SATS project report

Tingyu Qian

2023-09-28

Introduction

Statistics is a fundamental and omnipresent component of modern life, influencing decision-making processes in diverse fields, from healthcare to economics, and from politics to education. Yet, for many individuals, statistics can be an intimidating and perplexing subject. This is where the Survey of Attitudes Towards Statistics, or SATS, comes into play.

SATS is a meticulously crafted instrument designed to delve into the intricate realm of human attitudes and perceptions regarding statistics. It is an invaluable tool used in the realms of social sciences, education, and research to gauge how individuals from all walks of life relate to and understand statistical concepts. By examining these attitudes and beliefs, SATS provides critical insights into the factors that may hinder or facilitate one's ability to comprehend and utilize statistics effectively.

This survey is not merely an academic exercise; it has profound implications for the way statistics is taught and applied. In education, understanding students' attitudes towards statistics can inform pedagogical approaches, helping educators tailor their teaching methods to promote a more positive learning experience. Moreover, in practical contexts, such as market research or public policy analysis, an appreciation of how individuals perceive statistical information can enhance the effectiveness of communication and decision-making processes.

In the dataset "sats_sci", there are 3730 observations, and 190 variables. "sats_sci" dataset contains SATS course data, SATS instructor data and SATS student data. SATS student dataset not only contains information such as age, major, etc., but also contains a survey called SATS_36. The SATS-36 comprise six subscales, Affect (6 items), Cognitive Competence (6items), Value (6 items), Difficulty (7 items), Interest (4 items) and Effort (4 items). SATS_36 has two versions, pre-course version and a post-course version. Each subscales contain some questions that can get to know students accurately about this subscales. However, we do not care about Effort (4 items).

The Affect subscale measures positive and negative feelings toward statistics.

The Cognitive Competence subscale measures participants' attitudes regarding their perception of their ability to mentally comprehend statistics.

The Value subscale measures participants' perceptions of the usefulness and worth of statistics.

The Difficulty (perceived easiness) subscale is measured by items which collectively asked about participants' attitudes regarding how difficult statistics is/was.

The Interest subscale measured how much interest a participant has in statistics.

For instructor data that contained in dataset "sats_sci", the survey filled out annually. For course data that contained in dataset "sats_sci", the survey filled out for each course.

Data Cleaning

In this report, we will only concerned about attitude components' relationship with grades, so we can extract the column that are interested in first.

Before we do data cleaning, we can first look at how many rows there are in total that contain missing values.

```
## Number of rows with missing values:  1661
```

We can find that there are a total of 1661 rows with missing values, but since we only have a total of 3730 observations, if we remove all rows that contain missing values, then we will have one third less data, which has a negative impact on our data. Therefore, in order to solve this problem, we can first remove the rows that have missing value in either all of 5 pre-course attitude components or all of 5 post-course attitude components. This way we can not only retain our data, but also allow us to do a perfect data cleaning.

```
#pre-course attitude components
columns_to_check <- c("Pre.Affect", "Pre.CogComp", "Pre.Value", "Pre.Diff", "Pre.Interest")
na_count <- rowSums(is.na(sats_sci_use[columns_to_check]))
threshold <- 5
sats_sci_clean <- sats_sci_use[na_count < threshold, ]

#post-course attitude components
columns_to_check <- c("Post.Affect", "Post.CogComp", "Post.Value", "Post.Diff", "Post.Interest")
na_count <- rowSums(is.na(sats_sci_clean[columns_to_check]))
threshold <- 5
sats_sci_clean <- sats_sci_clean[na_count < threshold, ]
```

After we remove the rows that have missing value in either all of 5 pre-course attitude components or all of 5 post-course attitude components, because there are still individual missing values in pre-course attitude components or post-course attitude components, we can fill in the missing values based on the value of other pre-course attitude components or post-course attitude components. Because components of Affect, CogComp, Value, Diff, Interest, the larger the number, the better the student feels about statistics. In the component of Difficulty, the large the number, the simpler the student thinks statistics is. In other words, the better this student learn statistics, so we can use these data to help fill in missing values.

```
imp_model <- mice(data = sats_sci_clean[, c("Pre.Affect", "Pre.CogComp", "Pre.Value", "Pre.Diff", "Pre.
                  method = "pmm",
                  m = 5,
                  maxit = 5,
                  seed = 123)

# Impute missing values
imputed_data <- complete(imp_model)

# Specify the columns to update in the original dataset
columns_to_update <- c("Pre.Affect", "Pre.CogComp", "Pre.Value", "Pre.Diff", "Pre.Interest")

# Update the original dataset with the imputed values
sats_sci_clean[columns_to_update] <- imputed_data[columns_to_update]

imp_model <- mice(data = sats_sci_clean[, c("Post.Affect", "Post.CogComp", "Post.Value", "Post.Diff", "
                  method = "pmm",
                  m = 5,
                  maxit = 5,
                  seed = 123)

# Impute missing values
imputed_data <- complete(imp_model)

# Specify the columns to update in the original dataset
```

```r
columns_to_update <- c("Post.Affect", "Post.CogComp", "Post.Value", "Post.Diff", "Post.Interest")

# Update the original dataset with the imputed values
sats_sci_clean[columns_to_update] <- imputed_data[columns_to_update]
```

The method I used above is called Multivariate Imputation. Multivariate Imputation is a powerful statistical technique used to handle missing data in datasets. Unlike univariate imputation methods that impute missing values one variable at a time, Multivariate Imputation considers the relationships between variables and imputes missing values for multiple variables simultaneously. Predictive Mean Matching (PMM) is a robust imputation method used in the context of multiple imputation to address missing data in statistical analysis. PMM is designed to make imputations by predicting missing values from observed data and matching them to similar observed values. PMM imputes missing data through a series of steps. It constructs a predictive model based on the available data and generates predicted values for the missing observations. These predictions are matched to similar observed values from the dataset. PMM doesn't yield a single imputed value but generates multiple imputed datasets. Each dataset incorporates different imputed values for missing data points. This multiple imputation approach accounts for the inherent uncertainty in imputation. PMM respects the distribution and relationships within the observed data, resulting in more accurate imputations that retain the structure of the original data.

After we fill the missing values in pre-course version and post-course version of attitude constructs, let us check if other columns have missing value.

```
## Column ' Pre.exgrade ' has  9  missing values.
## Column ' major ' has  21  missing values.
## Column ' gpa ' has  126  missing values.
## Column ' Post.exgrade ' has  10  missing values.
```

Based on other complete data in these columns, we can find that there is a lot of uncertainty between these data and other data. For example, some people expected that their grade will be B after finishing the course, but the student's true grade is A. Some students expected that their grade will be C, but the actual grade is indeed C. Because of these uncertainties, we cannot accurately estimate the true number of these missing values, so after knowing that there are not many missing values left, in order to ensure the accuracy of data analysis, I chose to delete the rows with these missing values.

```r
sats_sci_clean <- sats_sci_clean[!(sats_sci_clean$Grade.Letter %in% c("w", " ")), ]
sats_sci_final <- na.omit(sats_sci_clean)
```

The first 10 rows in the dataset sats_sci_final are shown below:

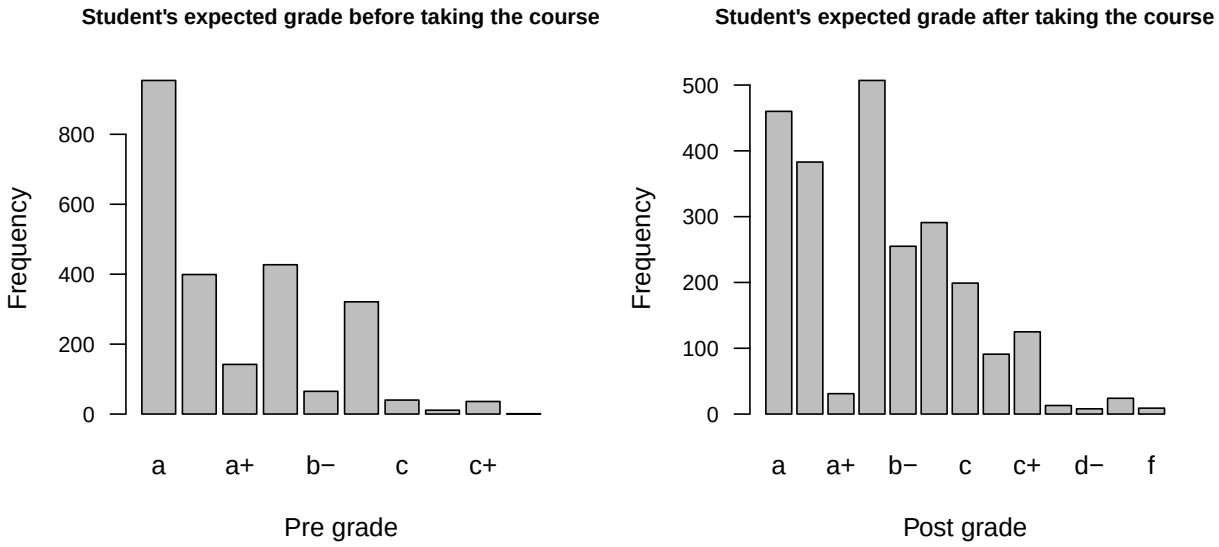
```r
#display the first 10 observations
display_first_10_observations <- function(data) {
  if (is.data.frame(data)) {
    head(data, 10)
  } else {
    cat("Input is not a data frame.\n")
  }
}

df <- sats_sci_final
display_first_10_observations(df)
```

```
##    StudentID CourseID InstructorID  Semester Pre.Consent Pre.Affect Pre.CogComp
```

```
## 1          1        1            1 Fall 2007           1   4.500000    4.166667
## 2          2        1            1 Fall 2007           1   5.166667    4.833333
## 3          3        1            1 Fall 2007           1   6.166667    6.500000
## 4          4        1            1 Fall 2007           1   5.000000    5.833333
## 5          5        1            1 Fall 2007           1   4.500000    6.333333
## 6          6        1            1 Fall 2007           1   5.833333    6.000000
## 7          7        1            1 Fall 2007           1   6.500000    6.666667
## 8          9        1            1 Fall 2007           1   5.333333    5.333333
## 9         10        1            1 Fall 2007           1   6.333333    6.166667
## 10        11        1            1 Fall 2007           1   6.166667    7.000000
##    Pre.Value Pre.Diff Pre.Interest Pre.exgrade major  gpa Post.Consent
## 1   4.888889 2.285714         5.75           2    12 0.00            1
## 2   5.888889 3.000000         7.00           2    12 3.21            1
## 3   5.444444 4.571429         5.25           4    12 0.00            1
## 4   6.444444 3.428571         6.75           4    12 3.00            1
## 5   5.555556 2.714286         6.50           2     6 0.00            1
## 6   5.666667 3.857143         6.00           4    12 0.00            1
## 7   5.111111 3.428571         5.50           4    12 3.80            1
## 8   6.000000 3.857143         6.00           4    12 3.20            1
## 9   6.333333 2.714286         6.75           2    12 0.00            1
## 10  5.222222 5.000000         6.25           2    12 0.00            1
##    Post.Affect Post.CogComp Post.Value Post.Diff Post.Interest Post.exgrade
## 1     5.666667     5.000000   5.111111  4.000000          6.00            5
## 2     4.500000     5.333333   6.888889  2.714286          6.75            4
## 3     6.333333     6.833333   5.222222  4.857143          4.25            5
## 4     5.666667     6.333333   6.777778  3.000000          7.00            5
## 5     5.500000     5.833333   6.666667  2.714286          6.50            3
## 6     6.500000     6.333333   6.444444  3.571429          6.00            4
## 7     6.166667     6.500000   6.555556  2.714286          6.00            5
## 8     6.833333     6.500000   6.777778  3.714286          7.00            4
## 9     6.000000     6.666667   6.888889  2.571429          6.50            3
## 10    5.166667     6.333333   5.888889  4.000000          5.50            5
##    Grade.Letter
## 1            b-
## 2            a-
## 3            c+
## 4             b
## 5            a-
## 6             b
## 7            b-
## 8            b+
## 9             a
## 10            b
```

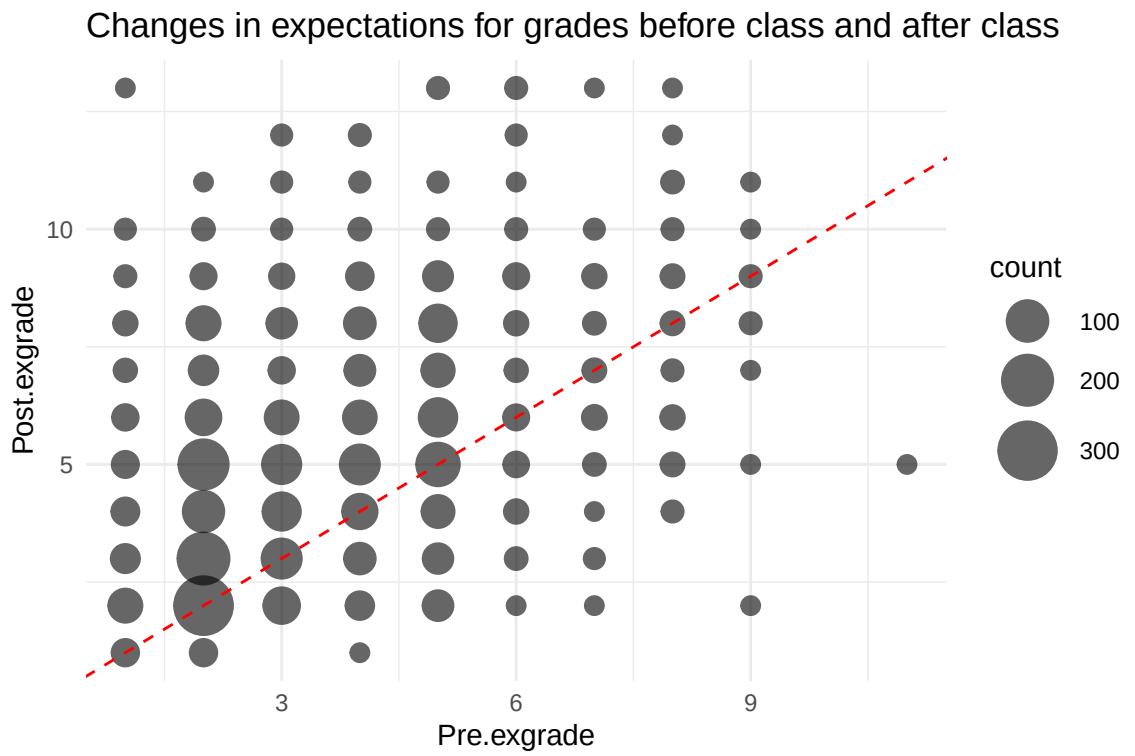
After we finish the data cleaning, we can do the data analysis next. Let's first take a look at the distribution chart (histogram) of student scores (Student's expected grade before taking course and Student's expected grade after taking course). Because when doing surveys, students use numbers to represent letter grades, so in order to facilitate viewing, we need to express the numbers with corresponding letters first.

The accuracy of students predicting their own grade.



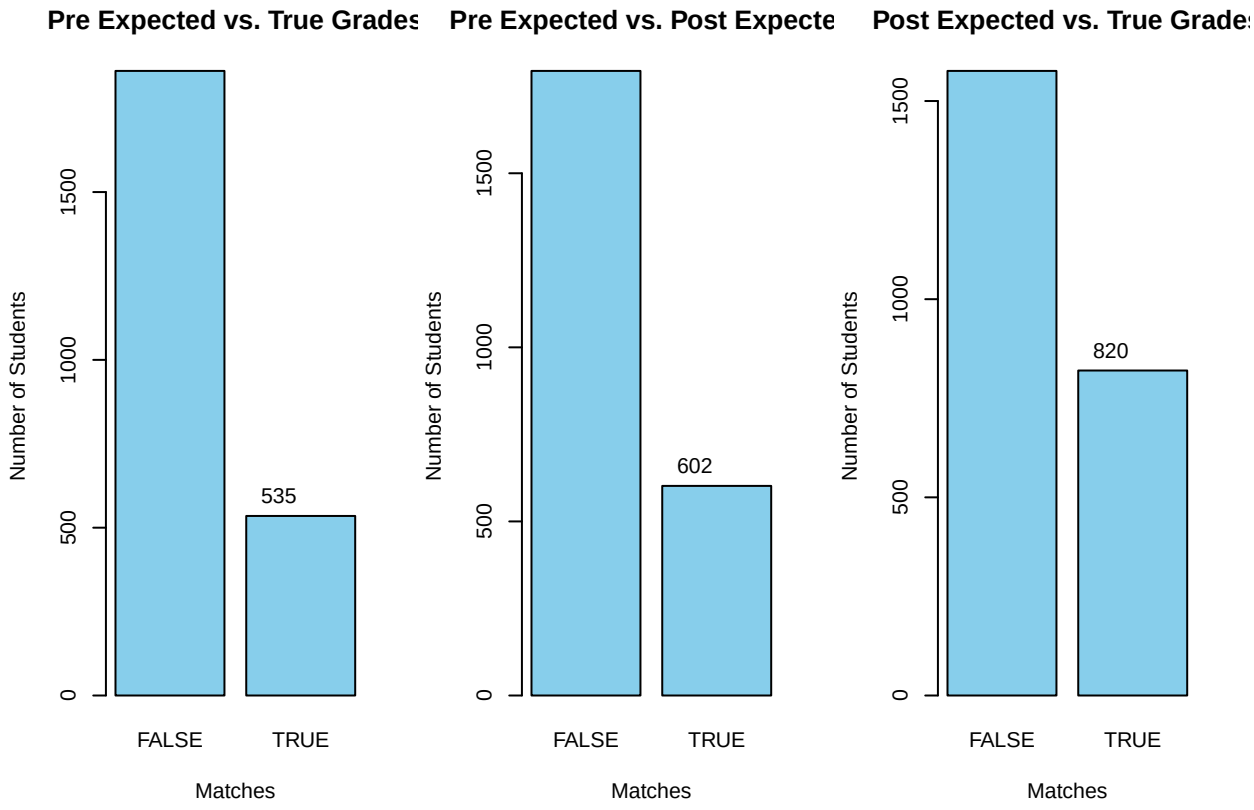
The plots above are the bar plot of the Student's expected grade before taking the course and the bar plot of Student's expected grade after taking the course. A bar plot, also known as a bar chart or bar graph, is a widely used data visualization tool that presents categorical data in the form of rectangular bars. Each bar represents a specific category, and the length or height of the bar corresponds to the quantity or value associated with that category. One of the primary advantages of using bar plots is their simplicity and effectiveness in conveying information. They provide a straightforward and intuitive way to compare different categories or groups, making it easy to identify trends, patterns, and variations in the data. Bar plots are especially useful for displaying discrete data, such as survey results or any data that can be grouped into distinct categories. Their visual appeal and ability to provide a quick, at-a-glance summary of data. Therefore, the x-axis of the bar plot above is the student's expected grade in different categories (A, A+, A-,. . .), and the y-axis of the bar plot above is the frequency of a certain grade category appears in data. However, we have less students to do the survey after taking the class than students before taking the class. Therefore, we can only conclude from these two plots that after finishing the class, d-, d+ and f appeared in the expected grades. In order to better see the changes in students' expectations for grades before class and after class, we need to use scatter plots.

Student's expected grade (1 = A+, 2 = A, 3 = A-, 4 = B+, 5 = B, 6 = B-, 7 = C+, 8 = C, 9 = C-, 10 = D+, 11 = D, 12 = D-, 13 = F)



The red diagonal line in the scatter plot represents a 1:1 relationship between the x and y variables. This means that any point that falls directly on this line has the same value for both the x and y axes. Therefore, in the scatter plot, this line is a visual aid that helps us quickly identify points where the pre-course and post-course expected grades are the same. The scatterplot above use the size of the dots to represent the sample size on each dot. We can clearly see from the scatter plot above that the number of dots in the upper left corner is greater than the number of dots in the lower right corner. We also know that the smaller the number, the better the score, so we can know from the scatter plot above that most students have better expectations for their grades after taking the class than they did before taking the class.

Let's first look at how many students' expected grades (both of pre course version and post course version) were the same as their actual grades, and how many students change their expected grade after taking the course.



From the three bar plots above, we can know that there are 2396 observations in total, there were 568 students whose expected scores before class were consistent with their actual scores. After class, only 602 students did not change their expectations of their scores. From the last plots, we can know that the expected scores of 861 students after finishing the class are consistent with their actual scores. Therefore, we can draw a data conclusion that the accuracy of students' expected grades before class is `568/2396 = 23.706%`, and the accuracy of students' expected grades after class is `861/2396 = 35.93%`. Therefore, we can conclude that after taking the course, students can understand the difficulty of the course and their adaptability to the course, so they can have a more accurate expectation of their grades.

Five attitude components' (Affect, Cognitive Competence, Value, Difficulty, Interest) relationship with grades.

Affect

There are 6 items for students to answer in order to know students' feelings concerning statistics accurately. Each question has a scale from 1 to 7 (Strongly Disagree to Strongly Agree), questions with stars means reverse coded. However, we do not concerned about the individual items and their structure with the constructs, so what we do is to use a "summary average score" to do the analyse. In the dataset, we have a column called Pre.Affect, the numbers in this column are the average of students' answers to each question below, meaning the larger the number, the better students' feelings concerning statistics. The questions are the following:

3. I will like statistics.

4.* I will feel insecure when I have to do statistics problems.

15.* I will get frustrated going over statistics tests in class.

18.* I will be under stress during statistics class.

19. I will enjoy taking statistics courses.

28.* I am scared by statistics.

Cognitive Competence

There are 6 items for students to answer in order to know students' attitudes about their intellectual knowledge and skills when applied to statistics accurately. Each question has a scale from 1 to 7 (Strongly Disagree to Strongly Agree), questions with stars means reverse coded. However, we do not concerned about the individual items and their structure with the constructs, so what we do is to use a "summary average score" to do the analyse. In the dataset, we have a column called Pre.CogComp, the numbers in this column are the average of students' answers to each question below, meaning the larger the number, the better students' attitudes about their intellectual knowledge and skills when applied to statistics. The questions are the following:

5.* I will have trouble understanding statistics because of how I think.

11.* I will have no idea of what's going on in this statistics course.

26.* I will make a lot of math errors in statistics.

31. I can learn statistics.

32. I will understand statistics equations.

35.* I will find it difficult to understand statistical concepts.

Value

There are 9 items for students to answer in order to know students' attitudes about the usefulness, relevance, and worth of statistics in personal and professional life accurately. Each question has a scale from 1 to 7 (Strongly Disagree to Strongly Agree), questions with stars means reverse coded. However, we do not concerned about the individual items and their structure with the constructs, so what we do is to use a "summary average score" to do the analyse. In the dataset, we have a column called Pre.Value, the numbers in this column are the average of students' answers to each question below, meaning the larger the number, the better students' attitudes about the usefulness, relevance, and worth of statistics in personal and professional life. The questions are the following:

7.* Statistics is worthless.

9. Statistics should be a required part of my professional training.

10. Statistical skills will make me more employable.

13.* Statistics is not useful to the typical professional.

16.* Statistical thinking is not applicable in my life outside my job.

17. I use statistics in my everyday life.

21.* Statistics conclusions are rarely presented in everyday life.

25.* I will have no application for statistics in my profession.

33.* Statistics is irrelevant in my life.

Difficulty

There are 7 items for students to answer in order to know students' attitudes about the difficulty of statistics as a subject accurately. Each question has a scale from 1 to 7 (Strongly Disagree to Strongly Agree), questions with stars means reverse coded. However, we do not concerned about the individual items and their structure with the constructs, so what we do is to use a "summary average score" to do the analyse. In the dataset, we have a column called Pre.Diff, the numbers in this column are the average of students' answers to each question below, meaning the larger the number, the easier the students think statistic is. The questions are the following:

6. Statistics formulas are easy to understand.

8.* Statistics is a complicated subject.

22. Statistics is a subject quickly learned by most people.

24.* Learning statistics requires a great deal of discipline.

30.* Statistics involves massive computations.

34.* Statistics is highly technical.

36.* Most people have to learn a new way of thinking to do statistics.

Interest

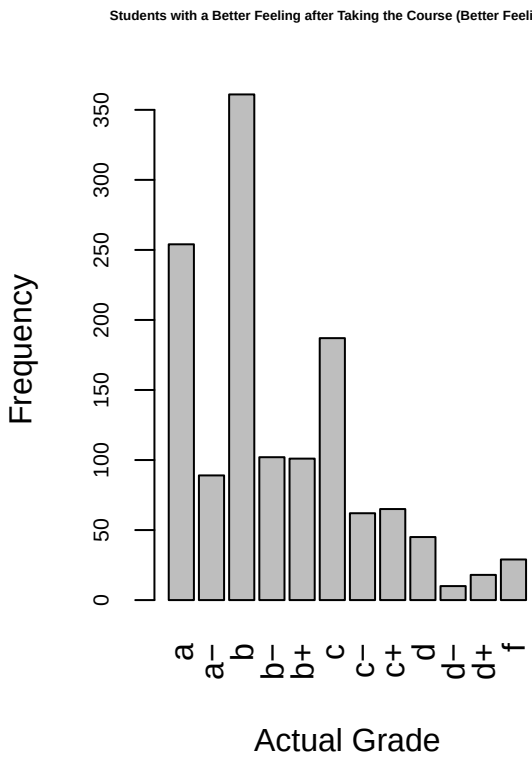
There are 4 items for students to answer in order to know students' level of individual interest in statistics accurately. Each question has a scale from 1 to 7 (Strongly Disagree to Strongly Agree), questions with stars means reverse coded. However, we do not concerned about the individual items and their structure with the constructs, so what we do is to use a "summary average score" to do the analyse. In the dataset, we have a column called Pre.Interest, the numbers in this column are the average of students' answers to each question below, meaning the larger the number, the higher students' level of individual interest in statistics. The questions are the following:

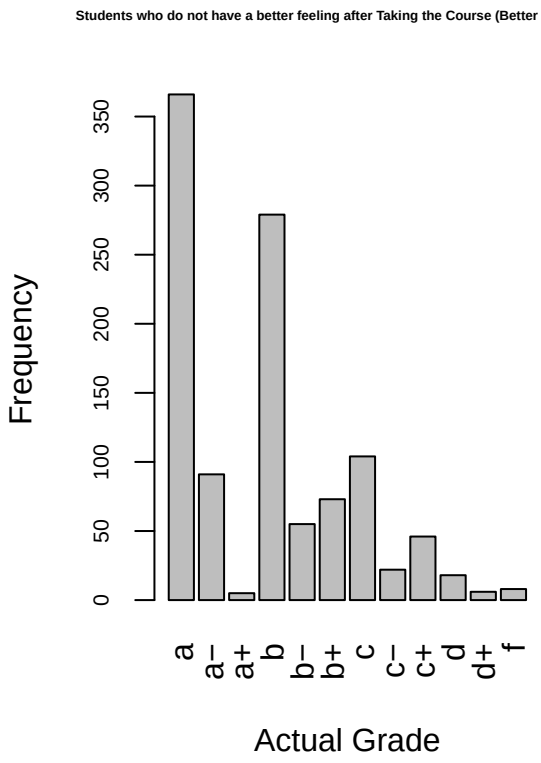
12. I am interested in being able to communicate statistical information to others.
13. I am interested in using statistics.
14. I am interested in understanding statistical information.
15. I am interested in learning statistics.

Due to the positive relationship between components of Affect, Cognitive Competence, Value, Interest, we can use 3.5 as a dividing line. Greater than 3.5 means that students feeling good about statistics overall, and less than 3.5 means that students feeling not good about statistics overall. Therefore, we can calculate the average of these 4 components and use 3.5 as a dividing line. To better analyze the accuracy of students' predicted grades, we can use grades with numbers and calculate the difference.

We can firstly see how many students have a better feeling about statistics overall after taking the course than before taking the course.

```
## There are 1323 students have a better feeling about statistics overall after taking the course than
## before taking the course.
```



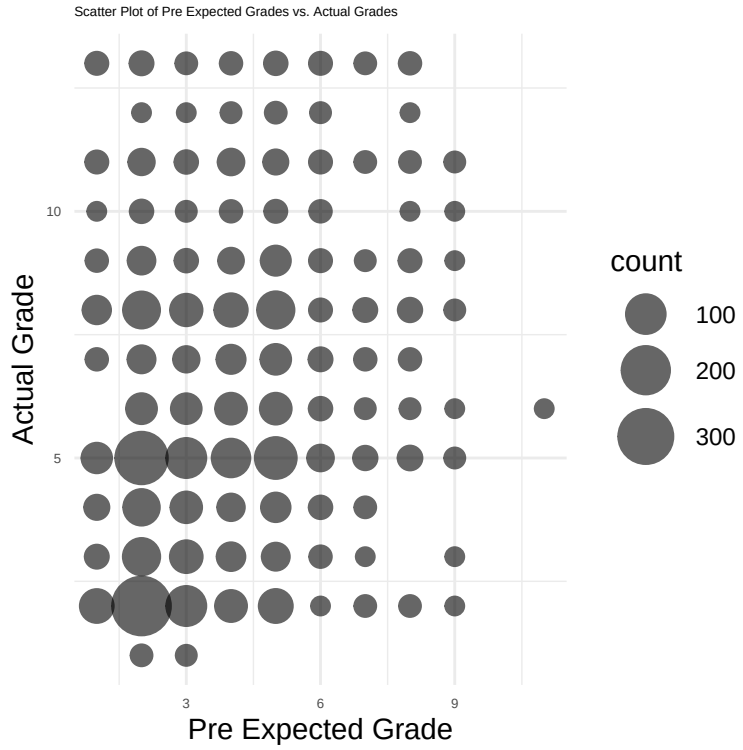


The 2 plots above are distribution of actual grade of students who have a better feeling after taking the course and distribution of actual grade of students who do not have a better feeling after taking the course.

We can look at the first plot and the second plot together. From the plots, we can find that students who have a poor overall feeling about statistics have better statistics scores than students who have a good overall feeling about statistics. From the first plot, we can see that about 250 students who have a good overall feeling about statistics got an A, but 350 students who did not have a good overall feeling about statistics got an A. There are more students who get F than students who don't feel good about statistics overall. What is even more surprising is that none of the students who had a good overall feeling about statistics got an A+, but about 10 students who had a poor overall feeling about statistics got an A+.

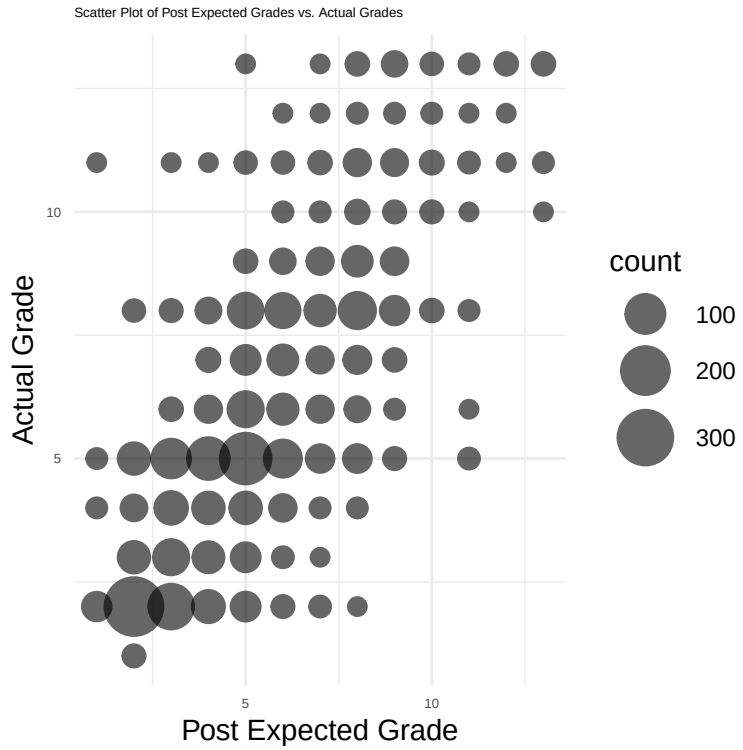
Everyone thinks that students who have a good overall feeling about statistics mean they are interested in statistics, and therefore can achieve better results. However, through data analysis, we can know that the real grades of students who do not have a better feeling about statistics after taking the class will be higher than those of students who have a better feeling about statistics after taking the class. [Because there are 1323 students have a better feeling about statistics overall after taking the course than before taking the course, and 1073 students do not have a better feeling about statistics overall after taking the course than before taking the course, which means we have a almost same sample size.]

Correlation between Actural grades and Pre expected grade




```
## Correlation coefficient: 0.3182277
```

Correlation between Actural grades and Post expected grade

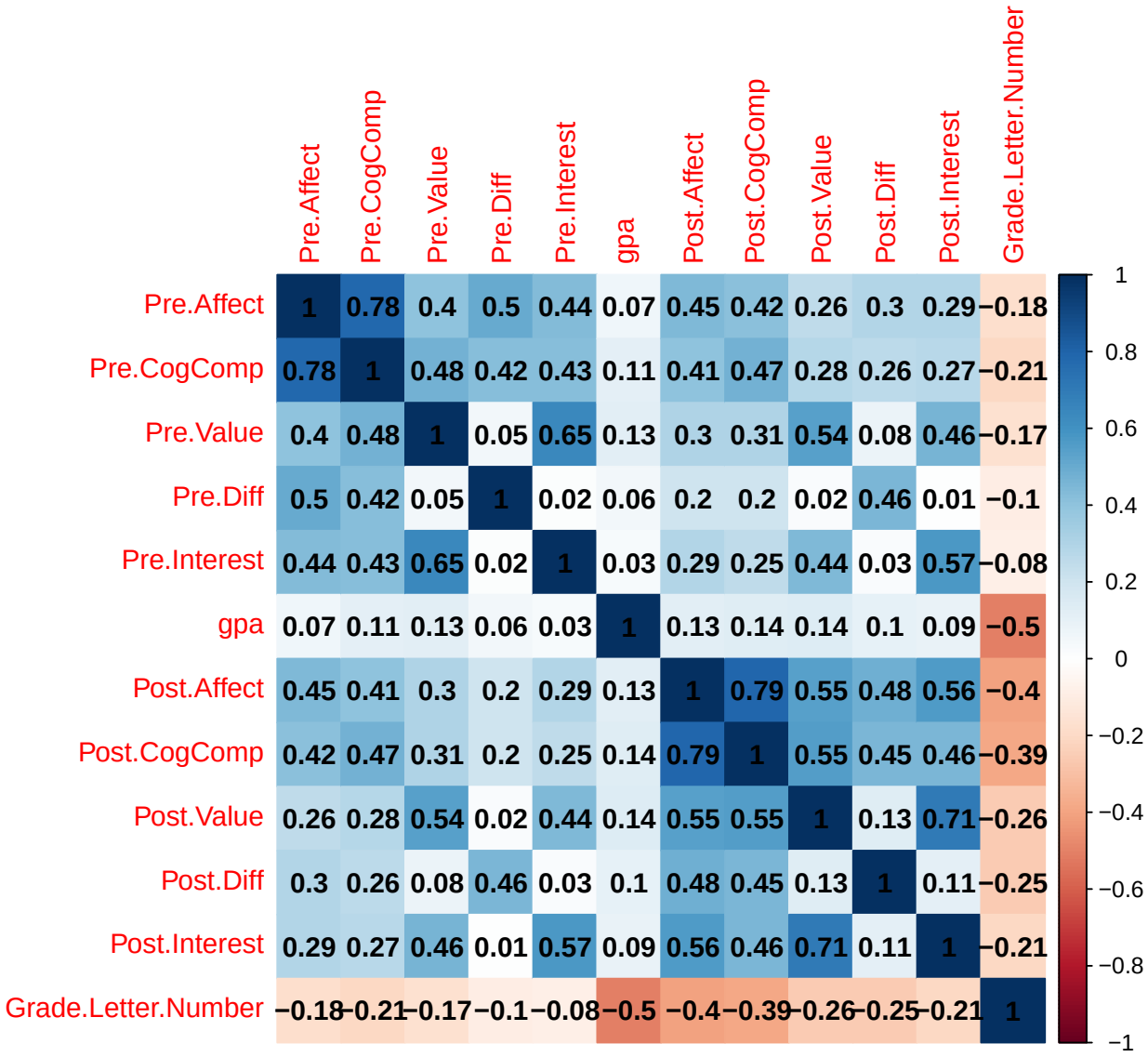


```
## Correlation coefficient: 0.7946889
```

Based on the two scatter plots above, we can know that the correlation coefficient between pre-course expected grades and real grades is 0.3182277, and the correlation coefficient between post-course expected grades and real grades is 0.7946889, that is to say before class, there is a positive but relatively weak relationship between expected grades and real grades. After class, there is a strong and positive relationship between expected grades and real grades. This suggests that students' expectations are a better predictor of their actual grades after the class than they are before the class , where the relationship is weaker.

To determine marginal relationships between variables and true grade, we chose to use the Spearman rank correlation coefficient. The calculation principle of Spearman rank correlation coefficient is as follows: 1. Rank the values of each variable separately, from lowest to highest, assigning them a rank. If there are ties (i.e., multiple values with the same value), assign the average rank to all tied values. 2. Calculate the difference between the ranks for each pair of data points for both variables. 3. Square these differences and calculate the sum of the squared differences. 4. Use the formula for Spearman's rank correlation coefficient: $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$ Where: ρ is Spearman's rank correlation coefficient. $\sum d^2$ is the sum of the squared rank differences. n is the number of data points.

Spearman's rank correlation coefficient was chosen to look at marginal relationships because it robust to outliers, easy to compute, and because Spearman's rank correlation does not assume that the data follows a specific probability distribution (such as the normal distribution), so it robust to deviations from normality.



The picture above is the correlation matrix, a table that displays the Pearson correlation coefficients between many variables. Each cell in the table shows the correlation between two specific variables. Correlation coefficients measure the strength and direction of a linear relationship between two continuous variables. The value of a correlation coefficient ranges from -1 to 1. A correlation coefficient of 1 indicates a strong positive linear relationship. A correlation coefficient of -1 indicates a strong negative linear relationship. From the correlation matrix above, different color represent different levels of correlation. The darker the blue, the stronger the positive correlation between the two variables, and the darker the red, the stronger the negative correlation between the two variables. By looking the correlation matrix, we can find that True grade has a strong negative correlation with gpa, which is -0.50. Pre.CogComp has a strong positive correlation with Pre.Affect, which is 0.78.

We can use the true grade as the response variable, and then use the other variables I used above as explanatory variables and generate a multiple linear regression model. In a multiple linear regression model, we have more than one independent variable (explantory variables) to predict a single dependent (response) variable. The model assumes a linear relationship between the dependent variable and all the independent

variables. Equation for multiple linear regression model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$, where $\beta_0, \beta_1, ..., \beta_p$ are coefficients or parameters associated with each independent variable, representing the change in Y for a one-unit change in each corresponding X while holding all other variables constant. $X_1, X_2, \ldots, X_p$ are the independent variables (explantory variables) that influence Y. ϵ is the error term, representing the unexplained or random variation in Y.

```
regression_model_all <- lm(Grade.Letter.Number ~ Pre.Affect + Pre.CogComp +
                           Pre.Value + Pre.Diff + Pre.Interest +
                           gpa +
                           Post.Affect + Post.CogComp + Post.Value + Post.Diff + Post.Interest, data = sa
summary(regression_model_all)
```

```
##
## Call:
## lm(formula = Grade.Letter.Number ~ Pre.Affect + Pre.CogComp +
##     Pre.Value + Pre.Diff + Pre.Interest + gpa + Post.Affect +
##     Post.CogComp + Post.Value + Post.Diff + Post.Interest, data = sats_sci_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1694 -1.5477 -0.2846  1.3788  8.9364
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.93044    0.38579  30.925  < 2e-16 ***
## Pre.Affect     0.06842    0.08077   0.847  0.39704
## Pre.CogComp   -0.04285    0.08377  -0.511  0.60906
## Pre.Value     -0.18311    0.07268  -2.519  0.01182 *
## Pre.Diff       0.04206    0.08570   0.491  0.62367
## Pre.Interest   0.12425    0.06119   2.030  0.04242 *
## gpa           -0.77223    0.05030 -15.352  < 2e-16 ***
## Post.Affect   -0.50350    0.07189  -7.004 3.22e-12 ***
## Post.CogComp  -0.35971    0.07770  -4.629 3.87e-06 ***
## Post.Value    -0.03837    0.06817  -0.563  0.57360
## Post.Diff     -0.18184    0.06978  -2.606  0.00922 **
## Post.Interest  0.08267    0.05525   1.496  0.13474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.293 on 2384 degrees of freedom
## Multiple R-squared:  0.2556, Adjusted R-squared:  0.2521
## F-statistic:  74.4 on 11 and 2384 DF,  p-value: < 2.2e-16
```

The linear model is:

True grade = 11.93044 + 0.06842*Pre.Affect* - *0.04285*Pre.CogComp - 0.18311*Pre.Value* + *0.04206*Pre.Diff + 0.12425*Pre.Interest* - *0.77223*gpa - 0.50350*Post.Affect* - *0.35971*Post.CogComp - 0.03837*Post.Value* - *0.18184*Post.Diff + 0.08267*Post.Interest

If the p-value is less than a chosen significance level (e.g., 0.05), it suggests that the corresponding independent variable has a statistically significant effect on the dependent variable.

If the p-value is greater than the significance level, it suggests that the variable may not have a statistically significant effect.

"Post.Diff" has a small p-value (0.00922), which is less than 0.05, indicating that it is likely a statistically significant predictor. "Pre.Interest" and "gpa" also have small p-values, suggesting that they are highly significant predictors. On the other hand, "Pre.CogComp," "Pre.Diff," "Post.Value," and "Post.Interest" have p-values greater than 0.05, indicating that they may not be statistically significant predictors in this model.

To make the model results more accurate, we can use the Akaike Information Criterion (AIC). The AIC is a tool for comparing the goodness of fit of different statistical models to a given dataset. It is based on the principle of finding a balance between the goodness of fit of the model and the complexity of the model. The formula of AIC is `AIC = -2 * log-likelihood + 2 * k`

log-likelihood is the maximized value of the likelihood function for the model, given the data.

k is the number of parameters in the model.

Our goal is to maximize the `log-likelihood`, and because we have -2 before the `log-likelihood`, so our goal is to minimize the `-2*log-likelihood`. The smaller value of AIC, the better the model fits the data.

AIC has lots of advantages:(1) Model selection:It balances the trade-off between model complexity and goodness of fit, helping us choose the model that best explains the data without overfitting.(2) Generalizability: Models selected using AIC tend to have better generalizability to new, unseen data because they are less likely to be overfitting the training data.

```
AIC_all = step (regression_model_all)
```

```
## Start:  AIC=3989.46
## Grade.Letter.Number ~ Pre.Affect + Pre.CogComp + Pre.Value +
##     Pre.Diff + Pre.Interest + gpa + Post.Affect + Post.CogComp +
##     Post.Value + Post.Diff + Post.Interest
##
##                  Df Sum of Sq   RSS    AIC
## - Pre.Diff        1      1.27 12540 3987.7
## - Pre.CogComp     1      1.38 12540 3987.7
## - Post.Value      1      1.67 12540 3987.8
## - Pre.Affect      1      3.77 12543 3988.2
## <none>                       12539 3989.5
## - Post.Interest   1     11.77 12551 3989.7
## - Pre.Interest    1     21.68 12560 3991.6
## - Pre.Value       1     33.38 12572 3993.8
## - Post.Diff       1     35.71 12575 3994.3
## - Post.CogComp    1    112.71 12652 4008.9
## - Post.Affect     1    258.02 12797 4036.3
## - gpa             1   1239.55 13778 4213.3
##
## Step:  AIC=3987.7
## Grade.Letter.Number ~ Pre.Affect + Pre.CogComp + Pre.Value +
##     Pre.Interest + gpa + Post.Affect + Post.CogComp + Post.Value +
##     Post.Diff + Post.Interest
##
##                  Df Sum of Sq   RSS    AIC
## - Pre.CogComp     1      1.06 12541 3985.9
## - Post.Value      1      1.52 12542 3986.0
## - Pre.Affect      1      6.30 12546 3986.9
## <none>                       12540 3987.7
## - Post.Interest   1     11.79 12552 3988.0
## - Pre.Interest    1     20.59 12561 3989.6
```

```
## - Pre.Value      1     35.13 12575 3992.4
## - Post.Diff      1     36.53 12577 3992.7
## - Post.CogComp   1    114.96 12655 4007.6
## - Post.Affect    1    264.72 12805 4035.8
## - gpa            1   1238.90 13779 4211.4
##
## Step:  AIC=3985.9
## Grade.Letter.Number ~ Pre.Affect + Pre.Value + Pre.Interest +
##     gpa + Post.Affect + Post.CogComp + Post.Value + Post.Diff +
##     Post.Interest
##
##                 Df Sum of Sq   RSS    AIC
## - Post.Value     1      1.27 12542 3984.1
## - Pre.Affect     1      6.06 12547 3985.1
## <none>                       12541 3985.9
## - Post.Interest  1     12.01 12553 3986.2
## - Pre.Interest   1     20.68 12562 3987.8
## - Post.Diff      1     36.39 12578 3990.8
## - Pre.Value      1     41.94 12583 3991.9
## - Post.CogComp   1    131.42 12673 4008.9
## - Post.Affect    1    264.05 12805 4033.8
## - gpa            1   1240.53 13782 4209.9
##
## Step:  AIC=3984.14
## Grade.Letter.Number ~ Pre.Affect + Pre.Value + Pre.Interest +
##     gpa + Post.Affect + Post.CogComp + Post.Diff + Post.Interest
##
##                 Df Sum of Sq   RSS    AIC
## - Pre.Affect     1      6.77 12549 3983.4
## <none>                       12542 3984.1
## - Post.Interest  1     11.17 12554 3984.3
## - Pre.Interest   1     22.01 12564 3986.3
## - Post.Diff      1     35.43 12578 3988.9
## - Pre.Value      1     53.12 12596 3992.3
## - Post.CogComp   1    149.87 12692 4010.6
## - Post.Affect    1    264.88 12807 4032.2
## - gpa            1   1250.97 13793 4209.9
##
## Step:  AIC=3983.44
## Grade.Letter.Number ~ Pre.Value + Pre.Interest + gpa + Post.Affect +
##     Post.CogComp + Post.Diff + Post.Interest
##
##                 Df Sum of Sq   RSS    AIC
## - Post.Interest  1      8.65 12558 3983.1
## <none>                       12549 3983.4
## - Post.Diff      1     32.14 12581 3987.6
## - Pre.Interest   1     34.20 12583 3988.0
## - Pre.Value      1     49.57 12599 3990.9
## - Post.CogComp   1    146.80 12696 4009.3
## - Post.Affect    1    258.51 12808 4030.3
## - gpa            1   1253.07 13802 4209.5
##
## Step:  AIC=3983.09
## Grade.Letter.Number ~ Pre.Value + Pre.Interest + gpa + Post.Affect +
```

```
##     Post.CogComp + Post.Diff
##
##                 Df Sum of Sq   RSS    AIC
## <none>                       12558 3983.1
## - Post.Diff      1     39.72 12598 3988.7
## - Pre.Value      1     47.53 12605 3990.1
## - Pre.Interest   1     57.20 12615 3992.0
## - Post.CogComp   1    146.53 12704 4008.9
## - Post.Affect    1    264.50 12822 4031.0
## - gpa            1   1247.88 13806 4208.1
```

After we use the AIC, we can see all the printout values, and the started AIC value is 3928.03. The AIC column in the picture represents the how will AIC value change after we drop the corresponding value. From the picture and the results above, we can find that as long as we drop off the variable "Pre.Diff", "Pre.CogComp", "Post.Value", "Pre.Affect", "Post.Interest", our AIC value become smaller, which is a good result (the smaller the AIC value, the better the model is). However, if we drop off other variables, our AIC value will increase. So we can only drop "Pre.Diff", "Pre.CogComp", "Post.Value", "Pre.Affect", "Post.Interest" predictor, and then we are having a small AIC value. We can see the summarized statistic on my final model after the variable selection below:

```
summary(AIC_all)
```

```
##
## Call:
## lm(formula = Grade.Letter.Number ~ Pre.Value + Pre.Interest +
##     gpa + Post.Affect + Post.CogComp + Post.Diff, data = sats_sci_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1689 -1.5513 -0.3043  1.3538  9.0302
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.98333    0.34587  34.647  < 2e-16 ***
## Pre.Value    -0.19474    0.06476  -3.007 0.002665 **
## Pre.Interest  0.17321    0.05251   3.299 0.000985 ***
## gpa          -0.77262    0.05015 -15.408  < 2e-16 ***
## Post.Affect  -0.45280    0.06383  -7.094 1.72e-12 ***
## Post.CogComp -0.37724    0.07145  -5.280 1.41e-07 ***
## Post.Diff    -0.17000    0.06185  -2.749 0.006026 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.293 on 2389 degrees of freedom
## Multiple R-squared:  0.2544, Adjusted R-squared:  0.2526
## F-statistic: 135.9 on 6 and 2389 DF,  p-value: < 2.2e-16
```

True grade = 11.98333 - 0.19474 * Pre.Value + 0.17321 * Pre.Interest - 0.77262 * gpa - 0.45280 * Post.Affect - 0.37724 * Post.CogComp - 0.17000 * Post.Diff

Some people may wonder why there is a negative relationship between real grades and gpa and other variables. This is because our real grades range from 1 to 13, 1 represents A+, and 13 represents F, so the smaller the number, the better the real grade.

The p-values provide strong evidence that these predictor variables are significantly associated with changes in "Grade.Letter.Number." The t-values measure the significance of these associations, and the coefficients (estimates) indicate the direction and strength of the relationship between each predictor and the outcome variable.