Output Report

SHANGWEN YAN | sy2160 | N17091204

Background

dataset:
the browsing records from mobile Telecom users in a specific period

purpose:
Process data to analyze user's favorate app, website, who spent most time online and the places where users surf most frequently, etc.

Methods and Outputs

1. Data Wragling

use R to do data wragling, codes in script "1.data_wragling.R". After dropping columns that will not be used and adding column names, we get the clean data in "./data/clean_data.csv", partly shown as below:

	A	B	C	D	E	F	G	H	I	J
1	region	user_ip	user_port	domain_nam	app_type	app_sub_ty	start_time	end_time	ul_data	dl_data
2	NA	10.83.124.	60914	559955.com	15	999	1.4096E+12	1.4096E+12	734	329
3	NA	10.83.124.	60914	559955.com	15	999	1.4096E+12	1.4096E+12	729	369
4	NA	10.83.124.	60915	559955.com	15	999	1.4096E+12	1.4096E+12	734	329
5	NA	10.83.124.	60915	559955.com	15	999	1.4096E+12	1.4096E+12	731	369
6	NA	10.83.124.	60916	559955.com	15	999	1.4096E+12	1.4096E+12	734	329
7	NA	10.83.124.	60916	559955.com	15	999	1.4096E+12	1.4096E+12	731	369
8	NA	10.83.124.	60917	559955.com	15	999	1.4096E+12	1.4096E+12	733	329
9	NA	10.83.124.	60917	559955.com	15	999	1.4096E+12	1.4096E+12	621	265
10	NA	10.83.124.	60917	559955.com	15	999	1.4096E+12	1.4096E+12	727	369
11	NA	10.83.124.	60918	559955.com	15	999	1.4096E+12	1.4096E+12	733	329
12	NA	10.83.124.	60918	559955.com	15	999	1.4096E+12	1.4096E+12	730	368
13	NA	10.83.124.	60919	559955.com	15	999	1.4096E+12	1.4096E+12	732	327
14	NA	10.83.124.	60919	559955.com	15	999	1.4096E+12	1.4096E+12	731	369
15	NA	10.83.124.	60920	559955.com	15	999	1.4096E+12	1.4096E+12	731	327
16	NA	10.83.124.	60920	559955.com	15	999	1.4096E+12	1.4096E+12	731	369
17	133833226	10.95.140.	34614	shop.m.taol	15	3	1.4096E+12	1.4096E+12	1322	1121
18	133833226	10.95.140.	34614	shop.m.taol	15	3	1.4096E+12	1.4096E+12	1222	1021
19	133833226	10.95.140.	34614	shop.m.taol	15	3	1.4096E+12	1.4096E+12	1305	1102
20	NA	10.83.106.	32809	pass.koube	18	184	1.4096E+12	1.4096E+12	990	1890
21	NA	10.83.106.	32812	pass.aliyu	18	100	1.4096E+12	1.4096E+12	990	1889
22	NA	10.83.106.	32814	pass.xiami.	6	10	1.4096E+12	1.4096E+12	989	1876
23	NA	10.77.47.1	38724	ju.mmstat.	15	246	1.4096E+12	1.4096E+12	1213	1093
24	NA	10.77.47.1	38740	log.mmstat.	15	246	1.4096E+12	1.4096E+12	1916	895
25	NA	10.78.159.	61705	log.mmstat.	15	246	1.4096E+12	1.4096E+12	1505	726
26	NA	10.78.159.	61721	log.mmstat.	15	246	1.4096E+12	1.4096E+12	1592	726
27	NA	10.78.200.	51913	log.mmstat.	15	246	1.4096E+12	1.4096E+12	1683	887
28	NA	10.78.200.	51949	log.mmstat.	15	246	1.4096E+12	1.4096E+12	1601	887
29	NA	10.90.78.1	4524	log.mmstat.	15	246	1.4096E+12	1.4096E+12	684	508
30	NA	10.77.6.72	55381	wwc.taobao	15	3	1.4096E+12	1.4096E+12	596	2261
31	NA	10.77.6.72	55381	wwc.taobao	15	3	1.4096E+12	1.4096E+12	536	3053

NA value in region means the region code is not available, and we will not use those rows when analyzing against region.

2. Analysis

use SQL to do data analysis, codes in script "2.data_analysis.sql". Write the output to several files in ./data/,partly shown as below:

(a) which region has more users (region, count(user))

132007946	96
131976249	71
133914142	58
133729802	52
134173204	51
134013450	48
134000394	46
131753738	45
133886484	37
133899028	37
131554836	37
131975452	37
131546890	36
131438397	35
131695370	32
131986708	32
131982643	32
131966238	31
133812756	31
132011284	31
131699230	30
131700500	29
133765140	28
131751454	28
133750046	28
131510794	27
201223225	26
133776660	26
131967498	24
131716412	24

(b) which region uses more traffic during this period (region, traffic/KB)

131982121	13183605
131758652	5568411
133951498	5137521
132007946	4707168
133899028	4302994
131753738	3821016
133942302	2996821
134158100	2833674
133776660	2610127
133900574	2555552
131574558	2329408
201222704	2224536
133732874	1946638
133871646	1843972
133709588	1611007
131533360	1258541
131667988	1251014
133726474	1184898
134013450	1155554
131674644	1150789
131524638	1130178
131512340	1072308
131637514	1063600
131685396	892518
133754644	892393
131699230	854813
131547422	851485
133759242	846009
133766164	783567
133903626	732944

(c) which region's users spent more time during this period (region, time/s)

133808404	241178
134149386	206734
131972362	180750
133861652	174578
133994260	157956
131516682	157601
132007946	155112
131699230	149459
133785098	142600
134000394	139328
131741460	130893
131753738	130568
133822484	128257
131555614	123914
133720842	123154
131701524	119681
133606713	116300
131638814	110374
131523860	107558
131699210	102637
134207506	102405
131957002	100182
132014858	98247
131679006	96930
133827358	96859
131736084	88673
131976249	88538
131967498	82477
131724574	81555
133979934	81508

(d) which kind of apps are more popular, that use more traffic (app_type, traffic/KB)

15	199199237
5	32648267
1	29494529
7	13980994
18	11212622
4	8074811
2	3277654
6	2226690
3	382549
12	348182
17	231554
8	116736
11	62363
9	60031
13	51628
16	5434
0	0

(e) which kind of apps are more popular, that users spent more time on it (app_type, time/s)

15	11996671
1	6870269
4	1878377
2	1821030
18	1429381
5	1091559
6	707087
7	530656
8	308025
17	194343
12	153841
3	104417
9	68832
11	63451
13	27972
16	5997
0	0

(f) for app_type=1, indicating instant messaging，which sub_apps people use most(app_sub_type,traffic/KB)

```
5       24806724
9       4384381
114     213868
53      39045
26      28828
8       13718
80      4627
33      1247
76      1199
79      892
```

(g) for app_type=1, indicating instant messaging，which sub_apps that users spent more time on it (app_sub_type, time/s)

```
5       5251408
9       1344052
114     123180
8       118405
53      26890
26      4239
80      1184
76      495
79      236
33      180
```

(h) which domain name uses most traffic(domain_name,traffic/KB)

```
www.icbc.com.cn 48981010
f1.market.xiaomi.com    30539403
apk.mmarket.com 13145209
223.82.245.66   10952802
183.218.12.138  8563718
apilocate.amap.com      7662055
vs4.wx.u3.ucweb.com:8080        7297329
120.203.229.32  7281186
gw.alicdn.com   6608221
120.203.230.150 5844825
122.72.99.82    5357409
223.82.254.211  5136415
122.72.99.80    3623392
a.ali.dongting.com:80   3321607
120.203.230.198 3286358
tyst.migu.cn    3264777
ww2.sinaimg.cn  3250943
static.miaozhiwei.net   2995372
218.205.79.88   2843042
asp.cntv.lxdns.com      2755699
120.203.230.200 2615662
ktv.a.yximgs.com        2610382
game.ds.qq.com  2572233
10.0.0.172      2555187
mmbiz.qpic.cn   2247200
223.82.254.217  2224536
223.82.254.210  2059381
120.203.229.3   1999835
223.82.245.67   1999204
122.72.99.84    1983633
```

(i) which domain name people spent more time on it(domain_name,time/s)

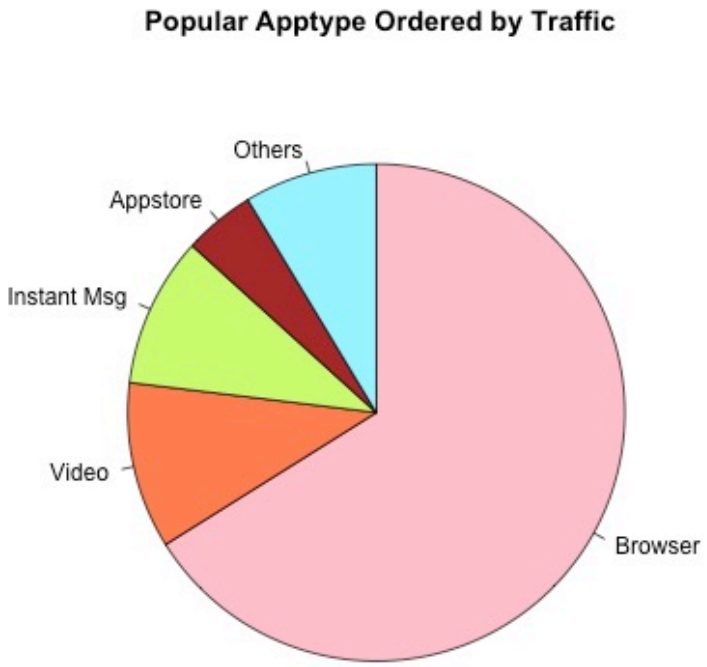
```
apilocate.amap.com       1586718
ichannel.snssdk.com      693513
api.m.taobao.com         691883
ossweb-img.qq.com        567054
qzonestyle.gtimg.cn      532366
ah.zsgjs.com     484583
ic.snssdk.com    417520
zxpic.gtimg.com 415913
proxy.music.qq.com       385230
gw.alicdn.com    328400
monitor.uu.qq.com        308260
gbres.dfcfw.com 271029
vs4.wx.u3.ucweb.com:8080         264529
nsclick.baidu.com        261276
mmsns.qpic.cn    247349
adash.m.taobao.com       224669
p-log.ykimg.com 221551
wfqqreader.3g.qq.com     212246
pingfore.qq.com 211216
mmbiz.qpic.cn    195909
d.ifengimg.com   195197
opensdk.uu.qq.com        191485
www.icbc.com.cn 184872
info.3g.qq.com   184123
btrace.qq.com    176785
10.0.0.172       173754
p.qpic.cn        168307
vs7.wxct.u3.gtm.ucweb.com        167958
u5.mm-img.com    167465
p1.pstatp.com    167197
```

3. Visualization: select three outputs to do the ploting

(d)

Plot: which kind of apps are more popular, that use more traffic (app_type, traffic/KB)

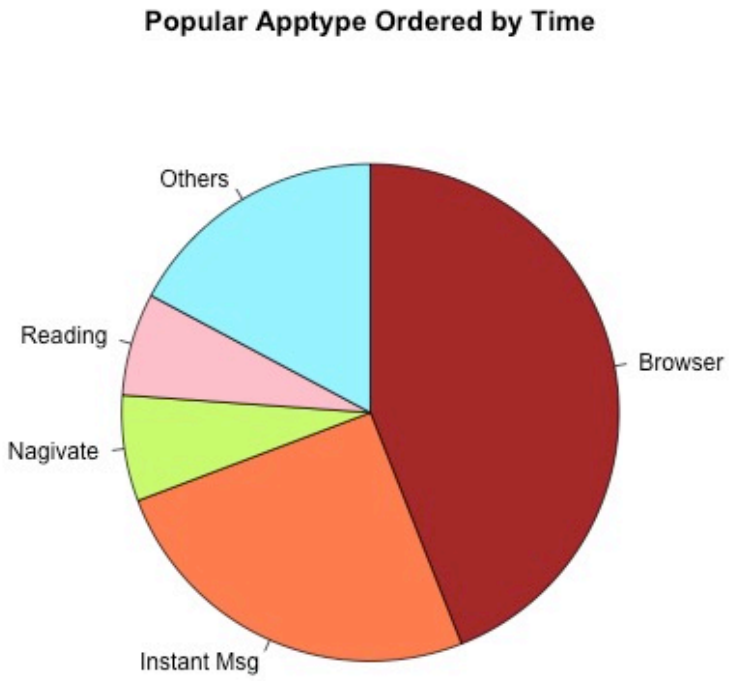
Analysis: People spend most traffic on browsing and watching videoes



(e)

Plot:which kind of apps are more popular, that users spent more time on it (app_type, time/s)

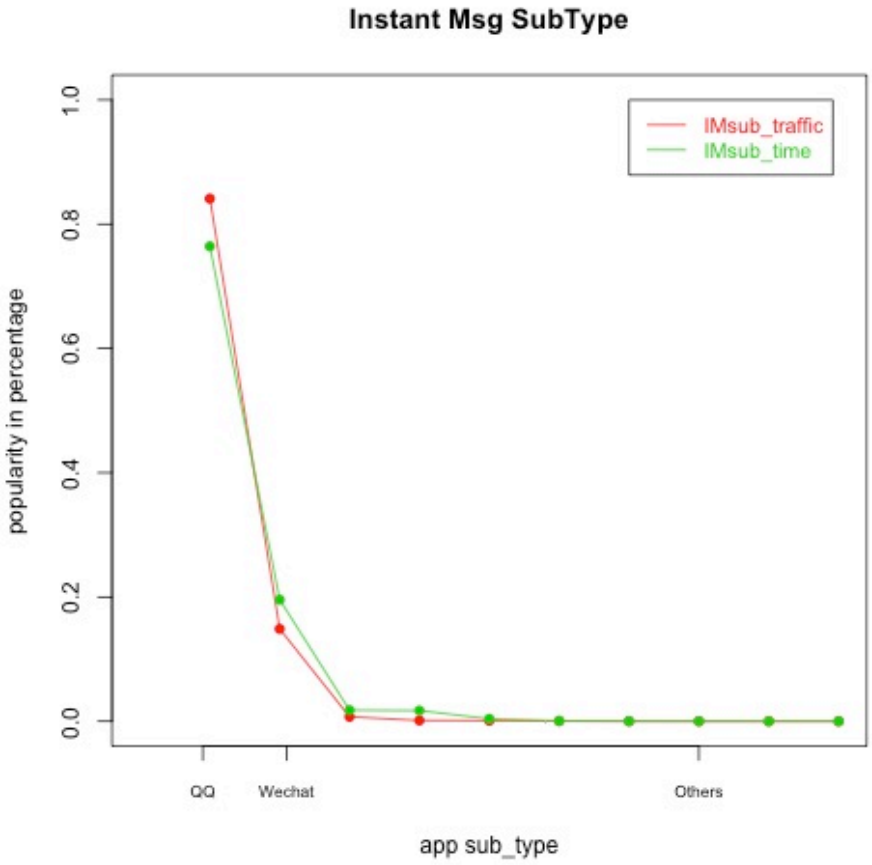
Analysis: Compared with the last picture, we can find people spend a lot of time on instant messaging, however it takes less traffic compared with browsing or watching video.



(f)

Plot:for app_type=1, which indicates instant msg， find the sub_apps people use most(two lines respectively stands for traffic and time)

Analysis: No matter using traffic or time, QQ and WeChat are the two most popular apps people would like to use to do instant messaging.


Instant Msg SubType
IMsub_traffic
IMsub_time
popularity in percentage
0.0
0.2
0.4
0.6
0.8
1.0
QQ
Wechat
Others
app sub_type