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)

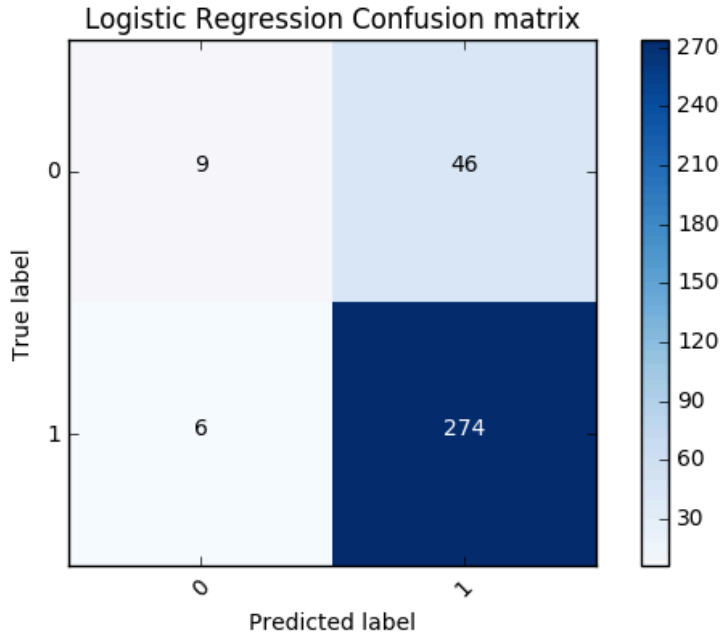
When it came to building models, we used 3 months worth of data (Jan – March 2014) and used it to predict April 2014 churn.

Training data: **827 rows 75 columns**
Test data: **: 335 rows 75 columns**

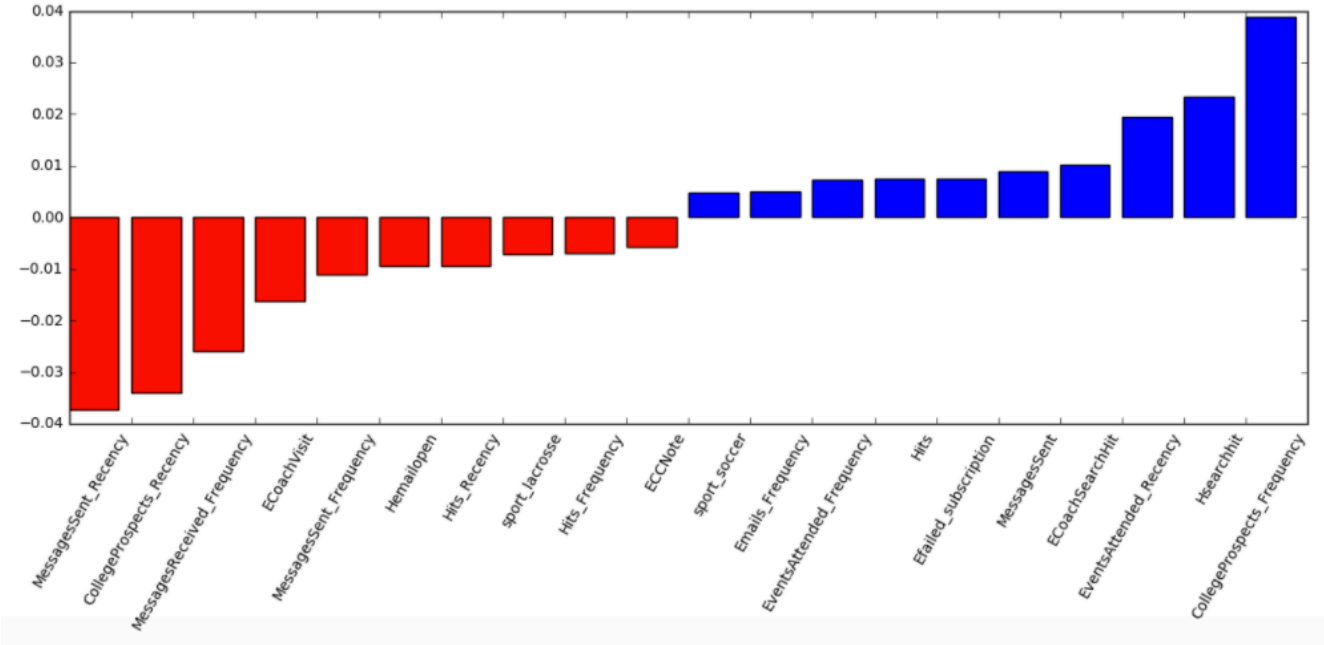
Summary of Precision Recall Values

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

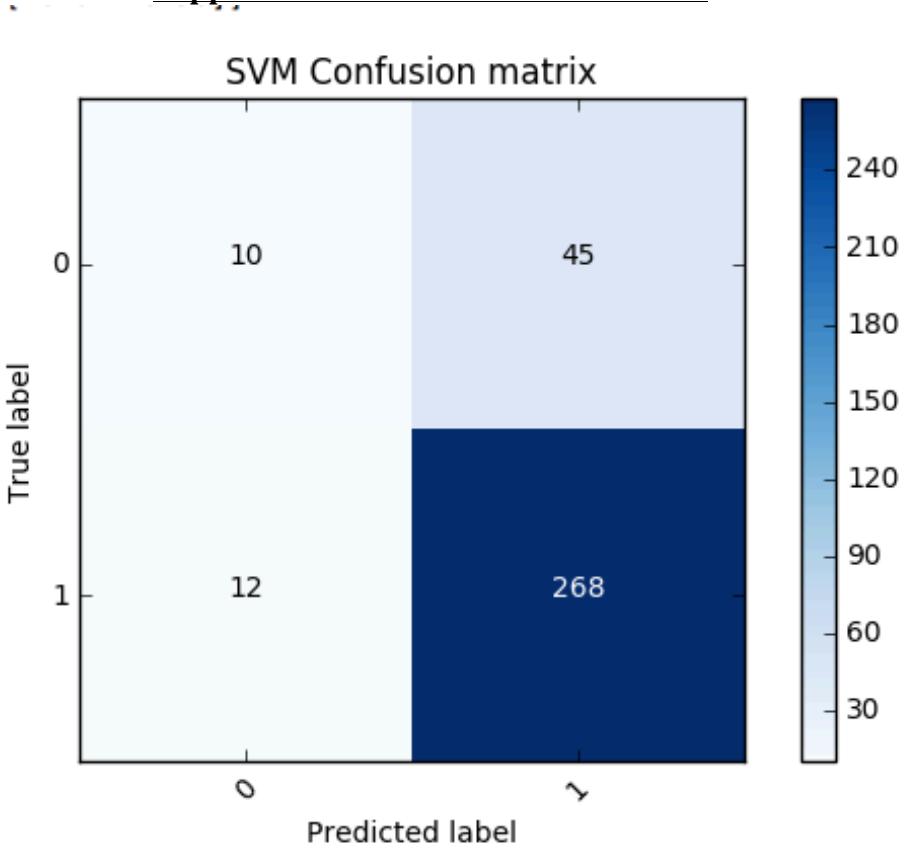
Logistic Regression Visualization



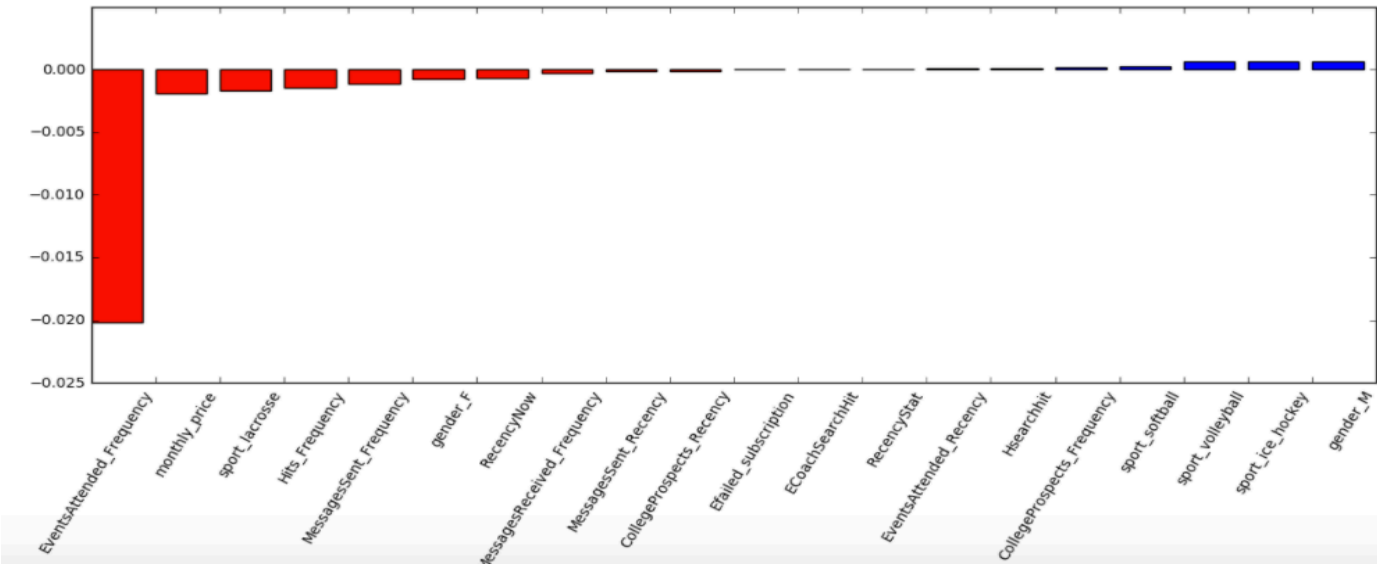
Important Features According to Logistic Regression



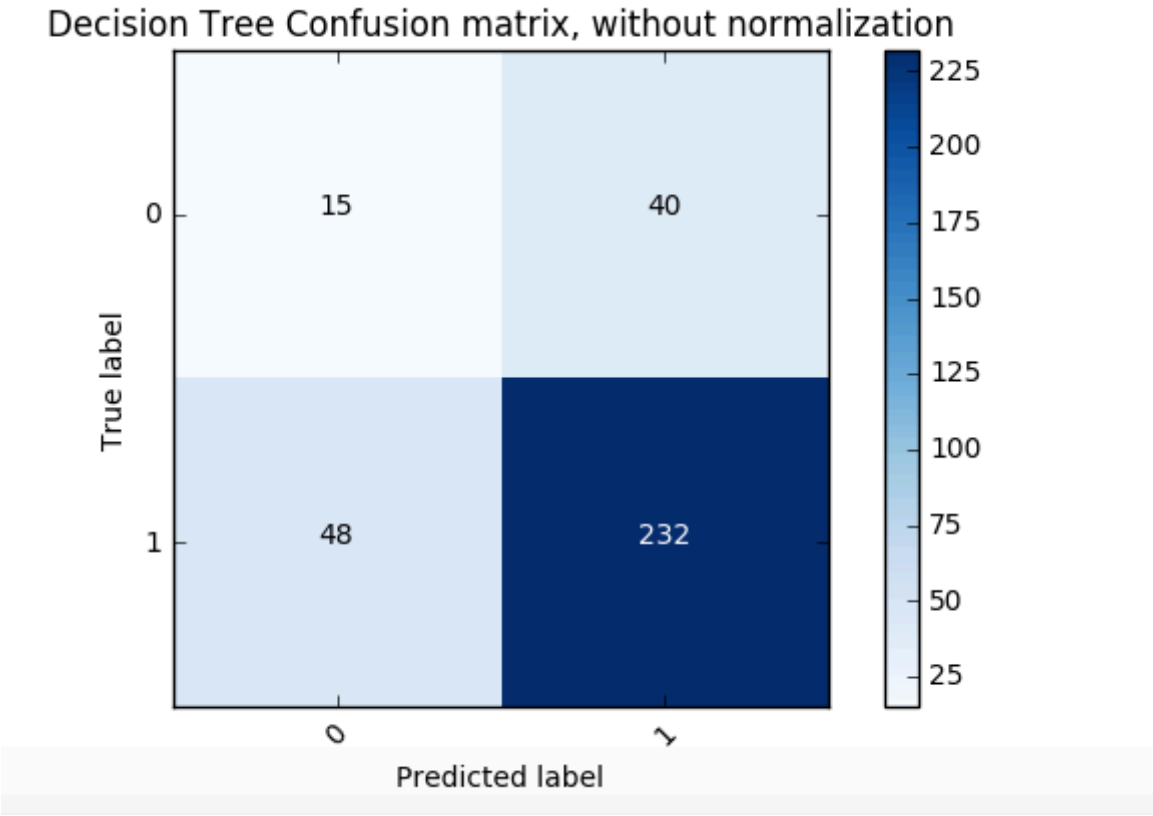
Support Vector Machine Visualization



Important Features According to SVM



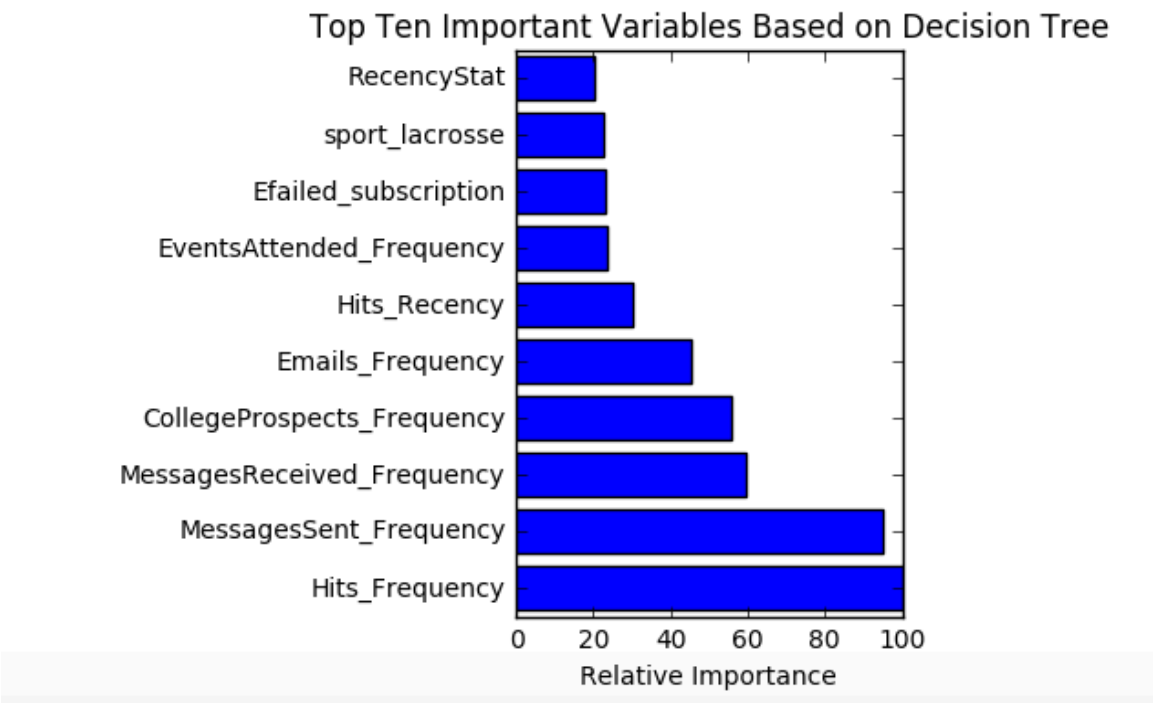
Decision Tree Visualization



Feature Importance

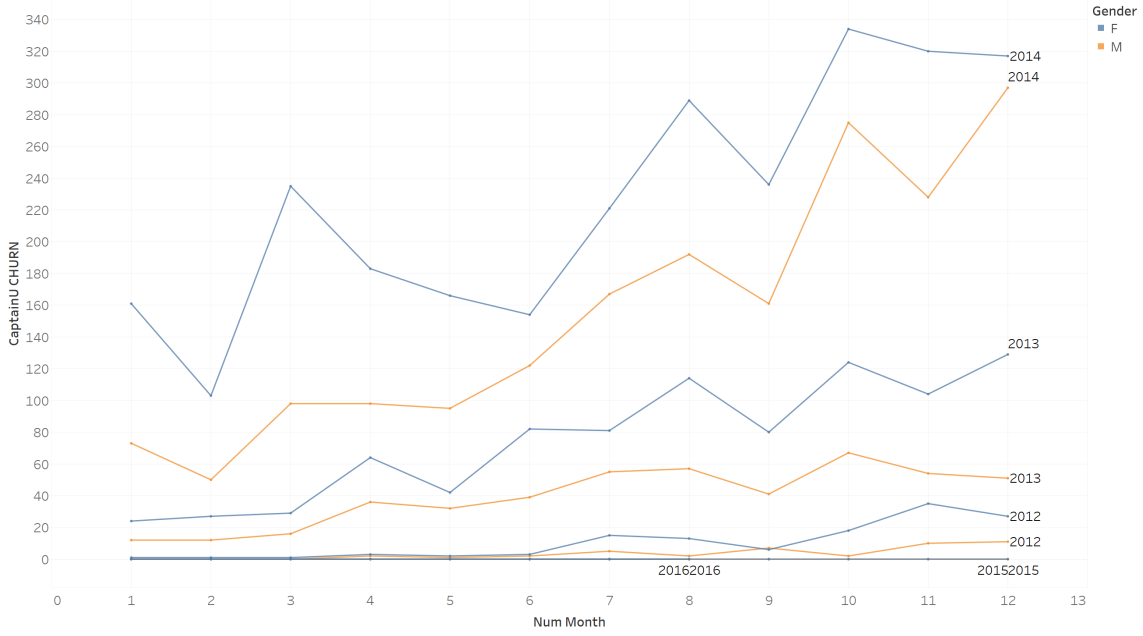
The Decision Tree Model deemed the following features important:

- 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

Garph showing the number of churns made from 2012 to 2016 with respect to gender



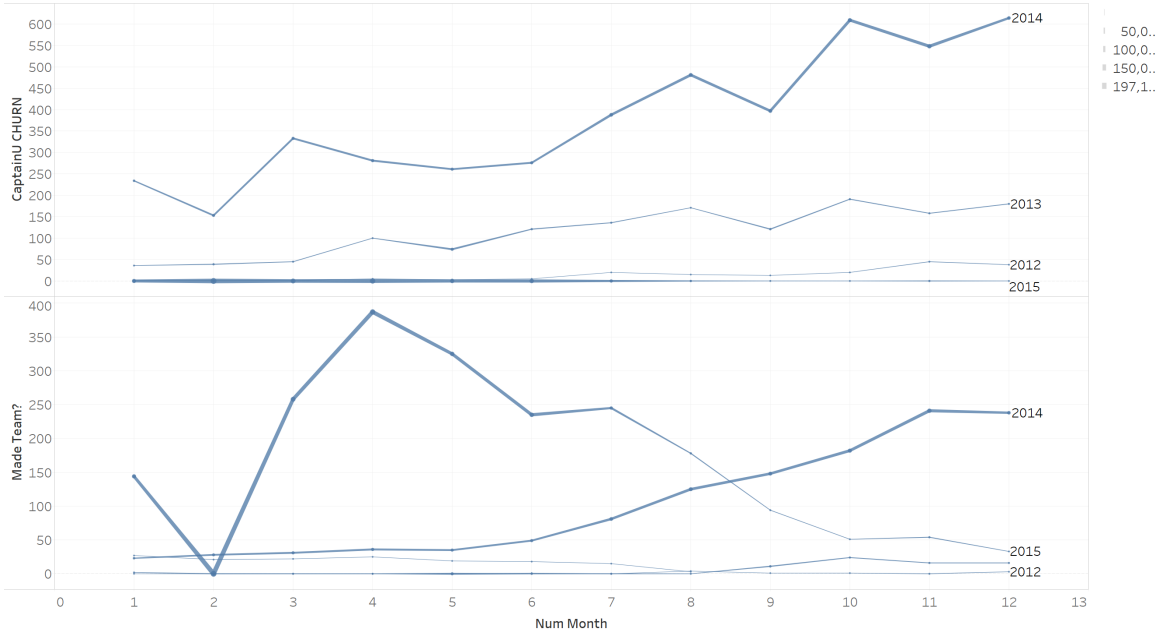
The trend of sum of CaptainU CHURN for Num Month. Color shows details about Gender. The marks are labeled by Num Year.

Graph showing the all the details of the churn

Gender, Num Year, sum of subscription Ended, sum of Subscription Started and sum of CaptainU CHURN. Color shows sum of CaptainU CHURN. Size shows sum of CaptainU CHURN. The marks are labeled by Gender, Num Year, sum of subscription Ended, sum of Subscription Started and sum of CaptainU CHURN. The view is filtered on Gender, which keeps F and M.

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.

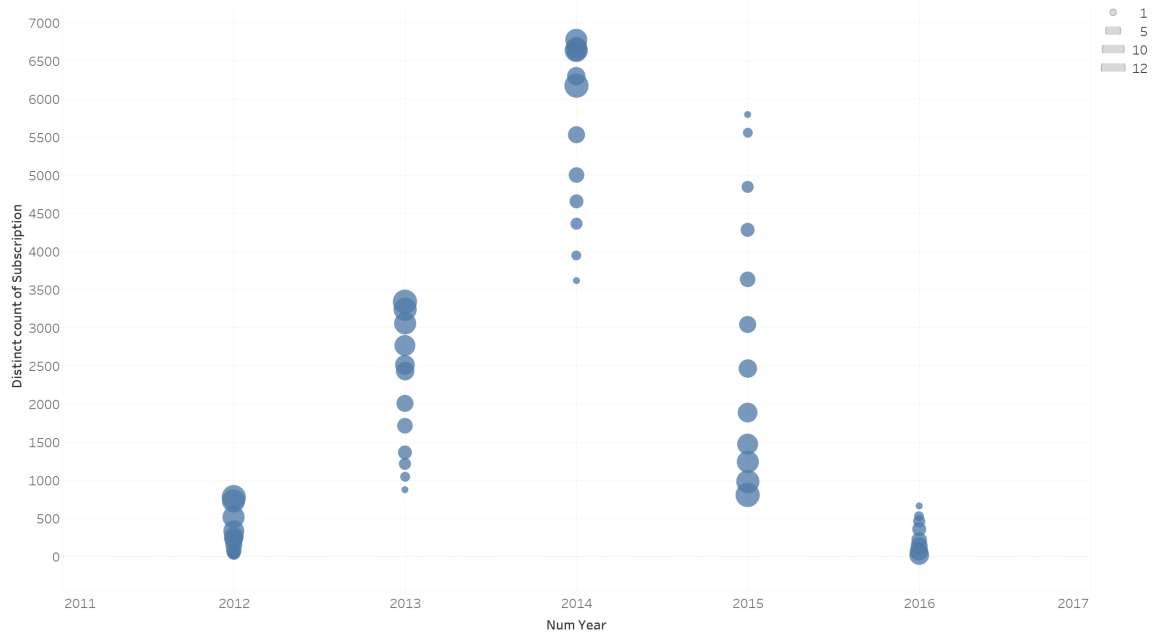
Garph showing the number of churns and made team from 2012 to 2016 according to the hsearchhit



The trends of sum of CaptainU CHURN and sum of Made Team? for Num Month. Size shows sum of Hsearchhit. The marks are labeled by Num Year.

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

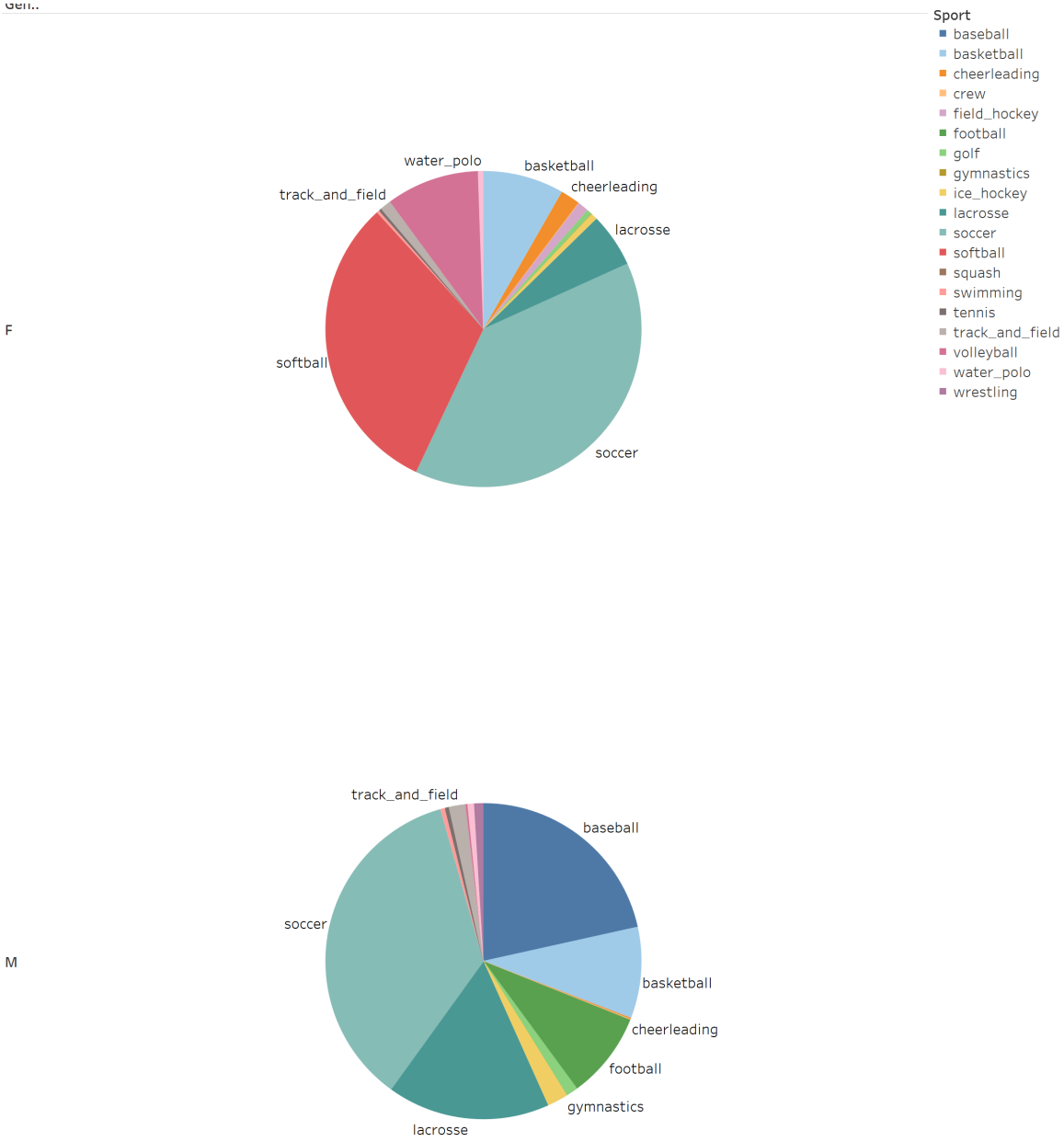
Graph showing number of subscriptions across the time scale



The plot of distinct count of Subscription for Num Year. Size shows details about Num Month.

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

Pie chart showing the number of subscriptions started based on gender



Sport broken down by Gender. Color shows details about Sport. The marks are labeled by Sport. The view is filtered on Gender, which keeps F and M.

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.