趙友誠 H24101060

2024-09-26

Table of contents

Brief introduction to the data 1

check missing 2

Descriptive statistic 3

Preprocessing 5

1. 分析所有候選人的知名度、支持度
2. 請提供3號候選人的競選策略(需在何地、對何人進行拉票)
3. 請建立3號候選人支持率的預測模式

```
library(haven)
library(Hmisc)
pollsav <- read_sav("poll.sav")
write.csv(pollsav, file = "poll.csv", row.names = FALSE)
pollcsv <- read.csv("poll.csv")
```

Brief introduction to the data

Dimension of the Data : **1671 samples × 15 columns**

Variables	Explanation
V1、V2、V3	District and Li
V4_1~V4_8	Popularity

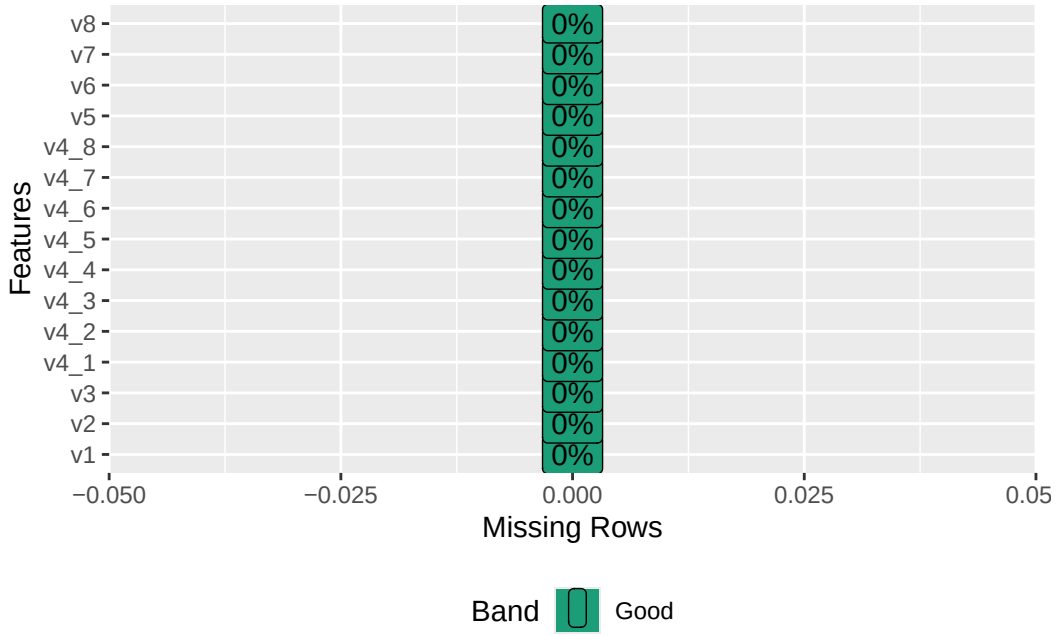
Variables	Explanation
V5	Support level
V6	Age
V7	Education level
V8	Sex

All variables are given in numeric format. For convenience, variables are converted to factor format.

```
str(pollcsv)
pollcsv[] <- lapply(pollcsv, function(item) return(as.factor(item)))
str(pollcsv)
```

check missing

```
library(DataExplorer)
plot_missing(pollcsv)
```



There is no missing value in this data. We then go to next step.

Descriptive statistic

```
latex(describe(pollcsv, "Public Opinion"),file = "", size = "normalsize")
```

Public Opinion
15 Variables 1671 Observations

---

v1

n	missing	distinct
1671	0	2

Value	1	2
Frequency	1107	564
Proportion	0.662	0.338

---

v2

n	missing	distinct
1671	0	36

lowest : 1 2 3 4 5 , highest: 32 33 44 98 99

---

v3

n	missing	distinct
1671	0	23

lowest : 1 2 3 4 5 , highest: 19 20 44 98 99

---

v4_1

n	missing	distinct
1671	0	12

Value	1	2	3	4	5	6	7	8	9	10	91	98
Frequency	328	5	214	43	27	38	47	4	1	11	14	939
Proportion	0.196	0.003	0.128	0.026	0.016	0.023	0.028	0.002	0.001	0.007	0.008	0.562

---

v4_2

n	missing	distinct
1671	0	10

Value	2	3	4	5	6	7	8	9	10	99
Frequency	6	189	59	32	75	99	2	4	15	1190
Proportion	0.004	0.113	0.035	0.019	0.045	0.059	0.001	0.002	0.009	0.712

---

v4_3

n	missing	distinct
1671	0	9

Value	3	4	5	6	7	8	9	10	99
Frequency	6	60	36	61	91	1	2	19	1395
Proportion	0.004	0.036	0.022	0.037	0.054	0.001	0.001	0.011	0.835

---

v4_4

n	missing	distinct
1671	0	8

Value	4	5	6	7	8	9	10	99
Frequency	4	28	41	52	3	4	20	1519
Proportion	0.002	0.017	0.025	0.031	0.002	0.002	0.012	0.909

---

v4_5

n	missing	distinct
1671	0	7

Value	5	6	7	8	9	10	99
Frequency	3	14	38	4	3	15	1594
Proportion	0.002	0.008	0.023	0.002	0.002	0.009	0.954

---

v4_6

n	missing	distinct
1671	0	6

Value	6	7	8	9	10	99
Frequency	3	12	6	7	20	1623
Proportion	0.002	0.007	0.004	0.004	0.012	0.971

---

v4_7

n	missing	distinct
1671	0	5

Value	7	8	9	10	99
Frequency	3	2	3	12	1651
Proportion	0.002	0.001	0.002	0.007	0.988

---

v4_8

n	missing	distinct
1671	0	3

Value	8	10	99
Frequency	1	4	1666
Proportion	0.001	0.002	0.997

---

v5

n	missing	distinct
1671	0	13

Value	1	2	3	4	5	6	7	8	9	10	91	98	99
Frequency	158	9	205	79	33	98	195	6	8	53	10	269	548
Proportion	0.095	0.005	0.123	0.047	0.020	0.059	0.117	0.004	0.005	0.032	0.006	0.161	0.328

---

v6

n	missing	distinct
1671	0	6

Value	1	2	3	4	5	6
Frequency	52	94	201	336	946	42
Proportion	0.031	0.056	0.120	0.201	0.566	0.025

---

v7

n	missing	distinct
1671	0	6

Value	1	2	3	4	5	95
Frequency	292	165	431	198	520	65
Proportion	0.175	0.099	0.258	0.118	0.311	0.039

---

v8

n	missing	distinct
1671	0	2

Value	1	2
Frequency	682	989
Proportion	0.408	0.592

Preprocessing

```
pollcsv$v2[pollcsv$v2==99] <- "-"
```

```
Warning in `[<-.factor`(`*tmp*`, pollcsv$v2 == 99, value = structure(c(NA, :
invalid factor level, NA generated
```

```
pollcsv$v3[pollcsv$v3==99] <- "-"
```

```
Warning in `[<-.factor`(`*tmp*`, pollcsv$v3 == 99, value = structure(c(5L, :
invalid factor level, NA generated
```

```
table(pollcsv$v2)
```

```

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
 14  12   4   8  35  10  39  70  34  38  25  24  12  19  32  18  16  23  40  21
 21  22  23  24  25  26  27  28  29  30  31  32  33  44  98  99
 41  33  29  12  32  42  57  39  12  28  37  31  42 161  17   0
```

```
table(pollcsv$v3)
```

```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 44 98 99
19 25 14 51 42 20 14 24  8 15  9 29 34 18  6 19 43 22 35 15 84 18  0
```

```
DistnLi <- apply(pollcsv[,1:3],MARGIN = 1, FUN = function(row){
  if(row[2]=="-") return(paste0(row[1],row[2],row[3]))
  if(row[3]=="-") return(paste0(row[1],row[3],row[2]))
})
```

```
Error in if (row[2] == "-") return(paste0(row[1], row[2], row[3])): missing value where TRUE,
```

```
pollcsv[,1:3] <- NULL
pollcsv <- data.frame(
  DistnLi = DistnLi,
  pollcsv
)
```

```
Error in eval(expr, envir, enclos): object 'DistnLi' not found
```

```
pollcsv <- data.frame(t(apply(pollcsv,MARGIN = 1, FUN = function(row){
  row[row==99] <- 0
  return(row)
})))
```