DSP110_Fall_2023_Project1

December 11, 2023

#

Project (Ch. 1-6): Exploring Variation

[1]:
```
# Load the CourseKata library
suppressPackageStartupMessages({
    library(coursekata)
})
```

1 Appendix

- **Initial equation:** `SAT = GPA + Other Stuff`
- **Refined equation:** `SAT = GPA + Smoking + Other Stuff`
- **Hypothesis:** There is a significant dependence between a student's SAT score and their GPA. I anticipate that students with higher GPAs will tend to achieve higher SAT scores, suggesting a positive correlation between these two variables.
- **Results:** Owing to the limited sample size employed in this study, it is imperative to acknowledge that the outcomes derived from the generated data hold relevance exclusively within the confines of the observed sample. Generalizing these results to broader populations, whether at a regional or global scale, is not methodologically justified. The inherent limitations of the sample size underscore the necessity for caution in extending these findings beyond the specific group surveyed, emphasizing the localized applicability of the generated data and its corresponding results.
- **Conclusion:** The research endeavor unfolds through three distinct phases: firstly, a thorough scrutiny of the dataset; secondly, a refinement of the research trajectory by delving into dataset variation and modeling intricacies; and finally, an acknowledgment of the intricate interplay among explanatory variables, particularly within the domain of SAT scores. The initial emphasis on the correlation between SAT scores and GPA evolved into a more nuanced approach, introducing the hypothesis of a multifaceted relationship that extends to sub-variables influencing SAT scores. Furthermore, a conjecture was made regarding the intricate yet non-hierarchical relationship between a student's GPA status and their smoking habits, highlighting the complexity embedded within the dataset.

2 Introduction

For this project, I have opted to utilize the `StudentSurvey` data frame for several compelling reasons. Primarily, its prominence stems from being the most expansive dataset among the available options. By harnessing a substantial dataset, we enhance our ability to closely mirror real-world

scenarios, particularly within the context of professional employment. This is particularly salient as many industries commonly deal with vast datasets, rendering `StudentSurvey` an ideal choice for honing data science skills. Furthermore, the dataset offers an abundance of information, affording me a richer and more intricate landscape for my analytical endeavors. The data for this dataset has been provided to both my group and myself by my DSP 110 professor, Mr. Harrison Dekker. Additionally, the custodial responsibility for the dataset lies with Robin Lock. It is important to underscore that the overarching aim of this data collection effort is to glean comprehensive demographic insights about the student population, thereby facilitating a deeper understanding of this particular cohort.

In the `StudentSurvey` data frame, the cases are representative of students enrolled in the introductory statistics course, while the variables encompass a wide array of general demographic attributes and information relevant to each student. The data frame exemplifies a comprehensive collection of information that collectively characterizes the student population within the context of this survey.

[2]:
```
str(StudentSurvey)
```

```
'data.frame':   362 obs. of  18 variables:
 $ Year      : Factor w/ 5 levels "","FirstYear",..: 4 5 2 3 5 5 2 5 3 2 …
 $ Gender    : Factor w/ 2 levels "F","M": 2 1 2 2 1 1 1 2 1 1 …
 $ Smoke     : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 …
 $ Award     : Factor w/ 3 levels "Academy","Nobel",..: 3 1 2 2 2 2 3 3 2 2 …
 $ HigherSAT : Factor w/ 3 levels "","Math","Verbal": 2 2 2 2 3 3 2 2 3 2 …
 $ Exercise  : num  10 4 14 3 3 5 10 13 3 12 …
 $ TV        : int  1 7 5 1 3 4 10 8 6 1 …
 $ Height    : int  71 66 72 63 65 65 66 74 61 60 …
 $ Weight    : int  180 120 208 110 150 114 128 235 NA 115 …
 $ Siblings  : int  4 2 2 1 1 2 1 1 2 7 …
 $ BirthOrder: int  4 2 1 1 1 2 1 1 2 8 …
 $ VerbalSAT : int  540 520 550 490 720 600 640 660 550 670 …
 $ MathSAT   : int  670 630 560 630 450 550 680 710 550 700 …
 $ SAT       : int  1210 1150 1110 1120 1170 1150 1320 1370 1100 1370 …
 $ GPA       : num  3.13 2.5 2.55 3.1 2.7 3.2 2.77 3.3 2.8 3.7 …
 $ Pulse     : int  54 66 130 78 40 80 94 77 60 94 …
 $ Piercings : int  0 3 0 0 6 4 8 0 7 2 …
 $ Sex       : Factor w/ 2 levels "Female","Male": 2 1 2 2 1 1 1 2 1 1 …
```

The outcome variable that I aim to investigate is the SAT scores. SAT scores are particularly intriguing for my analysis as they tend to exhibit substantial variation among students, making them an excellent candidate for a deeper exploration within the realm of statistics. To illustrate this point, let's consider the distribution of SAT scores:

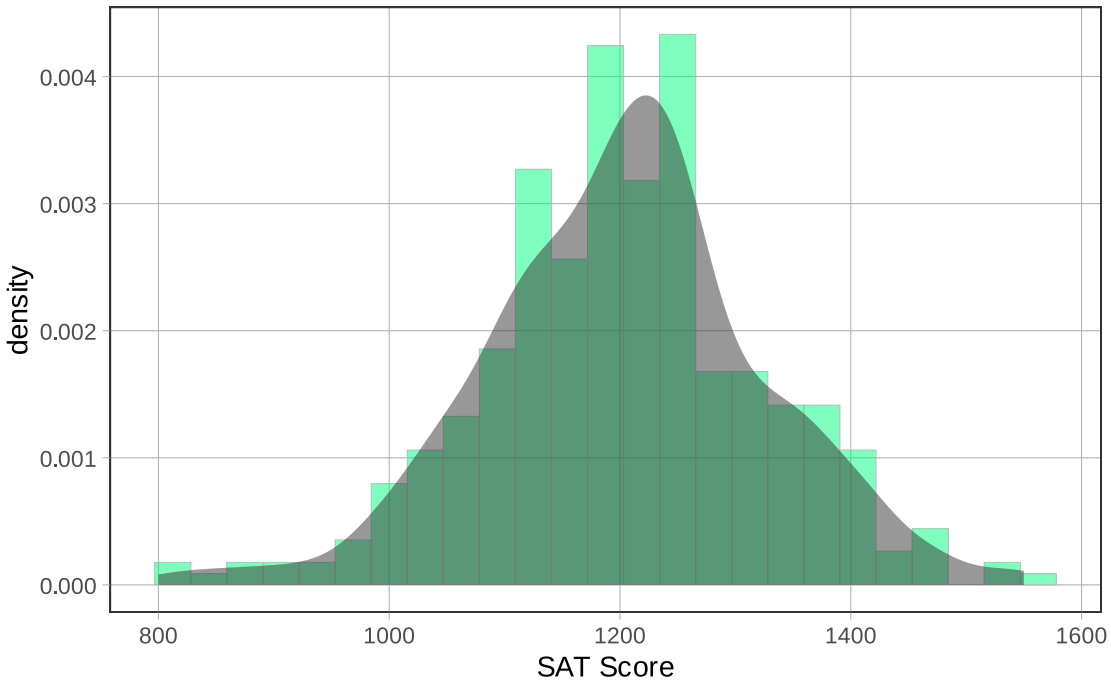
[3]:
```
gf_dhistogram(~ SAT, data = StudentSurvey, color = "gray", fill =␣
 ↪"springgreen") %>%
   gf_labs(title = "Figure 1. Distribution of SAT Scores", x = "SAT Score") %>%
       gf_density()

favstats(~ SAT, data = StudentSurvey)
```

A data.frame: 1 × 9

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
800	1130	1200	1270	1550	1203.627	121.2852	362	0

Figure 1. Distribution of SAT Scores



The distribution of SAT scores (Figure 1.) is expected to exhibit a wide range of values, reflecting the diverse academic backgrounds and aptitudes of the surveyed students. This inherent variability provides an excellent opportunity for me to apply statistical techniques and uncover meaningful insights within the dataset.

The research question I intend to investigate centers around the potential relationship between a student's SAT scores and their GPA. Specifically, I aim to discern whether there exists a dependency between these two variables. In this exploration, I will seek to elucidate whether a student's academic performance, as measured by their GPA, can serve as a predictor of their performance in the SAT examination. This question delves into the intricate interplay between academic achievement and standardized testing, shedding light on how these factors may influence each other.

Hypothesis: My initial hypothesis regarding this situation posits that there is a significant dependence between a student's SAT score and their GPA. I anticipate that students with higher GPAs will tend to achieve higher SAT scores, suggesting a positive correlation between these two variables.

Given the acknowledgment that other unidentified factors may also influence SAT scores, I have formulated the following word equation to holistically represent the relationship between the two variables:

```
SAT = GPA + Other Stuff
```

This equation encapsulates the idea that a student's SAT score is a product of their GPA and a set of additional, yet unspecified, factors that collectively determine their performance in the SAT examination. By incorporating the notion of `Other Stuff`, this equation acknowledges the complexity of the relationship while emphasizing the role of GPA as a key component in this context.

3 Exploring Variation

[4]:
```
gf_point(SAT ~ GPA, data = StudentSurvey) %>%
  # Creates the blue median line
  gf_lm() %>%
    gf_labs(title = "Figure 2. SAT Scores to GPA")

# Compare GPA to smoking
gf_histogram(~ GPA, data = StudentSurvey) %>%
  gf_facet_grid(Smoke ~ .) %>%
    gf_labs(title = "Figure 3. GPA to Student Smokers")
```

```
Warning message:
"Removed 17 rows containing non-finite values (`stat_lm()`)."
Warning message:
"Removed 17 rows containing missing values (`geom_point()`)."
Warning message:
"Removed 17 rows containing non-finite values (`stat_bin()`)."
```

Figure 2. SAT Scores to GPA

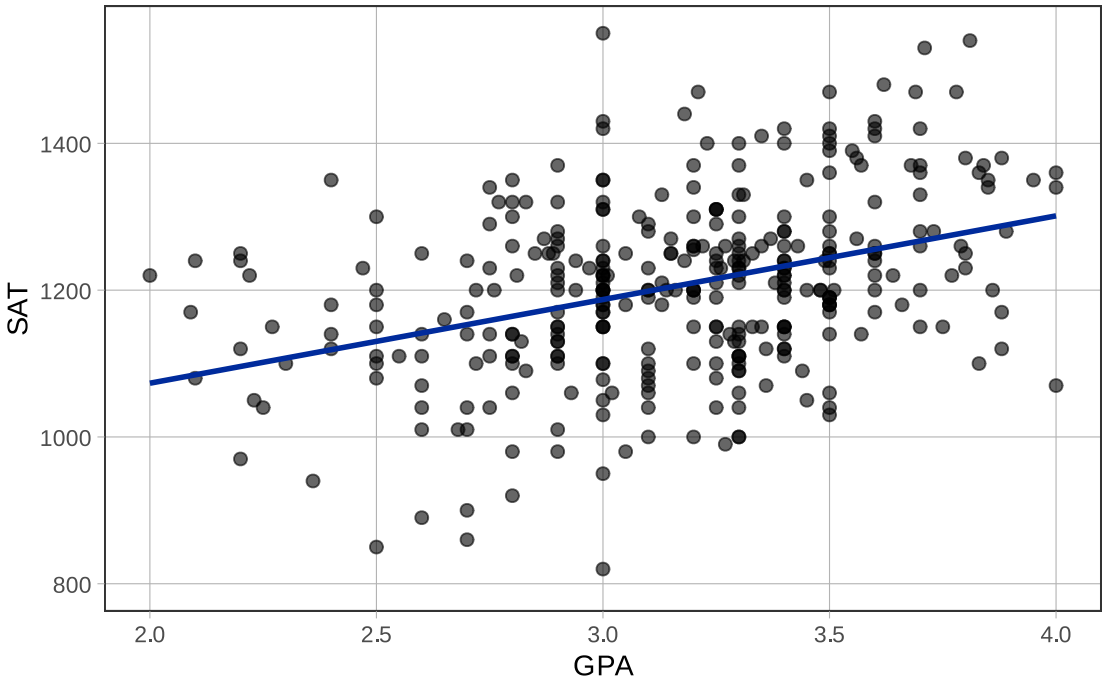
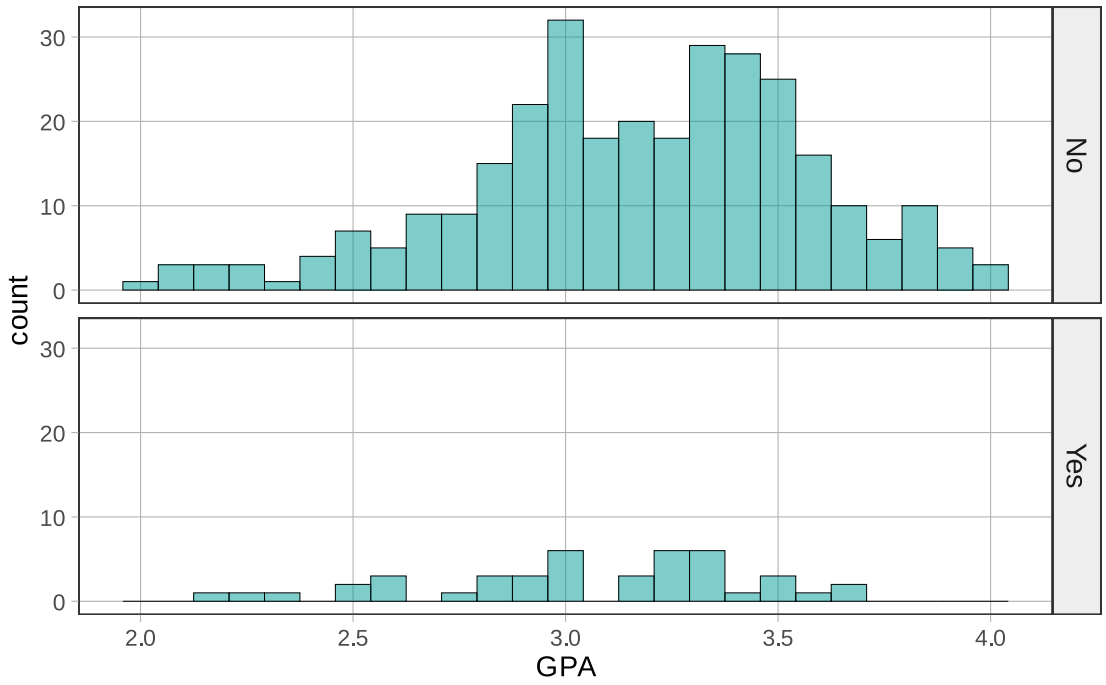


Figure 3. GPA to Student Smokers



Analysis: According to the data depicted in Figure 2, it is evident that the predictive capacity of SAT Scores cannot be exclusively attributed to the variable of GPA. Conversely, Figure 3 delineates an intricate investigation into the interplay between a student's smoking habits and the likelihood of attaining a high GPA. This analysis subsequently enables us to infer whether a high GPA is a significant determinant in achieving a successful SAT score.

Randomness/Chance: This dataset does indeed exhibit a certain level of stochasticity. However, the extent of this inherent randomness remains elusive without a meticulous scrutiny of intricate details.

Note: The sources of variability encompass factors such as the student's quality of sleep before the examination, their innate proficiency in test-taking, the levels of stress experienced before, during, and after the exam, and even extraneous events such as conflicts or arguments occurring the night before the assessment. These multifaceted variables collectively contribute to the complex and nuanced nature of the dataset's observed randomness.

DGP: The Data Generating Process (DGP) in this context is linked to a student survey administered to a specific class. Given the limited scope of this small sample size, it is imperative to recognize that the outcomes derived from both the collected data and subsequent research analysis lack generalizability to the broader population of students worldwide or even within the United States. It is crucial to acknowledge that the findings are applicable solely to the subset of individuals surveyed and do not extend to a more expansive demographic, underscoring the importance of cautious interpretation and restrained extrapolation of the results.

Refined Research Question/Hypothesis: As gleaned from the preliminary insights provided by these visualizations, it is apparent that SAT scores exhibit a discernible association with the presence or absence of a high GPA. Furthermore, it can be inferred that the attainment of a high GPA is contingent upon the student's smoking status, as this binary factor appears to significantly influence academic performance. These findings suggest an intricate interplay among these variables, forming a chain of relationships within the dataset. However, it is essential to employ rigorous statistical analysis to confirm and quantify these relationships, accounting for potential confounding variables and providing a robust foundation for these deductions.

In light of the aforementioned constraints the refined theoretical equation is tailored to reflect the focus of the interconnectivity of Smoking to GPA:

`SAT = GPA + Smoking + Other Stuff`

The presented equation encapsulates the concept that a student's SAT score is intricately influenced by their GPA, which in turn is shaped by the binary variable denoting whether or not the student is a smoker. Additionally, the equation incorporates a set of supplementary factors denoted as "Other Stuff," which, although unspecified, are acknowledged for their potential impact on the overall model. The explicit inclusion of the `Smoking` variable underscores the recognition of the intricate relationship between smoking habits and a student's GPA, while simultaneously emphasizing the pivotal role of GPA as a central determinant in the context of SAT scores within this specific analytical framework.

4 Model Variation

[5]:
```
# Creates and empty model and saves it into an R object called GPA_model
GPA_model <- lm(GPA ~ NULL, data = StudentSurvey)
GPA_model

# Saves the favstats of the variable GPA into an R object called GPA_stats
GPA_stats <- favstats(~ GPA, data = StudentSurvey)
supernova(GPA_model)

# Creates and empty model and saves it into an R object called SAT_model
SAT_model <- lm(SAT ~ NULL, data = StudentSurvey)
SAT_model

# Saves the favstats of the variable GPA into an R object called GPA_stats
SAT_stats <- favstats(~ SAT, data = StudentSurvey)
supernova(SAT_model)
```

```
Call:
lm(formula = GPA ~ NULL, data = StudentSurvey)

Coefficients:
(Intercept)
      3.158
```

```
Refitting to remove 17 cases with missing value(s)
  lm(formula = GPA ~ NULL, data = listwise_delete(StudentSurvey,
c("GPA")))


 Analysis of Variance Table (Type III SS)
 Model: GPA ~ NULL

                            SS  df    MS   F PRE   p
 ----- --------------- | ------ --- ----- --- --- ---
 Model (error reduced) |    --- ---   --- --- --- ---
 Error (from model)    |    --- ---   --- --- --- ---
 ----- --------------- | ------ --- ----- --- --- ---
 Total (empty model)   | 54.579 344 0.159


Call:
lm(formula = SAT ~ NULL, data = StudentSurvey)

Coefficients:
(Intercept)
       1204


 Analysis of Variance Table (Type III SS)
 Model: SAT ~ NULL

                                SS  df        MS   F PRE   p
 ----- --------------- | ---------- --- --------- --- --- ---
 Model (error reduced) |        --- ---       --- --- --- ---
 Error (from model)    |        --- ---       --- --- --- ---
 ----- --------------- | ---------- --- --------- --- --- ---
 Total (empty model)   | 5310346.655 361 14710.102
```

[6]:
```
gf_histogram(~ GPA, data = StudentSurvey, bins = 8) %>%
    gf_vline(xintercept = 3.158) %>%
        gf_labs(title = "Figure 4. Student GPA Model with Marked Average")

gf_histogram(~ SAT, data = StudentSurvey, bins = 8) %>%
    gf_vline(xintercept = 1204) %>%
        gf_labs(title = "Figure 5. Student SAT Model with Marked Average")
```

```
Warning message:
"Removed 17 rows containing non-finite values (`stat_bin()`)."
```

Figure 4. Student GPA Model with Marked Average

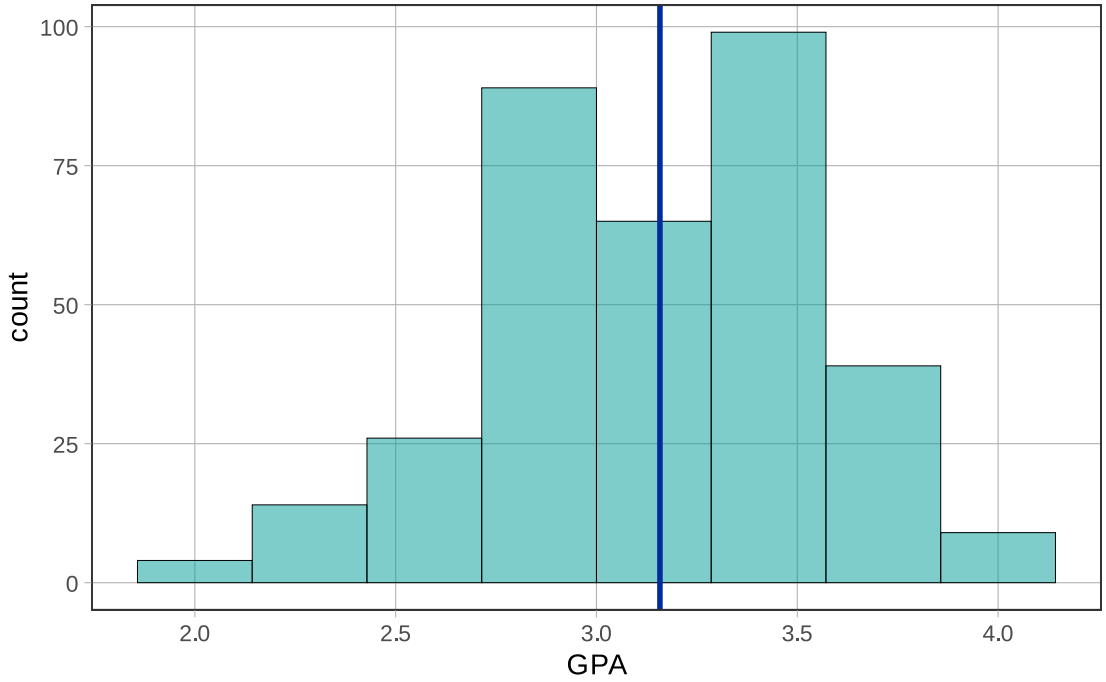
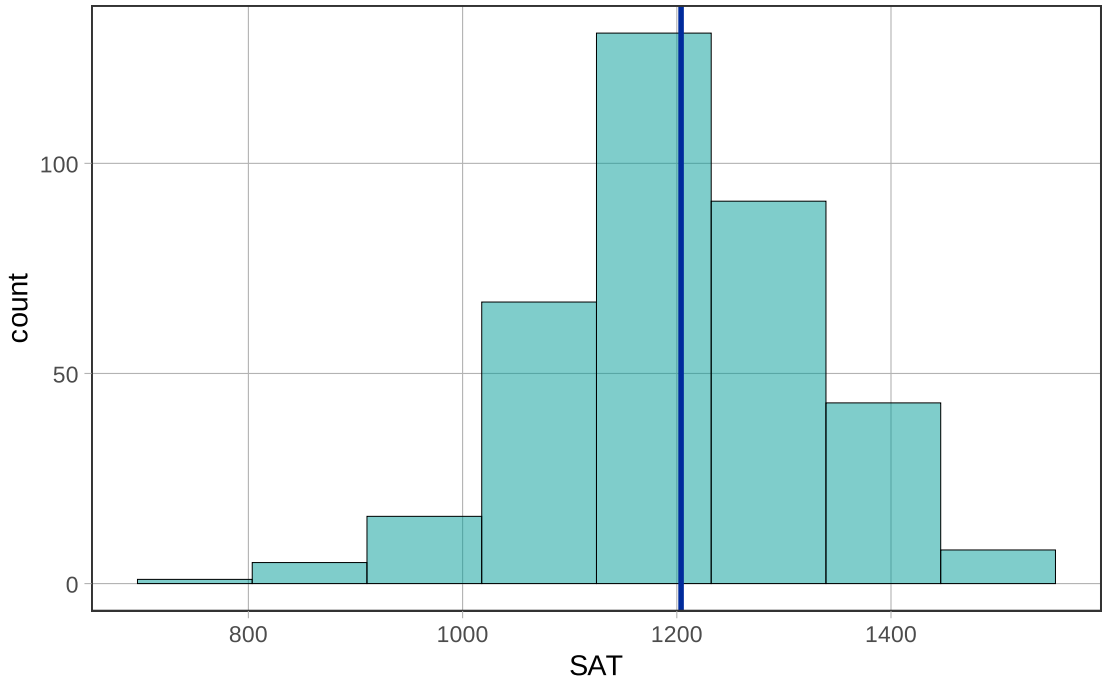


Figure 5. Student SAT Model with Marked Average



5 Conclusion

This undertaking can be systematically dissected into three distinct phases. The initial step entailed a comprehensive examination of the dataset in question. Subsequent to this preliminary data exploration, a more refined and focused research direction became apparent by exploring the variation within the dataset and modeling that same data. Initially, the research interest was centered on a comparative analysis between SAT scores and GPA status. However, as the investigation progressed, it became evident that SAT scores were not solely influenced by a single explanatory variable but comprised a complex set of explanatory variables, each with its own sub-variables that held pertinent information within the dataset. This realization necessitated a more nuanced and comprehensive approach to the research. In light of this complexity, the formulated hypothesis posited that the status of SAT scores was contingent upon a multifaceted relationship involving not only a high GPA but also an intricate interplay of sub-variables within the SAT scores. Furthermore, it was conjectured that the status of a student's GPA was intricately unlinked to their smoking habits, thereby implying a non-hierarchical relationship between these key elements within the dataset.