HW5_YQ

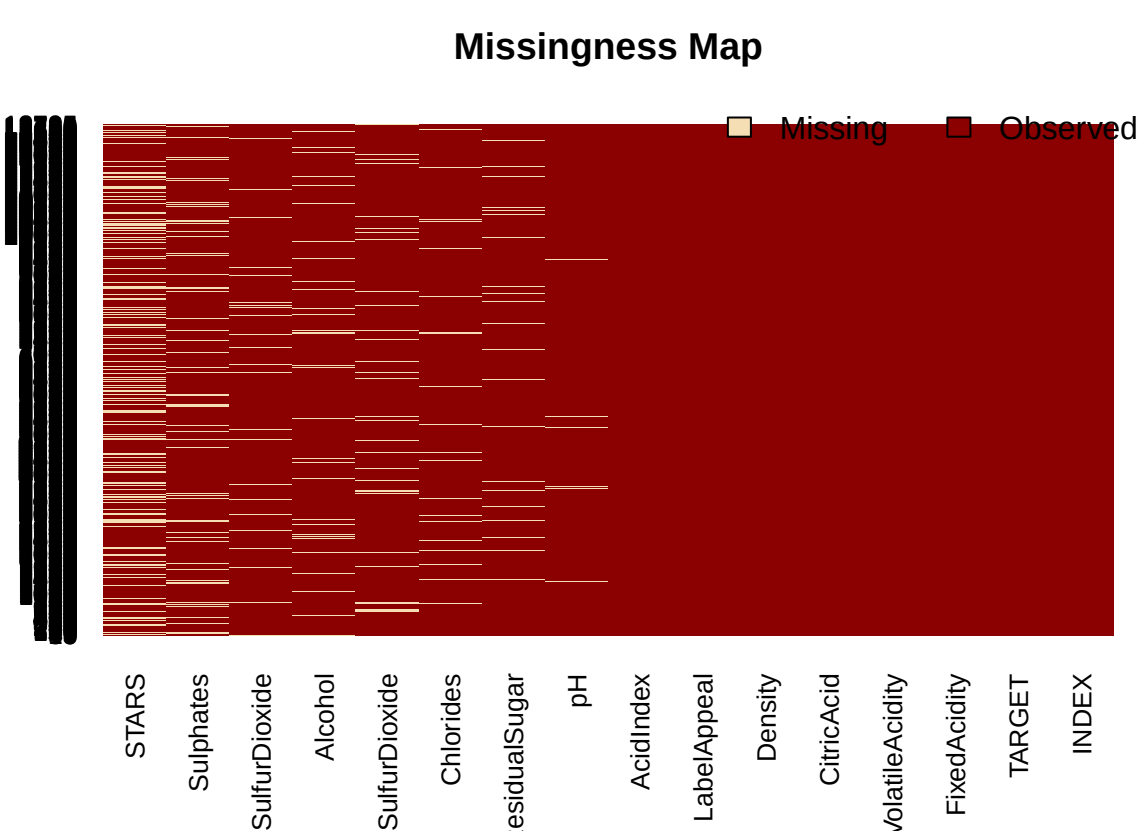
Youqing Xiang

July 16, 2016

Data Exploration

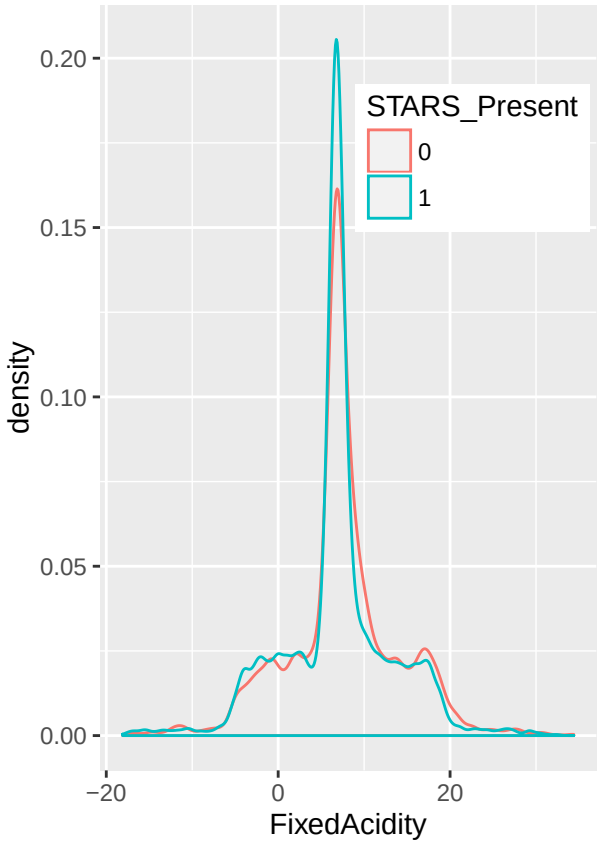
Variable	NAs	Percent_NAs
TARGET	0	0.000
FixedAcidity	0	0.000
VolatileAcidity	0	0.000
CitricAcid	0	0.000
ResidualSugar	616	0.048
Chlorides	638	0.050
FreeSulfurDioxide	647	0.051
TotalSulfurDioxide	682	0.053
Density	0	0.000
pH	395	0.031
Sulphates	1210	0.095
Alcohol	653	0.051
LabelAppeal	0	0.000
AcidIndex	0	0.000
STARS	3359	0.263

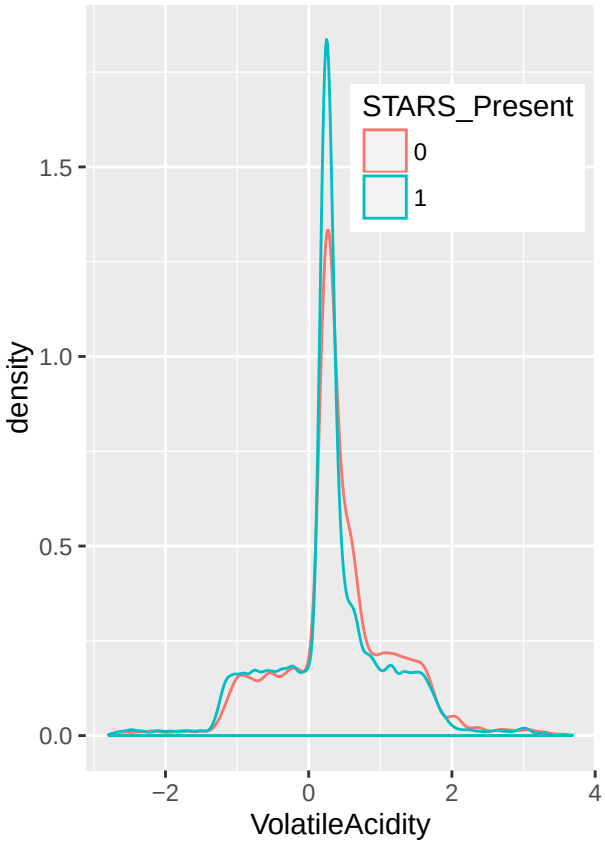
This dataset includes 12795 observations with 15 variables (`Index` columns excluded) in total. As the table shows above, there are a fair number of NAs and `STARS` variable has the most NAs.

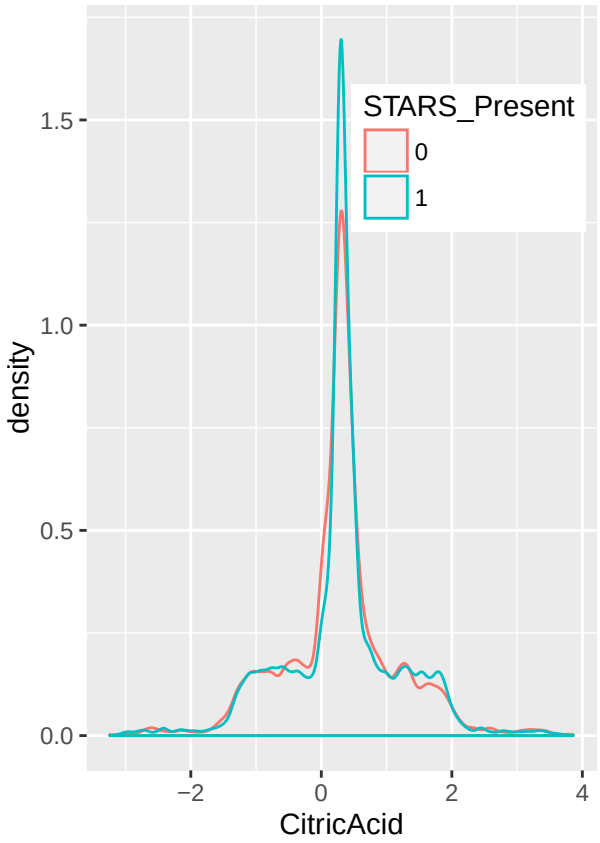


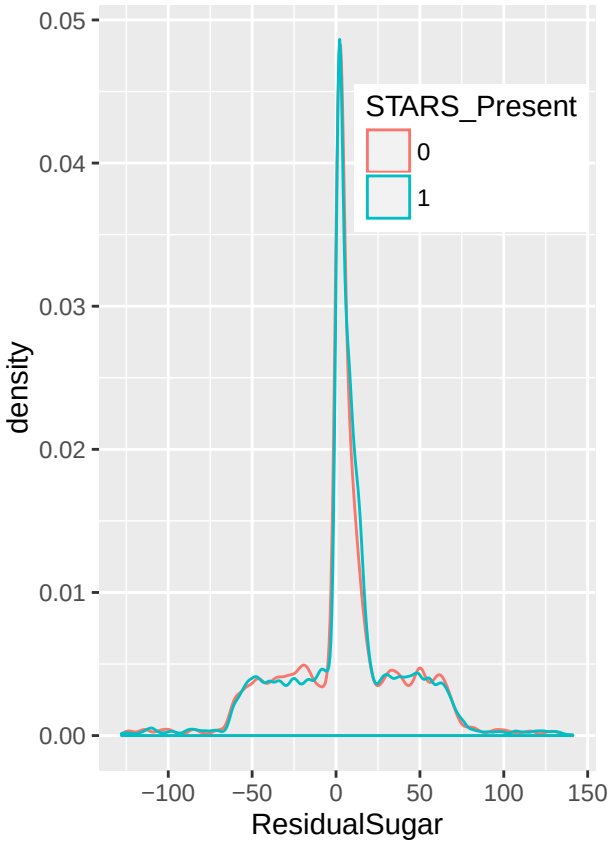
The above matrix of the NAs doesn't show any pattern. So, we could consider to replace the NA values with certain method to avoid potentially losing a lardge amount to data if we just simply drop off the NA values during data preparation process.

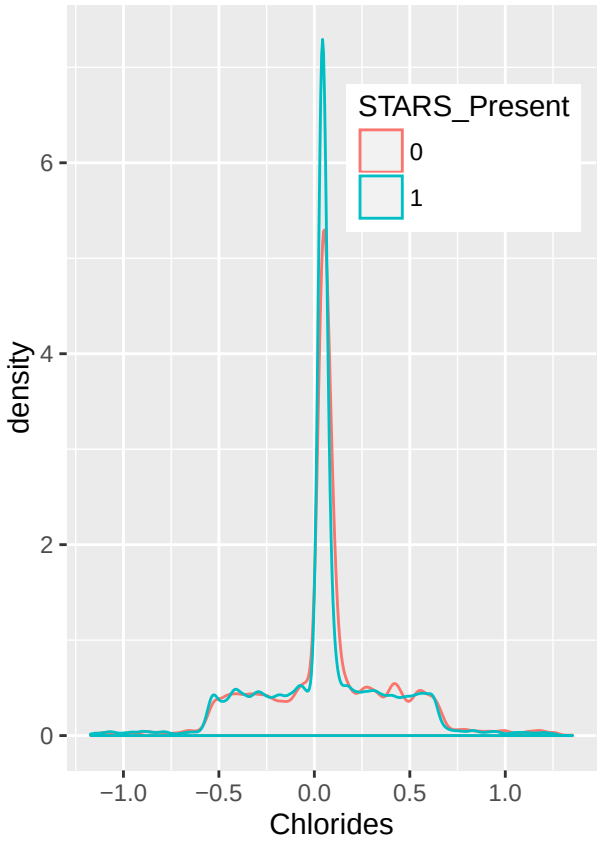
Since `STARS` variable has the most NAs, which accounts for 26.3% of data points, it is worthwhile to check how `STARS` variable NAs affect other variables, especially the `TARGET` variable. Here I created a new varialbe: `STARS_Present`, which is a categorical variable, equals to 0 when `STARS` value is missing and 1 when `STARS` value is present. And then I show density plots for each variable grouped by `STARS_Present`.

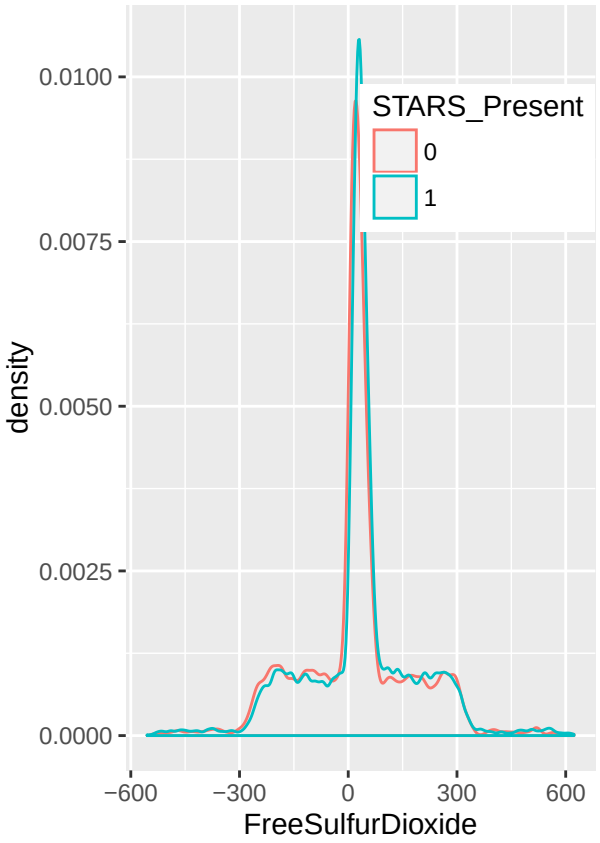

0.20
0.15
0.10
0.05
0.00
density
STARS_Present
0
1
−20
0
20
FixedAcidity

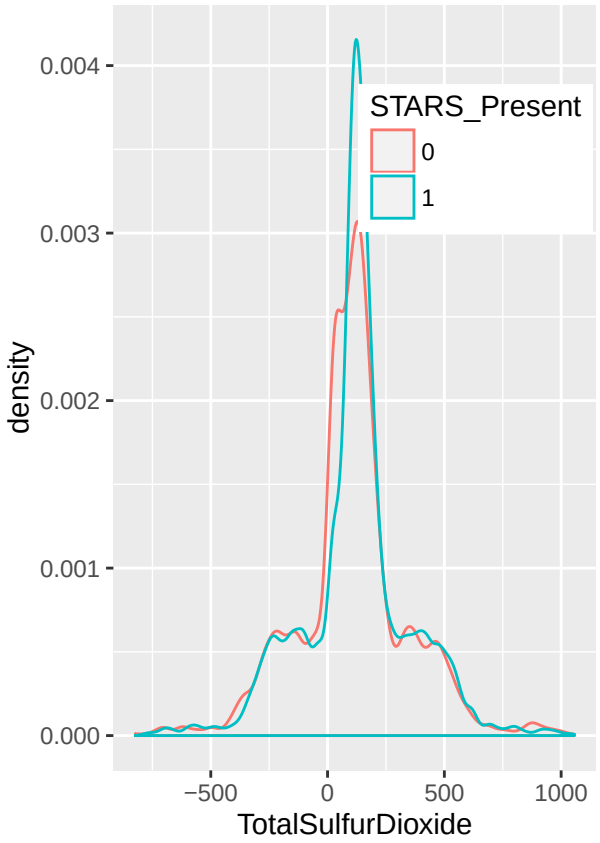

1.5
1.0
0.5
0.0
density
STARS_Present
0
1
−2
0
2
4
VolatileAcidity

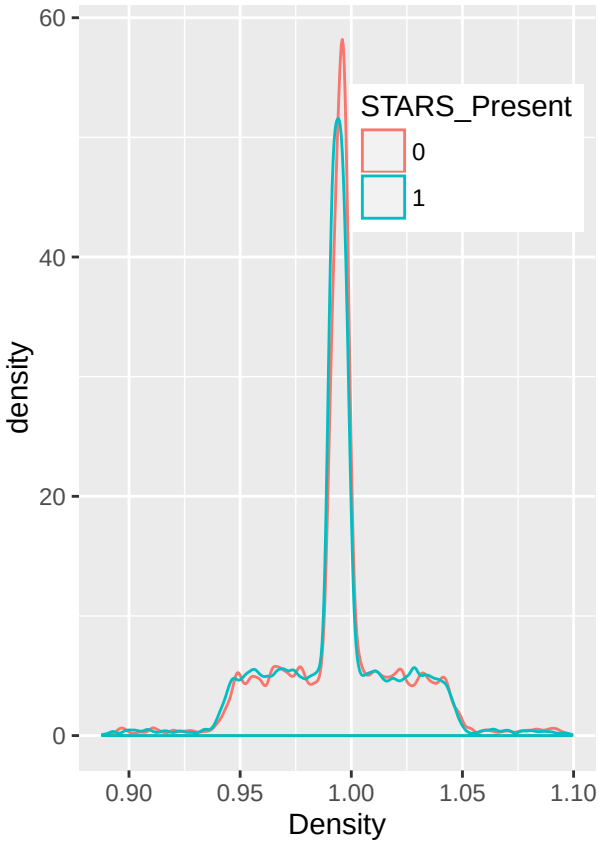

STARS_Present
0
1
density
1.5
1.0
0.5
0.0
−2
0
2
4
CitricAcid

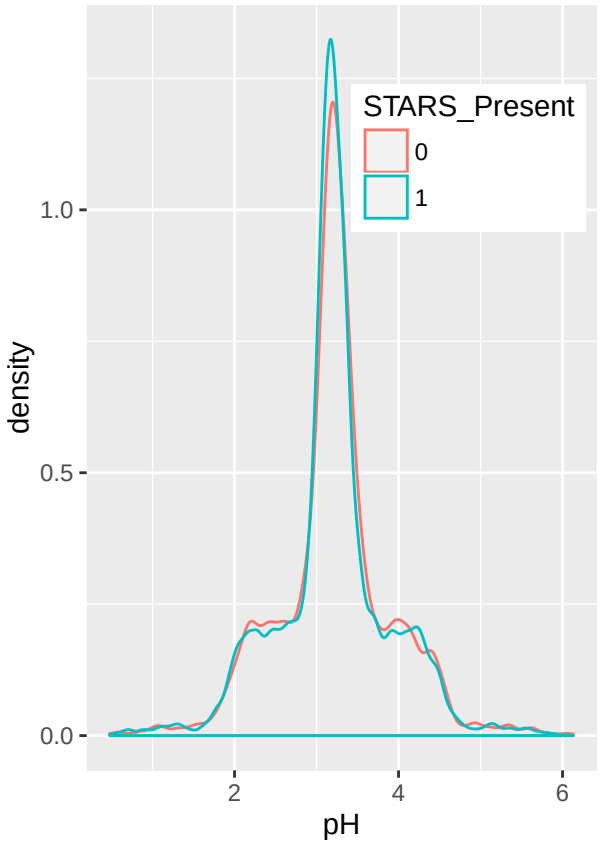

STARS_Present
0
1
density
0.05
0.04
0.03
0.02
0.01
0.00
−100
−50
0
50
100
150
ResidualSugar

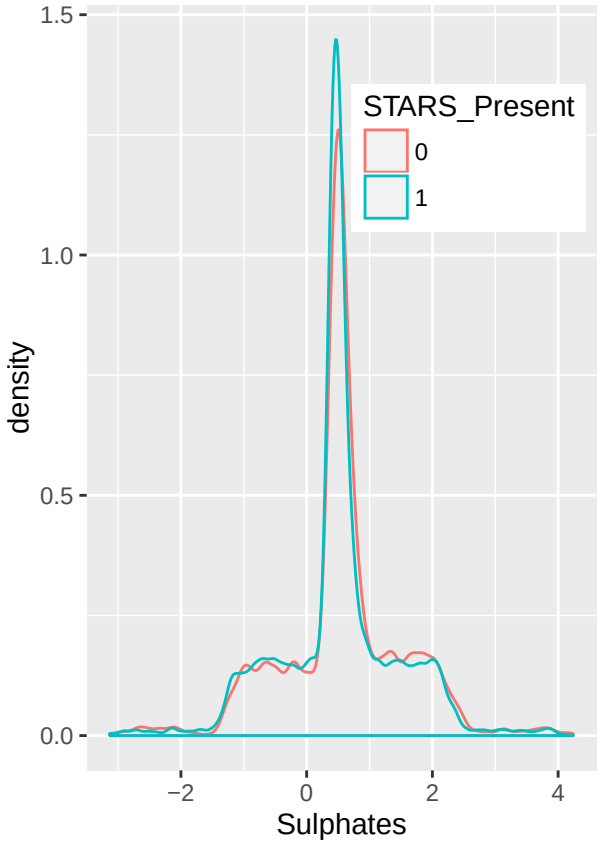

STARS_Present
0
1
density
6
4
2
0
−1.0
−0.5
0.0
0.5
1.0
Chlorides

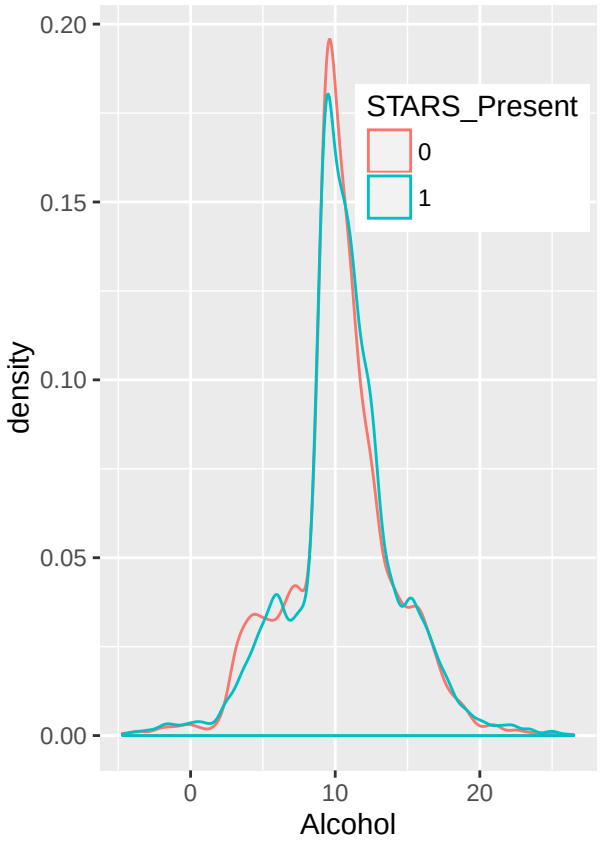

STARS_Present
0
1
density
0.0100
0.0075
0.0050
0.0025
0.0000
−600
−300
0
300
600
FreeSulfurDioxide

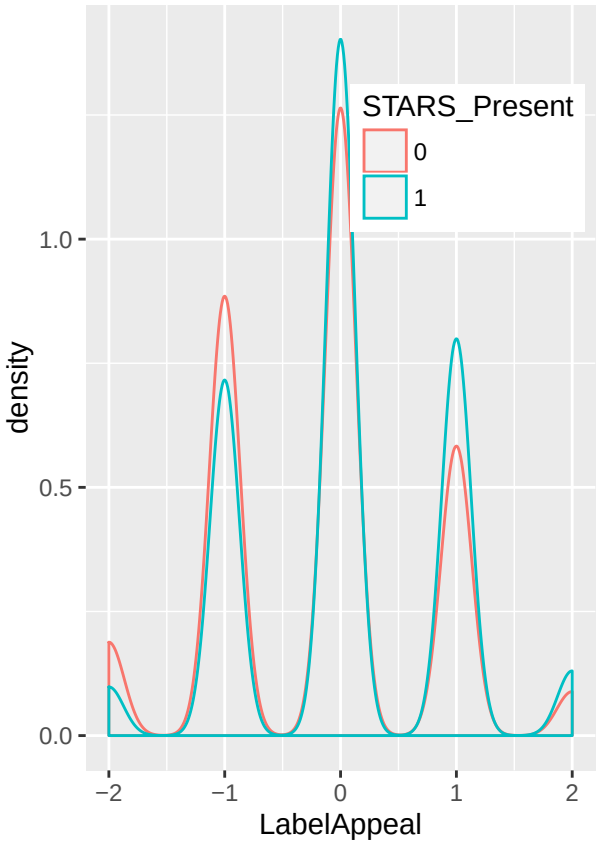

STARS_Present
0
1
density
0.004
0.003
0.002
0.001
0.000
−500
0
500
1000
TotalSulfurDioxide

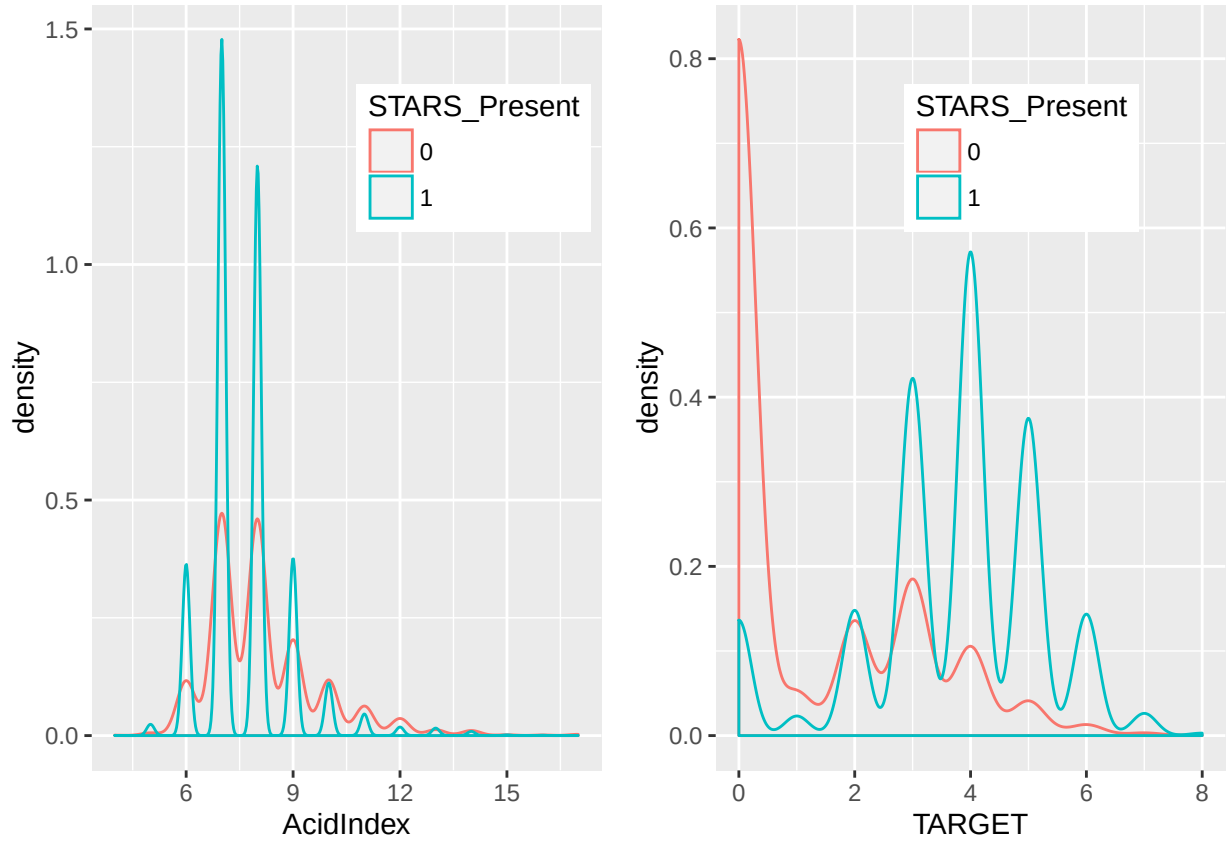

STARS_Present
0
1
density
60
40
20
0
0.90
0.95
1.00
1.05
1.10
Density


STARS_Present
0
1
density
1.0
0.5
0.0
2
4
6
pH


STARS_Present
0
1
density
1.5
1.0
0.5
0.0
−2
0
2
4
Sulphates


0.20
0.15
0.10
0.05
0.00
density
0
10
20
Alcohol
STARS_Present
0
1


1.0
0.5
0.0
density
−2
−1
0
1
2
LabelAppeal
STARS_Present
0
1



From the above density plots, we can see that `STARS` variable NAs have big effect on `TARGET` distribution but no obvious effects on other variables. And we can also conclude that **STARS variable NAs** is actually predictive of the target. Overall, this part of analysis suggests us that we should treat **STARS variable NAs** as category variable.