Hmk4-Q3

Zhida Zhang (zz2578)

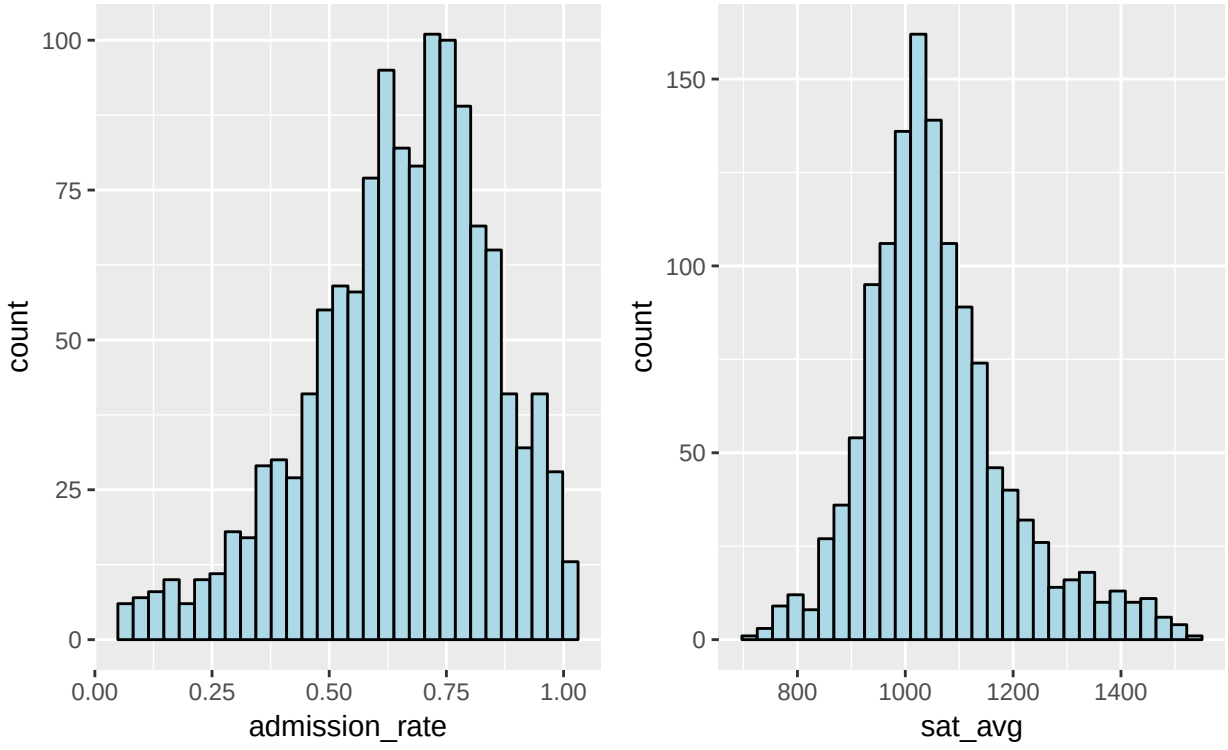
November 14, 2018

Admission Rate and SAT Average

To get a sense of how admission and SAT scores are distributed respectively, we plot the histogram on this two variables. From the histogram of admission rate, we could see that most schools have the admission rate range between 0.5 to 0.8. Only a small proportion of schools have admission rate below 0.25. From the SAT average histogram, we see that majority of students enrolled into universities have scores in the range of 900 to 1200. The high the score is, the less quantity of students there are.

However, how do we know the relationship between admission rate and SAT average? Do we have the intuitive pattern like higher score and lower admission rate?

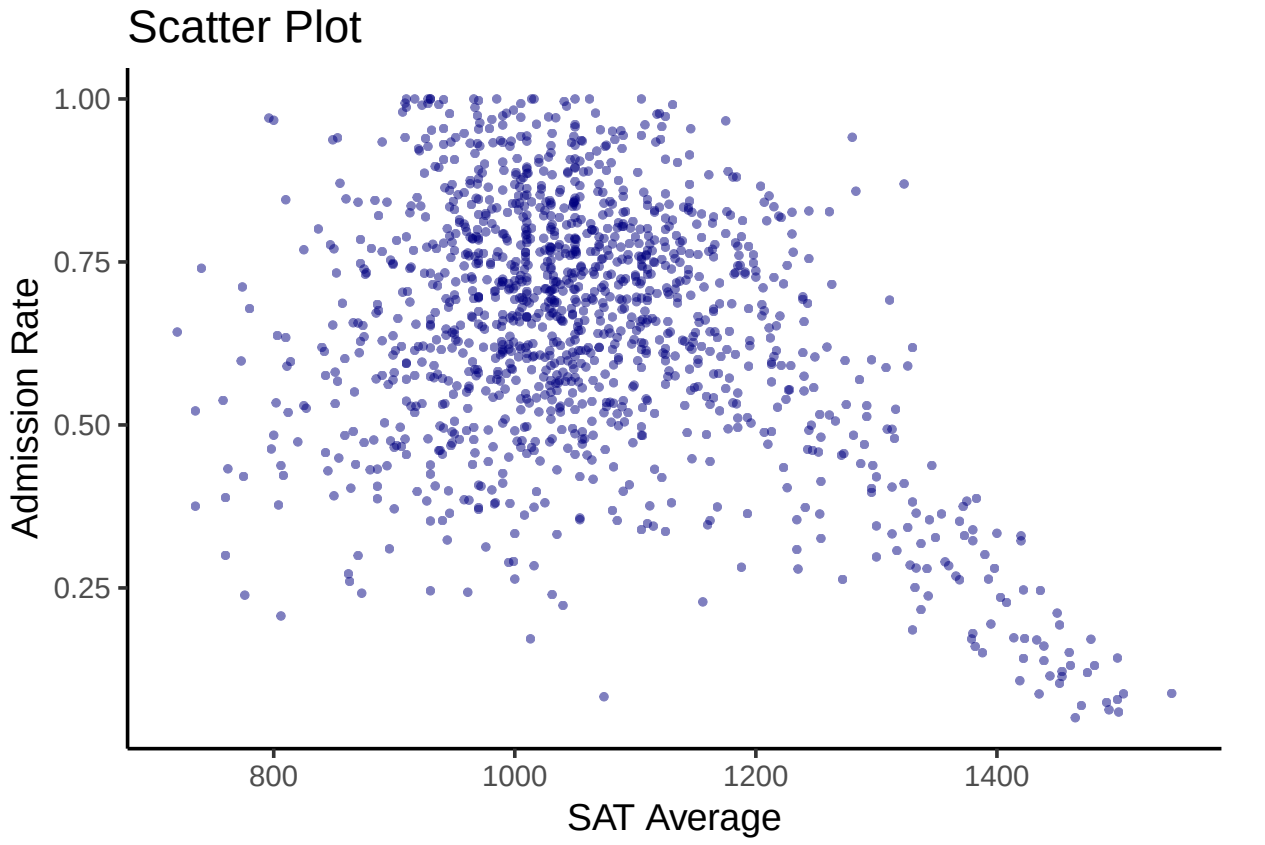
```
library(ggplot2)
library(gridExtra)

college <- readRDS('college.rds')
college_no_na <- college[with(college, (!is.na(sat_avg)) & (!is.na(admission_rate))), ]

plot_admi <- ggplot(college_no_na, aes(x = admission_rate)) +
  geom_histogram(fill = "lightblue", color = "black")

plot_sat <- ggplot(college_no_na, aes(x = sat_avg)) +
  geom_histogram(fill = "lightblue", color = "black")

grid.arrange(plot_admi, plot_sat, ncol = 2)
```



Scatter plot

We try scatter plot to explore the relationship between admission rate and SAT average. As we ccan see on the scatter plot below, there are mainly two main patterns. For those schools with SAT average below 1200, the admission rate seems not tight related with SAT average scores. That is, the SAT average is not the main factor when university enrolling students. On the other hand, for those university with SAT average higher than 1200, we could clear see a trend, The higher SAT average corresponds to low admission rate. However, we could only see the relationship, and we cannot infer the internal reason for that. Based on our daily experience, the reason for this pattern is that when a school has very high SAT score requirement, it will lead to lower admission rate. This explanation based on our experience could well count for this pattern.

```
college_no_na <- college[with(college, (!is.na(sat_avg)) & (!is.na(admission_rate))), ]
sat_avg <- college_no_na$sat_avg
admission_rate <- college_no_na$admission_rate
baseplt <- ggplot(college_no_na, aes(x = sat_avg, y = admission_rate))
baseplt + geom_point(alpha= 0.5, color = "navy", stroke = 0) +
  theme_classic(14) +
  labs(title = "Scatter Plot",
       x = "SAT Average",
       y = "Admission Rate")
```



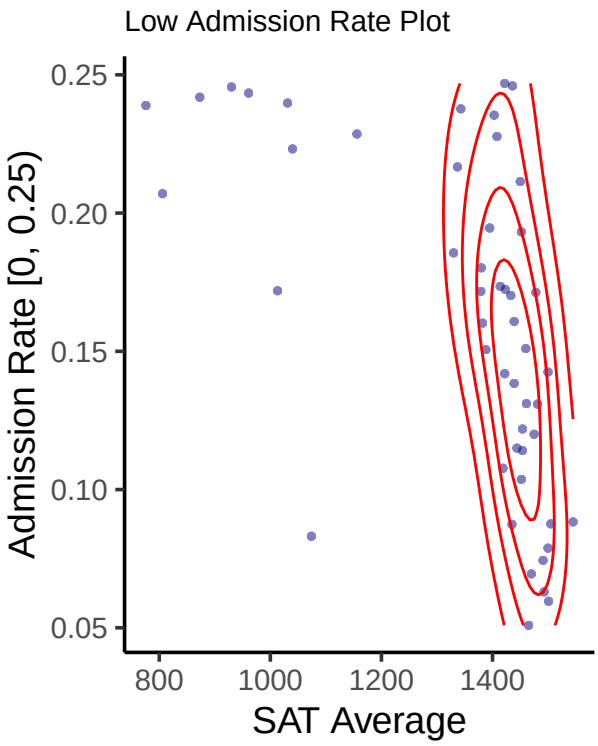
Low and Regular Admission Rate

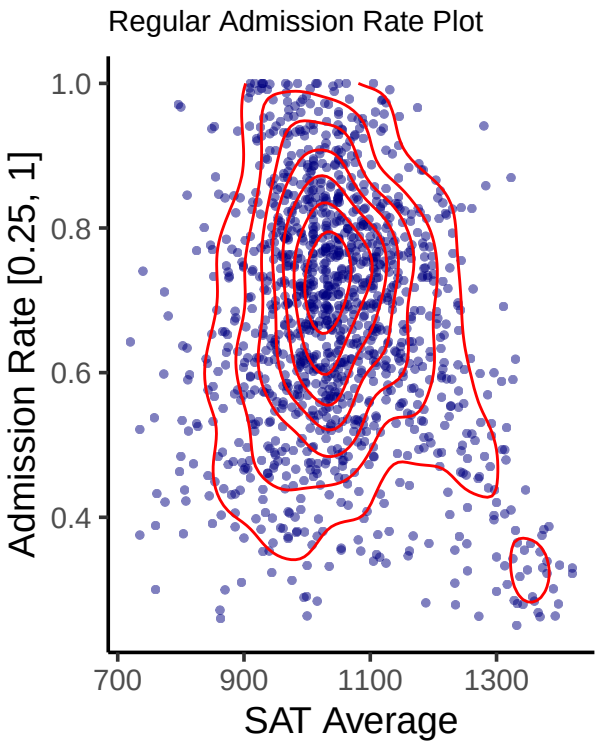
To follow on exploring the two patterns we found on the last graph, we plot the density contour lines based on the scatter plots.

From the left handside graph, we see there is a sharp decline pattern, that is, as SAR average become slighly higher, the admission rate immediately drops a lot. The reason for this may be that score higher than 1400 are very rare and there is not much space for a obvious higher score, but there is still space for lower admission rate.

Meanwhile, we see some university with relatively lower SAT average, but they still have very low admission rate. It is possibly based on the fact that SAT score is only one single requirement for universities to enroll good students. These universities on the left part of the plot may have more expectation on students with internship experience, sports talents, etc. Or, we can guess that students with very high SAT scores are not willing to go to these universities due to some reasons and thereby the SAT average becomes relatively lower. Those two reasons discussed above could explain that they have low admission rate even though they don't have very high requirement on SAT scores.

```
college_no_na_low <- college[with(college, (!is.na(sat_avg)) &
                                  (!is.na(admission_rate)) & (admission_rate < 0.25)), ]
college_no_na_high <- college[with(college, (!is.na(sat_avg)) &
                                  (!is.na(admission_rate)) & (admission_rate >= 0.25)), ]
sat_avg <- college_no_na$sat_avg
admission_rate <- college_no_na$admission_rate

plot_low <- ggplot(college_no_na_low, aes(x = sat_avg, y = admission_rate))+
  geom_point(alpha= 0.5, color = "navy", stroke = 0) +
  theme_classic(14) + theme(plot.title = element_text(size = rel(0.8))) +
  geom_density_2d(color = "red", bins = 5) +
  labs(title = "Low Admission Rate Plot",
       x = "SAT Average",
       y = "Admission Rate [0, 0.25)")

plot_high <- ggplot(college_no_na_high, aes(x = sat_avg, y = admission_rate))+
  geom_point(alpha= 0.5, color = "navy", stroke = 0) +
  theme_classic(14) + theme(plot.title = element_text(size = rel(0.8))) +
  geom_density_2d(color = "red", bins = 8) +
  labs(title = "Regular Admission Rate Plot",
       x = "SAT Average",
       y = "Admission Rate [0.25, 1]")

grid.arrange(plot_low, plot_high, ncol = 2)
```


Low Admission Rate Plot
Admission Rate [0, 0.25)
0.25
0.20
0.15
0.10
0.05
800
1000
1200
1400
SAT Average


Regular Admission Rate Plot
Admission Rate [0.25, 1]
1.0
0.8
0.6
0.4
700
900
1100
1300
SAT Average