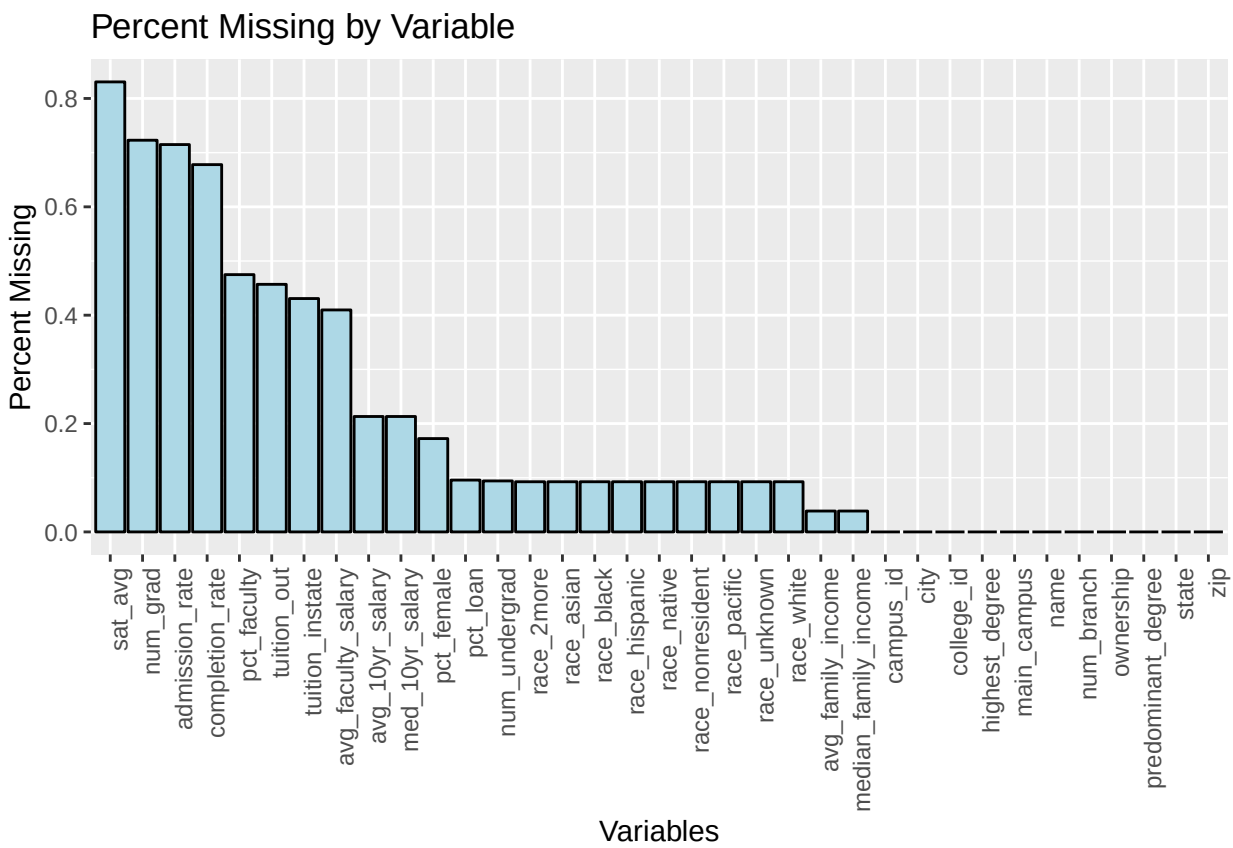
MissingValue

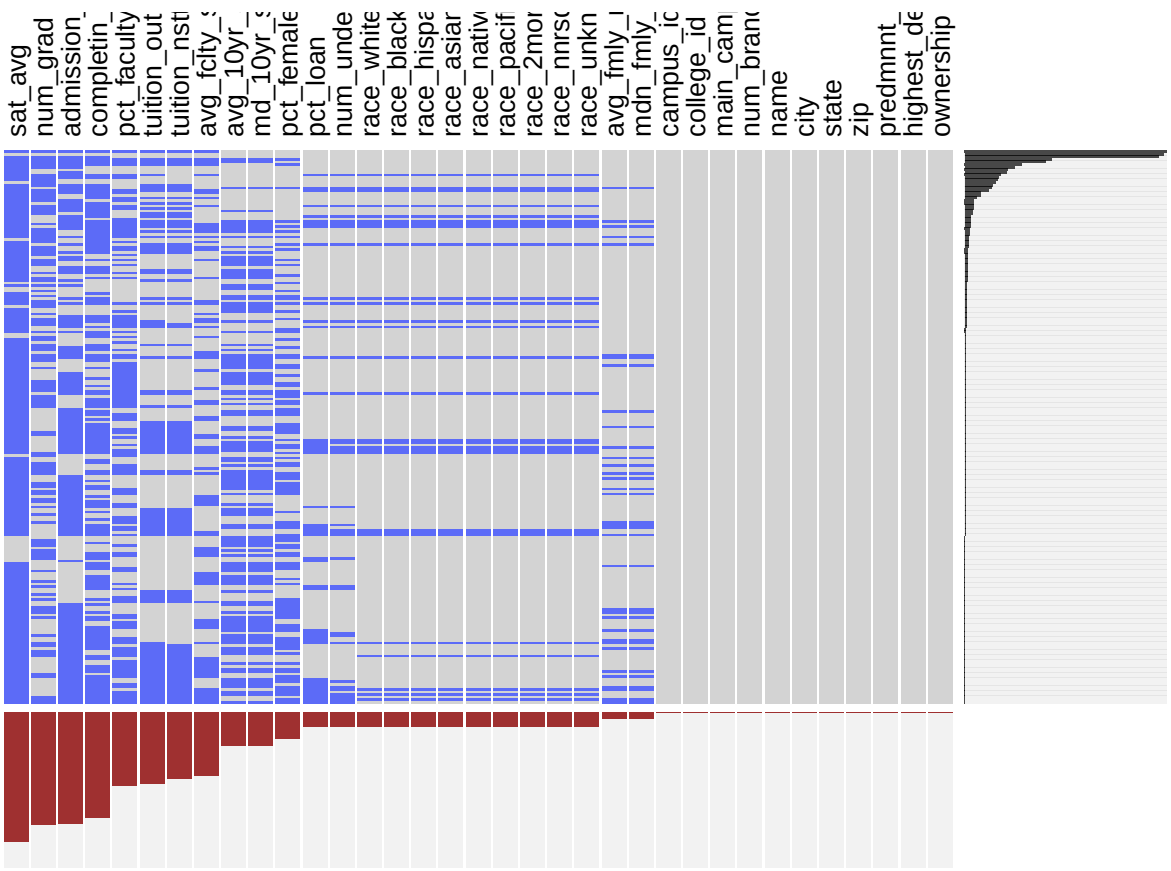
```
setwd(("/Users/Mavis1/Documents/ColumbiaUniv/Fall2018/EDAV/"))
library(tidyverse)
library(gridExtra)
college = read_rds("college.rds")
college[college=='NULL'] <- NA
a=data.frame(apply(college, 2, function(col)sum(is.na(col))/length(col)))
a$variables = rownames(a)
names(a) <- c("missing","variables")
rownames(a) <- NULL

ggplot(a, aes(x=reorder(factor(a$variables),-a$missing), y=a$missing))+
  geom_bar(stat="identity",fill = "lightblue", color = "black")+
  ggtitle("Percent Missing by Variable")+
  labs(x = "Variables", y = "Percent Missing") +
  theme(axis.text.x = element_text(size = rel(1),angle = 90, hjust = 1),
        axis.text.y = element_text(size = rel(1)),
        strip.text = element_text(size = rel(1)))
```



For most schools, average SAT scores are missing. Number of graduate students, admission rate, and completion rate have missing value over 50%. Percent of full time faculty, in state tuition, out-of-state tuition, and average faculty salary have missing value a little under 50%. Other variables have lower than 12% missing value.

```
library(extracat)
visna(college,sort="b")
```



The most frequent missing pattern is missing average SAT, number of graduate, admission rate, completion rate, percent of full time faculty, in-state tuition, out-of-state tuition, and average faculty salary at the same time. The eight variables are also the ones with most missing values. For race, missing race percentages are the same among different race types, and the missing pattern for race percentage is the same across data, meaning when one race percent is missing, all are missing.