1

1.1 该数据集有多少记录？多少变量？变量名称是什么？它们有意义吗？每个变量是什么类型？每个变量有多少个唯一值？什么值出现的频率最高，多久出现一次？有缺失值吗？如果有，这种情况发生的频率有多高？

```
melbcv<-read.csv("melbcv.csv")
str(melbcv)
```

```
## 'data.frame':    720 obs. of  15 variables:
##  $ Date                     : chr  "9/1/2012" "9/1/2012" "9/1/2012" "
9/1/2012" ...
##  $ Year                     : int  2012 2012 2012 2012 2012 2012 2012
 2012 2012 2012 ...
##  $ Month                    : int  9 9 9 9 9 9 9 9 9 9 ...
##  $ Mdate                    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Weekday_End              : int  20 20 20 20 20 20 20 20 20 20 ...
##  $ Day                      : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ Hour                     : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Town_Hall.West           : int  758 428 273 231 116 85 81 166 423
888 ...
##  $ Collins.Place.South      : int  169 37 32 11 6 29 37 86 121 177 ..
.
##  $ Australia.on.Collins     : int  90 61 34 60 23 29 26 80 249 580 ..
.
##  $ Bourke.Street.Mall.South: int  227 143 82 62 28 28 40 54 269 743
...
##  $ Bourke.Street.Mall.North: int  200 104 80 60 30 14 25 51 286 624
...
##  $ Melbourne.Central        : int  885 523 263 263 58 74 66 194 389 5
23 ...
##  $ Flagstaff.Station        : int  36 26 9 8 10 13 15 26 47 64 ...
##  $ State.Library            : int  252 182 92 91 39 28 33 54 185 356
...
```

从上面结果看出，数据 720 个记录，15 个变量，

变量名称分别是: Date、Year、Month、Mdate、Weekday_End、Day、Hour、Town_Hall.West、Collins.Place.South、Australia.on.Collins、Bourke.Street.Mall.South、Bourke.Street.Mall.North、Melbourne.Central、Flagstaff.Station、State.Library。

每个变量是什么类型:除了 Date 为 str 类型其余变量均为 int 类型。

每个变量有多少个唯一值:

```
sapply(melbcv,function(x){length(unique(x))})
```

```
##              Date              Year                     Month
##              30                 1                         1
##             Mdate        Weekday_End                       Day
##              30                 2                         7
##              Hour      Town_Hall.West      Collins.Place.South
##              24                613                       485
##     Australia.on.Collins Bourke.Street.Mall.South Bourke.Street.Mall
.North
##              498               552                       548
##        Melbourne.Central  Flagstaff.Station       State.Library
##              598               392                       308
```

什么值出现的频率最高，多久出现一次？

```
sapply(melbcv,function(x){sort(table(x),decreasing = T)[1]})
```

```
##               Date.9/1/2012                   Year.2012
##                         24                         720
##                    Month.9                     Mdate.1
##                        720                          24
##               Weekday_End.10                     Day.1
##                        480                         120
##                     Hour.0           Town_Hall.West.35
##                         30                           5
##      Collins.Place.South.10     Australia.on.Collins.16
##                         12                          10
## Bourke.Street.Mall.South.12  Bourke.Street.Mall.North.4
##                          9                           6
##        Melbourne.Central.32        Flagstaff.Station.10
##                          4                          15
##            State.Library.24
##                          5
```

有缺失值吗？如果有,这种情况发生的频率有多高？

```
sum(!complete.cases(melbcv))
```

```
## [1] 357
```

```
sapply(melbcv,function(x){sum(is.na(x))})
```

```
##              Date                      Year                     Month
##              0                         0                         0
##             Mdate               Weekday_End                       Day
##              0                         0                         0
##              Hour           Town_Hall.West      Collins.Place.South
##              0                         0                         0
##     Australia.on.Collins Bourke.Street.Mall.South Bourke.Street.Mall
.North
##              0                         0                         0
```

```
##        Melbourne.Central        Flagstaff.Station             State.L
ibrary
##                        0                        0                      357
```

有 357 个缺失值，都是在 State.Library 字段。

1.2 根据 1.1 的初步探索，你知道了数据集的基本情况，接下来要做感兴趣的探索，通过对变量的观察及描述统计，判断该数据集是否有变量可以忽略，如果有是什么？其中一个变量 Weekday_End，其有多少唯一取值的，根据数据集相关变量及常识说明取值的含义；

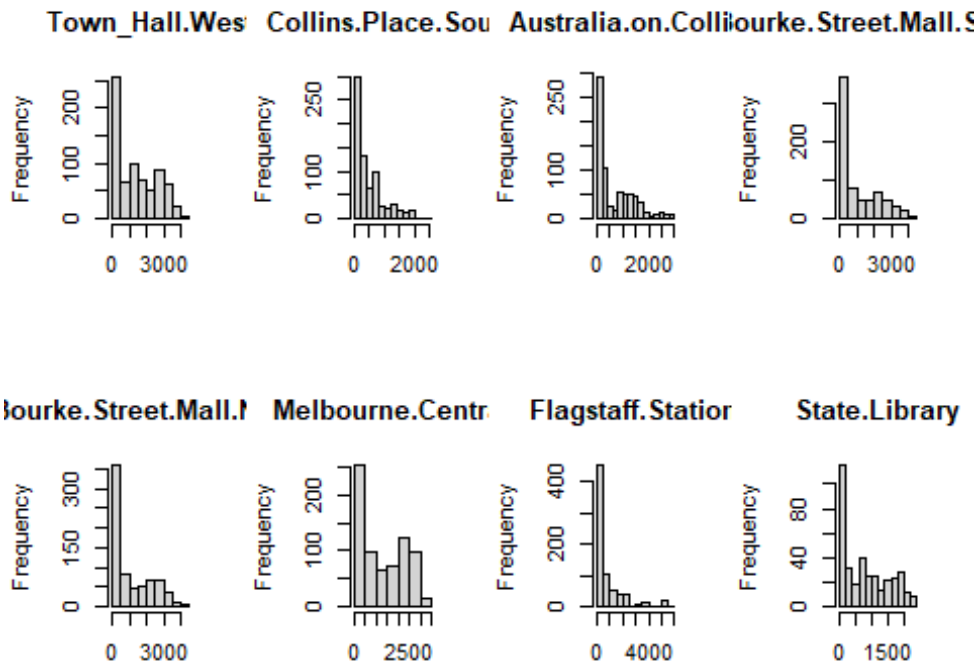
答：Year，Month 都是一样的，而且 Data 中已包含 Year 和 Month，且他们的值都是 2012 年 9 月，所以 Year,Month 这两个变量可以忽略； Weekday_End 只有两个不同的取值，即取值 10，20，可能一个代表工作日一个代表周末。

1.3 分别利用描述统计量（5 数）和图（选择适当的图）描述各条人行横道监控数据的特征；

```
summary(melbcv[8:15])
```

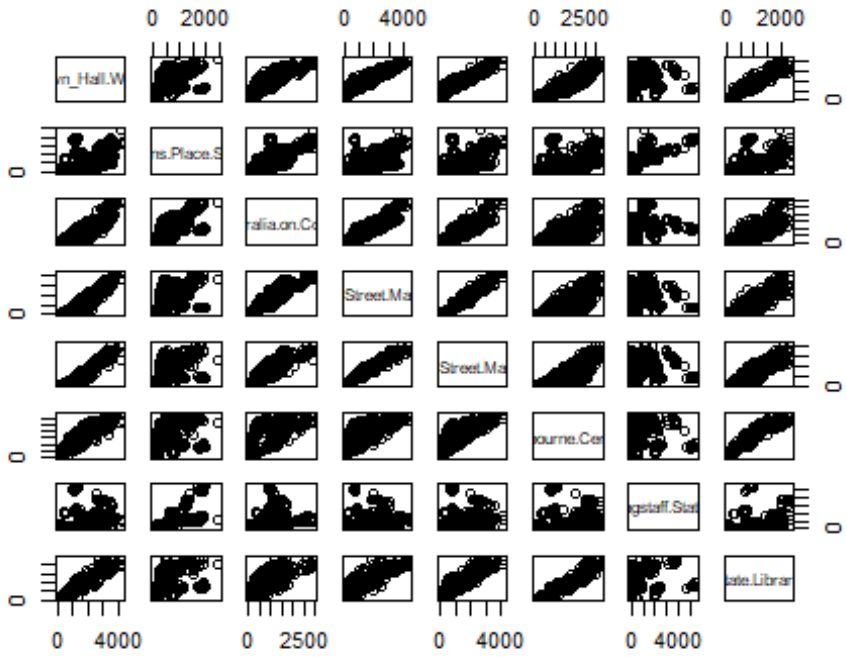
```
##  Town_Hall.West   Collins.Place.South Australia.on.Collins
##  Min.   :  13.0   Min.   :   1.00     Min.   :   0
##  1st Qu.: 239.8   1st Qu.:  57.75     1st Qu.:  70
##  Median :1166.0   Median : 275.00     Median : 333
##  Mean   :1380.1   Mean   : 455.98     Mean   : 683
##  3rd Qu.:2448.2   3rd Qu.: 686.75     3rd Qu.:1210
##  Max.   :4295.0   Max.   :2445.00     Max.   :2944
##
##  Bourke.Street.Mall.South Bourke.Street.Mall.North Melbourne.Central
##  Min.   :   0.00          Min.   :   1.00          Min.   :  14.0
##  1st Qu.:  83.75          1st Qu.:  76.75          1st Qu.: 235.8
##  Median : 447.00          Median : 515.00          Median :1089.5
##  Mean   :1040.90          Mean   :1045.48          Mean   :1264.5
##  3rd Qu.:1961.00          3rd Qu.:1980.75          3rd Qu.:2259.8
##  Max.   :4372.00          Max.   :4285.00          Max.   :3255.0
##
##  Flagstaff.Station State.Library
##  Min.   :   0.0    Min.   :   5.0
##  1st Qu.:  24.0    1st Qu.:  93.5
##  Median :  78.0    Median : 676.0
##  Mean   : 689.2    Mean   : 800.4
##  3rd Qu.: 965.5    3rd Qu.:1455.5
##  Max.   :5561.0    Max.   :2395.0
##                    NA's   :357
```

```
par(mfrow=c(2,4))
for (i in 8:15) {
  hist(melbcv[,names(melbcv)[i]],main=names(melbcv)[i],xlab="")
}
```



1.4 用监控点数据绘制散点矩阵图（**scatter matrix plot**），观察各条人行道监控数据间是否存在相关性，哪些变量间存在相关性？

```
plot(melbcv[8:15])
```



Town_Hall.West 与除 Collins.Place.South，Flagstaff.Station 外的人行道均相关；

Collins.Place.South 与其余人行道未发现明显相关；

Australia.on.Collins 其他人行道均有一定的相关性；

Bourke.Street.Mall.South 与除 Flagstaff.Station 外的人行道均有一定的相关性；

Bourke.Street.Mall.North 与除 Flagstaff.Station 外的人行道流量均相关；

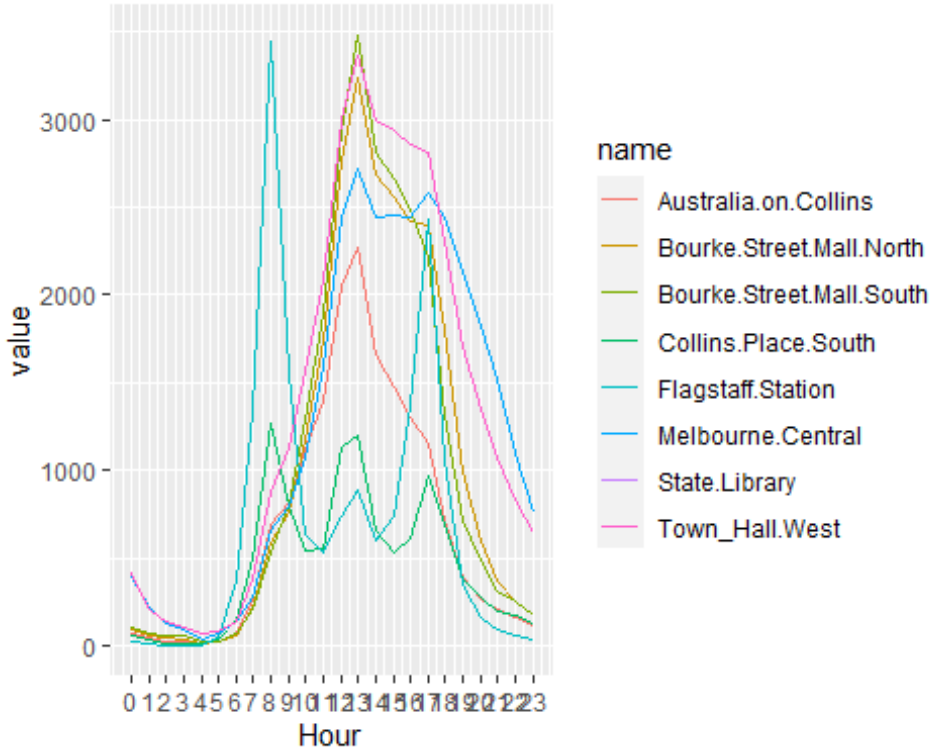
Melbourne.Central 与除 Flagstaff.Station 外的人行道流量均相关；

Flagstaff.Station 与其余人行道均未发现明显相关；

State.Library 与除 Flagstaff.Station 外的人行道流量均相关；

1.5 依据统计数据回答各个监控点一天的哪些时段是行人通过的高峰？所有监控点是否有一致的人流高峰时段？若有，是哪个（些）时段？

```
library(tidyverse)
library(ggplot2)
aggregate(melbcv[8:15],list(Hour=melbcv$Hour),mean) %>%
  pivot_longer(-1) %>%
  ggplot(aes(Hour,value))+
  geom_line(aes(colour=name))+
  scale_x_continuous(breaks = 0:23)
```



绘制对比图并观察得出，上午 8 点是 Flagstaff.Station,Collins.Place.South 的行人流量高峰时间；13 点是除 Flagstaff.Station 和 Collins.Place.South 的路口的行人流量高峰时间。结论：中午时间段各人行道人流量都比较大，早上时间段只有 Flagstaff.Station 人流量比较大。

哪个路口人流量最大？哪个路口人流量最小？依据什么得出的结论？

```
aggregate(melbcv[8:15],list(melbcv$Month),mean,na.rm=T) %>% t
```

```
##                                [,1]
## Group.1                      9.0000
## Town_Hall.West            1380.1250
## Collins.Place.South        455.9819
## Australia.on.Collins       682.9903
## Bourke.Street.Mall.South  1040.8972
## Bourke.Street.Mall.North  1045.4778
## Melbourne.Central         1264.5208
## Flagstaff.Station          689.1722
## State.Library              800.4077
```

从上面的统计结果看出，Town_Hall.West 的人流量最大，Collins.Place.South 的人流量最小。

人流量的最大值出现在哪个监控点的哪天的哪个时段？

```
melbcv %>% pivot_longer(8:15) %>%
  filter(value==max(value,na.rm = T))
```

```
## # A tibble: 1 x 9
##   Date       Year Month Mdate Weekday_End   Day  Hour name
    value
##   <chr>     <int> <int> <int>       <int> <int> <int> <chr>
    <int>
## 1 9/20/2012  2012     9    20          10     5     8 Flagstaff.Stat
ion  5561
```

所以人流量的最大值出现在 Flagstaff.Station 的 9/20/2012 的 8 时。

1.6 是否存在缺失值？出现在哪个（些）变量中，可否忽略？如不能，说明原因，并尝试对其进行插补，采用怎样的插补策略比较合适？

```
sapply(melbcv,function(x){sum(is.na(x))})
```

```
##                       Date                       Year
 Month
##                          0                          0
     0
##                      Mdate               Weekday_End
   Day
##                          0                          0
     0
##                       Hour           Town_Hall.West      Collins.Place
.South
##                          0                          0
     0
##     Australia.on.Collins Bourke.Street.Mall.South Bourke.Street.Mall
.North
##                          0                          0
     0
##        Melbourne.Central        Flagstaff.Station            State.L
ibrary
##                          0                          0
   357
```

存在。出现在 State.Library 中，有 357 个缺失值，其余变量无缺失值，不能忽略，可以用平均值对缺失值进行填充。

```
melbcv$State.Library<-ifelse(is.na(melbcv$State.Library),mean(melbcv$St
ate.Library,na.rm = T),melbcv$State.Library)
```

1.7 对有明显相关性的变量尝试建立相应的统计模型，并通过作图检验模型。

相关系数矩阵

```
round(cor(melbcv[8:15]),2)
```

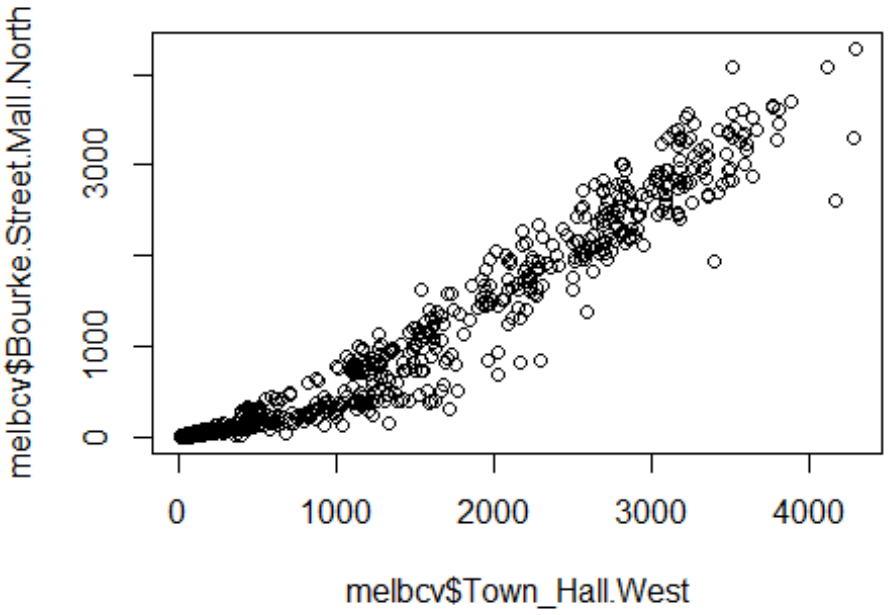
```
##                          Town_Hall.West Collins.Place.South
## Town_Hall.West                     1.00                0.67
```

```
## Collins.Place.South                0.67                1.00
## Australia.on.Collins               0.90                0.78
## Bourke.Street.Mall.South           0.95                0.64
## Bourke.Street.Mall.North           0.97                0.65
## Melbourne.Central                  0.95                0.56
## Flagstaff.Station                  0.34                0.82
## State.Library                      0.70                0.46
##                          Australia.on.Collins Bourke.Street.Mall.Sou
th
## Town_Hall.West                           0.90                     0.
95
## Collins.Place.South                      0.78                     0.
64
## Australia.on.Collins                     1.00                     0.
95
## Bourke.Street.Mall.South                 0.95                     1.
00
## Bourke.Street.Mall.North                 0.93                     0.
99
## Melbourne.Central                        0.76                     0.
84
## Flagstaff.Station                        0.41                     0.
28
## State.Library                            0.59                     0.
66
##                          Bourke.Street.Mall.North Melbourne.Central
## Town_Hall.West                               0.97              0.95
## Collins.Place.South                          0.65              0.56
## Australia.on.Collins                         0.93              0.76
## Bourke.Street.Mall.South                     0.99              0.84
## Bourke.Street.Mall.North                     1.00              0.89
## Melbourne.Central                            0.89              1.00
## Flagstaff.Station                            0.32              0.25
## State.Library                                0.68              0.70
##                          Flagstaff.Station State.Library
## Town_Hall.West                        0.34          0.70
## Collins.Place.South                   0.82          0.46
## Australia.on.Collins                  0.41          0.59
## Bourke.Street.Mall.South              0.28          0.66
## Bourke.Street.Mall.North              0.32          0.68
## Melbourne.Central                     0.25          0.70
## Flagstaff.Station                     1.00          0.25
## State.Library                         0.25          1.00
```

从相关系数矩阵来看，各站点的人流量均有一定的相关性，例如 Town_Hall.West 与 Bourke.Street.Mall.North 就高度相关。

```
cor.test(melbcv$Town_Hall.West,melbcv$Bourke.Street.Mall.North)
```

```
##
##  Pearson's product-moment correlation
##
## data:  melbcv$Town_Hall.West and melbcv$Bourke.Street.Mall.North
## t = 102.04, df = 718, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9621339 0.9716086
## sample estimates:
##      cor
## 0.967206
plot(melbcv$Town_Hall.West,melbcv$Bourke.Street.Mall.North)
```



p-value < 2.2e-16，说明二者有较强的线性相关关系。

2

2.1 选取除区站号和与气压相关以外的变量存放到对象 bjmoeda 中，通过 str（）函数、描述统计和适当的图查看 bjmoeda 数据基本特征，bjmoeda 有多少变量？每个变量有多少个唯一值？什么值出现的频率最高，多久出现一次？有缺失值吗？如果有，这种情况发生的频率有多高？

```
bjmo<-read.csv("bjmo.csv",encoding = "UTF-8")
bjmoeda<-subset(bjmo,select = -c(区站号,平均本站气压,平均水汽压,日最低本站气压,日最高本站气压))
str(bjmoeda)
```

```
## 'data.frame':    1551 obs. of  17 variables:
##  $ 年              : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ 月              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ 日              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X20.20时降水量: int  0 4 99 1 0 0 0 0 0 0 ...
##  $ 大型蒸发量      : int  32766 32766 32766 32766 32766 32766 32766 32766 32766 32766 ...
##  $ 极大风速        : int  62 57 110 179 134 61 51 42 72 51 ...
##  $ 极大风速的风向: int  3 4 5 16 2 2 3 10 3 4 ...
##  $ 平均风速        : int  16 21 27 50 23 15 13 9 16 19 ...
##  $ 平均气温        : int  -50 -45 -81 -103 -125 -112 -97 -98 -74 -70 ...
##  $ 平均相对湿度    : int  36 79 86 39 41 41 41 60 57 58 ...
##  $ 日照时数        : int  34 0 0 82 80 78 72 0 72 42 ...
##  $ 日最低气温      : int  -102 -63 -99 -134 -156 -167 -135 -121 -134 -110 ...
##  $ 日最高气温      : int  -16 -21 -44 -85 -75 -59 -44 -74 -28 -28 ...
##  $ 小型蒸发量      : int  12 2 0 13 10 7 8 4 7 6 ...
##  $ 最大风速        : int  34 34 57 77 60 33 27 22 43 33 ...
##  $ 最大风速的风向: int  3 4 5 16 1 3 2 9 1 3 ...
##  $ 最小相对湿度    : int  23 27 72 29 28 25 25 42 33 39 ...
```

```
summary(bjmoeda)
```

```
##        年             月               日         X20.20时降水量     大型蒸发量
##  Min.   :2010   Min.   : 1.00   Min.   : 1.00   Min.   :    0   Min.   :    9
##  1st Qu.:2011   1st Qu.: 3.00   1st Qu.: 8.00   1st Qu.:    0   1st Qu.:   40
##  Median :2012   Median : 6.00   Median :16.00   Median :    0   Median :   70
##  Mean   :2012   Mean   : 6.26   Mean   :15.72   Mean   : 2864   Mean
```

```
   :14938
##  3rd Qu.:2013   3rd Qu.: 9.00   3rd Qu.:23.00   3rd Qu.:    1   3rd
Qu.:32766
##  Max.   :2014   Max.   :12.00   Max.   :31.00   Max.   :32766   Max.
   :32766
##      极大风速       极大风速的风向        平均风速         平均气温
##  Min.   : 25.00   Min.   : 1.000   Min.   : 6.00   Min.   : -125.0
##  1st Qu.: 56.00   1st Qu.: 4.000   1st Qu.:16.00   1st Qu.:   15.0
##  Median : 73.00   Median : 9.000   Median :20.00   Median :  132.0
##  Mean   : 80.63   Mean   : 8.369   Mean   :21.99   Mean   :  145.2
##  3rd Qu.:100.50   3rd Qu.:11.000   3rd Qu.:26.00   3rd Qu.:  234.0
##  Max.   :228.00   Max.   :16.000   Max.   :66.00   Max.   :32766.0
##    平均相对湿度       日照时数         日最低气温         日最高气温
##  Min.   : 9.00   Min.   :  0.00   Min.   :-167.00   Min.   :-85.0
##  1st Qu.:34.00   1st Qu.: 31.50   1st Qu.: -27.00   1st Qu.: 67.0
##  Median :52.00   Median : 77.00   Median :  76.00   Median :189.0
##  Mean   :51.23   Mean   : 65.92   Mean   :  77.38   Mean   :174.8
##  3rd Qu.:67.00   3rd Qu.: 97.00   3rd Qu.: 186.00   3rd Qu.:280.0
##  Max.   :97.00   Max.   :141.00   Max.   : 292.00   Max.   :406.0
##     小型蒸发量        最大风速       最大风速的风向      最小相对湿度
##  Min.   :    0   Min.   : 17.00   Min.   : 1.000   Min.   : 4.00
##  1st Qu.:   24   1st Qu.: 35.50   1st Qu.: 4.000   1st Qu.:16.00
##  Median :32766   Median : 46.00   Median : 9.000   Median :26.00
##  Mean   :18096   Mean   : 49.06   Mean   : 8.397   Mean   :30.85
##  3rd Qu.:32766   3rd Qu.: 61.00   3rd Qu.:11.000   3rd Qu.:43.00
##  Max.   :32766   Max.   :120.00   Max.   :16.000   Max.   :85.00
```

bjmoeda 有 17 个变量,

每个变量有多少个唯一值？

```
sapply(bjmoeda,function(x)length(unique(x)))
```

```
##       年              月              日  X20.20 时降水量      大型蒸发量
##       5             12             31            136             87
##      极大风速  极大风速的风向        平均风速         平均气温    平均相对湿度
##       156             16             54            393             85
##      日照时数      日最低气温      日最高气温       小型蒸发量        最大风速
##       137            394            410             87             90
## 最大风速的风向    最小相对湿度
##       16             82
```

什么值出现的频率最高，

```
sapply(bjmoeda,function(x){sort(table(x),decreasing = T)[1]})
```

```
##    年.2012              月.1              日.1  X20.20 时降水量.0
##    366              155               51             1136
##  大型蒸发量.32766       极大风速.56 极大风速的风向.10       平均风速.16
##     706               29              199              99
```

```
##     平均气温.277    平均相对湿度.65          日照时数.0     日最低气温.-39
##     13                  39                  266                 12
##     日最高气温.278   小型蒸发量.32766        最大风速.34 最大风速的风向.10
##     15                  856                 51                  254
##    最小相对湿度.11
##     58
```

有缺失值吗？如果有，这种情况发生的频率有多高？

```
sum(!complete.cases(bjmoeda))

## [1] 0

sapply(bjmoeda,function(x){sum(is.na(x))})

##   年              月              日  X20.20 时降水量       大型蒸发量
##   0               0               0               0               0
##   极大风速  极大风速的风向        平均风速        平均气温    平均相对湿度
##   0               0               0               0               0
##   日照时数       日最低气温       日最高气温       小型蒸发量         最大风速
##   0               0               0               0               0
## 最大风速的风向    最小相对湿度
##   0               0
```

数据没有缺失值

2.2 阅读“气象数据集说明文档”，检查并处理缺失值，将缺失值替换成NA；

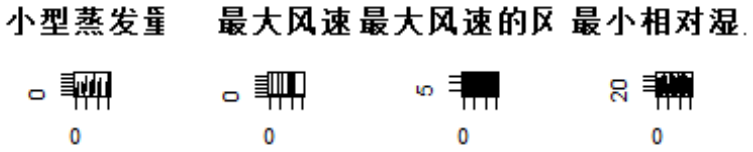
```
bjmoeda<-lapply(bjmoeda,function(x){
  ifelse(x %in%c(32744,32700,32766),NA,x)}) %>% data.frame()
```

检查除缺失值外的其他需要处理的数据编码，给出将其转换为正常值的策略；

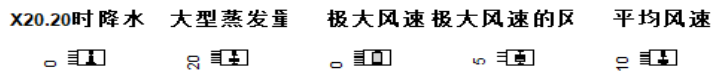
```
#风速,当风速≥xxx 值时，在原值上加"1000"
bjmoeda$最大风速<-ifelse(bjmoeda$最大风速>=100,bjmoeda$最大风速+1000,bjmo
eda$最大风速)
bjmoeda$极大风速<-ifelse(bjmoeda$极大风速>=100,bjmoeda$极大风速+1000,bjmo
eda$极大风速)
#风向
```

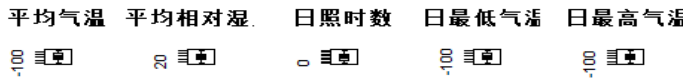
2.3 用折线图和箱线图描述各个观测变量值的变化情况，

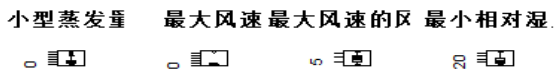
```
par(mfrow=c(3,5))
for (i in 4:17) {
  plot(1:nrow(bjmoeda),bjmoeda[,names(bjmoeda)[i]],type="l",
       main =names(bjmoeda)[i],xlab = "",ylab = "")
}
```




```
par(mfrow=c(3,5))
for (i in 4:17) {
  boxplot(bjmoeda[,names(bjmoeda)[i]],main =names(bjmoeda)[i])
}
```







计算各个变量的月平均值，最大值，最小值以及中值，

```
aggregate(bjmoeda[4:17],list(month=bjmoeda[,'月']),mean,na.rm=T)
```

```
##    month X20.20时降水量 大型蒸发量 极大风速 极大风速的风向 平均风速    平均气温
## 1      1      0.9060403        NaN 347.6387       8.316129 21.33548
```

```
-36.194805
## 2      2      1.7238806        NaN 297.0567       7.354610 21.12766
 -8.496454
## 3      3      2.8859060        NaN 447.5097       8.851613 25.37419
 68.896774
## 4      4      8.3454545   47.18333 620.1333       9.141667 28.70000
138.075000
## 5      5      9.6545455   56.75806 451.1048       9.145161 25.89516
219.459677
## 6      6     38.5576923   44.94017 273.8750       8.050000 21.73333
249.641667
## 7      7     76.5700935   42.64706 217.7419       8.032258 19.46774
277.241935
## 8      8     44.7118644   42.57377 225.5645       8.080645 19.26613
265.233871
## 9      9     25.6120690   36.63333 134.7167       8.108333 18.90000
208.208333
## 10    10     10.7264957   28.96748 229.1855       8.346774 17.96774
140.056452
## 11    11      9.1517857        NaN 386.3500       8.333333 20.44167
 58.925000
## 12    12      1.1666667        NaN 433.8145       8.709677 23.11290
-15.395161
##     平均相对湿度 日照时数 日最低气温 日最高气温 小型蒸发量 最大风速
## 1      43.94839 60.03871  -76.58065   12.54839   15.28387 46.47097
## 2      46.57447 54.43262  -51.36170   41.56028   19.90780 54.00000
## 3      38.68387 72.58710   16.15484  123.43226   48.65806 62.23226
## 4      39.04167 81.30833   80.65833  194.62500        NaN 62.60000
## 5      42.91935 85.06452  158.70161  277.03226        NaN 65.78226
## 6      60.62500 64.12500  202.89167  300.05833        NaN 83.85833
## 7      68.54032 57.84677  236.63710  321.00000        NaN 52.11290
## 8      68.83065 67.40323  225.25000  310.04032        NaN 43.20161
## 9      62.67500 67.20833  163.06667  258.40833        NaN 42.70833
## 10     58.46774 63.66129   90.03226  195.05645        NaN 42.01613
## 11     48.01667 60.63333   12.87500  110.12500   29.11667 54.74167
## 12     42.17742 58.40323  -53.51613   31.20968   18.21774 65.91935
##     最大风速的风向 最小相对湿度
## 1        8.161290     26.20000
## 2        7.567376     25.34043
## 3        8.980645     20.79355
## 4        9.425000     20.37500
## 5        9.758065     22.09677
## 6        7.466667     38.71667
## 7        7.919355     47.64516
## 8        8.096774     46.26613
## 9        7.683333     39.35000
## 10       7.524194     34.06452
## 11       9.241667     27.54167
## 12       8.967742     26.27419
```

```
aggregate(bjmoeda[4:17],list(month=bjmoeda[,'月']),max,na.rm=T)
```

```
##    month X20.20时降水量 大型蒸发量 极大风速 极大风速的风向 平均风速 平均气温
## 1      1            99       -Inf     1193             16       58       44
## 2      2            45       -Inf     1165             16       47       61
## 3      3            86       -Inf     1214             16       47      184
## 4      4           267         98     1192             16       64      223
## 5      5           206        117     1228             16       53      267
## 6      6           698         98     1213             16       37      306
## 7      7           842        100     1152             16       32      345
## 8      8           789         85     1167             16       41      317
## 9      9           636         85     1172             16       52      266
## 10    10           231         78     1171             16       46      203
## 11    11           489       -Inf     1184             16       66      143
## 12    12            52       -Inf     1171             16       63       64
##    平均相对湿度 日照时数 日最低气温 日最高气温 小型蒸发量 最大风速
## 1            91       95          0        128         44       98
## 2            86      104         34        157         60     1113
## 3            91      115        132        259        102     1110
## 4            83      127        167        303       -Inf       99
## 5            86      137        224        350       -Inf     1107
## 6            91      137        267        380       -Inf     1120
## 7            97      141        292        406       -Inf     1101
## 8            91      127        265        361       -Inf       87
## 9            92      122        240        326       -Inf       95
## 10           87      109        165        264       -Inf       88
## 11           89       98        126        203         66     1108
## 12           90       87         34        147         50     1110
##    最大风速的风向 最小相对湿度
## 1              16           82
## 2              16           72
## 3              16           76
## 4              16           67
## 5              16           63
## 6              15           84
## 7              16           83
```

```
## 8              16           84
## 9              16           83
## 10             16           75
## 11             16           77
## 12             16           85

aggregate(bjmoeda[4:17],list(month=bjmoeda[,'月']),min,na.rm=T)

##    month X20.20时降水量 大型蒸发量 极大风速 极大风速的风向 平均风速 平均
气温
## 1      1              0        Inf       30              1        6
    -125
## 2      2              0        Inf       32              1        7
     -78
## 3      3              0        Inf       39              1       11
     -28
## 4      4              0          9       48              1       10
      63
## 5      5              0         15       49              1       10
     156
## 6      6              0         12       36              1       12
     190
## 7      7              0         10       31              1        7
     209
## 8      8              0         12       35              1       10
     213
## 9      9              0         10       37              1        6
     148
## 10    10              0          9       27              1        8
      49
## 11    11              0        Inf       25              1        6
     -22
## 12    12              0        Inf       29              1        7
     -95
##    平均相对湿度 日照时数 日最低气温 日最高气温 小型蒸发量 最大风速
## 1            16        0       -167        -85          0       20
## 2             9        0       -128        -27          1       20
## 3            12        0        -67         -4          2       27
## 4             9        0          8         98        Inf       28
## 5            13        0         82        198        Inf       34
## 6            23        0        146        211        Inf       25
## 7            21        0        195        248        Inf       22
## 8            35        0        159        235        Inf       24
## 9            25        0         79        179        Inf       25
## 10           21        0         -2         91        Inf       19
## 11           14        0        -52          1          2       17
## 12           17        0       -137        -58          0       20
##    最大风速的风向 最小相对湿度
## 1               1            8
## 2               1            6
```

```
## 3               1            6
## 4               1            5
## 5               1            4
## 6               1           10
## 7               1           11
## 8               1           11
## 9               1           12
## 10              1            9
## 11              1            7
## 12              1           11
```

```
aggregate(bjmoeda[4:17],list(month=bjmoeda$月),median,na.rm=T)
```

```
##    month X20.20时降水量 大型蒸发量 极大风速 极大风速的风向 平均风速 平均气温
## 1      1              0         NA     65.0              9     18.0   -36.5
## 2      2              0         NA     70.0              7     19.0    -9.0
## 3      3              0         NA     87.0             10     24.0    68.0
## 4      4              0       46.5   1101.0             10     27.0   141.0
## 5      5              0       55.5     90.0             10     25.0   219.0
## 6      6              0       46.0     76.0              8     21.0   253.5
## 7      7              0       44.0     69.0              8     19.0   277.0
## 8      8              0       41.0     65.0              9     19.0   266.5
## 9      9              0       34.0     62.0              9     18.0   208.5
## 10    10              0       27.0     61.5              9     16.0   139.5
## 11    11              0         NA     65.0              9     16.5    52.5
## 12    12              0         NA     68.5             10     19.0   -15.5
##    平均相对湿度 日照时数 日最低气温 日最高气温 小型蒸发量 最大风速
## 1          40.0     75.0      -78.0        8.0         14     42.0
## 2          49.0     70.0      -50.0       37.0         18     44.0
## 3          36.0     86.0       14.0      117.0         48     53.0
## 4          36.0     98.5       77.0      198.0         NA     62.5
## 5          42.5     97.5      160.0      278.0         NA     54.5
## 6          62.0     76.0      203.0      306.0         NA     46.0
## 7          69.0     55.0      236.0      323.0         NA     42.0
## 8          70.5     82.0      227.0      311.0         NA     41.0
## 9          64.5     81.5      171.0      259.0         NA     40.0
## 10         59.5     81.0       94.5      202.5         NA     39.0
```

```
## 11          47.5     77.0        8.5      109.5          26     40.0
## 12          36.5     73.0      -53.0       25.5          16     43.5
##    最大风速的风向 最小相对湿度
## 1                9         19.0
## 2                8         21.0
## 3               10         15.0
## 4               10         14.0
## 5               10         19.0
## 6                7         35.5
## 7                9         47.0
## 8                9         46.0
## 9                9         39.0
## 10               9         30.0
## 11              10         23.0
## 12              10         20.0
```

找出 5 年间北京地区气候特征，最高温、最低温、 平均温的极值分别出现在哪年哪月哪日，值是多少？

```
subset(bjmoeda,日最高气温==max(日最高气温),select=c(年,月,日,日最高气温))
```

```
##       年 月 日 日最高气温
## 186 2010  7  5        406
```

```
subset(bjmoeda,日最低气温==max(日最低气温),select=c(年,月,日,日最低气温))
```

```
##       年 月 日 日最低气温
## 212 2010  7 31        292
```

```
subset(bjmoeda,平均气温==max(平均气温,na.rm = T),select=c(年,月,日,平均气温))
```

```
##       年 月 日 平均气温
## 187 2010  7  6      345
```

2.4

这 5 年间北京的大风（10m/s 以上）天气有多少天？

```
sum(bjmoeda$平均风速>10)
```

```
## [1] 1484
```

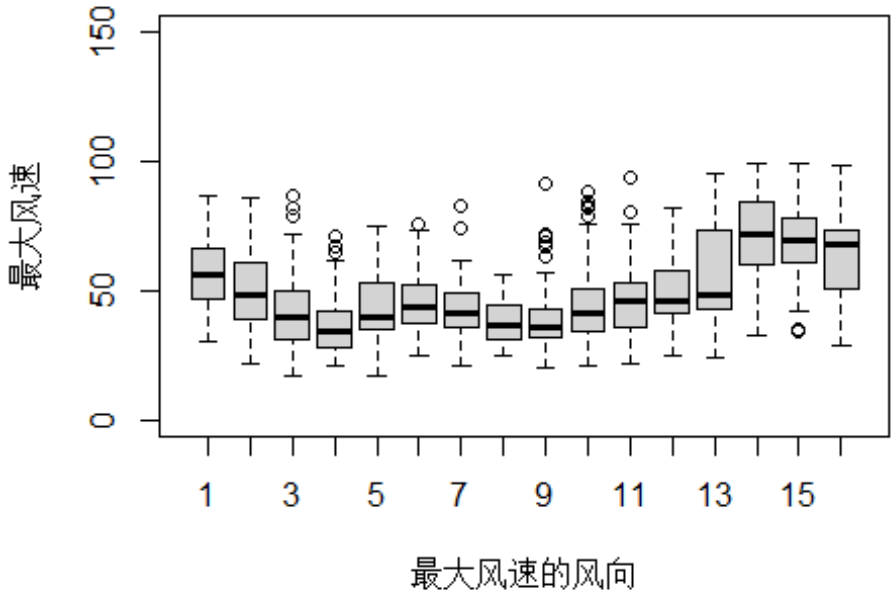
最大风速是多少？

```
max(bjmoeda$平均风速)
```

```
## [1] 66
```

探索风向和风速的关系，是否存在相关性？ 若存在，相关性如何？

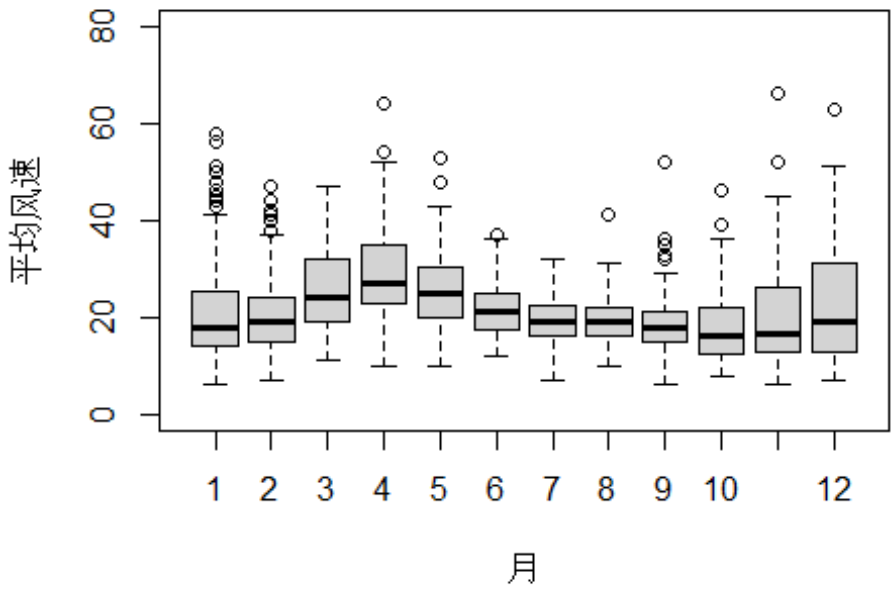
```
boxplot(最大风速~最大风速的风向,data=bjmoeda,ylim=c(0,150))
```



最大风速跟风向有相关性，风向为 14 时的风速更大。

大风极端天气出现在什么季节？

```
boxplot(平均风速~月,data=bjmoeda,ylim=c(0,80))
```



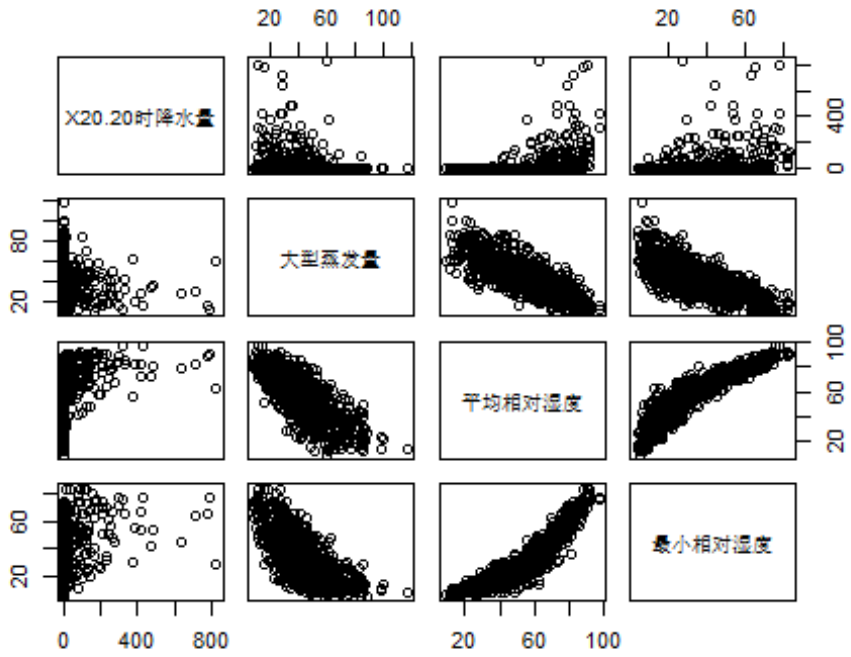
从上图看出，3，4 的平均风速更大。

2.5 探索降水量和蒸发量以及湿度之间的关系；用适当的图描述。

```
cor(bjmoeda[complete.cases(bjmoeda[c(4,5,10,17)]),c(4,5,10,17)]) %>% ro
und(2)
```

```
##                  X20.20 时降水量 大型蒸发量 平均相对湿度 最小相对湿度
## X20.20 时降水量             1.00      -0.18         0.32         0.31
## 大型蒸发量                 -0.18       1.00        -0.76        -0.71
## 平均相对湿度                0.32      -0.76         1.00         0.91
## 最小相对湿度                0.31      -0.71         0.91         1.00
```

```
plot(bjmoeda[complete.cases(bjmoeda[c(4,5,10,17)]),c(4,5,10,17)])
```



3

3.1 用探索一个新的数据集的一般策略对其进行探索，用描述统计方法对数据结构进行概述；

```
spam<-read.csv("spam.csv")
str(spam)
```

```
## 'data.frame':    2171 obs. of  21 variables:
##  $ isuid      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ day.of.week: chr  "Thu" "Thu" "Thu" "Thu" ...
##  $ time.of.day: int  0 0 14 3 3 4 4 5 6 6 ...
##  $ size.kb    : int  7 2 3 3 4 4 3 2 3 1 ...
##  $ box        : chr  "no" "no" "no" "yes" ...
##  $ domain     : chr  "com" "com" "edu" "de" ...
##  $ local      : chr  "no" "no" "yes" "no" ...
##  $ digits     : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ name       : chr  "name" "name" "name" "name" ...
##  $ cappct     : num  0.1935 0.1915 0.0741 0.2 0.4348 ...
##  $ special    : int  1 5 2 0 2 1 1 0 1 0 ...
##  $ credit     : chr  "no" "no" "no" "no" ...
##  $ sucker     : chr  "no" "no" "no" "no" ...
##  $ porn       : chr  "no" "no" "no" "no" ...
##  $ chain      : chr  "no" "no" "yes" "no" ...
```

```
##  $ username   : chr  "no" "no" "no" "no" ...
##  $ large.text : chr  "no" "no" "no" "no" ...
##  $ spampct    : int  NA NA 13 NA NA NA NA NA NA NA ...
##  $ category   : chr  "news" "com" "list" "ord" ...
##  $ spam       : chr  "no" "yes" "no" "no" ...
```

检查缺失值

```
sapply(spam,function(x)sum(is.na(x)))

##       isuid          id day.of.week time.of.day     size.kb
box
##           0           0           0           0           0
  0
##      domain       local      digits        name      cappct     spec
ial
##           0           0           0           0           0
  0
##      credit      sucker        porn       chain    username  large.t
ext
##           0           0           0           0           0
  0
##     spampct    category        spam
##        1353           0           0

summary(spam)

##      isuid              id         day.of.week         time.of.day
##  Min.   : 1.000   Min.   :   1.0   Length:2171        Min.   : 0.00
##  1st Qu.: 4.000   1st Qu.:  33.0   Class :character   1st Qu.: 9.00
##  Median : 9.000   Median :  62.0   Mode  :character   Median :12.00
##  Mean   : 9.234   Mean   : 201.2                      Mean   :12.26
##  3rd Qu.:14.000   3rd Qu.: 108.0                      3rd Qu.:16.00
##  Max.   :19.000   Max.   :3470.0                      Max.   :23.00
##
##     size.kb            box               domain             local
##  Min.   :   0.00   Length:2171        Length:2171        Length:2171
##  1st Qu.:   2.00   Class :character   Class :character   Class :char
acter
##  Median :   4.00   Mode  :character   Mode  :character   Mode  :char
```

```
acter
##  Mean   :  16.49

##  3rd Qu.:   7.00

##  Max.   :1337.00

##

##      digits           name               cappct          special

##  Min.   : 0.000   Length:2171        Min.   :0.0000   Min.   : 0.000

##  1st Qu.: 0.000   Class :character   1st Qu.:0.0600   1st Qu.: 0.000

##  Median : 0.000   Mode  :character   Median :0.1280   Median : 1.000

##  Mean   : 0.591                      Mean   :0.1584   Mean   : 1.397

##  3rd Qu.: 0.000                      3rd Qu.:0.2000   3rd Qu.: 2.000

##  Max.   :23.000                      Max.   :1.0000   Max.   :35.000

##

##     credit             sucker              porn              chain

##  Length:2171        Length:2171        Length:2171        Length:217
1
##  Class :character   Class :character   Class :character   Class :cha
racter
##  Mode  :character   Mode  :character   Mode  :character   Mode  :cha
racter
##

##

##

##

##    username          large.text           spampct        category

##  Length:2171        Length:2171        Min.   : 0.00   Length:2171

##  Class :character   Class :character   1st Qu.:11.00   Class :charac
ter
##  Mode  :character   Mode  :character   Median :47.50   Mode  :charac
```

```
ter
##                                          Mean   :44.63

##                                          3rd Qu.:76.00

##                                          Max.   :99.00

##                                          NA's   :1353

##      spam
##  Length:2171
##  Class :character
##  Mode  :character
##
##
##
##
```

3.2 建立一个新变量 domain.reduced 用以减少域名的 domain 类别为：“edu” , “com” , “gov” , “org” , “net” , 和 “other.”

```
spam$domain.reduced<-ifelse(spam$domain %in% c("edu","com","gov","org",
"net"),spam$domain,"other")
table(spam$domain.reduced)

##
##   com   edu   gov   net   org other
##   807  1037    10   118    26   173
```

3.3 将 spam 作为分类变量, 用解释变量： day of week, time of day, size.kb, box, domain.reduced, local, digits, name, capct, special, credit, sucker, porn, chain, username, 和 large text, 构建一个随机森林分类器，令 mtry = 2.

```
spam$spam<-factor(spam$spam)
library(randomForest)
rf<-randomForest(spam~.,data = spam[-c(1,2,7,19,20)],mtry=2,importance=
T)
```

3.4 给出各个变量重要性排序结果；

```
importance<-rf$importance
importance[order(importance[,4],decreasing = T),]

##                            no          yes MeanDecreaseAccuracy MeanDe
creaseGini
## box             0.0303242559 0.2115446507         0.0896711666
180.527029
## domain.reduced 0.0327537385 0.0940848120         0.0528620119
123.693349
## local           0.0102573112 0.1022927734         0.0405176197
```

```
 99.316548
## large.text     0.0259049380 0.0373661579         0.0296702089
 82.539944
## sucker         0.0237821829 0.0184360152         0.0220295412
 52.675398
## size.kb        0.0066234088 0.0154643274         0.0095421279
 38.521756
## time.of.day    0.0051623269 0.0101608413         0.0068133597
 37.501054
## digits         0.0106490378 0.0132419147         0.0114736723
 35.237780
## name           0.0081831845 0.0107268843         0.0090226150
 29.342458
## cappct         0.0045895487 0.0058602455         0.0050000862
 28.014404
## special        0.0036993502 0.0027169048         0.0033845787
 15.773144
## day.of.week    0.0023136557 0.0053415034         0.0033152809
 15.646682
## credit         0.0054730711 0.0055646684         0.0055042214
 14.839089
## username       0.0027832310 0.0031349208         0.0029032820
  8.977702
## porn           0.0014869449 0.0025149006         0.0018221903
  5.162671
## chain          0.0001886244 0.0006894426         0.0003551951
  2.003343
```

3.5 有多少非垃圾邮件被错判成垃圾邮件 spam?

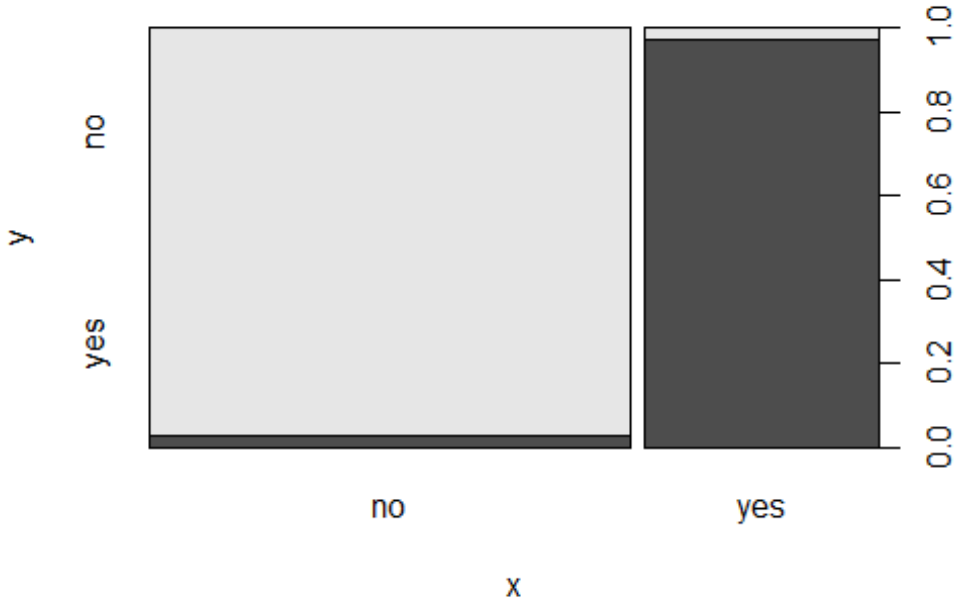
```
pred<-predict(rf,spam[-c(1,2,7,19,20)])
table(spam$spam,pred,dnn=c("actual","predicted"))

##       predicted
## actual   no  yes
##    no  1419   42
##    yes   21  689
```

有 43 条非垃圾邮件被错判为垃圾邮件。

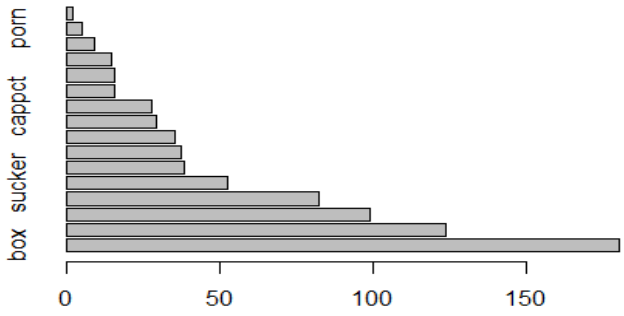
3.6 作出预测类别对真实类别的散点图,

```
plot(spam$spam,pred)
```



作出按随机森林返回的变量重要性排序作出解释变量的平行坐标图（a parallel coordinate plot）

```
barplot(importance[order(importance[,4],decreasing = T),4],horiz=T)
```



在 ggobi 中刷出非垃圾邮件被错分成垃圾邮件的案例，标出这些邮件的信息（如，全都来自 local box，字节少等等），然后观察垃圾邮件被正确分类的数据，它们有什么特殊之处？

```
spam2<-spam
spam2$pred<-pred
subset(spam2,spam=='no'&pred=='yes')
```

```
##      isuid  id day.of.week time.of.day size.kb box domain local digi
ts   name
## 1        1   1         Thu           0       7  no    com    no
 0   name
## 49       1  49         Fri          11       6  no    com    no
 0   name
## 66       1  66         Sun           9       2  no    com    no
 4  empty
## 82       1  82         Mon           7       3  no    com    no
 0   name
## 97       1  97         Mon          19       5  no    com    no
 0   name
## 101      1 101         Mon          20       3  no    com    no
 0  empty
## 103      1 103         Mon          20       2  no    com    no
 0  empty
## 415      2 283         Wed          12       3  no    com    no
 0   name
## 501      3  60         Thu          12       7  no    com    no
 0   name
## 753      6  82         Sun          21       2  no    com    no
 0   name
## 853      7  44         Thu           6       6  no    com    no
 0   name
## 879      8  11         Wed          23       4  no    com    no
 0   name
## 884      8  16         Thu          16       4  no    com    no
 4  empty
## 895      8  27         Sat           8       3  no    com    no
 0  empty
## 899      8  31         Sun          17       5  no    com    no
 0 single
## 912      8  44         Wed           9       6  no    com    no
 0  empty
## 916      8  48         Wed          18       9  no    com    no
 0   name
## 1114     9 130         Sun           0       3  no    com    no
 0   name
## 1214    10  95         Sun          21       2  no    com    no
 0   name
## 1301    11  31         Sun          11      10  no    com    no
 0   name
## 1308    11  38         Mon           8       2  no    com    no
 0 single
## 1347    11  77         Wed          20       7  no    com    no
 0   name
```

```
## 1358    11  88         Thu           18       8  no    com    no
 0   name
## 1361    11  91         Thu           22      35  no    com    no
 0  empty
## 1375    11 105         Fri           14       4  no    com    no
 0 single
## 1390    11 120         Sun            2       3  no    com    no
 0   name
## 1414    11 144         Wed            7      45  no    com    no
 0   name
## 1416    11 146         Wed           12       7  no    com    no
 0 single
## 1474    12  57         Wed            2       2  no    com    no
 0  empty
## 1479    12  62         Fri           17      18  no    com    no
 0 single
## 1558    14  13         Tue           20       2  no    com    no
 0  empty
## 1580    14  35         Fri           15       4  no    com    no
 0   name
## 1732    15  96         Thu           15       4  no    com    no
 0  empty
## 1839    16  57         Sun           16       1  no    com    no
 4  empty
## 1846    16  64         Thu           16       8  no    com    no
 0  empty
## 1930    17  84         Tue            0       6  no    com    no
 1   name
## 1940    17  94         Tue           13       4  no    com    no
 0   name
## 1977    17 131         Wed            0       5  no    com    no
 0   name
## 1978    17 132         Wed           17       4  no    com    no
 3   name
## 2004    18   8         Thu            9       6  no    com    no
 0   name
## 2005    18   9         Thu           16       3  no    com    no
11   name
## 2006    18  10         Thu           16       6  no    com    no
 0 single
##          cappct special credit sucker porn chain username large.text
 spampct
## 1    0.19354840       1     no     no   no    no       no         no
      NA
## 49   0.17241380       1     no     no   no    no       no         no
      NA
## 66   0.55319150       2     no     no   no    no       no         no
      NA
## 82   0.00000000       0     no     no   no    no       no         no
      NA
```

```
## 97   0.13953490       2     no     no   no     no        no          no
      28
## 101  0.04166667       0     no     no   no     no        no          no
      21
## 103  0.08333333       1     no     no   no     no        no          no
      42
## 415  0.04000000       1     no     no   no     no        no          no
      20
## 501  0.18000000       1     no     no   no     no        no          no
      NA
## 753  0.05882353       1     no     no   no     no        no          no
      NA
## 853  0.19354839       1     no     no   no     no        no          no
       0
## 879  0.00000000       0     no     no   no     no        no          no
       0
## 884  0.00000000       0    yes     no   no     no        no          no
       0
## 895  0.04347826       1     no     no   no     no        no          no
      19
## 899  0.02564103       7    yes     no   no     no        no          no
       7
## 912  0.15384615       4     no     no   no     no        no          no
      19
## 916  0.21052632       2     no     no   no     no        no          no
      61
## 1114 0.08300000       0     no     no   no     no        no          no
       0
## 1214 0.05882353       1     no     no   no     no        no          no
       0
## 1301 0.25806452       0     no     no   no     no       yes          no
      45
## 1308 0.05882353       0     no     no   no     no        no          no
      42
## 1347 0.11904762       4     no    yes   no     no        no         yes
      NA
## 1358 0.12820513       2     no    yes   no     no        no         yes
      NA
## 1361 0.25000000       1     no     no   no     no        no          no
      NA
## 1375 0.13157895       5     no     no   no     no        no          no
      NA
## 1390 0.13888889       1     no     no   no     no        no         yes
      NA
## 1414 0.13636364       0     no     no   no     no        no          no
      NA
## 1416 0.10000000       1     no    yes   no     no        no          no
      NA
## 1474 0.23076923       0     no     no   no     no        no          no
      NA
```

```
## 1479 0.12500000       1     no     no   no     no        no         yes
      NA
## 1558 0.00000000       0     no     no   no     no        no         yes
      NA
## 1580 0.00000000       0     no     no   no     no       yes          no
      NA
## 1732 0.09375000       0     no     no   no     no        no          no
       0
## 1839 0.16100000       1     no     no   no     no        no          no
      NA
## 1846 0.16000000       4     no     no   no     no        no         yes
      NA
## 1930 0.50000000       2     no     no   no     no        no          no
      22
## 1940 0.31578947       0     no     no   no     no        no          no
      NA
## 1977 0.10714286       0     no     no   no     no        no          no
       3
## 1978 0.00000000       3     no     no   no     no        no          no
      67
## 2004 0.16700000       5     no     no   no     no        no         yes
      47
## 2005 0.08000000       0     no     no   no     no        no          no
      13
## 2006 0.23000000       2     no     no   no     no        no          no
      63
##      category spam domain.reduced pred
## 1        news   no            com  yes
## 49       list   no            com  yes
## 66        ord   no            com  yes
## 82       list   no            com  yes
## 97        ord   no            com  yes
## 101       ord   no            com  yes
## 103       ord   no            com  yes
## 415      list   no            com  yes
## 501      list   no            com  yes
## 753       ord   no            com  yes
## 853      list   no            com  yes
## 879      list   no            com  yes
## 884      list   no            com  yes
## 895      list   no            com  yes
## 899      list   no            com  yes
## 912      list   no            com  yes
## 916      news   no            com  yes
## 1114     list   no            com  yes
## 1214      ord   no            com  yes
## 1301     list   no            com  yes
## 1308      ord   no            com  yes
## 1347     list   no            com  yes
## 1358     list   no            com  yes
```

```
## 1361     list   no            com  yes
## 1375     list   no            com  yes
## 1390     list   no            com  yes
## 1414     list   no            com  yes
## 1416     list   no            com  yes
## 1474      com   no            com  yes
## 1479      com   no            com  yes
## 1558      ord   no            com  yes
## 1580      ord   no            com  yes
## 1732      ord   no            com  yes
## 1839      ord   no            com  yes
## 1846      ord   no            com  yes
## 1930      ord   no            com  yes
## 1940     list   no            com  yes
## 1977     list   no            com  yes
## 1978     list   no            com  yes
## 2004     news   no            com  yes
## 2005     news   no            com  yes
## 2006     news   no            com  yes
```

这些邮件 box=no,domain=com,local=no,chain=no

3.7 检查 Spam (真实类别) 和 Spam.Prob (可能被当成垃圾邮件的). 有多少不是垃圾邮件的被认定超过 50% 可能是垃圾邮件（spam）?

```
spam.prob<-predict(rf,spam[-c(1,2,7,19,20)],type = "prob")[,2]
sum(spam$spam=='no' &spam.prob>0.5)
```

```
## [1] 42
```

3.8 检查非垃圾邮件被随机森林分成垃圾邮件的案例的概率排名（probability rating）.用一段文字描述具有高概率被当成垃圾邮件且被随机森林认为非常像垃圾邮件的案例。

```
spam2$spam.prob<-spam.prob
head(spam2[order(spam2$spam.prob,decreasing = T),])
```

```
##     isuid   id day.of.week time.of.day size.kb box domain local digi
ts   name
## 265     2  133         Fri          14       9  no    com    no
 4 single
## 714     6   95         Tue           8      17  no    com    no
 8   name
## 730     6 3398         Mon          21       6  no    com    no
 0 single
## 731     6 3397         Mon          16      18  no    com    no
 0 single
## 732     6 3396         Mon          16      18  no    com    no
 0 single
## 780     6   56         Fri           3       8  no    com    no
 0   name
```

```
##         cappct special credit sucker porn chain username large.text
spampct
## 265 0.11000000       0     no    yes   no    no       no         yes
     82
## 714 0.15094340       0     no    yes   no    no       no         yes
     83
## 730 0.18750000       2     no    yes   no    no       no         yes
     51
## 731 0.18750000       0     no    yes   no    no       no         yes
     39
## 732 0.18750000       0     no    yes   no    no       no         yes
     39
## 780 0.07692308       0     no    yes   no    no       no         yes
     63
##     category spam domain.reduced pred spam.prob
## 265      com  yes            com  yes     1.000
## 714     news  yes            com  yes     0.998
## 730      com  yes            com  yes     0.998
## 731      com  yes            com  yes     0.998
## 732      com  yes            com  yes     0.998
## 780      com  yes            com  yes     0.998
```

box=no,domain=com,local=no,credit=no,sucker=yes,porn=no,chain=no,username=no,large.text=no 的非垃圾邮件容易被分类为垃圾邮件。

3.9 哪个用户的非垃圾邮件最有可能被当成垃圾邮件？

```
subset(spam2,spam.prob==max(spam.prob),select = c(isuid,id))
```

```
##     isuid  id
## 265     2 133
```

3.10 根据你对数据的分析，你认为哪些变量在判断一个电子邮件是否垃圾邮件时最重要？

```
importance[order(importance[,4],decreasing = T),3:4]
```

```
##                MeanDecreaseAccuracy MeanDecreaseGini
## box                      0.0896711666       180.527029
## domain.reduced           0.0528620119       123.693349
## local                    0.0405176197        99.316548
## large.text               0.0296702089        82.539944
## sucker                   0.0220295412        52.675398
## size.kb                  0.0095421279        38.521756
## time.of.day              0.0068133597        37.501054
## digits                   0.0114736723        35.237780
## name                     0.0090226150        29.342458
## cappct                   0.0050000862        28.014404
## special                  0.0033845787        15.773144
## day.of.week              0.0033152809        15.646682
## credit                   0.0055042214        14.839089
```

```
## username                0.0029032820         8.977702
## porn                    0.0018221903         5.162671
## chain                   0.0003551951         2.003343
```

根据随机森林输出的特征重要性排序看，box，domain.reduce,local,larg.text 是更重要的变量。