The Insurance Company - Data Exploration

Atanu Choudhury

2021-05-19

Contents

1 Introduction

This problem focuses on the prediction of probable customers buying caravan insurance. The data set provided was part of the CoIL challenge in 2000. There are two main components in the problem. Firstly, identifying the customers who would like to buy the caravan insurance, and secondly an explanation of the customer behaviour which helped us in predicting the above behaviour.

As the data consists of real world data, it has 86 variables, half of those relate to socio-demographic data whereas the other half relates to product ownership data. The training set consists of 5822 records, including the information of whether the customers hold a caravan insurance. The dataset for predictions have 4000 records, where the target variable is missing. The target variable for the predictions is present in another file.

For the prediction task it is expected to find the set of 800 customers out of the 4000 who are more likely to buy the caravan insurance policy.

For the description task, it is expected to be explaninable to a marketing professional, who is not expected to have any information about machine learning. The final outcome of Machine Learning is its profiatbility in business scenarios. Thus, an explanatory model is expected from a business perspective.

The data dictionary explains the variables that were used in the dataset.

1.1 Importing libraries

```
library(psych)
source("read_data.R")
```

2 Data Exploration

2.1 Reading the train data

```
#caravan_data=read.table("ticdata2000.txt")
head(caravan_data)
```

```
##    V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19
   V20 V21
## 1 33  1  3  2  8  0  5  1  3   7   0   2   1   2   6   1   2   7   1
   0   1
## 2 37  1  2  2  8  1  4  1  4   6   2   2   0   4   5   0   5   4   0
   0   0
## 3 37  1  2  2  8  0  4  2  4   3   2   4   4   4   2   0   5   4   0
   0   0
## 4  9  1  3  3  3  2  3  2  4   5   2   2   2   3   4   3   4   2   4
   0   0
## 5 40  1  4  2 10  1  4  1  4   7   1   2   2   4   4   5   4   0   0
   5   4
## 6 23  1  2  1  5  0  5  0  5   0   6   3   3   5   2   0   5   4   2
   0   0
##   V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38
   V39 V40
## 1   2   5   2   1   1   2   6   1   1   8   8   0   1   8   1   0   4
     5   0
## 2   5   0   4   0   2   3   5   0   2   7   7   1   2   6   3   2   0
     5   2
## 3   7   0   2   0   5   0   4   0   7   2   7   0   2   9   0   4   5
     0   0
## 4   3   1   2   3   2   1   4   0   5   4   9   0   0   7   2   1   5
     3   0
## 5   0   0   0   9   0   0   0   0   4   5   6   2   1   5   4   0   0
     9   0
## 6   4   2   2   2   2   2   4   2   9   0   5   3   3   9   0   5   2
     3   0
##   V41 V42 V43 V44 V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57
   V58 V59
## 1   0   4   3   0   0   0   6   0   0   0   0   0   0   0   0   0   0
     0   5
## 2   0   5   4   2   0   0   0   0   0   0   0   0   0   0   0   0   0
     0   2
## 3   0   3   4   2   0   0   6   0   0   0   0   0   0   0   0   0   0
     0   2
## 4   0   4   4   0   0   0   6   0   0   0   0   0   0   0   0   0   0
     0   2
## 5   0   6   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0
     0   6
## 6   0   3   3   0   0   0   6   0   0   0   0   0   0   0   0   0   0
     0   0
##   V60 V61 V62 V63 V64 V65 V66 V67 V68 V69 V70 V71 V72 V73 V74 V75 V76
   V77 V78
## 1   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
     0   0
## 2   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0
```

```
       0    0
## 3    0    0    0    0    0    1    0    0    1    0    0    0    0    0    0    0    0
       0    0
## 4    0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0
       0    0
## 5    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
       0    0
## 6    0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0
       0    0
##     V79  V80  V81  V82  V83  V84  V85  V86
## 1    0    1    0    0    0    0    0    0
## 2    0    1    0    0    0    0    0    0
## 3    0    1    0    0    0    0    0    0
## 4    0    1    0    0    0    0    0    0
## 5    0    1    0    0    0    0    0    0
## 6    0    0    0    0    0    0    0    0
```

2.2 Checking the dimensions of the data

We can see that there are 5822 records and 86 columns as expected based on the description of the task.

```
dim(caravan_data)
```

```
## [1] 5822   86
```

2.3 Checking the structure of the data

On evaluating the structure of the dataframe we see that all the columns have discrete values, and are of type int. It would be preferable to convert the target response to a factor as we have to use classification algorithms to predict whether the customer is a propective buyer or not.

```
str(caravan_data)
```

```
## 'data.frame':    5822 obs. of  86 variables:
##  $ V1 : int  33 37 37 9 40 23 39 33 33 11 ...
##  $ V2 : int  1 1 1 1 1 1 2 1 1 2 ...
##  $ V3 : int  3 2 2 3 4 2 3 2 2 3 ...
##  $ V4 : int  2 2 2 3 2 1 2 3 4 3 ...
##  $ V5 : int  8 8 8 3 10 5 9 8 8 3 ...
##  $ V6 : int  0 1 0 2 1 0 2 0 0 3 ...
##  $ V7 : int  5 4 4 3 4 5 2 7 1 5 ...
##  $ V8 : int  1 1 2 2 1 0 0 0 3 0 ...
##  $ V9 : int  3 4 4 4 4 5 5 2 6 2 ...
##  $ V10: int  7 6 3 5 7 0 7 7 6 7 ...
##  $ V11: int  0 2 2 2 1 6 2 2 0 0 ...
##  $ V12: int  2 2 4 2 2 3 0 0 3 2 ...
##  $ V13: int  1 0 4 2 2 3 0 0 3 2 ...
##  $ V14: int  2 4 4 3 4 5 3 5 3 2 ...
##  $ V15: int  6 5 2 4 4 2 6 4 3 6 ...
##  $ V16: int  1 0 0 3 5 0 0 0 0 0 ...
##  $ V17: int  2 5 5 4 4 5 4 3 1 4 ...
##  $ V18: int  7 4 4 2 0 4 5 6 8 5 ...
##  $ V19: int  1 0 0 4 0 2 0 2 1 2 ...
```

```
##  $ V20: int  0 0 0 0 5 0 0 0 1 0 ...
##  $ V21: int  1 0 0 0 4 0 0 0 0 0 ...
##  $ V22: int  2 5 7 3 0 4 4 2 1 3 ...
##  $ V23: int  5 0 0 1 0 2 1 5 8 3 ...
##  $ V24: int  2 4 2 2 0 2 5 2 1 3 ...
##  $ V25: int  1 0 0 3 9 2 0 2 1 1 ...
##  $ V26: int  1 2 5 2 0 2 1 1 1 2 ...
##  $ V27: int  2 3 0 1 0 2 4 2 0 1 ...
##  $ V28: int  6 5 4 4 0 4 5 5 8 4 ...
##  $ V29: int  1 0 0 0 0 2 0 2 1 2 ...
##  $ V30: int  1 2 7 5 4 9 6 0 9 0 ...
##  $ V31: int  8 7 2 4 5 0 3 9 0 9 ...
##  $ V32: int  8 7 7 9 6 5 8 4 5 6 ...
##  $ V33: int  0 1 0 0 2 3 0 4 2 1 ...
##  $ V34: int  1 2 2 0 1 3 1 2 3 2 ...
##  $ V35: int  8 6 9 7 5 9 9 6 7 6 ...
##  $ V36: int  1 3 0 2 4 0 0 3 2 3 ...
##  $ V37: int  0 2 4 1 0 5 4 2 7 2 ...
##  $ V38: int  4 0 5 5 0 2 3 5 2 3 ...
##  $ V39: int  5 5 0 3 9 3 3 3 1 3 ...
##  $ V40: int  0 2 0 0 0 0 0 0 0 1 ...
##  $ V41: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V42: int  4 5 3 4 6 3 3 3 2 4 ...
##  $ V43: int  3 4 4 4 3 3 5 3 3 7 ...
##  $ V44: int  0 2 2 0 0 0 0 0 0 2 ...
##  $ V45: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V46: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V47: int  6 0 6 6 0 6 6 0 5 0 ...
##  $ V48: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V49: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V50: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V51: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V52: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V53: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V54: int  0 0 0 0 0 0 0 3 0 0 ...
##  $ V55: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V56: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V57: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V58: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V59: int  5 2 2 2 6 0 0 0 0 3 ...
##  $ V60: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V61: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V62: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V63: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V64: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V65: int  0 2 1 0 0 0 0 0 0 1 ...
##  $ V66: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V67: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V68: int  1 0 1 1 0 1 1 0 1 0 ...
##  $ V69: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V70: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V71: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V72: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V73: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ V74: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V75: int  0 0 0 0 0 0 0 1 0 0 ...
##  $ V76: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V77: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V78: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V79: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V80: int  1 1 1 1 1 0 0 0 0 1 ...
##  $ V81: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V82: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V83: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V84: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V85: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V86: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

2.4 Convert target to factor

```
caravan_data$V86=as.factor(caravan_data$V86)
```

2.5 Summary of the data

```
summary(caravan_data)
```

```
##        V1              V2               V3              V4
##  Min.   : 1.00   Min.   : 1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:10.00   1st Qu.: 1.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :30.00   Median : 1.000   Median :3.000   Median :3.000
##  Mean   :24.25   Mean   : 1.111   Mean   :2.679   Mean   :2.991
##  3rd Qu.:35.00   3rd Qu.: 1.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :41.00   Max.   :10.000   Max.   :5.000   Max.   :6.000
##        V5               V6               V7              V8
##  Min.   : 1.000   Min.   :0.0000   Min.   :0.000   Min.   :0.00
##  1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.:4.000   1st Qu.:0.00
##  Median : 7.000   Median :0.0000   Median :5.000   Median :1.00
##  Mean   : 5.774   Mean   :0.6965   Mean   :4.627   Mean   :1.07
##  3rd Qu.: 8.000   3rd Qu.:1.0000   3rd Qu.:6.000   3rd Qu.:2.00
##  Max.   :10.000   Max.   :9.0000   Max.   :9.000   Max.   :5.00
##        V9              V10              V11               V12
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.00
##  1st Qu.:2.000   1st Qu.:5.000   1st Qu.:0.0000   1st Qu.:1.00
##  Median :3.000   Median :6.000   Median :1.0000   Median :2.00
##  Mean   :3.259   Mean   :6.183   Mean   :0.8835   Mean   :2.29
##  3rd Qu.:4.000   3rd Qu.:7.000   3rd Qu.:1.0000   3rd Qu.:3.00
##  Max.   :9.000   Max.   :9.000   Max.   :7.0000   Max.   :9.00
##       V13              V14             V15            V16              V17
##  Min.   :0.000   Min.   :0.00   Min.   :0.0   Min.   :0.000   Min.
   :0.000
##  1st Qu.:0.000   1st Qu.:2.00   1st Qu.:3.0   1st Qu.:0.000   1st Qu
   .:2.000
##  Median :2.000   Median :3.00   Median :4.0   Median :1.000   Median
   :3.000
##  Mean   :1.888   Mean   :3.23   Mean   :4.3   Mean   :1.461   Mean
   :3.351
```

```
##  3rd Qu.:3.000   3rd Qu.:4.00   3rd Qu.:6.0   3rd Qu.:2.000   3rd Qu
    .:4.000
##  Max.   :9.000   Max.   :9.00   Max.   :9.0   Max.   :9.000   Max.
    :9.000
##        V18             V19             V20             V21
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:3.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :5.000   Median :2.000   Median :0.000   Median :0.0000
##  Mean   :4.572   Mean   :1.895   Mean   :0.398   Mean   :0.5223
##  3rd Qu.:6.000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :9.000   Max.   :9.000   Max.   :5.000   Max.   :9.0000
##        V22             V23            V24             V25
    V26
##  Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.
    :0.000
##  1st Qu.:2.000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:0.000   1st Qu
    .:1.000
##  Median :3.000   Median :2.00   Median :2.000   Median :1.000   Median
    :2.000
##  Mean   :2.899   Mean   :2.22   Mean   :2.306   Mean   :1.621   Mean
    :1.607
##  3rd Qu.:4.000   3rd Qu.:3.00   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu
    .:2.000
##  Max.   :9.000   Max.   :9.00   Max.   :9.000   Max.   :9.000   Max.
    :9.000
##        V27             V28             V29             V30
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:2.000
##  Median :2.000   Median :4.000   Median :1.000   Median :4.000
##  Mean   :2.203   Mean   :3.759   Mean   :1.067   Mean   :4.237
##  3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:7.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##        V31             V32            V33             V34
    V35
##  Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.
    :0.000
##  1st Qu.:2.000   1st Qu.:5.00   1st Qu.:0.000   1st Qu.:1.000   1st Qu
    .:5.000
##  Median :5.000   Median :6.00   Median :1.000   Median :2.000   Median
    :7.000
##  Mean   :4.772   Mean   :6.04   Mean   :1.316   Mean   :1.959   Mean
    :6.277
##  3rd Qu.:7.000   3rd Qu.:7.00   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu
    .:8.000
##  Max.   :9.000   Max.   :9.00   Max.   :7.000   Max.   :9.000   Max.
    :9.000
##        V36             V37             V38             V39
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :2.000   Median :2.000   Median :4.000   Median :3.000
##  Mean   :2.729   Mean   :2.574   Mean   :3.536   Mean   :2.731
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##        V40              V41              V42             V43
```

```
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:3.000  
##  Median :0.0000   Median :0.0000   Median :4.000   Median :4.000  
##  Mean   :0.7961   Mean   :0.2027   Mean   :3.784   Mean   :4.236  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:6.000  
##  Max.   :9.0000   Max.   :9.0000   Max.   :9.000   Max.   :8.000  
##       V44              V45                V46                V47      
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00  
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :5.00  
##  Mean   :0.7712   Mean   :0.04002   Mean   :0.07162   Mean   :2.97  
##  3rd Qu.:2.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:6.00  
##  Max.   :3.0000   Max.   :6.00000   Max.   :4.00000   Max.   :8.00  
##       V48                V49              V50                 V51          
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.000000   Min.   :0.00000  
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000  
##  Median :0.00000   Median :0.0000   Median :0.000000   Median :0.00000  
##  Mean   :0.04827   Mean   :0.1754   Mean   :0.009447   Mean   :0.02096  
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.00000  
##  Max.   :7.00000   Max.   :7.0000   Max.   :9.000000   Max.   :5.00000  
##       V52                V53                V54             V55        
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.0000  
##  Median :0.00000   Median :0.00000   Median :0.000   Median :0.0000  
##  Mean   :0.09258   Mean   :0.01305   Mean   :0.215   Mean   :0.1948  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:0.0000  
##  Max.   :6.00000   Max.   :6.00000   Max.   :6.000   Max.   :9.0000  
##       V56                V57                V58                V59       
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000  
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :2.000  
##  Mean   :0.01374   Mean   :0.01529   Mean   :0.02353   Mean   :1.828  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:4.000  
##  Max.   :6.00000   Max.   :3.00000   Max.   :7.00000   Max.   :8.000  
##       V60                  V61                V62                V63          
##  Min.   :0.0000000   Min.   :0.00000   Min.   :0.00000   Min.   
   :0.00000  
##  1st Qu.:0.0000000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu
   .:0.00000  
##  Median :0.0000000   Median :0.00000   Median :0.00000   Median 
   :0.00000  
##  Mean   :0.0008588   Mean   :0.01889   Mean   :0.02525   Mean   
   :0.01563  
##  3rd Qu.:0.0000000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu
   .:0.00000  
##  Max.   :3.0000000   Max.   :6.00000   Max.   :1.00000   Max.   
   :6.00000  
##       V64                V65             V66                V67          
##  Min.   :0.00000   Min.   :0.000   Min.   :0.00000   Min.   :0.00000  
##  1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.00000  
##  Median :0.00000   Median :0.000   Median :0.00000   Median :0.00000  
##  Mean   :0.04758   Mean   :0.403   Mean   :0.01477   Mean   :0.02061  
##  3rd Qu.:0.00000   3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :5.00000   Max.   :2.000   Max.   :5.00000   Max.   :1.00000  
```

```
##        V68                V69                 V70                 V71
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :1.0000   Median :0.00000   Median :0.00000   Median :0.000000
##  Mean   :0.5622   Mean   :0.01048   Mean   :0.04105   Mean   :0.002233
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :7.0000   Max.   :4.00000   Max.   :8.00000   Max.   :3.000000
##        V72                V73                 V74                 V75
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.000000   Median :0.00000
##  Mean   :0.01254   Mean   :0.03367   Mean   :0.006183   Mean   :0.07042
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :3.00000   Max.   :4.00000   Max.   :6.000000   Max.   :2.00000
##        V76                V77                 V78                 V79
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.000000   Min.   
    :0.000000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu
    .:0.000000
##  Median :0.00000   Median :0.000000   Median :0.000000   Median 
    :0.000000
##  Mean   :0.07661   Mean   :0.005325   Mean   :0.006527   Mean   
    :0.004638
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu
    .:0.000000
##  Max.   :8.00000   Max.   :1.000000   Max.   :1.000000   Max.   
    :2.000000
##        V80                V81                 V82                 V83
##  Min.   :0.0000   Min.   :0.0000000   Min.   :0.000000   Min.   
    :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000000   1st Qu.:0.000000   1st Qu
    .:0.00000
##  Median :1.0000   Median :0.0000000   Median :0.000000   Median 
    :0.00000
##  Mean   :0.5701   Mean   :0.0005153   Mean   :0.006012   Mean   
    :0.03178
##  3rd Qu.:1.0000   3rd Qu.:0.0000000   3rd Qu.:0.000000   3rd Qu
    .:0.00000
##  Max.   :7.0000   Max.   :1.0000000   Max.   :2.000000   Max.   
    :3.00000
##        V84                 V85            V86
##  Min.   :0.000000   Min.   :0.00000   0:5474
##  1st Qu.:0.000000   1st Qu.:0.00000   1: 348
##  Median :0.000000   Median :0.00000
##  Mean   :0.007901   Mean   :0.01426
##  3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :2.000000   Max.   :2.00000
```

2.5.1 Detailed summary

```
round(describe(caravan_data), 3)
```

```
##       vars    n   mean    sd median trimmed   mad min max range  skew
    kurtosis
```

```
## V1       1 5822 24.25 12.85      30   24.98 11.86   1  41     40 -0.44
   -1.35
## V2       2 5822  1.11  0.41       1    1.00  0.00   1  10      9  7.42
   99.98
## V3       3 5822  2.68  0.79       3    2.64  1.48   1   5      4  0.18
       0.01
## V4       4 5822  2.99  0.81       3    2.95  0.00   1   6      5  0.47
       0.62
## V5       5 5822  5.77  2.86       7    5.90  2.96   1  10      9 -0.33
   -1.34
## V6       6 5822  0.70  1.00       0    0.52  0.00   0   9      9  2.24
       8.62
## V7       7 5822  4.63  1.72       5    4.63  1.48   0   9      9  0.07
       0.45
## V8       8 5822  1.07  1.02       1    0.96  1.48   0   5      5  0.90
       0.79
## V9       9 5822  3.26  1.60       3    3.32  1.48   0   9      9 -0.13
   -0.03
## V10     10 5822  6.18  1.91       6    6.33  1.48   0   9      9 -0.72
       0.68
## V11     11 5822  0.88  0.97       1    0.76  1.48   0   7      7  1.32
       2.76
## V12     12 5822  2.29  1.72       2    2.14  1.48   0   9      9  0.69
       0.71
## V13     13 5822  1.89  1.80       2    1.66  1.48   0   9      9  0.97
       0.82
## V14     14 5822  3.23  1.62       3    3.22  1.48   0   9      9  0.18
       0.40
## V15     15 5822  4.30  2.00       4    4.26  1.48   0   9      9  0.18
   -0.21
## V16     16 5822  1.46  1.62       1    1.20  1.48   0   9      9  1.36
       1.99
## V17     17 5822  3.35  1.76       3    3.33  1.48   0   9      9  0.19
       0.21
## V18     18 5822  4.57  2.30       5    4.58  2.96   0   9      9 -0.05
   -0.61
## V19     19 5822  1.90  1.80       2    1.64  1.48   0   9      9  1.17
       1.42
## V20     20 5822  0.40  0.78       0    0.23  0.00   0   5      5  2.85
   11.09
## V21     21 5822  0.52  1.06       0    0.27  0.00   0   9      9  2.83
   10.38
## V22     22 5822  2.90  1.84       3    2.80  1.48   0   9      9  0.66
       0.80
## V23     23 5822  2.22  1.73       2    2.06  1.48   0   9      9  0.68
       0.32
## V24     24 5822  2.31  1.69       2    2.18  1.48   0   9      9  0.67
       0.57
## V25     25 5822  1.62  1.72       1    1.33  1.48   0   9      9  1.64
       3.41
## V26     26 5822  1.61  1.33       2    1.49  1.48   0   9      9  1.11
       3.03
## V27     27 5822  2.20  1.53       2    2.12  1.48   0   9      9  0.38
   -0.19
```

```
## V28      28 5822  3.76  1.94       4     3.74  1.48    0    9       9  0.19
         0.09
## V29      29 5822  1.07  1.30       1     0.84  1.48    0    9       9  1.42
         1.98
## V30      30 5822  4.24  3.09       4     4.17  4.45    0    9       9  0.15
    -1.30
## V31      31 5822  4.77  3.09       5     4.84  4.45    0    9       9 -0.16
    -1.30
## V32      32 5822  6.04  1.55       6     6.04  1.48    0    9       9 -0.24
         0.62
## V33      33 5822  1.32  1.20       1     1.18  1.48    0    7       7  0.77
         0.31
## V34      34 5822  1.96  1.60       2     1.83  1.48    0    9       9  0.73
         0.86
## V35      35 5822  6.28  1.98       7     6.45  1.48    0    9       9 -0.69
         0.20
## V36      36 5822  2.73  1.98       2     2.56  1.48    0    9       9  0.68
         0.18
## V37      37 5822  2.57  2.09       2     2.39  2.96    0    9       9  0.60
    -0.17
## V38      38 5822  3.54  1.88       4     3.53  1.48    0    9       9  0.18
         0.17
## V39      39 5822  2.73  1.93       3     2.61  1.48    0    9       9  0.66
         0.71
## V40      40 5822  0.80  1.16       0     0.56  0.00    0    9       9  1.91
         4.76
## V41      41 5822  0.20  0.55       0     0.07  0.00    0    9       9  4.21
    28.86
## V42      42 5822  3.78  1.32       4     3.68  1.48    0    9       9  0.82
         1.44
## V43      43 5822  4.24  2.01       4     4.20  1.48    1    8       7  0.22
    -0.88
## V44      44 5822  0.77  0.96       0     0.71  0.00    0    3       3  0.48
    -1.71
## V45      45 5822  0.04  0.36       0     0.00  0.00    0    6       6 10.27
    116.60
## V46      46 5822  0.07  0.50       0     0.00  0.00    0    4       4  6.99
    47.91
## V47      47 5822  2.97  2.92       5     2.95  1.48    0    8       8 -0.01
    -1.96
## V48      48 5822  0.05  0.53       0     0.00  0.00    0    7       7 10.99
    119.69
## V49      49 5822  0.17  0.90       0     0.00  0.00    0    7       7  5.13
    25.46
## V50      50 5822  0.01  0.24       0     0.00  0.00    0    9       9 26.91
    754.32
## V51      51 5822  0.02  0.21       0     0.00  0.00    0    5       5 11.73
    159.54
## V52      52 5822  0.09  0.60       0     0.00  0.00    0    6       6  6.82
    47.82
## V53      53 5822  0.01  0.23       0     0.00  0.00    0    6       6 19.22
    398.36
## V54      54 5822  0.22  0.81       0     0.00  0.00    0    6       6  3.70
    12.62
```

```
## V55     55 5822  0.20  0.90        0     0.00  0.00    0   9      9  4.88
    24.25
## V56     56 5822  0.01  0.21        0     0.00  0.00    0   6      6 18.62
    404.12
## V57     57 5822  0.01  0.19        0     0.00  0.00    0   3      3 13.04
    174.68
## V58     58 5822  0.02  0.38        0     0.00  0.00    0   7      7 15.99
    255.43
## V59     59 5822  1.83  1.88        2     1.68  2.96    0   8      8  0.39
    -1.23
## V60     60 5822  0.00  0.04        0     0.00  0.00    0   3      3 60.61
    3987.56
## V61     61 5822  0.02  0.27        0     0.00  0.00    0   6      6 15.91
    269.40
## V62     62 5822  0.03  0.16        0     0.00  0.00    0   1      1  6.05
    34.62
## V63     63 5822  0.02  0.20        0     0.00  0.00    0   6      6 16.65
    330.21
## V64     64 5822  0.05  0.41        0     0.00  0.00    0   5      5  8.82
    78.19
## V65     65 5822  0.40  0.49        0     0.38  0.00    0   2      2  0.42
    -1.75
## V66     66 5822  0.01  0.13        0     0.00  0.00    0   5      5 14.33
    365.23
## V67     67 5822  0.02  0.14        0     0.00  0.00    0   1      1  6.75
    43.52
## V68     68 5822  0.56  0.60        1     0.51  1.48    0   7      7  0.98
        3.61
## V69     69 5822  0.01  0.13        0     0.00  0.00    0   4      4 16.73
    354.31
## V70     70 5822  0.04  0.23        0     0.00  0.00    0   8      8 10.95
    268.08
## V71     71 5822  0.00  0.06        0     0.00  0.00    0   3      3 33.84
    1304.72
## V72     72 5822  0.01  0.13        0     0.00  0.00    0   3      3 12.22
    187.68
## V73     73 5822  0.03  0.24        0     0.00  0.00    0   4      4  9.45
    111.64
## V74     74 5822  0.01  0.12        0     0.00  0.00    0   6      6 29.44
    1121.91
## V75     75 5822  0.07  0.26        0     0.00  0.00    0   2      2  3.74
    13.67
## V76     76 5822  0.08  0.38        0     0.00  0.00    0   8      8  6.70
    65.75
## V77     77 5822  0.00  0.07        0     0.00  0.00    0   1      1 13.59
    182.75
## V78     78 5822  0.01  0.08        0     0.00  0.00    0   1      1 12.25
    148.16
## V79     79 5822  0.00  0.08        0     0.00  0.00    0   2      2 18.71
    389.67
## V80     80 5822  0.57  0.56        1     0.55  0.00    0   7      7  0.75
        3.97
## V81     81 5822  0.00  0.02        0     0.00  0.00    0   1      1 44.01
    1935.00
```

```
## V82     82 5822  0.01  0.08      0    0.00  0.00   0   2     2 14.62
   236.35
## V83     83 5822  0.03  0.21      0    0.00  0.00   0   3     3  7.54
   63.14
## V84     84 5822  0.01  0.09      0    0.00  0.00   0   2     2 11.80
   146.72
## V85     85 5822  0.01  0.12      0    0.00  0.00   0   2     2  8.49
   73.24
## V86*    86 5822  1.06  0.24      1    1.00  0.00   1   2     1  3.71
   11.79
##          se
## V1    0.17
## V2    0.00
## V3    0.01
## V4    0.01
## V5    0.04
## V6    0.01
## V7    0.02
## V8    0.01
## V9    0.02
## V10   0.03
## V11   0.01
## V12   0.02
## V13   0.02
## V14   0.02
## V15   0.03
## V16   0.02
## V17   0.02
## V18   0.03
## V19   0.02
## V20   0.01
## V21   0.01
## V22   0.02
## V23   0.02
## V24   0.02
## V25   0.02
## V26   0.02
## V27   0.02
## V28   0.03
## V29   0.02
## V30   0.04
## V31   0.04
## V32   0.02
## V33   0.02
## V34   0.02
## V35   0.03
## V36   0.03
## V37   0.03
## V38   0.03
## V39   0.03
## V40   0.01
## V41   0.01
## V42   0.02
## V43   0.03
```

```
## V44   0.01
## V45   0.00
## V46   0.01
## V47   0.04
## V48   0.01
## V49   0.01
## V50   0.00
## V51   0.00
## V52   0.01
## V53   0.00
## V54   0.01
## V55   0.01
## V56   0.00
## V57   0.00
## V58   0.00
## V59   0.03
## V60   0.00
## V61   0.00
## V62   0.00
## V63   0.00
## V64   0.00
## V65   0.01
## V66   0.00
## V67   0.00
## V68   0.01
## V69   0.00
## V70   0.00
## V71   0.00
## V72   0.00
## V73   0.00
## V74   0.00
## V75   0.00
## V76   0.00
## V77   0.00
## V78   0.00
## V79   0.00
## V80   0.01
## V81   0.00
## V82   0.00
## V83   0.00
## V84   0.00
## V85   0.00
## V86*  0.00
```

2.5.2 Observations

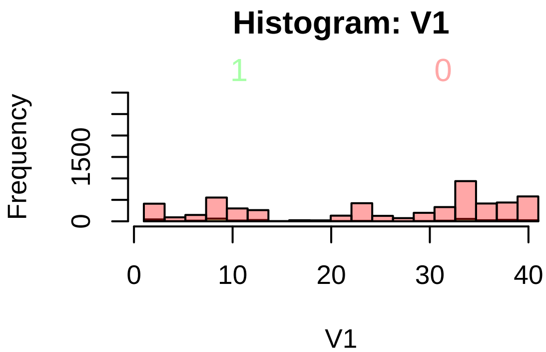
- Some of the predictors have a very high kurtosis value which means that the predictors have a heavy tail as compared to a normal distribution resulting in a lot of outliers. We will look at each variable later with respect to their distribution and outliers.
- V1, i.e. Customer Subtype as mentioned in the data dictionary has 41 different categories detailed in the link. Similarly, V4, i.e. Avg age can be identified as categories as mentioned in the data dictionary.
- One major observation from the dataset is that all the predictors have discretised by the insurance company so there is not much of wrangling to be done prioir to the model buildidng process.
- We can see that there is pattern in which the predictors have been observed.

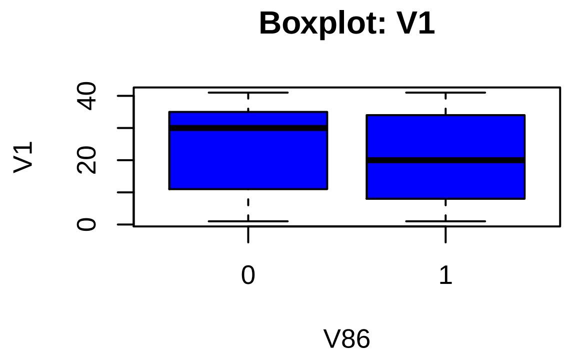
- V1 has been identified as 41 different categories,
- V2 (number of houses) ranges from 1 to 10, and is heavy tailed towards the right which means the dataset has customers with more houses than a normally distributed one.
- V3 (avg size household) ranges from 1-6,
- V4 (avg age) is identified as 6 different categories,
- V5 (customer main type) L2 is identified as 10 different categories
- V6 to V43, i.e. the rest of the socio-demographic data is identified based on zipcodes, and as these are precentages mentioned in the data dictionary, it is likely to explain the percentage of people belong to that particular category in that customer's zipcode. For example, if V6 explains Roman Catholic as 7, it means that, 76-88% of people in that zipcode are Roman Catholic
- V44 to V64 (number of policies), means that the number of policies in that category held by the customer in the range of 1 to 12, few of the predictors in this bracket are also heavy tailed.
- V65 to V85 (contribution to policies) means the amount category contributed by a customer as part of that policy held. Similar to the previous predictors, this is also heavy tailed as compared to a normal distribution, which implies that customers tend to contribute more to certain categories of insurance.

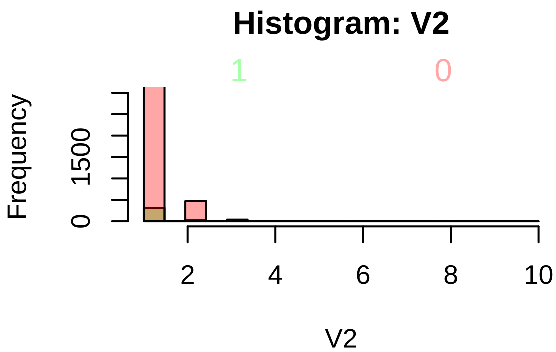
2.6 Distributions of the predictors with respect to the target

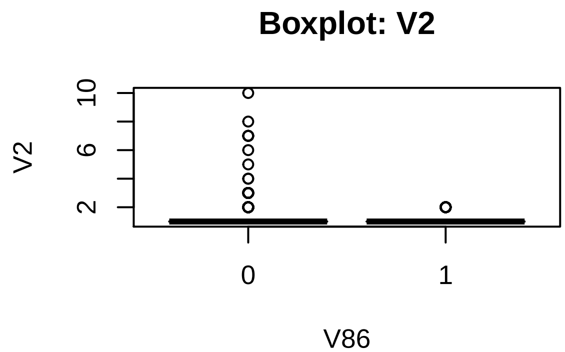
2.6.1 Plots

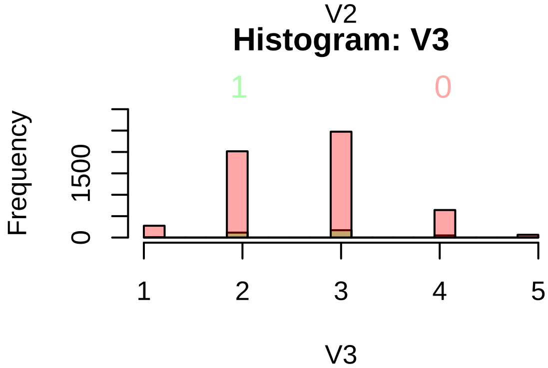
```
# Define a two-row by two-column plotting area.
par(mfrow = c(2, 2))
d=caravan_data
# Plot a histogram and box plot for each of the predictors,
# by response.
for (x in colnames(d[-ncol(d)])) {
    min_d <- min(d[ , x])
    max_d <- max(d[ , x])
    b <- seq(min_d, max_d, length.out = 20)

    hist(d[ , x][d$V86 == 1], col = rgb(0, 1, 0, 0.35), breaks = b,
         main = paste("Histogram:␣", x, sep = ""), xlab = x,ylim=c(0,3000)
             )

    hist(d[ , x][d$V86 == 0], col = rgb(1, 0, 0, 0.35), breaks = b,
         add = TRUE,ylim=c(0,3000))

    mtext(c("1", "0"), adj = c(0.25, 0.75), col = c(rgb(0, 1, 0, 0.35),rgb
        (1, 0, 0, 0.35)))

    boxplot(d[ , x] ~ d$V86,
            col="blue",
            main = paste("Boxplot:␣", x, sep = ""),
            xlab="V86",
            ylab=x)
}
```

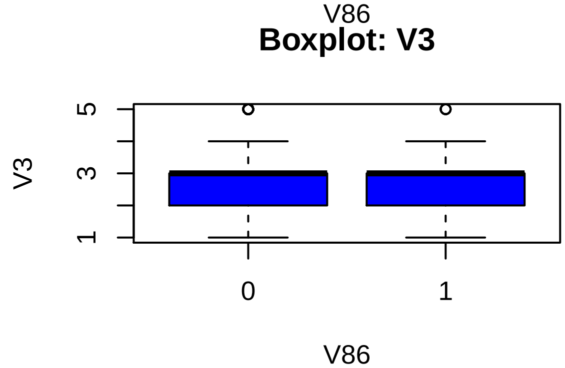

Histogram: V1
1
0
Frequency
0
1500
0
10
20
30
40
V1

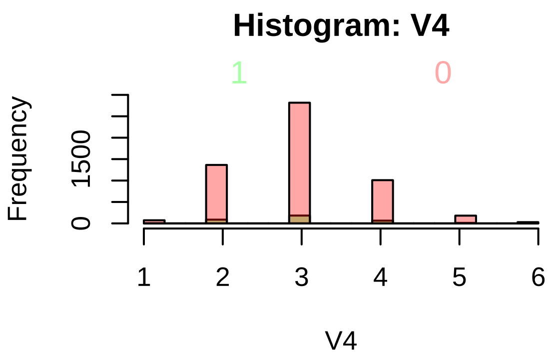

Boxplot: V1
V1
0
20
40
0
1
V86

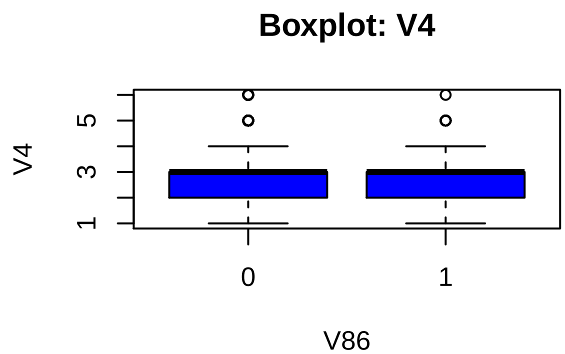

Histogram: V2
1
0
Frequency
0
1500
2
4
6
8
10
V2

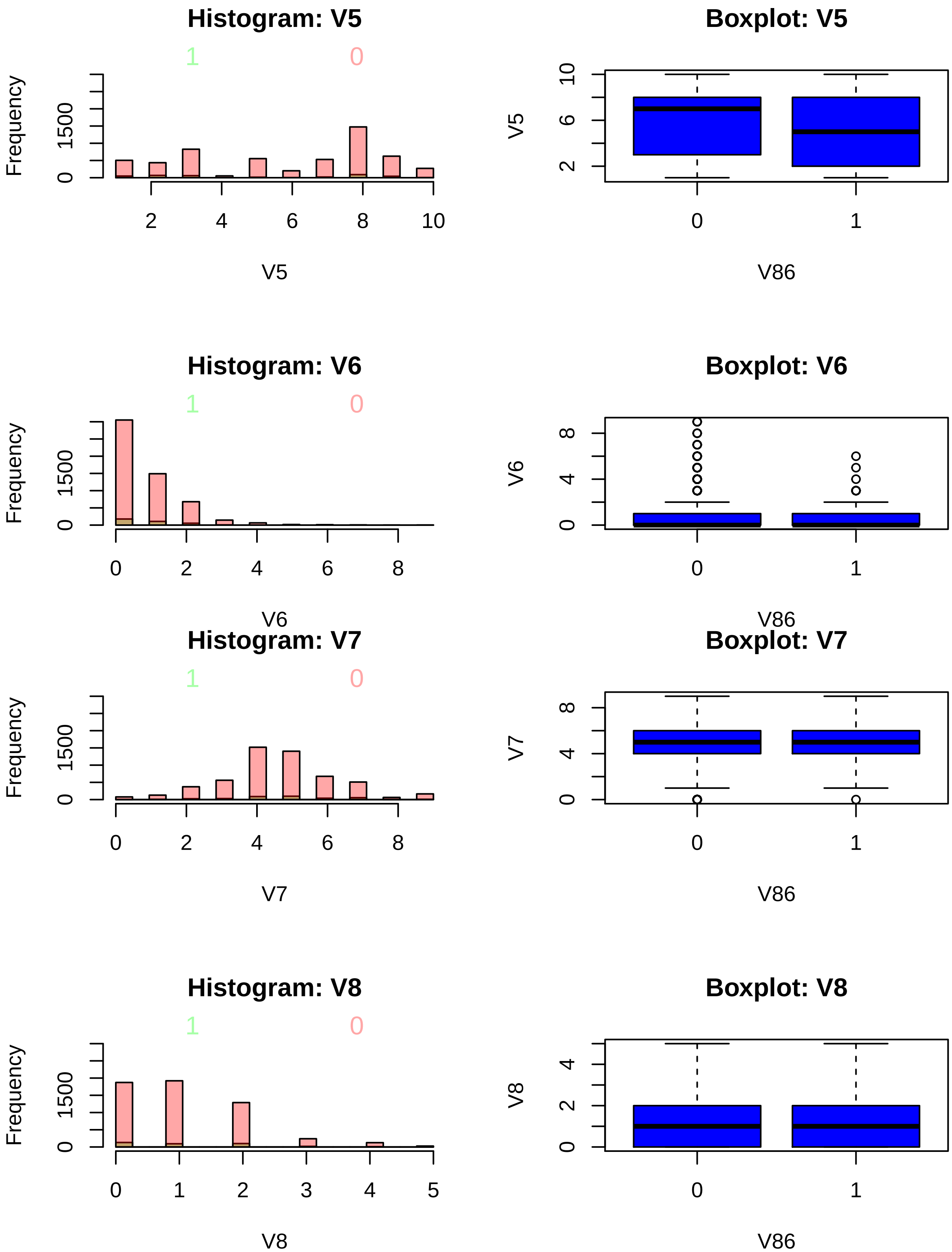

Boxplot: V2
V2
2
6
10
0
1
V86

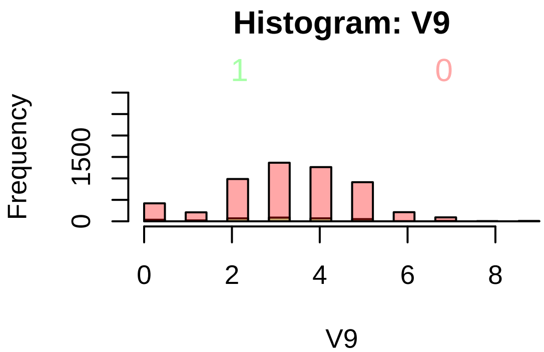

Histogram: V3
1
0
Frequency
0
1500
1
2
3
4
5
V3

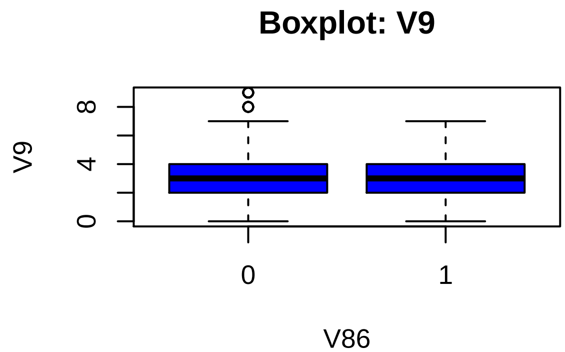

Boxplot: V3
V3
1
3
5
0
1
V86

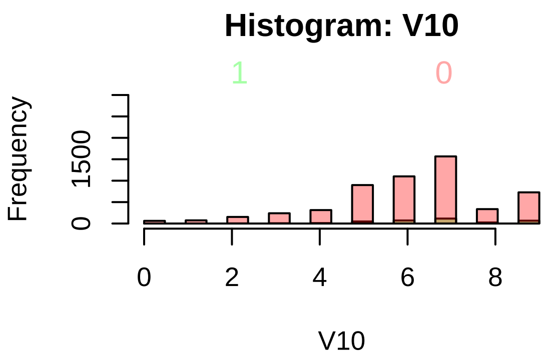

Histogram: V4
1
0
Frequency
0
1500
1
2
3
4
5
6
V4

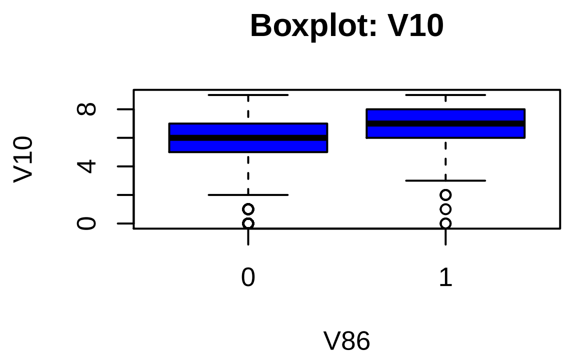

Boxplot: V4
V4
1
3
5
0
1
V86

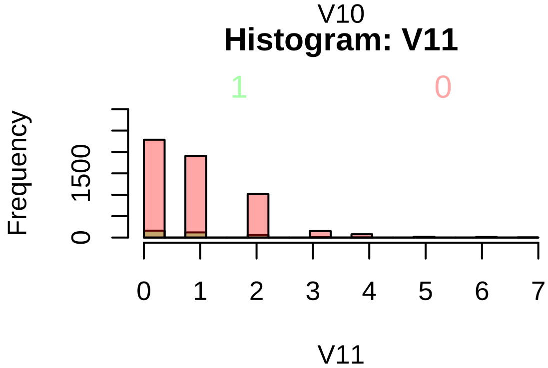

Histogram: V5
1
0
Frequency
0
1500
2
4
6
8
10
V5
Boxplot: V5
V5
2
6
10
0
1
V86
Histogram: V6
1
0
Frequency
0
1500
0
2
4
6
8
V6
Boxplot: V6
V6
0
4
8
0
1
V86
Histogram: V7
1
0
Frequency
0
1500
0
2
4
6
8
V7
Boxplot: V7
V7
0
4
8
0
1
V86
Histogram: V8
1
0
Frequency
0
1500
0
1
2
3
4
5
V8
Boxplot: V8
V8
0
2
4
0
1
V86

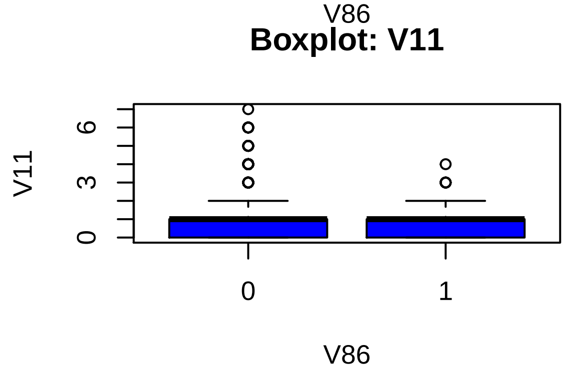

Histogram: V9
1
0
Frequency
0
1500
0
2
4
6
8
V9

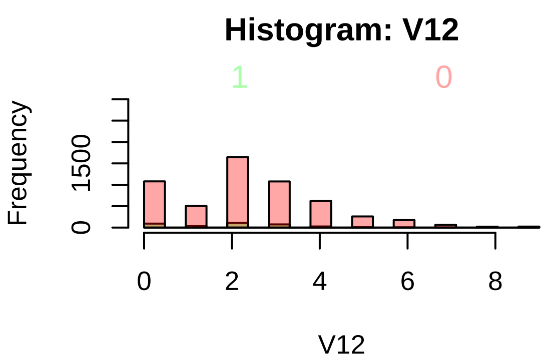

Boxplot: V9
V9
0
4
8
0
1
V86

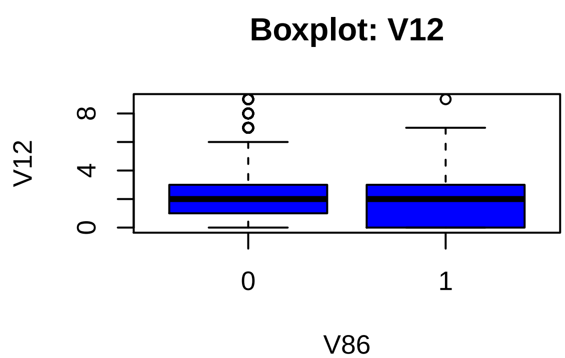

Histogram: V10
1
0
Frequency
0
1500
0
2
4
6
8
V10

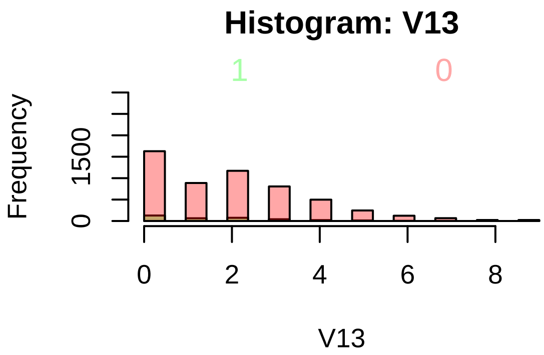

Boxplot: V10
V10
0
4
8
0
1
V86

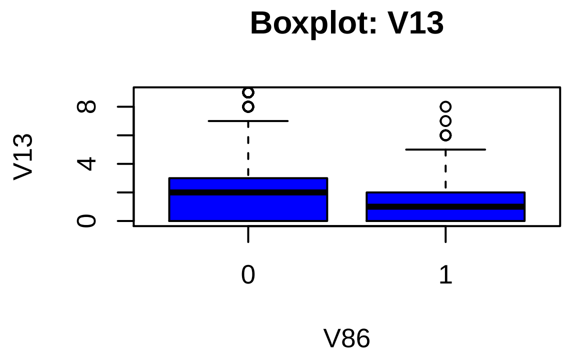

Histogram: V11
1
0
Frequency
0
1500
0
1
2
3
4
5
6
7
V11

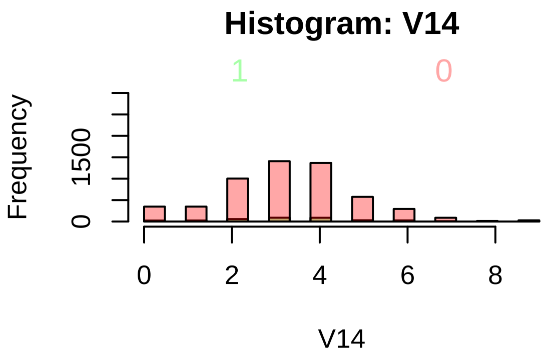

Boxplot: V11
V11
0
3
6
0
1
V86

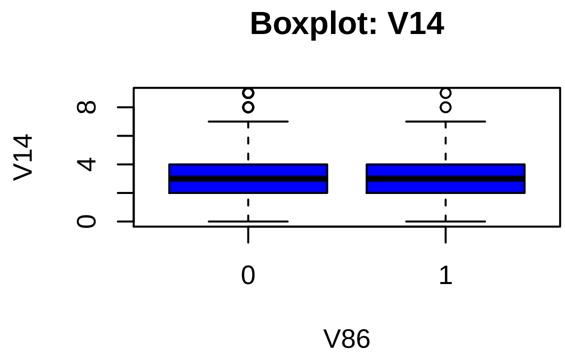

Histogram: V12
1
0
Frequency
0
1500
0
2
4
6
8
V12

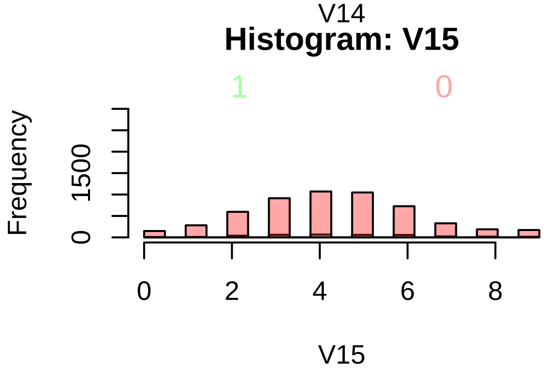

Boxplot: V12
V12
0
4
8
0
1
V86

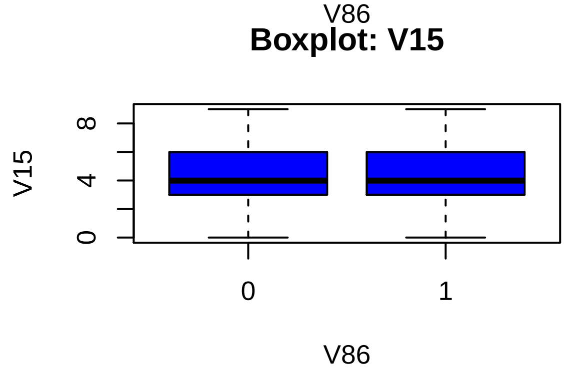

Histogram: V13
1
0
Frequency
0
1500
0
2
4
6
8
V13

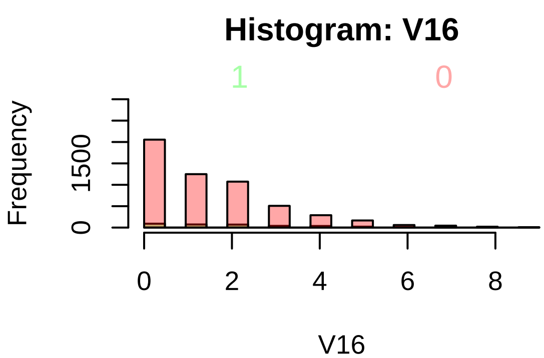

Boxplot: V13
V13
0
4
8
0
1
V86

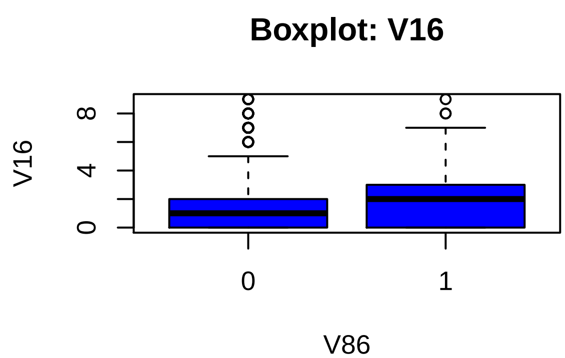

Histogram: V14
1
0
Frequency
0
1500
0
2
4
6
8
V14

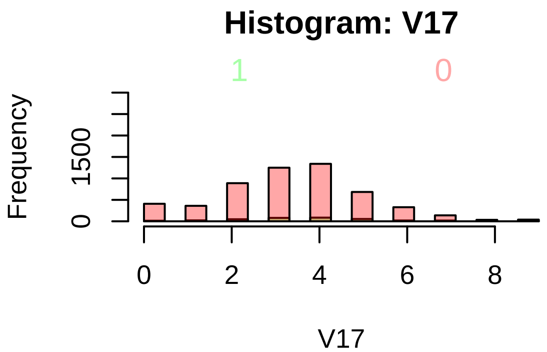

Boxplot: V14
V14
0
4
8
0
1
V86

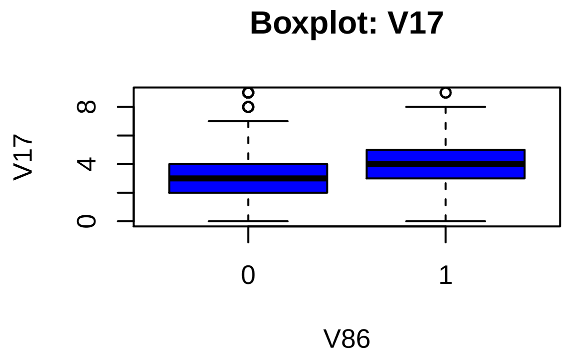

Histogram: V15
1
0
Frequency
0
1500
0
2
4
6
8
V15

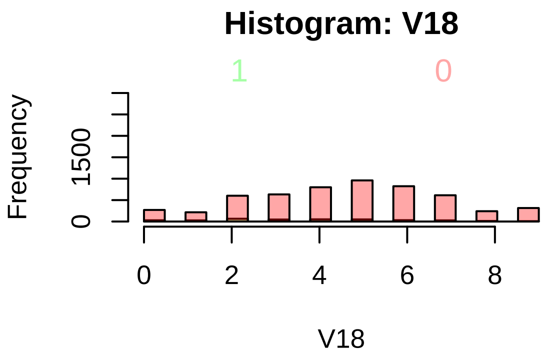

Boxplot: V15
V15
0
4
8
0
1
V86

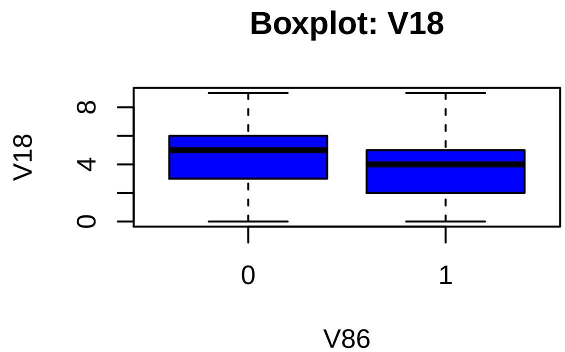

Histogram: V16
1
0
Frequency
0
1500
0
2
4
6
8
V16

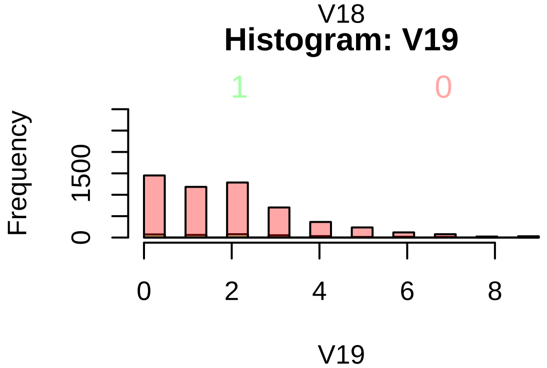

Boxplot: V16
V16
0
4
8
0
1
V86

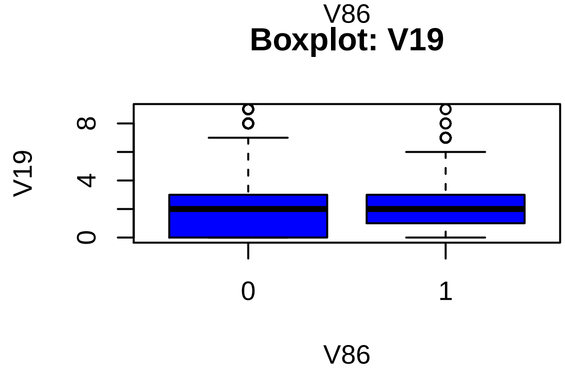

Histogram: V17
1
0
Frequency
0
1500
0
2
4
6
8
V17

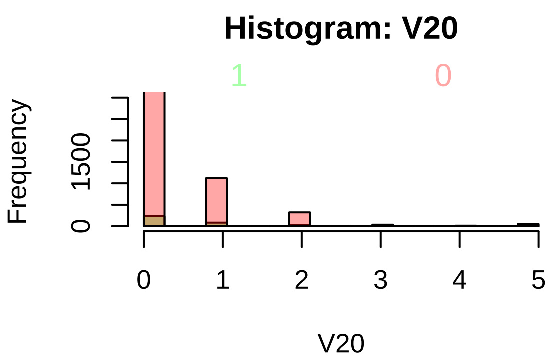

Boxplot: V17
V17
0
4
8
0
1
V86

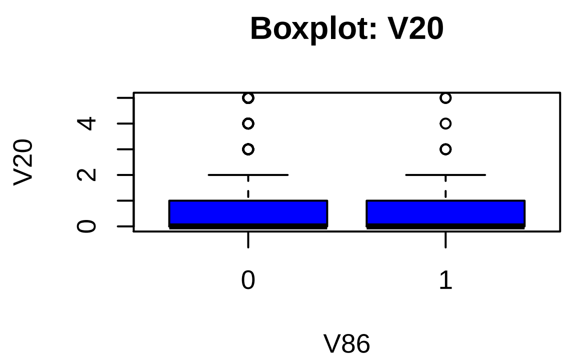

Histogram: V18
1
0
Frequency
0
1500
0
2
4
6
8
V18

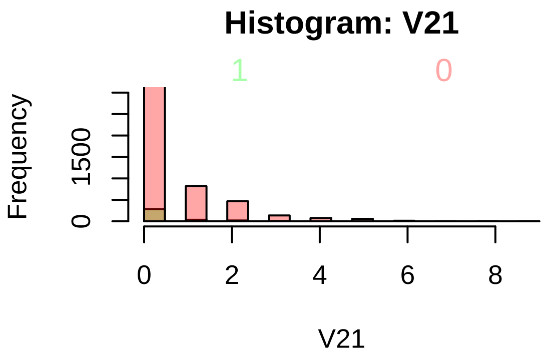

Boxplot: V18
V18
0
4
8
0
1
V86

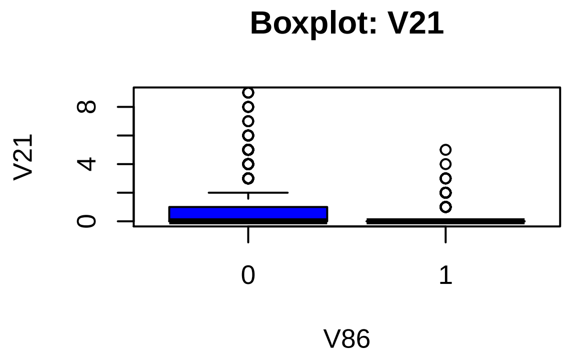

Histogram: V19
1
0
Frequency
0
1500
0
2
4
6
8
V19

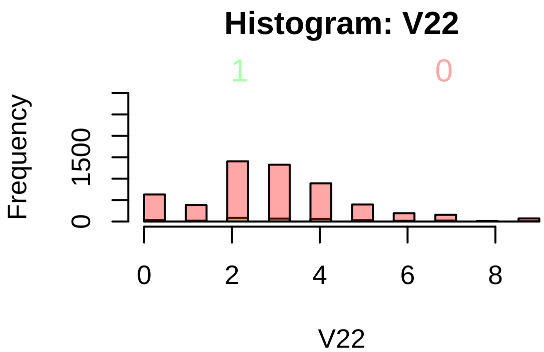

Boxplot: V19
V19
0
4
8
0
1
V86

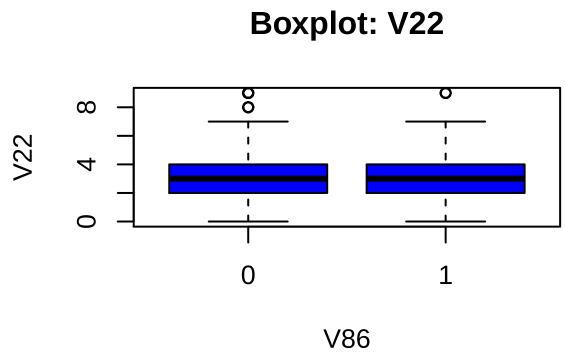

Histogram: V20
1
0
Frequency
0
1500
0
1
2
3
4
5
V20

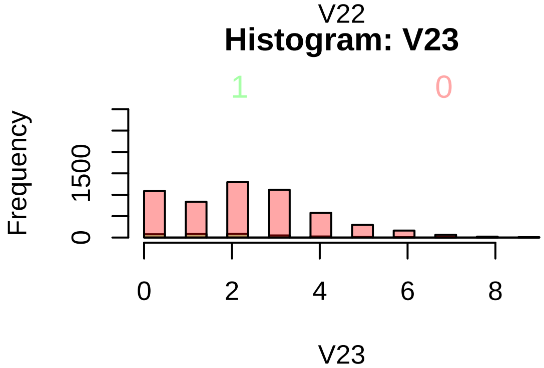

Boxplot: V20
V20
0
2
4
0
1
V86

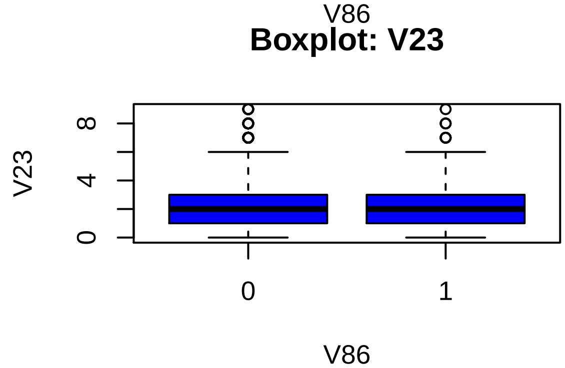

Histogram: V21
1
0
Frequency
0
1500
0
2
4
6
8
V21

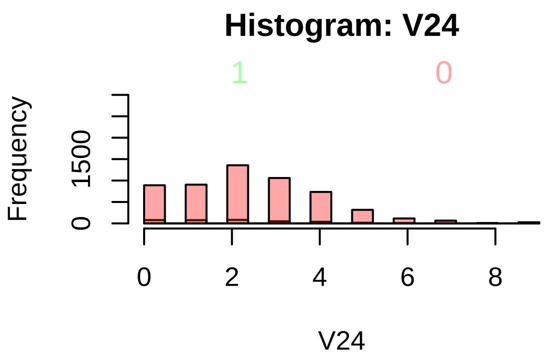

Boxplot: V21
V21
0
4
8
0
1
V86

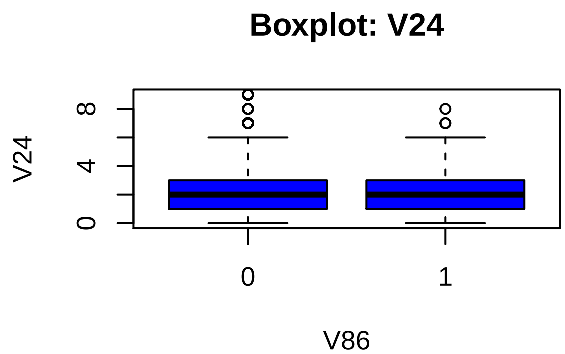

Histogram: V22
1
0
Frequency
0
1500
0
2
4
6
8
V22

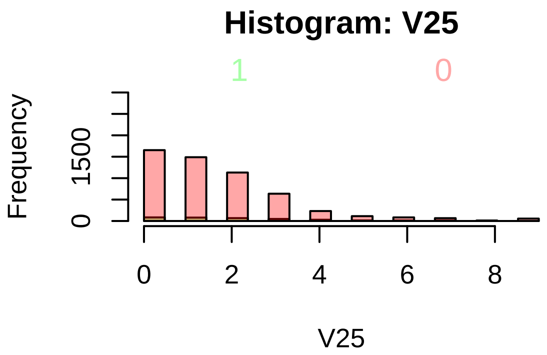

Boxplot: V22
V22
0
4
8
0
1
V86

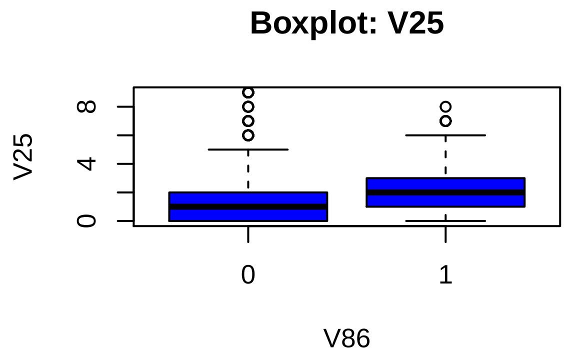

Histogram: V23
1
0
Frequency
0
1500
0
2
4
6
8
V23

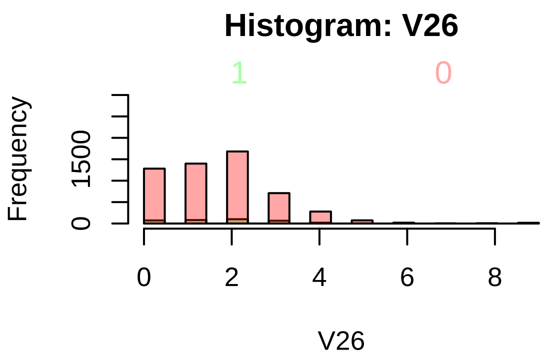

Boxplot: V23
V23
0
4
8
0
1
V86

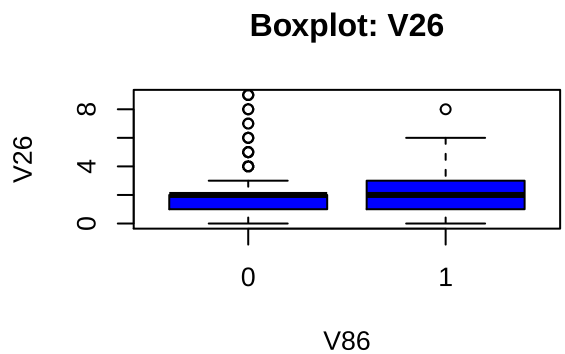

Histogram: V24
1
0
Frequency
0
1500
0
2
4
6
8
V24

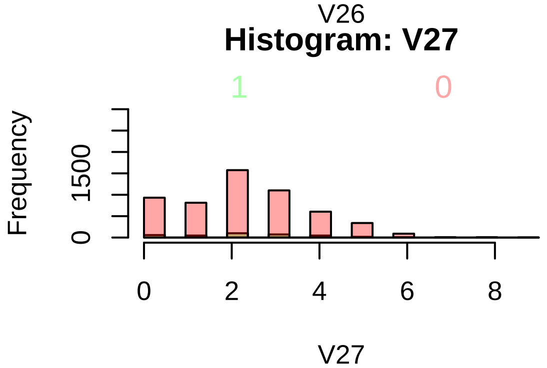

Boxplot: V24
V24
0
4
8
0
1
V86

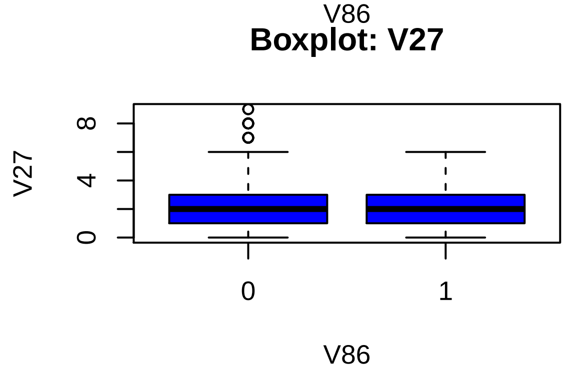

Histogram: V25
1
0
Frequency
0
1500
0
2
4
6
8
V25

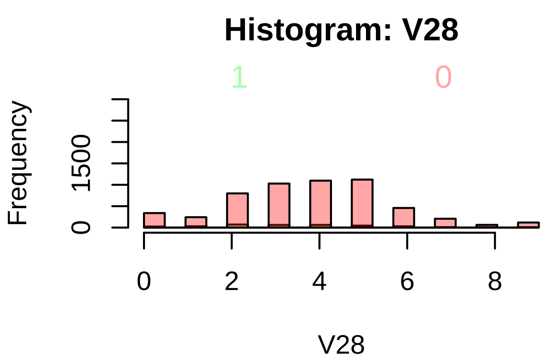

Boxplot: V25
V25
0
4
8
0
1
V86

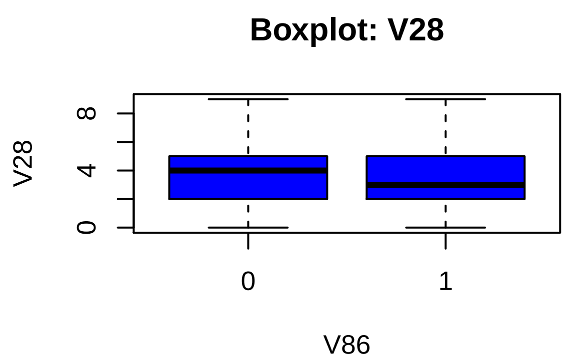

Histogram: V26
1
0
Frequency
0
1500
0
2
4
6
8
V26

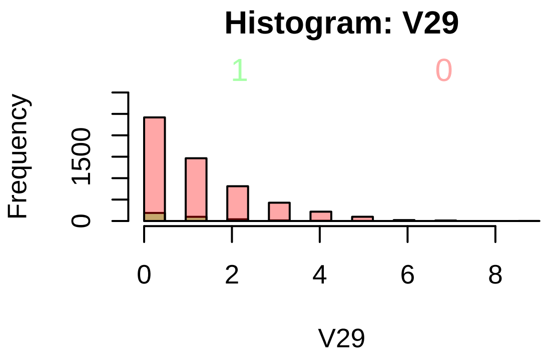

Boxplot: V26
V26
0
4
8
0
1
V86

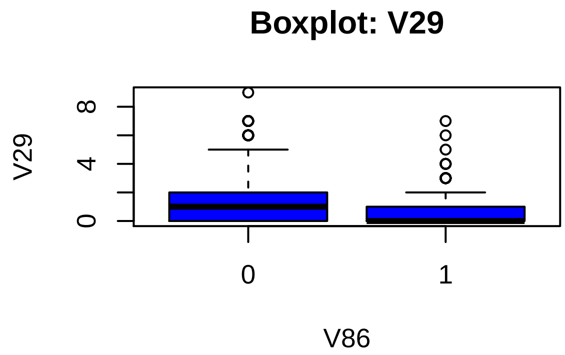

Histogram: V27
1
0
Frequency
0
1500
0
2
4
6
8
V27

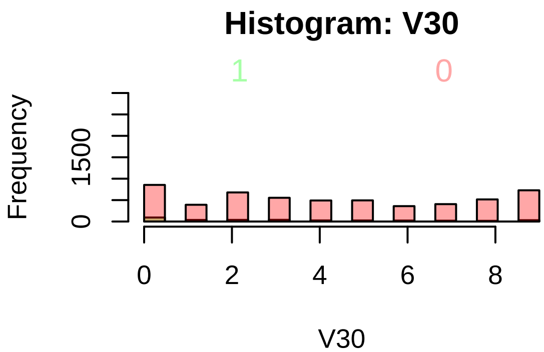

Boxplot: V27
V27
0
4
8
0
1
V86

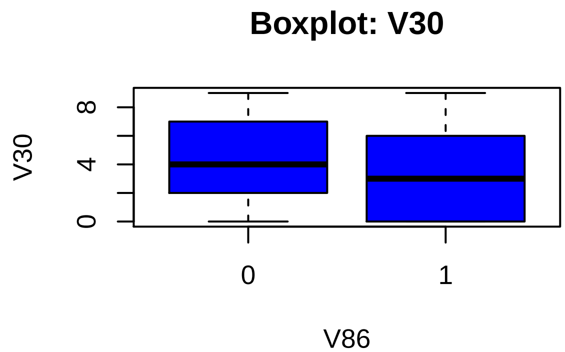

Histogram: V28
1
0
Frequency
0
1500
0
2
4
6
8
V28

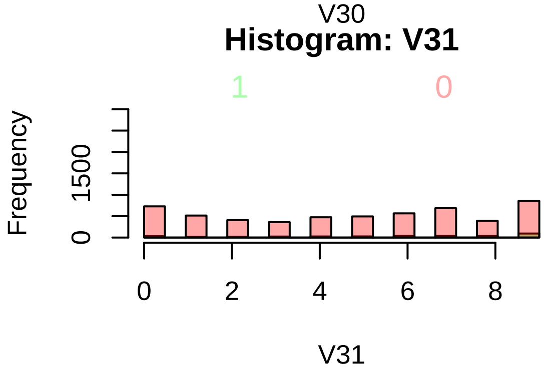

Boxplot: V28
V28
0
4
8
0
1
V86

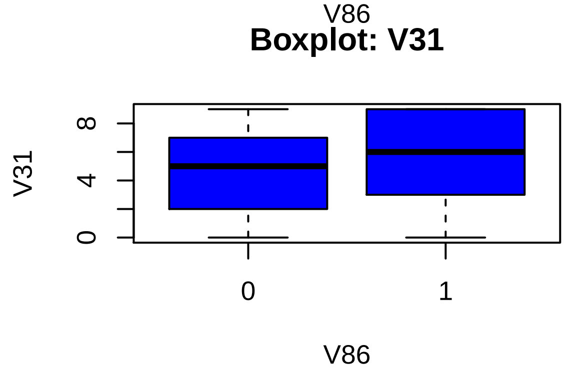

Histogram: V29
1
0
Frequency
0
1500
0
2
4
6
8
V29

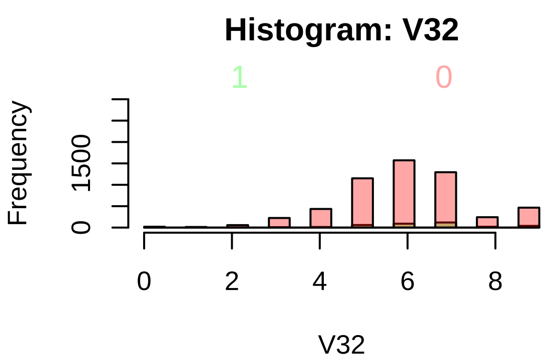

Boxplot: V29
V29
0
4
8
0
1
V86

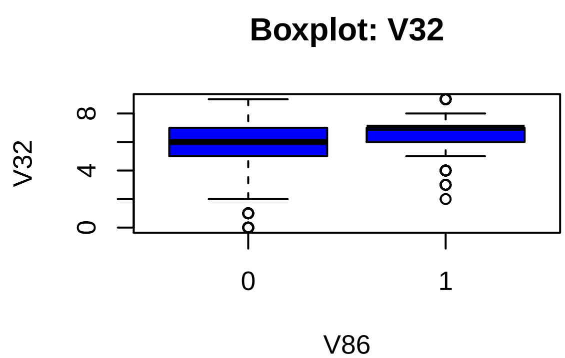

Histogram: V30
1
0
Frequency
0
1500
0
2
4
6
8
V30

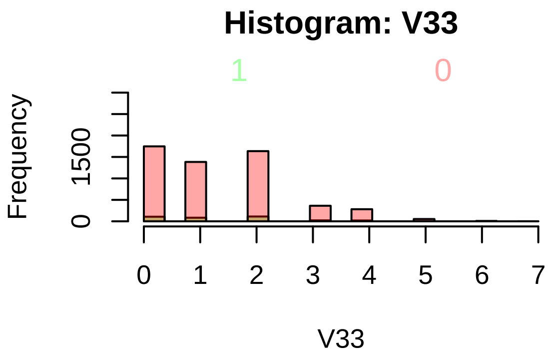

Boxplot: V30
V30
0
4
8
0
1
V86

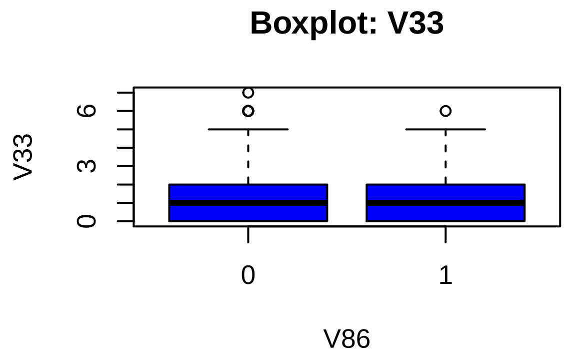

Histogram: V31
1
0
Frequency
0
1500
0
2
4
6
8
V31

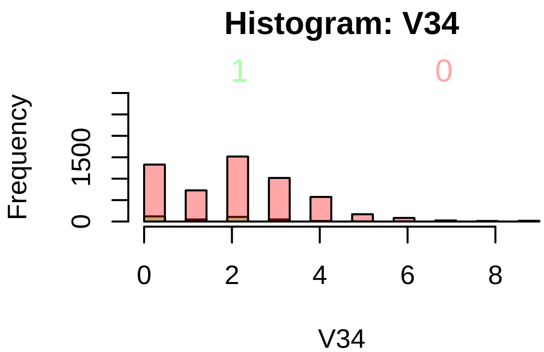

Boxplot: V31
V31
0
4
8
0
1
V86

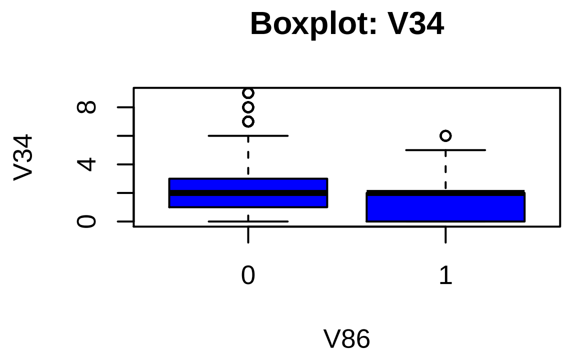

Histogram: V32
1
0
Frequency
0
1500
0
2
4
6
8
V32

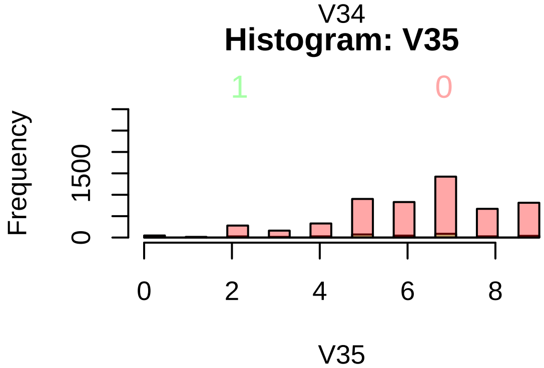

Boxplot: V32
V32
0
4
8
0
1
V86

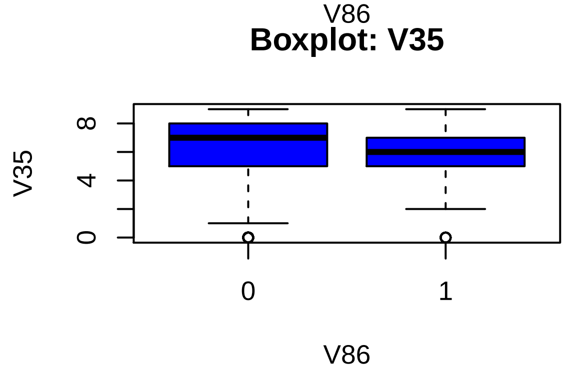

Histogram: V33
1
0
Frequency
0
1500
0 1 2 3 4 5 6 7
V33

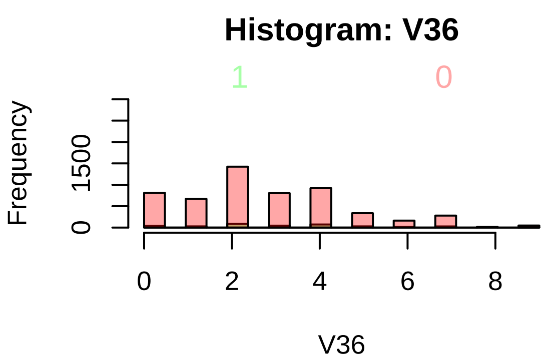

Boxplot: V33
V33
0
3
6
0 1
V86

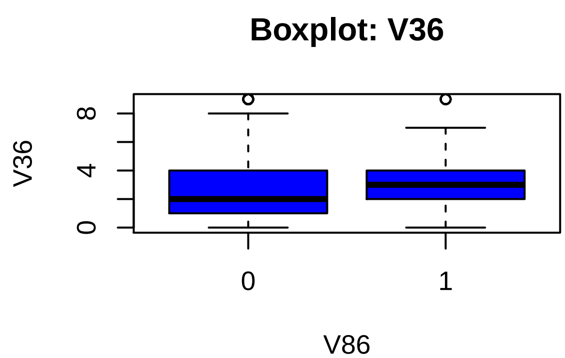

Histogram: V34
1
0
Frequency
0
1500
0 2 4 6 8
V34

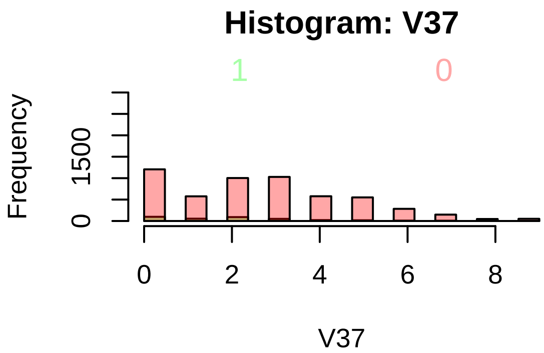

Boxplot: V34
V34
0
4
8
0 1
V86

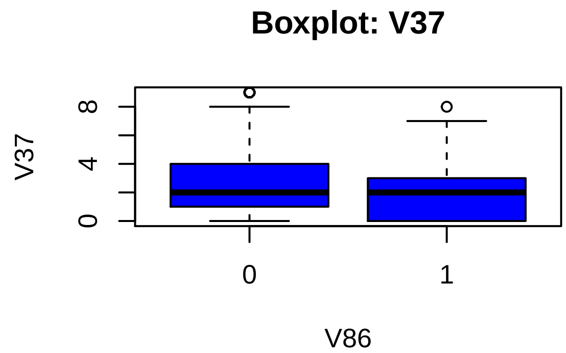

Histogram: V35
1
0
Frequency
0
1500
0 2 4 6 8
V35

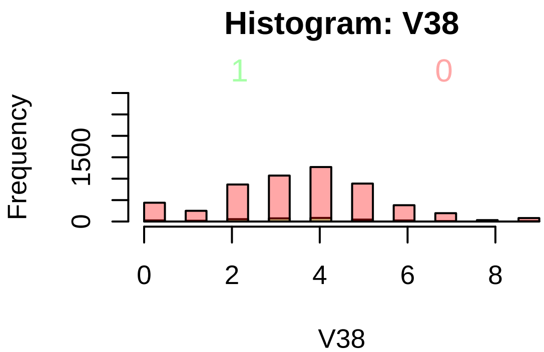

Boxplot: V35
V35
0
4
8
0 1
V86

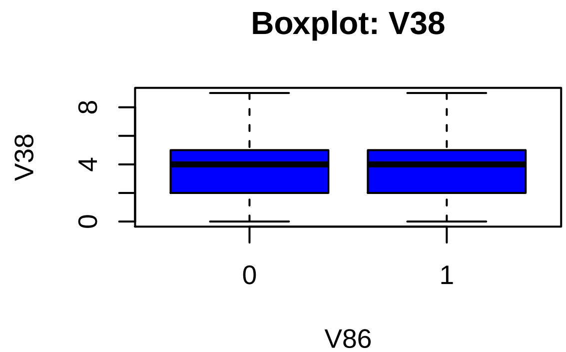

Histogram: V36
1
0
Frequency
0
1500
0 2 4 6 8
V36

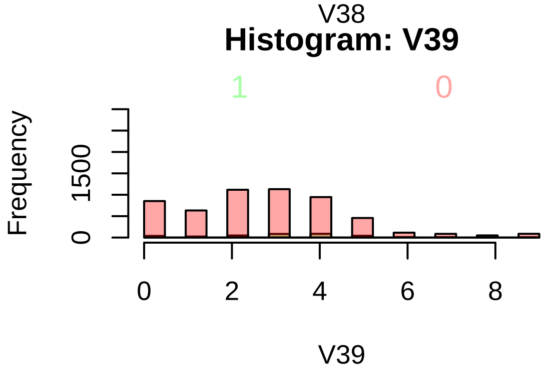

Boxplot: V36
V36
0
4
8
0 1
V86

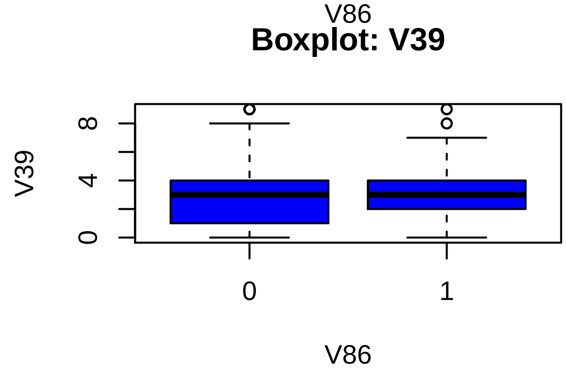

Histogram: V37
1
0
Frequency
0
1500
0
2
4
6
8
V37

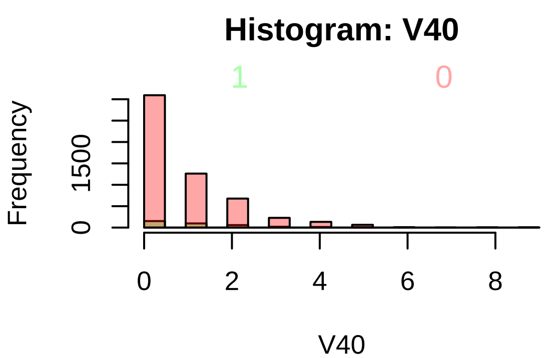

Boxplot: V37
V37
0
4
8
0
1
V86

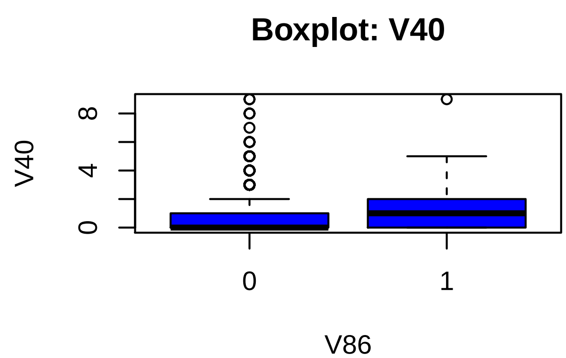

Histogram: V38
1
0
Frequency
0
1500
0
2
4
6
8
V38

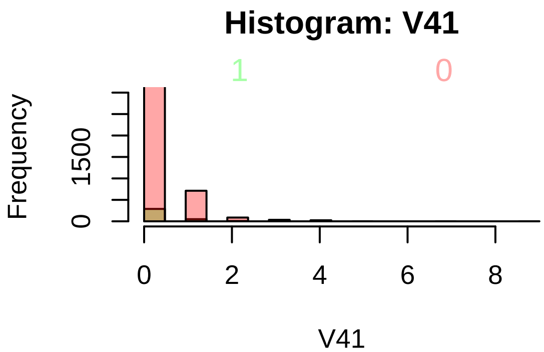

Boxplot: V38
V38
0
4
8
0
1
V86

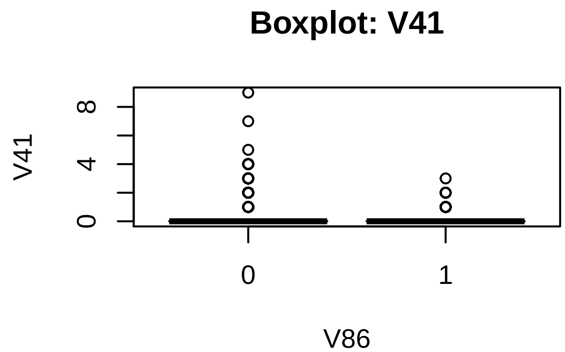

Histogram: V39
1
0
Frequency
0
1500
0
2
4
6
8
V39

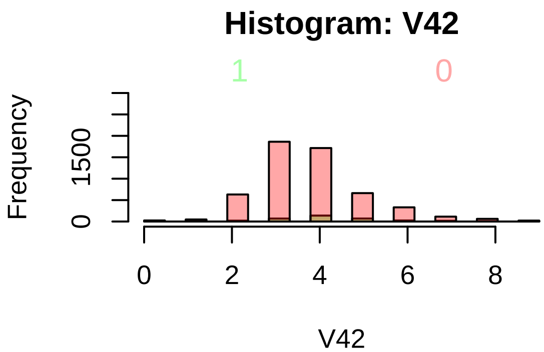

Boxplot: V39
V39
0
4
8
0
1
V86

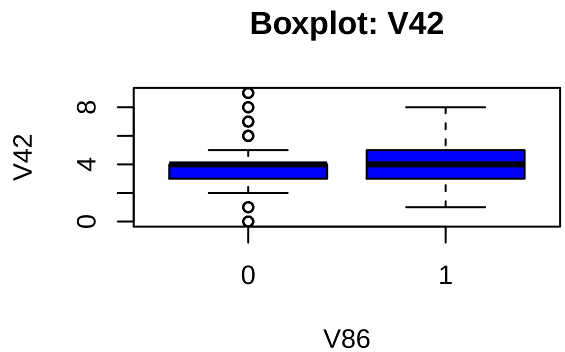

Histogram: V40
1
0
Frequency
0
1500
0
2
4
6
8
V40

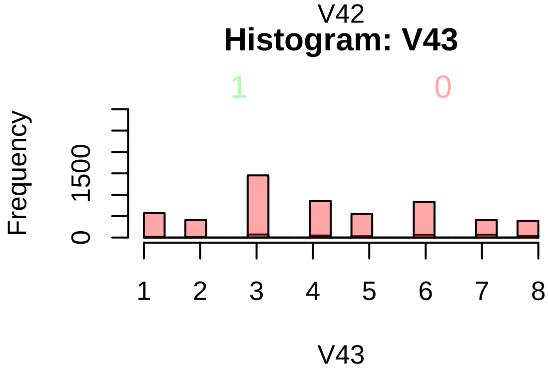

Boxplot: V40
V40
0
4
8
0
1
V86

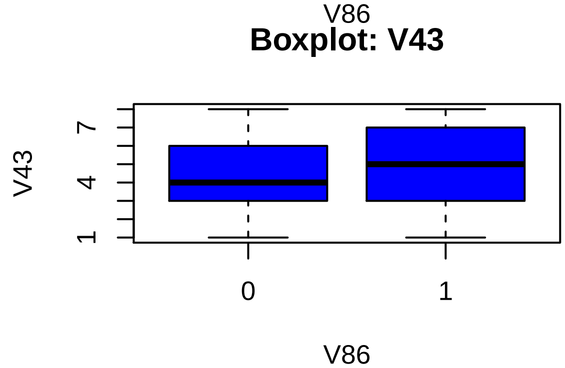

Histogram: V41
1
0
Frequency
0
1500
0 2 4 6 8
V41

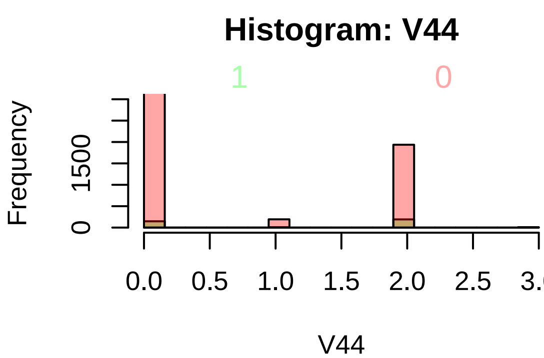

Boxplot: V41
V41
0
4
8
0 1
V86

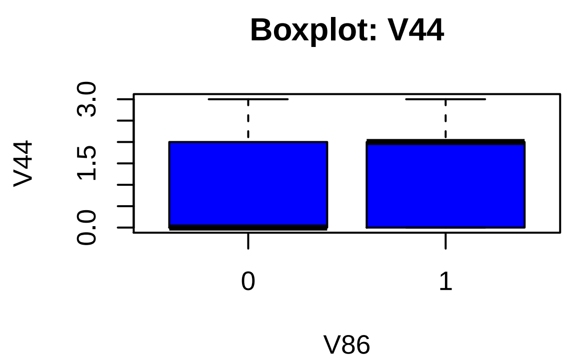

Histogram: V42
1
0
Frequency
0
1500
0 2 4 6 8
V42

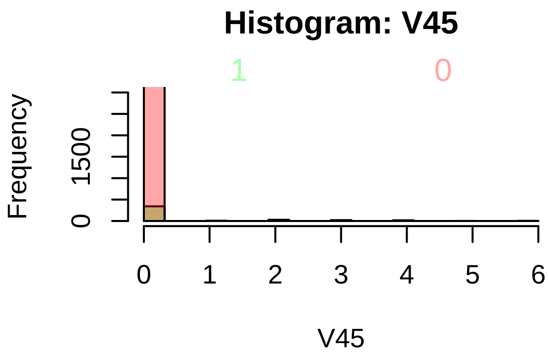

Boxplot: V42
V42
0
4
8
0 1
V86

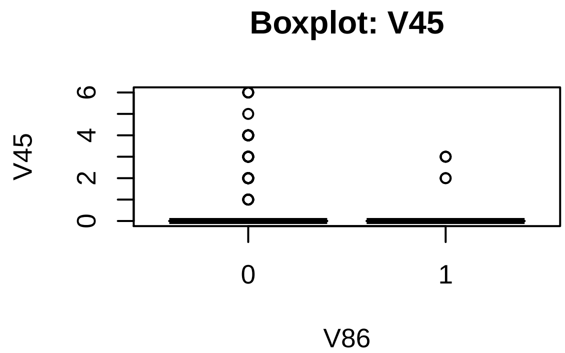

Histogram: V43
1
0
Frequency
0
1500
1 2 3 4 5 6 7 8
V43

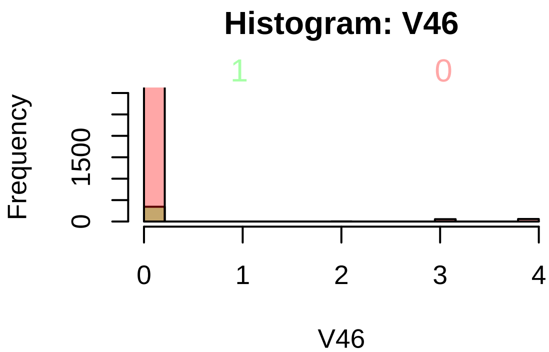

Boxplot: V43
V43
1
4
7
0 1
V86

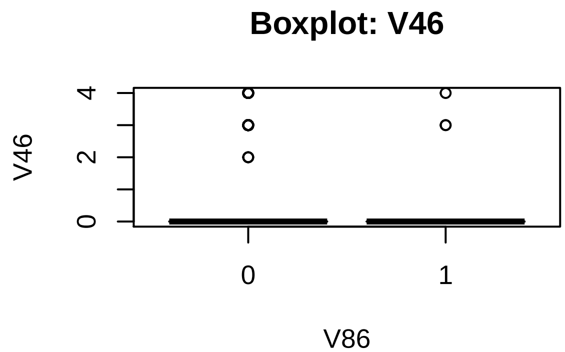

Histogram: V44
1
0
Frequency
0
1500
0.0 0.5 1.0 1.5 2.0 2.5 3.0
V44

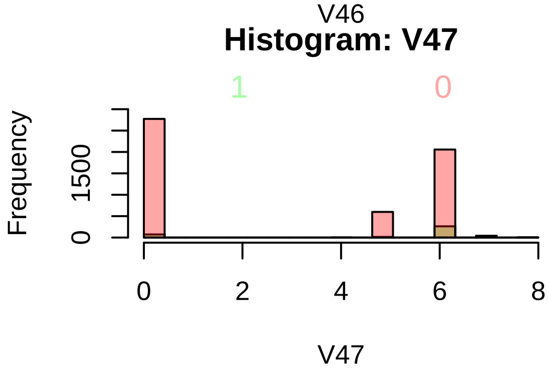

Boxplot: V44
V44
0.0
1.5
3.0
0 1
V86

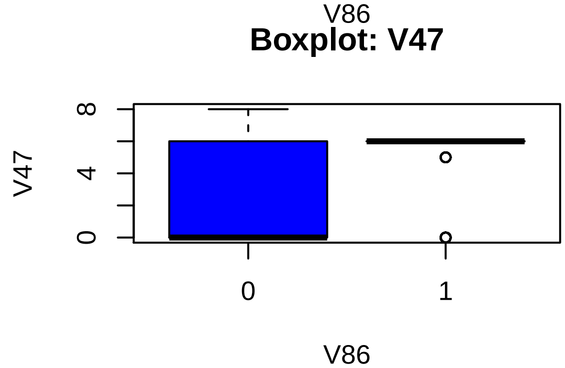

Histogram: V45
1
0
Frequency
0
1500
0 1 2 3 4 5 6
V45

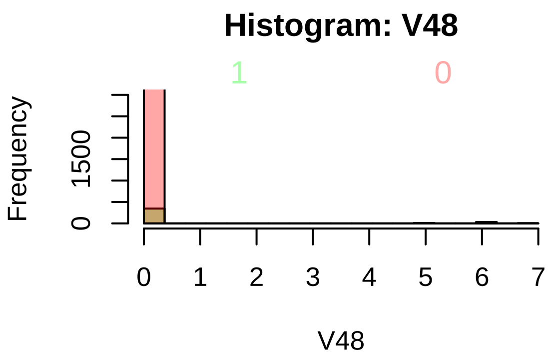

Boxplot: V45
V45
0 2 4 6
0 1
V86

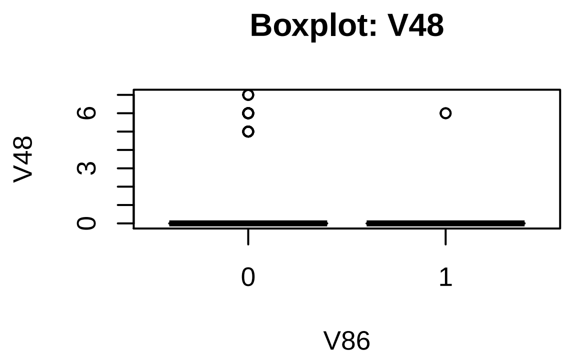

Histogram: V46
1
0
Frequency
0
1500
0 1 2 3 4
V46

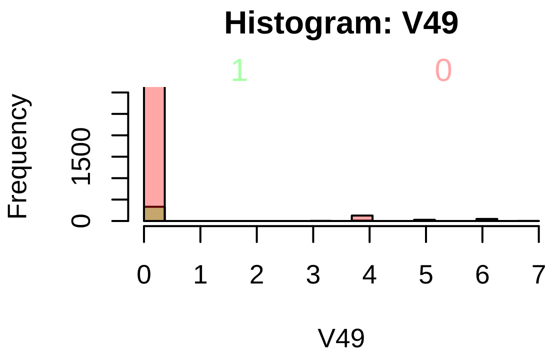

Boxplot: V46
V46
0 2 4
0 1
V86

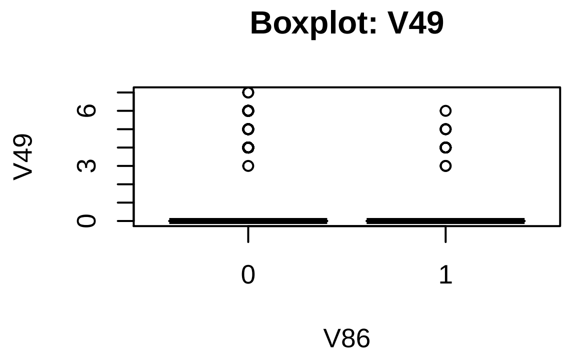

Histogram: V47
1
0
Frequency
0
1500
0 2 4 6 8
V47

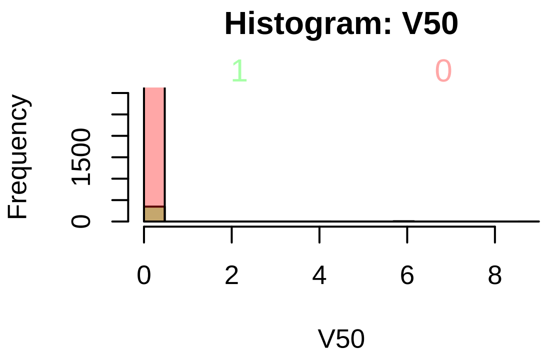

Boxplot: V47
V47
0 4 8
0 1
V86

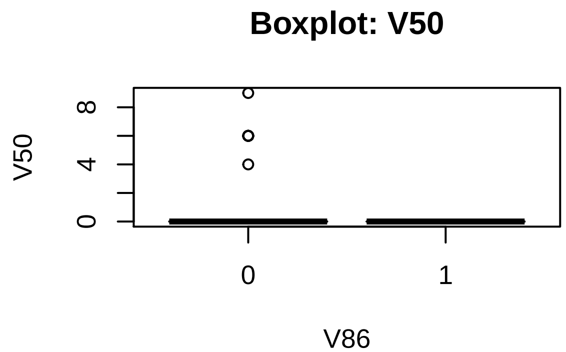

Histogram: V48
1
0
Frequency
0
1500
0 1 2 3 4 5 6 7
V48

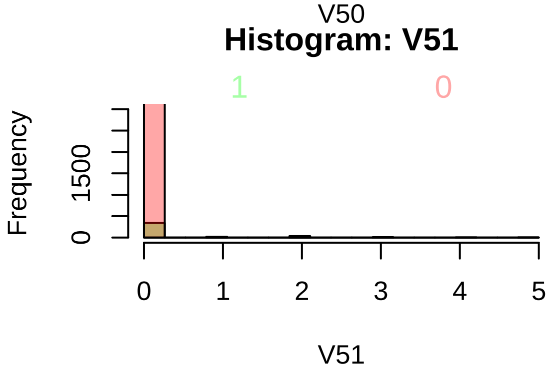

Boxplot: V48
V48
0 3 6
0 1
V86

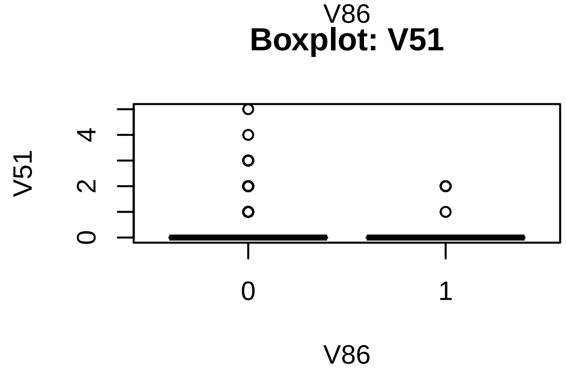

Histogram: V49
1
0
Frequency
0
1500
0 1 2 3 4 5 6 7
V49

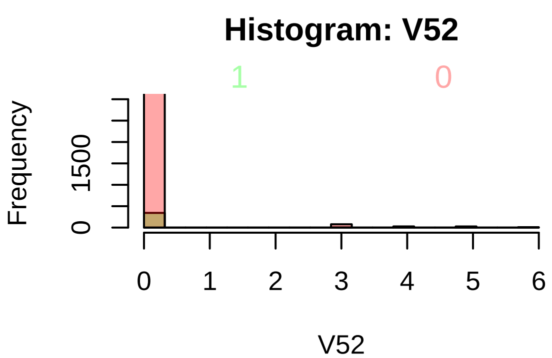

Boxplot: V49
V49
0
3
6
0
1
V86


Histogram: V50
1
0
Frequency
0
1500
0 2 4 6 8
V50

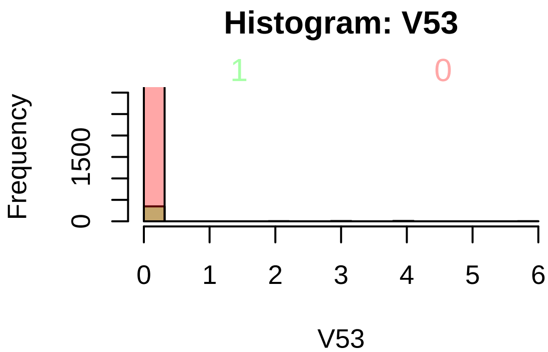

Boxplot: V50
V50
0
4
8
0
1
V86

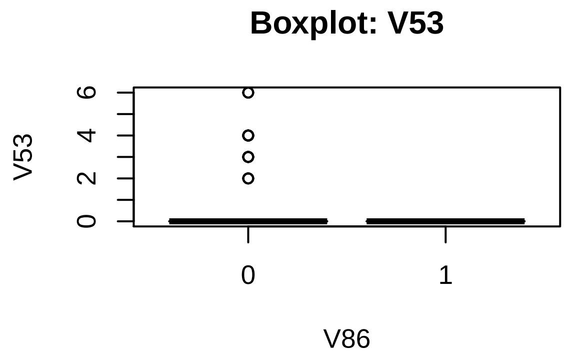

Histogram: V51
1
0
Frequency
0
1500
0 1 2 3 4 5
V51

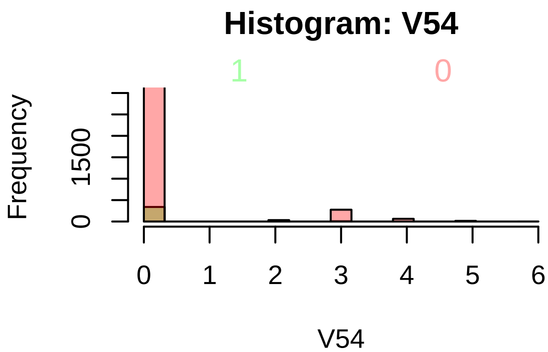

Boxplot: V51
V51
0
2
4
0
1
V86

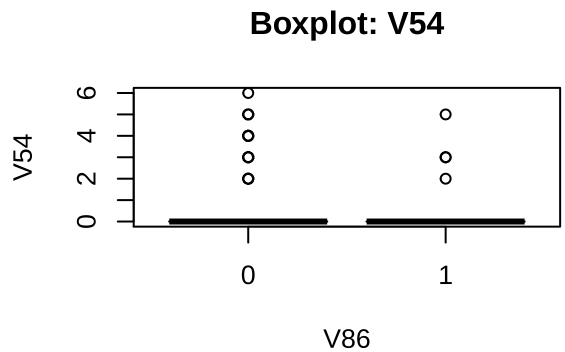

Histogram: V52
1
0
Frequency
0
1500
0 1 2 3 4 5 6
V52

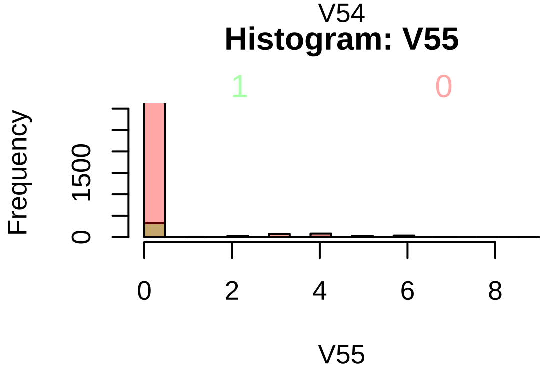

Boxplot: V52
V52
0
2
4
6
0
1
V86

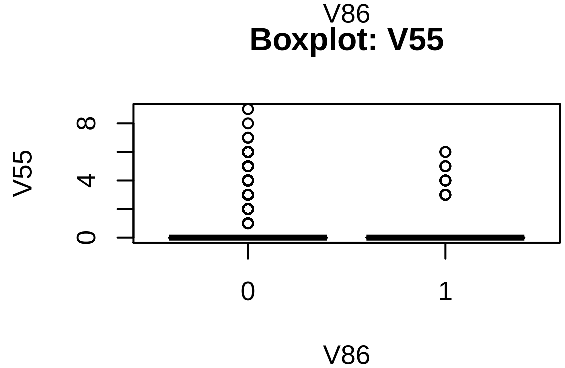

Histogram: V53
1
0
Frequency
0
1500
0 1 2 3 4 5 6
V53

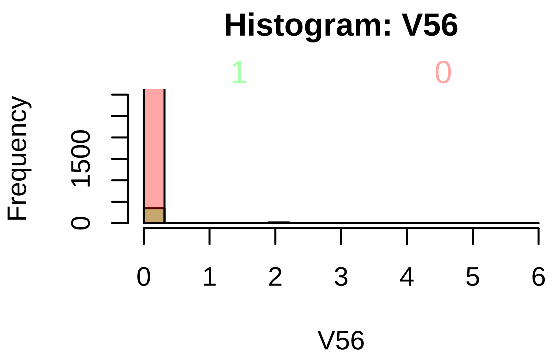

Boxplot: V53
V53
0 2 4 6
0 1
V86

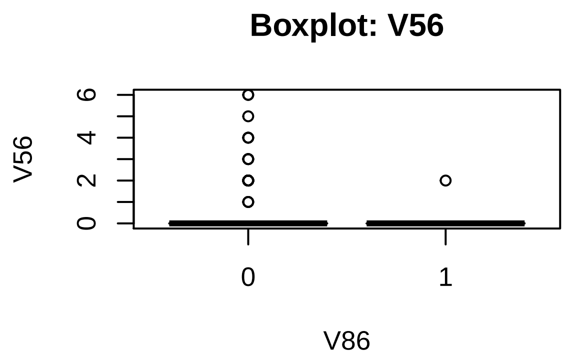

Histogram: V54
1
0
Frequency
0
1500
0 1 2 3 4 5 6
V54

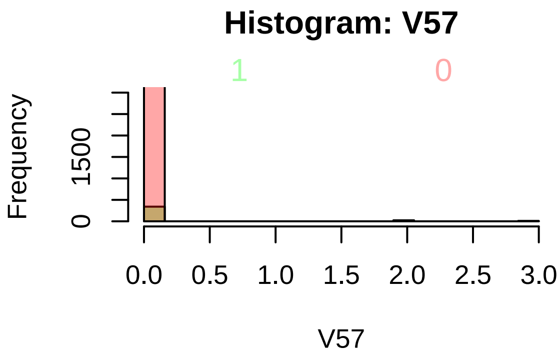

Boxplot: V54
V54
0 2 4 6
0 1
V86

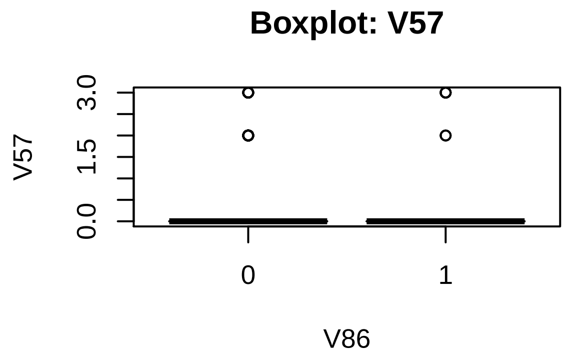

Histogram: V55
1
0
Frequency
0
1500
0 2 4 6 8
V55

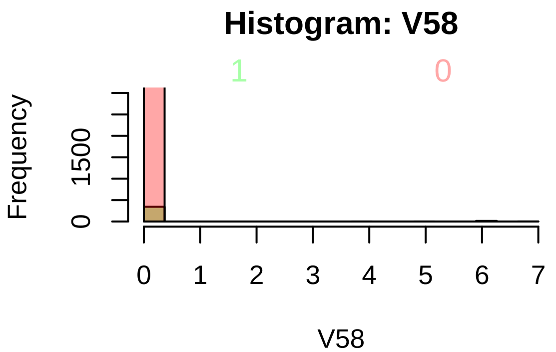

Boxplot: V55
V55
0 4 8
0 1
V86

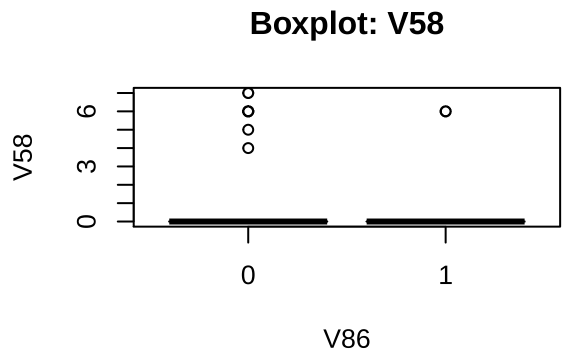

Histogram: V56
1
0
Frequency
0
1500
0 1 2 3 4 5 6
V56

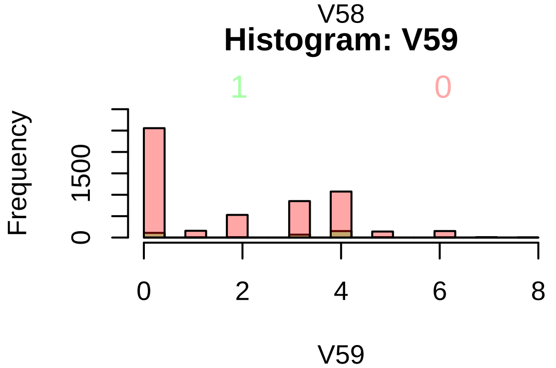

Boxplot: V56
V56
0 2 4 6
0 1
V86

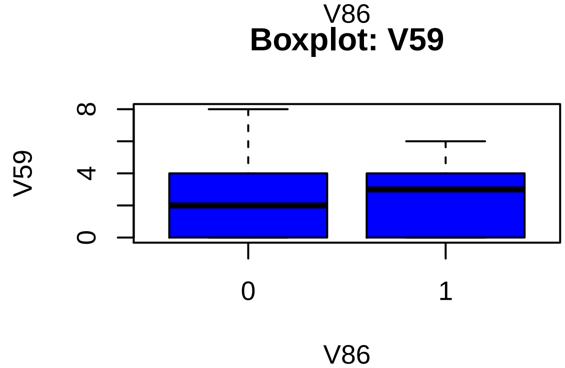

Histogram: V57
1
0
Frequency
0
1500
0.0 0.5 1.0 1.5 2.0 2.5 3.0
V57

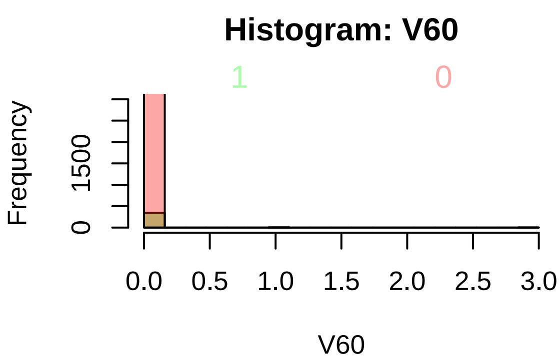

Boxplot: V57
V57
0.0 1.5 3.0
0
1
V86

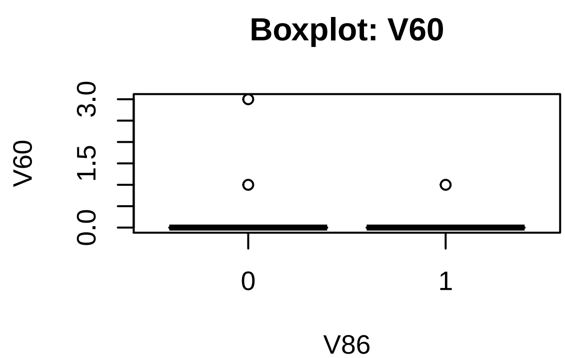

Histogram: V58
1
0
Frequency
0
1500
0 1 2 3 4 5 6 7
V58

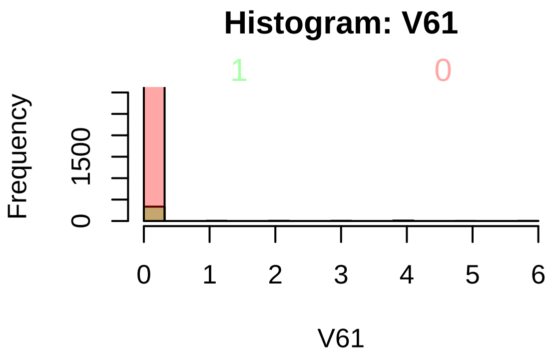

Boxplot: V58
V58
0 3 6
0
1
V86

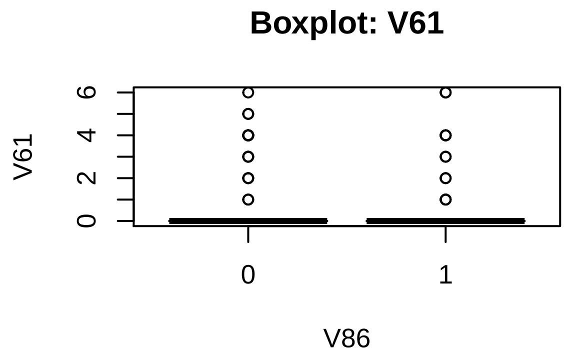

Histogram: V59
1
0
Frequency
0
1500
0 2 4 6 8
V59

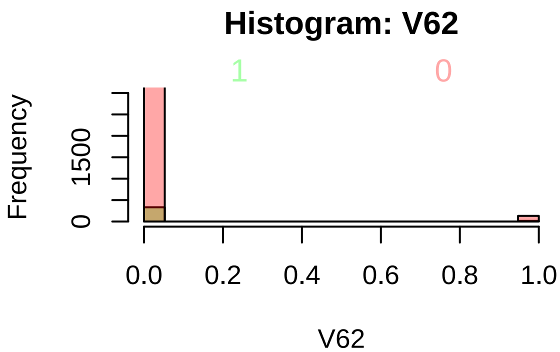

Boxplot: V59
V59
0 4 8
0
1
V86

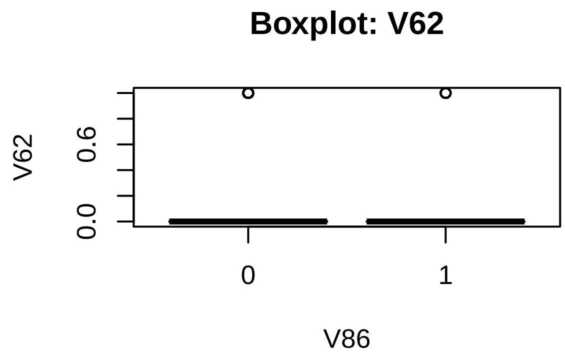

Histogram: V60
1
0
Frequency
0
1500
0.0 0.5 1.0 1.5 2.0 2.5 3.0
V60

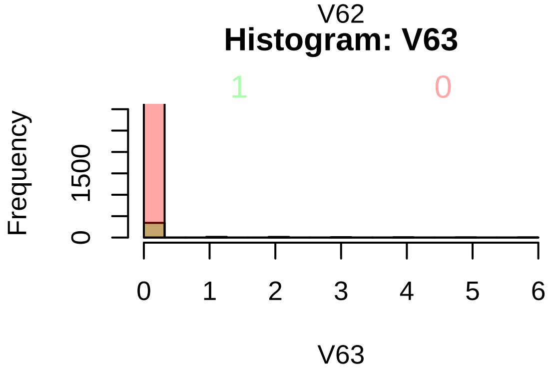

Boxplot: V60
V60
0.0 1.5 3.0
0
1
V86

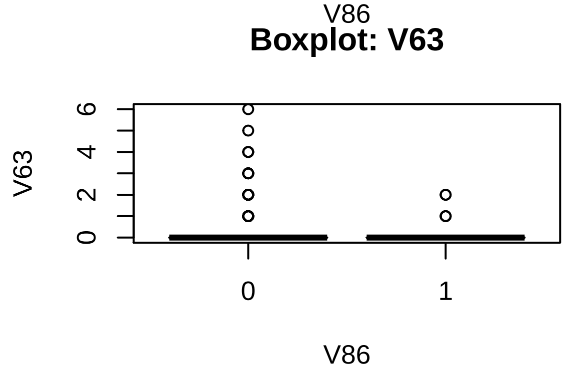

Histogram: V61
1
0
Frequency
0
1500
0
1
2
3
4
5
6
V61

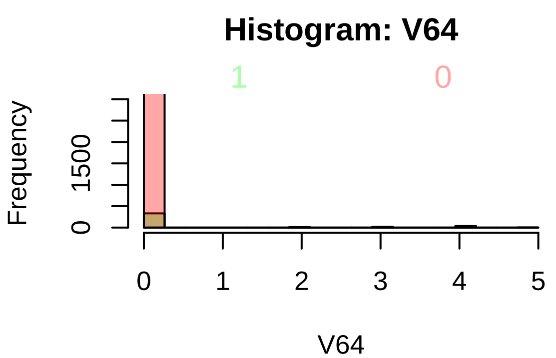

Boxplot: V61
V61
0
2
4
6
0
1
V86

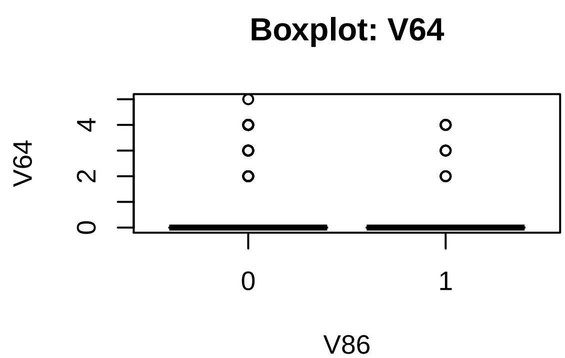

Histogram: V62
1
0
Frequency
0
1500
0.0
0.2
0.4
0.6
0.8
1.0
V62

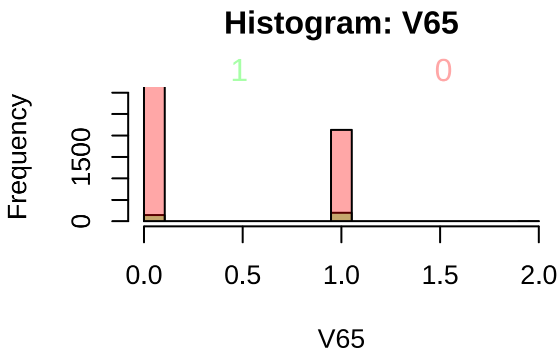
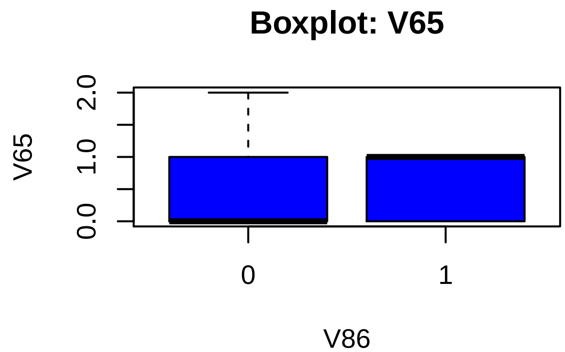
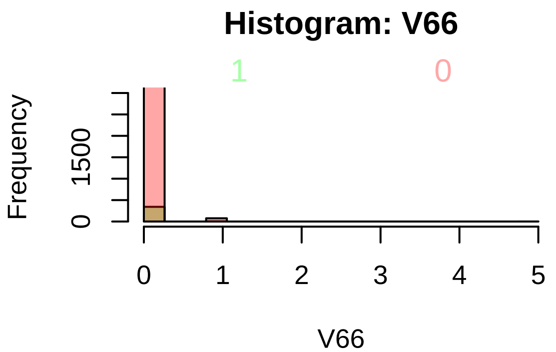
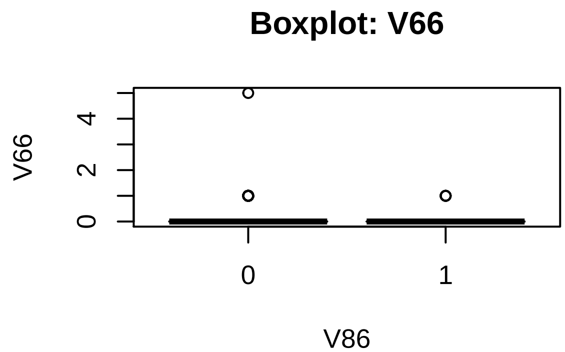
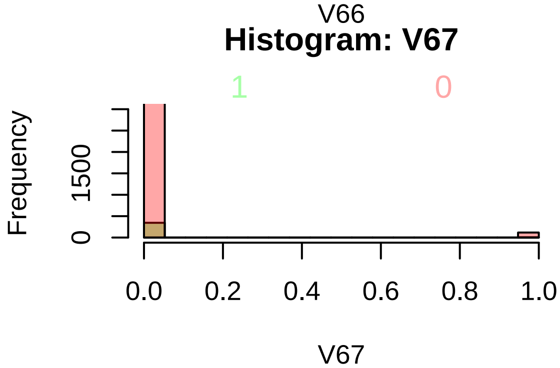
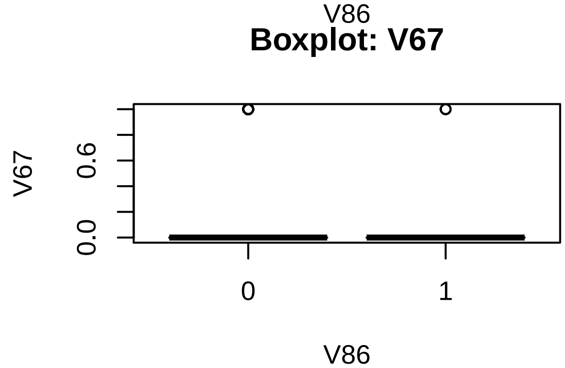
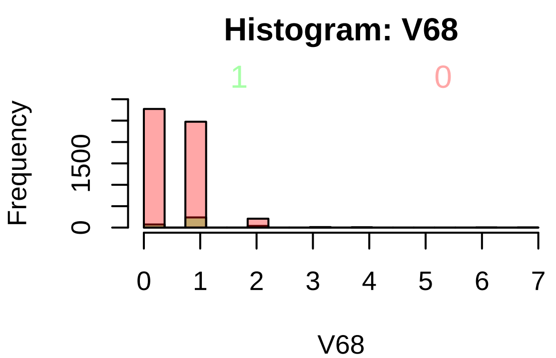
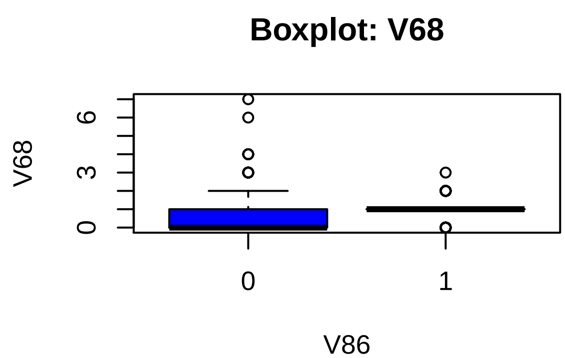

Boxplot: V62
V62
0.0
0.6
0
1
V86

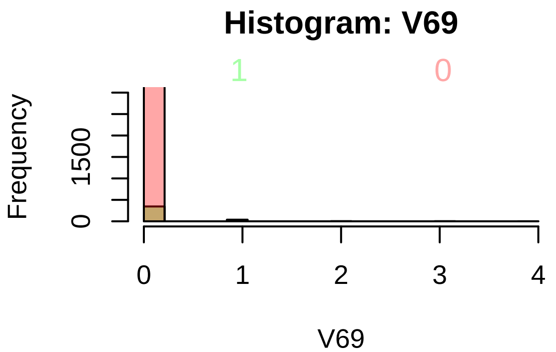

Histogram: V63
1
0
Frequency
0
1500
0
1
2
3
4
5
6
V63

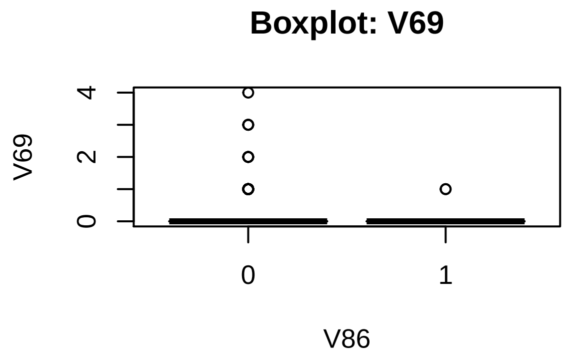

Boxplot: V63
V63
0
2
4
6
0
1
V86

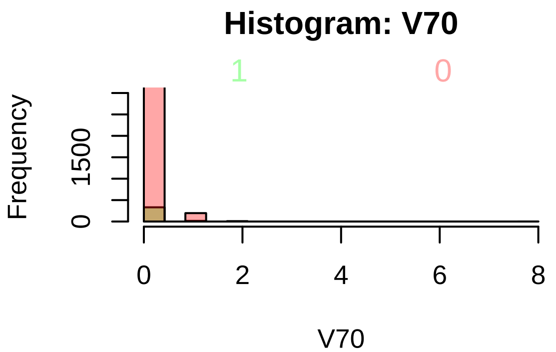

Histogram: V64
1
0
Frequency
0
1500
0
1
2
3
4
5
V64

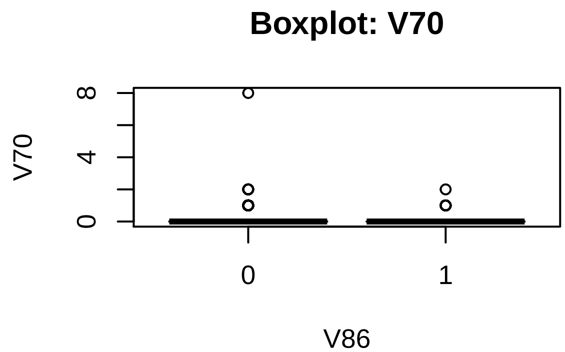

Boxplot: V64
V64
0
2
4
0
1
V86

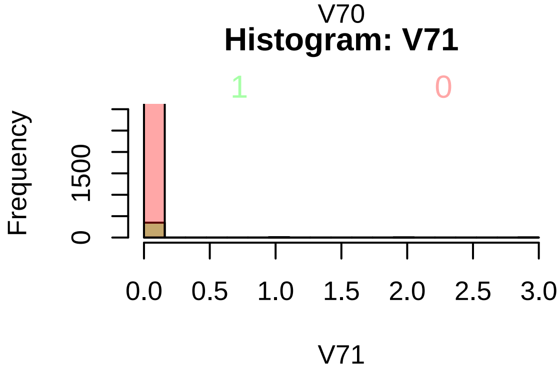

Histogram: V65
1
0
Frequency
0
1500
0.0
0.5
1.0
1.5
2.0
V65

Boxplot: V65
V65
0.0
1.0
2.0
0
1
V86

Histogram: V66
1
0
Frequency
0
1500
0
1
2
3
4
5
V66

Boxplot: V66
V66
0
2
4
0
1
V86

Histogram: V67
1
0
Frequency
0
1500
0.0
0.2
0.4
0.6
0.8
1.0
V67

Boxplot: V67
V67
0.0
0.6
0
1
V86

Histogram: V68
1
0
Frequency
0
1500
0
1
2
3
4
5
6
7
V68

Boxplot: V68
V68
0
3
6
0
1
V86

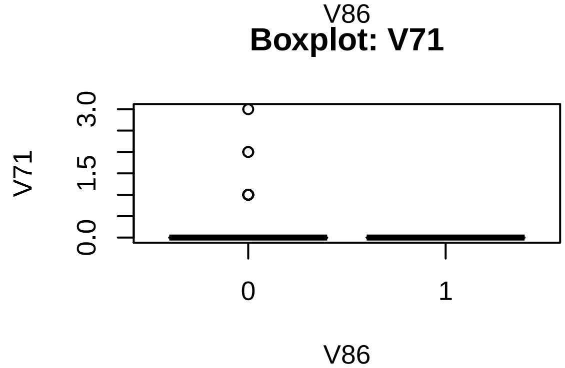

Histogram: V69
1
0
Frequency
0
1500
0
1
2
3
4
V69

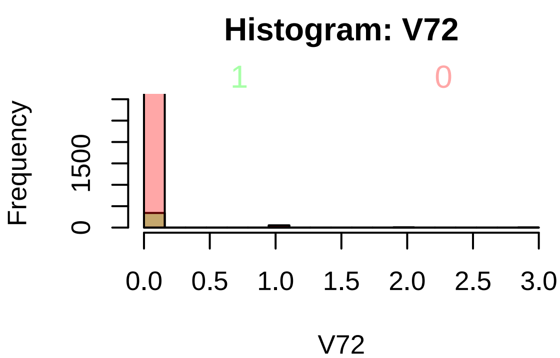

Boxplot: V69
V69
0
2
4
0
1
V86

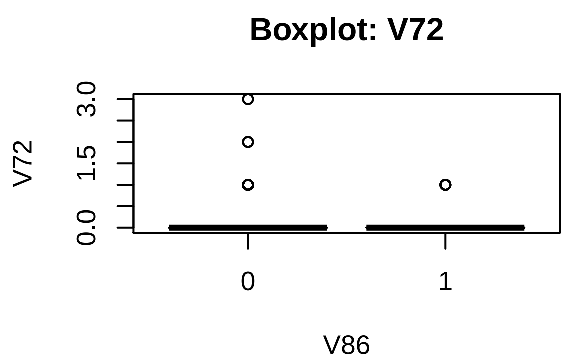

Histogram: V70
1
0
Frequency
0
1500
0
2
4
6
8
V70

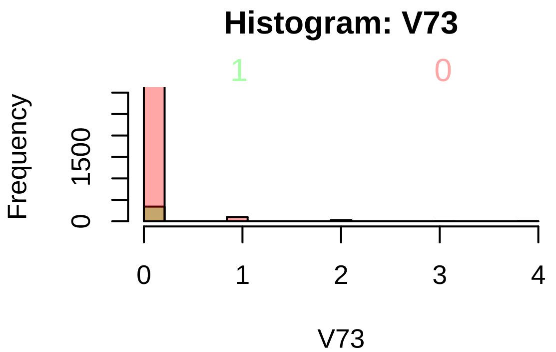

Boxplot: V70
V70
0
4
8
0
1
V86

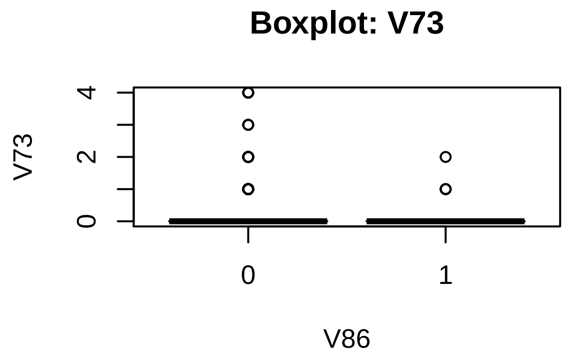

Histogram: V71
1
0
Frequency
0
1500
0.0
0.5
1.0
1.5
2.0
2.5
3.0
V71

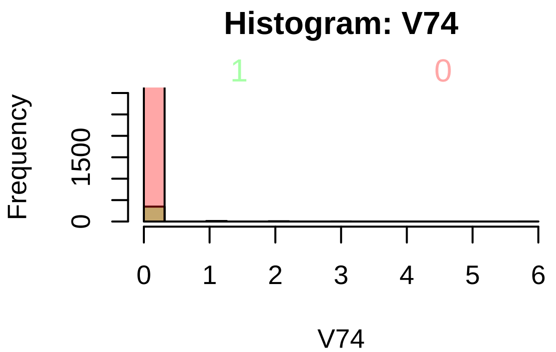

Boxplot: V71
V71
0.0
1.5
3.0
0
1
V86

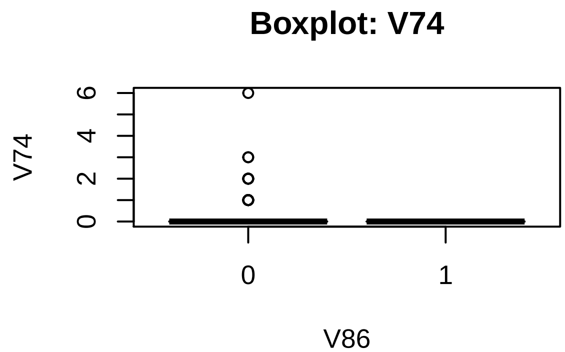

Histogram: V72
1
0
Frequency
0
1500
0.0
0.5
1.0
1.5
2.0
2.5
3.0
V72

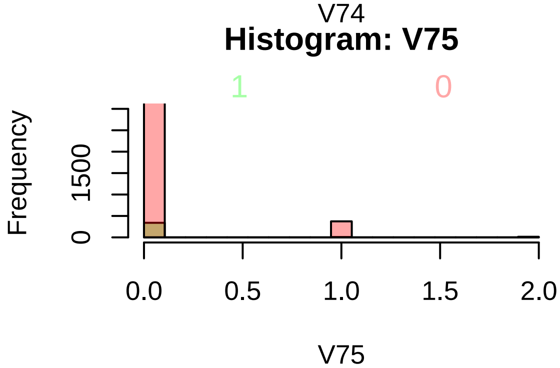

Boxplot: V72
V72
0.0
1.5
3.0
0
1
V86

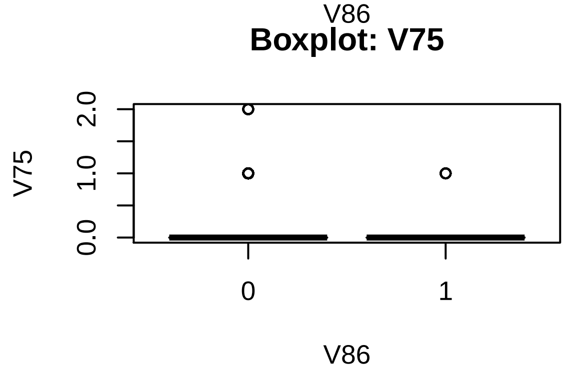

Histogram: V73
1
0
Frequency
0
1500
0
1
2
3
4
V73

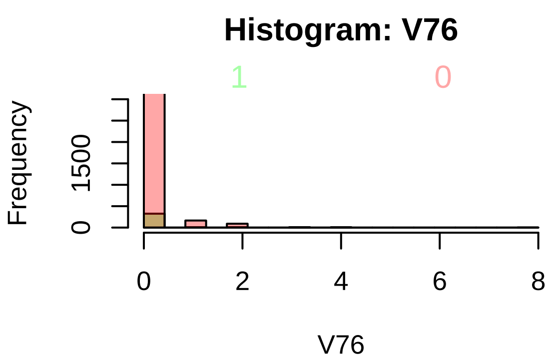

Boxplot: V73
V73
0
2
4
0
1
V86

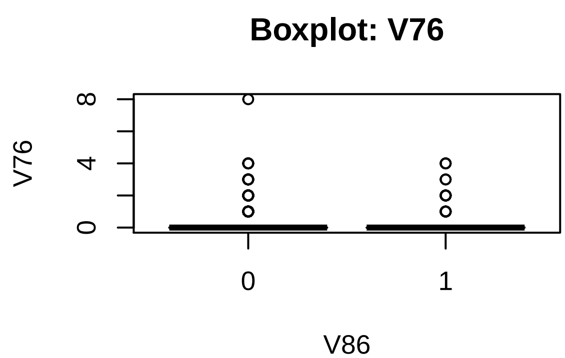

Histogram: V74
1
0
Frequency
0
1500
0
1
2
3
4
5
6
V74

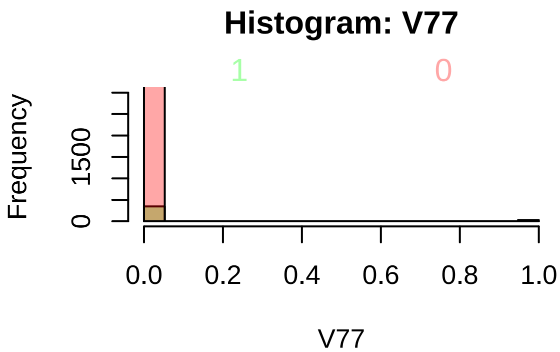

Boxplot: V74
V74
0
2
4
6
0
1
V86

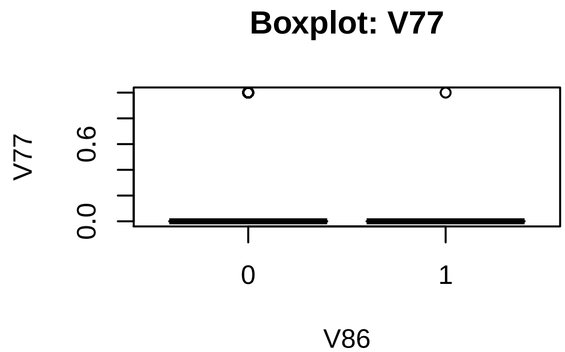

Histogram: V75
1
0
Frequency
0
1500
0.0
0.5
1.0
1.5
2.0
V75

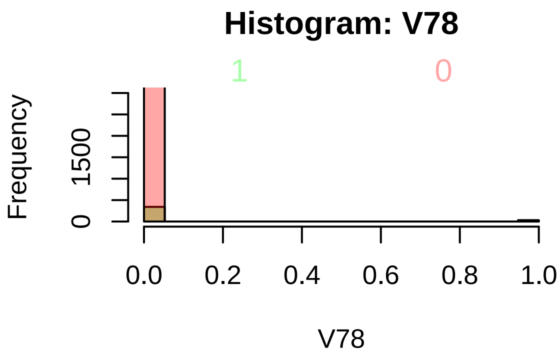

Boxplot: V75
V75
0.0
1.0
2.0
0
1
V86

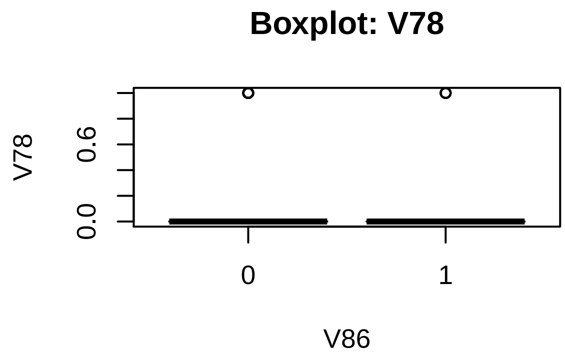

Histogram: V76
1
0
Frequency
0
1500
0
2
4
6
8
V76

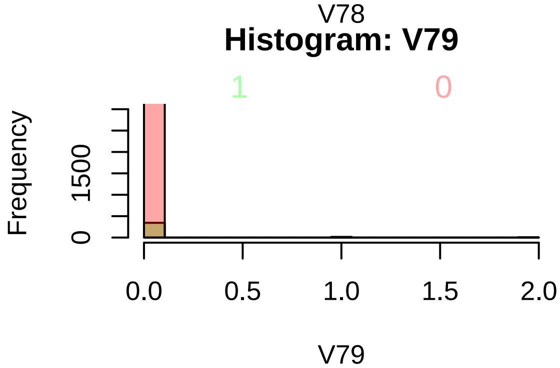

Boxplot: V76
V76
0
4
8
0
1
V86

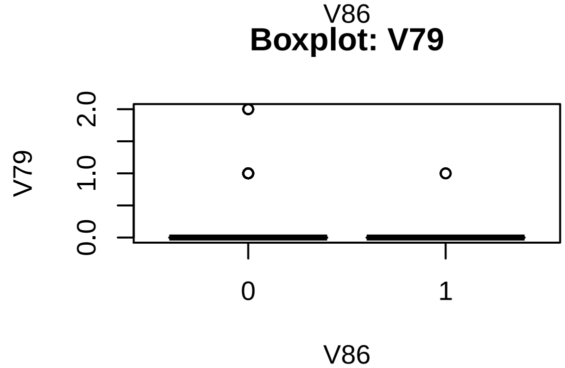

Histogram: V77
1
0
Frequency
0
1500
0.0
0.2
0.4
0.6
0.8
1.0
V77

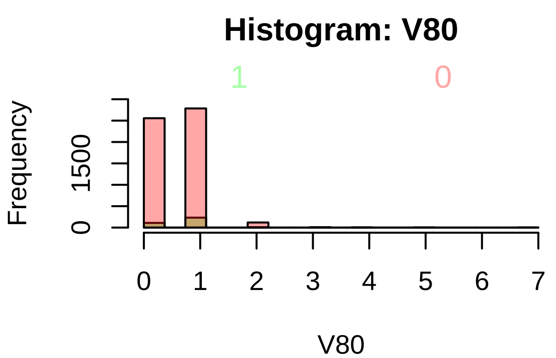

Boxplot: V77
V77
0.0
0.6
0
1
V86

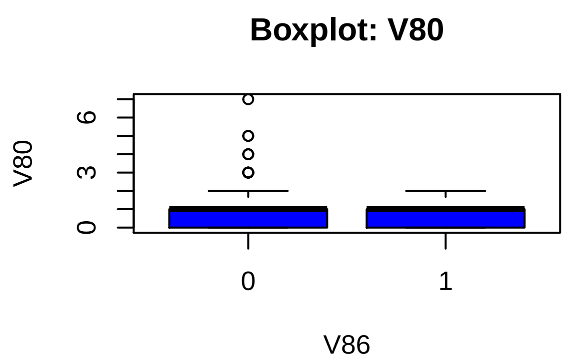

Histogram: V78
1
0
Frequency
0
1500
0.0
0.2
0.4
0.6
0.8
1.0
V78

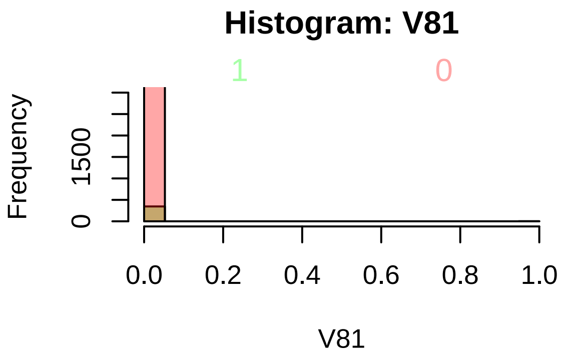

Boxplot: V78
V78
0.0
0.6
0
1
V86

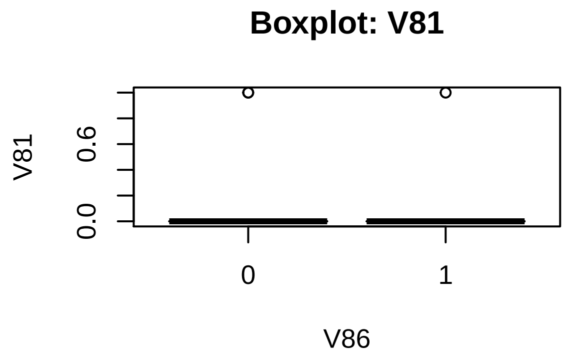

Histogram: V79
1
0
Frequency
0
1500
0.0
0.5
1.0
1.5
2.0
V79

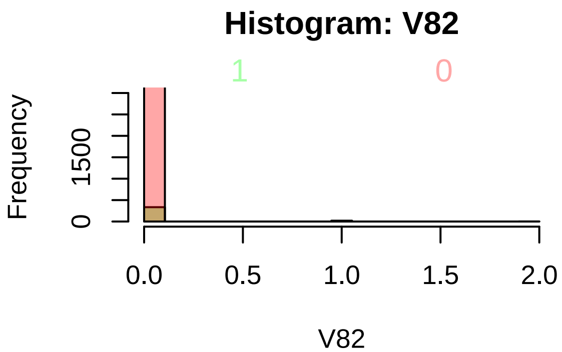

Boxplot: V79
V79
0.0
1.0
2.0
0
1
V86

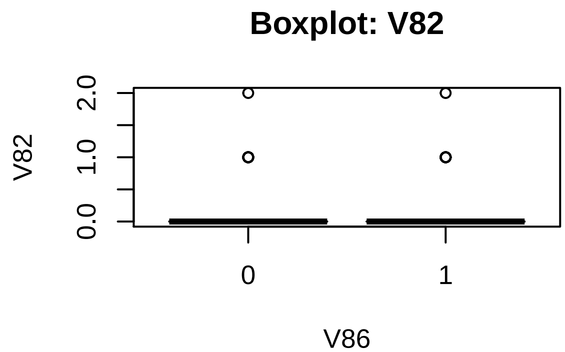

Histogram: V80
1
0
Frequency
0
1500
0
1
2
3
4
5
6
7
V80

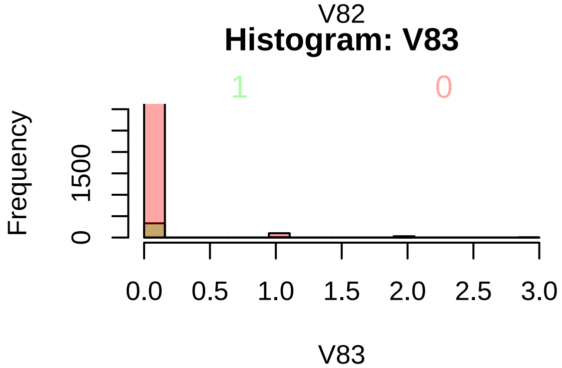

Boxplot: V80
V80
0
3
6
0
1
V86

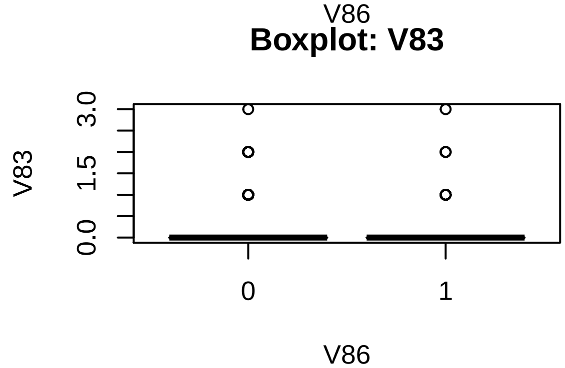

Histogram: V81
1
0
Frequency
0
1500
0.0
0.2
0.4
0.6
0.8
1.0
V81

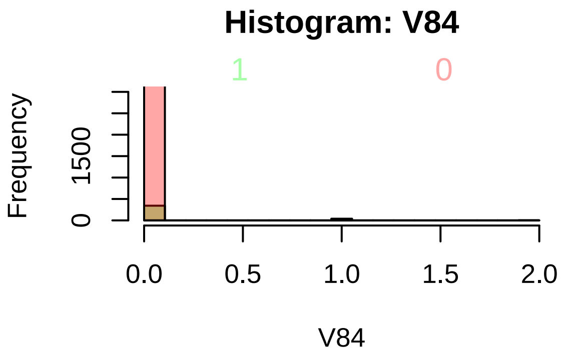

Boxplot: V81
V81
0.0
0.6
0
1
V86

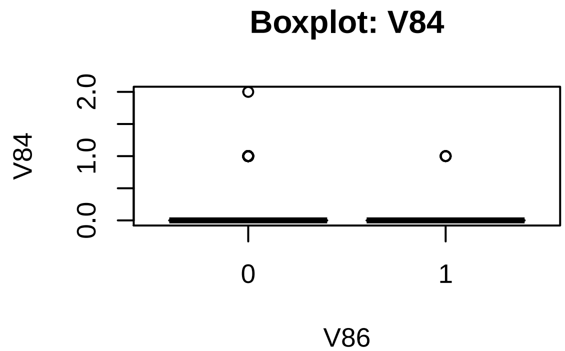

Histogram: V82
1
0
Frequency
0
1500
0.0
0.5
1.0
1.5
2.0
V82

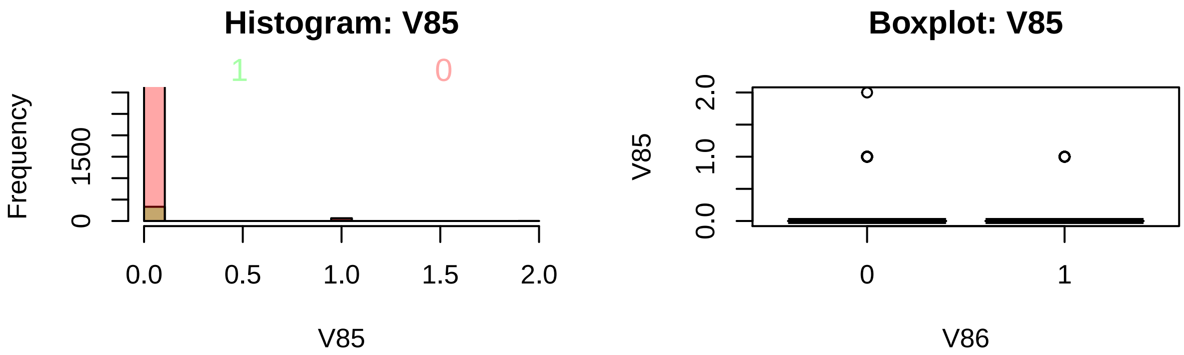

Boxplot: V82
V82
0.0
1.0
2.0
0
1
V86


Histogram: V83
1
0
Frequency
0
1500
0.0
0.5
1.0
1.5
2.0
2.5
3.0
V83


Boxplot: V83
V83
0.0
1.5
3.0
0
1
V86


Histogram: V84
1
0
Frequency
0
1500
0.0
0.5
1.0
1.5
2.0
V84


Boxplot: V84
V84
0.0
1.0
2.0
0
1
V86



2.6.2 Observations

- As we see from the above plots for each predictor's distribution the class of 1 is very small as compared to the class of 0, in caravan policy holders.
- Previous observations on summarising the data have shown that the few of the predictors have more outliers or are heavy tailed as we see in V2, V6, V11 and more. Let's analyse the reason behind it. Take V2 for example which shows that the number of houses owned by the customer. Most customers would generally own a single house, and the others are considered as outliers. Interesting thing would be to think of a customer having many houses, would he be interested in buying a policy for a mobile home, or rather would he own a mobile home. Diving deeper would unfold whether these actually affect the customer's decision in buying or not.
- For some of the predictors such as V18, v43 and similar ones show a rather even distribution of its values corresponding to the classes. The imbalance is similar, but most variables like V41, V44 and similar show a much higher imbalance.

2.7 Reading the test data

As the test data is separated into two files, we at first read them, combine the two and convert the target as factors for modelling purposes

```
caravan_eval=read.table("ticeval2000.txt")
caravan_tgts=read.table("tictgts2000.txt")
caravan_test <- cbind(caravan_eval, V86 = caravan_tgts$V1)
caravan_test$V86=as.factor(caravan_test$V86)
dim(caravan_test)
```

```
## [1] 4000   86
```

2.7.1 CLass imbalance

```
x <- table(caravan_test$V86)
labels <- c("0", "1")
pct <- round(x/sum(x)*100,2)
lbls <- paste(pct,"%",sep="") # ad % to labels
pie(x,labels = lbls, col=rainbow(length(lbls)),main="customers␣of␣caravan␣
   policy")
legend("topright", labels, cex=0.8,fill=rainbow(length(x)))
```

See Figure 1.

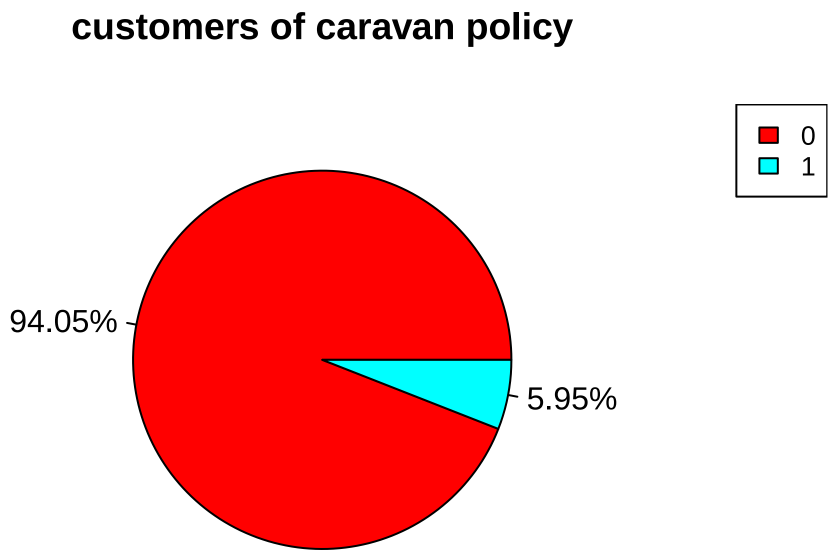


Figure 1: Imbalanced class in Test Data