Cap Dap Report

Drew

12/14/2022

By Drew Racioppa

Introduction

- For this project I looked at frog surveying data on St. Lawrence Campus from the year 2017.
- This data was student collected for a Global Amphibian Decline class, to be used in comparison of the surveying data used the year prior, and pulled from ArcGIS software.
- The data is observational field data collected by several different students, using different methods of surveying.
- I wanted to look into this data knowing that it could be incorrect or inaccurate to prove that the data collection methods for this class are lacking, and need to be more consistent across the board. Although there is no recorded data specific to this location, generalized models for all the statistical tests I will run, have been proven to have significant results. If the data anaylisis tests are not significant, it can be evidence to help showcase the data collection methods of this class, are inaccurate and need to be adressed. I personally took this class and found errors in the collection methods, however I think most tests will still show significant results since many of the errors were removed in the raw data table.

Analysis

```
rm (list = ls())
library(dplyr)
library(ggplot2)
library(here)
library(tidyverse)
```

Data Import and Formating

```
DF <- read.csv(here("Data", "Frog_Data2.csv"), stringsAsFactors = TRUE)
```

```
DF <- DF %>% na.omit()
view(DF)
```

```
DF <- DF %>%
  rename(Survey_Type = "SurveyType", Species_Name = "SpeciesName", Air_Temp_F = "AirTempF", Number_of_I
```

```
DF$Activity <- sub("^Moving.*", "Moving", DF$Activity)
DF$Activity <- sub("^Basking.*", "Basking", DF$Activity)
DF$Activity <- sub("^Thermoregulating.*", "Thermoregulating", DF$Activity)
DF$Activity <- sub("^Swimming.*", "Swimming", DF$Activity)
DF$Activity <- sub("^Calling.*", "Calling", DF$Activity)
DF$Activity <- sub("^Feeding.*", "Feeding", DF$Activity)
DF$Activity <- sub("^Jumping.*", "Jumping", DF$Activity)
DF$Activity <- sub("^Dead.*", "Dead", DF$Activity)
```

```
view(DF)
```

```
DF <- DF  %>%
  filter(Activity =='Moving'|Activity == 'Basking'| Activity == 'Thermoregulating'|Activity == 'Swimmin
view(DF)
```

Relationship 1

I will be exploring the relationship between Wind-Speed and Relative Humidity
My hypothesis is that the higher the wind speed, the less relative humidity there will be. This is from wind blowing away air molecules, decreasing the humidity. This is an initial test to first see if the data is at all accurate, and second to see if it correlates to how it theoretically should.

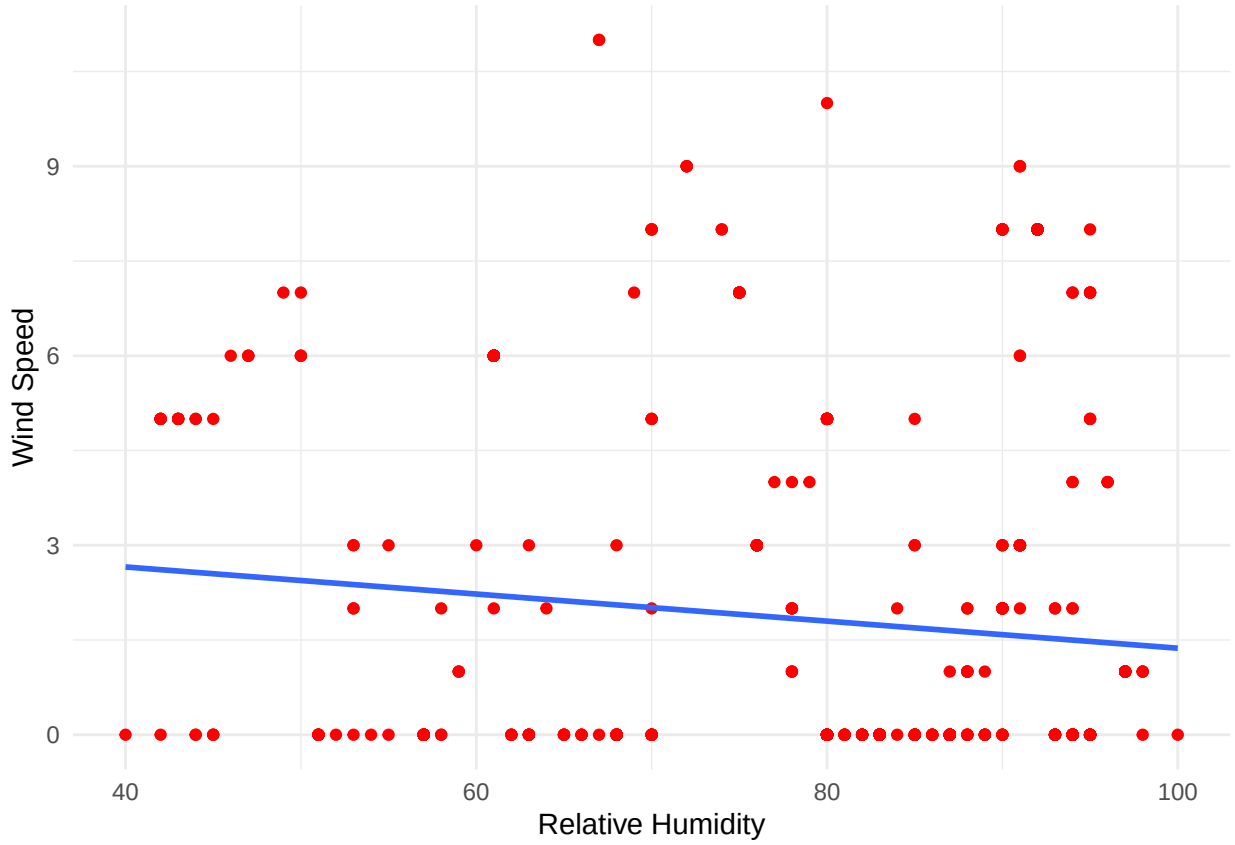
Hypothesis: Higher wind speed will result in a lower relative humidity

Null Hypothesis: No correlation between wind speed and relative humidity

Plot #1 Humidity and Wind Speed

```
DF_new <- DF%>%filter(Wind_Speed<15 & Rel_Humid>25)

ggplot(DF_new, aes(x = Rel_Humid, y = Wind_Speed)) +
geom_point(color = "red")+
  geom_smooth(method = lm, se = FALSE)+
    xlab("Relative Humidity") +
    ylab("Wind Speed")+
  theme_minimal()
```



Data Analysis: Simple Linear Regression

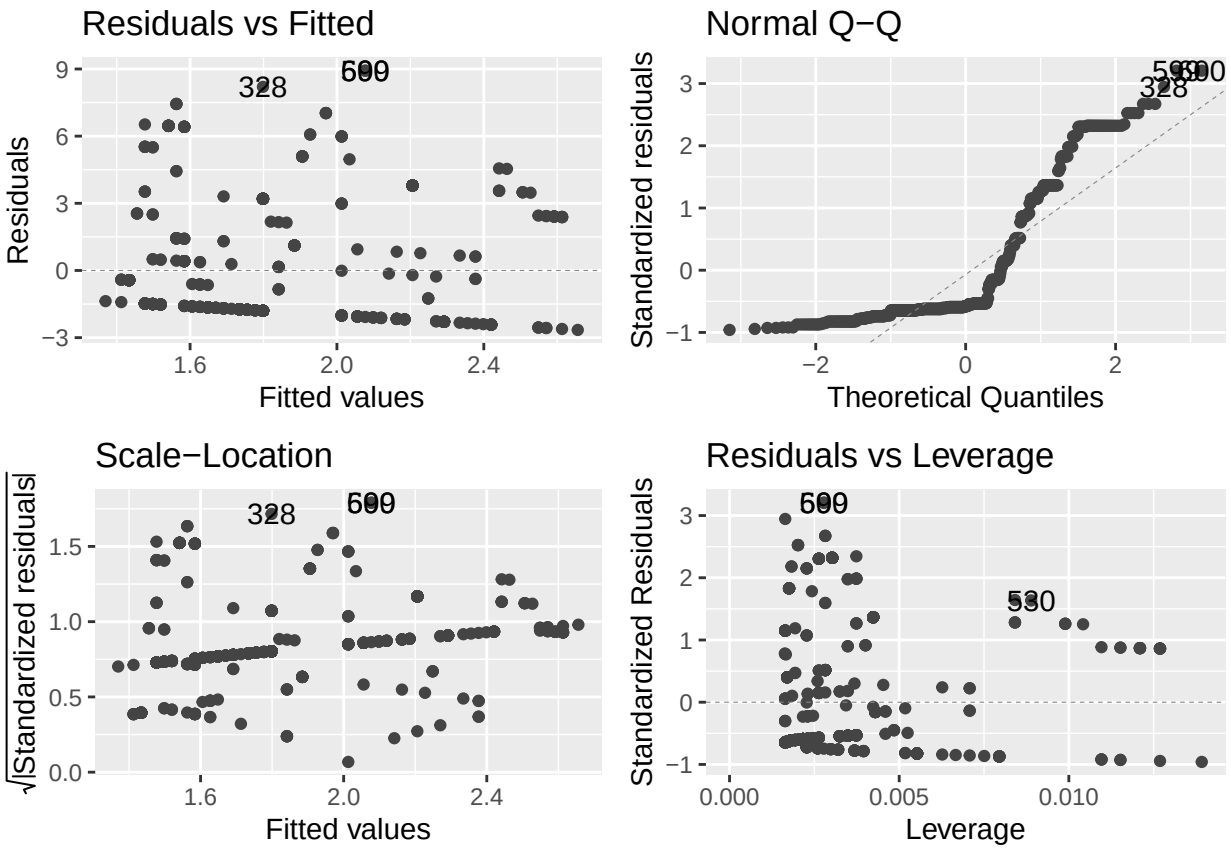
```
LM1 <- lm(Wind_Speed ~ Rel_Humid,
data = DF_new)

summary(LM1)
```

```
##
## Call:
## lm(formula = Wind_Speed ~ Rel_Humid, data = DF_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.656 -1.798 -1.606  1.416  8.923
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.513379   0.636602   5.519 5.04e-08 ***
## Rel_Humid   -0.021436   0.007942  -2.699  0.00714 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.787 on 611 degrees of freedom
## Multiple R-squared:  0.01178,    Adjusted R-squared:  0.01017
## F-statistic: 7.285 on 1 and 611 DF,  p-value: 0.007145
```

Plot #2

```
library(ggfortify)
autoplot(LM1, smooth.colour = NA)
```




```
  relation <- lm(DF$Wind_Speed~DF$Rel_Humid)
```

Results: The p-value: 0.007145 shows that the data is statistically significant so we can reject the null hypothesis. Although it is significant the study area and location could have been impacts of the data. The instruments to measure the data was also unstated so there could have been a discrepancy over the measurements.

Relationship 2

I will be exploring the relationship between what specific species were surveyed at specific relative humidity. My hypothesis is that amphibians who reside in more aquatic landscapes like the green frog, bullfrog, and wood frog, will have a higher average than the rest of the amphibians.

Hypothesis: Aquatic residing amphibians will have a higher mean average Null Hypothesis: No correlation between species and relative humidity

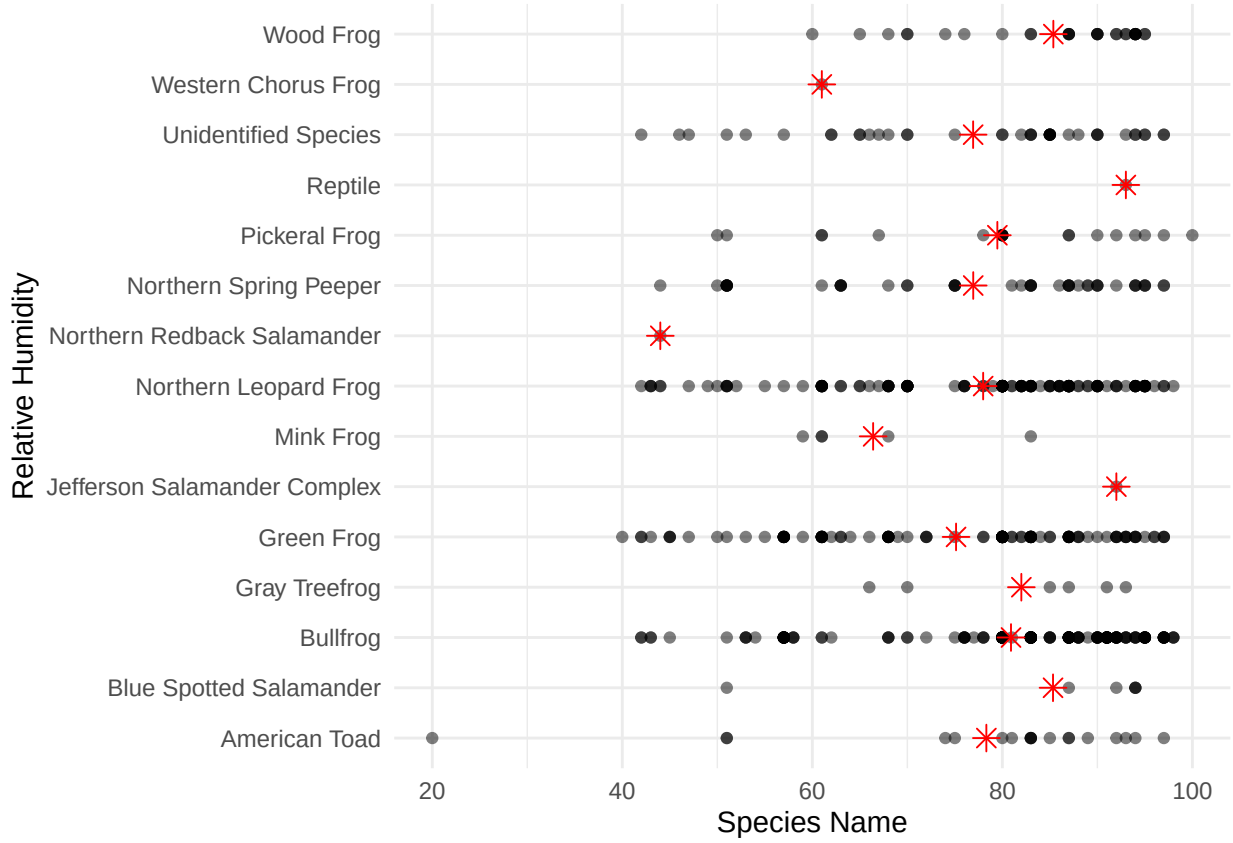
Finding mean of relative humidity by species name

```
DF2 <- DF %>%
  filter(Rel_Humid > 15)

Mean_DFIND <- DF2%>%
group_by(Species_Name) %>%
  summarise(Mean_Num = mean(Rel_Humid))
```

Plot #3 Relative Humidity by Species

```
ggplot(DF2, aes(x = Species_Name, y = Rel_Humid))+
  geom_point(alpha = 0.5) +
  geom_point(data = Mean_DFIND, aes(y = Mean_Num), color = "red", size = 3, shape = 8) +
  coord_flip() +
    xlab("Relative Humidity") +
  ylab("Species Name")+
  theme_minimal()
```



Data Analysis: ANOVA Test

```
LM2 <- lm(Rel_Humid ~ Species_Name, data = DF2)

anova(LM2)
```

```
## Analysis of Variance Table
##
## Response: Rel_Humid
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Species_Name  14   7051  503.64  2.4331 0.002465 **
## Residuals    607 125644  206.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(LM2)
```

```
##
## Call:
## lm(formula = Rel_Humid ~ Species_Name, data = DF2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.316  -7.981   4.078  10.086  21.872
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             78.3158     3.3007  23.727   <2e-16
## Species_NameBlue Spotted Salamander      7.0175     6.7374   1.042   0.2980
## Species_NameBullfrog                     2.5944     3.4834   0.745   0.4567
## Species_NameGray Treefrog                3.6842     6.7374   0.547   0.5847
## Species_NameGreen Frog                  -3.1878     3.5426  -0.900   0.3686
## Species_NameJefferson Salamander Complex 13.6842    14.7610   0.927   0.3543
## Species_NameMink Frog                   -11.9158     7.2314  -1.648   0.0999
## Species_NameNorthern Leopard Frog        -0.3351     3.4971  -0.096   0.9237
## Species_NameNorthern Redback Salamander -34.3158    14.7610  -2.325   0.0204
## Species_NameNorthern Spring Peeper       -1.3810     3.9235  -0.352   0.7250
## Species_NamePickeral Frog                 1.1579     4.6678   0.248   0.8042
## Species_NameReptile                      14.6842    14.7610   0.995   0.3202
## Species_NameUnidentified Species         -1.3908     4.0086  -0.347   0.7288
## Species_NameWestern Chorus Frog          -17.3158    14.7610  -1.173   0.2412
## Species_NameWood Frog                     7.0509     4.2183   1.671   0.0951
##
## (Intercept)                             ***
## Species_NameBlue Spotted Salamander
## Species_NameBullfrog
## Species_NameGray Treefrog
## Species_NameGreen Frog
## Species_NameJefferson Salamander Complex
## Species_NameMink Frog                   .
## Species_NameNorthern Leopard Frog
## Species_NameNorthern Redback Salamander  *
## Species_NameNorthern Spring Peeper
## Species_NamePickeral Frog
## Species_NameReptile
## Species_NameUnidentified Species
## Species_NameWestern Chorus Frog
## Species_NameWood Frog                   .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 607 degrees of freedom
## Multiple R-squared:  0.05314,    Adjusted R-squared:  0.0313
## F-statistic: 2.433 on 14 and 607 DF,  p-value: 0.002465
```

Results: The p-value: 0.002465, although statistically significant, does not paint the entire picture. This is because an ANOVA test measures every Coefficient against the Intercept and if one species is found to be signifcant, then the p-value can lower. Since there was only one Red back Salamander surveyed and it happened to be found at one of the extreme ends of the range, its small p-value is low. We dont know if the data would still be significant without this point, but the averages of most other amphibians are within a range of 20%

Relationship 3

I will be exploring the relationship between temperature and amphibian activity. My hypothesis is that amphibians who were surveyed in the extreme temperature ranges, (High and Low) are more likely to have been surveyed with passive behavior. Amphibians who were surveyed doing active activities would have been found in the median temperature.

Hypothesis: Passive amphibians will be surveyed in the extreme temperatures Null Hypothesis: There is no relationship between activity and surveyed temperature

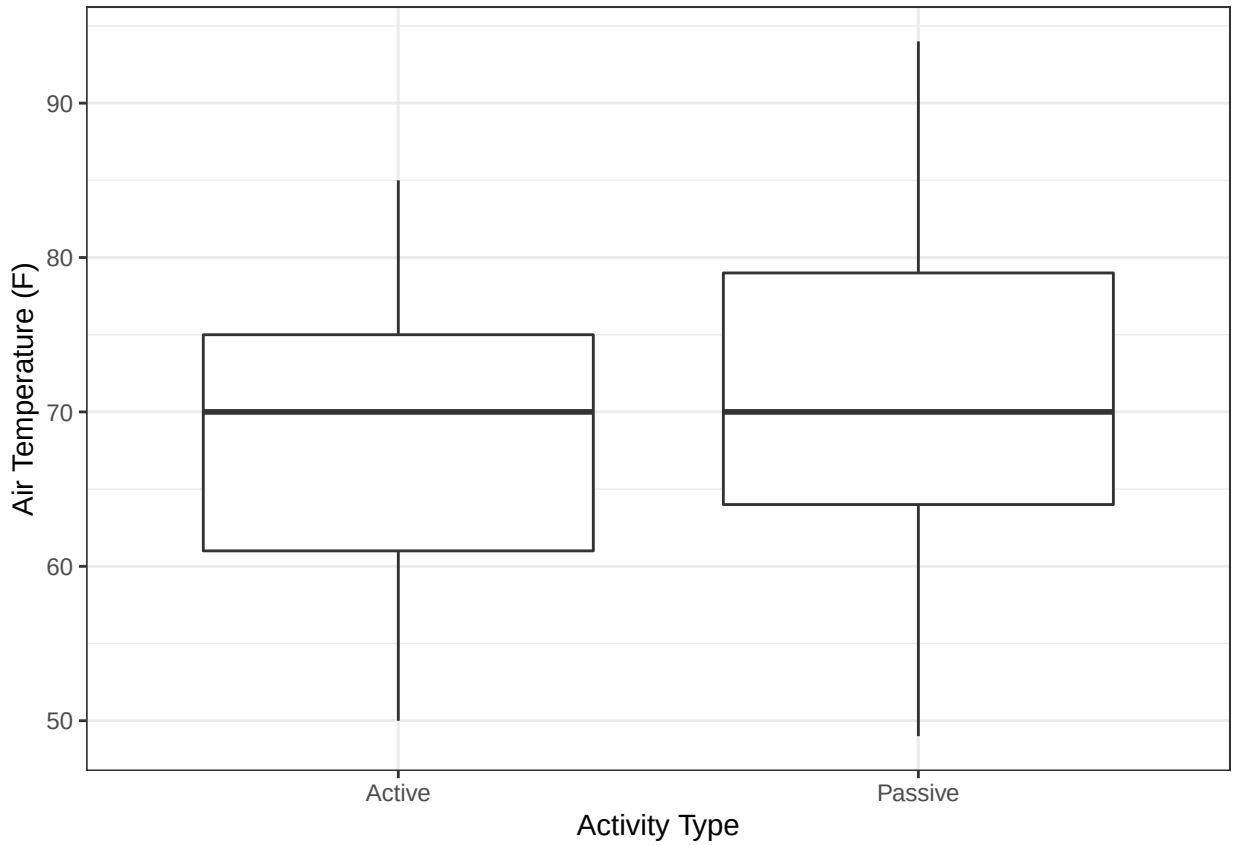
Creating Passive and Active catagories for different activities

```
active_vec <- c("Moving", "Jumping")
passive_vec <- c("Basking", "Thermoregulating")

DF_activity <- DF %>%
  filter(Activity %in% active_vec |
           Activity %in% passive_vec) %>%
  mutate(activity_type = case_when(Activity %in% active_vec ~ "Active",
                                   Activity %in% passive_vec ~ "Passive")) %>%
  filter(Air_Temp_F > 25)
```

Plot #4 Activity by Airtemp

```
ggplot(DF_activity, aes(x = activity_type, y = Air_Temp_F)) +
  geom_boxplot()+
  theme_bw()+
    xlab("Activity Type") +
  ylab("Air Temperature (F)")
```




```
t.test(Air_Temp_F ~ activity_type, data = DF_activity)
```

```
##
##  Welch Two Sample t-test
##
## data:  Air_Temp_F by activity_type
## t = -2.399, df = 141.91, p-value = 0.01774
## alternative hypothesis: true difference in means between group Active and group Passive is not equal
## 95 percent confidence interval:
##  -4.9230121 -0.4749623
## sample estimates:
##  mean in group Active mean in group Passive
##              68.26214              70.96112
```

Results: The p-value: 0.01774 shows that the data is significant but not very. Part of the reason for this result is because of a lack of splitting up passive the passive temperature range. This was an incorrect graph to visualize the data, even though the means of the average were very similar. My hypothesis was correct in assuming that the passive section held more extremities yet averaged the same as the active group.

Biological Summary Overall the data collected was very hard to work with and much of it was omitted from having either a lack of data or incorrect data (outlines that were impossible). The data that was viable gave statistically significant results for all three tests, however they were mostly just barley significant at the best case scenario. The goal of this project was to analyze data to see if it was viable to continue collecting data in this manner for future classes, and I have concluded that there needs to be significant changes. The initial data sheet needs data validation for all the categorical data, and all rows need to be filled in. There is

room for some leniency since this is student collected data, however several of these suggestions are more than capable for students.

Reflection I would showcase Relationship 3 in a different graph that splits the temperatures ranges and have side by side comparisons of each 10 degree range of both active and passive activity types. I would also try and find data from a more established scientific study, although seeing how data is analyzed and collected at the school I attend is very fascinating. I found the most challenging part of this project to be my hypothesis. Since the data relationships I choose have previously studied and recorded significance, my hypothesis was based around my own knowledge of the class itself. I often felt too close to the issue. I found that statistical analysis, is not what always proves an argument, and can be a factor in determining a larger issue. Since data can be manipulated, the results can vary based on how you use these factors to prove or disprove the hypothesis.