BABYLIST

Email Metrics Project

Peggy Lin

Master of Statistics, UC Davis

Nomura Research Institute

Outlines

Data description
From dim_user database, we got 1,500,000 users x 37 columns

---

Data overview

| 4 Email Lists	Email Related	Others
Baby Registry 101	1.Subscribe_at	ID
Newsletter	2.Delivered	Created_at
	3.Open	Original_arrival_date
Pregnancy_weekly	4.Click	referrer
Sponsored	5.Unsubscribe_at	Source...etc
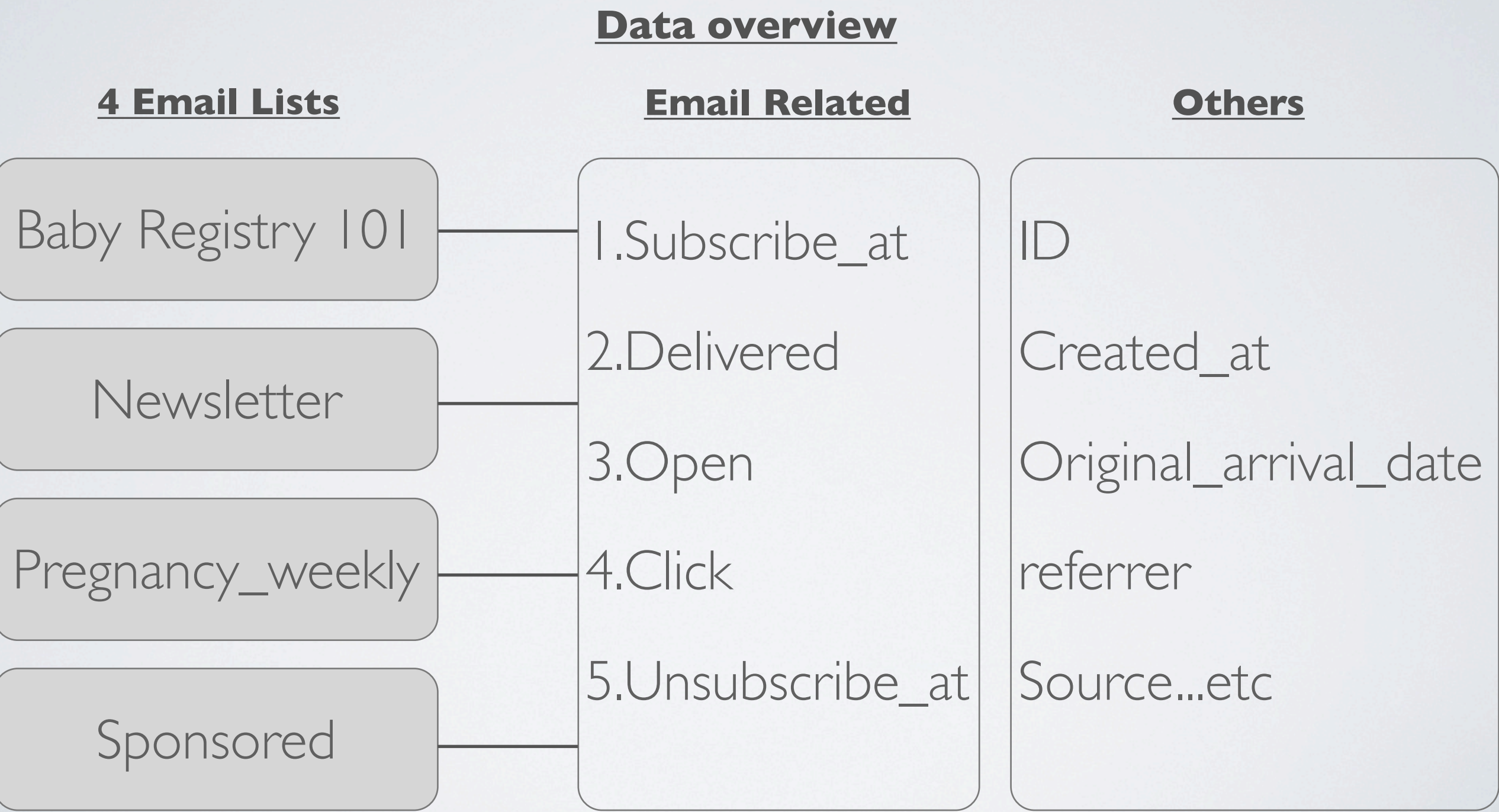
[Baby Registry 101] There is a stable growth in checklist. However, we see in September, there is a declining trend in this email list.
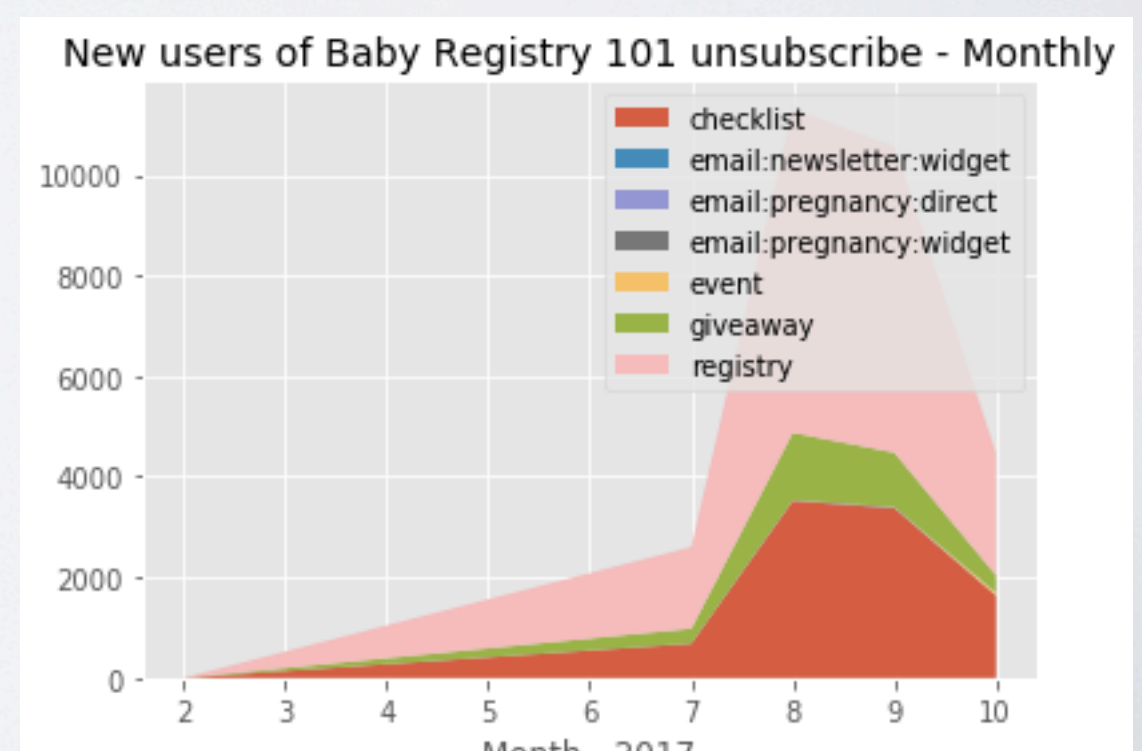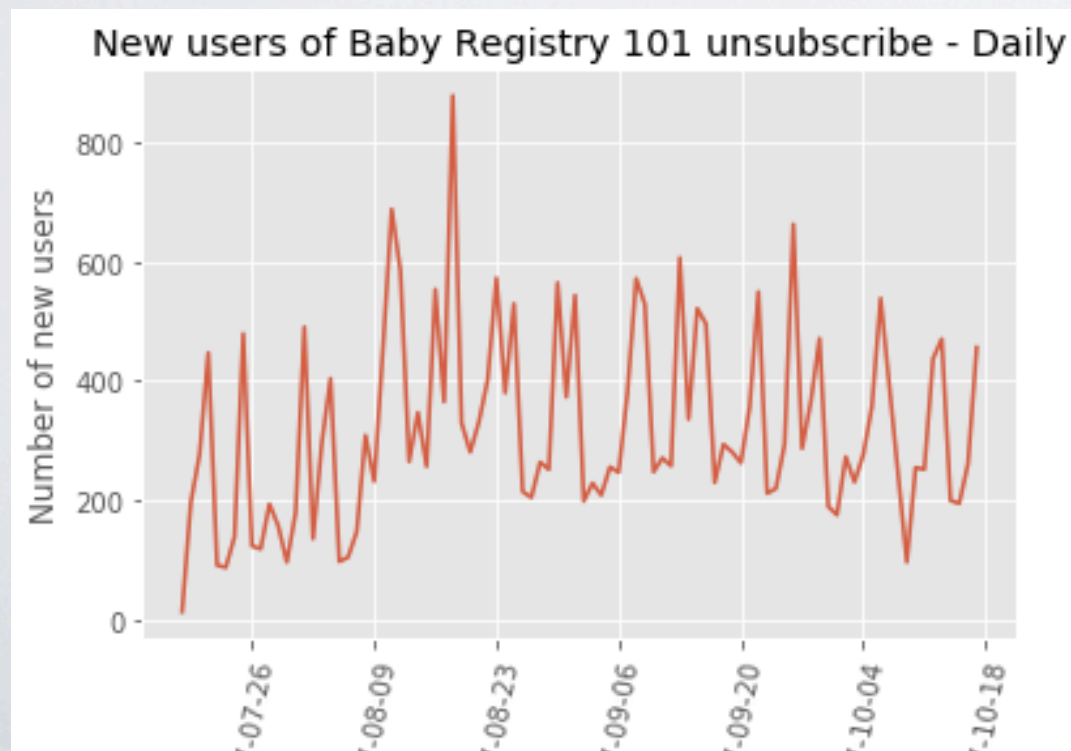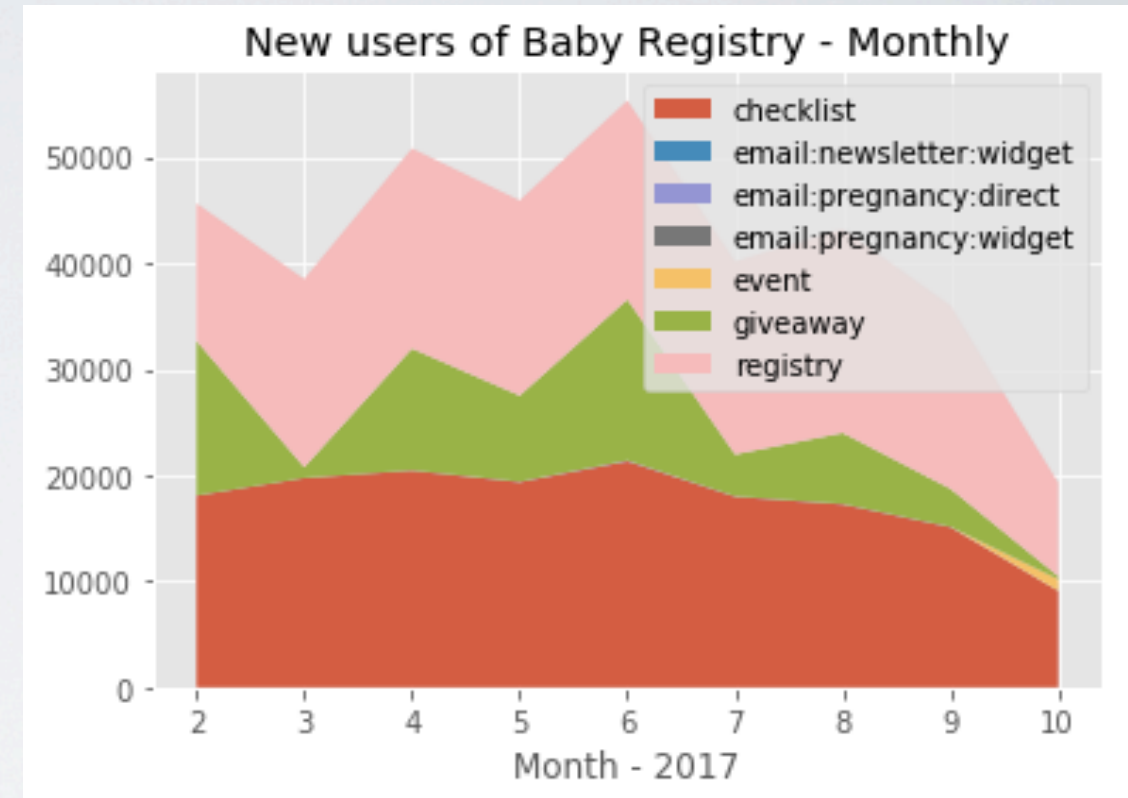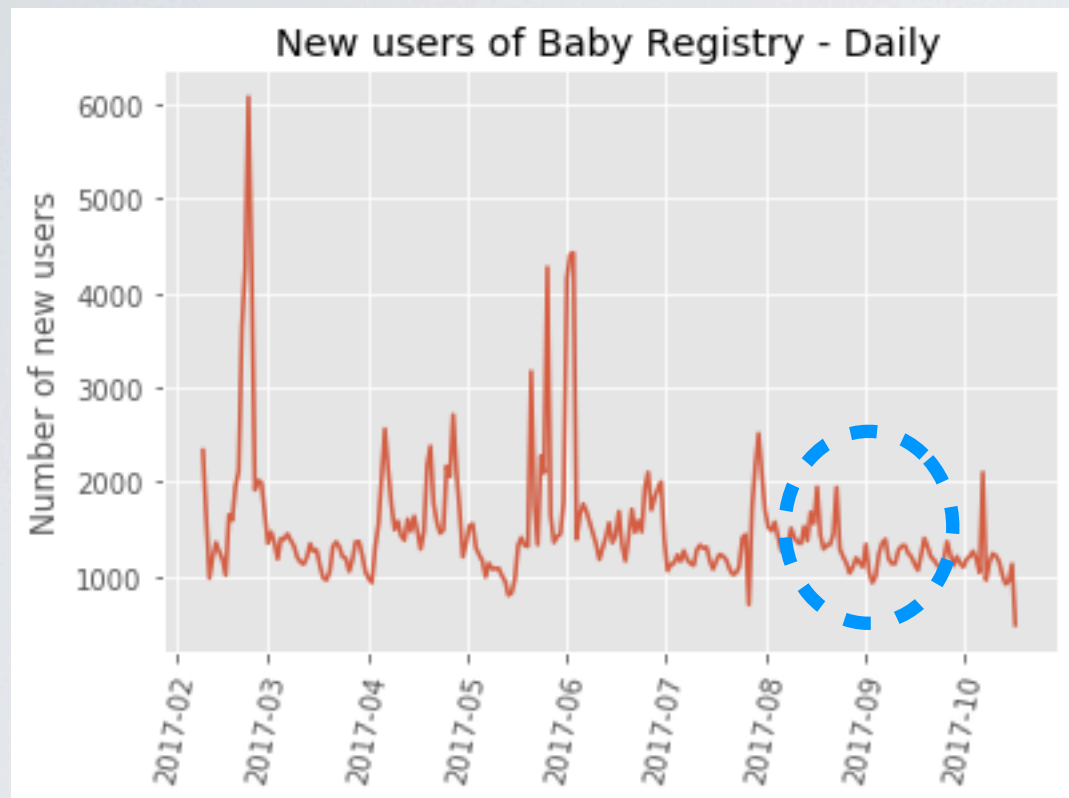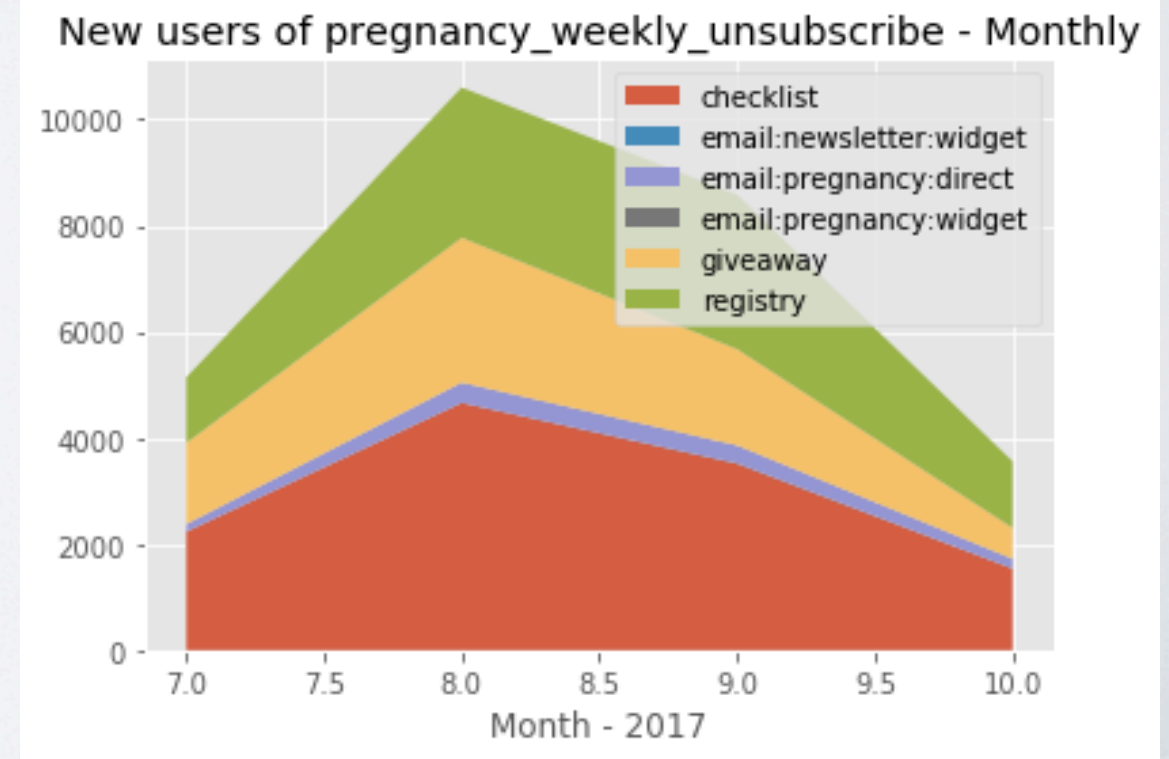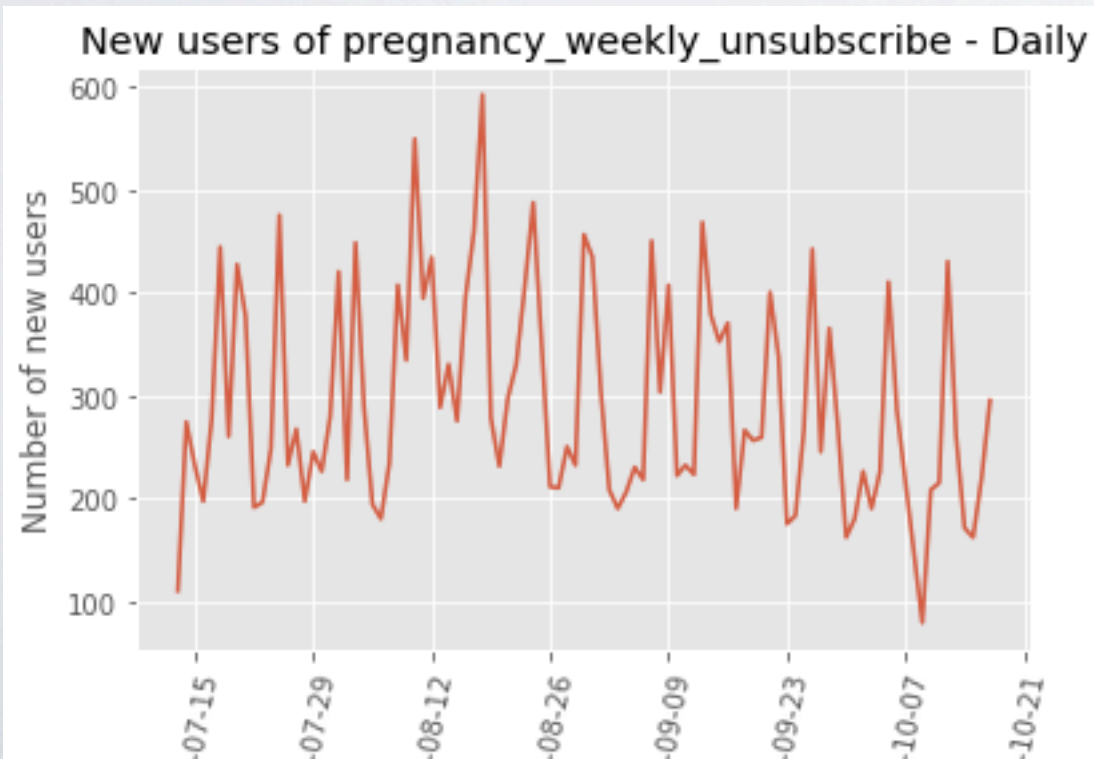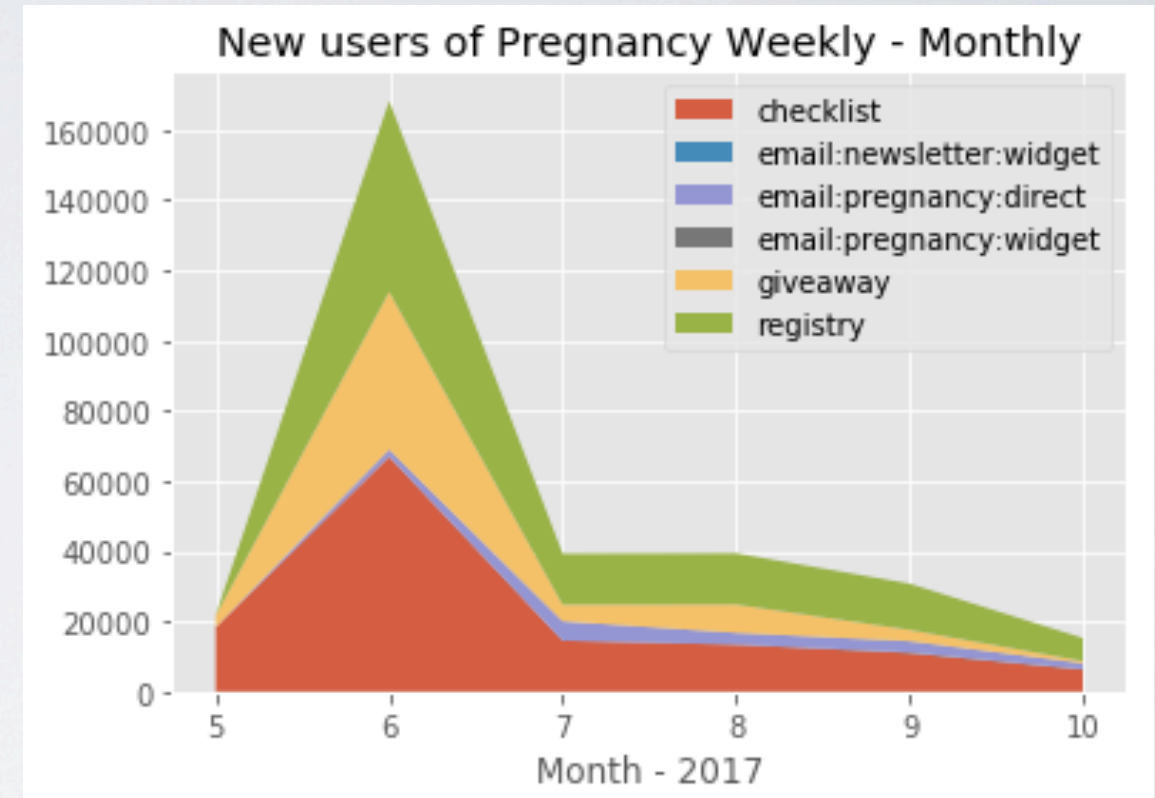
[Newsletter] We focus on data after 2017-08 and we can see the growth/churning source is mainly from registry/giveaway/checklist
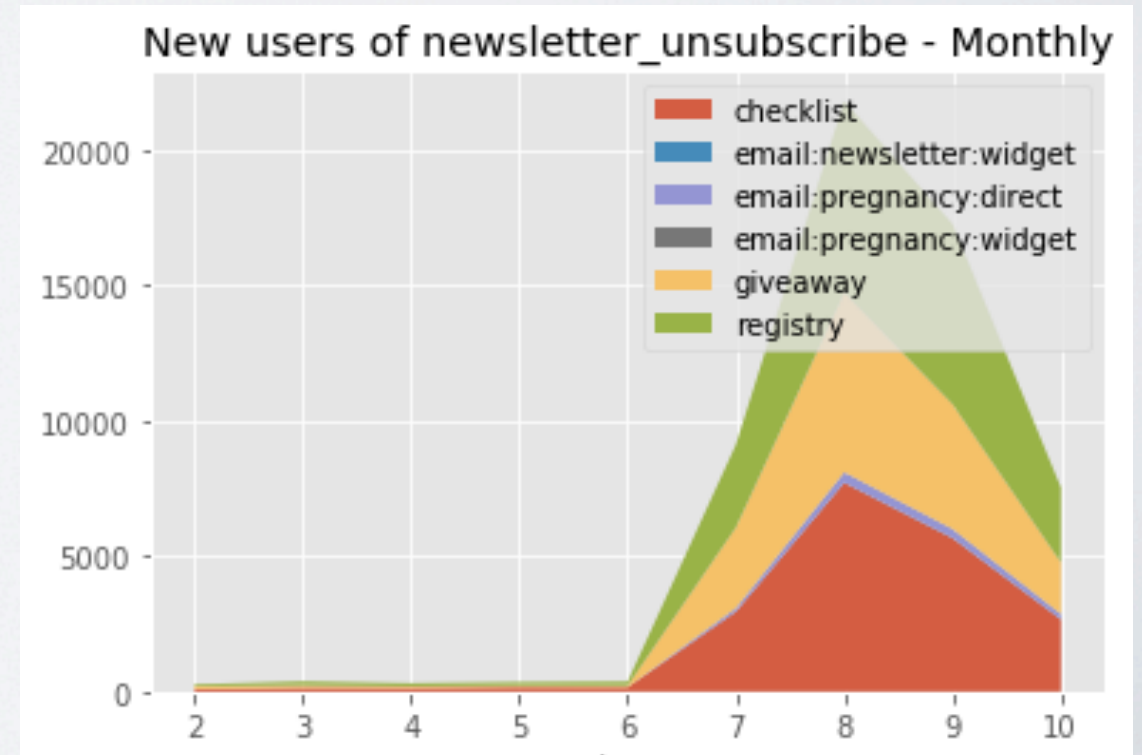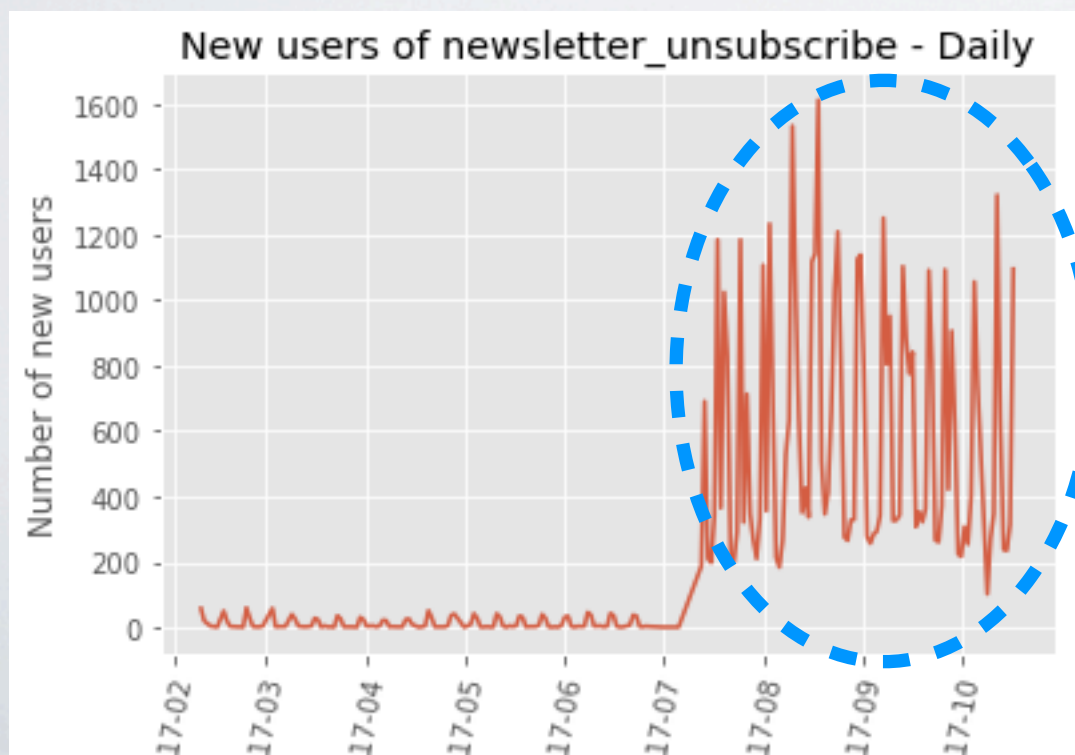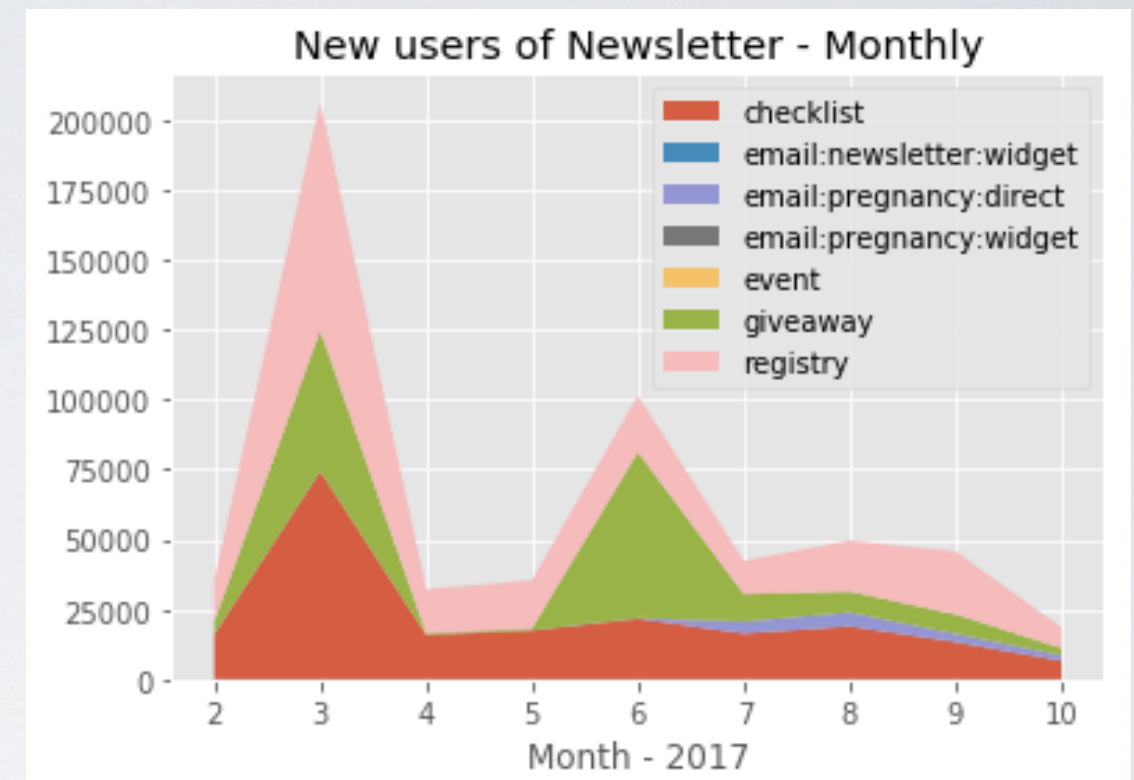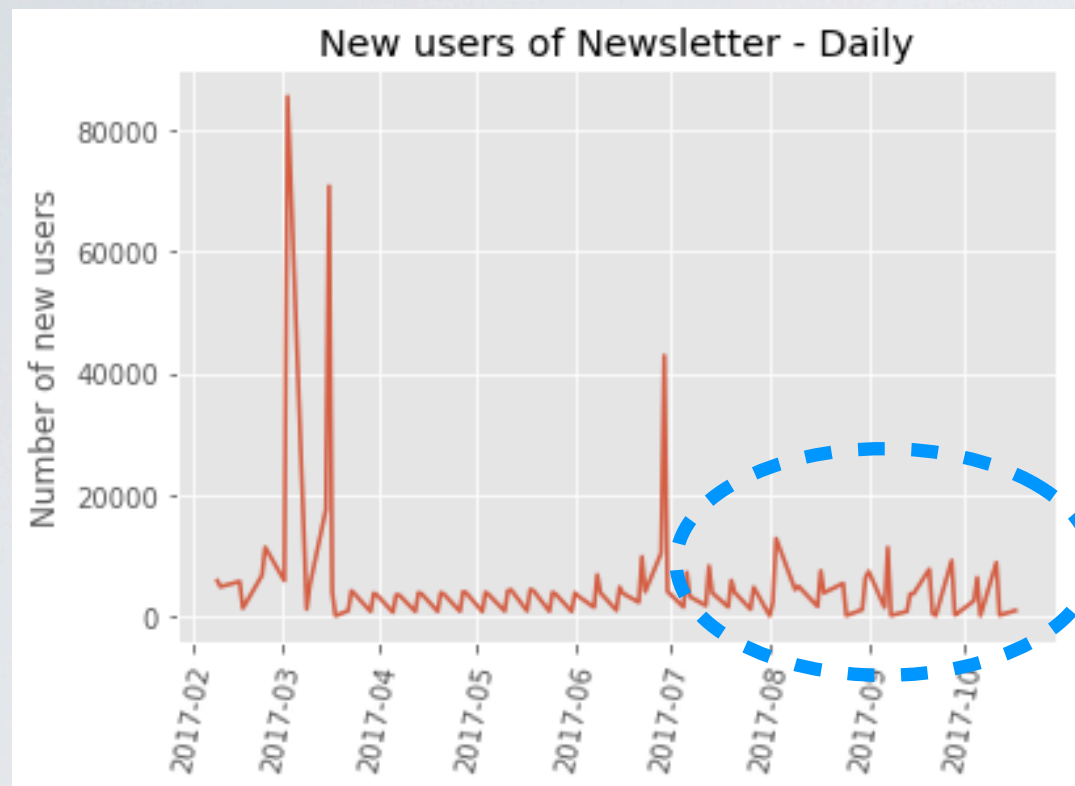
[Pregnancy Weekly] It is relatively stable and the growth/churning source is mainly from registry/giveaway/checklist
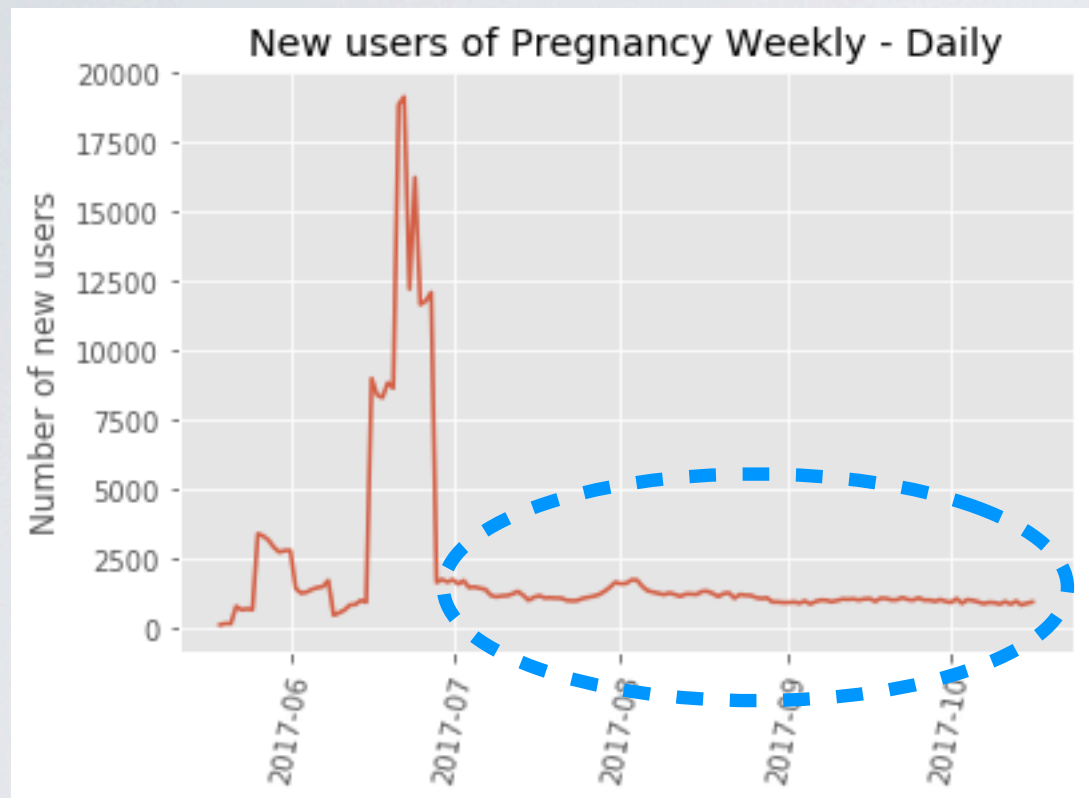
[Sponsored] This is a strategic e-mail list. We see a seasonal growth in this one.
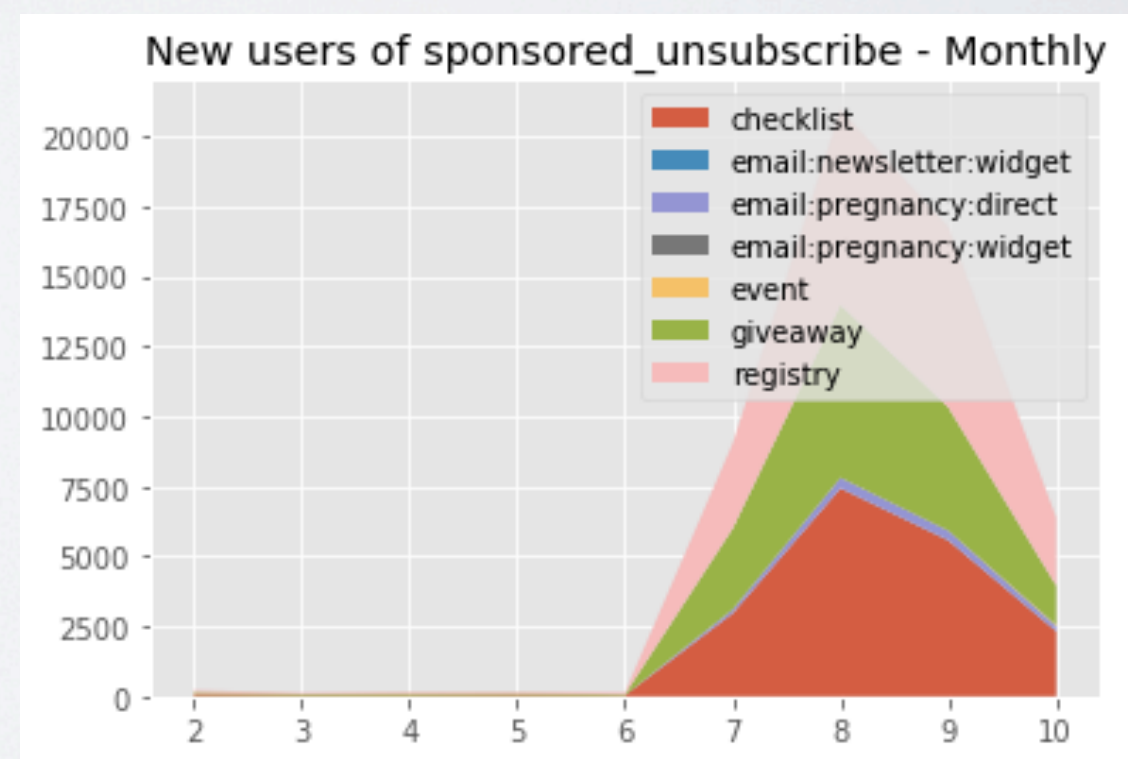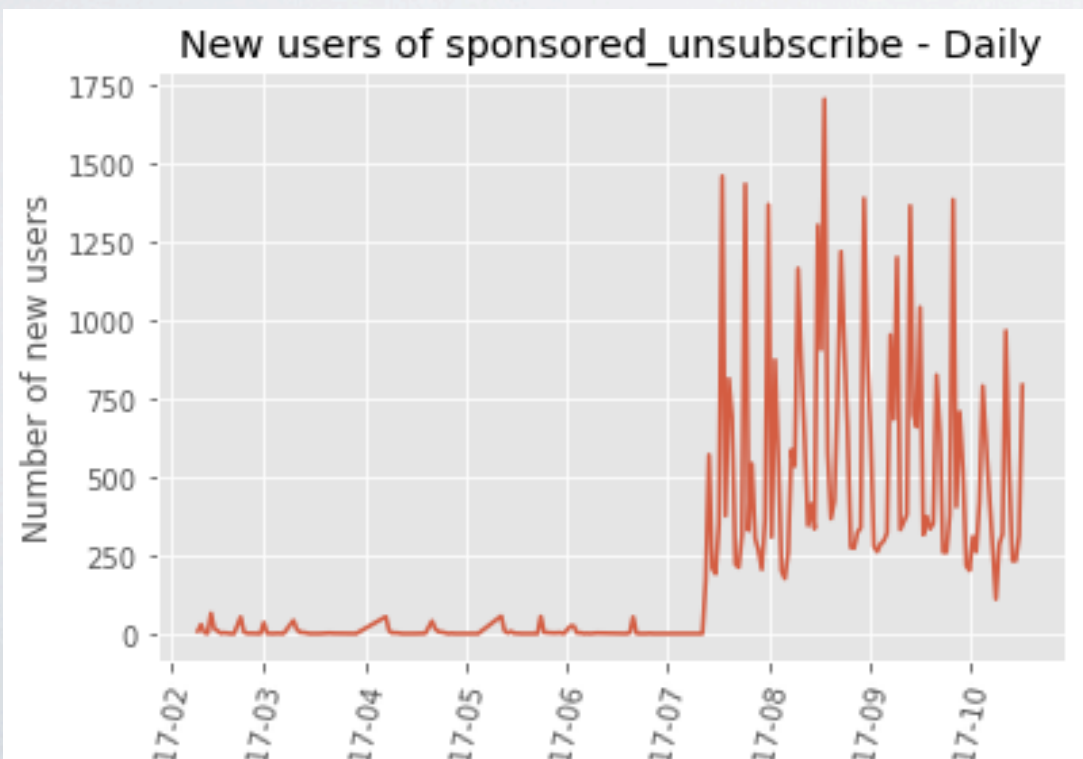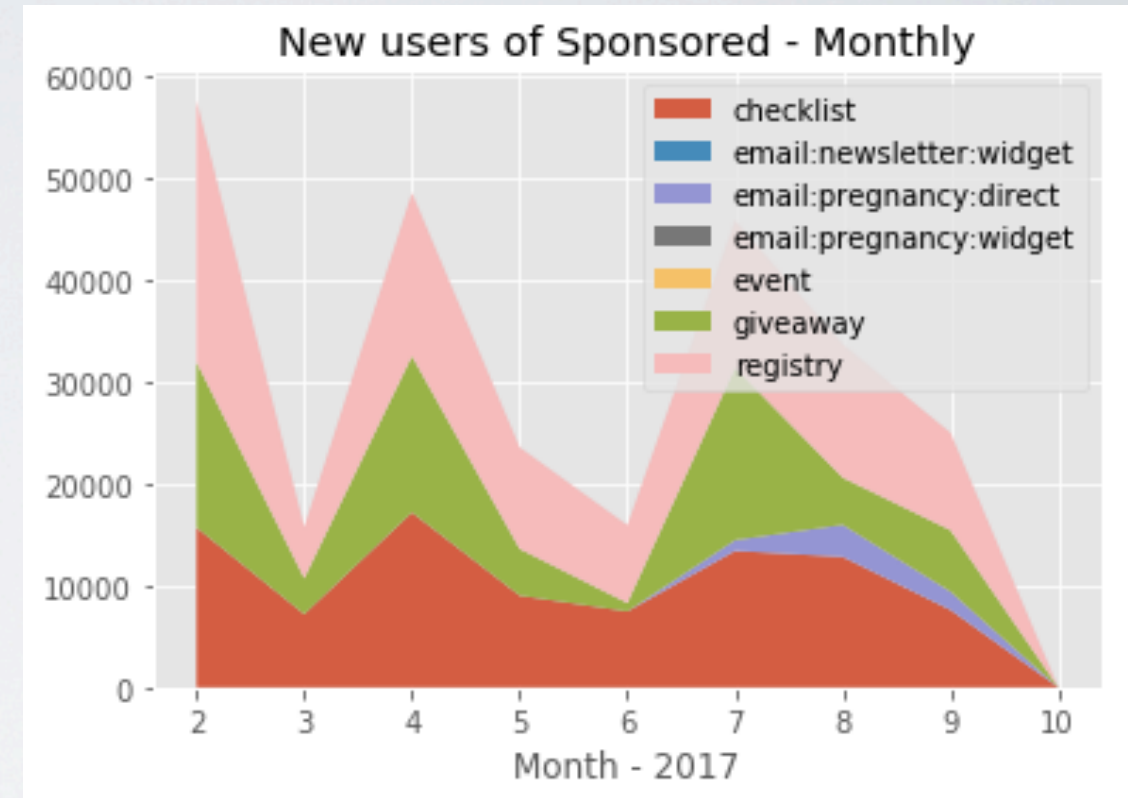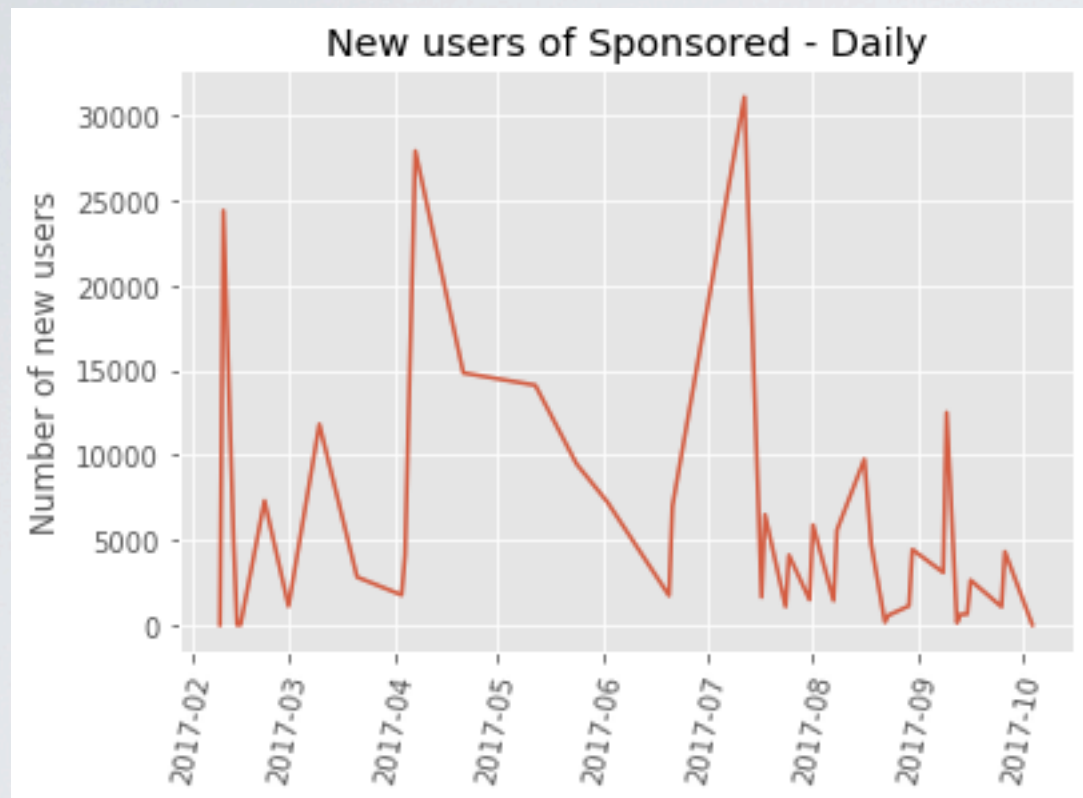
Outlines

EMAIL CHURN - RELATED QUESTIONS

- Can you describe how our email list(s) are churning?

- What factors correlate the most to churn? Due date?Source? Anything else?

- How would you describe our worst cohort?

- How would you describe our best cohort?

Users will unsubscribe email at any stage. We use "Deliver rate", "Open rate" and "Click rate" as our main parameters.

Churn Process

| Register | → | Deliver | → | Open | → | Click | → | Order |

| Deliver Frequency | Open rate | Click rate |
| = Deliver / Subscription Days | = Open / Deliver | = Click / Open |
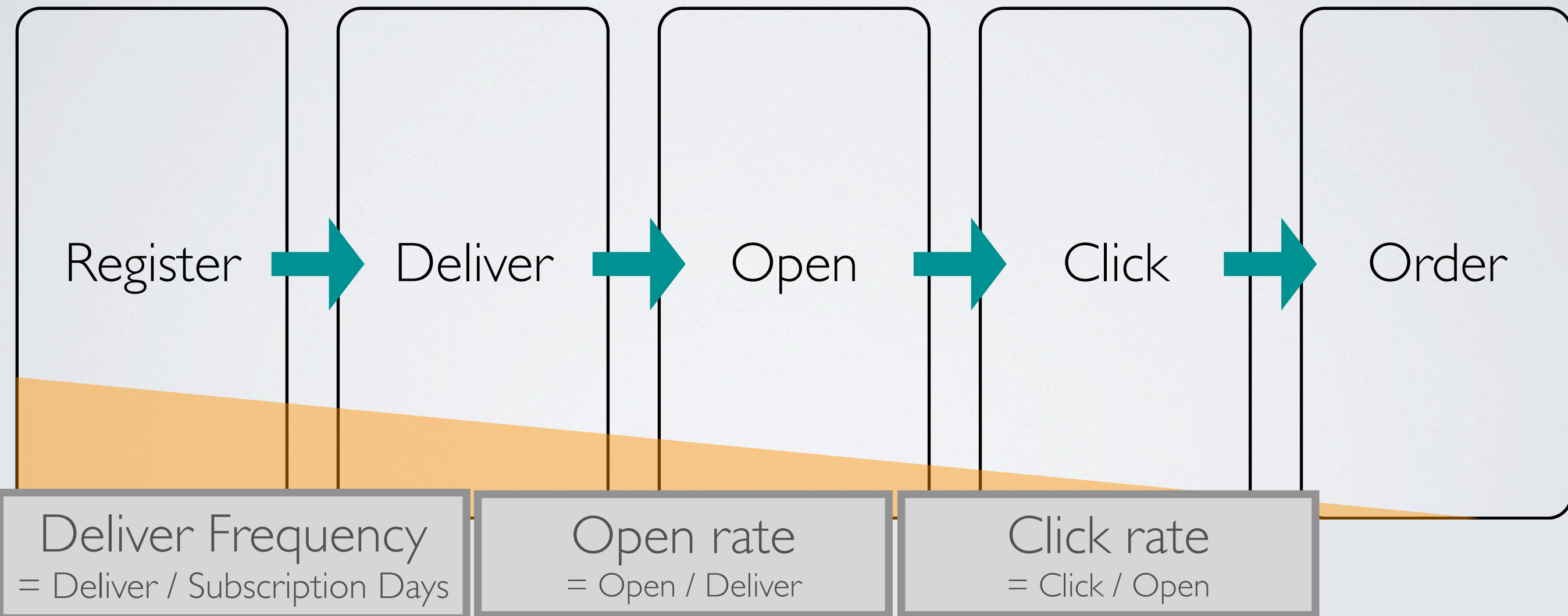
Goal: We want to segment users and show their profile in different groups with machine learning model (unsupervised)

Goal: Clustering (Expected)

Open Rate

Group A

Group B

Group C
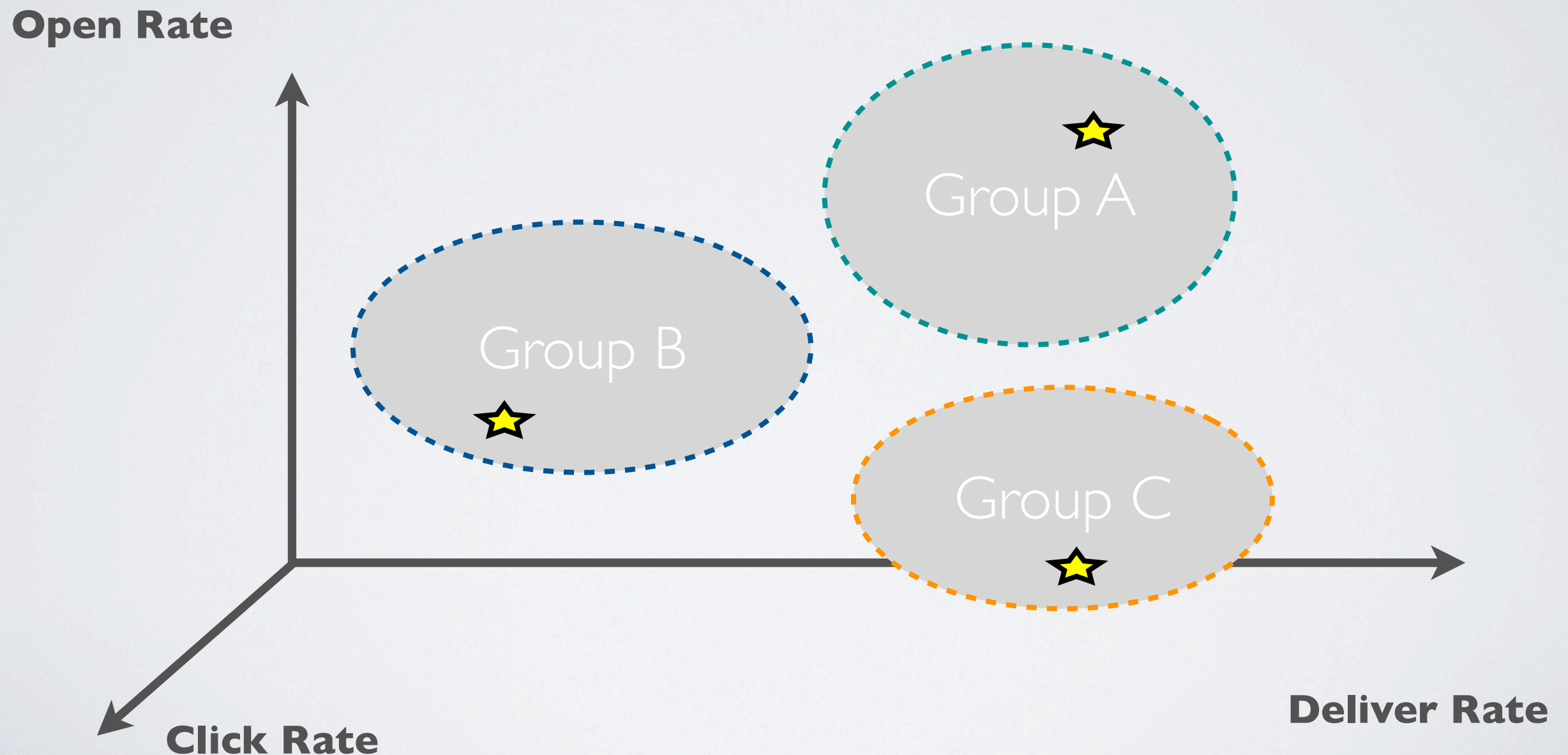
Click Rate

Deliver Rate

Model Settings: KMENAS Model unsupervised model with K = 3; 50-50 Testing / Training Data Set

Determain number of K

Train-Test split

Data Set

50%   50%

Training Data   Testing Data
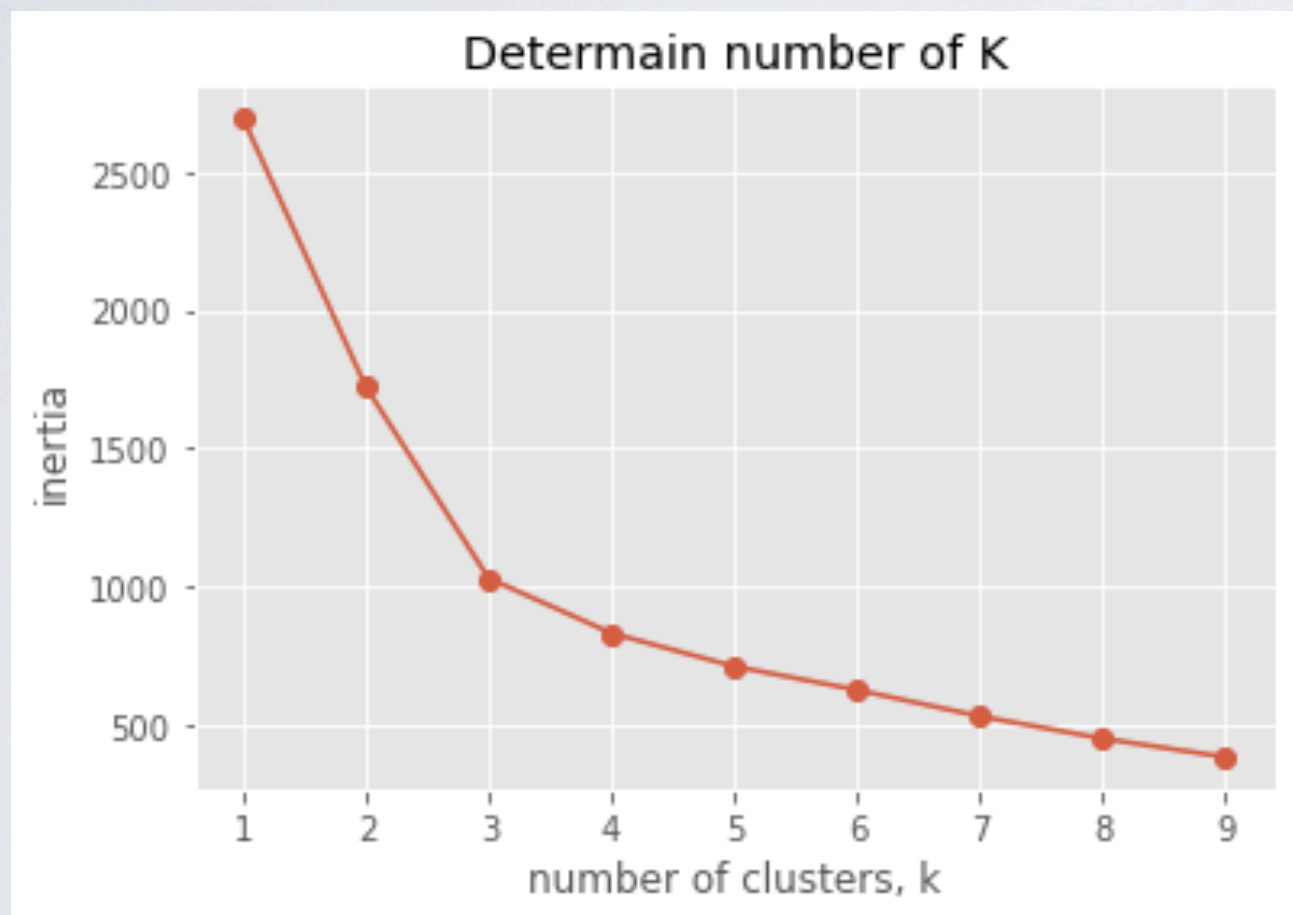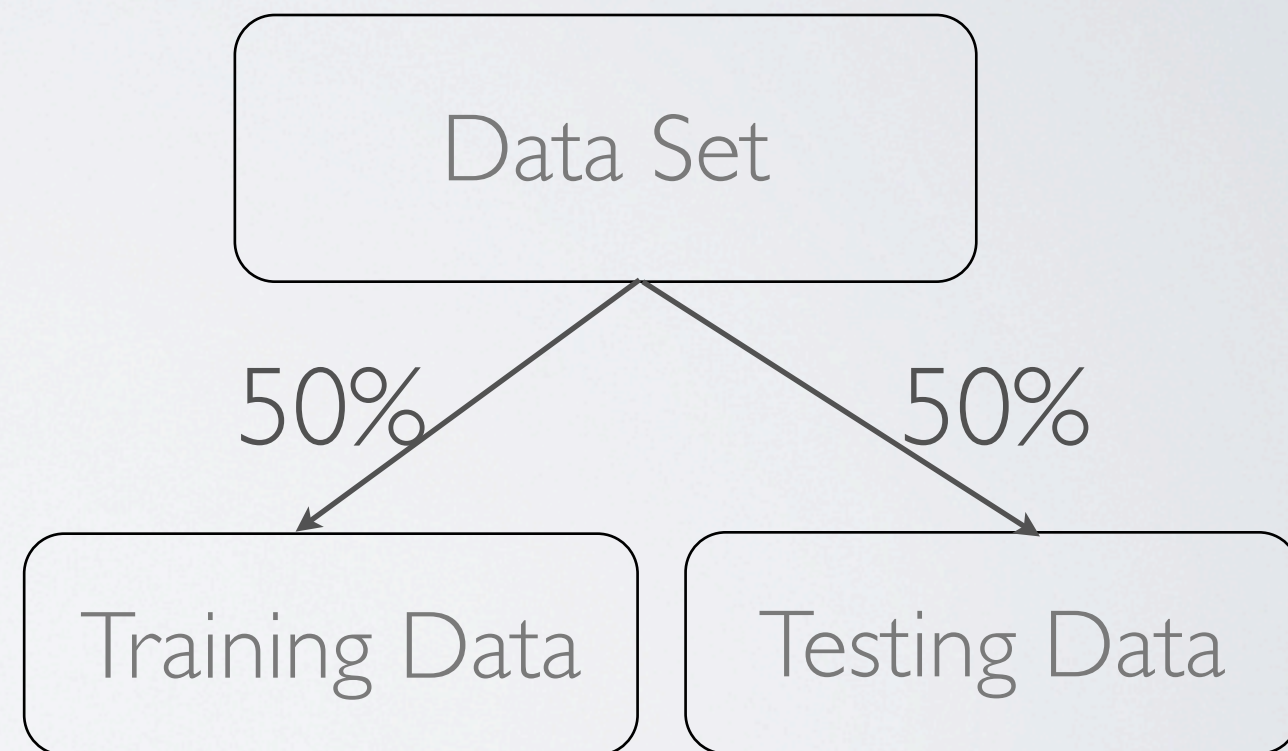
Deliver rate   Open rate   Click rate

Before modeling, we found several data need to be cleaned. I spent 50% of my time cleaning the data to make sure the quality of data

Data Cleaning

1. **Overwhelmed delivery**: Some users only register for 1 day, but got more than 10 times of delivery. We only keep deliver rate (deliver times / subscribe_days) < 1.
2. **No opened but clicked**: Some users didn't open an email, but click email. This may cause by some people using windows outlook, so they don't open the email but can click on the link. For those users, we make the open times = click times.
3. **Subscribe again**: Some users' subscribe_at date is later than unsubscribe date. We remove those with negative subscribe_days. Those might be still active users. (Unsubscribe email but then subscribe again)
4. **Arrival after 10 years**: Some people's baby may arrive before the year 2000 or after the year 2020. We only keep arrival date that makes more sense.

From machine learning model, we can cluster users into 3 different of groups and get their centroid position

Machine learning model output
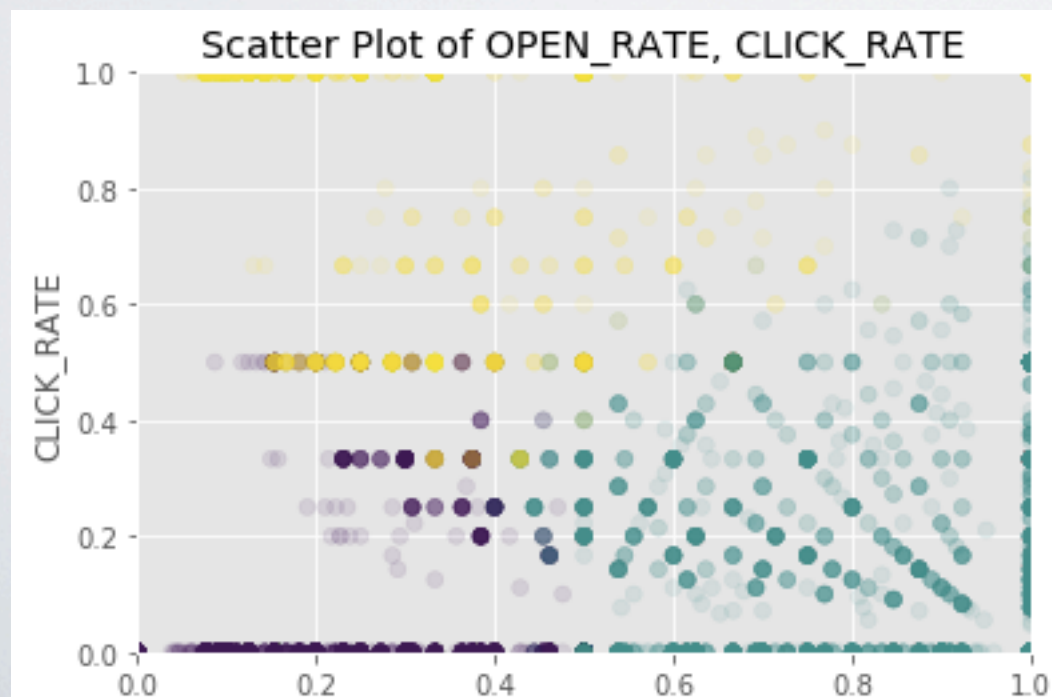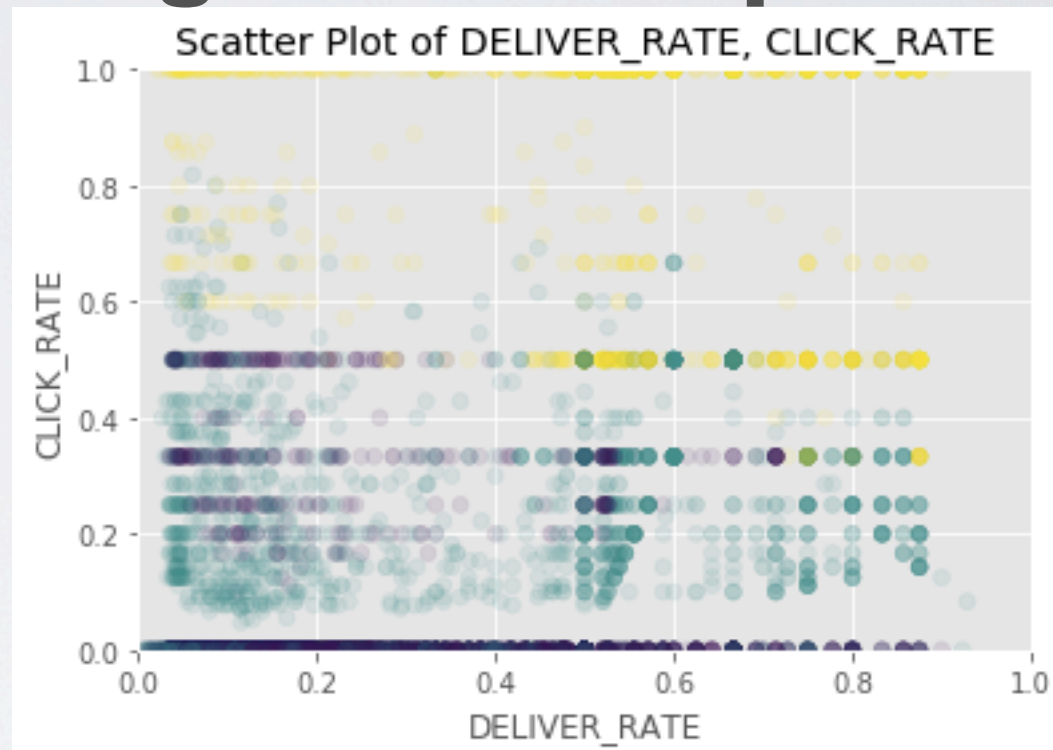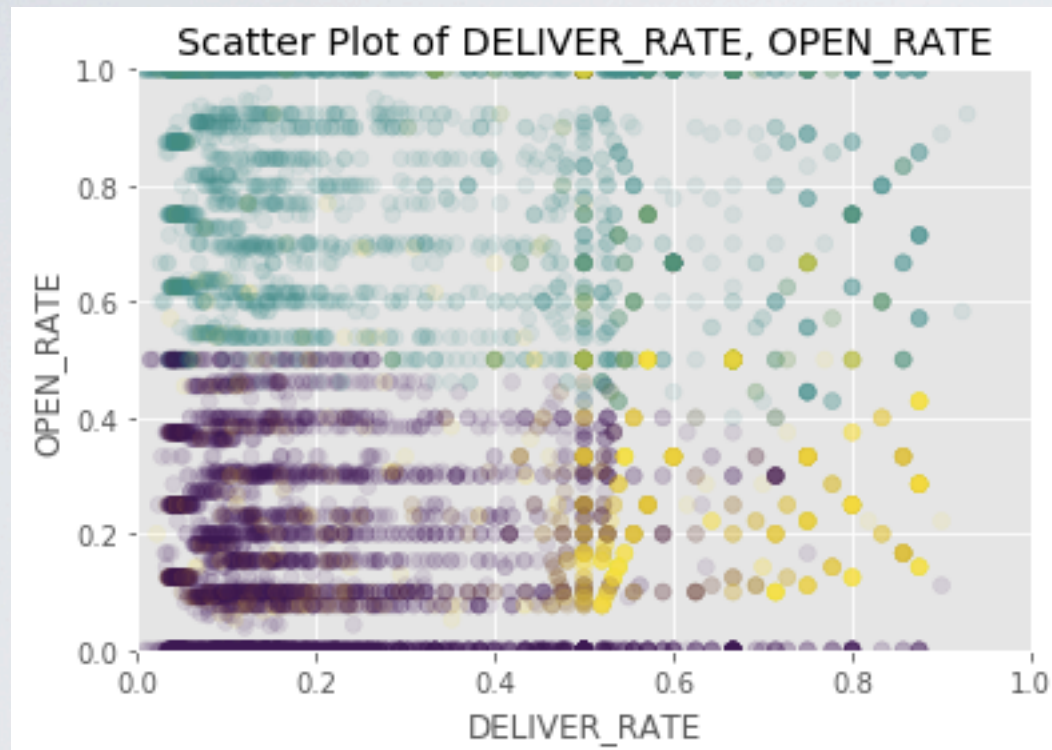
	DELIVER_RATE	OPEN_RATE	CLICK_RATE
0	0.228210	0.147665	0.030858
1	0.333370	0.817357	0.156528
2	0.552065	0.395322	0.908738

centroids

We found 3 types of users in Baby Registry 101.

User profile - Baby Registry 101

Type of User	Type 1	Type 2	Type 3
Deliver Frequency	0.23	0.33	0.55
Open Rate	0.14	0.81	0.39
Click Rate	0.03	0.16	0.91
Subscribe Days	82 Days ⭐	63 Days ⭐	25 Days ⭐
Urgency	126	122 ⭐	108 ⭐
Source	Registry (55%) Checklist (29%) Giveaway(16%)	Registry (60%) Checklist (28%) Giveaway(12%)	Registry (57%) Checklist (32%) Giveaway(9%)
Transaction	9	10 ⭐	5 ⭐

According to the classifier, we could determine relative marketing action plan for new users.

Scatter plot of new users action plan

Outlines

GOAL: Build a model to predict if a user is brought profit or not base on the current information.

Original
transactional_delivered distribution
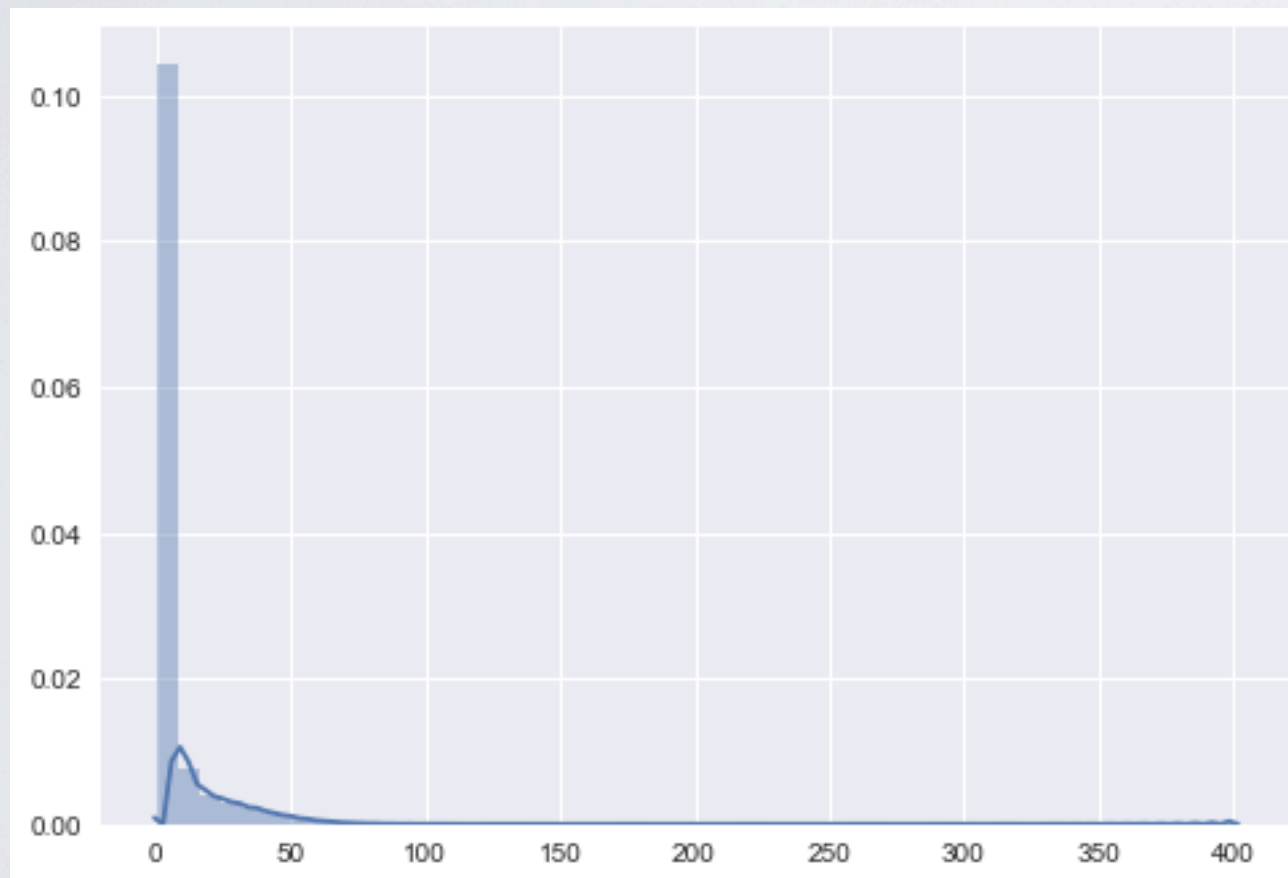
Transformed
transactional_delivered distribution
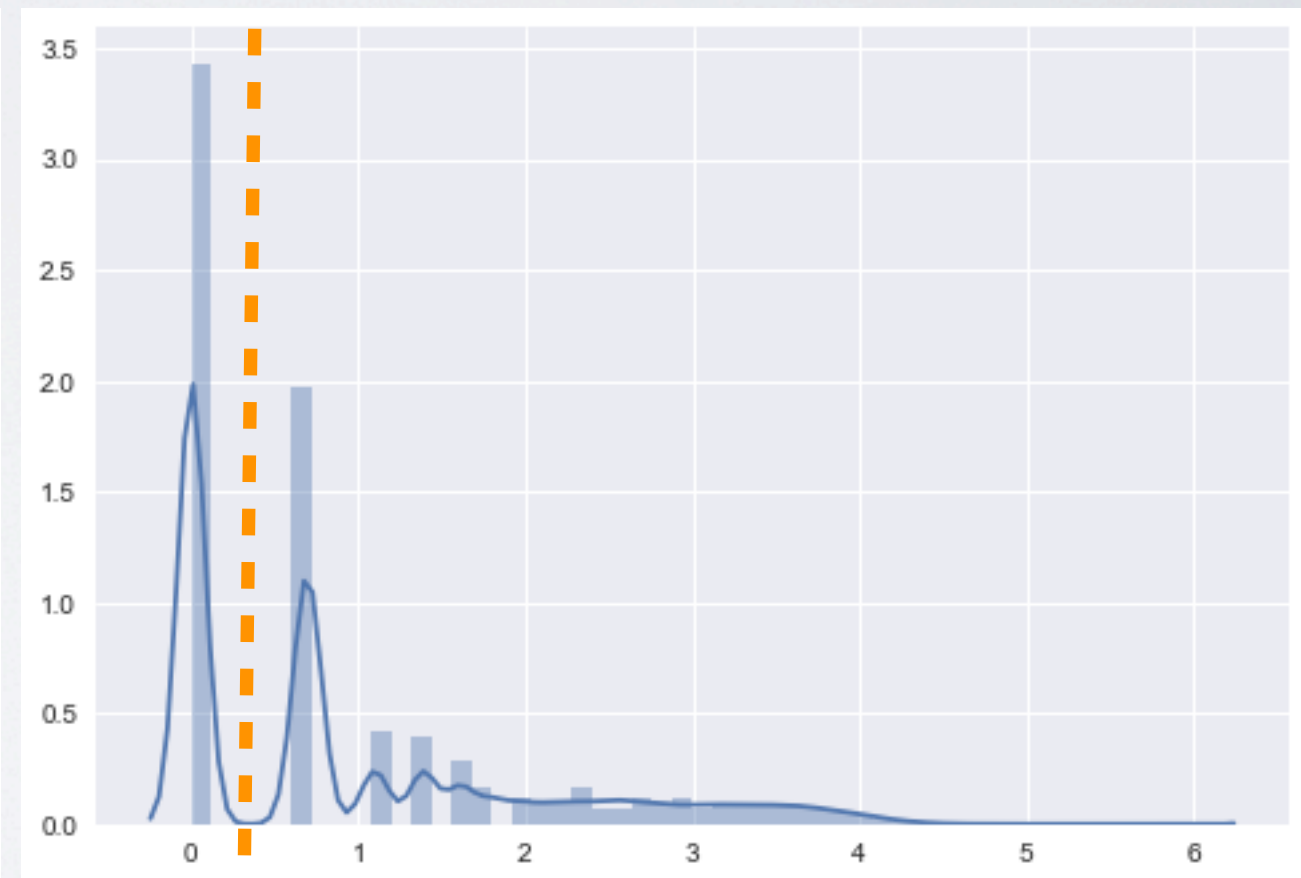
3: PROFIT PREDICTION MODEL

Here, we set diversity, urgency, and life-time as model variables, and profit/ not profit as predictor

Model Overview

Predictor	Variables
Not Profit (Deliver = 0)	1.Diversity : Subscribed E-mail lists
	2.Urgency : Original arrival date until now
Profit (Deliver > 0)	3.Life-time : How long did he/she be with us

Data

Training Data (Model Building)
Testing Data (Accuracy Rate)

3: PROFIT PREDICTION MODEL

We have 80% of accuracy and a great performance model to predict if our customer is profit or not.

Model accuracy result

Model	KNN	NB	Logistic regression	SVM
Accuracy Rate	0.83	0.63	0.81	N/A*
Run time	143 secs	0.27 secs	1.82 secs	5 hours

Note: Due to time and resource limit, I did only one time model building. For more accurate model, I suggest using pipeline tuning model and n-fold cross validation

Outlines

4: RECOMMENDATIONS

It would be great to have these recommendations in the future projects.

Recommendations

1. **Data cleaning** : This data base is merged with other data base, thus I recommend to add more data cleaning before modeling.

2. **Add transaction features** : Add more transaction data, product information into the data, so we could group users base on this information.

3. **Profit prediction model** : It would be great to have a project for profit prediction model.

Thank you!