Final Project

Ji Won Mok

2022-05-09

```
college_first6rows <- head(college)
```

[Introduction] ## Question of Interest This report will investigate on the relationship between SAT average score in a college and 3-years replayment rate of the loans. It is interesting to predict how many people can figure out how SAT score can be relevant to graduatation rate. I will use SAT_AVG(SAT average score) and COMPL_RPY_3YR_RT(3 years replayment rate). A linear model will be used to see the relationship between two variables.

[Preprocessing] ## Preprocessing i.Narrow down to 3 columns: SAT_AVG,LOCALE,COMPL_RPY_3YR_RT. They are renamed to user-friendly names.

```
college_reduced <- college %>% select(SAT_AVG, LOCALE,COMPL_RPY_3YR_RT) %>%
                               rename(Average_SAT_Score = SAT_AVG,
                                      Repayment_Rate = COMPL_RPY_3YR_RT,
                                      LOCATION = LOCALE)
```

ii. Use mutate to recode the location categorizing (‘11’,‘12’,‘13’) to “city”,(‘21’,‘22’,‘23’) to “suburb”,(‘31’,‘32’,‘33’) to “town”,(‘41’,‘42’,‘43’) to “rural”.

```
college_reduced_location <- college_reduced %>%
  mutate( recoded_location = recode ( LOCATION,
                                      '11' = "city",
                                      '12' = "city",
                                      '13' = "city",
                                      '21'= "suburb",
                                      '22'= "suburb",
                                      '23' = "suburb",
                                      '31'= "town",
                                      '32'= "town",
                                      '33' = "town",
                                      '41'= "rural",
                                      '42'= "rural",
                                      '43' = "rural"))
```

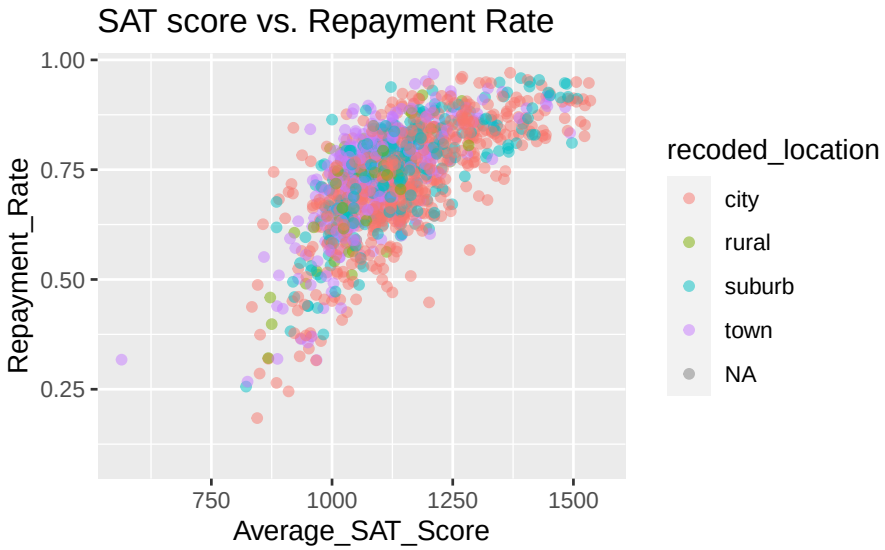
```
## Warning: Unreplaced values treated as NA as `.x` is not compatible.
## Please specify replacements exhaustively or supply `.default`.
```

[Visualization section] ## Visualization i. Scatter plot is used to see the relationship between average sat score of a students in colleges and 3 years repayment rate of the loan. Overall, it shows

the positive correlation between two variables, repayment rate and SAT score. The distribution of recorded_location(city, rural, suburb, town) also show the strong positive correlation as well.

```
college_reduced_location  %>%
  ggplot() +
  geom_point(mapping = aes( x = Average_SAT_Score,
                            y = Repayment_Rate,
                            color = recoded_location),
                            alpha = 0.5) +
  labs(title = "SAT score vs. Repayment Rate")
```

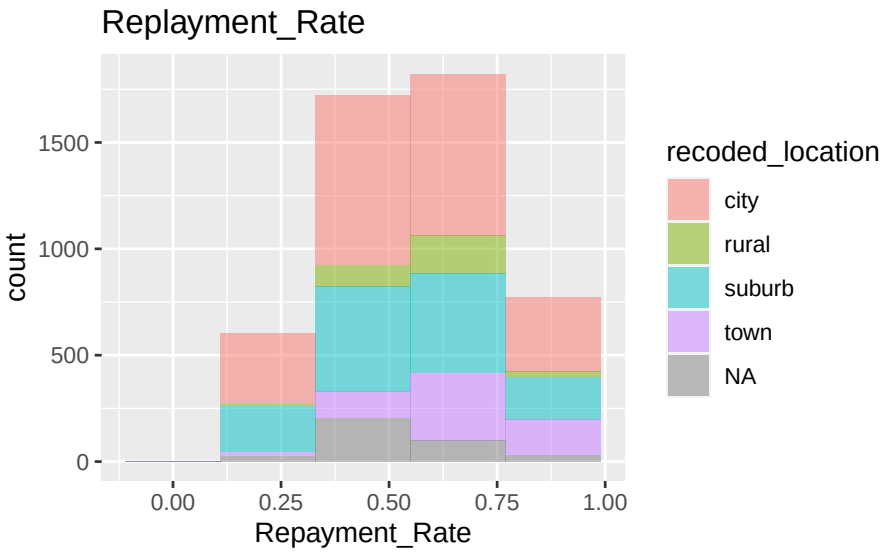
```
## Warning: Removed 5803 rows containing missing values (geom_point).
```



ii. I chose histogram since distribution is well organized in histogram plot. It is normally distributed since the center is 0.5 and distribution or replayment rate in recoded_location show the simliar distribution in the histogram. It proves that repayment rate and location are not related.

```
# 1. The relationship between repayment rate of education loans and cost of attendance
college_reduced_location  %>%
  ggplot() +
  geom_histogram(mapping = aes(x=Repayment_Rate, fill = recoded_location),
                 bins =5,
                 alpha = 0.5) +
  labs(title = "Replayment_Rate")
```

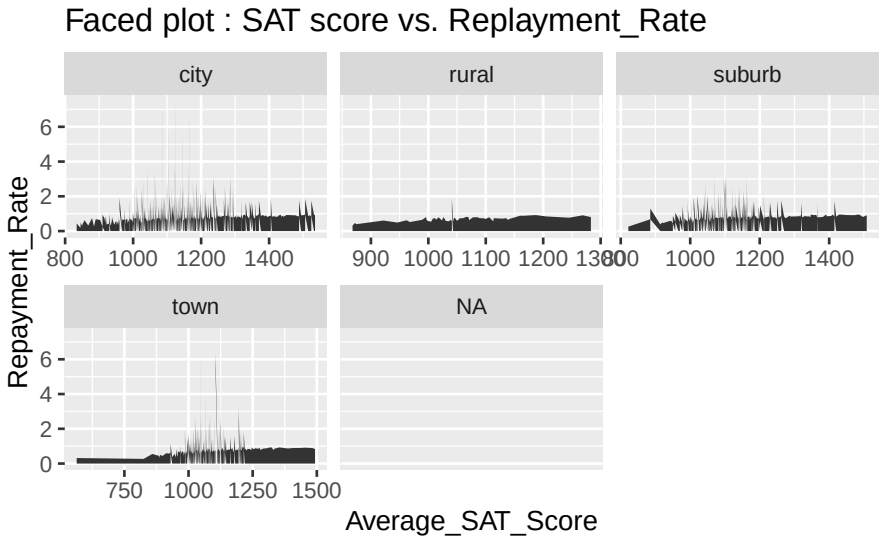
```
## Warning: Removed 2134 rows containing non-finite values (stat_bin).
```



iii. The distribution of the recoded location variable does not make difference depending on average sat score from 1000 to 1250. In city, suburb, and town, repayment rate is higher in that range. Moreover, if you see the rural area, it does not make any difference, and repayment rate is relatively lower than other type of recoded-location.

```
college_reduced_location  %>%
  ggplot() +
  geom_area(mapping = aes(x=Average_SAT_Score, y = Repayment_Rate)) +
  facet_wrap( ~ recoded_location, scales = "free_x")+
  labs(title = "Faced plot : SAT score vs. Replayment_Rate")
```

```
## Warning: Removed 5803 rows containing missing values (position_stack).
```



Summary Statistics

It shows the medan, median, standard deviation, minimum, and maximum value of average sat score. Specifically, mean is 1131.28, median is 1116, and standard deviation is 129.6887. The gap between minimum and max is pretty big as 564, 1558 while the median and mean are simliar.

```
college_reduced_location  %>%
  summarize(mean = mean(Average_SAT_Score, na.rm = TRUE),
            median = median(Average_SAT_Score, na.rm = TRUE),
            sd = sd(Average_SAT_Score, na.rm = TRUE),
            min = min(Average_SAT_Score, na.rm = TRUE),
            max(Average_SAT_Score, na.rm = TRUE)
            )
```

mean	median	sd	min	max(Average_SAT_Score, na.rm = TRUE)
1131.28	1116	129.6887	564	1558

It shows the medan, median, standard deviation, minimum, and maximum value of repayment rate. Specifically, mean and median are simliar as 0.56. However, surprisingly,minimum and max have huge gap, 0.09, and 0.97.

```
college_reduced_location  %>%
  summarize(mean = mean(Repayment_Rate, na.rm = TRUE),
            median = median(Repayment_Rate, na.rm = TRUE),
            sd = sd(Repayment_Rate, na.rm = TRUE),
            min = min(Repayment_Rate, na.rm = TRUE),
            max(Repayment_Rate, na.rm = TRUE)
            )
```

mean	median	sd	min	max(Repayment_Rate, na.rm = TRUE)
0.5615238	0.5642658	0.1824644	0.0905172	0.9706458

Data Analysis

Linear modelling and I use tidy to show estimate, standard deviation error, statistics, and p.value. The cofficient is -0.001

```
continuous_model <- lm(Repayment_Rate ~ Average_SAT_Score, data = college_reduced_location)
tidy(continuous_model)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0014276	0.0234947	-0.0607633	0.9515574
Average_SAT_Score	0.0006503	0.0000207	31.4298131	0.0000000

R.squared parameter manifests that it is closer to 0 more than 1 since it is 0.44. It shows that it has low variability in the repayment rate.

```
continuous_model %>%
  glance() %>%
  select(r.squared)
```

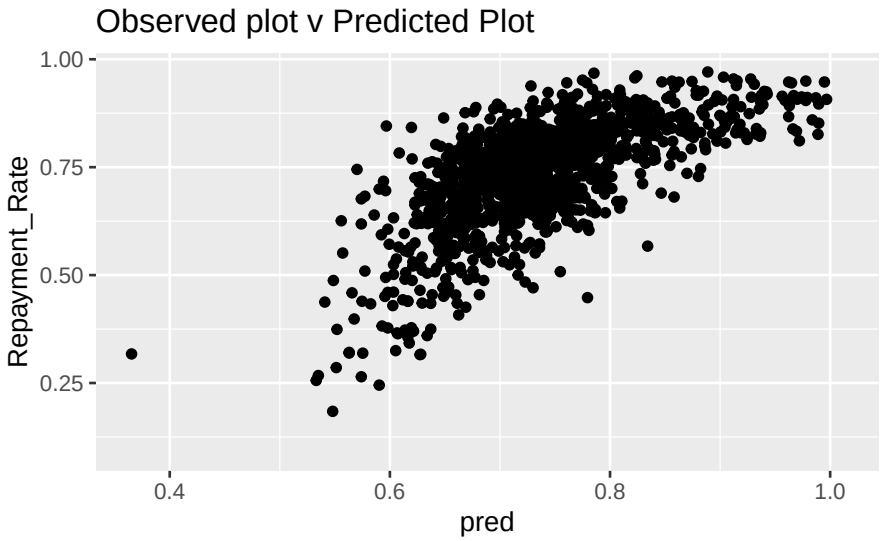
r.squared
0.440833

```
continuous_df <- college_reduced_location %>%
  add_predictions(continuous_model) %>%
  add_residuals(continuous_model)
```

This plot is "observed plot vs predicted plot", and it shows the positive correlation between prediction_model and repayment_rate.

```
continuous_df %>%
  ggplot()+
  geom_point(mapping = aes(x=pred, y= Repayment_Rate)) +
  labs(title = "Observed plot v Predicted Plot")
```

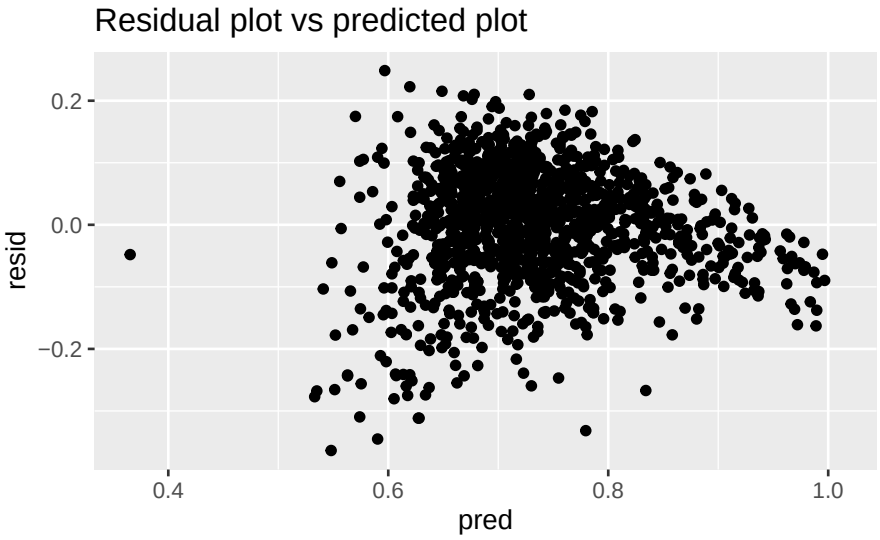
```
## Warning: Removed 5803 rows containing missing values (geom_point).
```



This plot is "Residual plot vs predicted plot". It can be said that it is not consistent since it is highly populated between 0.6 and 0.8.

```
continuous_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  labs(title = "Residual plot vs predicted plot")
```

```
## Warning: Removed 5803 rows containing missing values (geom_point).
```

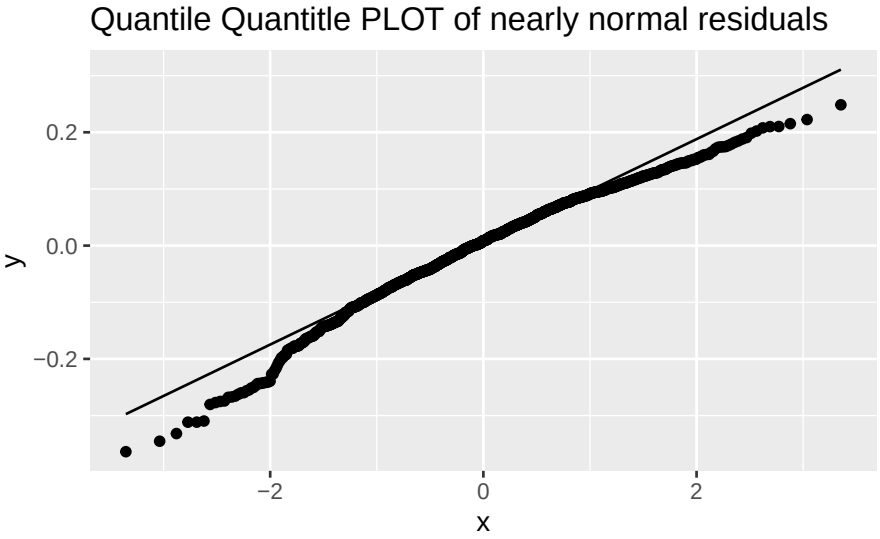


This QQplot shows that residuals are normal. It shows that it is mostly alinged with the line except some parts below -2 and after 2.

```
continuous_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid))+
  labs(title = "Quantile Quantitle PLOT of nearly normal residuals")
```

```
## Warning: Removed 5803 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 5803 rows containing non-finite values (stat_qq_line).
```



Conclusion

The covariation between the SAT score and repayment rate is highly related. Refer to the mean of two variables, SAT score higher 1131.28 can be considered as students with greater score than the average, and repayment rate above 0.56 is also higher than the average repayment rate. In data

anaylsis, r.squared was closer to 0, and it shows that it has low variability of repayment rate variable. Moreover, 3 linear modelling sohws that model's result is pretty consistent except the plot,'Residual plot vs predicted plot'. It violates the constant assumption. To sum up, SAT score and repayment rate has strong positive correlation. On top of that, the location is not a huge determinent variable except the fact that suburb shows exceptional case that it overall has low repayment rate regardless of SAT score.