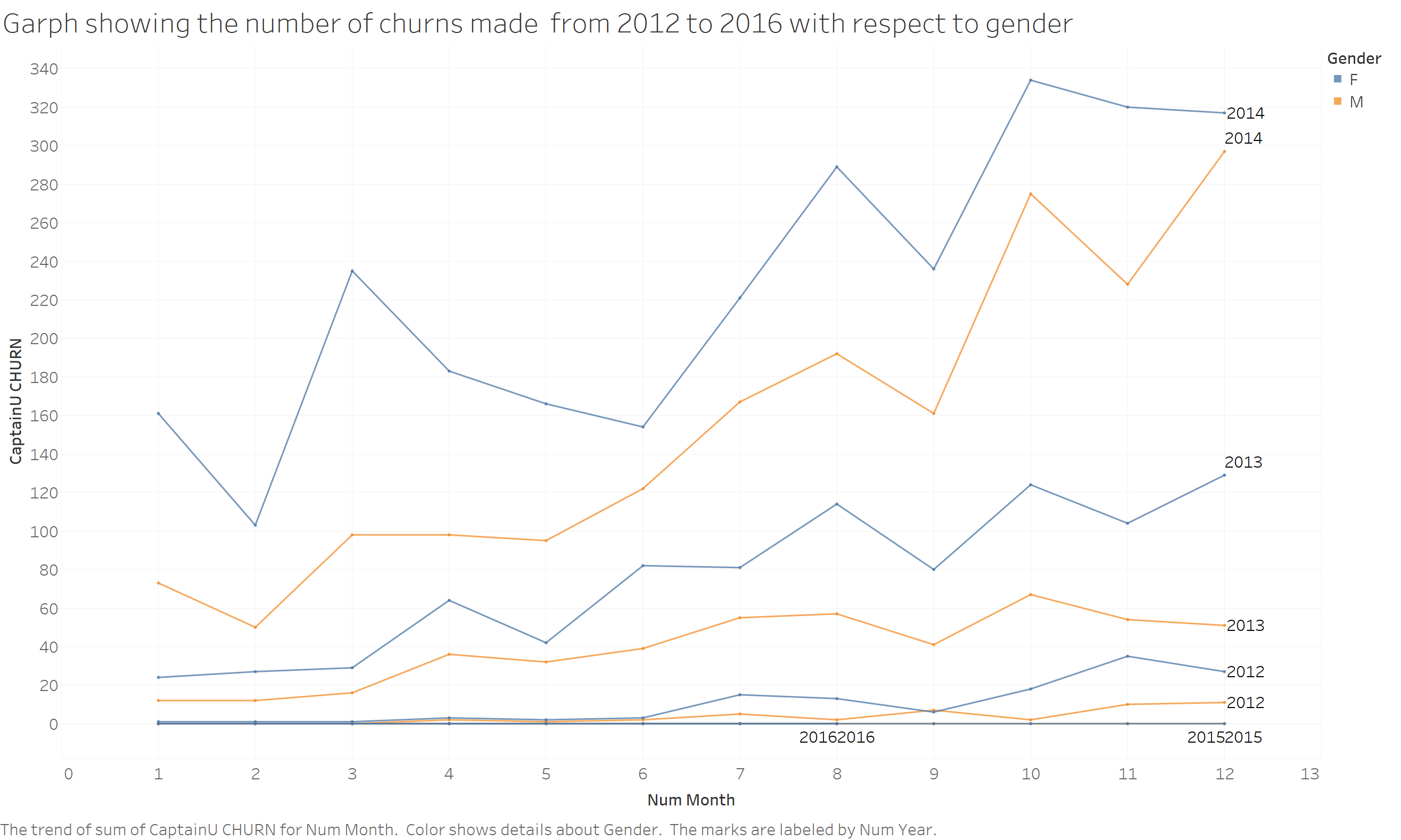
**Feature Importance**

The Decision Tree Model deemed the following features important:

* Hits Frequency
* Message\_Sent\_Frequency
* Message\_Received\_Frequency
* College\_Prospects\_Frequency
* Emails\_Frequency
* Hits\_Frequency
* EventsAttened\_Frequency
* Efailed\_subscription
* Sports\_lacrosse
* RecencyStat

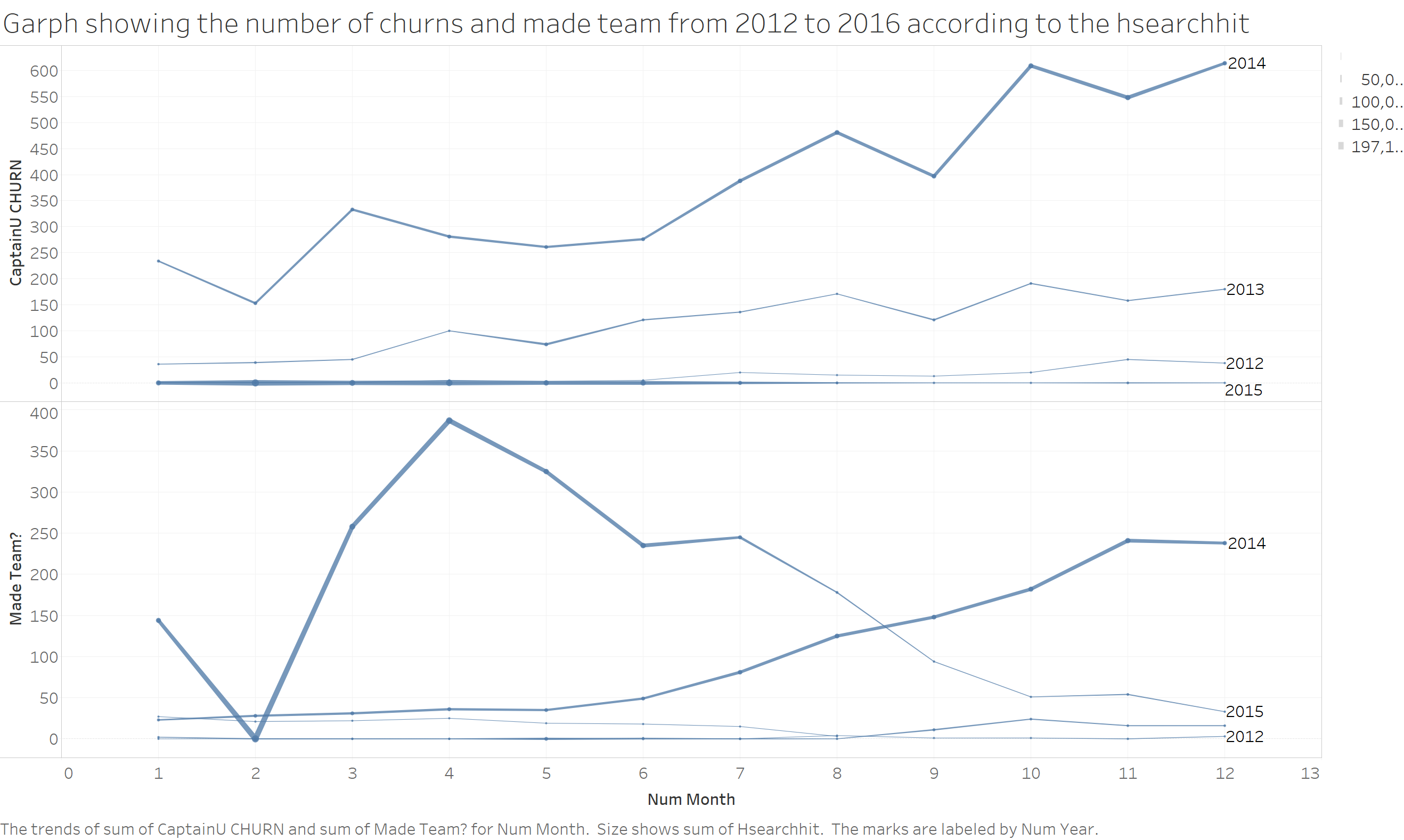


These are the few insights which we got from the whole data

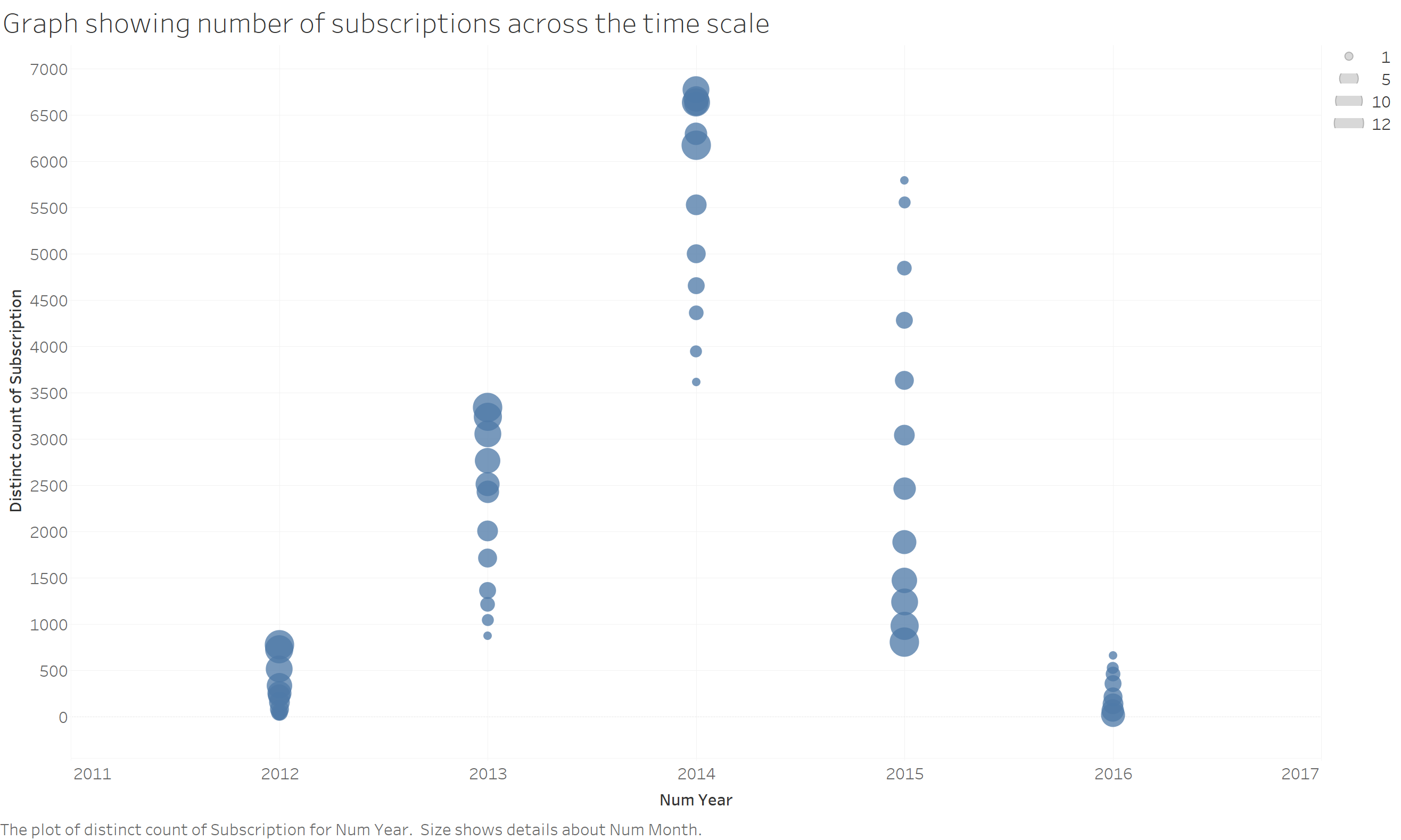




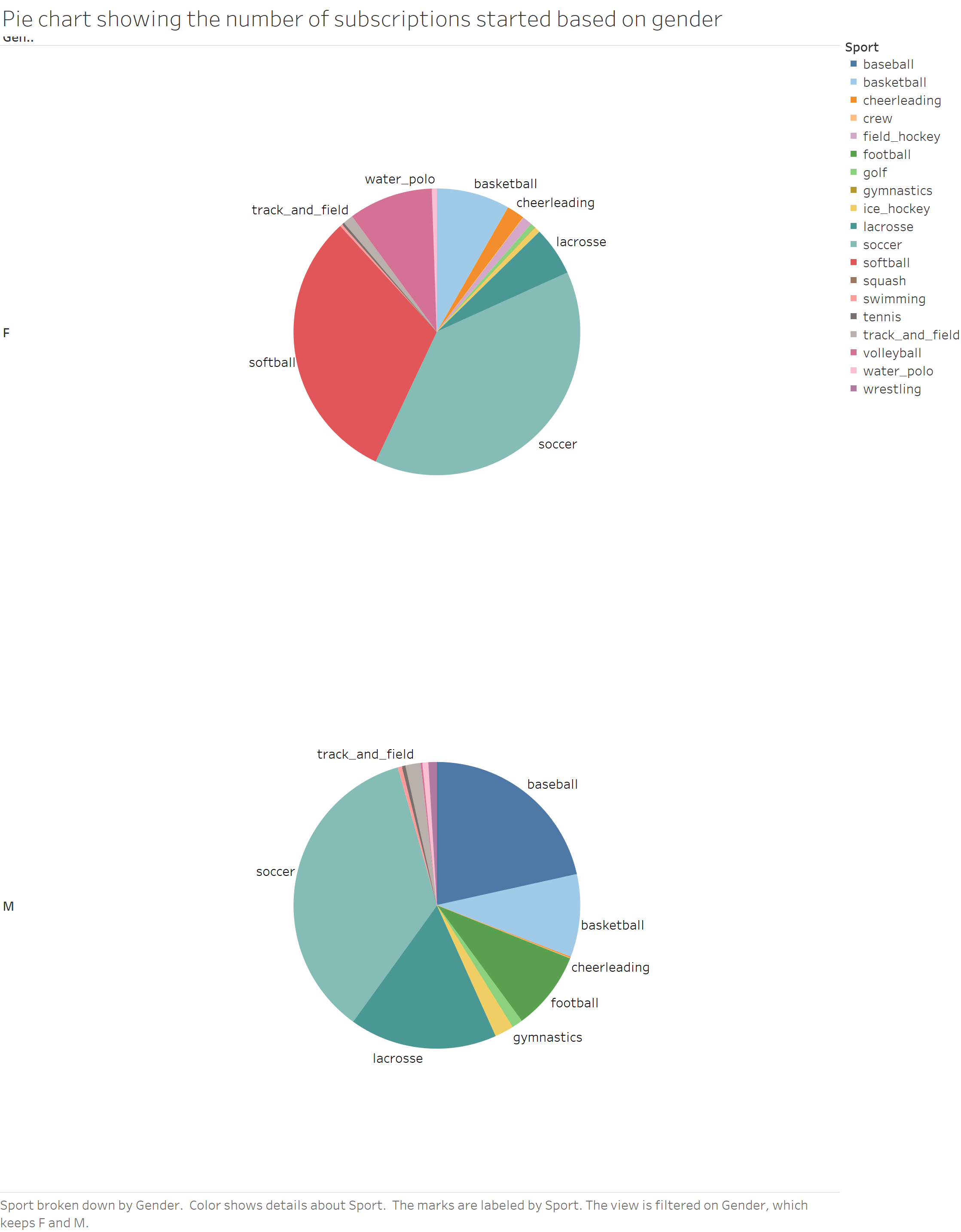
It is observed that the number of churns is more in 2014. And the second largest number of churns happened in 2013. One major observation is that the number of churners in 2015 and 2016 is almost zero.



The graph showing the number of churns and made team based of number of hsearchhits



It is observed that the number of subscriptions increased till 2014 and is decreased to a minimum from there on.



It is observed that the number of subscriptions is larger in the case of soccer and softball for the female athletes and the majority of the male athletes subscribed for the soccer, lacrosse and baseball.