【余额宝快报】平台数据

研究报告II

未完，待更新

2013-10-07

李良军 liliangun88@gmail.com

摘要：【余额宝快报】是基于微信，提供余额宝（天弘增利宝货币基金）收益播报，计算，查询及常见问题回答的微信公共平台。本文是对该平台后台数据的分析与挖据，主要内容包括：1.对基金每日收益进行分析；2.对用户及其订阅及取消订阅行为进行分析；3.对消息及其种类，发送时间等信息进行分析；4.对每日新增关注人数进行回归分析；5.对每日消息数进行挖据分析。本文使用 R 语言作为工具，使用 ggplot2 包进行绘图，使用 knitr+markdown 进行文档编辑。

观点：

1. 快报男性用户是女性用户的 2 倍，男性更爱“理财”
2. 很多用户通过向【快报】提问，获取余额宝相关的信息
3. 【快报】每天的活跃用户数并没有随着用户的增加而增加，用户活跃度度相在降低

1 目录

2 概述

2.1 报告所涉数据来源

本报告所涉主要数据均来自【余额宝快报】微信公众平台所收集的数据，时间范围为 2013 年 7 月 23 日起至 2013 年 10 月 2 日止。

【余额宝快报】微信公共平台于 7 月 13 号注册，后台开发于 7 月 21 日完成，从 23 日开始产生完整的信息数据记录，所以本报告以 2013 年 7 月 23 日起至 2013 年 10 月 2 日止，77 个自然日的数据为样本。

2.2 为什么要做这样的报告分析

【余额宝快报】微信公共平台推出以来受到用户欢迎，目前累计用户超过 6000，每天处理 2000 次以上的用户互动查询，但用户及每天互动的消息数增长速度并未达到预期。通过本报告，了解细化用户的需求，了解他们的活跃时间，推出个性化的内容及服务。

3 平台功能描述

本报告主要分析余额宝收益信息及用户与平台互动的消息数据，其中，用户与平台的交互信息又可以细分为计算收益，查询收益，请求帮助信息，提问互动信息，订阅/取消订阅信息，异常信息等。

余额宝快报微信公共平台系统提供主要功能如下：

- 输入 T，查询天气
- 输入 C ，查询收益
- 输入金额，计算收益

- 输入 H，请求帮助
- 输入 W，提交问题或建议
- 输入未定义信息，属于其他
- 系统运行错误，属异常

后台主要的数据有：

- 余额宝每日收益信息
- 用户（订阅者）信息
- 用户与平台互动的消息
- 平台提供的常见问题帮助信息

为了便于分析，数据分析之前，作者已通过 R 语言，将 mysql 数据库中的数据导入 RData 文件中。

4 数据分析与可视化

本节主要从不同的角度展现基金收益，用户，消息信息的数据分布特点及时间走势，通过各类可视化工具展现数据背后的信息。

4.1 基金收益分析

4.1.1 数据整体分析

首先，看下基金收益的数据格式，基金样例数据为：

```
## day profit rate updatetime
## 1 2013-07-16 1.221 4.634 2013-07-16 00:00:00
## 2 2013-07-17 1.183 4.582 2013-07-17 22:17:35
## 3 2013-07-18 1.190 4.543 2013-07-18 19:16:03
## 4 2013-07-19 1.203 4.518 2013-07-19 19:05:59
## 5 2013-07-20 1.196 4.494 2013-07-20 08:10:25
```

6 2013-07-21 1.196 4.471 2013-07-21 00:04:50

其中：

- day 是日期数据，一天一份，数据从 2013-07-16 开始
- profit 是增利宝（余额宝基金公司）每日公布的每万份收益额
- rate 是增利宝每日公布的七日年化收益率
- updatetime 是该收益的更新时间

对万份收益做简单的分析：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.15    1.21    1.22    1.25    1.26 1.51
```

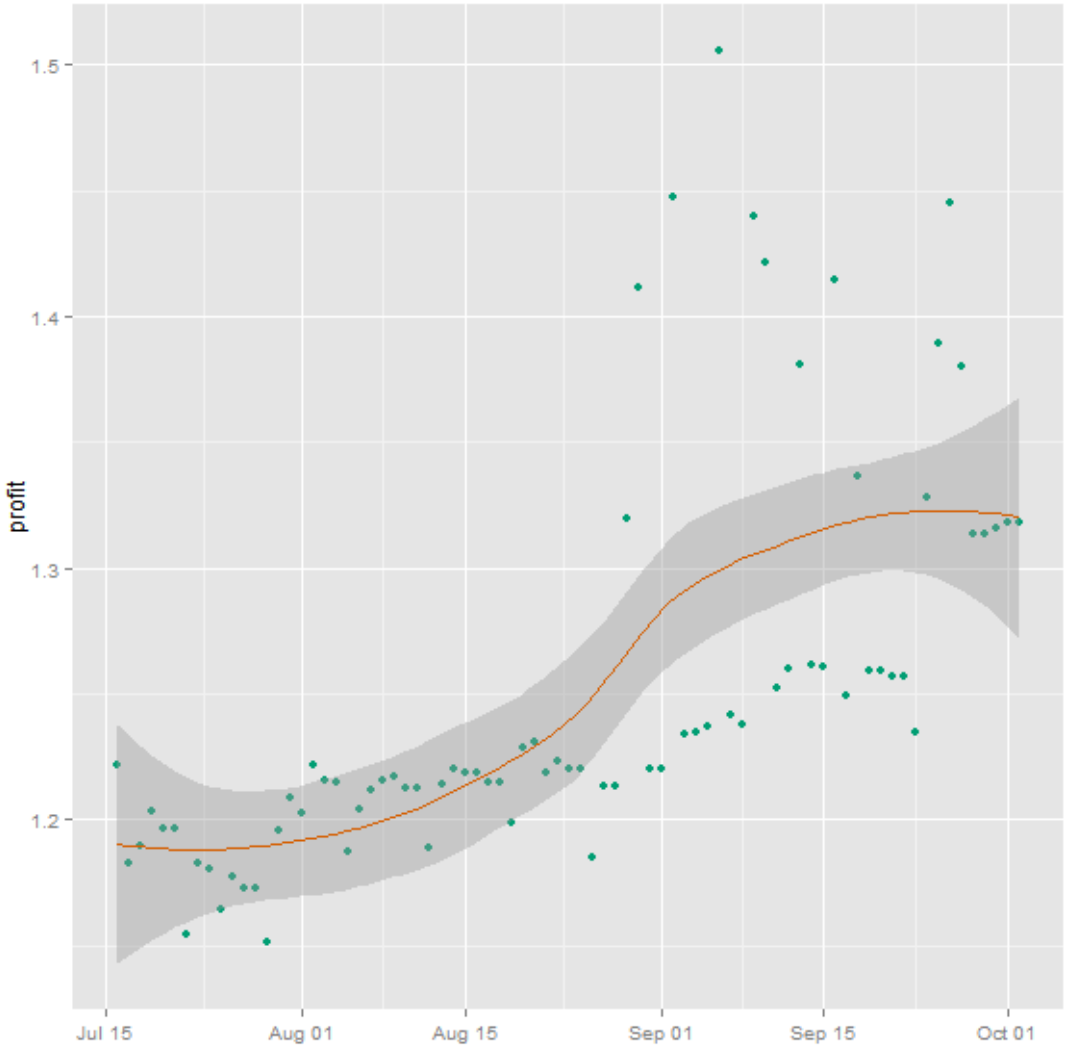
均值为 1.25 标准差为 0.7796

五分位为： 1.15 1.21 1.22 1.26 1.51

通过 shapiro 判断是否符合数据是否符合正态分布，p-value = 3.088e-08，明显小于 0.05，表明数据不符合正态分布。

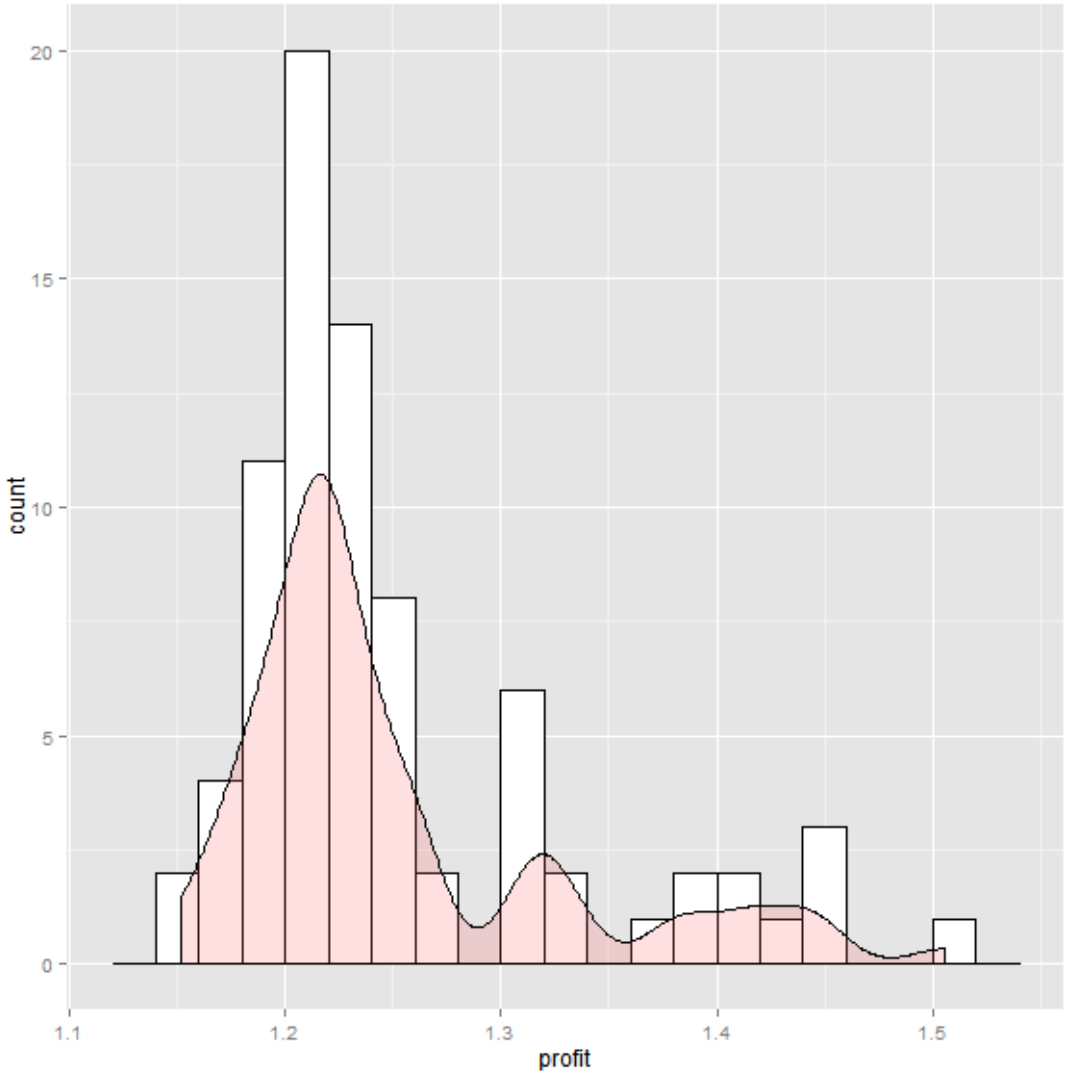
4.1.2 万分收益时间序列图

对万份收益做时间序列图，并通过 loess（局部加权回归散点平滑法 locally weighted scatterplot smoothing，LOWESS 或 LOESS）回归拟合，如下图所示：



图中阴影部分是置信度为 95%的置信区间值，可以看出在 8 月 25 日-9 月 25 日拟合较差，这期间收益涨跌幅度较大。

4.1.3 统计直方图



通过对万份收益数据的分布拟合，也表明样本数据不符合正态分布。

4.2 用户分析

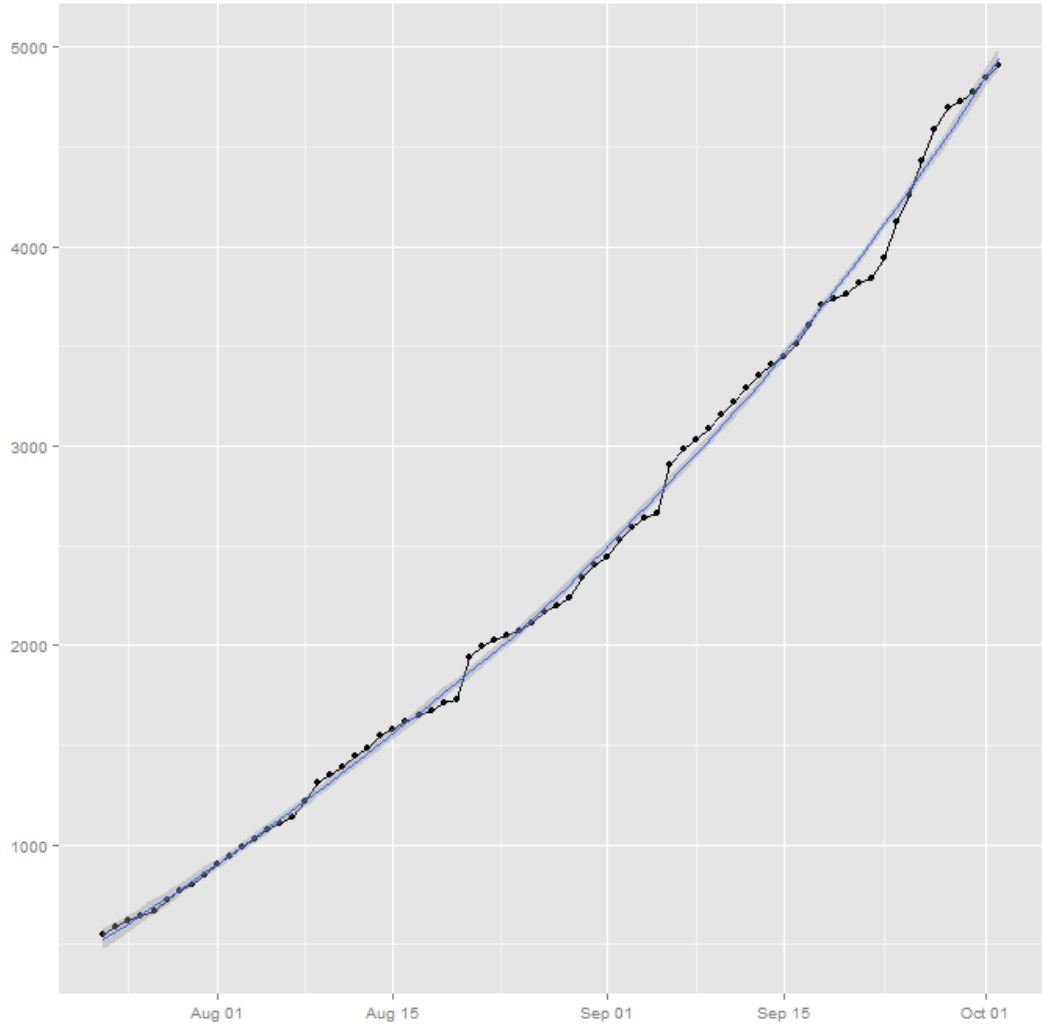
4.2.1 数据样例

```
##         day subscribe unsubscribe newuser totaluser
## 1 2013-07-23        50           9      41       547
## 2 2013-07-24        43           2      41       588
## 3 2013-07-25        38           3      35       623
## 4 2013-07-26        21           2      19       642
## 5 2013-07-27        36           9      27       669
## 6 2013-07-28        63           9      54       723
```

其中，

- day 是日期数据，一天一份，数据从 2013-07-23 开始
- subscribe 表示每日新关注人数
- unsubscribe 表示每日取消关注人数
- newuser 表示净增关注人数
- newuser totaluser 表示累计关注人数

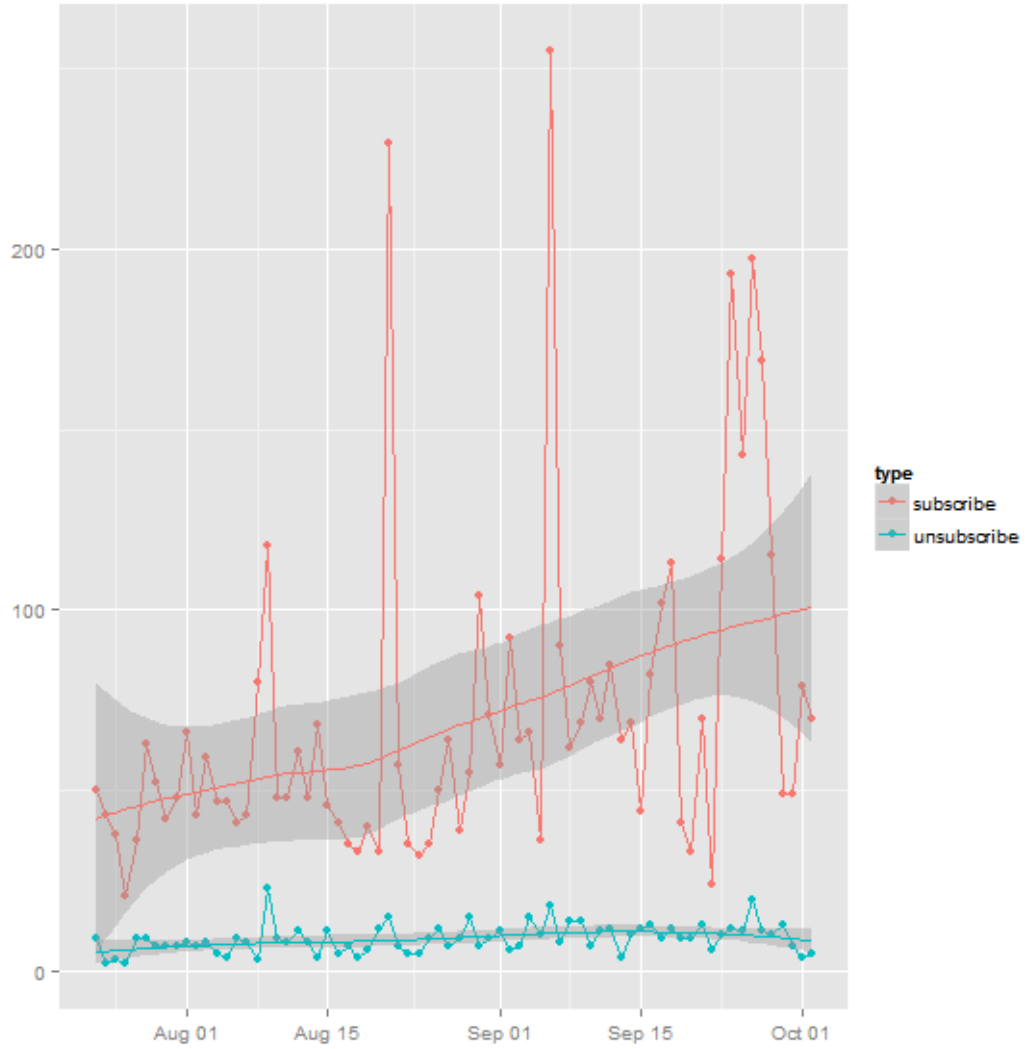
4.2.2 每日累计用户数时间序列图



可以看出，每日累计用户数在比较平缓稳定的增长，增加趋势也略有提升，但是增长幅度没有大幅提升。

4.2.3 每日新增关注人与取消关注用户对比

下图为将数据进行标准正态化后的比较，二者的关联关系较为明显。

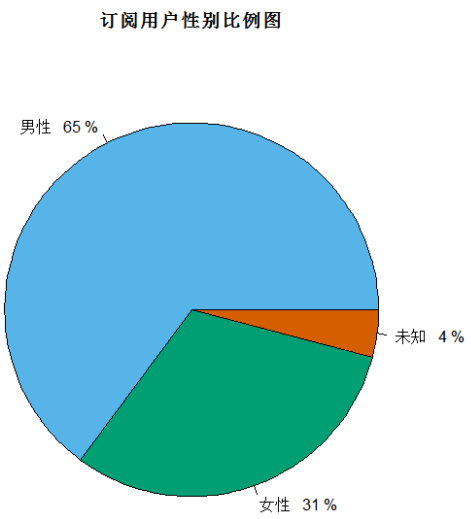


由上图可以大致看出，1）每日新增关注人数和每日取消关注人数有一定的关联关系，当新增用户数增加时，取消关注的人数也会增加；2）每日新增关注数（红色的拟合线）有增长的趋势，每日取消关注数在 9 月 15 号之后有下降的趋势，说明 9 月 15 号之后用户的满意度在提高。

系统在 9 月 15 号左右对部分功能进行了升级，调整了用户订阅后的欢迎语句，证明该调整是有明显效果的。

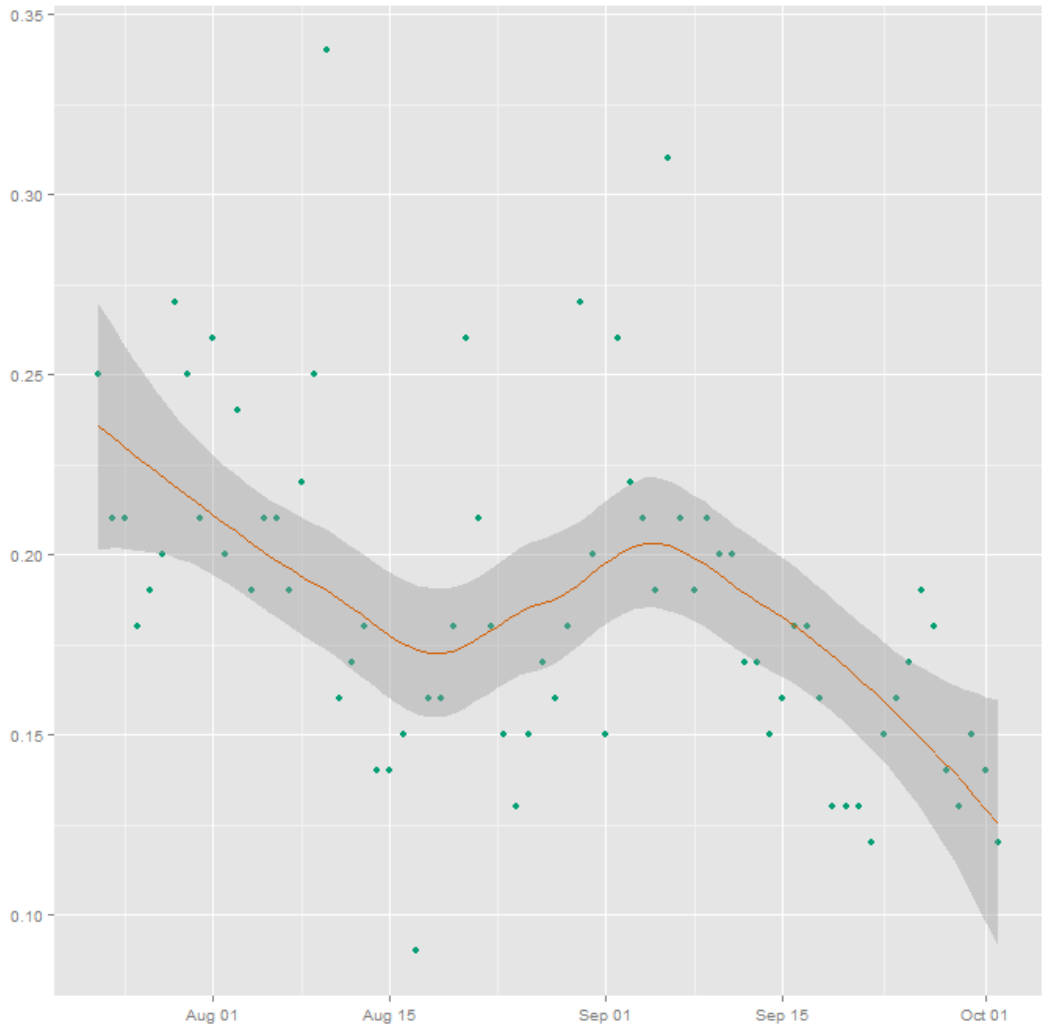
4.2.4 性别分析

所有关注者中，男性 3164 名，女性 1514 名，性别未知 199 名。男女比例为 2.1：1，详见下图：



4.2.5 用户活跃度分析

将每天发送消息 1 条及以上的用户定义为活跃用户，想将每天都活跃用户与当天累计用户的比值定义为用户活跃度。用户活跃度表征了用户对系统的使用频率及依赖程度。



有上图可以看出，用户的活跃度整体呈现向下滑动趋势，尽管在 8 月 18 号到 9 月 3 号之间有短暂上涨的趋势。

4.3 消息分析

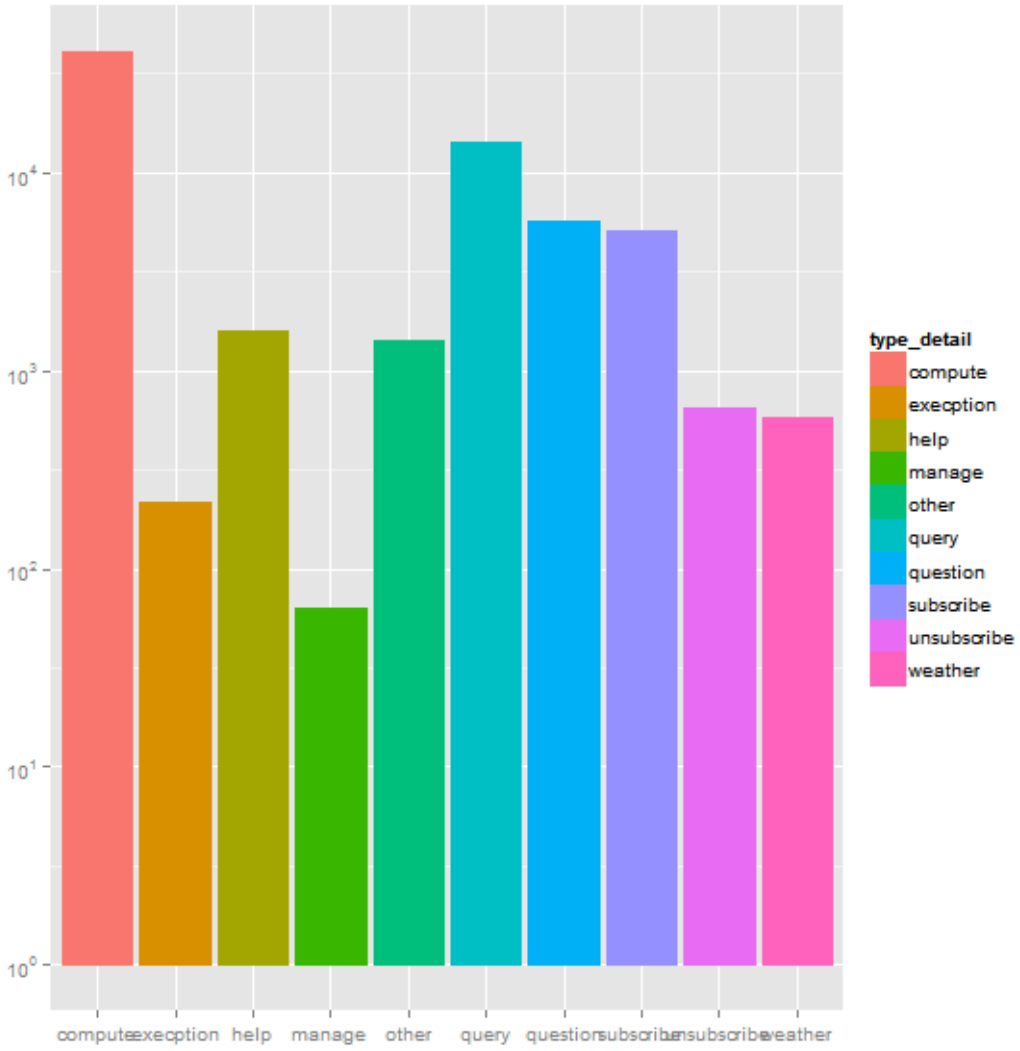
4.3.1 样例数据

```
##     id userid sex             msgtime        day hour msg_type type_detail
## 1  254      5   1 2013-07-23 00:24:03 2013-07-23    0        1       query
## 2  255      5   1 2013-07-23 00:38:35 2013-07-23    0        1       query
## 3  256      5   1 2013-07-23 00:38:41 2013-07-23    0        1       query
## 4  257      5   1 2013-07-23 00:45:03 2013-07-23    0        5 unsubscribe
## 5  258      5   1 2013-07-23 00:46:24 2013-07-23    0        5   subscribe
## 6  259     16   1 2013-07-23 01:46:28 2013-07-23    1        1     compute
## 7  260     16   1 2013-07-23 01:46:52 2013-07-23    1        1     compute
```

```
## 8  261    87   0 2013-07-23 03:30:59 2013-07-23    3       5  subscribe
## 9  262    88   0 2013-07-23 04:04:08 2013-07-23    4       5  subscribe
## 10 263     0   1 2013-07-23 05:01:19 2013-07-23    5       1    compute
```

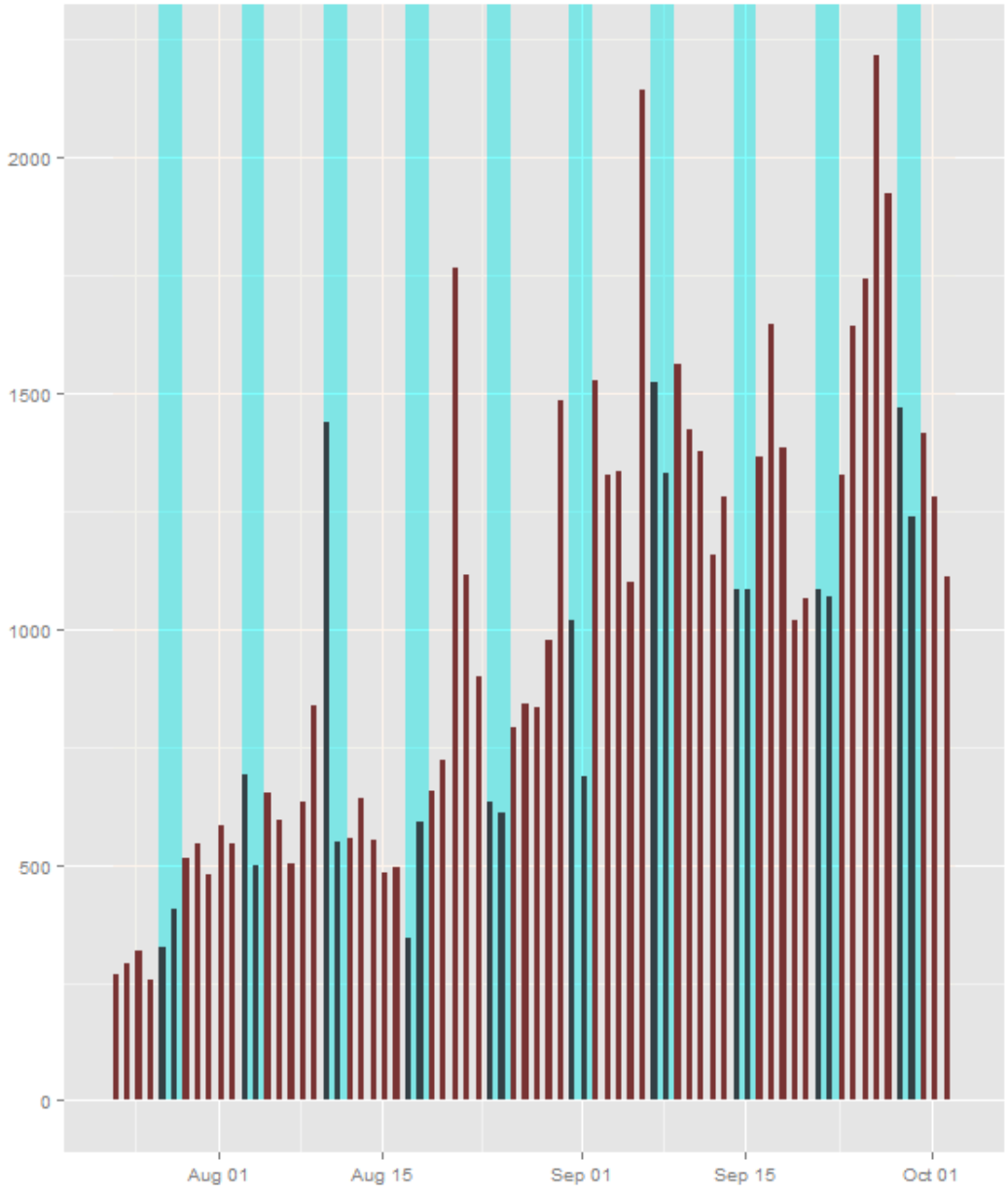
4.3.2 不同类型的消息数量对比

下图为不同类型的消息数量对图图，可以看出计算收益，查询收益是用户比较常用的 2 个功能。但排名第三位的消息类型是提问，可见，有大量的用户会通过快报提出余额宝相关的问题。

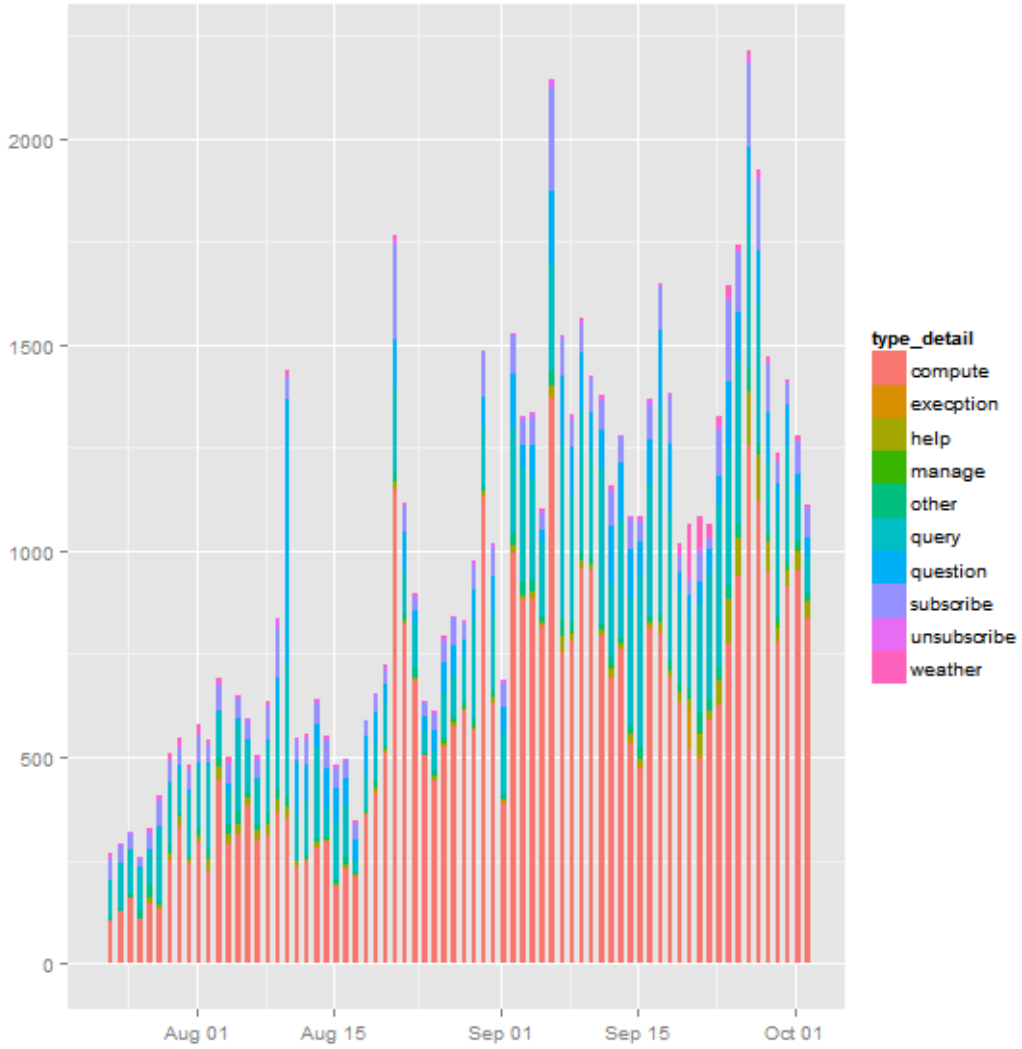


4.3.3 每天的消息数

下图为系统每天接收到消息数：



蓝色背景表示当天是周末，可以看到一般在周末，消息数量就会较少。下图为每天不同消息种类的对比图，也可以看出，计算和查询收益是 2 个常用的功能。



5 数据挖掘

本节，笔者将对部分数据进行挖掘，探索不同的变量之间是否存在着一定的关联关系。

5.1 每日新增关注人数分析

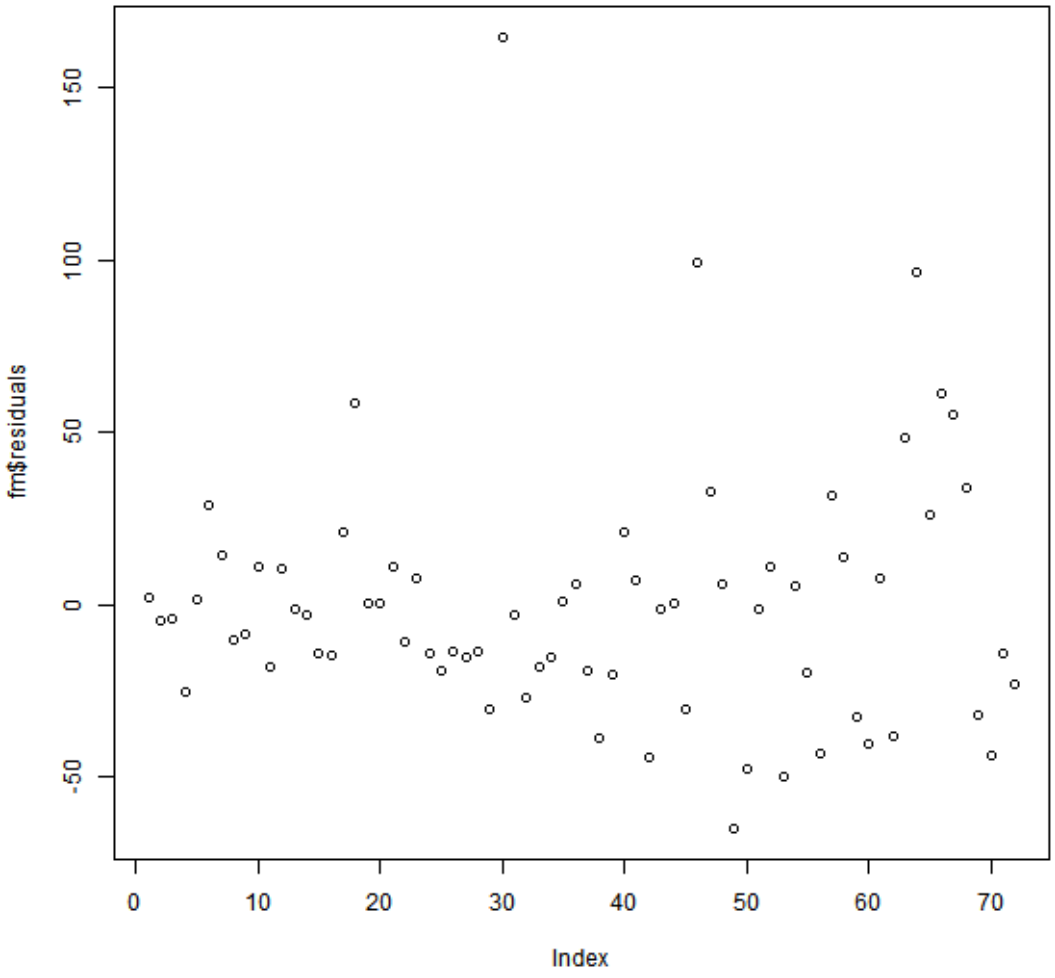
初步分析，每日新增人数和万分收益的时间走势图有一定的关联关系，和是否是周末也有一定的关系，下面先对新增关注人数做回归分析。

代码如下：

```
fm <- lm(formula = subscribe ~ profit + isweekend, data = subscribeanalysis)
summary(fm)
##
## Call:
## lm(formula = subscribe ~ profit + isweekend, data = subscribeanalysis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -64.9  -19.4   -3.0   10.9  164.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -345.85      72.13   -4.80  9.0e-06 ***
## profit          333.18      56.78    5.87  1.4e-07 ***
## isweekendTRUE   -10.52       9.96   -1.06     0.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.1 on 69 degrees of freedom
## Multiple R-squared:  0.364,  Adjusted R-squared:  0.346
## F-statistic: 19.7 on 2 and 69 DF,  p-value: 1.66e-07
```

分析结果可以看出，方差比例（R 平方值）为 34.6%，拟合值很低，表明用线性回归不合适或二者没有决定性的关联关系。

下图为残差图



残差图也显示，拟合很差，存在部分大的噪点，笔者通过去除噪点后分析，方差比例（R 平方值）为 36.4%，略有提高，但还是拟合很低。

通过分析有可以看出 isweekend 对减少模型误差的贡献最少，去除该因子继续做线性回归分析，方差比例（R 平方值）为 36.8%，去除是否是周末这个因子，对拟合效果并没有大的影响，反倒提高到方差比例。

决定每日新增用户数的因素很多，如订阅者向好友推荐，用户随机关注等都存在很大不确定性，很难对其做明确的分析结果。

5.2 对每天消息数做回归分析

初步分析，每日消息数和万分收益的时间走势图有一定的关联关系，和每日新增用户数也有一定的关系，下面对每日消息数做回归分析。

```
lm.msgana <- lm(message ~ profit + subscribe, data = msgana)
summary(lm.msgana)
##
## Call:
## lm(formula = message ~ profit + subscribe, data = msgana)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -451.9 -197.4  -55.5  152.6  695.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3404.200    553.784   -6.15  4.5e-08 ***
## profit       3246.395    464.898    6.98  1.4e-09 ***
## subscribe       4.359      0.802    5.44  7.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 249 on 69 degrees of freedom
## Multiple R-squared:  0.735,  Adjusted R-squared:  0.727
## F-statistic: 95.6 on 2 and 69 DF,  p-value: <2e-16
```

分析结果可以看出，方差比例（R 平方值）为 72.7%，线性回归公式为：

每日消息数 = 4.359 * 每日新增关注人数 +

3246.395 * 每日万分收益 -

3404.200

6 总结