2013-2019 School Attendance

By Jeremy and Frank

Background Information

Attendance data includes students in district 1-32, 75 (Special Education), district 79 (Alternative Schools & Programs), charter schools, home schooling. Home and hospital instruction are excluded. Pre-K data does not include NYC Early Education Centers or District Pre-K Centers therefore data is limited to those who attend K-12 schools that offer Pre-K. Transfer school counts are not in school level file. Attendance is registered to school student is attending at the time. If a student attend multiple schools in a school year the data will be reflected in multiple schools. Chronical absence is defined if a student has an attendance rate of less than 90 percent (students who are absent 10 percent or more of the total days).

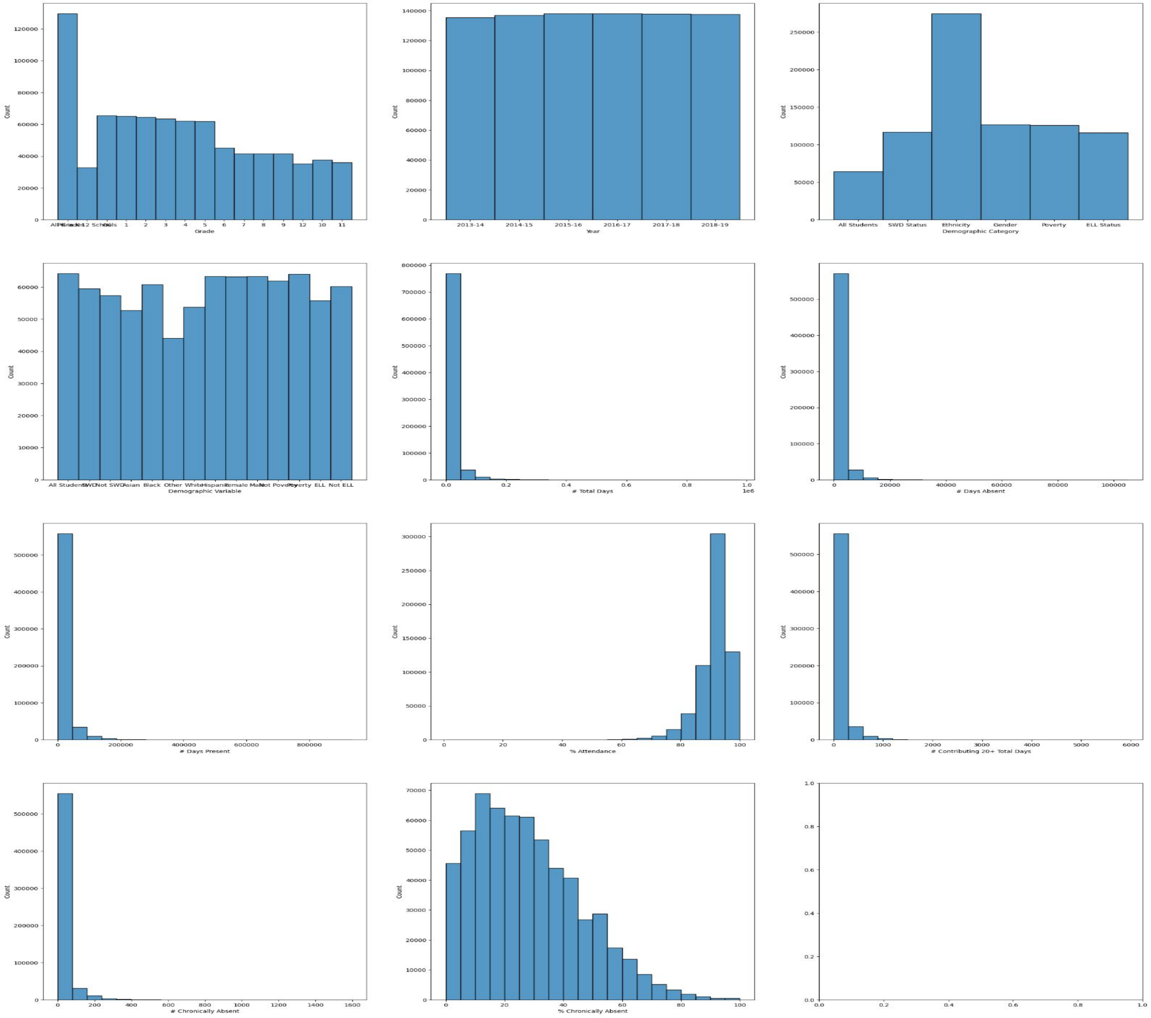
Link to Dataset: https://data.cityofnewyork.us/Education/2013-2019-Attendance-Results-School/vww9-qguh

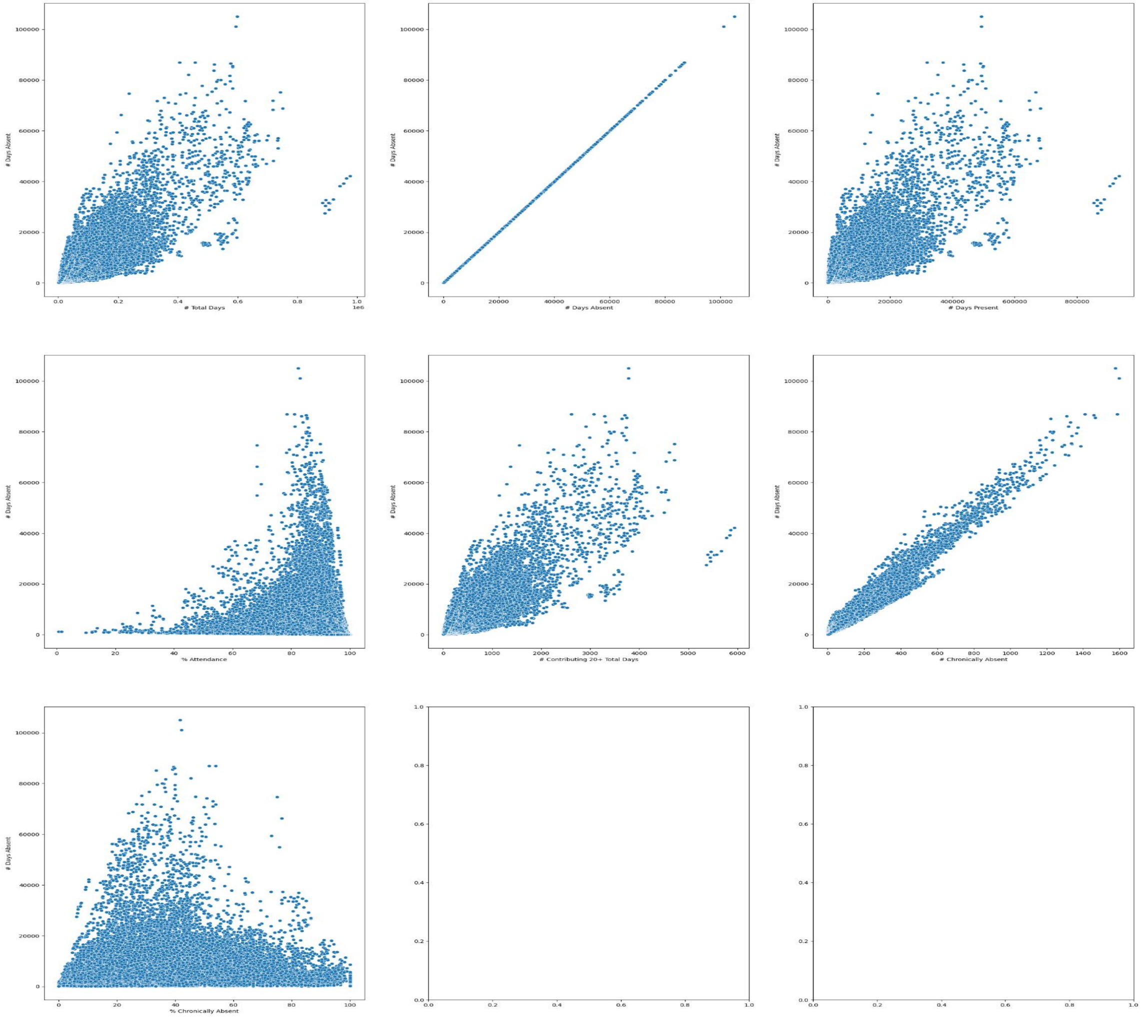
Features

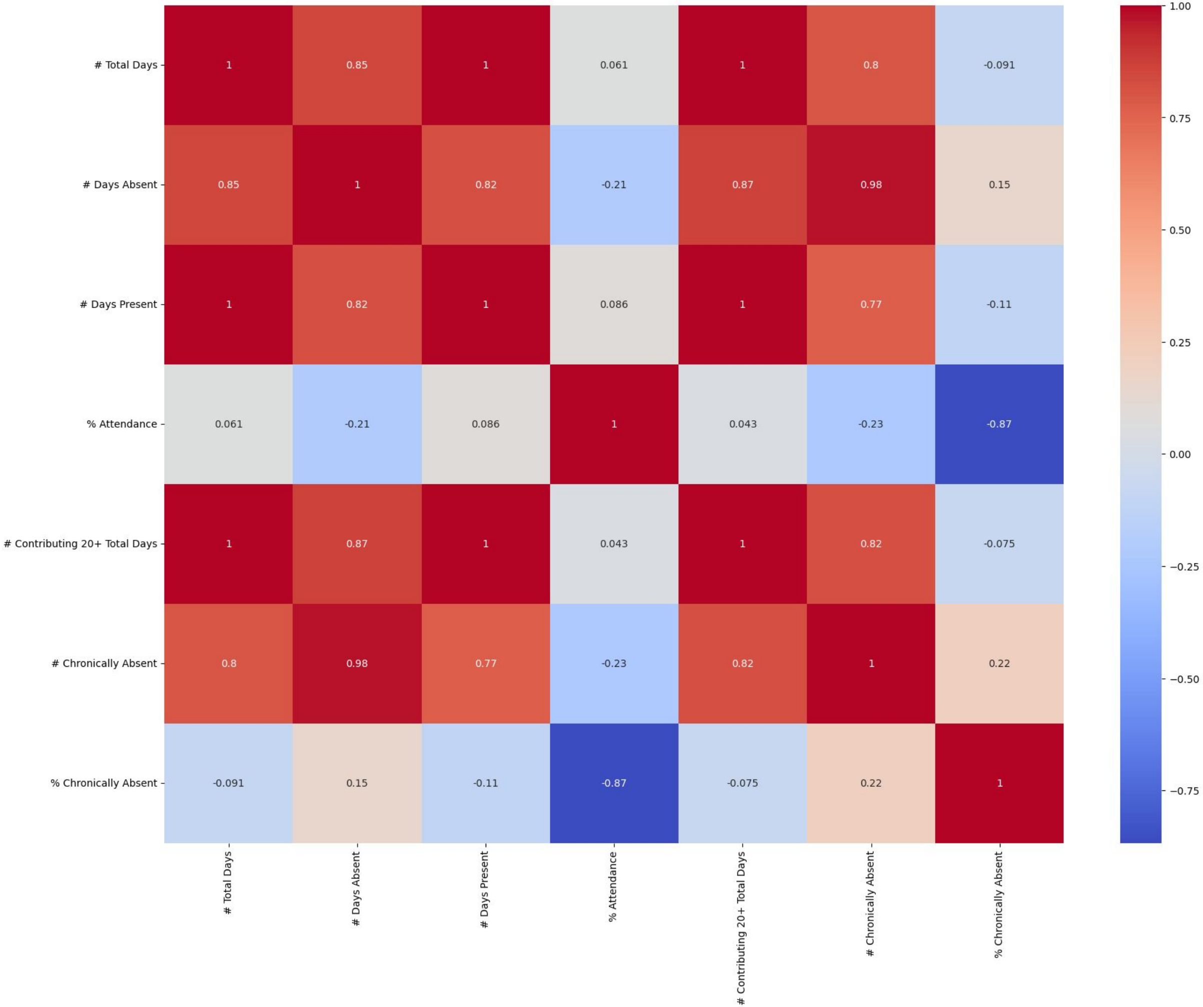
- #total day: The combination of #days present and # days absent. It is also the total number of days of school for all students on that row.
- #days absent: The total number of days all students on that row are absent.
- #days present: The total number of days that all students on that row are present.(unsure how data is collected for this, needs more research)
- % Attendance: Percentage of total days that all students on that row are present.
- # Contributing 20+ Total Days: The total number of students in the specific row that attended school for at least 20 days.
- # Chronically Absent: Total number of students who were chronically absent.
- % Chronically Absent: The percent of students who are chronically absent out of the total number of students.
- Demographics Category: Contains the types of demographics used when collecting the data. There is All students, SWD status, Ethnicity, Gender, Poverty and ELL status.

Demographic Categories

- All Students: This demographic consists of all students. It does not separate the student data into separate rows: Female and Male
- SWD status: Refers to whether a student has a disability or not.
- Ethnicity: Refers to the students’ ethnicity: white, asian, black, hispanic, and other.
- Gender: Separates student data based on gender, male and female.
- Poverty: Separates students into two different categories: poverty and not poverty
- ELL status: Separates students into whether English is their first language or not.
- DBN: The distribute borough number of the students at that row
- Year: The year of the students at that row
- Grade: The grade of the students at that row
- School: The school at which those students attend.


Count
0
20000
40000
60000
80000
100000
120000
All Grades
PK
K
K-12 Schools
0K
1
2
3
4
5
6
7
8
9
12
10
11
Grade
Count
0
20000
40000
60000
80000
100000
120000
140000
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
Year
Count
0
50000
100000
150000
200000
250000
All Students
SWD Status
Ethnicity
Gender
Poverty
ELL Status
Demographic Category
Count
0
10000
20000
30000
40000
50000
60000
All Students
SWD
Not SWD
Asian
Black
Other
White
Hispanic
Female
Male
Not Poverty
Poverty
ELL
Not ELL
Demographic Variable
Count
0
100000
200000
300000
400000
500000
600000
700000
800000
0.0
0.2
0.4
0.6
0.8
1.0
1e6
# Total Days
Count
0
100000
200000
300000
400000
500000
0
20000
40000
60000
80000
100000
# Days Absent
Count
0
100000
200000
300000
400000
500000
0
200000
400000
600000
800000
# Days Present
Count
0
50000
100000
150000
200000
250000
300000
0
20
40
60
80
100
% Attendance
Count
0
100000
200000
300000
400000
500000
0
1000
2000
3000
4000
5000
6000
# Contributing 20+ Total Days
Count
0
100000
200000
300000
400000
500000
0
200
400
600
800
1000
1200
1400
1600
# Chronically Absent
Count
0
10000
20000
30000
40000
50000
60000
70000
0
20
40
60
80
100
% Chronically Absent
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


# Days Absent
# Total Days
1e6
# Days Absent
# Days Present
% Attendance
# Contributing 20+ Total Days
# Chronically Absent
% Chronically Absent
0
20000
40000
60000
80000
100000
0.0
0.2
0.4
0.6
0.8
1.0
200000
400000
600000
800000
20
40
60
80
100
1000
2000
3000
4000
5000
6000
200
400
600
800
1000
1200
1400
1600



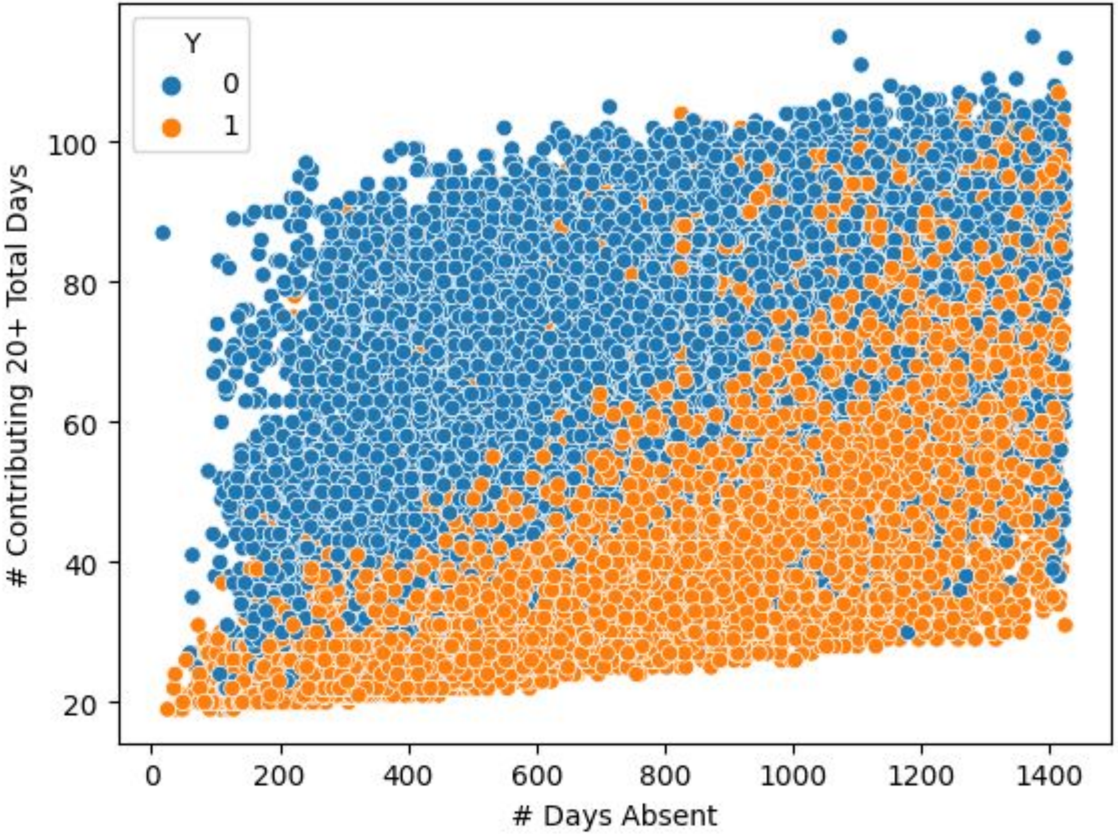
# Total Days
# Days Absent
# Days Present
% Attendance
# Contributing 20+ Total Days
# Chronically Absent
% Chronically Absent
1
0.85
1
0.061
1
0.8
-0.091
0.85
1
0.82
-0.21
0.87
0.98
0.15
1
0.82
1
0.086
1
0.77
-0.11
0.061
-0.21
0.086
1
0.043
-0.23
-0.87
1
0.87
1
0.043
1
0.82
-0.075
0.8
0.98
0.77
-0.23
0.82
1
0.22
-0.091
0.15
-0.11
-0.87
-0.075
0.22
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Objective

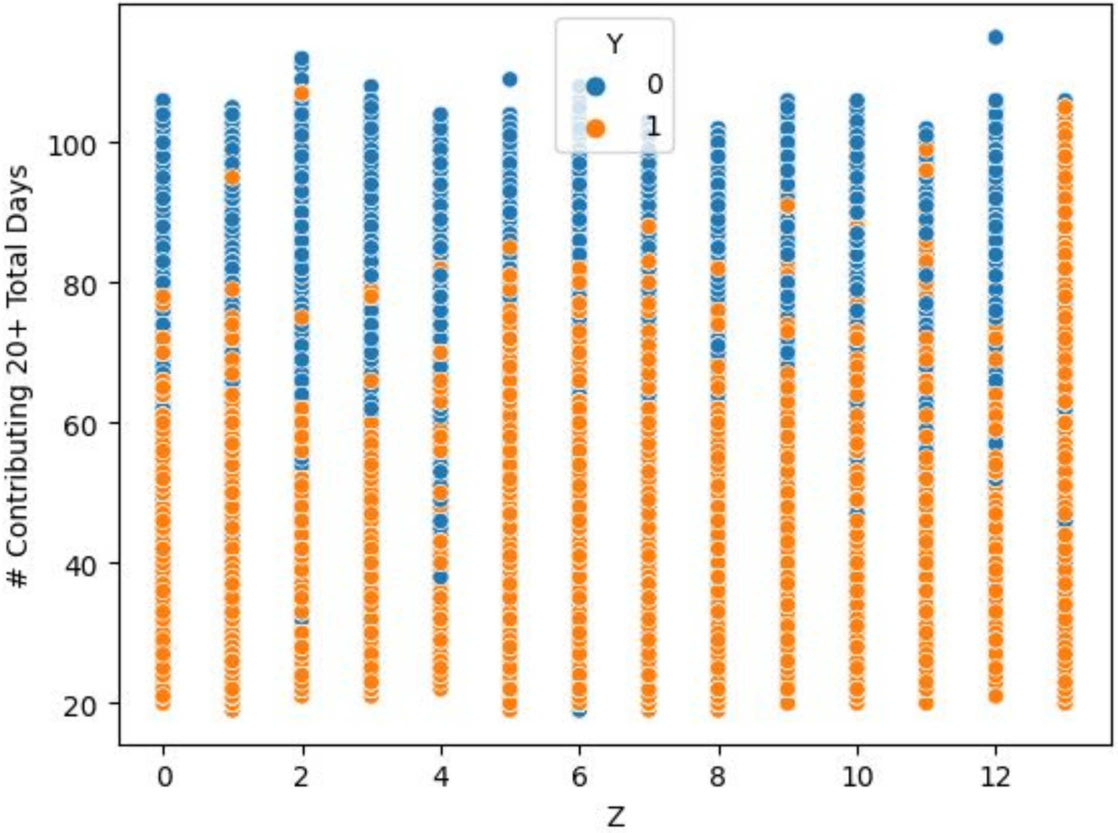
The goal is to find whether a student is classified as SWD(Student with Disability) or not SWD. The 3 independent variables are the number of days absent, the total number of students in the specific row that attended school for at least 20 days and the grade that group of students are in.

The dependent variable is a classification of a student who is classified as SWD or not SWD

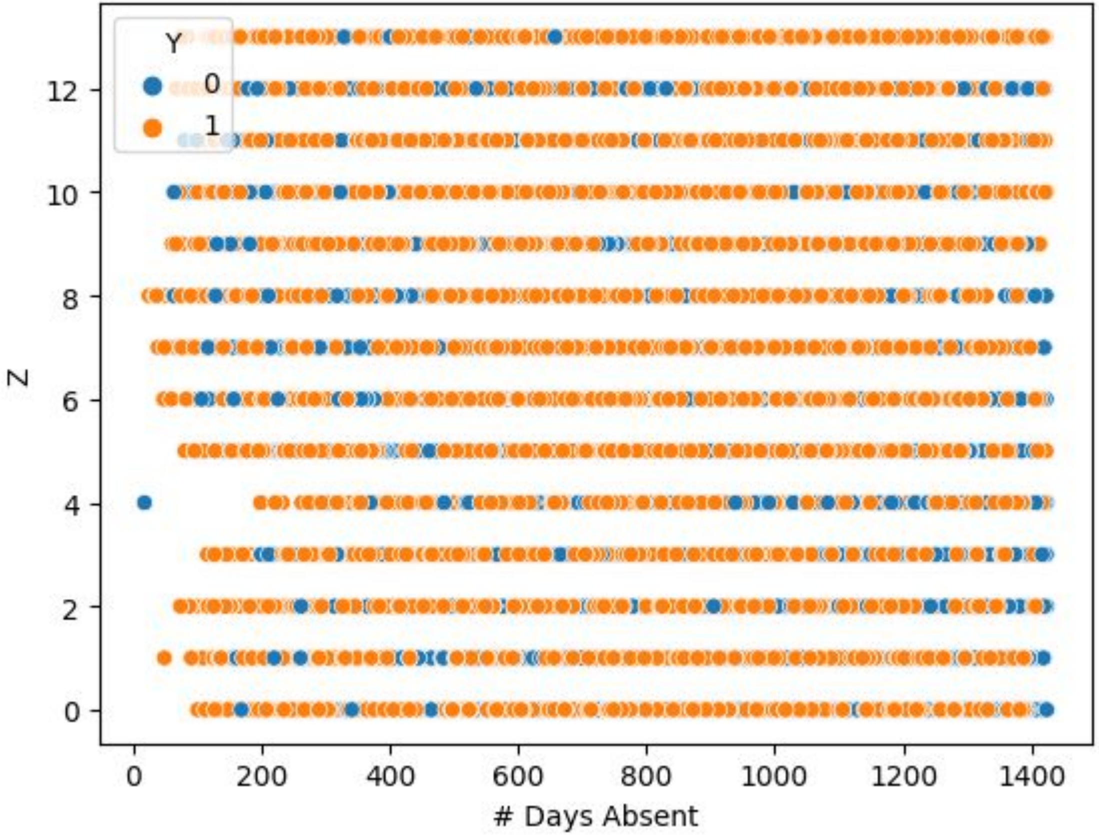
0 = Not SWD
1 = SWD



Z is an ordinal encoding of the Grade column

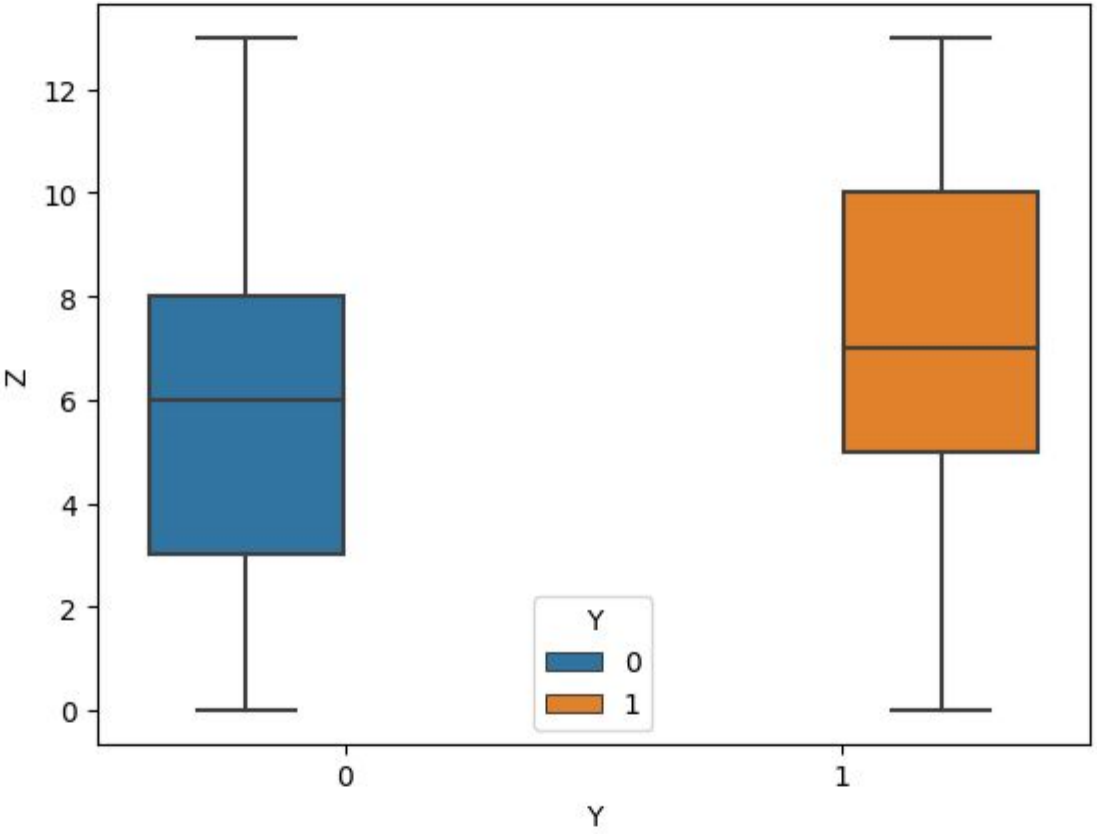


Z is an ordinal encoding of the Grade column

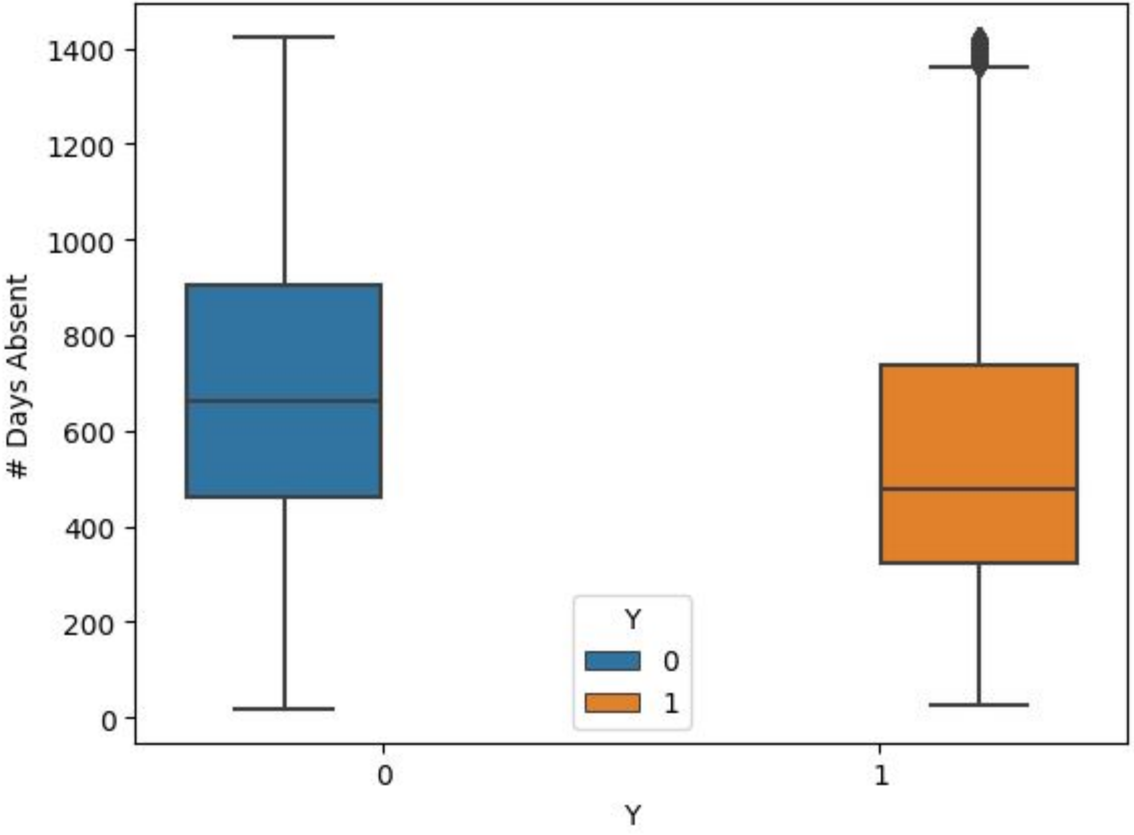


Z is an ordinal encoding of the Grade column

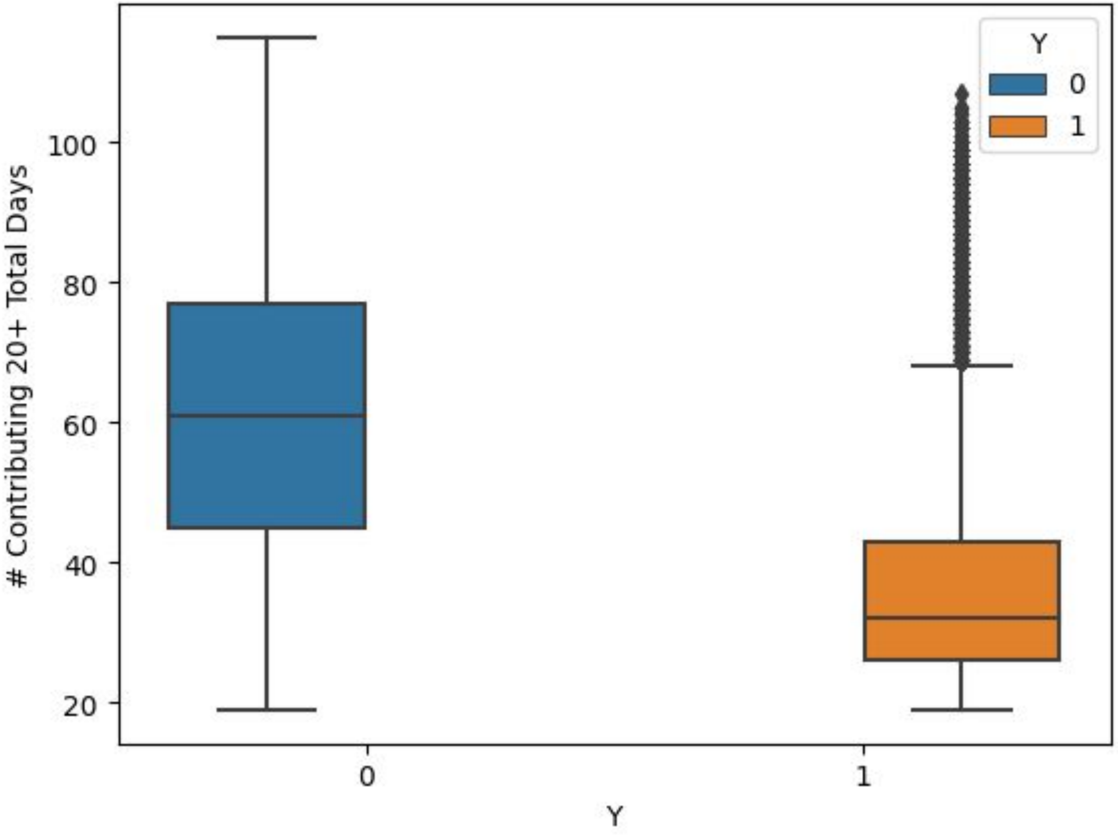
Y is a Binary Encoding of the SWD category in the Demographic Variable Column



Y is a Binary Encoding of the SWD category in the Demographic Variable Column



Y is a Binary Encoding of the SWD category in the Demographic Variable Column



Objective Model 1

We will use Linear Regression to portray the relationship between the three independent variables, the number of days absent, the total number of students in the specific row that attended school for at least 20 days and the grade that group of students are in. The dependent variable is the SWD status of that group of students surveyed.

Linear Regression Model

R-squared Score: 0.3547830570047862

Coefficients:

    # Days Absent: 0.00033351

    # Contributing 20+ Total Days: -0.01525098

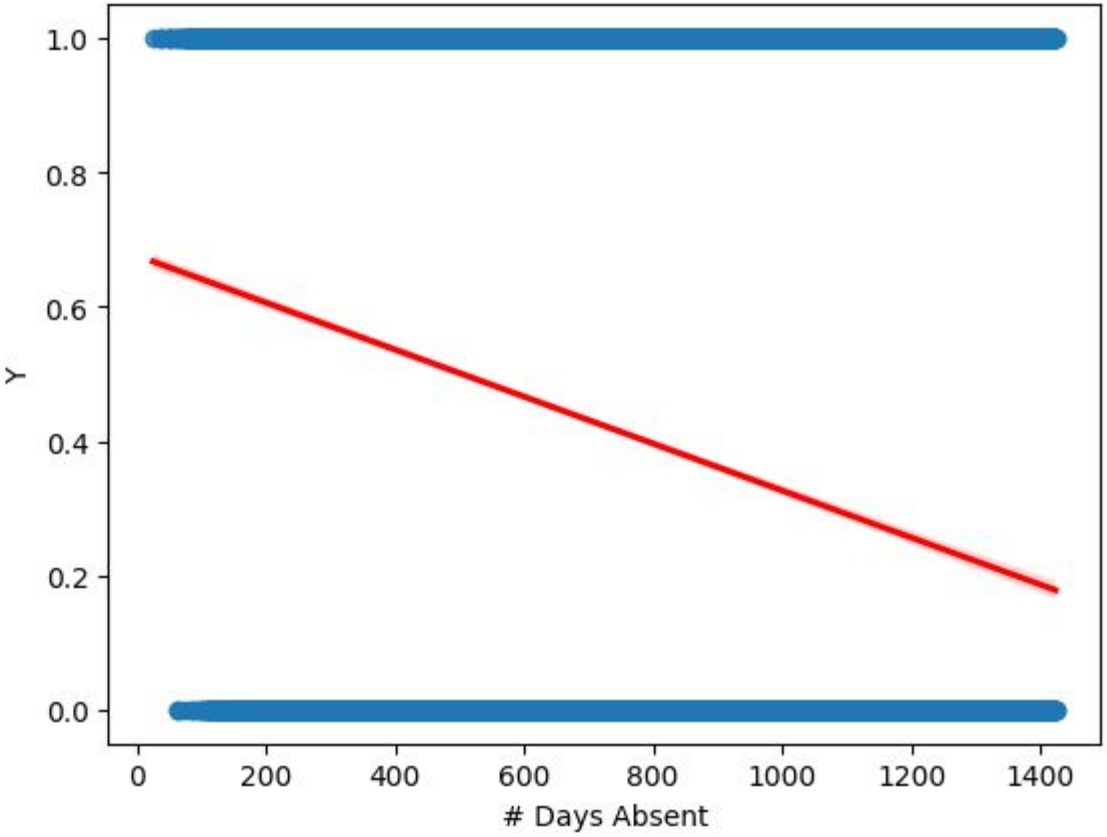
    Z: 0.02973281

Y Intercept: 0.8137906450124841

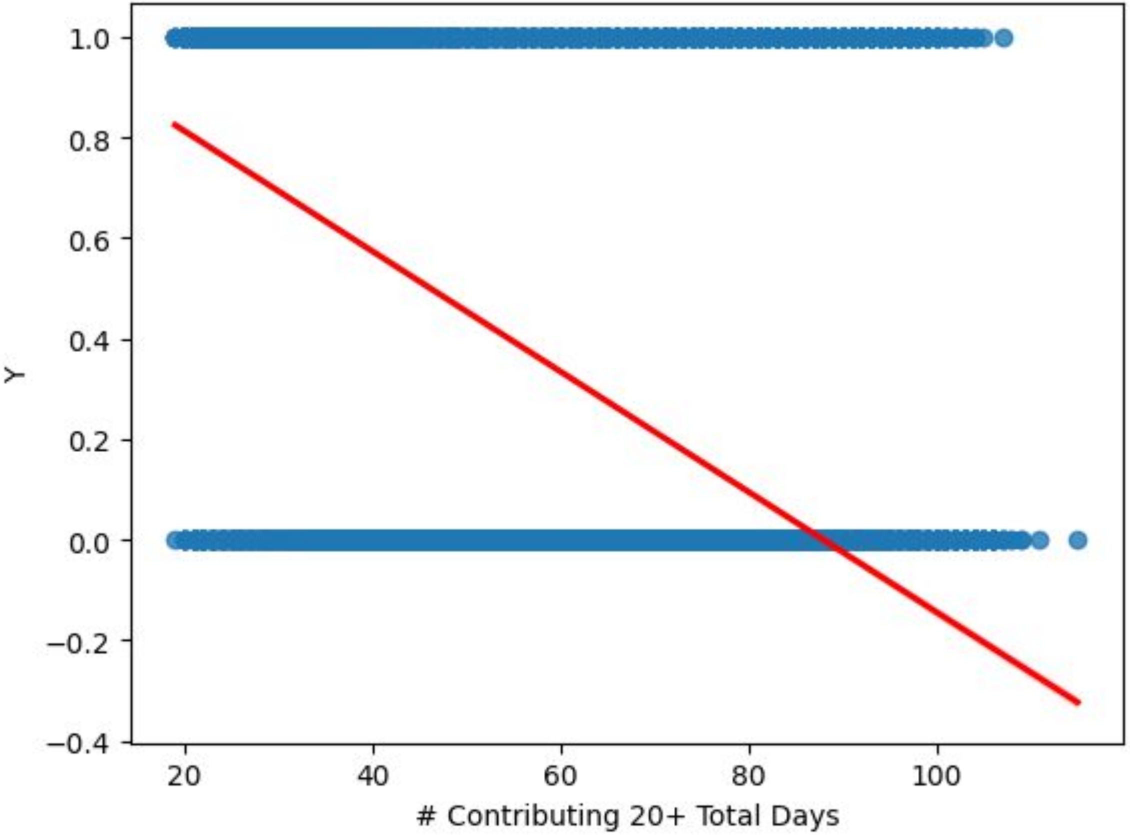
[ ]

```
X_train_LR, X_test_LR, y_train_LR,
y_test_LR = train_test_split(X, y,
test_size=0.3,)

from sklearn.linear_model import
LinearRegression
lr = LinearRegression()
model_LR=lr.fit(X_train_LR,
y_train_LR)

from sklearn.metrics import
r2_score

y_pred_LR =
model_LR.predict(X_test_LR)
r2_score_LR=r2_score(y_test_LR,
y_pred_LR)
slopes_LR=model_LR.coef_
intercept_LR=model_LR.intercept_
print(r2_score_LR)
print(slopes_LR)
print(intercept_LR)
```

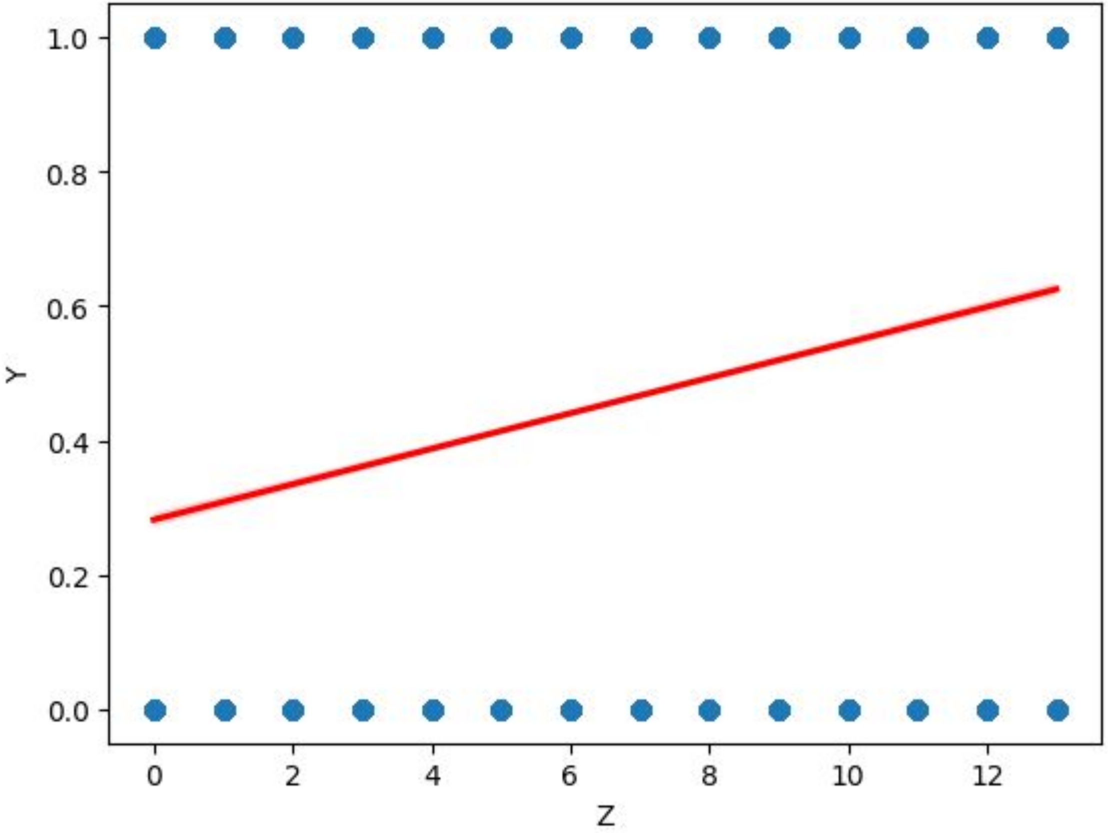
```
sns.regplot(x="# Days
Absent", y=y_train_LR,
data=X_train_LR,
line_kws={"color":
"red"})
```



```
sns.regplot(x="#
Contributing 20+ Total
Days", y=y_train_LR,
data=X_train_LR,
line_kws={"color":
"red"})
```



```
sns.regplot(x="Z",
y=y_train_LR,
data=X_train_LR,
line_kws={"color":
"red"})
```



Objective Model 2

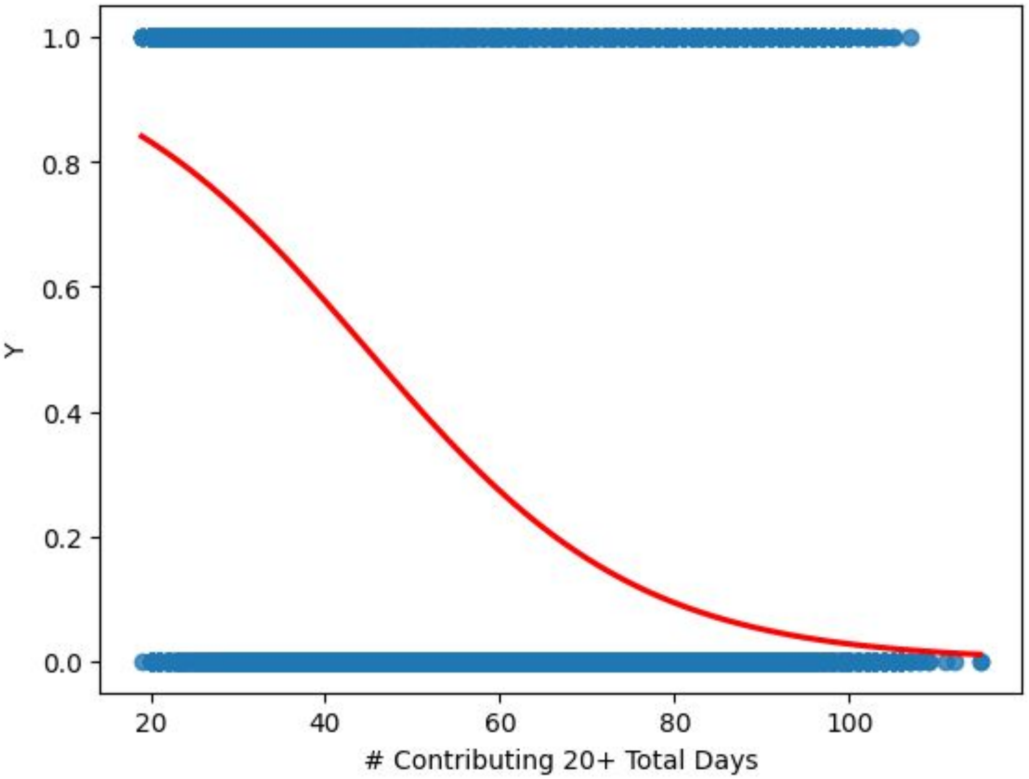
We will use Logistic Regression to portray the relationship between the three independent variables, the number of days absent, the total number of students in the specific row that attended school for at least 20 days and the grade that group of students are in. The dependent variable is the SWD status of that group of students surveyed.

Logistic Regression Model

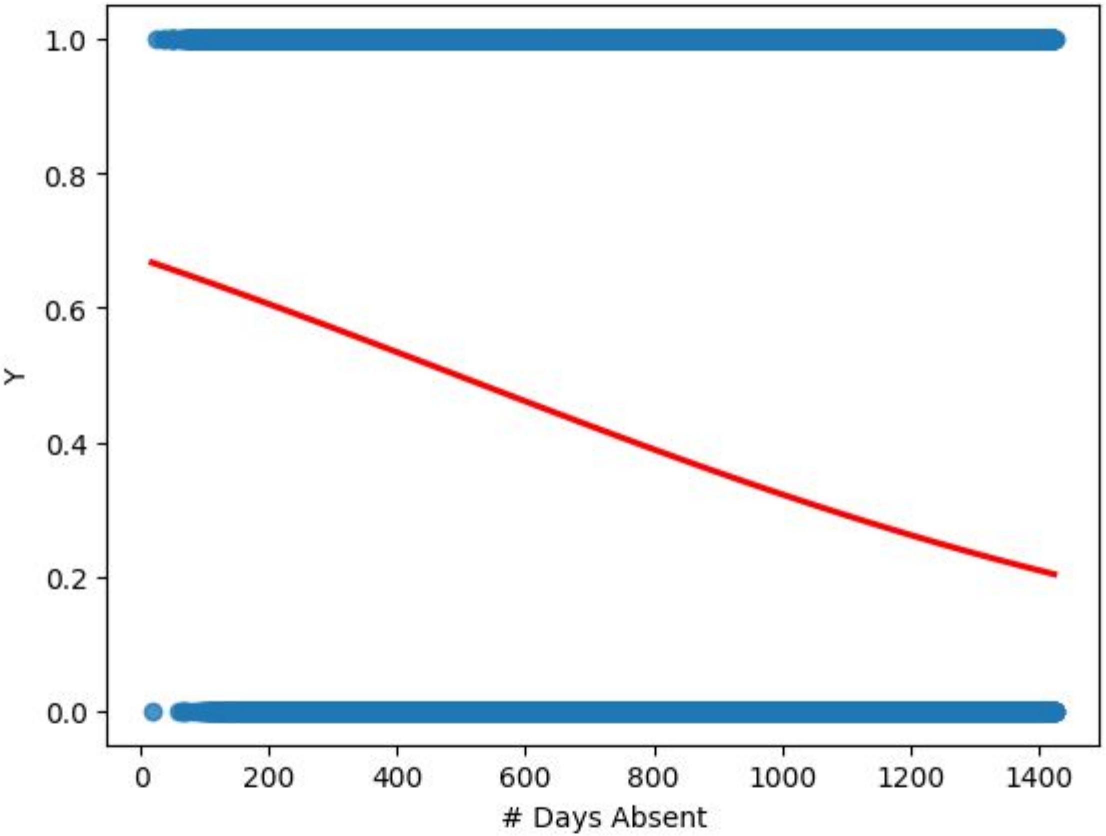
```
Accuracy Score:  0.7836212801583073

Coefficients:

     Z: 0.20169294

     # Days Absent: 0.00254206

     # Contributing 20+ Total Days: -0.1006931

Y Intercept:  1.67466691
```

```
from sklearn.model_selection import
train_test_split
X_train_LG, X_test_LG, y_train_LG, y_test_LG =
train_test_split(X, y, test_size=0.3,)

from sklearn.linear_model import
LogisticRegression
lgr = LogisticRegression()
model_LG=lgr.fit(X_train_LG, y_train_LG)

from sklearn.metrics import accuracy_score
y_pred_LG = model_LG.predict(X_test_LG)
accuracy_LG=accuracy_score(y_test_LG,y_pred_LG
,normalize=True)
slopes_LG=model_LG.coef_
intercept_LG=model_LG.intercept_
```

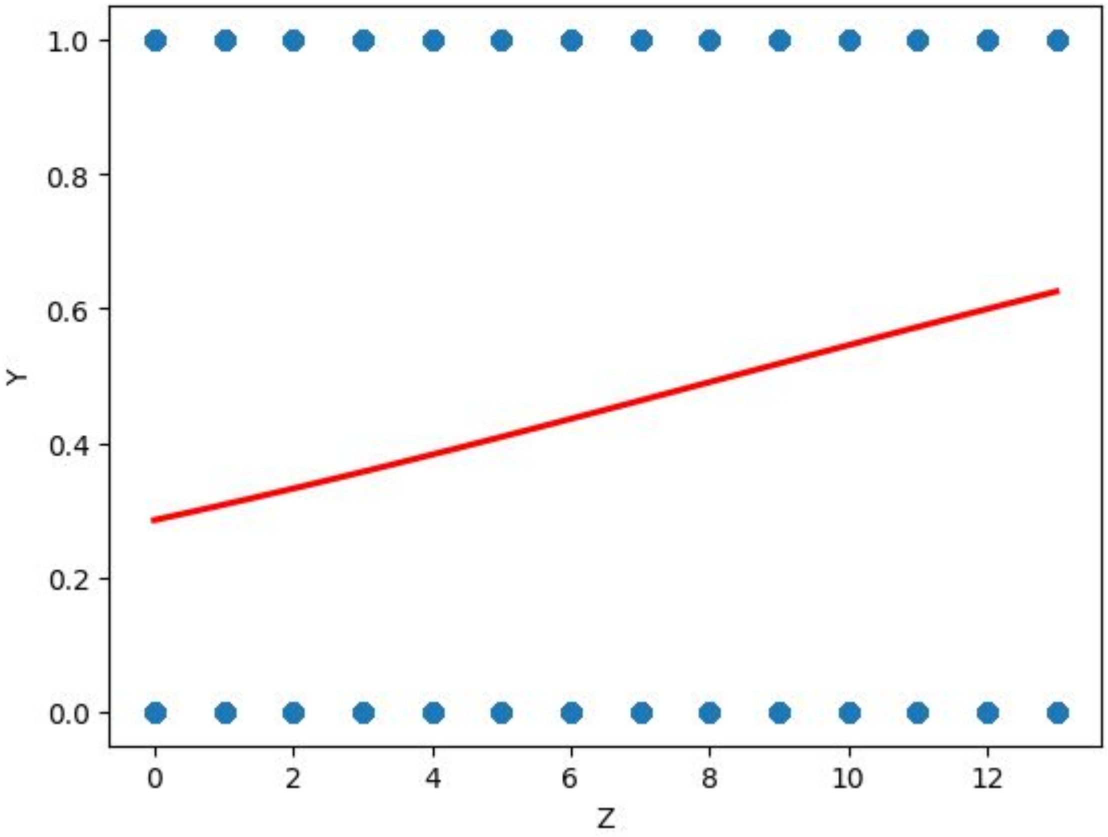
```
sns.regplot(x=log_data.loc[:,"#
Contributing 20+ Total Days"], y=y,
data=log_data, logistic=True,
ci=None, line_kws={"color": "red"})
```



```
sns.regplot(x=log_data.loc[:,"#
Days Absent"], y=y, data=log_data,
logistic=True, ci=None,
line_kws={"color": "red"})
```



```
sns.regplot(x=log_data.
loc[:,"Z"], y=y,
data=log_data,
logistic=True, ci=None,
line_kws={"color":
"red"})
```



Conclusion

The R2 score of Linear Regression is lower than 0.5 while the accuracy score of Logistic Regression is greater than 0.5. Not only is the accuracy score higher for Logistic Regression, the graph for Linear Regression shows that the model will eventually be useless as it goes outside of the upper and lower bounds (0, 1). Therefore, the Logistic Regression model is the better model to use to predict the probability that a group of students that were surveyed were SWD.