選舉民意調查資料分析報告

趙友誠 H24101060

2024-09-28

Table of contents

Brief introduction to the data

This is a complete data with no actual missing value while some might be labeled as missing. Dimension of the Data : ***1671 samples* × *15 columns***

Variables	Explanation
V1、V2、V3	District and Li
V4_1~V4_8	Popularity
V5	Support level
V6	Age
V7	Education level
V8	Sex

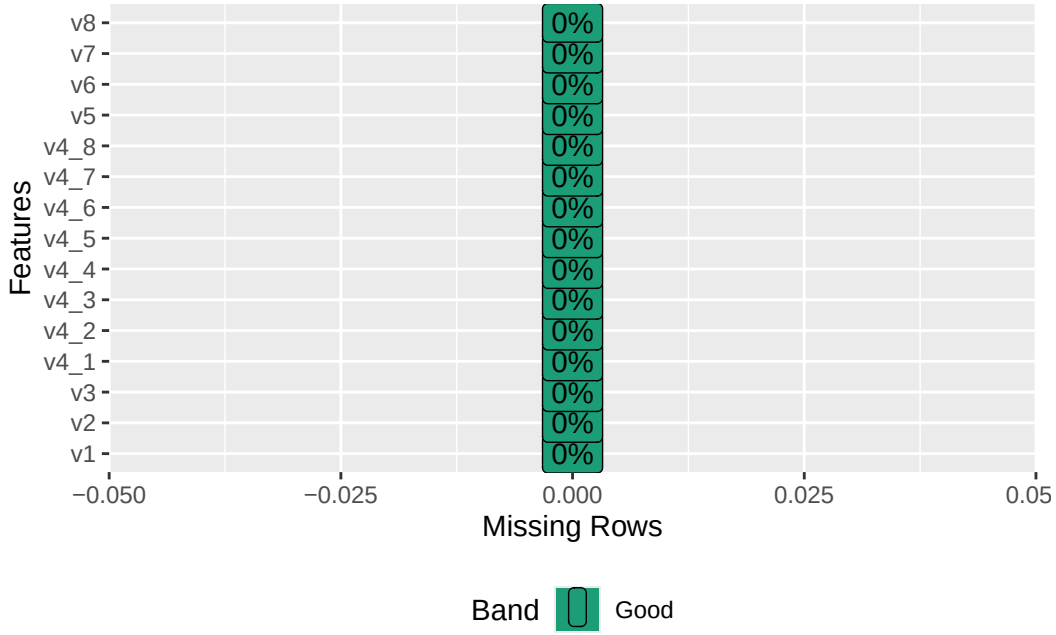
```r
library(haven)
library(Hmisc)
pollsav <- read_sav("poll.sav")
write.csv(pollsav, file = "poll.csv", row.names = FALSE)
pollcsv <- read.csv("poll.csv")
```

Preprocessing

First, we replace 91,95,98,99 by 0 and then converted variables from numeric format to factor one. As the graph shown below, there is no missing value in this data.

```r
pollcsv <- data.frame(
  t(apply(pollcsv,MARGIN = 1, FUN = function(row){
    row[row==99 | row==98 | row==91 | row==95] <- 0
    return(row)
  }))
)
pollcsv[] <- lapply(pollcsv, function(item) return(as.factor(item)))
n <- dim(pollcsv)[1]
```

```r
DataExplorer::plot_missing(pollcsv)
```



Descriptive statistic

This chunk is for the convenience of analysis, so it will not be shown here.

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

The definition of the popularity of a candidate in this analysis is the number of appearance of a candidate that a participant answered in the 4th question. And the definition of the support level of a candidate is the number of appearance of a candidate in 5th question divided by the number of the participants who specifically choose a name in 5th question. That is, participants who did not actually answered the question are removed from the calculation.

```
# 計算 1~10 號候選人在複選題出現的次數
count4 <- unlist(lapply(factor(1:10), function(x){
  return(length(unlist(apply(pollcsv[,4:11], MARGIN = 1, function(row) if(x %in% row) return
}))

# 計算 1~10 號候選人在第 5 題出現的次數
count5 <- unlist(lapply(factor(0:10),function(x){
  return(sum(pollcsv$v5==x))
} ))

n_prime <- n-count5[1]

# 知名度 = 出現次數/sample size
p <- data.frame(factor(1:10), popularity=round(count4/n,3), count4)

# 支持度 = 出現次數/(sample size-沒回答的)
s <- data.frame(factor(1:10), `support level`=round(count5[2:11]/n_prime,3), count5[2:11])

# 將候選人依據知名度與支持度排序
p <- p[order(p$popularity, decreasing = TRUE ),]
s <- s[order(s$support.level, decreasing = TRUE ),]

latex(data.table::data.table(cbind(p,s)),title="",file = "", booktabs = TRUE, cgroup = c('Po
```

Then we calculate the popularity and the support level grouped by district.

```
# 依據區將資料分成 Dist1 與 Dist2
Dist1 <-subset(pollcsv,pollcsv$v1==1)
Dist2 <-subset(pollcsv,pollcsv$v1==2)
```

	Popularity			Support level		
	candidate	rate	count	candidate	rate	count
1	3	0.245	409	3	0.243	205
2	7	0.205	342	7	0.231	195
3	1	0.196	328	1	0.187	158
4	6	0.139	232	6	0.116	98
5	4	0.099	166	4	0.094	79
6	5	0.075	126	10	0.063	53
7	10	0.069	116	5	0.039	33
8	8	0.014	23	2	0.011	9
9	9	0.014	24	9	0.009	8
10	2	0.007	11	8	0.007	6

```
n1 <- dim(Dist1)[1]
n2 <- dim(Dist2)[1]

# 第一區的計算
# 候選人在複選題出現的次數
count4_1 <- unlist(lapply(factor(1:10), function(x){
  return(length(unlist(apply(Dist1[,4:11], MARGIN = 1, function(row) if(x %in% row) return (
}))
# 候選人在第 5 題出現的次數
count5_1 <- unlist(lapply(factor(0:10),function(x){
  return(sum(Dist1$v5==x))
} ))

n1_prime <- n1-count5_1[1]

# 知名度 = 出現次數/sample size
p_1 <- data.frame(factor(1:10), popularity=round(count4_1/n1,3), count4_1)

# 支持度 = 出現次數/(sample size-沒回答的)
s_1 <- data.frame(factor(1:10), `support level`=round(count5_1[2:11]/n1_prime,3), count5_1[2

# 將候選人依據知名度與支持度排序
p_1 <- p_1[order(p_1$popularity, decreasing = TRUE ),]
s_1 <- s_1[order(s_1$support.level, decreasing = TRUE ),]

latex(data.table::data.table(cbind(p_1,s_1)),title="",file = "", booktabs = TRUE, cgroup = c
```

	Popularity			Support level		
	candidate	rate	count	candidate	rate	count
1	3	0.288	319	3	0.296	163
2	7	0.225	249	7	0.260	143
3	6	0.176	195	6	0.151	83
4	1	0.156	173	1	0.102	56
5	4	0.096	106	4	0.058	32
6	5	0.087	96	10	0.055	30
7	10	0.062	69	5	0.053	29
8	8	0.014	15	2	0.013	7
9	9	0.013	14	9	0.007	4
10	2	0.006	7	8	0.005	3

```
# 第二區的計算
# 候選人在複選題出現的次數
count4_2 <- unlist(lapply(factor(1:10), function(x){
  return(length(unlist(apply(Dist2[,4:11], MARGIN = 1, function(row) if(x %in% row) return (
}))

# 候選人在第 5 題出現的次數
count5_2 <- unlist(lapply(factor(0:10),function(x){
  return(sum(Dist2$v5==x))
} ))

n2_prime <- n2-count5_2[1]

# 知名度 = 出現次數/sample size
p_2 <- data.frame(factor(1:10), popularity=round(count4_2/n2,3), count4_2)

# 支持度 = 出現次數/(sample size-沒回答的)
s_2 <- data.frame(factor(1:10), `support level`=round(count5_2[2:11]/n2_prime,3), count5_2[2

# 將候選人依據知名度與支持度排序
p_2 <- p_2[order(p_2$popularity, decreasing = TRUE ),]
s_2 <- s_2[order(s_2$support.level, decreasing = TRUE ),]

latex(data.table::data.table(cbind(p_2,s_2)),title="",file = "", booktabs = TRUE, cgroup = c
```

	Popularity			Support level		
	candidate	rate	count	candidate	rate	count
1	1	0.275	155	1	0.347	102
2	7	0.165	93	7	0.177	52
3	3	0.160	90	4	0.160	47
4	4	0.106	60	3	0.143	42
5	10	0.083	47	10	0.078	23
6	6	0.066	37	6	0.051	15
7	5	0.053	30	5	0.014	4
8	9	0.018	10	9	0.014	4
9	8	0.014	8	8	0.010	3
10	2	0.007	4	2	0.007	2

候選人 3 的知名度與支持度 (分區、年齡層、性別): 後面三個還沒做

```
#1 區 +2 區的知名度與支持度
p[1,2]
```

```
[1] 0.245
```

```
s[1,2]
```

```
[1] 0.243
```

```
#1 區的知名度與支持度
p_1[1,2]
```

```
[1] 0.288
```

```
s_1[1,2]
```

```
[1] 0.296
```

```
#2 區的知名度與支持度
p_2[1,2]
```

```
[1] 0.275
```

```
s_2[1,2]
```

```
[1] 0.347
```