Here is the metadata for each of files:

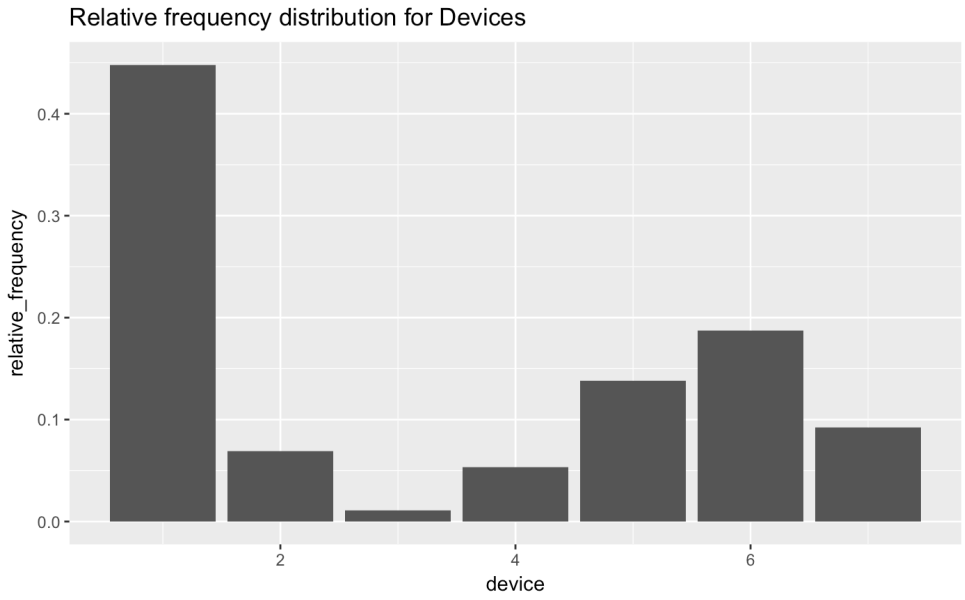
- Signups.csv
  - Uid – unique user_id (also the primary key)
  - Auth_type – there are 3 types; A,B,C. Let's assume this corresponds to whether the user signed up using social login or different email clients (gmail, yahoo etc)
  - Device – This is the device from which the user signed up. It has values from 1 – 7, each depicting some device type (eg. PC, ios app, android app etc.)
  - Signup_dt – Date when the user signed up

- Visits.csv
  - Uid – unique user_id
  - Dt – Date when the user visited our Houzz website/app

1) frequency distribution of users by different

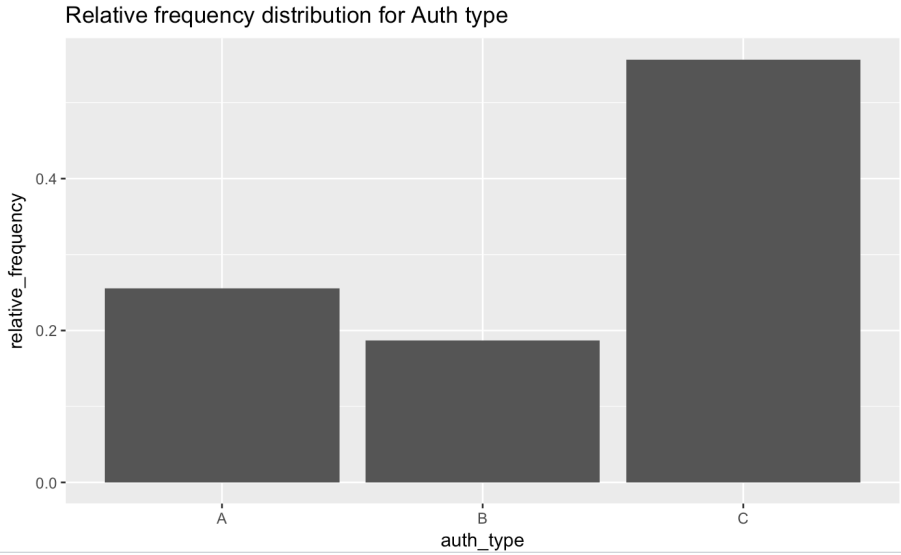
a. Devices

device <int>	cnt <int>	relative_frequency <dbl>
1	32245	0.44791565
2	4981	0.06919113
3	791	0.01098779
4	3859	0.05360541
5	9957	0.13831280
6	13497	0.18748698
7	6659	0.09250024



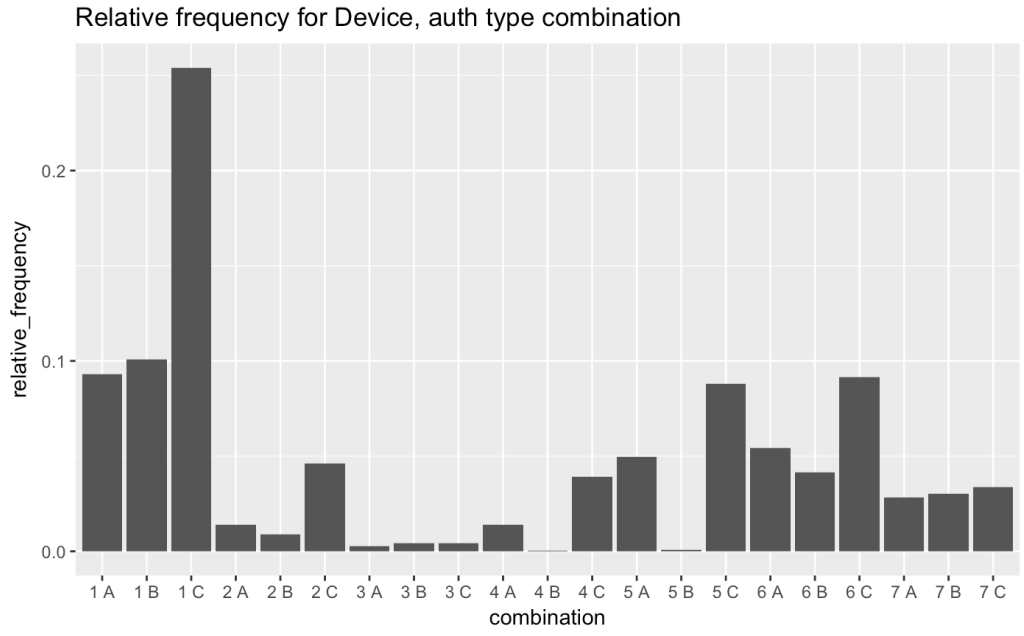
b. Auth type

auth_type <fctr>	cnt <int>	relative_frequency <dbl>
A	18409	0.2557196
B	13485	0.1873203
C	40095	0.5569601



c. Device, auth type combination

	device	auth_type	cnt	relative_frequency
1	1	A	6693	0.0929725375
2	1	B	7267	0.1009459779
3	1	C	18285	0.2539971385
4	2	A	1004	0.0139465752
5	2	B	643	0.0089319202
6	2	C	3334	0.0463126311
7	3	A	187	0.0025976191
8	3	B	309	0.0042923224
9	3	C	295	0.0040978483
10	4	A	1007	0.0139882482
11	4	B	22	0.0003056022
12	4	C	2830	0.0393115615
13	5	A	3571	0.0496048007
14	5	B	59	0.0008195697
15	5	C	6327	0.0878884274
16	6	A	3904	0.0542305074
17	6	B	2998	0.0416452514
18	6	C	6595	0.0916112184
19	7	A	2043	0.0283793357
20	7	B	2187	0.0303796413
21	7	C	2429	0.0337412660

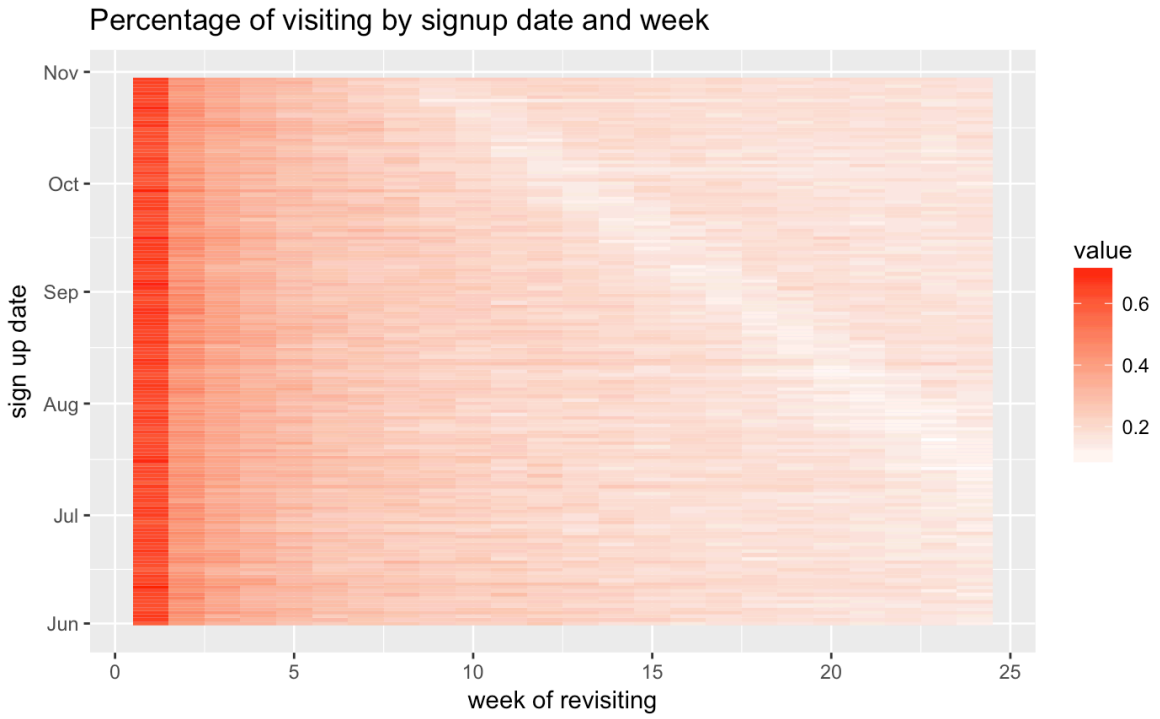


2) Let's build a table, and generate a heatmap – For customer retention rate

Part of the table:

	signup_dt	# signed up	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	2016-06-01	400	0.60	0.40	0.33	0.32	0.27	0.24	0.26	0.26	0.23	0.22	0.22	0.26	0.20	0.16	0.17	0.14	0.16	0.16	0.15	0.16	0.14	0.15	0.15	0.
2	2016-06-02	439	0.67	0.43	0.38	0.33	0.29	0.30	0.26	0.23	0.24	0.25	0.23	0.25	0.24	0.23	0.23	0.21	0.17	0.18	0.17	0.18	0.16	0.17	0.15	0.
3	2016-06-03	407	0.68	0.43	0.32	0.36	0.33	0.24	0.28	0.26	0.23	0.20	0.24	0.21	0.21	0.20	0.21	0.16	0.16	0.18	0.16	0.15	0.15	0.16	0.15	0.
4	2016-06-04	436	0.67	0.47	0.42	0.38	0.33	0.31	0.31	0.28	0.27	0.28	0.26	0.26	0.25	0.22	0.20	0.22	0.22	0.19	0.19	0.19	0.14	0.17	0.17	0.
5	2016-06-05	540	0.62	0.43	0.37	0.34	0.35	0.28	0.28	0.24	0.23	0.26	0.26	0.22	0.23	0.23	0.19	0.21	0.19	0.20	0.19	0.16	0.16	0.17	0.15	0.

Column '1' means percentage visiting in the 1st week
Column '2' means percentage visiting in the 2st week
…
Column '24' means percentage visiting in the 24st week



x-axis: Number of weeks after signing up
y-axis: Sign up date
fill: The darker, the higher percentage of customer visiting Houzz at the specific week after his/her sign up at the specific date.

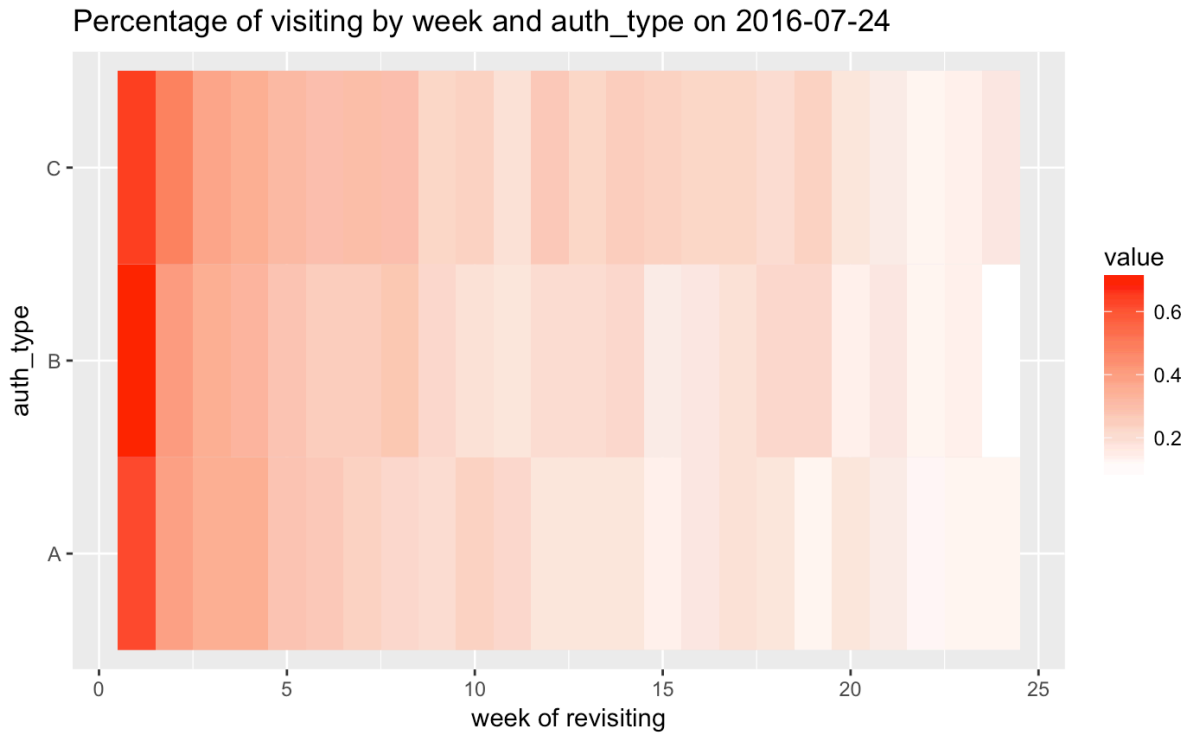
It reaches a steady state at the 10th week, before that time, the retention rate keeps decreasing (color becomes lighter) as week goes by. However, after the 10th week, we stop losing customer, those loyalty customers keep visiting us since then, and the number of them was relative steady.

In addition, there is a white slash in the heat map, which means the percentage of customers decreased sharply on that specific two weeks. The time of the white slash are Christmas week and it's following week, which explains the decrease of visiting.

3) Does the retention vary by different auth types? Let's build it for just two days – 24th July 2016 and 18th Aug 2016, and segmented by auth type.

For 2016-07-24:

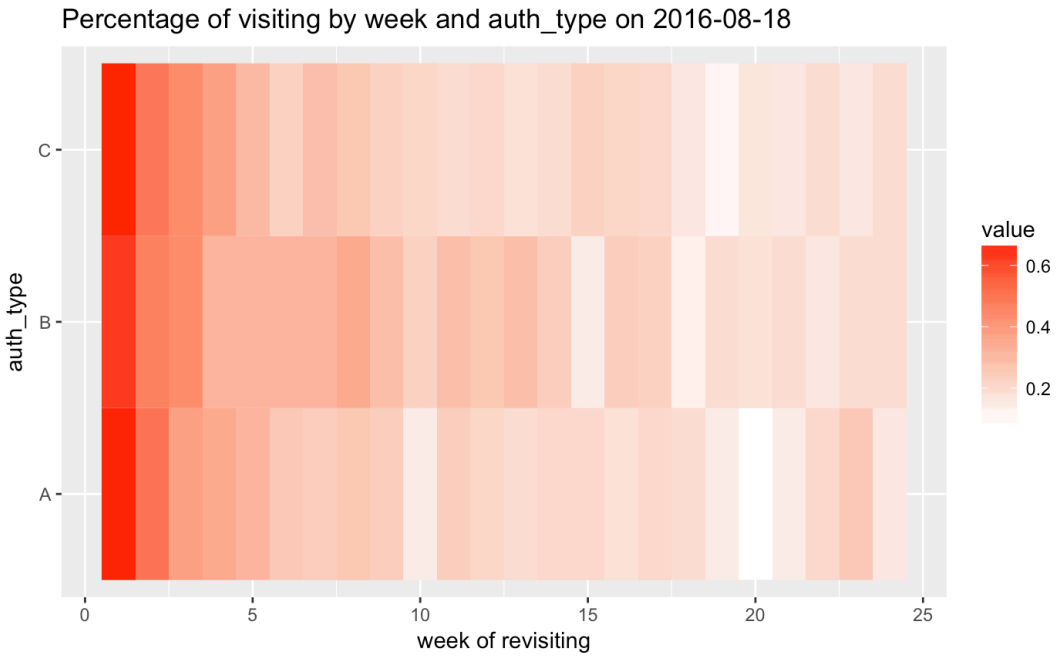
	auth_type	# signed up	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	A	145	0.63	0.39	0.35	0.35	0.28	0.26	0.24	0.21	0.20	0.23	0.21	0.17	0.17	0.17	0.14	0.16	0.19	0.17	0.12	0.17	0.15	0.11	0.12	0.12
2	B	102	0.72	0.41	0.35	0.33	0.28	0.25	0.25	0.27	0.21	0.19	0.17	0.20	0.20	0.21	0.15	0.16	0.19	0.21	0.21	0.14	0.16	0.12	0.14	0.09
3	C	283	0.65	0.49	0.38	0.35	0.31	0.29	0.30	0.29	0.22	0.23	0.19	0.26	0.22	0.25	0.23	0.22	0.22	0.20	0.23	0.17	0.15	0.12	0.13	0.16



Type C achieves the highest retention rate, especially for the first 20 weeks.

For 2016-08-18:

	auth_type	# signed up	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	2
1	A	102	0.69	0.51	0.38	0.35	0.32	0.25	0.24	0.26	0.24	0.14	0.23	0.21	0.19	0.20	0.20	0.17	0.20	0.18	0.14	0.08	0.14	0.20	0.25	
2	B	78	0.64	0.47	0.44	0.32	0.31	0.32	0.32	0.35	0.29	0.22	0.29	0.26	0.29	0.24	0.14	0.23	0.22	0.13	0.18	0.17	0.18	0.15	0.18	
3	C	229	0.70	0.49	0.44	0.38	0.30	0.22	0.28	0.26	0.22	0.21	0.19	0.20	0.17	0.18	0.22	0.21	0.20	0.15	0.10	0.16	0.15	0.18	0.15	



Type B achieves the highest retention rate, especially for the first 20 weeks.

Therefore, the retention varies by different auth types in different sign up date.

4) Let’s make a small twist to 2) - For users who signed up on Jun 1 2016, what proportion of them came back after signing up, for the first time within 1 week (Jun 2 – Jun 8 2016), first time within 2 weeks (Jun 2 – Jun 15 2016) , first time within 3 weeks (Jun 2 – 22 2016) etc. all the way up to within 24 weeks. Include all the signup dates until Oct 30 2016, such that you fill this table – On an avg, what proportion of users don’t come back even after 24 weeks?

Answer:
Part of the table:

	signup_dt	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	# signed up	percent_no_retur
1	2016-06-01	0.60	0.08	0.03	0.03	0.02	0.01	0.01	0.02	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	400	0.1
2	2016-06-02	0.67	0.06	0.05	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	439	0.1
3	2016-06-03	0.68	0.09	0.03	0.02	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	407	0.1
4	2016-06-04	0.67	0.08	0.03	0.03	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	436	0.1
5	2016-06-05	0.62	0.06	0.03	0.03	0.03	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	540	0.1

Column ‘1’ means percentage visiting first time in the 1st week
Column ‘2’ means percentage visiting first time in the 2st week
…
Column ‘24’ means percentage visiting first time in the 24st week
Column ‘percent_no_return’ means percentage of users don’t come back after 24 weeks’ time.

The whole table is saved in ‘question 4.csv’ file

The average proportion of users don’t come back even after 24 weeks’ time is 14.88%.