**Application**

**Return to the Vehicle data used in the previous lecture. Use the same split as**

**before.**

set . seed (46685326 , kind =" Mersenne - Twister ")

perm <- sample (x= nrow ( vehdata ))

set1 <- vehdata [ which ( perm <= 3\* nrow ( vehdata )/4) , ]

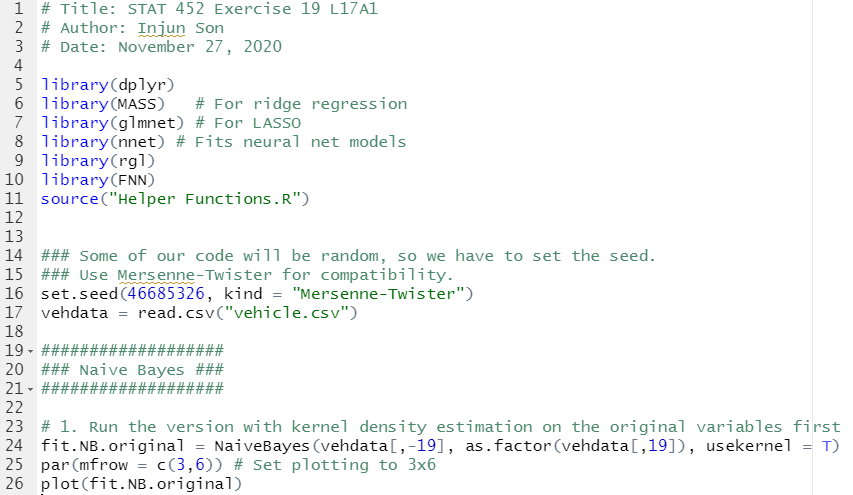
set2 <- vehdata [ which ( perm > 3\* nrow ( vehdata )/4) , ]

I already know that GAM will not fit these data. It uses enough degrees of freedom that it

creates a complete separation, and this ruins the logit model.

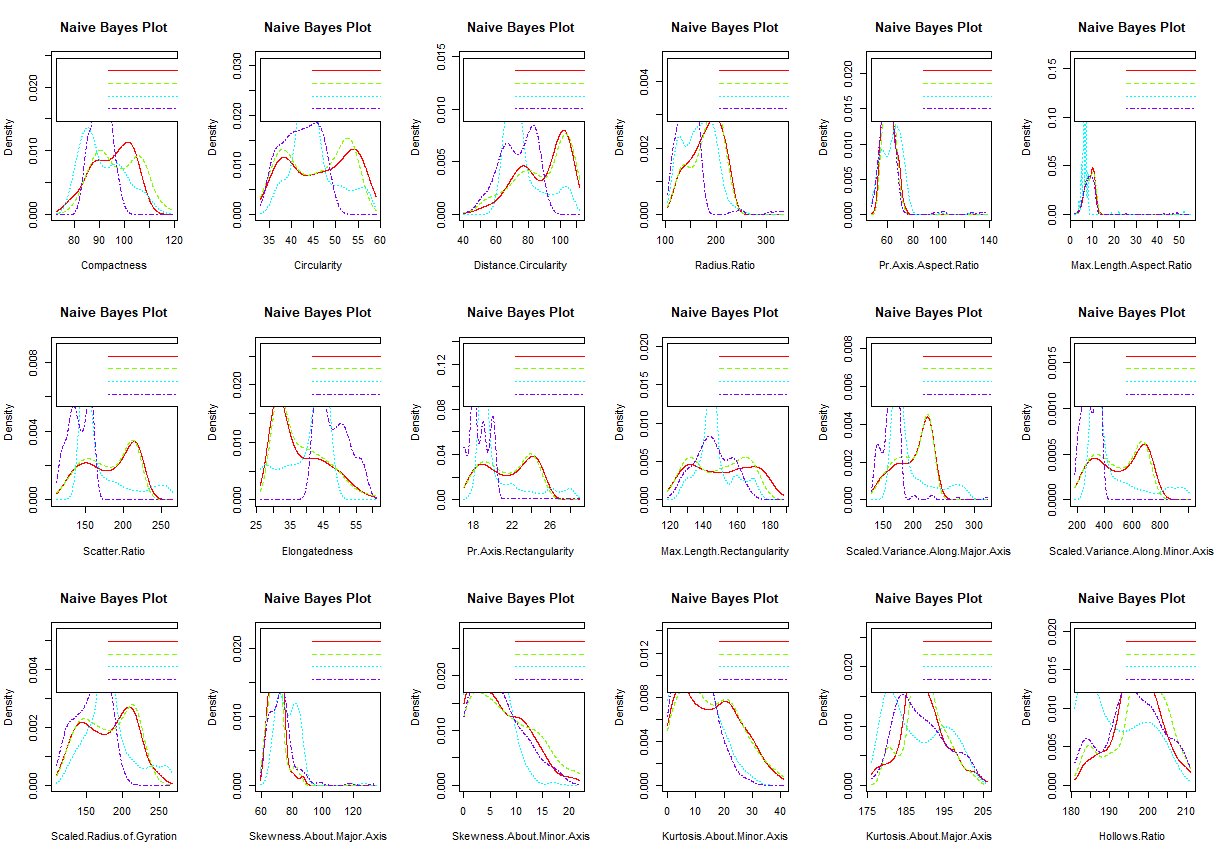
Instead, we will focus on Naive Bayes. You will fit the same four models that we used in

the example: with/without kernel, and with/without PC preprocessing.



1. Run the version with kernel density estimation on the original variables first

1. **Present plots of each variable’s density separated by classes.**



(b) Look at the plot:

i. Do many of the variables look like they have very skewed, multimodal, or otherwise non-normal distributions across the classes? If so, **name any that**

**seem pretty non-normal (but no more than three), and mention a word or two about what non-normal feature(s) each one has**.

* Yes, the graph with Pr.Axis.Aspect.Ratio and Max.Length.Aspect.Ratio looks very skewed and graph with Pr.Axis.Rectangularity looks multimodal.

ii. Do any of the variables look like they do a very good job of discriminating

among classes, particularly by separating their means? If so, **name any that**

**seem to separate the classes well (but no more than your top 3),**

**and mention which class(es) they seem to separate.**

* I think all of the graphs looks non-normal

In these questions, I mainly want to make sure you are absorbing the information

correctly and understanding what to look for. I do not have a specific three

variables that are “right” and consider all the rest “wrong”. Judgment may vary,

but the decisions should be made rationally.