

1 Etap I

```
df <- read.csv('churn.txt')

dim(df)

## [1] 3333   21

head(df)

##   State Account.Length Area.Code   Phone Int.l.Plan VMail.Plan VMail.Message
## 1    KS          128     415 382-4657       no      yes        25
## 2    OH          107     415 371-7191       no      yes        26
## 3    NJ          137     415 358-1921       no       no         0
## 4    OH           84     408 375-9999      yes      no         0
## 5    OK           75     415 330-6626      yes      no         0
## 6    AL          118     510 391-8027      yes      no         0
##   Day.Mins Day.Calls Day.Charge Eve.Mins Eve.Calls Eve.Charge Night.Mins
## 1    265.1      110    45.07   197.4      99    16.78    244.7
## 2    161.6      123    27.47   195.5     103    16.62    254.4
## 3    243.4      114    41.38   121.2     110    10.30    162.6
## 4    299.4       71    50.90   61.9       88     5.26    196.9
## 5    166.7      113    28.34   148.3     122    12.61    186.9
## 6    223.4      98    37.98   220.6     101    18.75    203.9
##   Night.Calls Night.Charge Intl.Mins Intl.Calls Intl.Charge CustServ.Calls
## 1         91     11.01     10.0       3      2.70        1
## 2        103     11.45     13.7       3      3.70        1
## 3        104      7.32     12.2       5      3.29        0
## 4         89      8.86      6.6       7      1.78        2
## 5        121      8.41     10.1       3      2.73        3
## 6        118      9.18      6.3       6      1.70        0
##   Churn.
## 1 False.
## 2 False.
## 3 False.
## 4 False.
## 5 False.
## 6 False.

str(df)

## 'data.frame': 3333 obs. of 21 variables:
## $ State : Factor w/ 51 levels "AK","AL","AR",...: 17 36 32 36 37 2 20 25 19 50 ...
## $ Account.Length: int 128 107 137 84 75 118 121 147 117 141 ...
## $ Area.Code : int 415 415 415 408 415 510 510 415 408 415 ...
## $ Phone : Factor w/ 3333 levels "327-1058","327-1319",...: 1927 1576 1118 1708 111 2254 1048 81
## $ Int.l.Plan : Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 2 1 2 ...
## $ VMail.Plan : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 1 1 2 ...
## $ VMail.Message : int 25 26 0 0 0 0 24 0 0 37 ...
## $ Day.Mins : num 265 162 243 299 167 ...
## $ Day.Calls : int 110 123 114 71 113 98 88 79 97 84 ...
## $ Day.Charge : num 45.1 27.5 41.4 50.9 28.3 ...
## $ Eve.Mins : num 197.4 195.5 121.2 61.9 148.3 ...
## $ Eve.Calls : int 99 103 110 88 122 101 108 94 80 111 ...
```

```

## $ Eve.Charge      : num 16.78 16.62 10.3 5.26 12.61 ...
## $ Night.Mins     : num 245 254 163 197 187 ...
## $ Night.Calls    : int 91 103 104 89 121 118 118 96 90 97 ...
## $ Night.Charge   : num 11.01 11.45 7.32 8.86 8.41 ...
## $ Intl.Mins      : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
## $ Intl.Calls     : int 3 3 5 7 3 6 7 6 4 5 ...
## $ Intl.Charge    : num 2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
## $ CustServ.Calls: int 1 1 0 2 3 0 3 0 1 0 ...
## $ Churn.          : Factor w/ 2 levels "False.", "True." : 1 1 1 1 1 1 1 1 1 1 ...

```

```

df$Area.Code <- as.factor(df$Area.Code)
df$Day.Charge <- as.integer(df$Day.Charge)

```

```

sapply(df, function(x) sum(is.na(x)))

##           State Account.Length       Area.Code        Phone Int.l.Plan
##             0            0            0            0            0
## VMail.Plan  VMail.Message       Day.Mins      Day.Calls Day.Charge
##             0            0            0            0            0
##   Eve.Mins    Eve.Calls    Eve.Charge    Night.Mins  Night.Calls
##             0            0            0            0            0
## Night.Charge    Intl.Mins    Intl.Calls    Intl.Charge CustServ.Calls
##             0            0            0            0            0
##           Churn.
##             0

```

```

df <- subset(df, select=-Phone)

```

```

sapply(df[, sapply(df, is.factor)], levels)

## $State
## [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "IA" "ID" "IL"
## [16] "IN" "KS" "KY" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT" "NC" "ND" "NE"
## [31] "NH" "NJ" "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT"
## [46] "VA" "VT" "WA" "WI" "WV" "WY"
##
## $Area.Code
## [1] "408" "415" "510"
##
## $Int.l.Plan
## [1] "no" "yes"
##
## $VMail.Plan
## [1] "no" "yes"
##
## $Churn.
## [1] "False." "True."

```

```

summary(df)

##      State    Account.Length  Area.Code Int.l.Plan VMail.Plan
##  WV     : 106   Min.    : 1.0    408: 838   no :3010   no :2411
##  MN     :  84   1st Qu.: 74.0   415:1655  yes: 323   yes: 922
##  NY     :  83   Median  :101.0   510: 840
##  AL     :  80   Mean    :101.1
##  OH     :  78   3rd Qu.:127.0
##  OR     :  78   Max.    :243.0
##  (Other):2824
## VMail.Message      Day.Mins      Day.Calls      Day.Charge
##  Min.   : 0.000   Min.   : 0.0   Min.   : 0.0   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:143.7  1st Qu.: 87.0  1st Qu.:24.00
##  Median : 0.000   Median  :179.4   Median :101.0   Median :30.00
##  Mean   : 8.099   Mean    :179.8   Mean   :100.4   Mean   :30.07
##  3rd Qu.:20.000   3rd Qu.:216.4  3rd Qu.:114.0  3rd Qu.:36.00
##  Max.   :51.000   Max.   :350.8   Max.   :165.0   Max.   :59.00
##
##      Eve.Mins      Eve.Calls      Eve.Charge      Night.Mins
##  Min.   : 0.0   Min.   : 0.0   Min.   : 0.00   Min.   : 23.2
##  1st Qu.:166.6  1st Qu.: 87.0  1st Qu.:14.16  1st Qu.:167.0
##  Median :201.4   Median :100.0   Median :17.12   Median :201.2
##  Mean   :201.0   Mean    :100.1   Mean   :17.08   Mean   :200.9
##  3rd Qu.:235.3  3rd Qu.:114.0  3rd Qu.:20.00  3rd Qu.:235.3
##  Max.   :363.7   Max.   :170.0   Max.   :30.91   Max.   :395.0
##
##      Night.Calls      Night.Charge      Intl.Mins      Intl.Calls
##  Min.   : 33.0   Min.   : 1.040   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 87.0   1st Qu.: 7.520   1st Qu.: 8.50   1st Qu.: 3.000
##  Median :100.0   Median : 9.050   Median :10.30   Median : 4.000
##  Mean   :100.1   Mean    : 9.039   Mean   :10.24   Mean   : 4.479
##  3rd Qu.:113.0   3rd Qu.:10.590   3rd Qu.:12.10  3rd Qu.: 6.000
##  Max.   :175.0   Max.   :17.770   Max.   :20.00   Max.   :20.000
##
##      Intl.Charge      CustServ.Calls      Churn.
##  Min.   :0.000   Min.   :0.000   False.:2850
##  1st Qu.:2.300   1st Qu.:1.000   True. : 483
##  Median :2.780   Median :1.000
##  Mean   :2.765   Mean    :1.563
##  3rd Qu.:3.270   3rd Qu.:2.000
##  Max.   :5.400   Max.   :9.000
##

```

2 Etap II

```

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg  ggplot2
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
```

```

##      filter, lag
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union

```

```

factors <- subset(df, select=sapply(df, is.factor))
numerics <- subset(df, select=sapply(df, function(x) !is.factor(x)))

```

a)

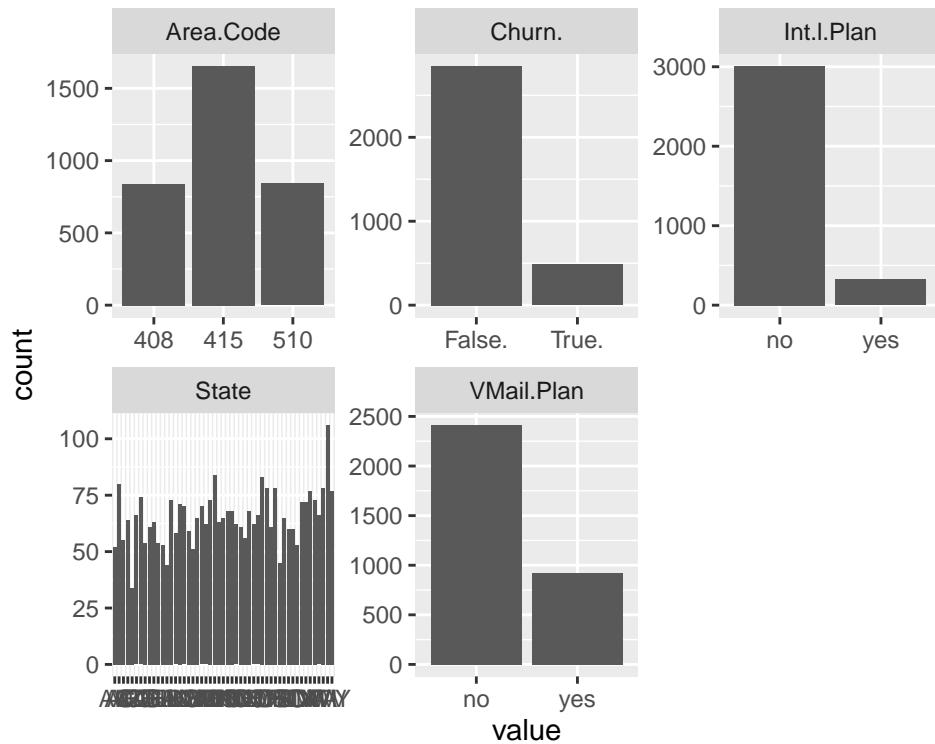
b)

```

ggplot(gather(factors), aes(value)) +
  geom_bar() +
  facet_wrap(~key, scales='free')

## Warning: attributes are not identical across measure variables;
## they will be dropped

```

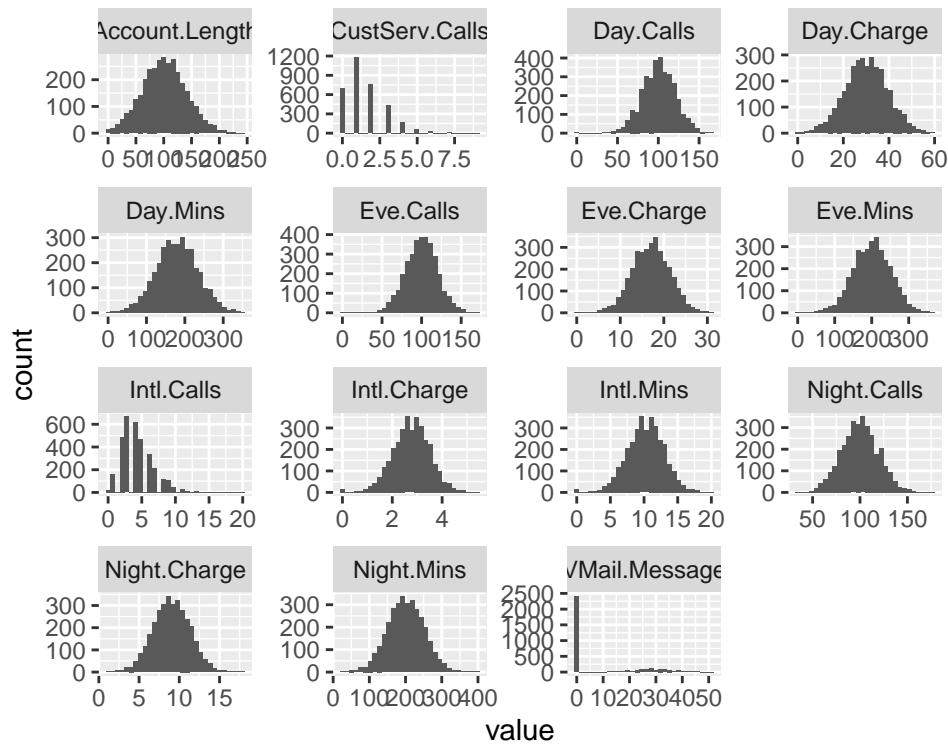


```

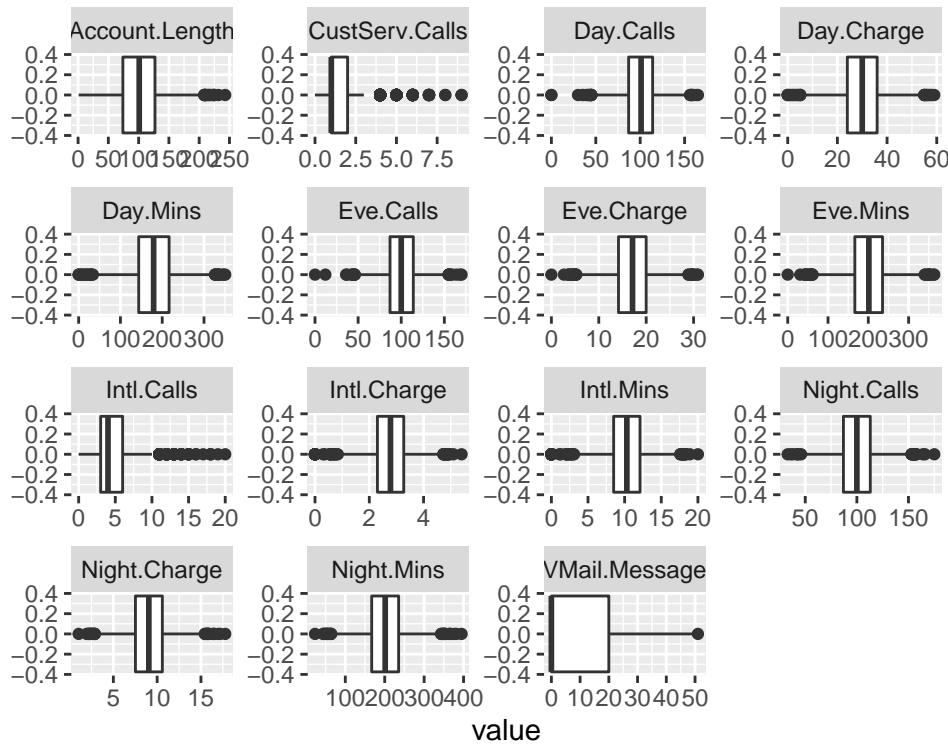
ggplot(gather(numerics), aes(value)) +
  geom_histogram() +
  facet_wrap(~key, scales='free')

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```

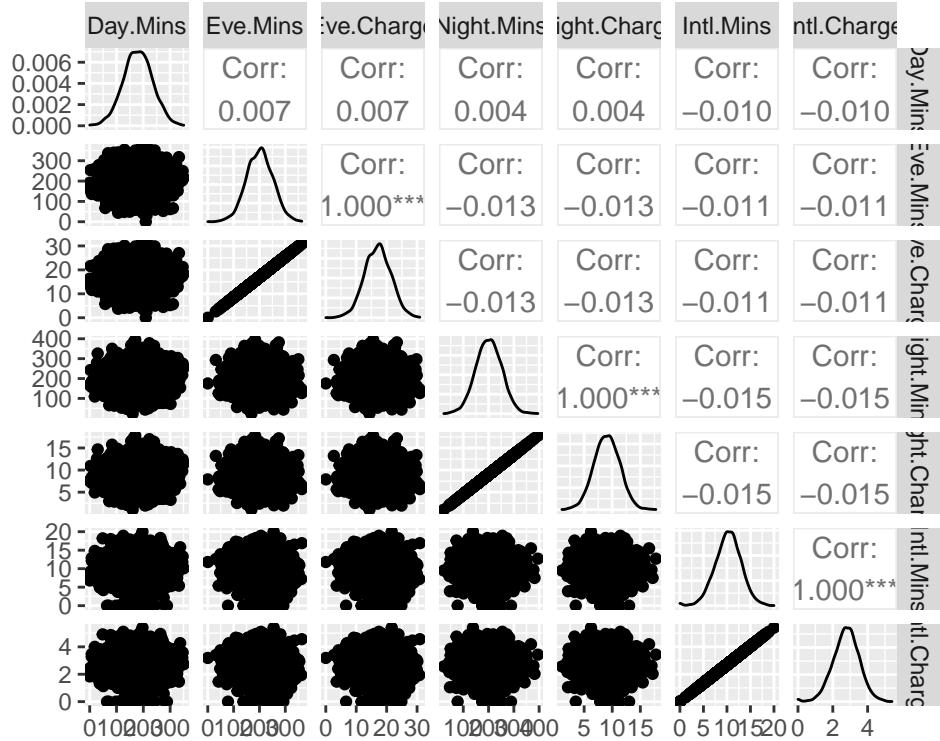


```
ggplot(gather(numerics), aes(value)) +
  geom_boxplot() +
  facet_wrap(~key, scales='free')
```



c)

```
continuous <- subset(numerics, select=sapply(numerics, function(x) !is.integer(x)))
ggpairs(continuous)
```



d)

```
sapply(numerics, range)

##      Account.Length VMail.Message Day.Mins Day.Calls Day.Charge Eve.Mins
## [1,]              1            0    0.0       0        0     0.0
## [2,]             243           51   350.8      165      59   363.7
##   Eve.Calls Eve.Charge Night.Mins Night.Calls Night.Charge Intl.Mins
## [1,]       0     0.00     23.2       33      1.04       0
## [2,]     170    30.91    395.0      175     17.77      20
##   Intl.Calls Intl.Charge CustServ.Calls
## [1,]       0     0.0         0
## [2,]     20      5.4         9
```

```
sapply(factors, levels)

## $State
## [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "IA" "ID" "IL"
## [16] "IN" "KS" "KY" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT" "NC" "ND" "NE"
## [31] "NH" "NJ" "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT"
## [46] "VA" "VT" "WA" "WI" "WV" "WY"
##
## $Area.Code
```

```

## [1] "408" "415" "510"
##
## $Int.l.Plan
## [1] "no"   "yes"
##
## $VMail.Plan
## [1] "no"   "yes"
##
## $Churn.
## [1] "False." "True."

```

```

library(moments)
sapply(numerics, skewness)

## Account.Length  VMail.Message      Day.Mins      Day.Calls      Day.Charge
##    0.096562812   1.264254335   -0.029063980  -0.111736324  -0.027264807
##    Eve.Mins       Eve.Calls       Eve.Charge     Night.Mins     Night.Calls
##   -0.023866709   -0.055538130   -0.023847250   0.008917276   0.032484942
##    Night.Charge    Intl.Mins      Intl.Calls    Intl.Charge  CustServ.Calls
##    0.008882237   -0.245025603   1.320883367  -0.245176105   1.090868260

```

```

cv <- function(X) sd(X) / mean(X)

sapply(numerics, cv)

## Account.Length  VMail.Message      Day.Mins      Day.Calls      Day.Charge
##    0.3940255     1.6901282     0.3029752     0.1998203     0.3080191
##    Eve.Mins       Eve.Calls       Eve.Charge     Night.Mins     Night.Calls
##    0.2523324     0.1989988     0.2523287     0.2517715     0.1954755
##    Night.Charge    Intl.Mins      Intl.Calls    Intl.Charge  CustServ.Calls
##    0.2517746     0.2727127     0.5494459     0.2726534     0.8417223

```

3 Etap III

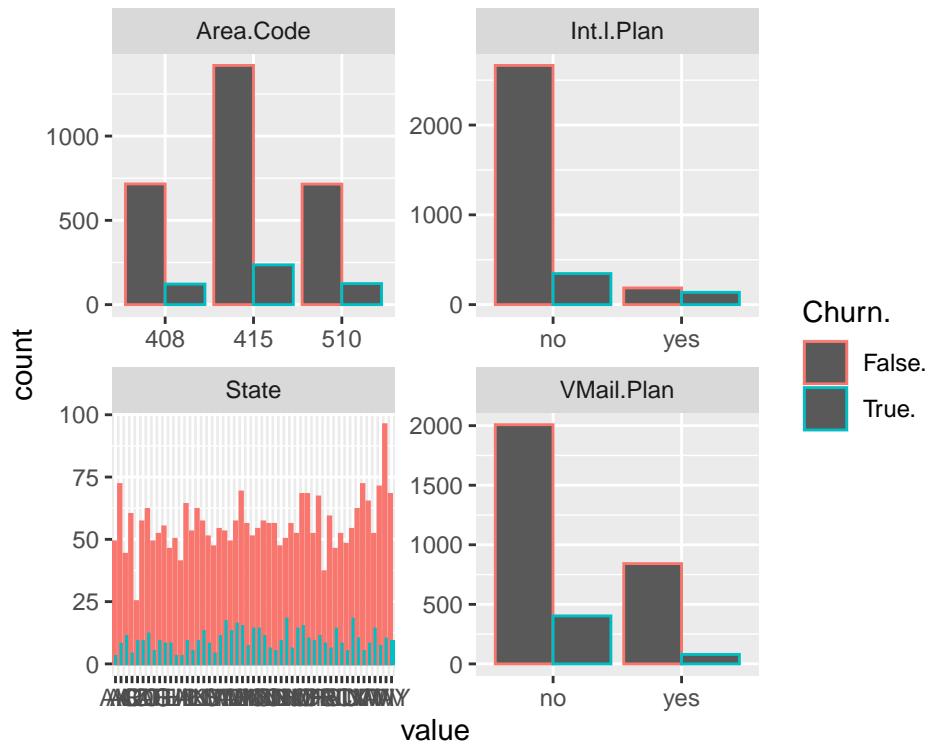
a)

```

ggplot(gather(factors, "key", "value", -Churn.), aes(value, color=Churn.)) +
  geom_bar(position='dodge') +
  facet_wrap(~key, scales='free')

## Warning: attributes are not identical across measure variables;
## they will be dropped

```



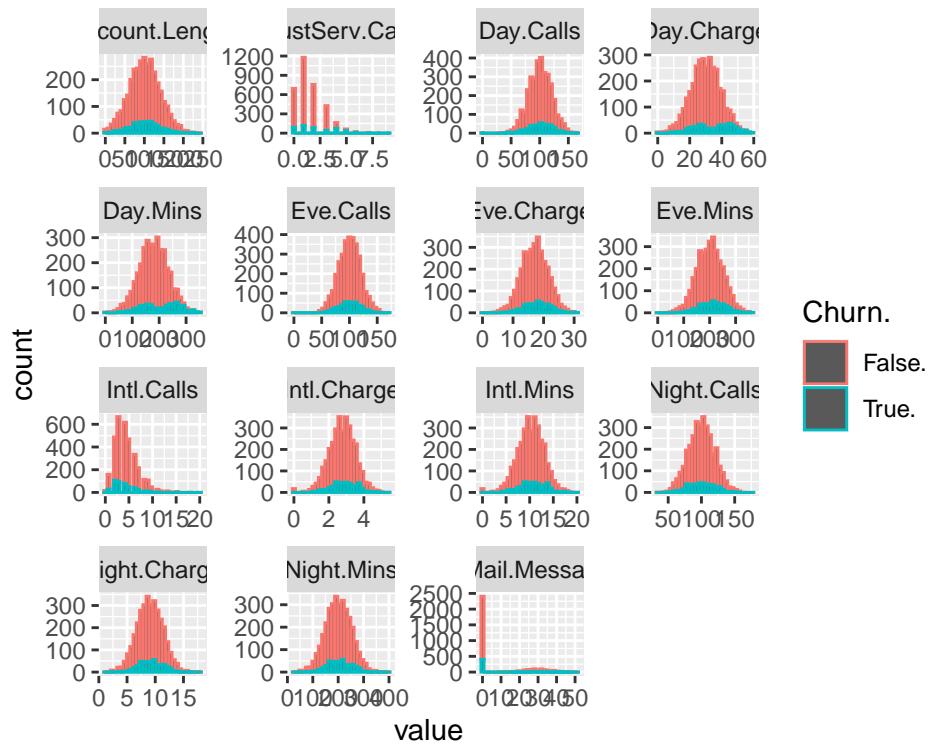
```

numerics <- data.frame(numerics, Churn. = df$Churn.)

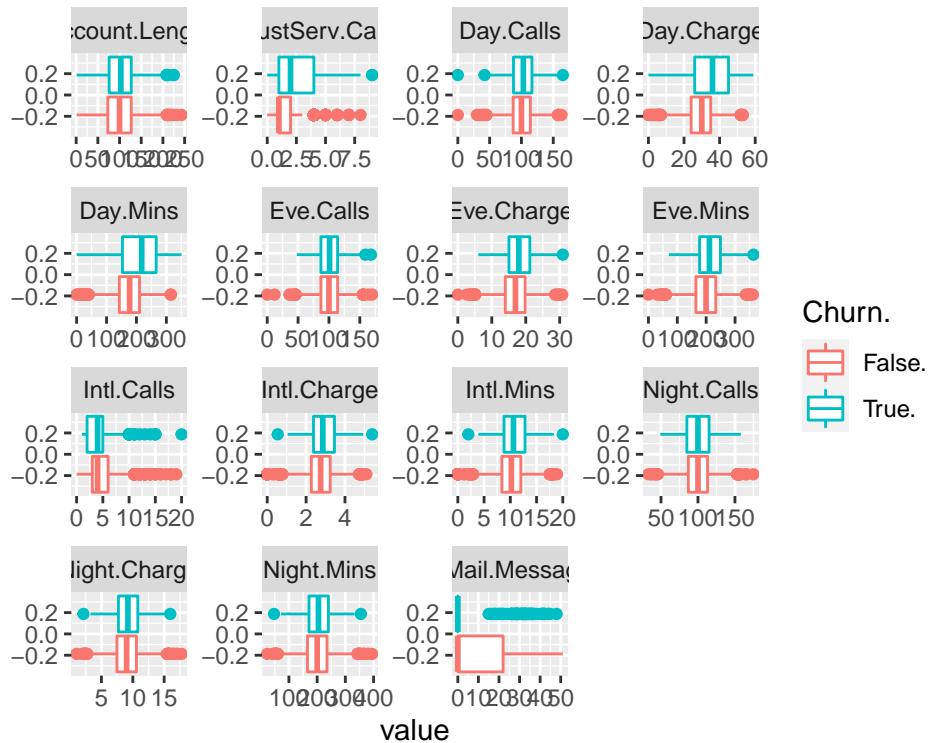
ggplot(gather(numerics, "key", "value", -Churn.), aes(value, color=Churn.)) +
  geom_histogram() +
  facet_wrap(~key, scales='free')

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```



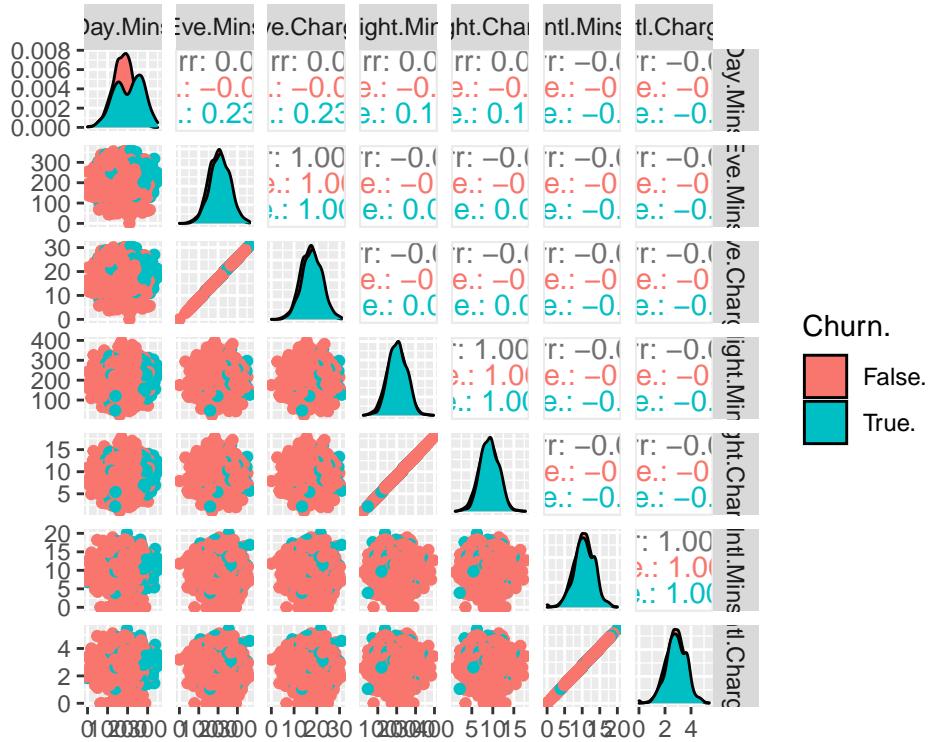
```
ggplot(gather(numerics, "key", "value", -Churn.), aes(value, color=Churn.)) +
  geom_boxplot() +
  facet_wrap(~key, scales='free')
```



```

continuous <- subset(numerics, select=sapply(numerics, function(x) !is.integer(x)))
continuous %>% ggpairs(,
  mapping = ggplot2::aes(color=Churn.),
  legend=1,
  columns=1:7)

```



```

numerics %>% group_by(Churn.) %>% group_map(~summarize(.x, across(everything(), range)))

## [[1]]
## # A tibble: 2 x 15
##   Account.Length VMail.Message Day.Mins Day.Calls Day.Charge Eve.Mins Eve.Calls
##       <int>           <int>    <dbl>     <int>      <dbl>    <dbl>     <int>
## 1          1            0        0        0        0        0        0
## 2         243           51      316.      163      53      362.     170
## # ... with 8 more variables: Eve.Charge <dbl>, Night.Mins <dbl>,
## #   Night.Calls <int>, Night.Charge <dbl>, Intl.Mins <dbl>, Intl.Calls <int>,
## #   Intl.Charge <dbl>, CustServ.Calls <int>
##
## [[2]]
## # A tibble: 2 x 15
##   Account.Length VMail.Message Day.Mins Day.Calls Day.Charge Eve.Mins Eve.Calls
##       <int>           <int>    <dbl>     <int>      <dbl>    <dbl>     <int>
## 1          1            0        0        0        0        70.9     48
## 2         225           48      351.      165      59      364.     168
## # ... with 8 more variables: Eve.Charge <dbl>, Night.Mins <dbl>,
## #   Night.Calls <int>, Night.Charge <dbl>, Intl.Mins <dbl>, Intl.Calls <int>,
## #   Intl.Charge <dbl>, CustServ.Calls <int>

```

```

numerics %>% group_by(Churn.) %>% group_map(~summarize(.x, across(everything(), skewness)))

## [[1]]
## # A tibble: 1 x 15
##   Account.Length VMail.Message Day.Mins Day.Calls Day.Charge Eve.Mins Eve.Calls
##       <dbl>           <dbl>     <dbl>      <dbl>      <dbl>      <dbl>
## 1        0.0905       1.17    -0.226    -0.0633    -0.224   -0.0431   -0.0692
## # ... with 8 more variables: Eve.Charge <dbl>, Night.Mins <dbl>,
## #   Night.Calls <dbl>, Night.Charge <dbl>, Intl.Mins <dbl>, Intl.Calls <dbl>,
## #   Intl.Charge <dbl>, CustServ.Calls <dbl>
## 
## [[2]]
## # A tibble: 1 x 15
##   Account.Length VMail.Message Day.Mins Day.Calls Day.Charge Eve.Mins Eve.Calls
##       <dbl>           <dbl>     <dbl>      <dbl>      <dbl>      <dbl>
## 1        0.137        2.03    -0.199    -0.352    -0.198    0.0330    0.0298
## # ... with 8 more variables: Eve.Charge <dbl>, Night.Mins <dbl>,
## #   Night.Calls <dbl>, Night.Charge <dbl>, Intl.Mins <dbl>, Intl.Calls <dbl>,
## #   Intl.Charge <dbl>, CustServ.Calls <dbl>

```

```

numerics %>% group_by(Churn.) %>% group_map(~summarize(.x, across(everything(), cv)))

## [[1]]
## # A tibble: 1 x 15
##   Account.Length VMail.Message Day.Mins Day.Calls Day.Charge Eve.Mins Eve.Calls
##       <dbl>           <dbl>     <dbl>      <dbl>      <dbl>      <dbl>
## 1        0.396        1.62    0.286     0.197     0.291     0.253     0.200
## # ... with 8 more variables: Eve.Charge <dbl>, Night.Mins <dbl>,
## #   Night.Calls <dbl>, Night.Charge <dbl>, Intl.Mins <dbl>, Intl.Calls <dbl>,
## #   Intl.Charge <dbl>, CustServ.Calls <dbl>
## 
## [[2]]
## # A tibble: 1 x 15
##   Account.Length VMail.Message Day.Mins Day.Calls Day.Charge Eve.Mins Eve.Calls
##       <dbl>           <dbl>     <dbl>      <dbl>      <dbl>      <dbl>
## 1        0.384        2.32    0.333     0.213     0.338     0.244     0.196
## # ... with 8 more variables: Eve.Charge <dbl>, Night.Mins <dbl>,
## #   Night.Calls <dbl>, Night.Charge <dbl>, Intl.Mins <dbl>, Intl.Calls <dbl>,
## #   Intl.Charge <dbl>, CustServ.Calls <dbl>

```

b)