Gwyneth Yuen
ITWS 4600 - Data Analytics
December 8, 2022

Term Project
Assignment 6

Abstract and Introduction

Although the pandemic has introduced online and hybrid schooling into the education industry, prior to this, absences in grade school, specifically elementary school from the age of five to the age of nine, have always been an issue for some individuals. Practically every school has some children who tend to have a lot more absences than other children. If a child is absent multiple times a month or even multiple times a week, it is important to determine why or if there is any correlation between that and any aspect of their home life, such as nutrition, physical activity, and health complications. With this information, schools have the opportunity to choose to provide anything that these children lack at home. They can provide healthier meals for lunch or extend any existing recess or physical education (P.E.) time during school hours or make any other adjustments that will benefit the children. I found this topic interesting because as someone who grew up in a household where not only food was available, but healthy food was available, I rarely had school absences. I noticed that there were always some of my peers who tended to be absent a lot, and I had always wondered the reason why.

According to a study done by Cambridge University, “food insecurity, poverty and family variables are also relevant indicators for chronic school absenteeism.” [1] This suggests that it is

[1] https://www.cambridge.org/core/journals/public-health-nutrition/article/school-absenteeism-is-linked-to-household-food-insecurity-in-school-catchment-areas-in-southern-nevada/78A2BCD58D4E3C5BF1BAB724FB40B212

important to continue the study of what factors, in addition to the listed ones, have the biggest impact on school absenteeism.

This leads to the hypothesis that parents who eat healthier and do more physical activity, as well as not having any health issues like diabetes and obesity, have children who have less absences in elementary school. The main variables to be considered and compared against school absenteeism are:

- Consumption of sugary drinks in adults
- Consumption of fruits and vegetables in adults
- Physical activity within the last 30 days in adults
- Obesity in adults
- Diabetes in adults

The predictions I made were that: (1) The consumption of sugary drinks, obesity, and diabetes would all have a positive correlation with school absences - an increase in any of the three values means there will likely be an increase in school absences, and (2) the consumption of fruits and vegetables and physical activity will have a negative correlation with school absences - if an adult consumes more nutritious food and does some sort of physical activity within 30 days will likely mean less school absences. This brings up the question: Do regions with adults who have poorer health and nutrition choices also have children with more school absences, and what is the correlation?

Data Description and Exploratory Data Analytics

The primary dataset I used was the Community Health Profiles (CHP) dataset, as follows:

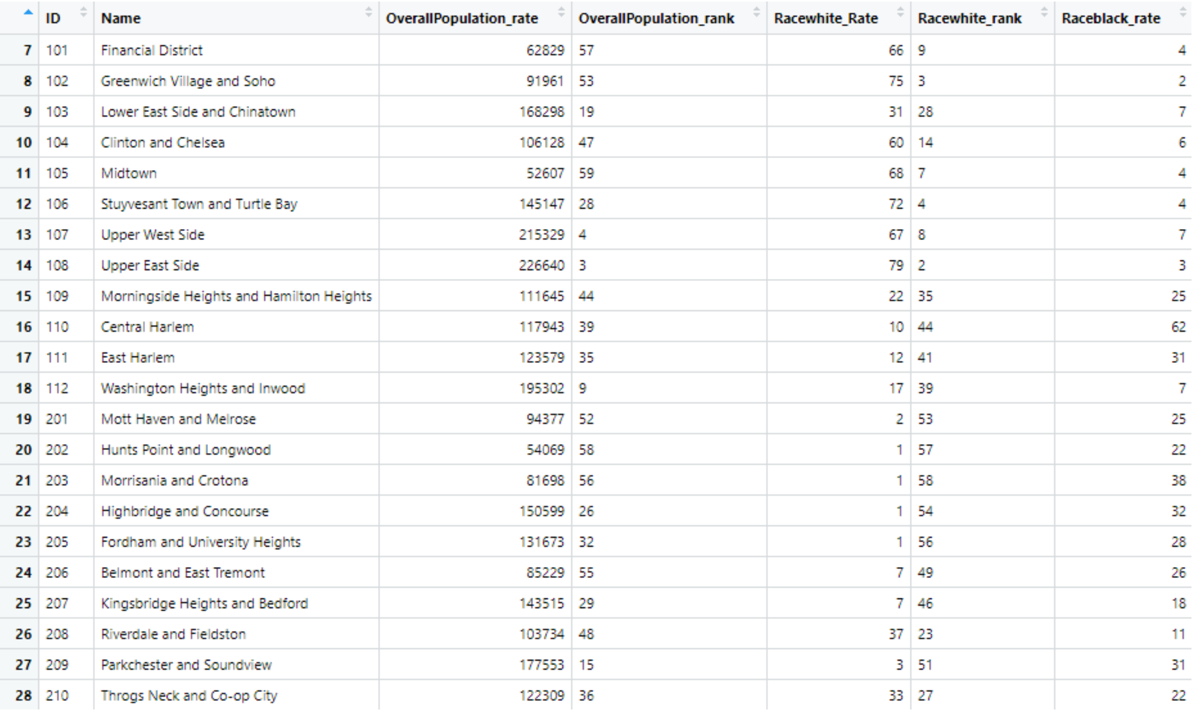
https://www.nyc.gov/site/doh/data/data-publications/profiles.page

The CHP data contains information from 59 community districts and 5 total boroughs in New York City, and is obtained through the Community Health Survey and information from the NYC department of education, specifically the FITNESSGRAM test that children from kindergarten to eighth grade are required to complete. The Community Health Survey is conducted annually of adult New Yorkers to get "estimates on a range of chronic diseases and behavioral risk factors."[2] The Community Health Survey is a "cross-sectional telephone survey with an annual sample of approximately 10,000 randomly selected adults aged 18 and older from all five boroughs."[3] The data collected is also self-reported. It is important to know the origin of the data and what it means so that it can be properly analyzed.

The original structure of the data can be seen in *Figure 1*. There are a total of 194 columns for a set of variables in which the data has been collected, including the name of the district as well as the assigned identification number. Below is the original structure of the data. This definitely needed to be cleaned up, as it is difficult to read and see which columns were actually used in my analysis. The clean up process and condensed data will be explained and shown in the Analysis portion of the report.

[2] https://www.nyc.gov/site/doh/data/data-sets/data-sets-and-tables.page

[3] https://www.nyc.gov/site/doh/data/data-sets/community-health-survey.page



	ID	Name	OverallPopulation_rate	OverallPopulation_rank	Racewhite_Rate	Racewhite_rank	Raceblack_rate
7	101	Financial District	62829	57	66	9	4
8	102	Greenwich Village and Soho	91961	53	75	3	2
9	103	Lower East Side and Chinatown	168298	19	31	28	7
10	104	Clinton and Chelsea	106128	47	60	14	6
11	105	Midtown	52607	59	68	7	4
12	106	Stuyvesant Town and Turtle Bay	145147	28	72	4	4
13	107	Upper West Side	215329	4	67	8	7
14	108	Upper East Side	226640	3	79	2	3
15	109	Morningside Heights and Hamilton Heights	111645	44	22	35	25
16	110	Central Harlem	117943	39	10	44	62
17	111	East Harlem	123579	35	12	41	31
18	112	Washington Heights and Inwood	195302	9	17	39	7
19	201	Mott Haven and Melrose	94377	52	2	53	25
20	202	Hunts Point and Longwood	54069	58	1	57	22
21	203	Morrisania and Crotona	81698	56	1	58	38
22	204	Highbridge and Concourse	150599	26	1	54	32
23	205	Fordham and University Heights	131673	32	1	56	28
24	206	Belmont and East Tremont	85229	55	7	49	26
25	207	Kingsbridge Heights and Bedford	143515	29	7	46	18
26	208	Riverdale and Fieldston	103734	48	37	23	11
27	209	Parkchester and Soundview	177553	15	3	51	31
28	210	Throgs Neck and Co-op City	122309	36	33	27	22

Figure 1: Original Data Structure

Some exploratory data analysis that I conducted included some scatter plots and determining the correlation coefficients for each variable as compared to elementary school absenteeism. The summaries for the variables are below:

```
> summary(by_district$Schoolabsent_rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   4.00   11.00   17.00   19.08   28.00   40.00 
> summary(by_district$Sugary_Drink)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.00   22.00   28.00   27.78   34.00   42.00 
> summary(by_district$Fruit_Veg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  77.00   83.00   88.00   87.08   91.00   95.00 
> summary(by_district$Exercise)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  67.00   73.00   77.00   77.15   79.00   90.00 
> summary(by_district$Obesity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   8.00   19.00   26.00   24.17   31.00   35.00 
> summary(by_district$Diabetes)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3.00    8.50   11.00   10.68   14.00   18.00 
```

From this I looked at the correlation coefficients for the five variables. The coefficients by borough are:

```
> cor(boroughs$Schoolabsent_rate, boroughs$Sugary_Drink)
[1] 0.6733038
> cor(boroughs$Schoolabsent_rate, boroughs$Fruit_Veg)
[1] -0.8699417
> cor(boroughs$Schoolabsent_rate, boroughs$Exercise)
[1] -0.3817032
> cor(boroughs$Schoolabsent_rate, boroughs$Obesity)
[1] 0.6533631
> cor(boroughs$Schoolabsent_rate, boroughs$Diabetes)
[1] 0.7409624
```

And the coefficients by district are:

```
> cor(districts$Schoolabsent_rate, districts$Sugary_Drink)
[1] 0.7965064
> cor(districts$Schoolabsent_rate, districts$Fruit_Veg)
[1] -0.7946673
> cor(districts$Schoolabsent_rate, districts$Exercise)
[1] -0.3255877
> cor(districts$Schoolabsent_rate, districts$Obesity)
[1] 0.8029293
> cor(districts$Schoolabsent_rate, districts$Diabetes)
[1] 0.7447489
```

By looking at the coefficients, we can get an idea of whether or not the hypotheses were correct before completing any detailed analysis. When we look at the district data, we can see that fruit and vegetable consumption have a high negative correlation with school absences. Looking at the same variables for each district separately, the correlation is still negative, but not as high of a correlation. On the other hand, obesity seems to have a higher positive correlation with school absenteeism for the district data. It is important to look at both because the borough data is combined data for all the districts in each borough in New York City.

Analysis

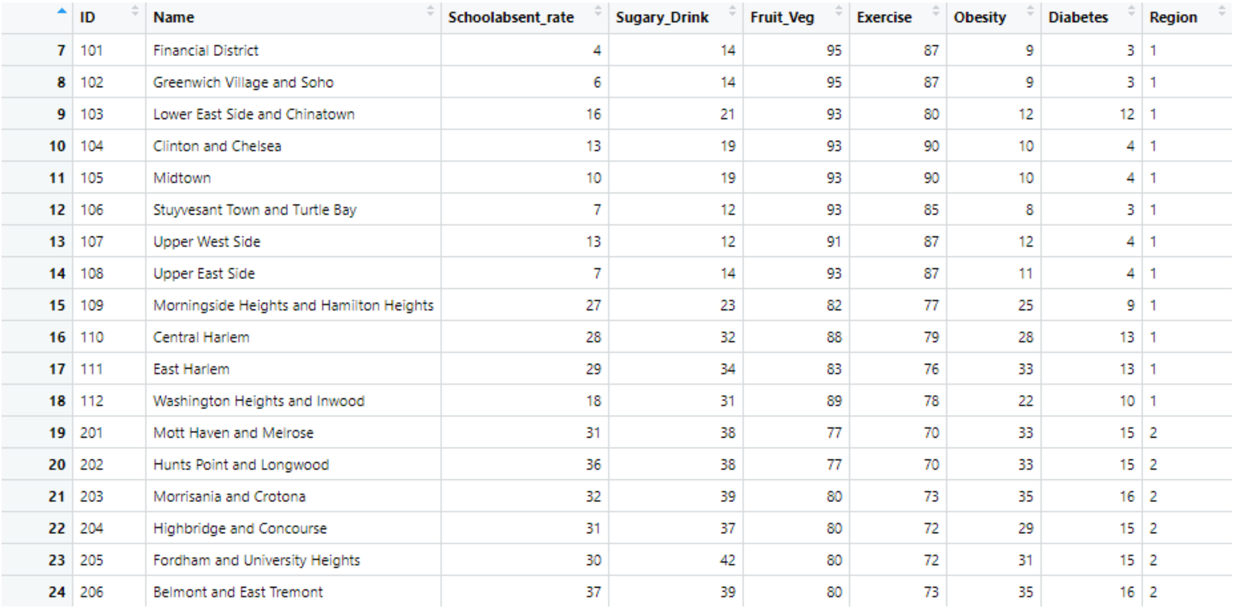
The first most important step of the analysis process is cleaning up the data for use. After reading in the comma-separated values (CSV) file, I separated the data in borough data and district data. The borough data was obtained by reading in the first six rows, which correlate to the five boroughs in New York City, plus compiled data for the whole city. The district data was the remaining rows, with one row per district. After properly dividing the data, I extracted the columns I wanted to use and removed the unnecessary ones. This was done for both sets of data. Finally, I added a column to the district data to identify which borough it was in, for future reference and coloring purposes. In this case, the boroughs were in order of: Manhattan, Bronx, Brooklyn, Queens, and Staten Island, each corresponding to a number from 1 to 5, respectively. The numbers were then assigned for the added borough column. After, I removed the last three rows of the data, as they were either completely null, or just notes on interpreting the information. There was no additional clean up necessary for the data, as there were no NA or null values for the specific columns I chose to use. In *Figure 2* is a sample of the cleaned up data by district. Below is the code written to read and clean up the data:

```r
# read data
by_borough <- read.csv(file="2015_CHP_PUD.csv", nrows=6)
by_district <- read.csv(file="2015_CHP_PUD.csv")

# remove borough and empty rows
by_district <- by_district[-(1:6),]
by_district <- head(by_district, -3)

# use specific columns
by_borough <- by_borough[c('ID', 'Name', 'Schoolabsent_rate', 'Sugary_Drink',
     'Fruit_Veg', 'Exercise', 'Obesity', 'Diabetes')]
by_district <- by_district[c('ID', 'Name', 'Schoolabsent_rate',
     'Sugary_Drink', 'Fruit_Veg', 'Exercise', 'Obesity', 'Diabetes')]

# add column for borough
by_district <- by_district%>% mutate(Borough = substr(ID, 0, 1))
```



	ID	Name	Schoolabsent_rate	Sugary_Drink	Fruit_Veg	Exercise	Obesity	Diabetes	Region
7	101	Financial District	4	14	95	87	9	3	1
8	102	Greenwich Village and Soho	6	14	95	87	9	3	1
9	103	Lower East Side and Chinatown	16	21	93	80	12	12	1
10	104	Clinton and Chelsea	13	19	93	90	10	4	1
11	105	Midtown	10	19	93	90	10	4	1
12	106	Stuyvesant Town and Turtle Bay	7	12	93	85	8	3	1
13	107	Upper West Side	13	12	91	87	12	4	1
14	108	Upper East Side	7	14	93	87	11	4	1
15	109	Morningside Heights and Hamilton Heights	27	23	82	77	25	9	1
16	110	Central Harlem	28	32	88	79	28	13	1
17	111	East Harlem	29	34	83	76	33	13	1
18	112	Washington Heights and Inwood	18	31	89	78	22	10	1
19	201	Mott Haven and Melrose	31	38	77	70	33	15	2
20	202	Hunts Point and Longwood	36	38	77	70	33	15	2
21	203	Morrisania and Crotona	32	39	80	73	35	16	2
22	204	Highbridge and Concourse	31	37	80	72	29	15	2
23	205	Fordham and University Heights	30	42	80	72	31	15	2
24	206	Belmont and East Tremont	37	39	80	73	35	16	2

Figure 2: Cleaned Up District Data

Next, I moved on to plotting the independent variables on a scatter plot with school absenteeism as the dependent variable. I grouped them by the predictions on which would have a positive and negative correlation with absences.

The first three variables plotted were: diabetes, obesity, and sugary drink consumption in adults. I had predicted that they would all have a positive correlation with the number of school absences, which is evident in *Figure 3* below. As we can see, even though all three variables do have a positive correlation, the diabetes correlation is less than that of the other two.

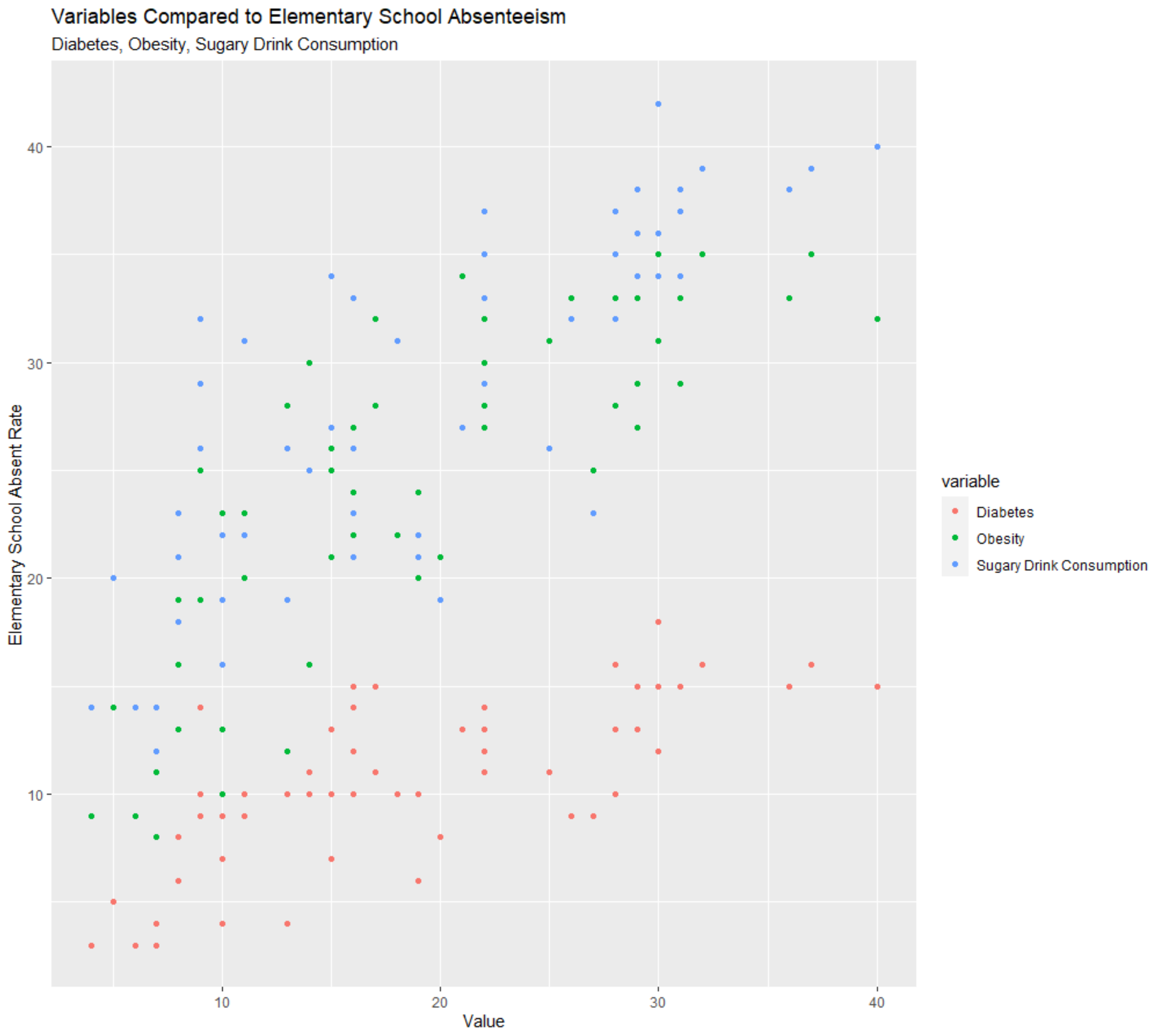


Figure 3: Diabetes, Obesity, Sugary Drink Consumption vs. Absenteeism

The next two variables plotted were: exercise and fruit and vegetable consumption. As opposed to the previous plot in *Figure 3*, exercise and fruit and vegetable consumption was predicted to have a negative correlation with school absences. Below is *Figure 4*, where it is evident that the correlation isn't as clear, but there are slight negative correlations for both variables. The data for exercise is more scattered than fruit and vegetable consumption when compared to the absenteeism value.

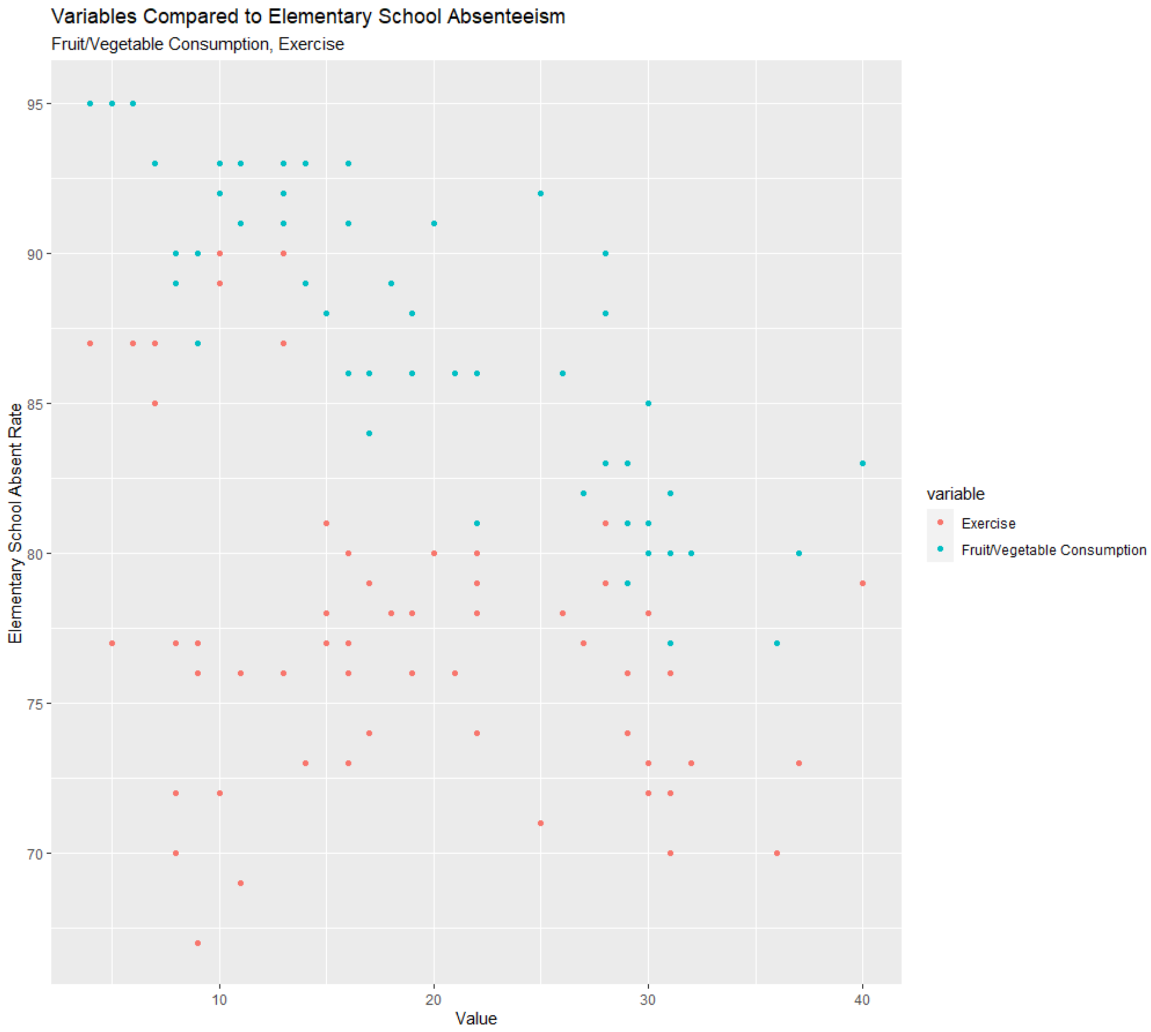
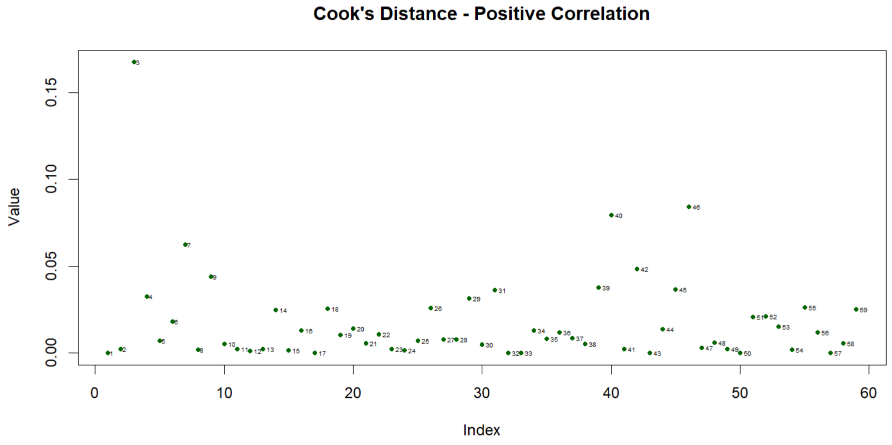


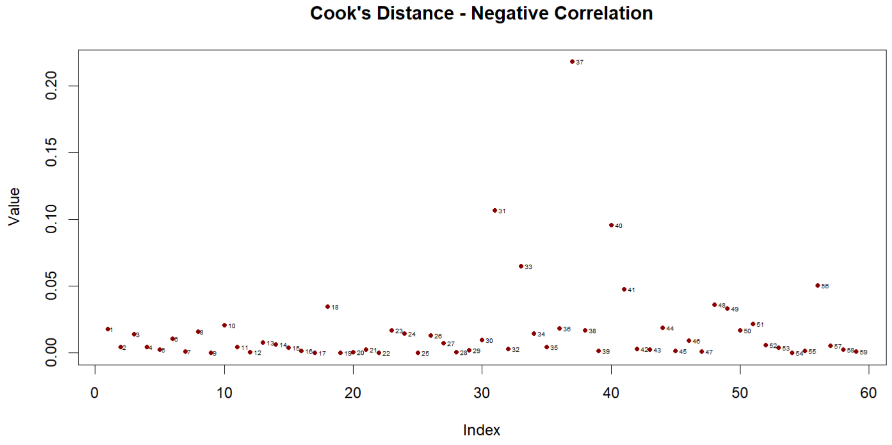
Figure 4: Exercise, Fruit/Vegetable Consumption vs. Absenteeism

Moving onto looking at the outliers and multivariate regression models, I looked at the variables predicted to have a positive and negative correlation separately, and then all of the variables together. The code for the multivariate regression models is as follows:

```
pos_cor <- lm(Schoolabsent_rate ~ Sugary_Drink + Obesity + Diabetes,
     data = by_district)
neg_cor <- lm(Schoolabsent_rate ~ Fruit_Veg + Exercise, data = by_district)
all_var <- lm(Schoolabsent_rate ~ Sugary_Drink + Fruit_Veg + Exercise +
     Obesity + Diabetes, data = by_district)
```

By creating these models, we can then use them for further data analysis. In this case, I used the models to look at the outliers using Cook's distance, `cooks.distance(model_name)`, for the three models addressed above.


Cook's Distance - Positive Correlation


Cook's Distance - Negative Correlation

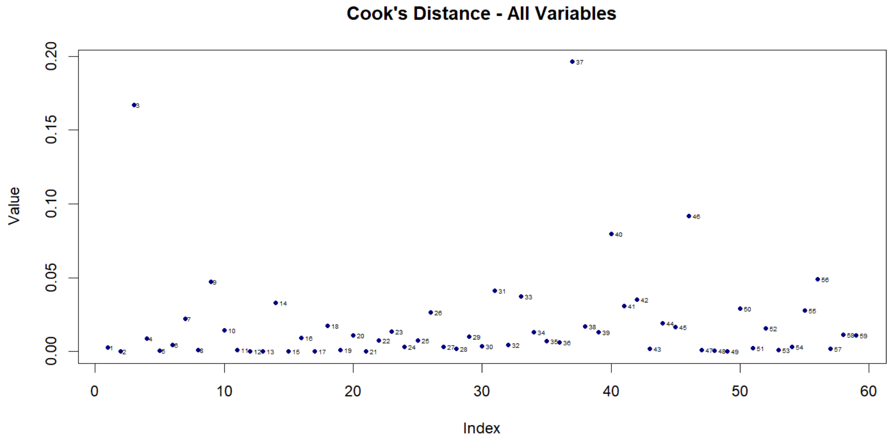

Cook's Distance - All Variables

Figure 5: Outliers - Positive (top), Negative (middle), All (bottom)

As can be seen in the three plots above, the specific districts that have outliers vary for the different variables. Looking at the outliers isn't as essential for model analysis, but more for exploring the data. If we look closely at the labeled data points, we can see that point 40, which is the district of Bay Ridge and Dyker Heights and derived from the 40th row, appears as an outlier on all three plots. Bay Ridge and Dyker Heights have a low absentee rate and a low diabetes rate, which are most likely the two variables contributing to the outliers on the plots, though every variable takes some part in the result. This was just one point that was interesting to mention because it appeared as an outlier on all three plots.

While exploring data, it is also important to look at any error, uncertainty, and bias factors. One factor for the Community Health Profiles data specifically is the fact that all of the data obtained from the Community Health Survey is a random subset of adults in New York City. This might make the data slightly inaccurate because we don't know the types of people surveyed. For example, although unlikely, there may be households who either don't have telephones to be pinged by the surveyors or who choose not to answer or even those who reported incorrect data because it is all self-reported. This may increase uncertainty as a whole. In terms of bias, there shouldn't be any bias because it is a randomized set of 10,000 individuals. Finally, for error, there is technically sampling error due to the fact that it is a sample of the whole population. Even by looking at the combined borough data and the district data, we can see there is some discrepancy. The non-sampling error would exist from the idea that the survey is targeted towards those who own landlines or mobile cell phones, as well as from self-reported data.

Model Development and Application of Models

The types of models developed in this project were: (1) *k-means clustering*, (2) *multivariate regression*, and (3) *decision tree*.

Multivariate Regression

As mentioned previously, the multivariate regression model was created to perform other types of operations with the data. It also helps to see any trend lines for the different variables and determine correlation. Below is the code for the multivariate regression models (which was also stated in the analysis portion of the report), including the summaries for each model:

```
# predicted positive correlation - sugary drink, obesity, diabetes
pos_cor <- lm(Schoolabsent_rate ~ Sugary_Drink + Obesity + Diabetes, data =
by_district)

#predicted negative correlation - fruit/veg, exercise
neg_cor <- lm(Schoolabsent_rate ~ Fruit_Veg + Exercise, data = by_district)

# all variables
all_var <- lm(Schoolabsent_rate ~ Sugary_Drink + Fruit_Veg + Exercise + Obesity +
Diabetes, data = by_district)
```

The summaries of the models with coefficients are as follows:

```
> summary(pos_cor)
Call:
lm(formula = Schoolabsent_rate ~ Sugary_Drink + Obesity + Diabetes,
    data = new_district)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5515  -3.9227   0.3591   3.4120  10.8172

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.22367    2.53802  -2.846  0.00621 **
Sugary_Drink  0.49338    0.18114   2.724  0.00863 **
Obesity       0.52868    0.16893   3.130  0.00280 **
Diabetes     -0.01641    0.38365  -0.043  0.96603
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 5.172 on 55 degrees of freedom
Multiple R-squared:  0.7011,     Adjusted R-squared:  0.6848 
F-statistic:    43 on 3 and 55 DF,  p-value: 1.909e-14

> summary(neg_cor)
Call:
lm(formula = Schoolabsent_rate ~ Fruit_Veg + Exercise, data = new_district)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1987  -3.6587   0.2754   2.6120  14.6022 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 143.28305   14.11434  10.152 2.64e-14 ***
Fruit_Veg    -1.50518    0.16764  -8.979 1.95e-12 ***
Exercise      0.08918    0.15896   0.561    0.577    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.676 on 56 degrees of freedom
Multiple R-squared:  0.6336,     Adjusted R-squared:  0.6205 
F-statistic: 48.41 on 2 and 56 DF,  p-value: 6.198e-13

> summary(all_var)
Call:
lm(formula = Schoolabsent_rate ~ Sugary_Drink + Fruit_Veg + Exercise + 
    Obesity + Diabetes, data = new_district)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0640  -3.0569  -0.0959   2.5200   9.4634 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)   22.1406    23.0788   0.959  0.34174   
Sugary_Drink   0.3543     0.1607   2.204  0.03186 * 
Fruit_Veg     -0.6531     0.1928  -3.387  0.00134 **
Exercise       0.4188     0.1415   2.959  0.00460 **
Obesity        0.4125     0.1554   2.655  0.01046 * 
Diabetes       0.1589     0.3481   0.457  0.64983   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.475 on 53 degrees of freedom
Multiple R-squared:  0.7844,     Adjusted R-squared:  0.7641 
F-statistic: 38.56 on 5 and 53 DF,  p-value: < 2.2e-16
```

When looking directly at the intercepts and coefficients, it is difficult to understand what they mean. Therefore, it is also essential to plot the data and regression lines. By plotting each variable with its regression line separately, we can clearly see the correlation for that specific variable. If we look at the summaries of each of these models above, we can see that the p-values for all three models is very small, which shows that they are statistically significant. It is essential to look at the p-values because it tells the statistical significance of the data in regards to the null hypothesis. The summary is also useful because it gives the coefficients of each of the variables, which is important for multivariate regression.

Then, I decided to look at the linear regression for each of the variables chosen. The first thing done was to do linear regression for each variable and looking at the summaries. It is important to see the *p*-values first to indicate whether or not the variable is statistically significant in relation to the school absentee rate. We obtain *p*-values of: `4.589e-14`, `2.021e-14`, `1.356e-11`, `5.772e-14`, and `0.01186`, respectively. All of the *p*-values are less than 0.05, which means that all of the variables are statistically significant and can be used to determine correlation.

```
# linear regression models
sugar <- lm(Schoolabsent_rate ~ Sugary_Drink, data = new_district)
summary(sugar)
obesity <- lm(Schoolabsent_rate ~ Obesity, data = new_district)
summary(obesity)
diabetes <- lm(Schoolabsent_rate ~ Diabetes, data = new_district)
summary(diabetes)
fruitveg <- lm(Schoolabsent_rate ~ Fruit_Veg, data = new_district)
summary(fruitveg)
exercise <- lm(Schoolabsent_rate ~ Exercise, data = new_district)
summary(exercise)
```

Below, you can see the scatterplots colored by borough along with the linear regression line for each borough. For this, I chose to look at sugary drink consumption and fruit and vegetable consumption specifically, because these had low *p*-values, and I predicted that they would affect the data the most. With this, the intention was to see the difference in regression for each of the boroughs, and whether one had a stronger correlation than the other. Because it was difficult to see the other lines with the confidence intervals, I did not include the confidence interval along with the regression lines.

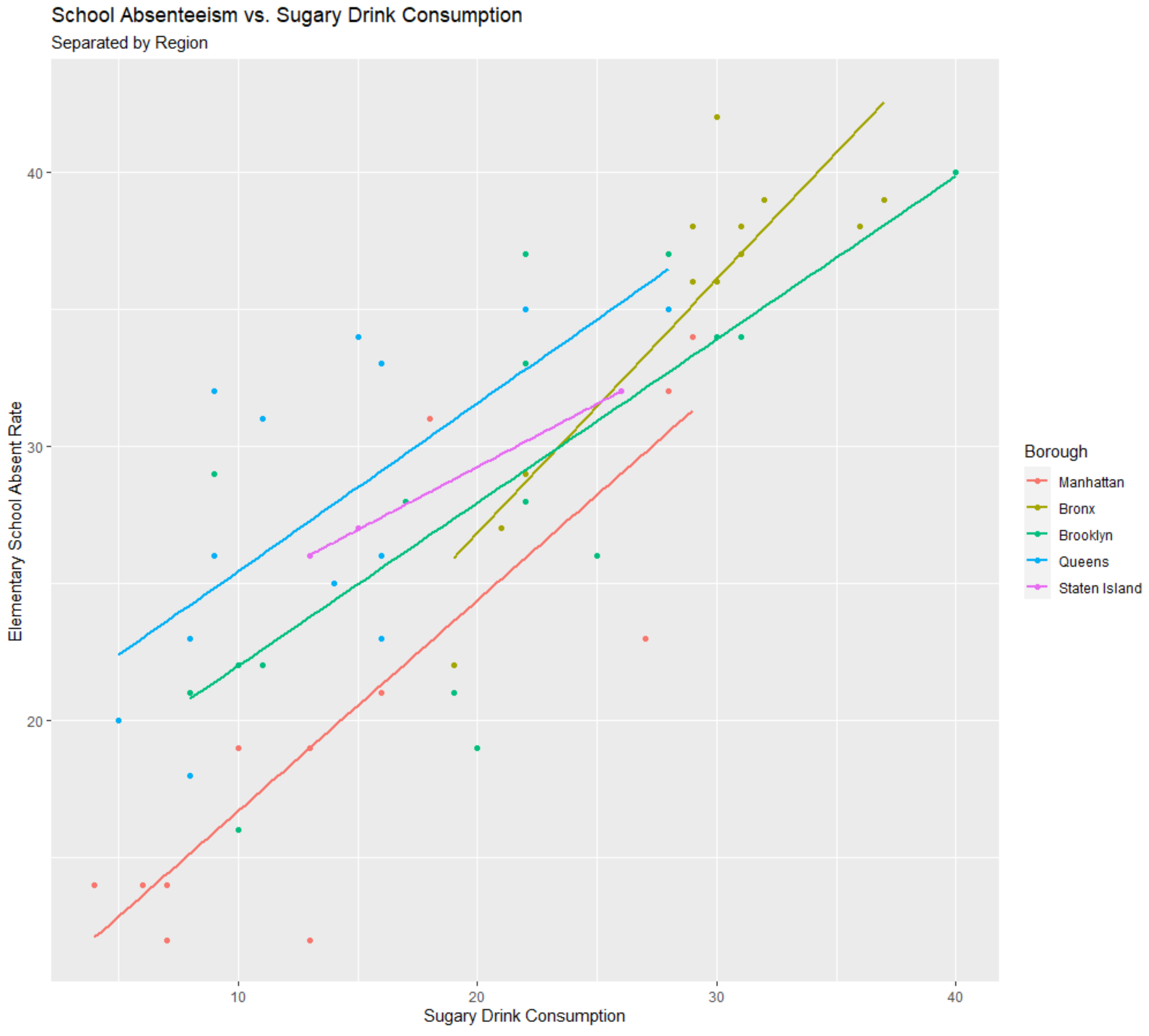


Figure 6: School Absenteeism vs. Sugary Drink Consumption

If we look closer at the data in *Figure 6*, we can see that Staten Island has the smallest slope, meaning it has the lowest positive correlation. It is a bit difficult to tell which has the highest correlation from the remaining four boroughs, but it looks as if Bronx has the highest positive correlation. However, the regression lines may also be affected by the number of points for each borough, as Staten Island has the least number of points at just two.

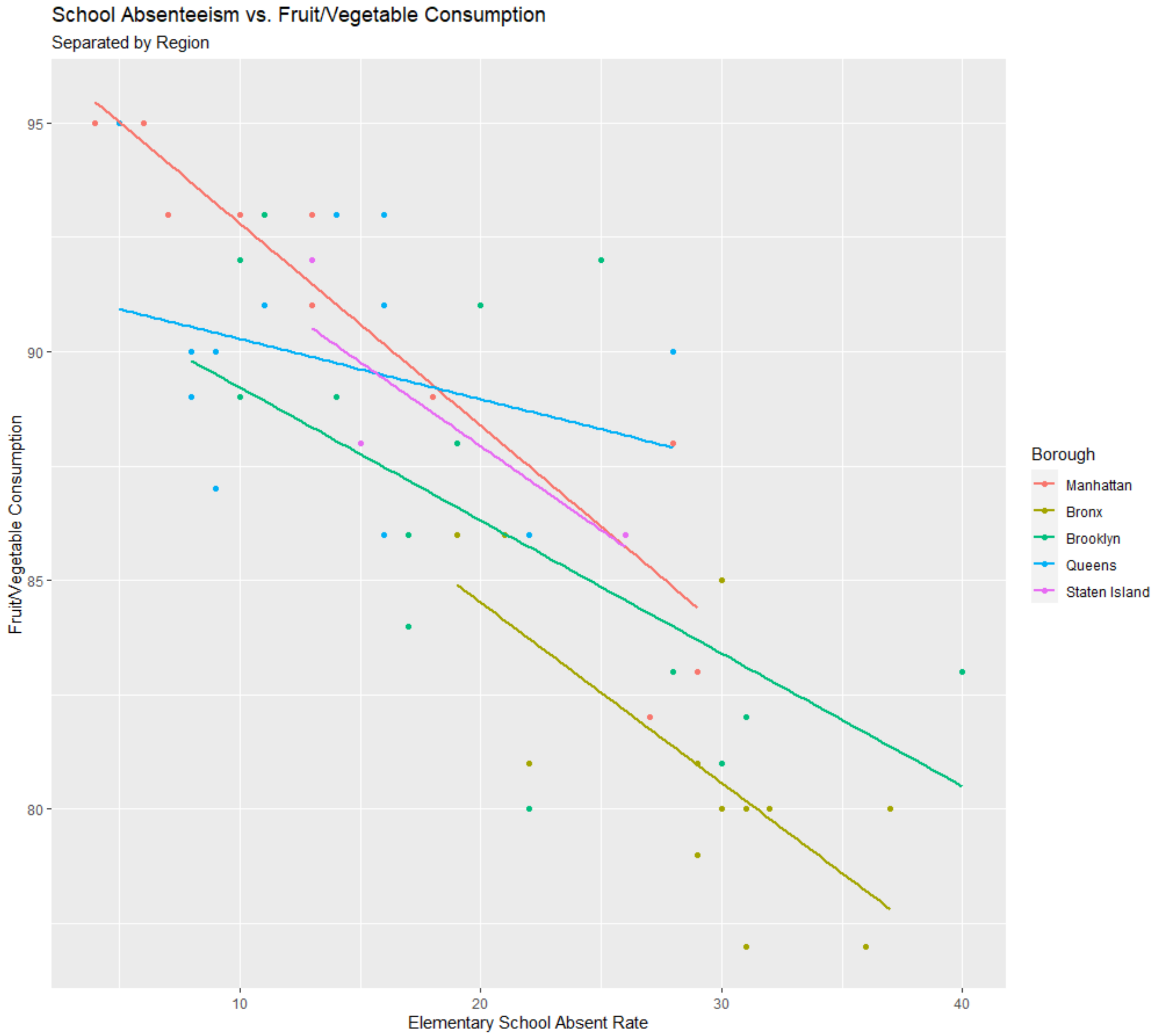


Figure 7: School Absenteeism vs. Fruit/Vegetable Consumption

As we look at *Figure 7* above, we can see that for each borough there is a negative correlation between the two variables. For fruit and vegetable consumption, Brooklyn appears to have the lowest negative correlation, with Manhattan appearing to have the highest negative correlation.

We know that both sugary drink consumption and fruit and vegetable consumption have a significant impact on the elementary school absent rate from the *p*-values determined earlier in the linear regression model analysis.

K-Means Clustering

The second model I looked into was k-means clustering. I thought this would be helpful to display the data for each district by clustering the boroughs to see if there was any separation or distinction between the boroughs. The clustering ellipses would also be helpful in the future, given a specific data point, we might be able to work from that and associate the point with a borough. As we obtain more data points, the model will get better at predicting either the borough that the point is in, or the absentee rate given the rest of the data.

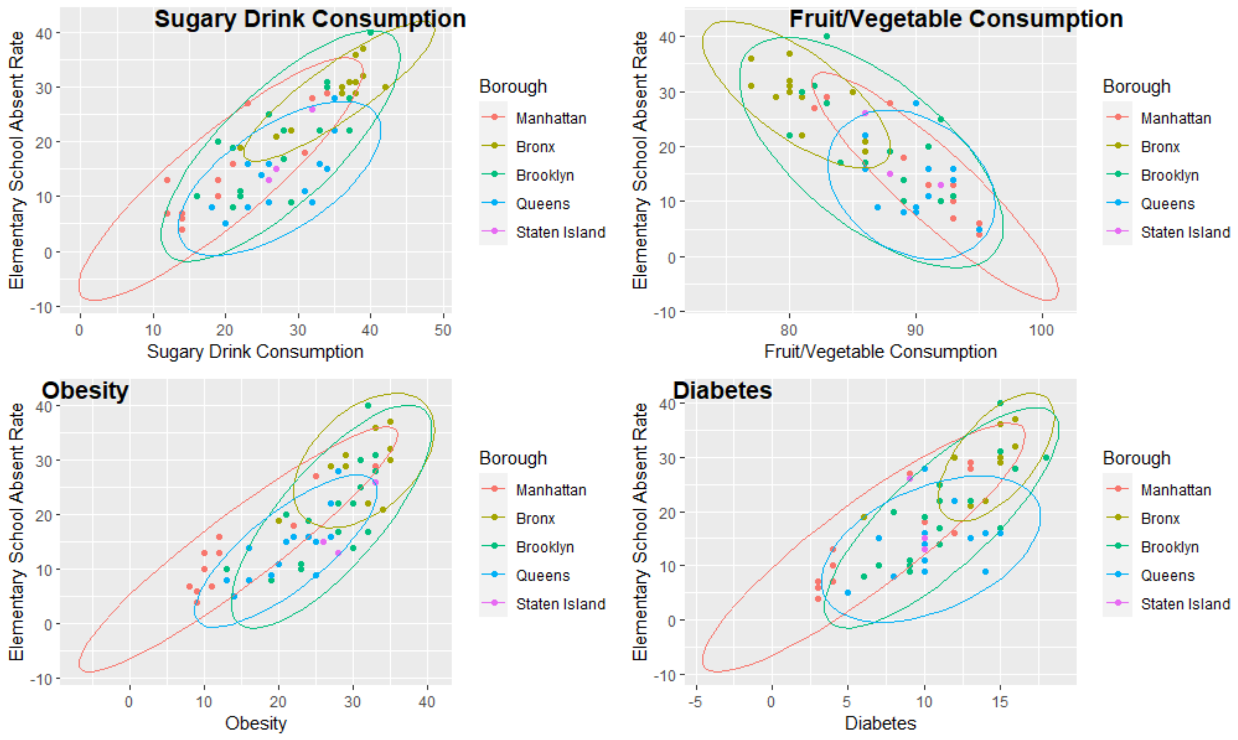


Figure 8a: K-Means Clustering Per Variable, Separated by Borough

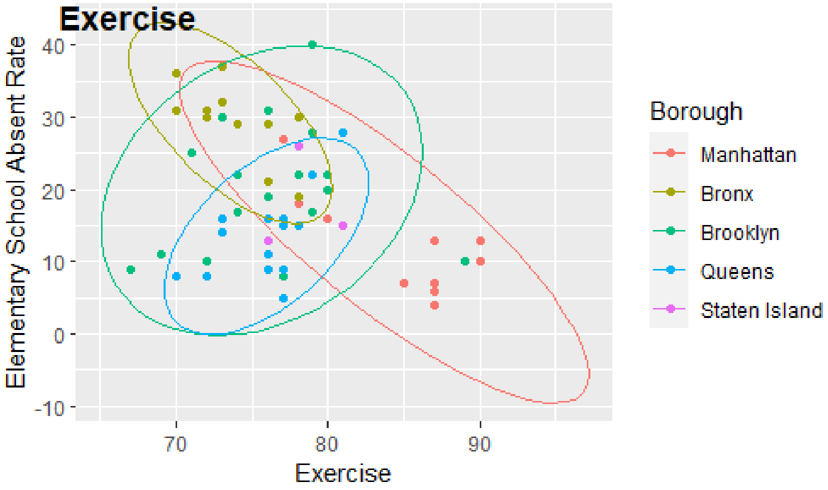


Figure 8b: K-Means Clustering Per Variable, Separated by Borough

The distribution in the five plots above are interesting. For the first four, sugary drink consumption, fruit and vegetable consumption, obesity, and diabetes, each borough follows the same trend - either positive or negative correlation. When we look at the exercise plot, however, Manhattan and Bronx have a negative correlation, while Brooklyn and Queens have a slight positive correlation or no correlation at all. It is interesting to see the distribution of points as well. For most of the Queens and Brooklyn data points, they tend to be more precise, with points falling closer together, and the cluster ellipse being closer to a circle than the other ones. On the other hand, the Manhattan data points are very spread apart, making it more of an ellipse. This would most likely be caused by having more data points overall, as well as potentially the region size. As you can see in the plots, there is no cluster for Staten Island. This is because there were too few points to create one. The code for the above plots is below:

```
kmeans_sugarydrink <- ggplot(new_district, aes(x = Sugary_Drink, y =
Schoolabsent_rate, col = Borough)) + geom_point() + stat_ellipse() +
  scale_color_discrete(labels=c('Manhattan', 'Bronx', 'Brooklyn', 'Queens', 'Staten
Island')) +
  ylab("Elementary School Absent Rate") +
  xlab("Sugary Drink Consumption")
```

```
kmeans_fruitveg <- ggplot(new_district, aes(x = Fruit_Veg, y = Schoolabsent_rate, col
= Borough)) + geom_point() + stat_ellipse() +
  scale_color_discrete(labels=c('Manhattan', 'Bronx', 'Brooklyn', 'Queens', 'Staten
Island')) +
  ylab("Elementary School Absent Rate") +
  xlab("Fruit/Vegetable Consumption")

kmeans_obesity <- ggplot(new_district, aes(x = Obesity, y = Schoolabsent_rate, col =
Borough)) + geom_point() + stat_ellipse() +
  scale_color_discrete(labels=c('Manhattan', 'Bronx', 'Brooklyn', 'Queens', 'Staten
Island')) +
  ylab("Elementary School Absent Rate") +
  xlab("Obesity")

kmeans_diabetes <- ggplot(new_district, aes(x = Diabetes, y = Schoolabsent_rate, col
= Borough)) + geom_point() + stat_ellipse() +
  scale_color_discrete(labels=c('Manhattan', 'Bronx', 'Brooklyn', 'Queens', 'Staten
Island')) +
  ylab("Elementary School Absent Rate") +
  xlab("Diabetes")

kmeans_exercise <- ggplot(new_district, aes(x = Exercise, y = Schoolabsent_rate, col
= Borough)) + geom_point() + stat_ellipse() +
  scale_color_discrete(labels=c('Manhattan', 'Bronx', 'Brooklyn', 'Queens', 'Staten
Island')) +
  ylab("Elementary School Absent Rate") +
  xlab("Exercise")

figure <- ggarrange(kmeans_sugarydrink, kmeans_fruitveg, kmeans_obesity,
kmeans_diabetes, kmeans_exercise, labels = c("Sugary Drink Consumption",
"Fruit/Vegetable Consumption", "Obesity", "Diabetes", "Exercise"), ncol = 2, nrow =
3)
```

In addition to clustering based on the borough, I also did clustering based on the level of absenteeism. The first thing I did was to split up the absentee rate into three sections. To do this, I took a look at the summary, specifically the minimum and maximum values, and divided that into three even sections - low, moderate, and high levels of absenteeism. With a minimum of 4 and a maximum of 40, I broke the sections up into groups of twelve and created a new column within the data frame accordingly.

```
summary(new_district$Schoolabsent_rate)
# min = 4, max = 40 -> 4:15, 16:28, 29:40
```

```
new_district$absent_distribution <- cut(new_district$Schoolabsent_rate, br=c(0, 15,
28, 40), labels=c("Low", "Moderate", "High"))
```

After labeling the data appropriately, I checked to make sure the column was added to the data frame. From this, I then decided to plot the Sugary Drink Consumption on the x-axis and then the Obesity rate on the y-axis, with the coloring as the newly created column for absentee level. The cluster ellipses were also drawn around. The purpose of creating this graph was to see if these two variables together had any correlation with the absentee level, which, as can be seen in *Figure 9* below, does have some correlation. The data points with a high absentee level according to the minimum and maximum, calculated earlier, do appear on the upper right side of the graph, showing that the obesity and sugary drink consumption correlate with the absentee rate.

```
# k means with absentee levels
absent_kmeans <- ggplot(new_district, aes(x=Sugary_Drink, y=Obesity, col =
absent_distribution)) + geom_point() + stat_ellipse() +
  scale_color_discrete(labels=c('Low', 'Moderate', 'High')) +
  ylab("Obesity") +
  xlab("Sugary Drink Consumption") +
  labs(title="Sugary Drink Consumption vs. Obesity", subtitle="Separated by Absentee
Level")
absent_kmeans
```

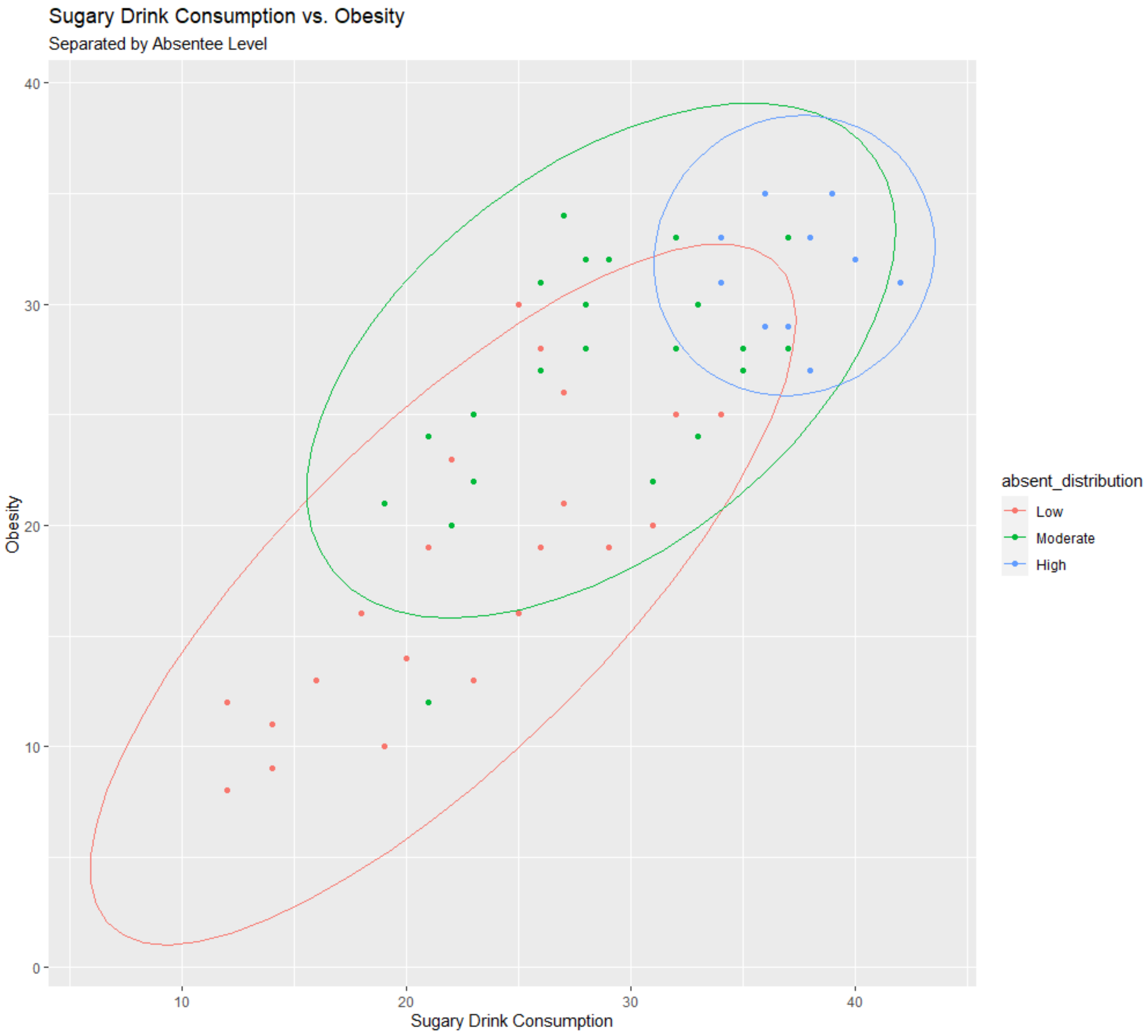


Figure 9: K-Means Clustering by Absentee Level

Decision Tree

For the decision tree, I used the new column created above. With this column, I was able to plot a decision tree based on the five variables mentioned, and see the different leaves in the tree. By running just `rpart()`, only two of the five variables appeared to be used. Due to this, I used the control parameter in `rpart()`, with `maxdepth = 30`, `minsplit = 1`, `minbucket = 1`, and `cp = 0`, as seen in the code snippets below.

```
# simple
rpart_data <- rpart(absent_distribution ~ Sugary_Drink + Fruit_Veg + Obesity +
Diabetes + Exercise, data=new_district, method="class")
rpart_data
summary(rpart_data)
rpart.plot(rpart_data)

# all variables
rpart_data2 <- rpart(absent_distribution ~ Sugary_Drink + Fruit_Veg + Obesity +
Diabetes + Exercise, data=new_district, control = rpart.control(maxdepth = 30,
minsplit = 1, minbucket = 1, cp=0))
rpart_data2
summary(rpart_data2)
rpart.plot(rpart_data2)
```

The results of the decision tree are as follows:

```
> rpart_data
n= 59

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 59 35 Low (0.4067797 0.3728814 0.2203390)
  2) Fruit_Veg>=86.5 33  9 Low (0.7272727 0.2727273 0.0000000) *
  3) Fruit_Veg< 86.5 26 13 Moderate (0.0000000 0.5000000 0.5000000)
    6) Sugary_Drink< 33.5 10  0 Moderate (0.0000000 1.0000000 0.0000000) *
    7) Sugary_Drink>=33.5 16  3 High (0.0000000 0.1875000 0.8125000) *
```

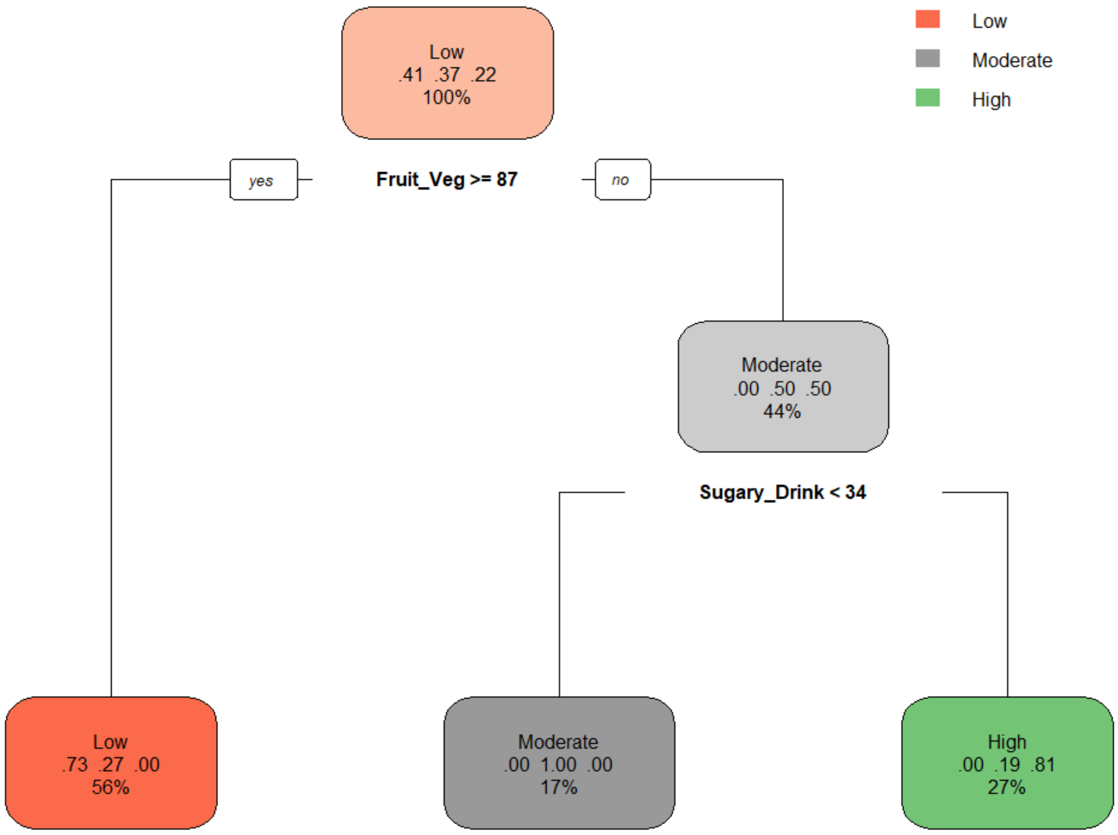


Figure 10: Decision Tree (no modifications)

```
> rpart_data2
n= 59

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 59 35 Low (0.40677966 0.37288136 0.22033898)
    2) Fruit_Veg>=86.5 33  9 Low (0.72727273 0.27272727 0.00000000)
      4) Obesity< 20.5 17  1 Low (0.94117647 0.05882353 0.00000000)
        8) Diabetes< 11 16  0 Low (1.00000000 0.00000000 0.00000000) *
        9) Diabetes>=11 1  0 Moderate (0.00000000 1.00000000 0.00000000) *
      5) Obesity>=20.5 16  8 Low (0.50000000 0.50000000 0.00000000)
       10) Sugary_Drink>=21.5 14  6 Low (0.57142857 0.42857143 0.00000000)
         20) Diabetes< 9.5 3  0 Low (1.00000000 0.00000000 0.00000000) *
         21) Diabetes>=9.5 11  5 Moderate (0.45454545 0.54545455 0.00000000)
           42) Obesity>=24.5 8  3 Low (0.62500000 0.37500000 0.00000000)
             84) Obesity< 27 3  0 Low (1.00000000 0.00000000 0.00000000) *
             85) Obesity>=27 5  2 Moderate (0.40000000 0.60000000 0.00000000)
              170) Sugary_Drink< 29 3  1 Low (0.66666667 0.33333333 0.00000000)
                340) Obesity< 30.5 2  0 Low (1.00000000 0.00000000 0.00000000) *
                341) Obesity>=30.5 1  0 Moderate (0.00000000 1.00000000 0.00000000) *
              171) Sugary_Drink>=29 2  0 Moderate (0.00000000 1.00000000 0.00000000)
*
           43) Obesity< 24.5 3  0 Moderate (0.00000000 1.00000000 0.00000000) *
       11) Sugary_Drink< 21.5 2  0 Moderate (0.00000000 1.00000000 0.00000000) *
    3) Fruit_Veg< 86.5 26 13 Moderate (0.00000000 0.50000000 0.50000000)
      6) Sugary_Drink< 33.5 10  0 Moderate (0.00000000 1.00000000 0.00000000) *
      7) Sugary_Drink>=33.5 16  3 High (0.00000000 0.18750000 0.81250000)
       14) Fruit_Veg>=85.5 2  0 Moderate (0.00000000 1.00000000 0.00000000) *
       15) Fruit_Veg< 85.5 14  1 High (0.00000000 0.07142857 0.92857143)
         30) Exercise>=78.5 2  1 Moderate (0.00000000 0.50000000 0.50000000)
           60) Sugary_Drink< 38.5 1  0 Moderate (0.00000000 1.00000000 0.00000000) *
           61) Sugary_Drink>=38.5 1  0 High (0.00000000 0.00000000 1.00000000) *
         31) Exercise< 78.5 12  0 High (0.00000000 0.00000000 1.00000000) *
```

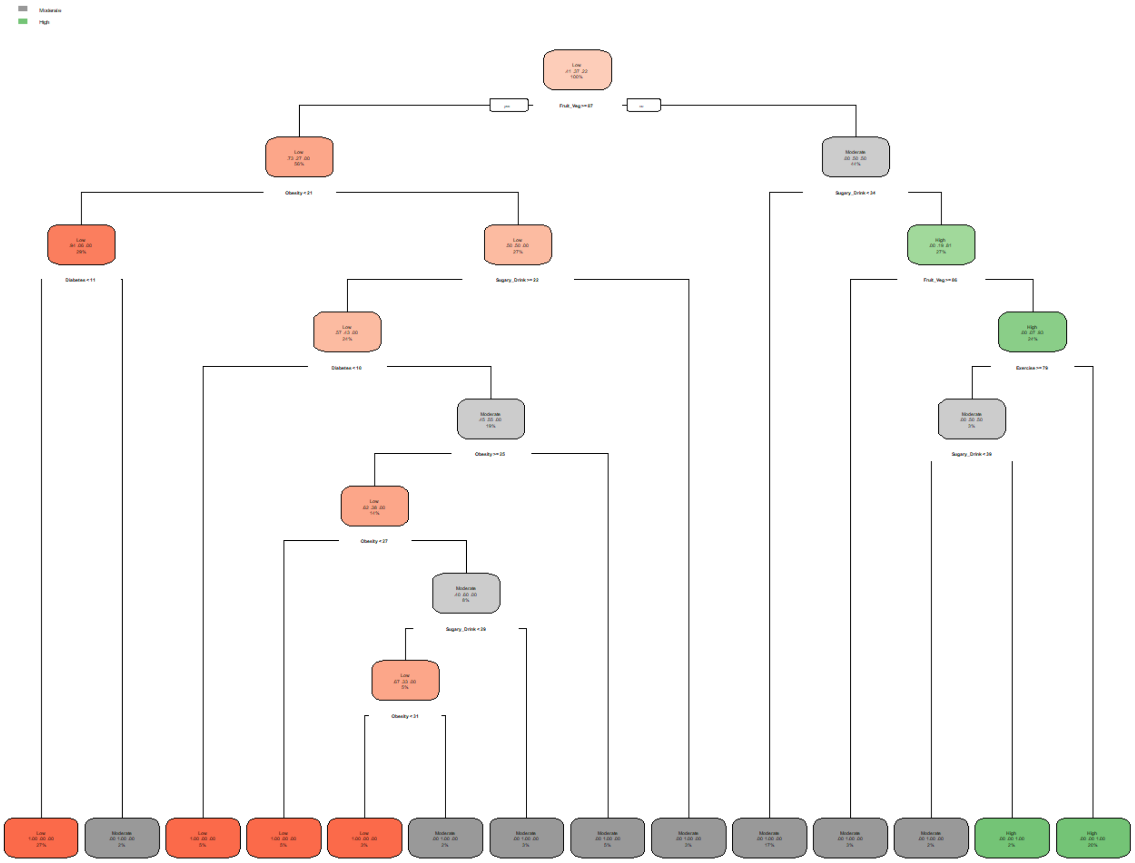


Figure 11: Decision Tree (modified)

Although this plot is a little difficult to see in the written document (above), it is attached in the `/plots` folder in the zip attached with the assignment. It is pretty simple to understand, as the data was fairly straightforward. The tree starts at the root, then looks at fruit and vegetable consumption, depending on if the result was yes or no, it then looks at obesity and sugary drink consumption. The rest of the tree can also be visualized in the summary above *Figure 11*.

Random Forest

Finally, the last model created was the random forest model. I chose to add this in because the decision tree was a little too simple, and the clustering wasn't as helpful as I had expected it to be. The code is as follows:

```
set.seed(324)
rf_train <- sample(nrow(new_district), 0.7 * nrow(new_district), replace = FALSE)
rf_train_set <- new_district[rf_train, ]
rf_valid_set <- new_district[-rf_train, ]
```

First, I set the seed so that the results use the same sets of data every time I run it. Then, the `sample()` function is used to get a sample set of the data for training and testing. Three iterations of the `randomForest()` function were run: **(rf1)** `ntree = 500`, `mtry = 2`, **(rf2)** `ntree = 350`, `mtry = 2`, and **(rf3)** `ntree = 80`, `mtry = 3`.

```
rf1:

OOB estimate of  error rate: 31.71%
Confusion matrix:
         Low Moderate High class.error
Low       12        5    0   0.2941176
Moderate   5        7    1   0.4615385
High       0        2    9   0.1818182
```

```
rf2:

OOB estimate of  error rate: 34.15%
Confusion matrix:
         Low Moderate High class.error
Low       12        5    0   0.2941176
Moderate   5        7    1   0.4615385
High       0        3    8   0.2727273
```

rf3:

```
OOB estimate of  error rate: 34.15%
Confusion matrix:
         Low Moderate High class.error
Low       13        4    0   0.2352941
Moderate   6        6    1   0.5384615
High       0        3    8   0.2727273
```

Although unexpected, the random forest for the original values, with 500 trees and 2 variables at each split, was the best in terms of error. `rf2` and `rf3` both had an out of bounds error rate of 34.15%, while `rf1` only had 31.71%. In all three of the confusion matrices, it is evident that the model is able to predict the low and high absentee rates well, but not the moderate ones. This could be due to the fact that many points fall in the moderate category, thus skewing the predictions because the points are not precise within that category. This model is not the best fit and should likely not be used to predict absentee levels.

Conclusions and Discussions

When looking at the regression lines, this data on which boroughs have a higher correlation for any of the variables will allow the New York City government and education departments to look further into why this is and make any necessary adjustments. It also is beneficial in just seeing the actual correlation between the variables, and was helpful in determining the next steps in the process, whether it was creating different types of models or choosing new variables. The k-means clustering wasn't very well thought out, and didn't really provide much information regarding the variables. The k-means plots in *Figure 8* were mainly for seeing the relationships between the variables and absentee rate in the different boroughs, and not so much for predicting data. The second k-means plot was a bit more useful, as it was clear that sugary drink consumption and obesity rates did somewhat correspond to the absentee level assigned by creating three equal levels, low, moderate, and high, using the minimum and maximum absentee rates. The decision tree was not too helpful, but when the values were modified, it was more complicated but used every variable. This might be worthwhile in looking at in the future, but for the given data, it wasn't necessary. The plot also became difficult to see with so many branches and leaves in the tree, so if an individual wanted to see a visual representation of the data, it would be difficult to see. This could also be fixed with some adjustments in R. Finally, the random forest was interesting to see the confusion matrix and see the error percentages for different random forest configurations. It was fascinating to see that the model predicted the low and high absentee rates correctly two times more than it did for the moderate absentee rates.

Choosing to look at k-means and the decision tree was not the best option for models, but through it, I learned that certain sets of data fit better for certain types of models. In addition, in a subsequent exploration, it would be helpful to dive deeper into how each variable has a correlation, especially for the random forest models.

All in all, the hypothesis that the consumption of sugary drinks, obesity, and diabetes would all have a positive correlation with school absences was incorrect. Instead, only the consumption of sugary drinks and the obesity rate had a direct positive correlation. The prediction that the consumption of fruits and vegetables and physical activity would have a negative correlation was also wrong - only the consumption of fruits and vegetables had a direct negative correlation. Only three of the five variables tested had direct correlations. Therefore, it can be concluded that regions with adults who have poorer health and nutrition choices do not necessarily have children with more school absences, but it may have some relation with specific variables only.

References

Alexina Cather, MPH. “The Impact of Food on Academic Behavior, Attendance, and Performance.” *NYC Food Policy Center (Hunter College)*, 18 Mar. 2021, https://www.nycfoodpolicy.org/resource-and-guide-the-impact-of-food-on-academic-behavior-attendance-performance-and-attrition/.

“Community Health Survey.” *Community Health Survey - NYC Health*, https://www.nyc.gov/site/doh/data/data-sets/community-health-survey.page.

Coughenour, Courtney, et al. “School Absenteeism Is Linked to Household Food Insecurity in School Catchment Areas in Southern Nevada: Public Health Nutrition.” *Cambridge Core*, Cambridge University Press, 15 Feb. 2021, https://www.cambridge.org/core/journals/public-health-nutrition/article/school-absenteeism-is-linked-to-household-food-insecurity-in-school-catchment-areas-in-southern-nevada/78A2BCD58D4E3C5BF1BAB724FB40B212.

“Datasets.” *Datasets - NYC Health*, https://www.nyc.gov/site/doh/data/data-sets/data-sets-and-tables.page.

“The Link Between School Attendance and Good Health.” *Publications.aap.org*, https://publications.aap.org/pediatrics/article/143/2/e20183648/37326/The-Link-Between-School-Attendance-and-Good-Health?autologincheck=redirected%3FnfToken.

Rodríguez-Escobar G;Vargas-Cruz SL;Ibáñez-Pinilla E;Matiz-Salazar MI;Jörgen-Overgaard H; “[Relationship between Nutritional Status and School Absenteeism among Students in Rural Schools].” *Revista De Salud Publica (Bogota, Colombia)*, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/28453140/.