The Effects of Physical Health on Elementary School Absenteeism

Gwyneth Yuen (yueng@rpi.edu), Rensselaer Polytechnic Institute, Troy, NY, United States

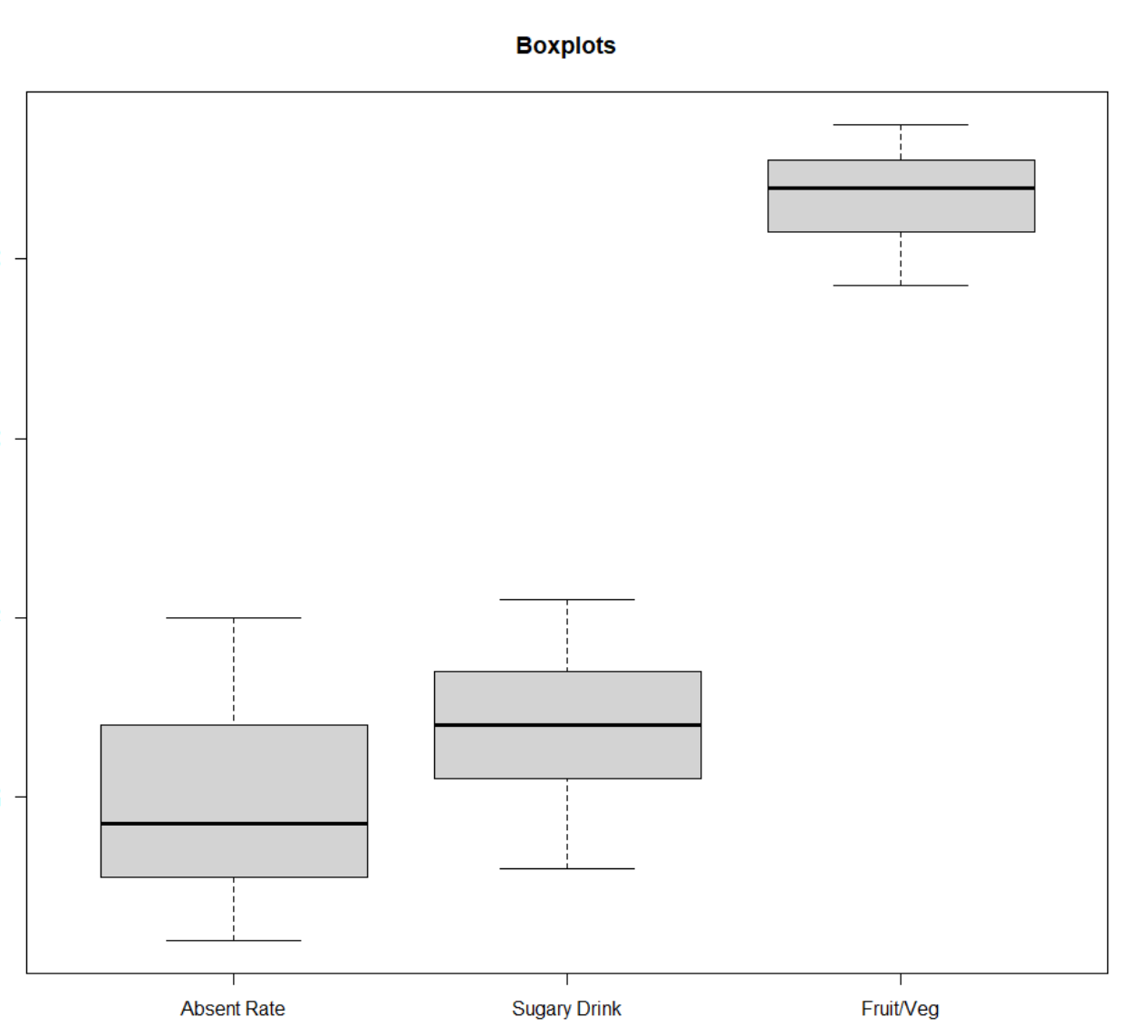
Abstract & Motivation

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. 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 provide healthier meals for lunch or extend any existing recess or physical education time during school hours that will benefit the children.

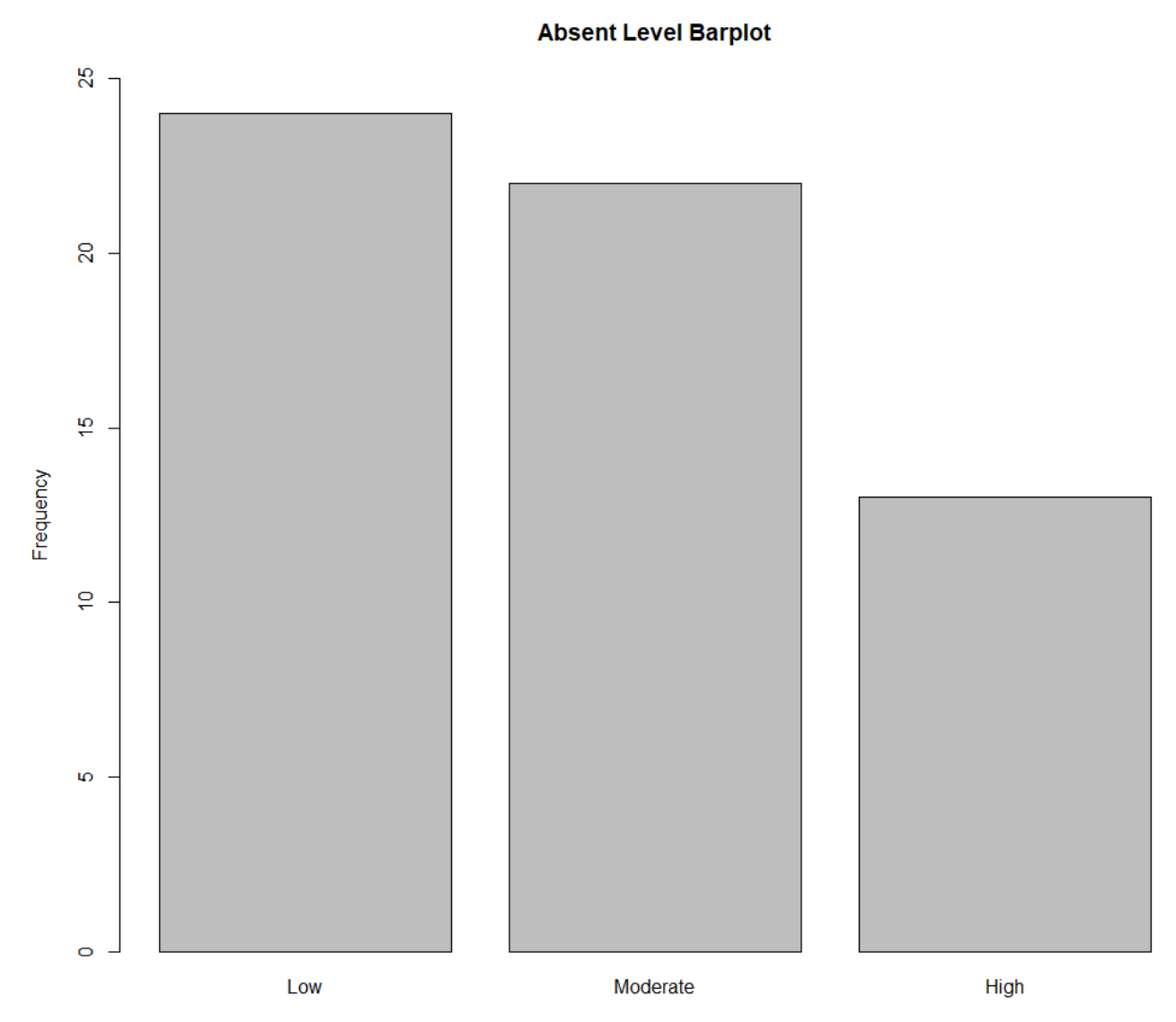
According to a study done by Cambridge University, "food insecurity, poverty and family variables are also relevant indicators for chronic school absenteeism." This brings up the question: Do regions with adults who have poorer health and nutrition choices also have children with more school absences, and what is the correlation? The predictions I made were that: (1) The consumption of sugary drinks, obesity, and diabetes would all have a positive correlation with school absences, and (2) the consumption of fruits and vegetables and physical activity will have a negative correlation with school absences.

Exploratory Data Analysis

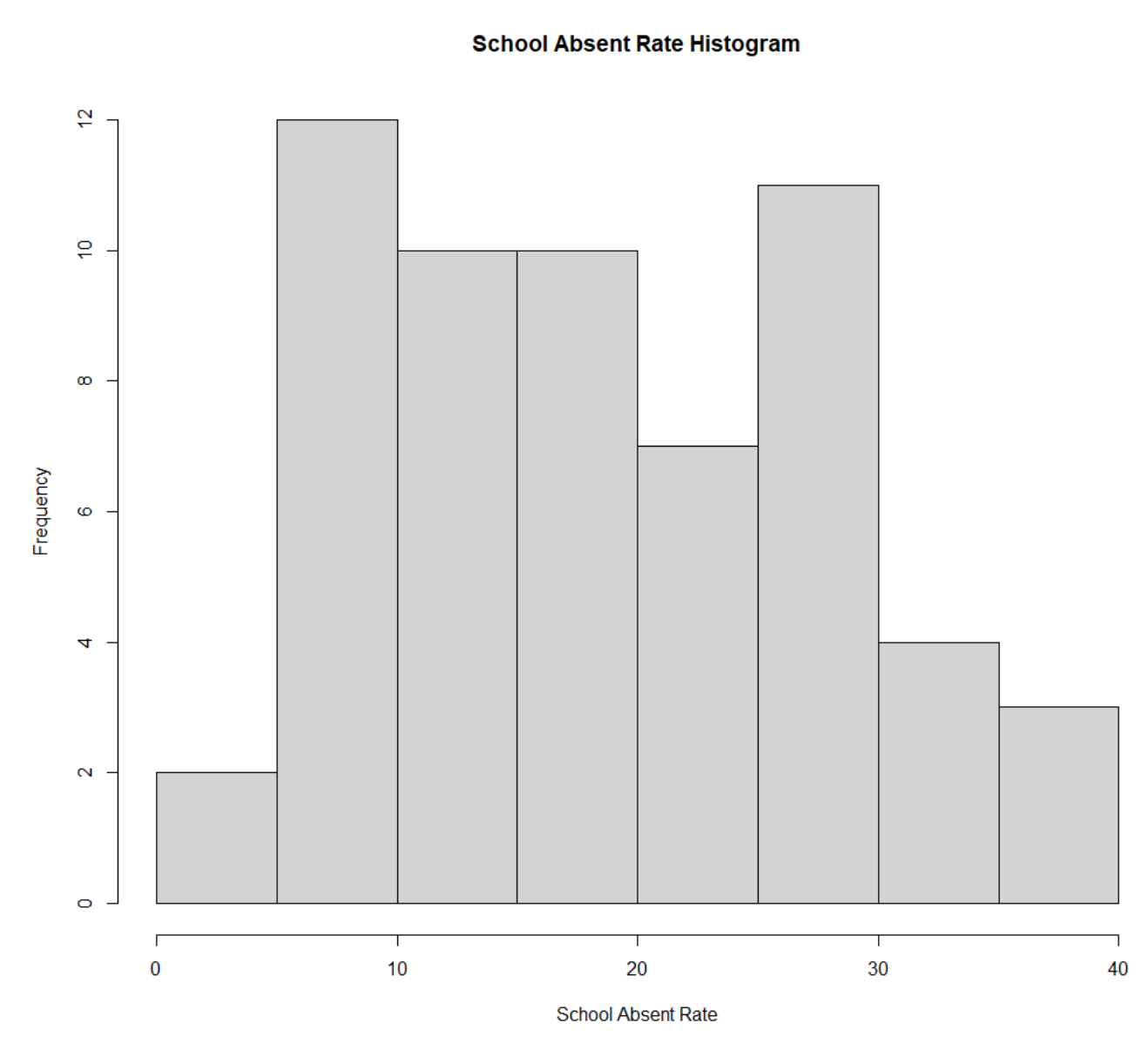
Below are some of the exploratory plots (box, bar, histogram) created to look at the data and the distribution of it. In addition to these, scatterplots were also generated to see the correlation between absent rate and five other variables.



Boxplots for: Absent Rate, Sugary Drink, Fruit/Veg



Bar graph of Absent Level



Histogram for school absent rate

It is important to understand the distribution of data points within each borough using the data for each district. With the scatterplots and correlation coefficients (not shown), it was evident that sugary drink consumption, diabetes, and obesity had positive correlations with absent rate, while fruit and vegetable consumption and exercise had negative correlations. The two variables with the highest correlations were sugary drink consumption and fruit and vegetable consumption, hence they are shown in the box plots.

To clean up the data, I isolated the six variables and created two columns: one for borough identification per district, and one for absent level based on min and max of Schoolabsent_rate.

Models

Four models were used: multivariate and linear regression, k-means clustering by borough and absentee level, random forest, and decision tree.

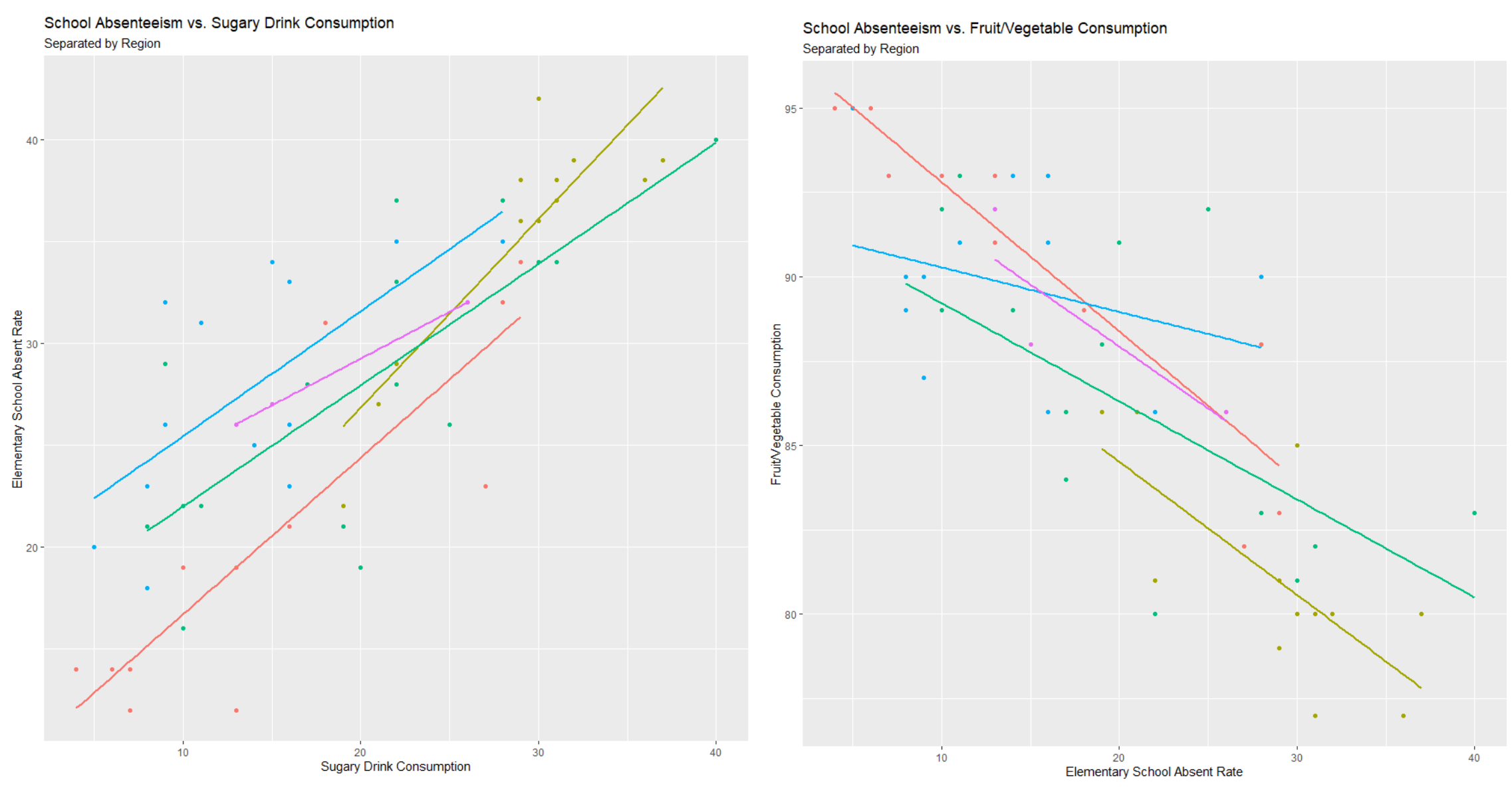


Figure 1: Linear Regression

Model 1: Multivariate + Linear Regression

Multivariate regression was used to compare the variables predicted to have positive and negative correlation separately. Below are the multivariate regression models:

```
# positive correlation
absent = -7.2 + 0.5*sugary_drink +
0.5*obesity – 0.01*diabetes
# negative correlation
absent = 143.3 – 1.5*fruit_veg +
0.09*exercise
# all variables
absent = 22.1 + 0.4*sugary_drink –
0.7*fruit_veg + 0.4*exercise + 0.4*obesity
+ 0.2*diabetes
```

To the left (Fig. 1) are the linear models of sugary drink consumption and fruit and vegetable consumption.

Model 2: K-means Clustering

Clustering was used to compare the absent rate to the five variables among the boroughs (Fig. 2). It was also used to see the absent level distributions, specifically low, moderate, and high, with sugary drink consumption and obesity.

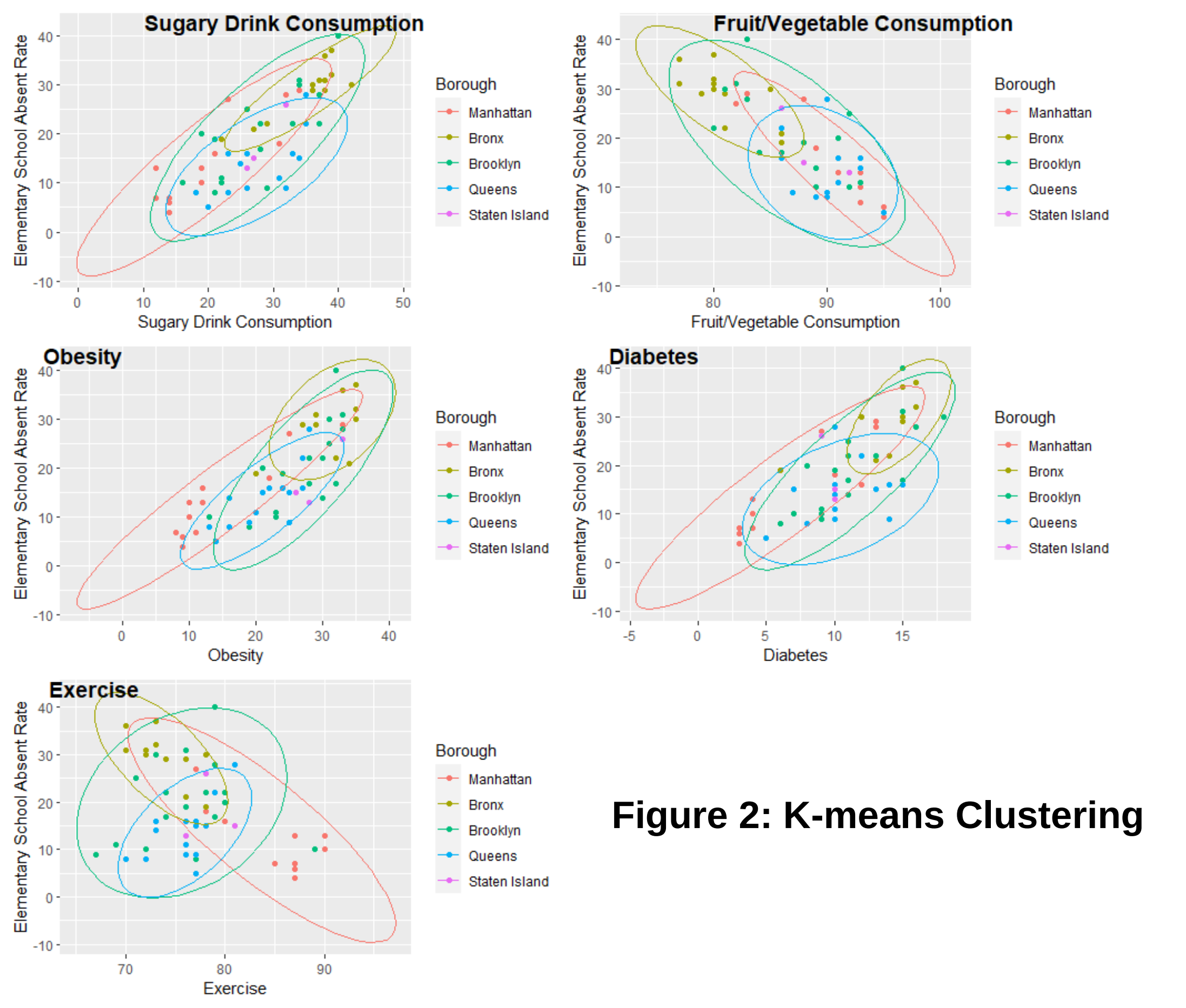


Figure 2: K-means Clustering

Model 3: Random Forest

The random forest models were created by obtaining a sample set of the given data and providing different parameters for the number of variables and trees at each split. Below is the confusion matrix and error for the default function with `ntree = 500` and `mtry = 2`.

```
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
```

Datasets & Cleaning

The primary dataset used was the New York City Community Health Profiles (CHP) dataset. The CHP data contains information from 59 community districts and 5 total boroughs in NYC, 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 a "cross-sectional telephone survey with an annual sample of approximately 10,000 randomly selected adults aged 18 and older from all five boroughs."

Column name	Description of data
Name	District name
School_absentrate	Rate of NYC public school students, grades K-5, that were chronically absent
Sugary_Drink	Percent of adults who drank one or more 12oz sugar-sweetened beverages per day
Fruit_Veg	Percent of adults who ate at least one serving of fruits or vegetables in the last day
Exercise	Percent of adults who got any exercise in the last 30 days
Obesity	Percent of adults that are obese (BMI of 30 or greater) based on height and weight
Diabetes	Percent of adults who been told by a healthcare professional that they have diabetes
Borough	Borough the district (row) is in (based on ID)
Absent_distribution	Low/Moderate/High level of absentee rate based on minimum and maximum of dataset

Results

The models indicated that there a relationship between three of the variables and school absent rate, sugary drink consumption, obesity, and fruit and vegetable consumption. However, more data is required to confirm the correlation and do any further analyses. In terms of the null hypotheses, the hypotheses for the variables mentioned are accepted, while the rest are rejected.

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.

Glossary:

Community health profiles (CHP) – dataset provided NYC with information on health and other variables within districts in New York City

FITNESSGRAM test – a fitness and activity assessment that children from K-8 are required to complete testing aerobic capacity, muscular strength, muscular endurance, flexibility, and body composition

Resources:

1. 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/.
2. "Community Health Survey." Community Health Survey - NYC Health, https://www.nyc.gov/site/doh/data/data-sets/community-health-survey.page.
3. "Datasets." Datasets - NYC Health, https://www.nyc.gov/site/doh/data/data-sets/data-sets-and-tables.page.
4. 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/.