Stat 652 Final

Kotomi Oda

Answer two questions:

1. What are the important variables identified by the Boruta algorithm from the Ozone data?

- The important variables identified by the Boruta algorithm are v1, v5, v7, v8, v10, v11, v12, and v13.

2. Try to run the code in parallel.

```
library(pacman)
p_load(tidyverse, janitor, naniar, Boruta, mlbench)
```

```
data(Ozone)

head(Ozone)
```

```
##   V1 V2 V3 V4   V5 V6 V7 V8    V9  V10 V11   V12 V13
## 1  1  1  4  3 5480  8 20 NA    NA 5000 -15 30.56 200
## 2  1  2  5  3 5660  6 NA 38    NA   NA -14    NA 300
## 3  1  3  6  3 5710  4 28 40    NA 2693 -25 47.66 250
## 4  1  4  7  5 5700  3 37 45    NA  590 -24 55.04 100
## 5  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
## 6  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
```

Note that the target variable is V4 = Daily maximum one-hour-average ozone reading

```
Ozone <- Ozone %>% mutate(
  V1 = as.integer(V1),
  V2 = as.integer(V2),
  V3 = as.integer(V3)
)

head(Ozone)
```

```
##   V1 V2 V3 V4   V5 V6 V7 V8    V9  V10 V11   V12 V13
## 1  1  1  4  3 5480  8 20 NA    NA 5000 -15 30.56 200
## 2  1  2  5  3 5660  6 NA 38    NA   NA -14    NA 300
## 3  1  3  6  3 5710  4 28 40    NA 2693 -25 47.66 250
## 4  1  4  7  5 5700  3 37 45    NA  590 -24 55.04 100
## 5  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
## 6  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
```

```
Ozone2 <- Ozone %>% clean_names()

head(Ozone2)
```

```
##   v1 v2 v3 v4   v5 v6 v7 v8    v9  v10 v11   v12 v13
## 1  1  1  4  3 5480  8 20 NA    NA 5000 -15 30.56 200
## 2  1  2  5  3 5660  6 NA 38    NA   NA -14    NA 300
## 3  1  3  6  3 5710  4 28 40    NA 2693 -25 47.66 250
## 4  1  4  7  5 5700  3 37 45    NA  590 -24 55.04 100
## 5  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
## 6  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
```

It is always a good idea to check for duplicate records/examples/rows in your dataset.

```
get_dupes(Ozone2)
```

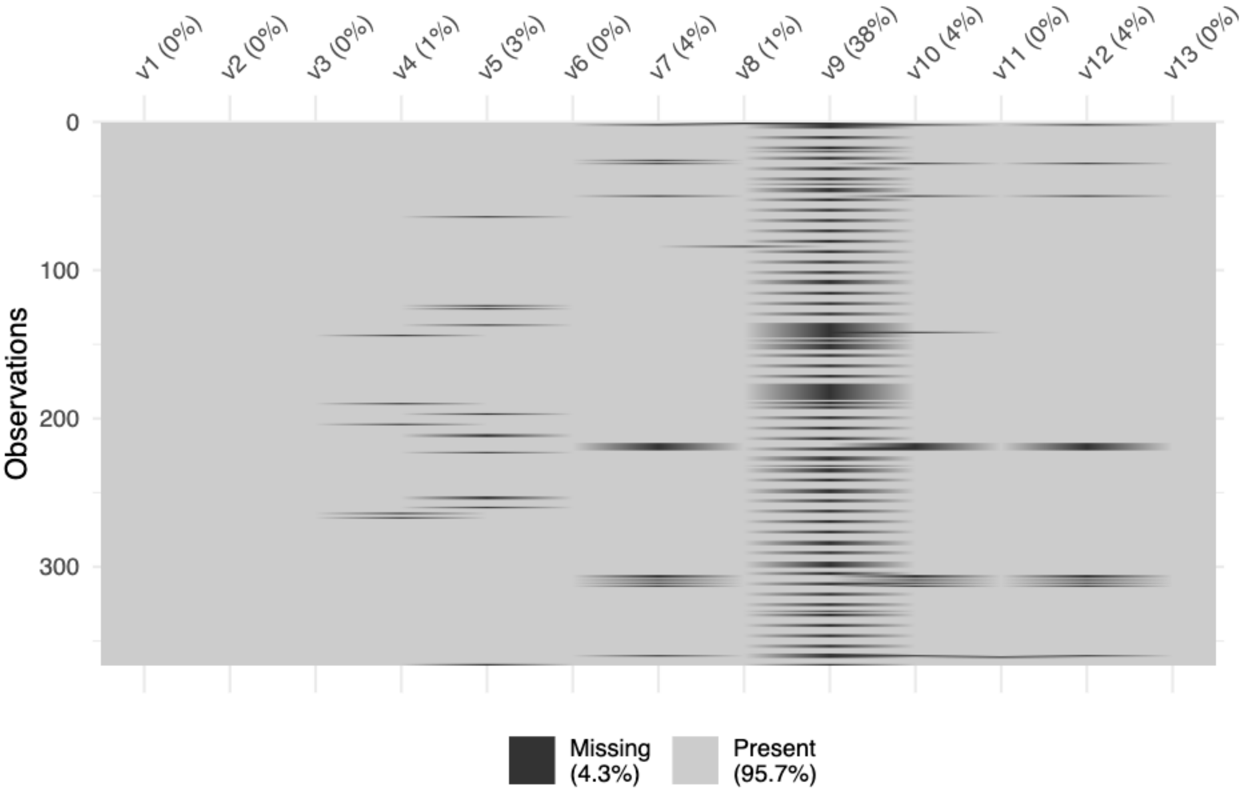
```
## No variable names specified - using all columns.
```

```
## No duplicate combinations found of: v1, v2, v3, v4, v5, v6, v7, v8, v9, ... and 4 other variables
```

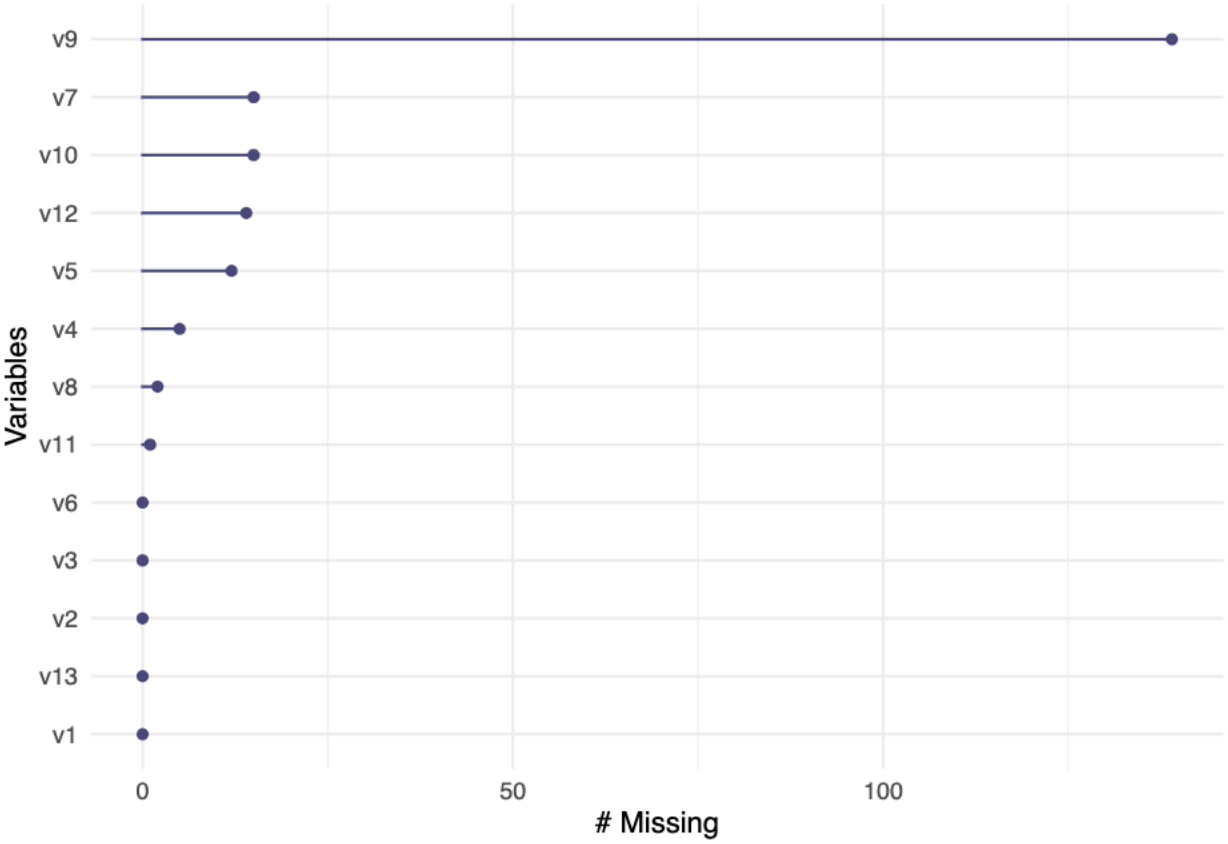
```
##  [1] v1         v2         v3         v4         v5         v6
##  [7] v7         v8         v9         v10        v11        v12
## [13] v13        dupe_count
## <0 rows> (or 0-length row.names)
```

Start by investigating the missing values and completeness of the features in the data. Note that the *age* variable contains some missing values.

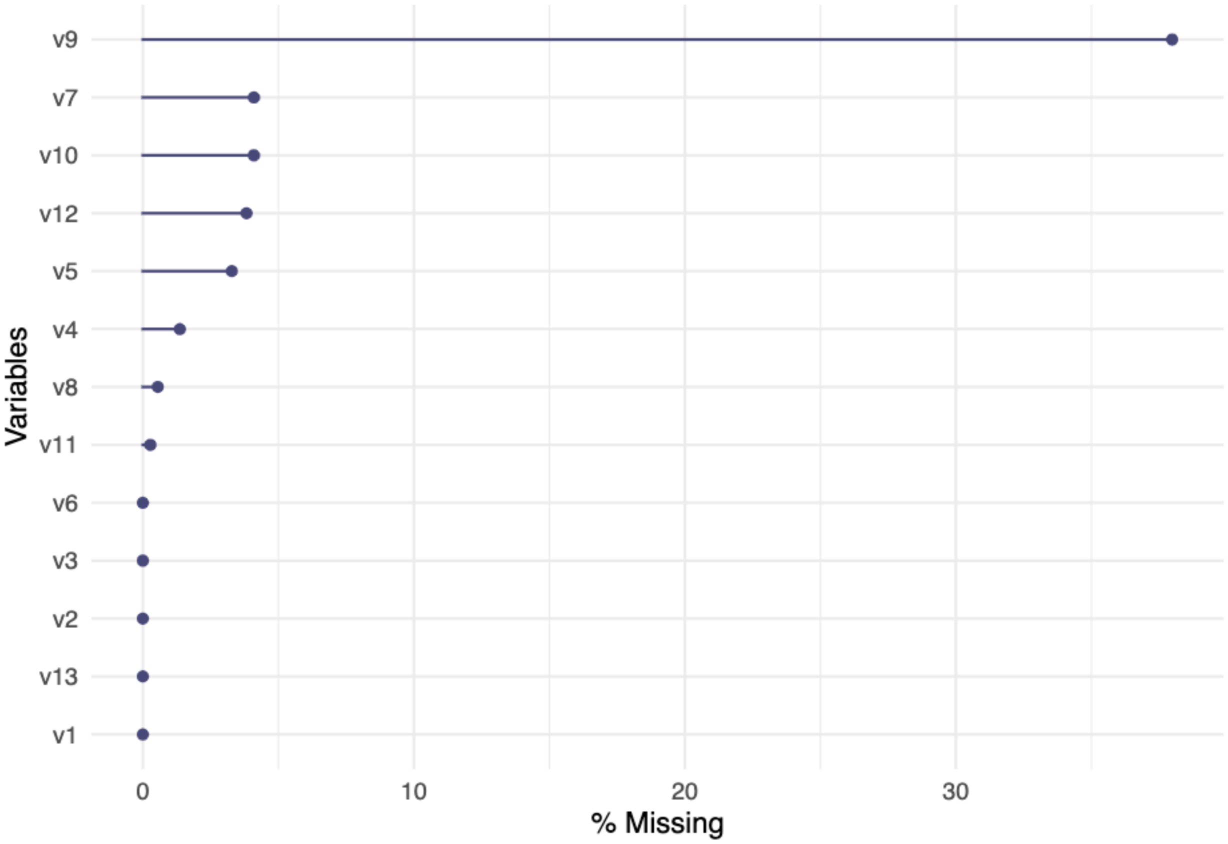
```
vis_miss(Ozone2)
```



```
gg_miss_var(Ozone2)
```



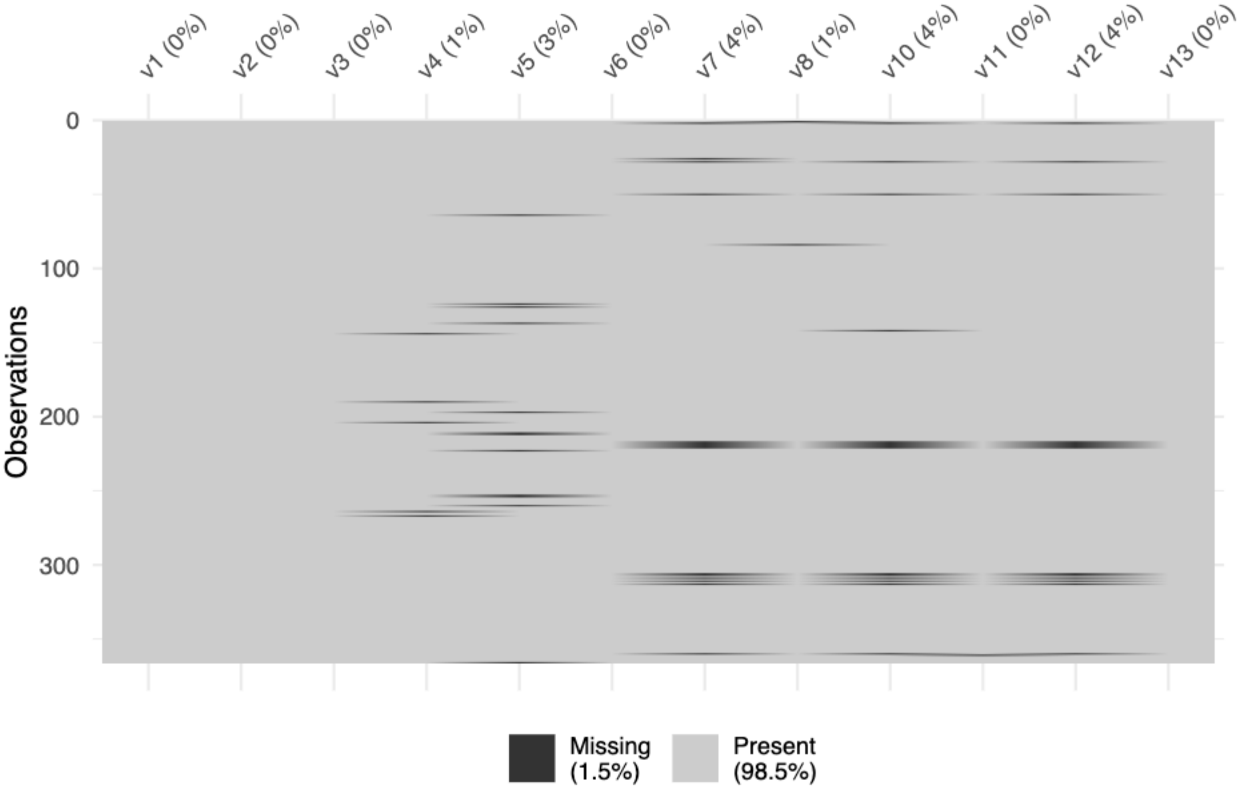
```
gg_miss_var(Ozone2, show_pct = TRUE)
```



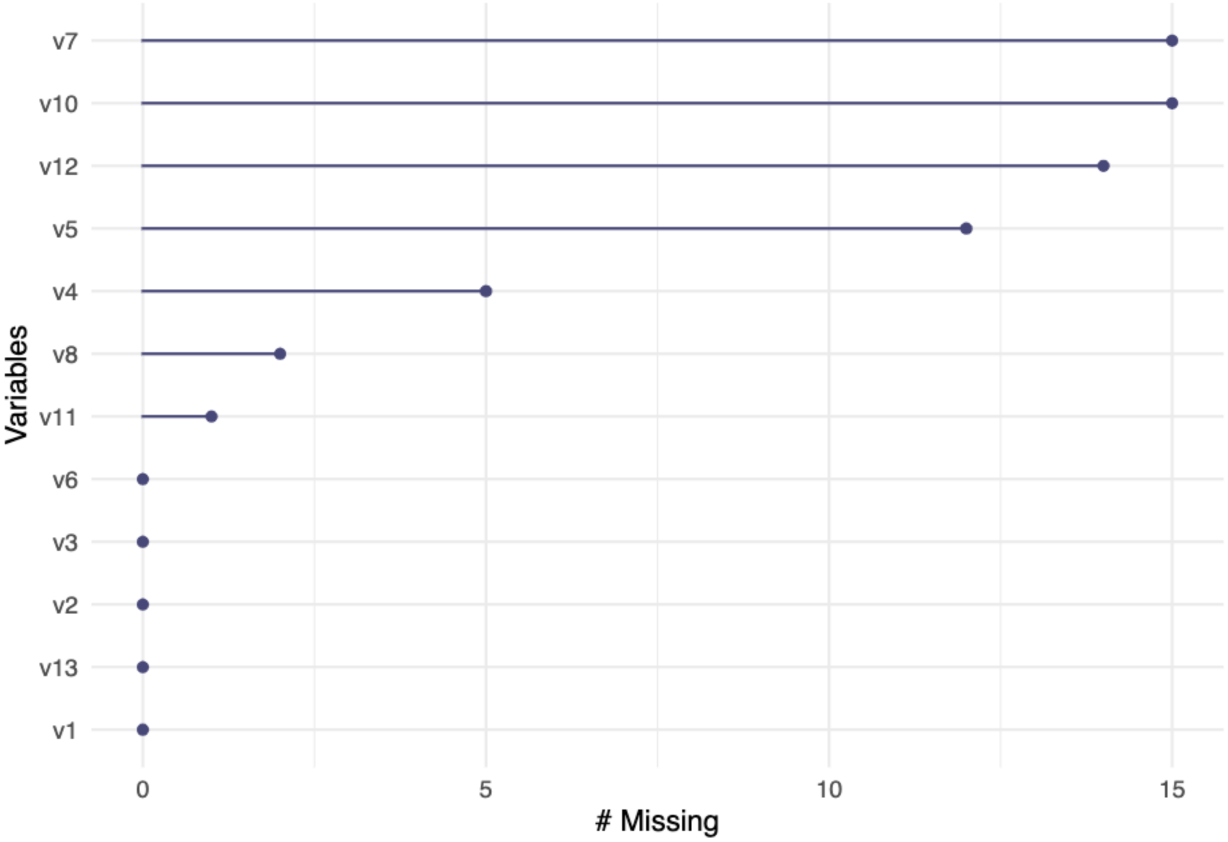
```
create_report(Ozone2, y = "v4", output_file = "report_Ozone.html", output_dir = getwd())
```

```
## Error in create_report(Ozone2, y = "v4", output_file = "report_Ozone.html", : could not find functio
```

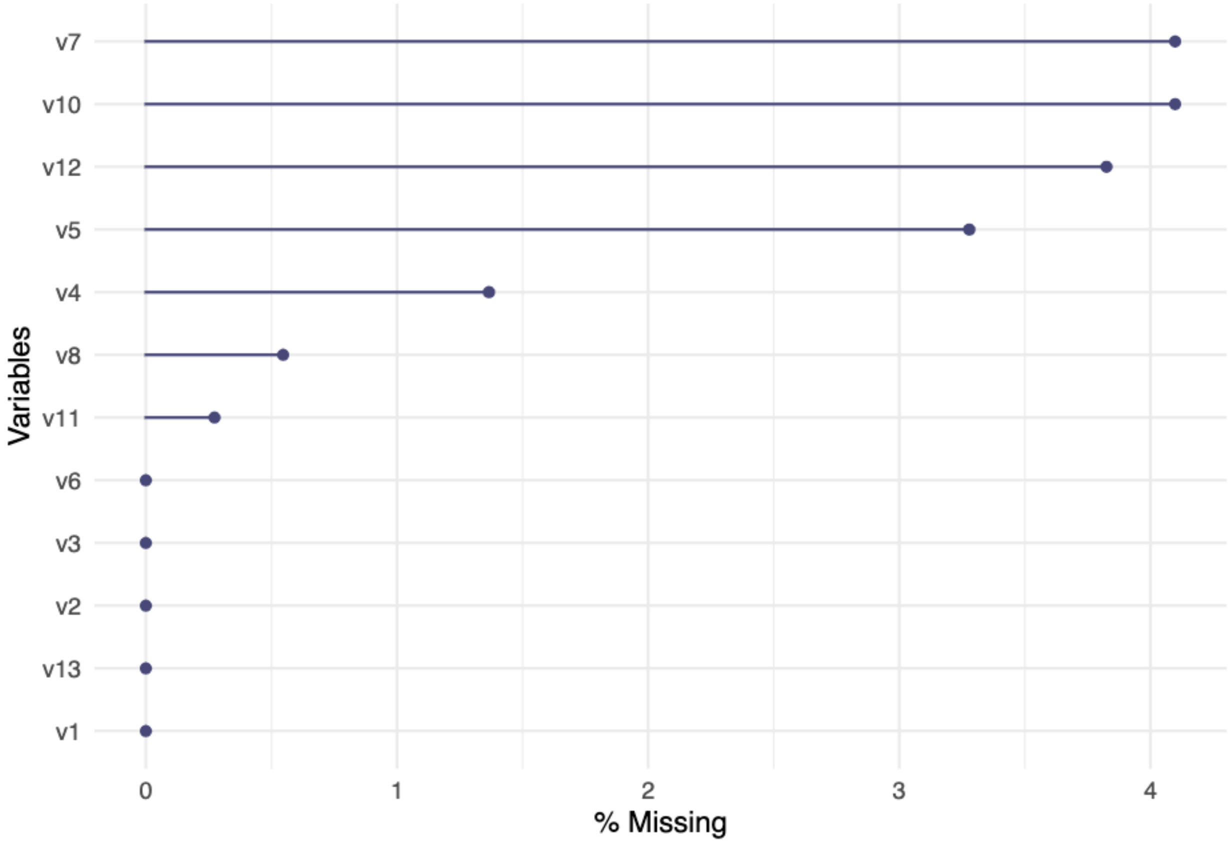
```
Ozone2 <- Ozone2 %>% select(-v9)

vis_miss(Ozone2)
```



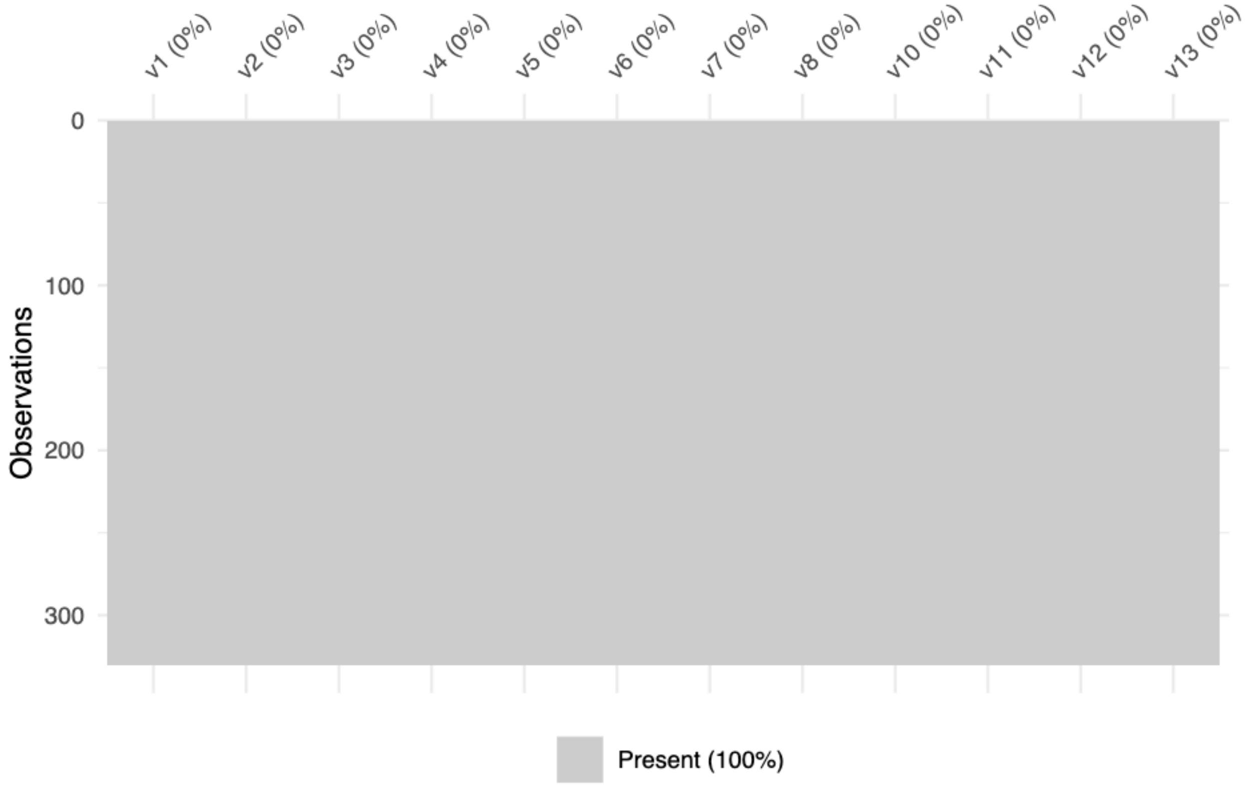
```
gg_miss_var(Ozone2)
```



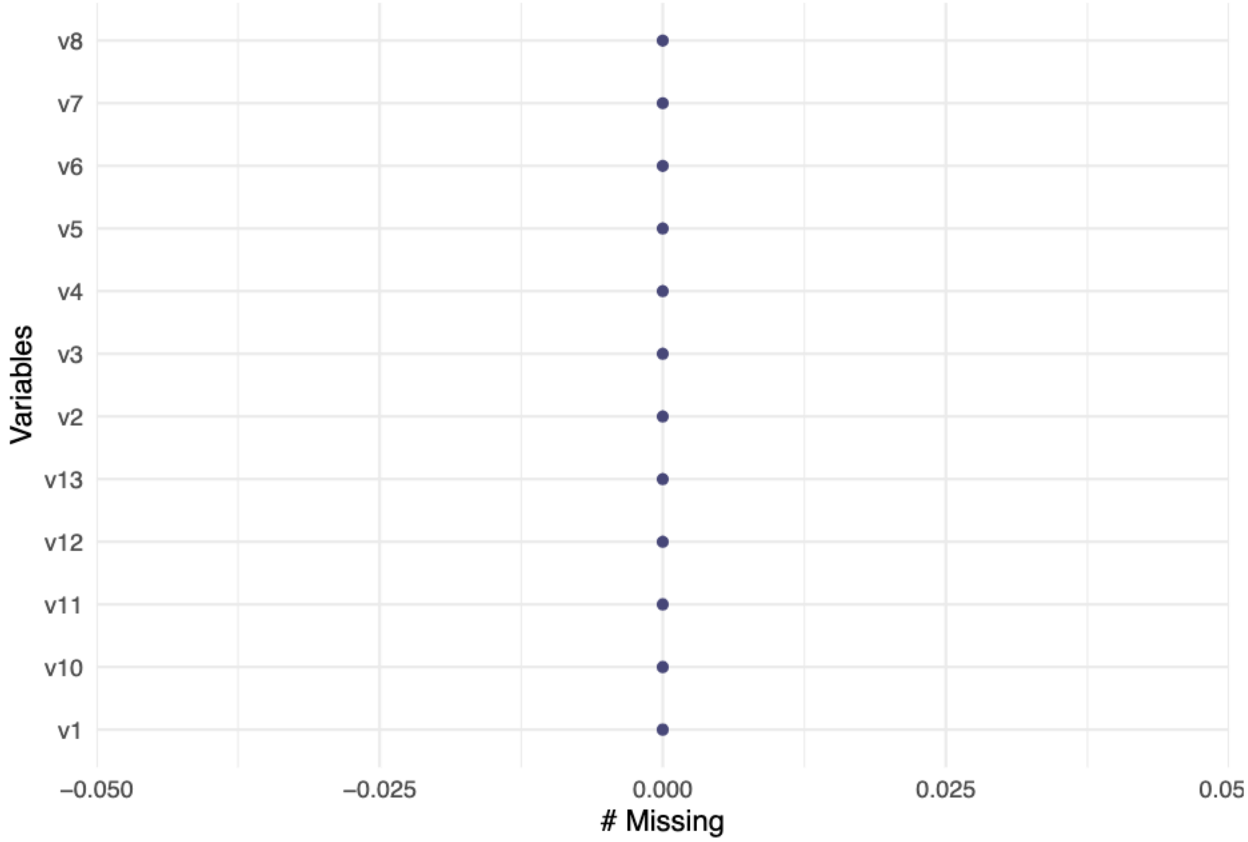
```r
gg_miss_var(Ozone2, show_pct = TRUE)
```



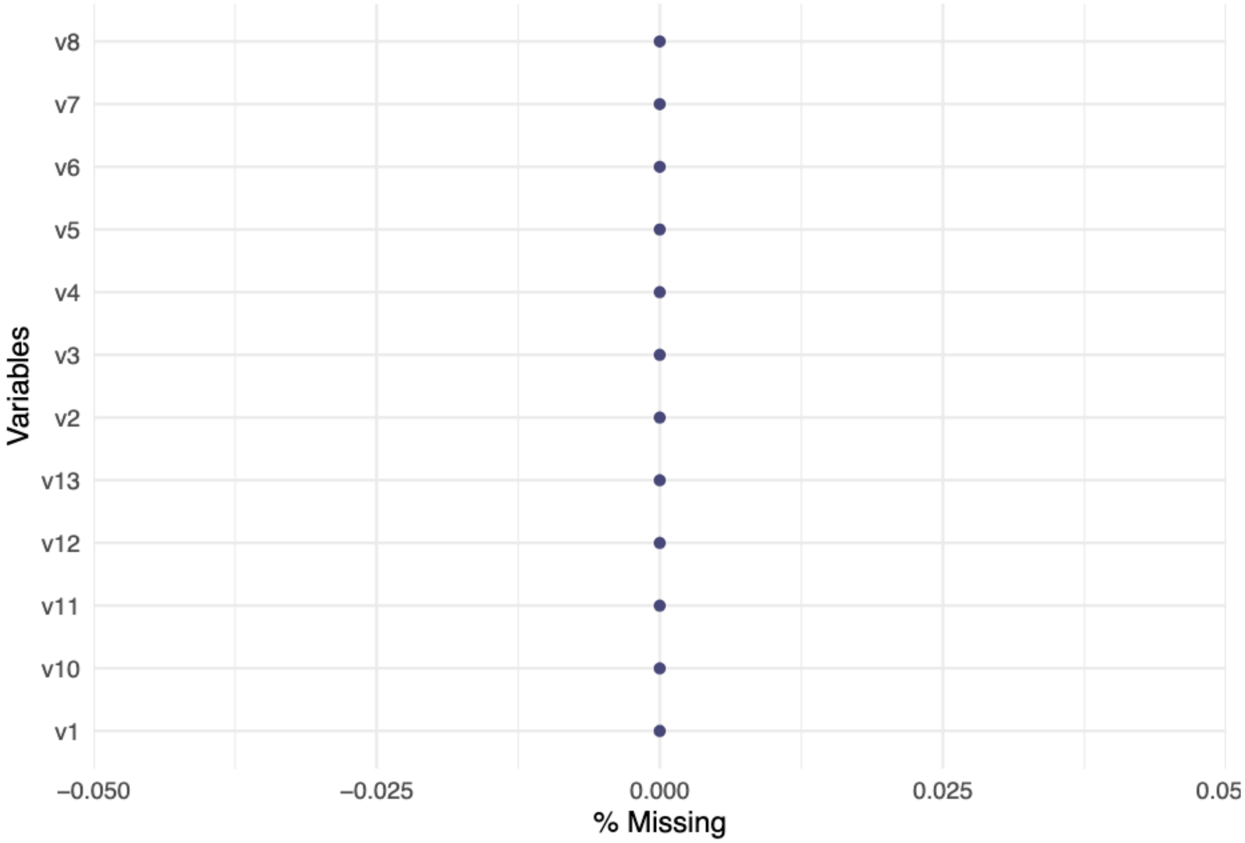
```r
Ozone2 <- drop_na(Ozone2)

vis_miss(Ozone2)
```



```
gg_miss_var(Ozone2)
```



```
gg_miss_var(Ozone2, show_pct = TRUE)
```



```
Boruta.Ozone <- Boruta(v4 ~ ., data = Ozone2, doTrace = 2, ntree = 500)
```

```
##  1. run of importance source...
##  2. run of importance source...
##  3. run of importance source...
##  4. run of importance source...
##  5. run of importance source...
##  6. run of importance source...
##  7. run of importance source...
##  8. run of importance source...
##  9. run of importance source...
##  10. run of importance source...
##  11. run of importance source...
## After 11 iterations, +0.88 secs:
##  confirmed 8 attributes: v1, v10, v11, v12, v13 and 3 more;
##  still have 3 attributes left.
##  12. run of importance source...
##  13. run of importance source...
##  14. run of importance source...
##  15. run of importance source...
## After 15 iterations, +1 secs:
##  rejected 1 attribute: v3;
##  still have 2 attributes left.
##  16. run of importance source...
##  17. run of importance source...
##  18. run of importance source...
##  19. run of importance source...
##  20. run of importance source...
##  21. run of importance source...
##  22. run of importance source...
##  23. run of importance source...
##  24. run of importance source...
##  25. run of importance source...
##  26. run of importance source...
##  27. run of importance source...
##  28. run of importance source...
##  29. run of importance source...
##  30. run of importance source...
##  31. run of importance source...
##  32. run of importance source...
##  33. run of importance source...
##  34. run of importance source...
##  35. run of importance source...
##  36. run of importance source...
##  37. run of importance source...
##  38. run of importance source...
##  39. run of importance source...
##  40. run of importance source...
##  41. run of importance source...
##  42. run of importance source...
##  43. run of importance source...
##  44. run of importance source...
##  45. run of importance source...
##  46. run of importance source...
##  47. run of importance source...
##  48. run of importance source...
##  49. run of importance source...
##  50. run of importance source...
##  51. run of importance source...
## After 51 iterations, +2.3 secs:
##  rejected 1 attribute: v2;
##  still have 1 attribute left.
##  52. run of importance source...
##  53. run of importance source...
##  54. run of importance source...
##  55. run of importance source...
##  56. run of importance source...
##  57. run of importance source...
##  58. run of importance source...
##  59. run of importance source...
##  60. run of importance source...
##  61. run of importance source...
##  62. run of importance source...
##  63. run of importance source...
##  64. run of importance source...
##  65. run of importance source...
##  66. run of importance source...
##  67. run of importance source...
##  68. run of importance source...
##  69. run of importance source...
##  70. run of importance source...
##  71. run of importance source...
##  72. run of importance source...
##  73. run of importance source...
##  74. run of importance source...
##  75. run of importance source...
##  76. run of importance source...
##  77. run of importance source...
##  78. run of importance source...
##  79. run of importance source...
##  80. run of importance source...
```

```
##  81. run of importance source...
```

```
## After 81 iterations, +3.3 secs:
```

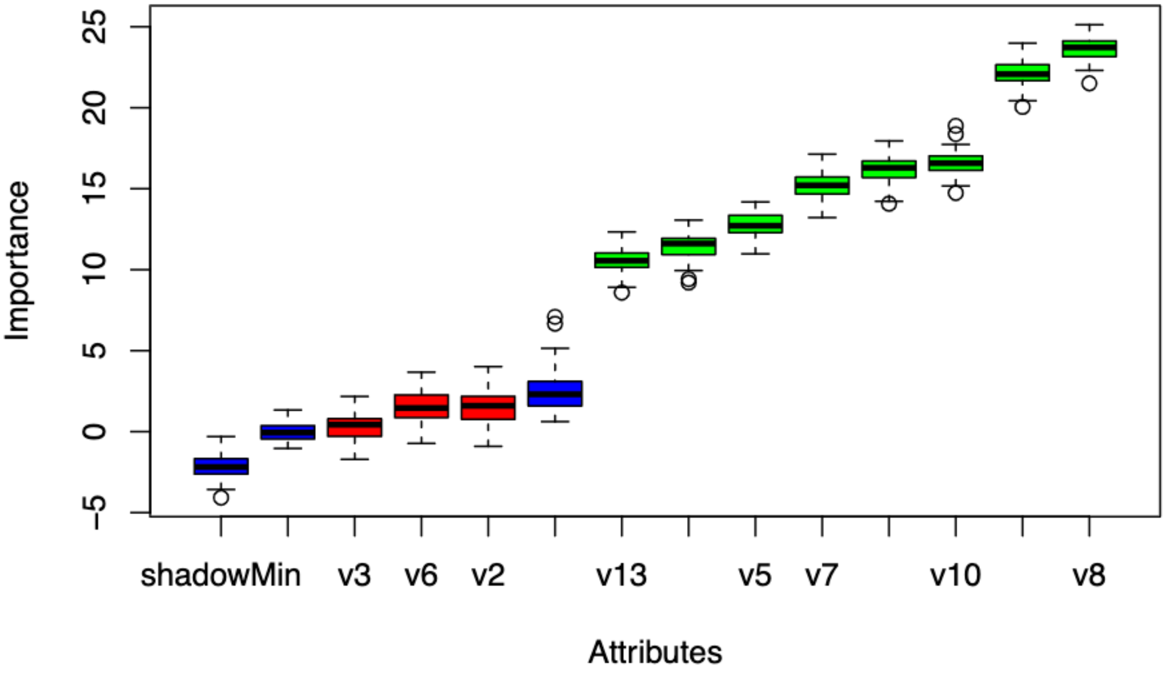
```
##  rejected 1 attribute: v6;
```

```
##  no more attributes left.
```

```
Boruta.Ozone
```

```
## Boruta performed 81 iterations in 3.319859 secs.
##  8 attributes confirmed important: v1, v10, v11, v12, v13 and 3 more;
##  3 attributes confirmed unimportant: v2, v3, v6;
```

```
plot(Boruta.Ozone)
```

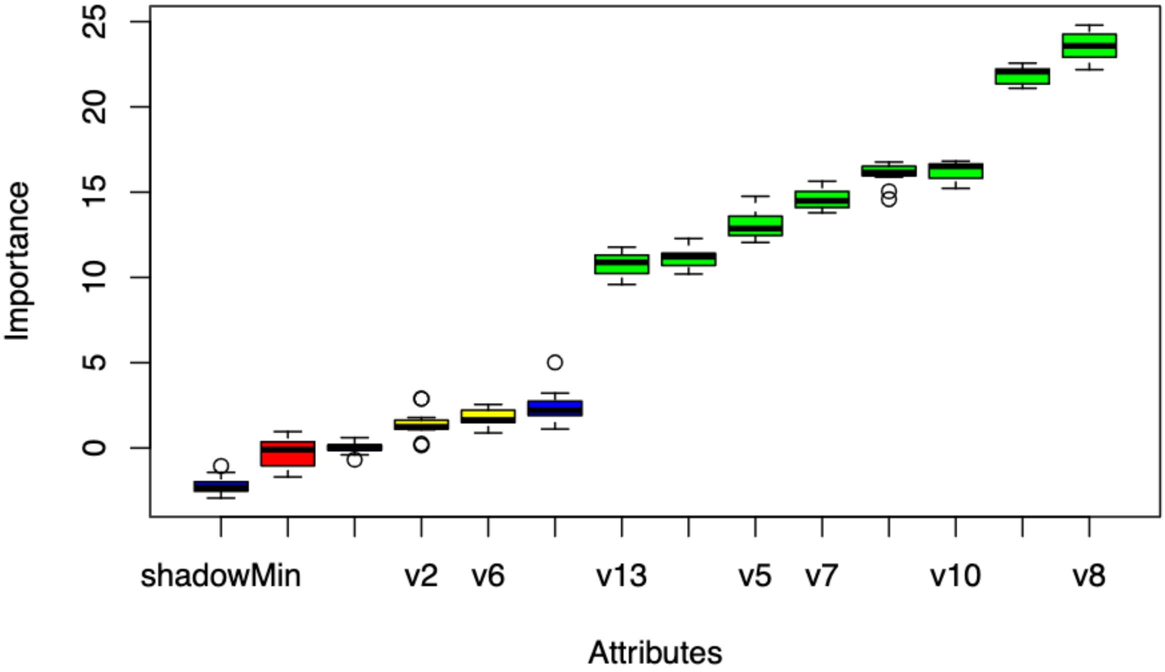


```
Boruta.Short <- Boruta(v4 ~ ., data = Ozone2, maxRuns = 12)
```

```
Boruta.Short
```

```
## Boruta performed 11 iterations in 0.402935 secs.
##  8 attributes confirmed important: v1, v10, v11, v12, v13 and 3 more;
##  1 attributes confirmed unimportant: v3;
##  2 tentative attributes left: v2, v6;
```

```
plot(Boruta.Short)
```



```
TentativeRoughFix(Boruta.Short)
```

```
## Boruta performed 11 iterations in 0.402935 secs.
## Tentatives roughfixed over the last 11 iterations.
##  8 attributes confirmed important: v1, v10, v11, v12, v13 and 3 more;
##  3 attributes confirmed unimportant: v2, v3, v6;
```

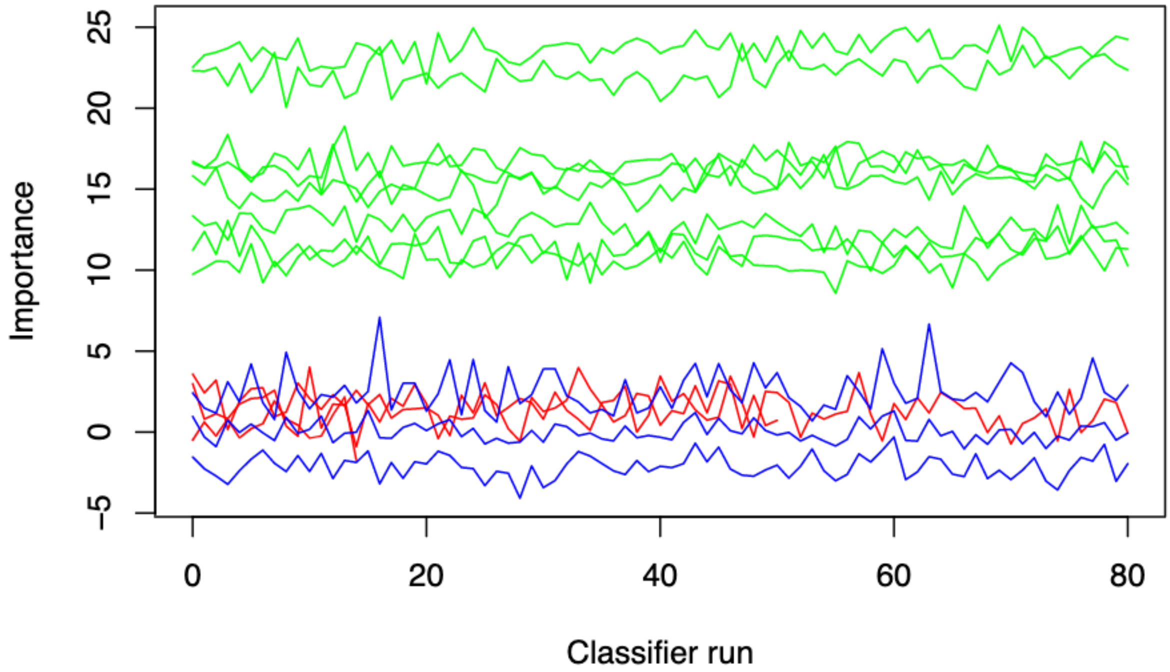
```
getConfirmedFormula(Boruta.Ozone)
```

```
## v4 ~ v1 + v5 + v7 + v8 + v10 + v11 + v12 + v13
## <environment: 0x13515e548>
```

```
attStats(Boruta.Ozone)
```

```
##        meanImp  medianImp     minImp    maxImp   normHits  decision
## v1  11.4107433 11.6143630  9.2038580 13.061376 1.00000000 Confirmed
## v2   1.4892821  1.5981335 -0.9060367  4.014716 0.17283951  Rejected
## v3   0.3438862  0.4303831 -1.7105361  2.173680 0.01234568  Rejected
## v5  12.7164579 12.7144143 10.9764705 14.182637 1.00000000 Confirmed
## v6   1.5033332  1.4575894 -0.7250994  3.670690 0.32098765  Rejected
## v7  15.1521678 15.2080500 13.2170177 17.133755 1.00000000 Confirmed
## v8  23.6783427 23.7225637 21.5064798 25.124967 1.00000000 Confirmed
## v10 16.5832420 16.5759119 14.7377945 18.882630 1.00000000 Confirmed
## v11 16.2682615 16.2910084 14.0736665 17.948995 1.00000000 Confirmed
## v12 22.1093730 22.0742165 20.0493032 23.986057 1.00000000 Confirmed
## v13 10.5564179 10.5516133  8.5811450 12.332580 1.00000000 Confirmed
```

```
plotImpHistory(Boruta.Ozone)
```



```
library(doParallel)
```

```
## Loading required package: foreach
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

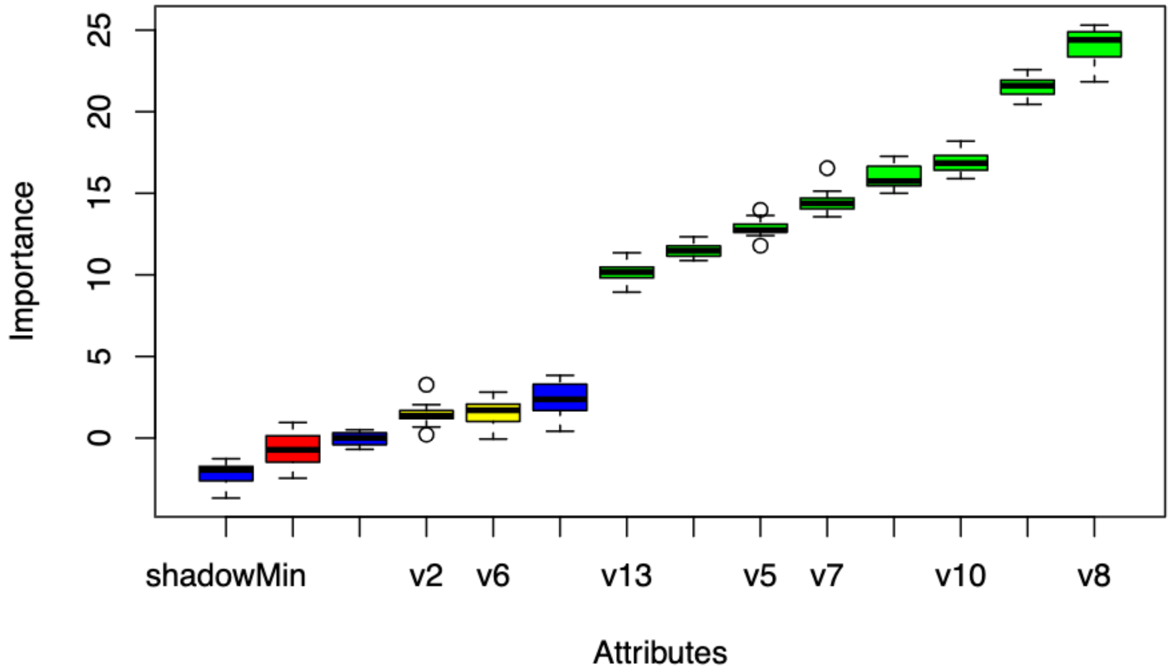
```
## Loading required package: iterators
```

```
## Loading required package: parallel
```

```
Boruta.Short <- Boruta(v4 ~ ., data = Ozone2, maxRuns = 12,doParallel = TRUE)
Boruta.Short
```

```
## Boruta performed 11 iterations in 0.4053612 secs.
##  8 attributes confirmed important: v1, v10, v11, v12, v13 and 3 more;
##  1 attributes confirmed unimportant: v3;
##  2 tentative attributes left: v2, v6;
```

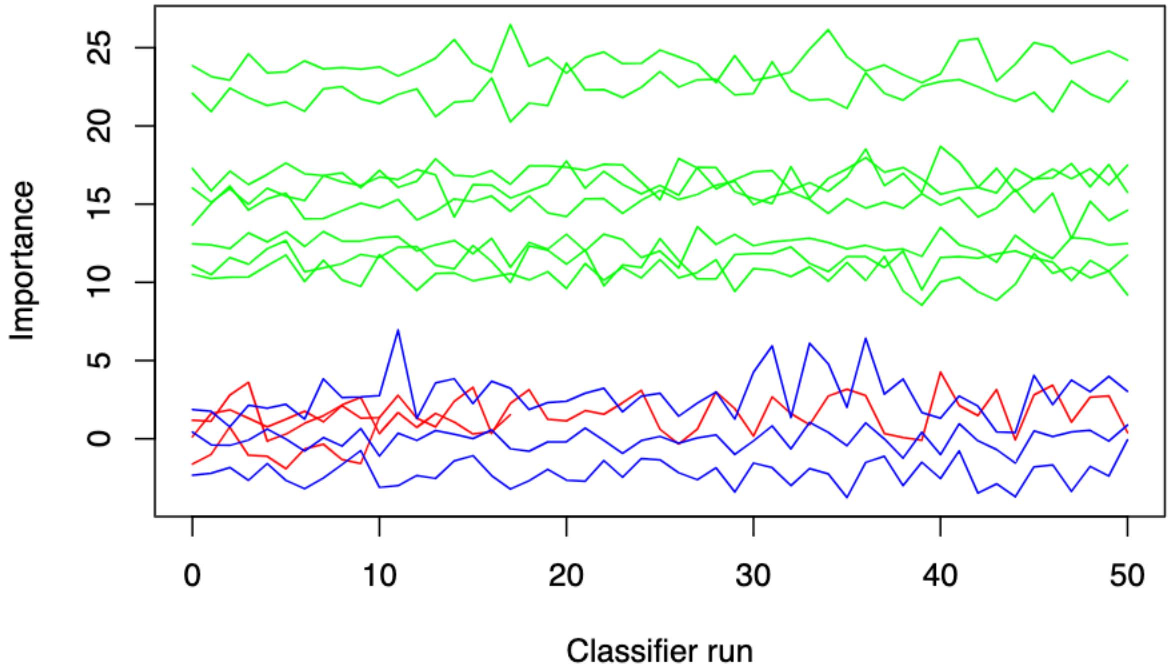
```
plot(Boruta.Short)
```



```
TentativeRoughFix(Boruta.Short)
```

```
## Boruta performed 11 iterations in 0.4053612 secs.
## Tentatives roughfixed over the last 11 iterations.
##  8 attributes confirmed important: v1, v10, v11, v12, v13 and 3 more;
##  3 attributes confirmed unimportant: v2, v3, v6;
```

```
Boruta.Ozone <- Boruta(v4 ~ ., data = Ozone2, doTrace = 2, ntree = 500,doParallel = TRUE)
plotImpHistory(Boruta.Ozone)
```


Importance
Classifier run