Ionosphere Dataset

1 Introduction

This radar data was collected by a system near Goose Bay, Labrador. This system is made up of a phased array of 16 high-frequency antennas with a total transmission power of around 6.4 kilowatts. The free electrons in the ionosphere were the targets. "Good" radar returns are those that show indications of ionosphere structure. Those with "poor" returns do not; their signals pass through the ionosphere.

The received signals were processed using an auto-correlation function with the pulse time and number as parameters. For the Goose Bay system, there were 17 pulse numbers. In this database, instances are defined by two characteristics per pulse number, which correspond to the complex values produced by the function as a result of the complex electromagnetic signal.

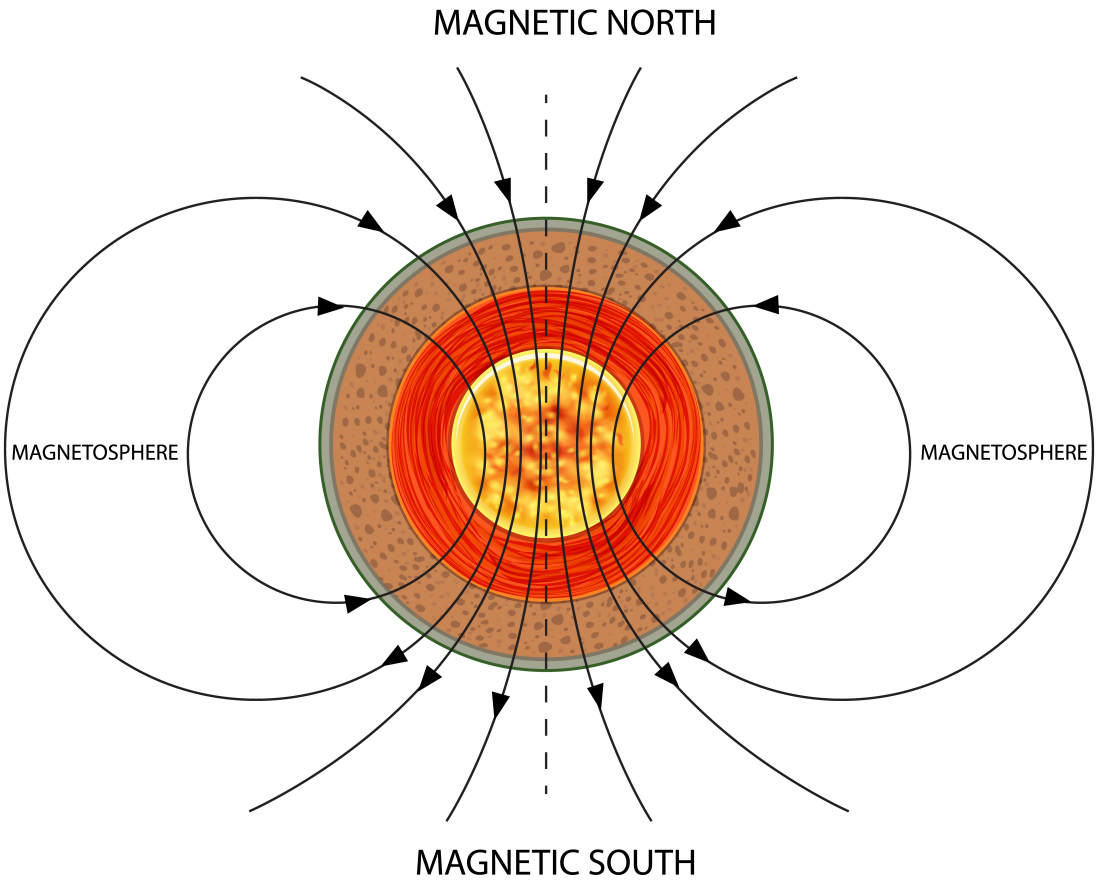


Figure 1: Source: brgfx (link)

Dataset basic information:

- `Class` (Target): "good"/"bad" radar signal.
- `V[1-34]`: 32 continuous variables + 2 factor variables, which represent 17 pulse numbers, each is characterized by 2 attributes.

To load the dataset, we use `mlbench`:

```
# load the dataset from mlbench
data(Ionosphere)
ionosphere <- Ionosphere %>% as_tibble() %>% dplyr::relocate(Class)
skimmed_ionosphere <- skimr::skim(ionosphere)
print(ionosphere)
```

```
## # A tibble: 351 x 35
##    Class V1    V2         V3       V4      V5       V6       V7     V8    V9
##    <fct> <fct> <fct>   <dbl>    <dbl>   <dbl>    <dbl>    <dbl>  <dbl> <dbl>
##  1 good  1     0      0.995  -0.0589   0.852   0.0231   0.834   -0.377 1
##  2 bad   1     0      1      -0.188    0.930  -0.362   -0.109   -0.936 1
##  3 good  1     0      1      -0.0336   1       0.00485  1       -0.121 0.890
##  4 bad   1     0      1      -0.452    1       1        0.712   -1     0
##  5 good  1     0      1      -0.0240   0.941   0.0653   0.921   -0.233 0.772
##  6 bad   1     0      0.0234 -0.00592 -0.0992 -0.119   -0.00763 -0.118 0.147
##  7 good  1     0      0.976  -0.106    0.946  -0.208    0.928   -0.284 0.860
##  8 bad   0     0      0       0        0       0        1       -1     0
##  9 good  1     0      0.964  -0.0720   1      -0.143    1       -0.213 1
## 10 bad   1     0     -0.0186 -0.0846   0       0        0        0     0.115
## # ... with 341 more rows, and 25 more variables: V10 <dbl>, V11 <dbl>,
## #   V12 <dbl>, V13 <dbl>, V14 <dbl>, V15 <dbl>, V16 <dbl>, V17 <dbl>,
## #   V18 <dbl>, V19 <dbl>, V20 <dbl>, V21 <dbl>, V22 <dbl>, V23 <dbl>,
## #   V24 <dbl>, V25 <dbl>, V26 <dbl>, V27 <dbl>, V28 <dbl>, V29 <dbl>,
## #   V30 <dbl>, V31 <dbl>, V32 <dbl>, V33 <dbl>, V34 <dbl>
```

2 Exploratory Data Analysis (EDA)

In this part, we will walk through a few characteristics of ionosphere dataset using library `skimr` and `DataExplorer`.

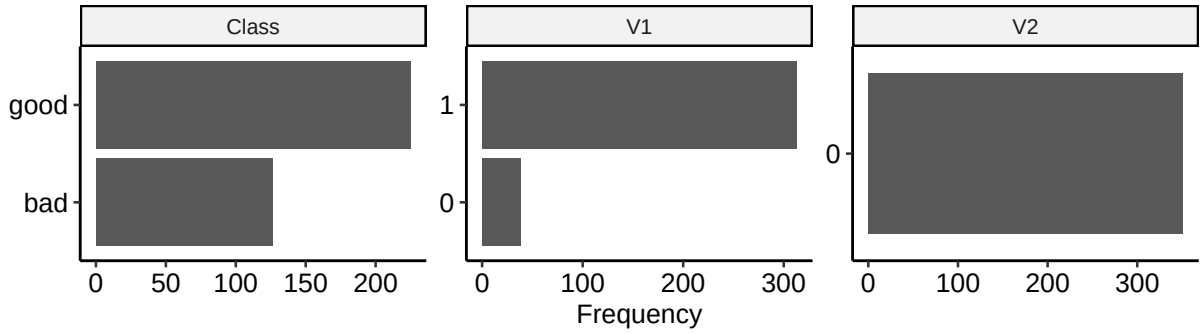
2.1 Factor variables

General statistics about factor variables from ionosphere dataset:

```
skimr::partition(skimmed_ionosphere)$factor %>%
        knitr::kable(format = 'latex', booktabs = TRUE) %>%
        kableExtra::kable_styling(latex_options = 'HOLD_position')
```

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Class	0	1	FALSE	2	goo: 225, bad: 126
V1	0	1	FALSE	2	1: 313, 0: 38
V2	0	1	FALSE	1	0: 351

```
DataExplorer::plot_bar(ionosphere, ggtheme = ggpubr::theme_pubr(base_size = 10))
```




```
DataExplorer::plot_bar(
        ionosphere %>% select("V1", "Class"),
        by = "Class",
        by_position = "dodge",
        ggtheme = ggpubr::theme_pubr(base_size = 10)
)
```

The dataset consists of 3 factor variables: `Class` (target), `V1` and `V2`. The three variables don't have missing values. `Class` is imbalanced with 225 observations labeled as `good` ($\approx 64\%$) and 126 as `bad` ($\approx 36\%$). The factor variable `V1` is even more imbalanced with 313 data points labeled as `1` ($\approx 89\%$) and only 38 as `0` ($\approx 11\%$). Noticeably, feature `V2` only has one label `0` for all data points.

From the bar plot of feature `V1` broken down by `Class`, it can be seen that observations with `V1=0` are all labeled `bad`, while with `V1=1`, the majority of the signals is `good`. This can be a useful feature for this classification task.

2.2 Numerical variables

General statistics about numerical variables from ionosphere dataset:

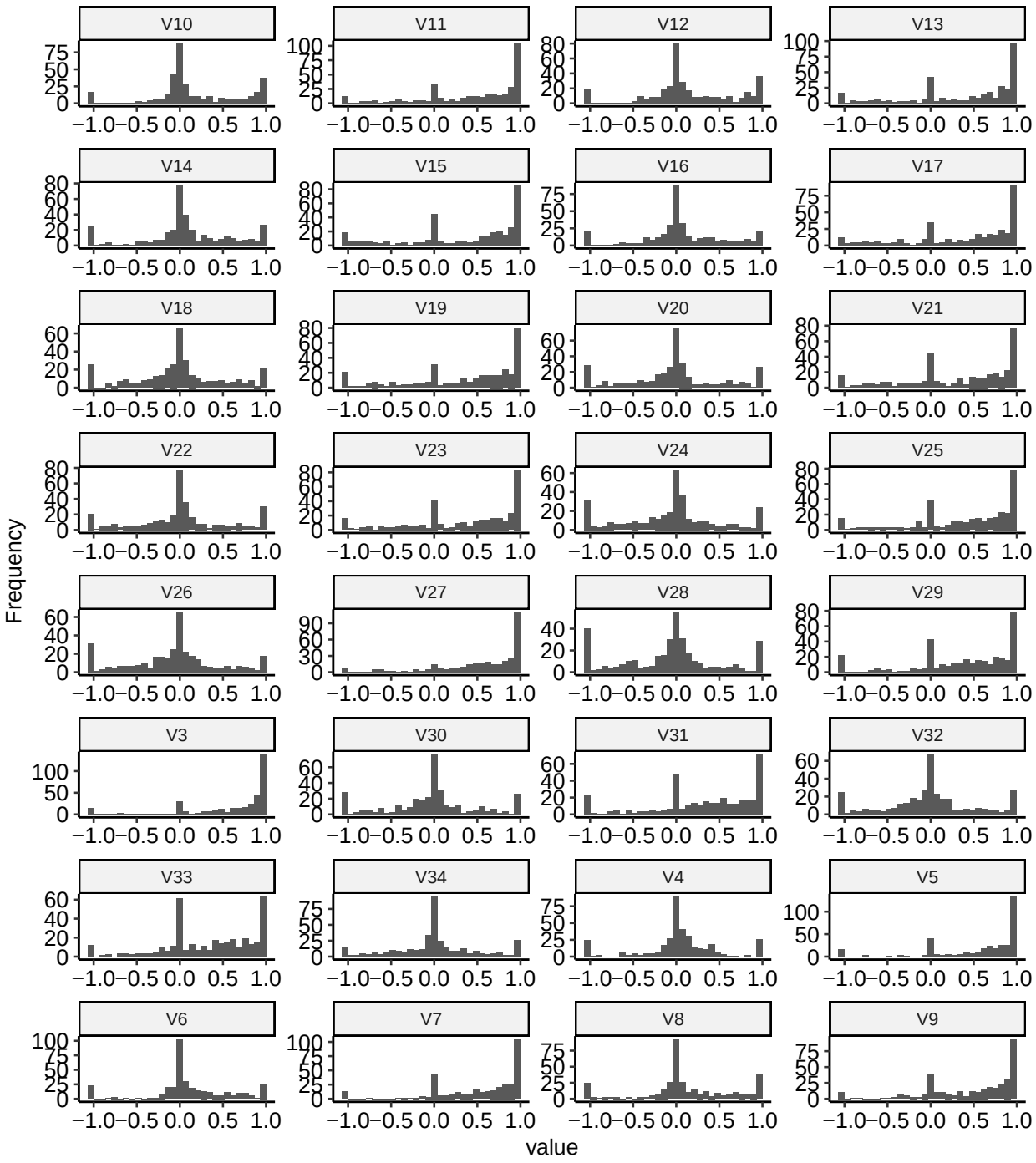
```
skimr::partition(skimmed_ionosphere)$numeric %>%
        knitr::kable(format = 'latex', booktabs = TRUE, digits = 2) %>%
        kableExtra::kable_styling(latex_options = 'HOLD_position')
```

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
V3	0	1	0.64	0.50	-1	0.47	0.87	1.00	1	
V4	0	1	0.04	0.44	-1	-0.06	0.02	0.19	1	
V5	0	1	0.60	0.52	-1	0.41	0.81	1.00	1	
V6	0	1	0.12	0.46	-1	-0.02	0.02	0.33	1	
V7	0	1	0.55	0.49	-1	0.21	0.73	0.97	1	
V8	0	1	0.12	0.52	-1	-0.05	0.01	0.45	1	
V9	0	1	0.51	0.51	-1	0.09	0.68	0.95	1	
V10	0	1	0.18	0.48	-1	-0.05	0.02	0.53	1	
V11	0	1	0.48	0.56	-1	0.02	0.67	0.96	1	
V12	0	1	0.16	0.49	-1	-0.07	0.03	0.48	1	
V13	0	1	0.40	0.62	-1	0.00	0.64	0.96	1	
V14	0	1	0.09	0.49	-1	-0.07	0.03	0.37	1	
V15	0	1	0.34	0.65	-1	0.00	0.60	0.92	1	
V16	0	1	0.07	0.46	-1	-0.08	0.00	0.31	1	
V17	0	1	0.38	0.62	-1	0.00	0.59	0.94	1	
V18	0	1	0.00	0.50	-1	-0.23	0.00	0.20	1	
V19	0	1	0.36	0.63	-1	0.00	0.58	0.90	1	
V20	0	1	-0.02	0.52	-1	-0.23	0.00	0.13	1	
V21	0	1	0.34	0.61	-1	0.00	0.50	0.89	1	
V22	0	1	0.01	0.52	-1	-0.24	0.00	0.19	1	
V23	0	1	0.36	0.60	-1	0.00	0.53	0.91	1	
V24	0	1	-0.06	0.53	-1	-0.37	0.00	0.16	1	
V25	0	1	0.40	0.58	-1	0.00	0.55	0.91	1	
V26	0	1	-0.07	0.51	-1	-0.33	-0.02	0.16	1	
V27	0	1	0.54	0.52	-1	0.29	0.71	1.00	1	
V28	0	1	-0.07	0.55	-1	-0.44	-0.02	0.15	1	
V29	0	1	0.38	0.58	-1	0.00	0.50	0.88	1	
V30	0	1	-0.03	0.51	-1	-0.24	0.00	0.15	1	
V31	0	1	0.35	0.57	-1	0.00	0.44	0.86	1	
V32	0	1	0.00	0.51	-1	-0.24	0.00	0.20	1	
V33	0	1	0.35	0.52	-1	0.00	0.41	0.81	1	
V34	0	1	0.01	0.47	-1	-0.17	0.00	0.17	1	

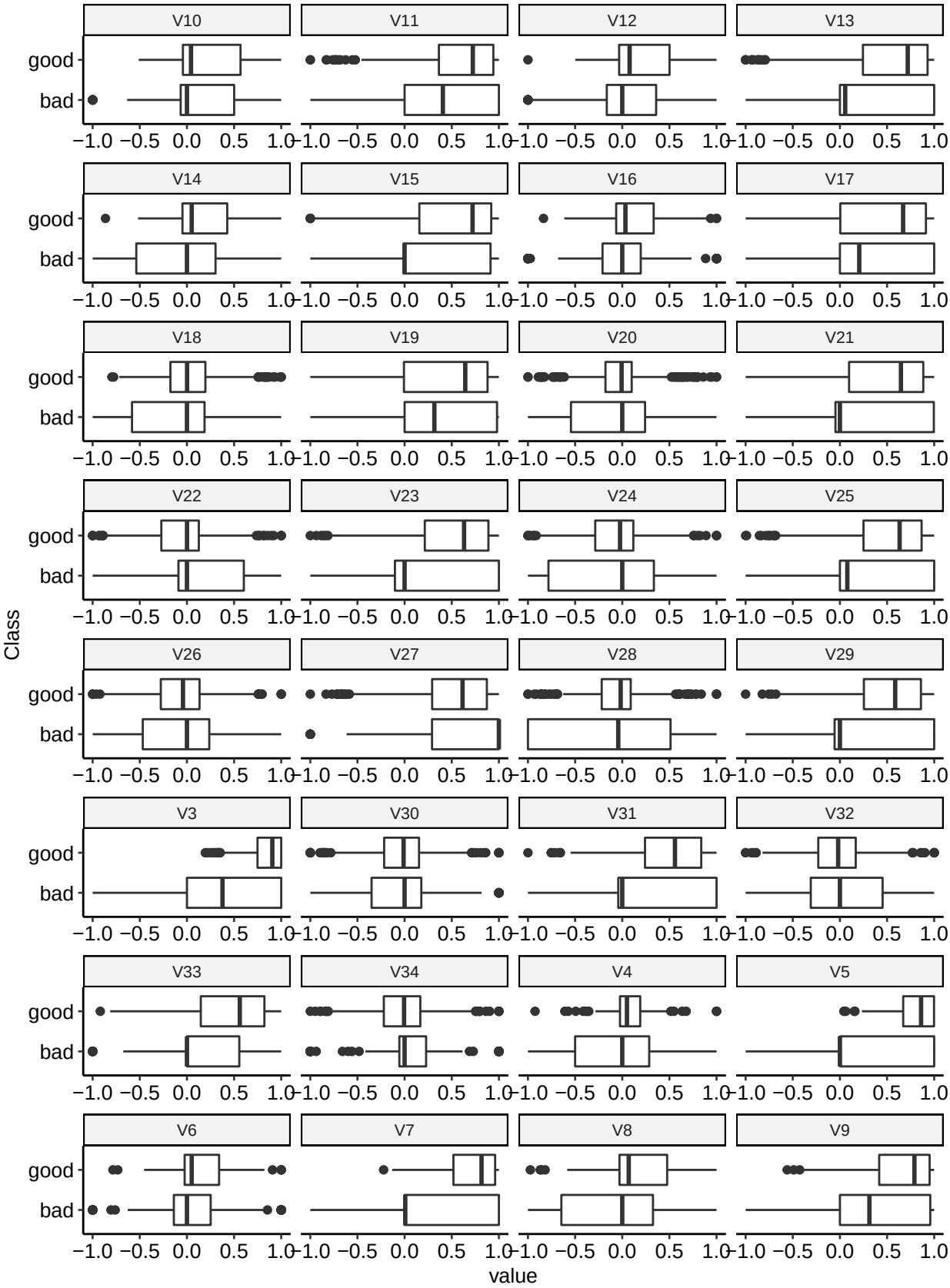
From the general statistics of the numerical features, it can be seen that 32 numerical features don't have missing values. Furthermore, all the numerical features share the same range of values: `[-1,1]`.

To have a better view at the distributions of these features, let's take a look at their histograms and their boxplots (broken down by class labels).

```
DataExplorer::plot_histogram(
        ionosphere,
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        ncol = 4, nrow = 8)
```




```
DataExplorer::plot_boxplot(
        ionosphere,
        by = "Class",
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        ncol = 4, nrow = 8)
```

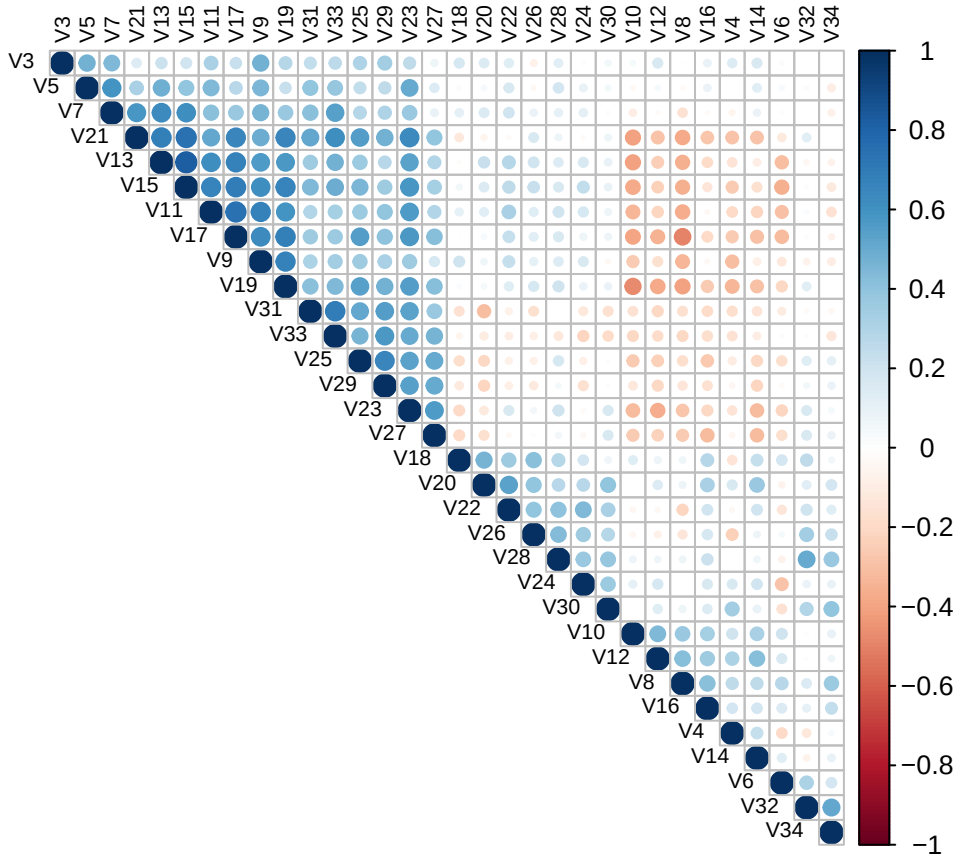

V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
V29
V3
V30
V31
V32
V33
V34
V4
V5
V6
V7
V8
V9
good
bad
−1.0 −0.5 0.0 0.5 1.0
Class
value

From the histograms, it can be seen that there are two patterns. The first pattern is that the majority of numerical features has fairly symmetrical distribution with the center and value with the highest frequency to be `0`, e.g. `V10`, `V12`, `V14`, `V16`, `V18`, etc. The second pattern is that other asymmetrically distributed numerical features appear to be highly left skewed with the mode to be `1`.

From the boxplots, there seems to be no strong relationship between any numerical feature and the class labels as the values of each feature are highly overlapping across the two class labels.

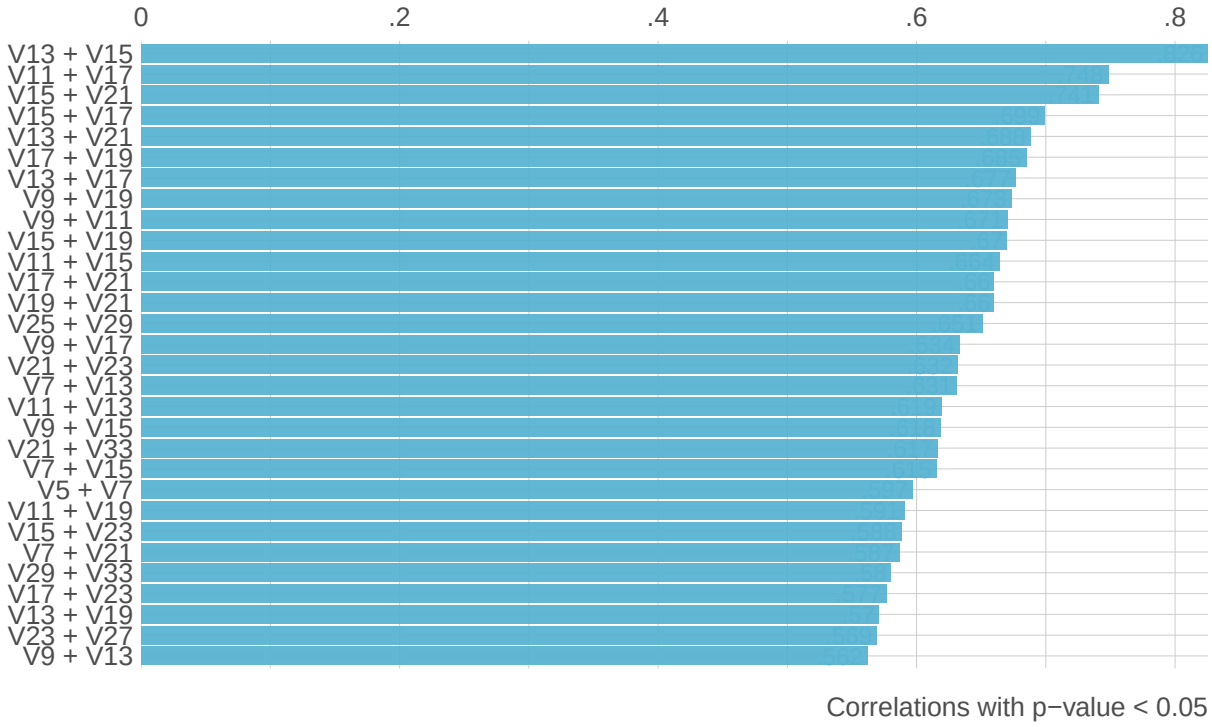
To understand more the linear relationship between the pairs of numerical variables, we create a correlation matrix:

```
ionosphere %>% select(where(is.numeric)) %>% cor() %>%
  corrplot(
   type = "upper",
   order = "hclust",
   tl.col = "black",
   tl.cex = 0.7
  )
```



From the correlation matrix, it can be seen that a lot of variables show strong linear relationship with each other. We can create a ranking of top 30 pairs of variables by the magnitude of correlation to interpret the result with `corr_cross` function from library `lares`:

```
corr_cross(ionosphere %>% select(where(is.numeric)),
  max_pvalue = 0.05, # display only significant correlations (at 5% level)
  top = 30 # display top 30 couples of variables (by correlation coefficient)
)
```



The top 30 pairs of features indicate that there are a lot of features that highly and positively correlated with each other. Noticeably, there are 3 pairs with correlation greater than 0.7, i.e. `V13-V15`, `V11-V17`, `V15-V21`, in which the pair `V15-V21` has the correlation surpassing 0.8.