Tony Min
Data Analytics 6600
December 8, 2023

Assignment 6

1. Abstract and Introduction
    a. **Abstract:**
    The COVID-19 pandemic has brought about unprecedented challenges in various sectors around the world, one of which including education. As schools grappled with implementing safety measures to stop the spread of the virus, one significant move was the mandates of masks. At the same time, vaccine companies aimed to mitigate the impact of the pandemic, which offered an avenue of going back to normal within schools. This study is to investigate the correlation between school attendance rates and vaccination rates in the state of Connecticut during the COVID-19 pandemic.
    b. **Introduction:**
    The response to the COVID-19 pandemic was swift and adaptive to set measures in order to stop spread. For the realm of education, schools faced the task of ensuring the safety of staff, students, and student families while attempting to maintain a normal learning environment. With the introduction of mask mandates, it aimed to stop the spread of the virus within school walls. However, was this mandate effective in keeping kids in school? The motivation for this study stems from the observation of the dynamic changes in the world of education during the COVID-19 pandemic. With schools implementing protocols to keep individuals safe, it becomes crucial to understand the potential correlation between these measures and attendance rates. Specifically, the focus was on exploring whether the vaccination rates in the state of Connecticut had any discernible influence on school attendance. To further this topic we also want to see if different race / ethnicity groups have different reactions to attendance rates and vaccination rates. My initial hypothesis driving this investigation was if there was a correlation between school attendance rates and vaccination rates. Then compare these results within each ethnicity group to see if a certain ethnicity is more likely to be attending school. The hypothesis of this is based on the idea that higher vaccination rates within a town or school district may contribute to a safer and more confident return to in-person learning which then would impact the attendance rates. On the other hand, communities with lower vaccination rates may experience concerns about going back in-person which could lead to a decrease in attendance.
2. Data Description and Exploratory Data Analytics

a. For the analysis of the correlation between school attendance rates and vaccination rates in Connecticut during the COVID-19 pandemic, two primary datasets were utilized. The first set of data contains vaccination rates within Connecticut, categorized by towns, and includes information on the race and ethnicity of vaccinated individuals, the status of vaccination were fully vaccinated or at least one dose. The second dataset focused on school attendance rates of students in Connecticut for the academic years of 2019-2020, 2020-2021, 2021-2022, also with information of student groups by ethnicity.

[Dataset 1: COVID-19 Vaccination by Town and Race/Ethnicity]

For this dataset why we chose this dataset was how it was sorted out already by Towns in Connecticut and also Race/Ethnicity. This dataset was driven by the need to understand the vaccination landscape at the local level. The data was obtained by the State of Connecticut. The criteria for inclusion were that the dataset needed to cover the towns within Connecticut and needed a breakdown of vaccination rates based on race / ethnicity. The dataset contains features about individuals who have received at least one dose of the COVID-19 vaccine, and also race and ethnicity, which was obtained from electronic health care records. This data is further separated by towns, which was verified by geocoding the reported address and mapping it to a certain town. In summary this dataset provided an insight into the vaccination landscape in Connecticut.
Now for some Exploratory Data Analysis on dataset1.

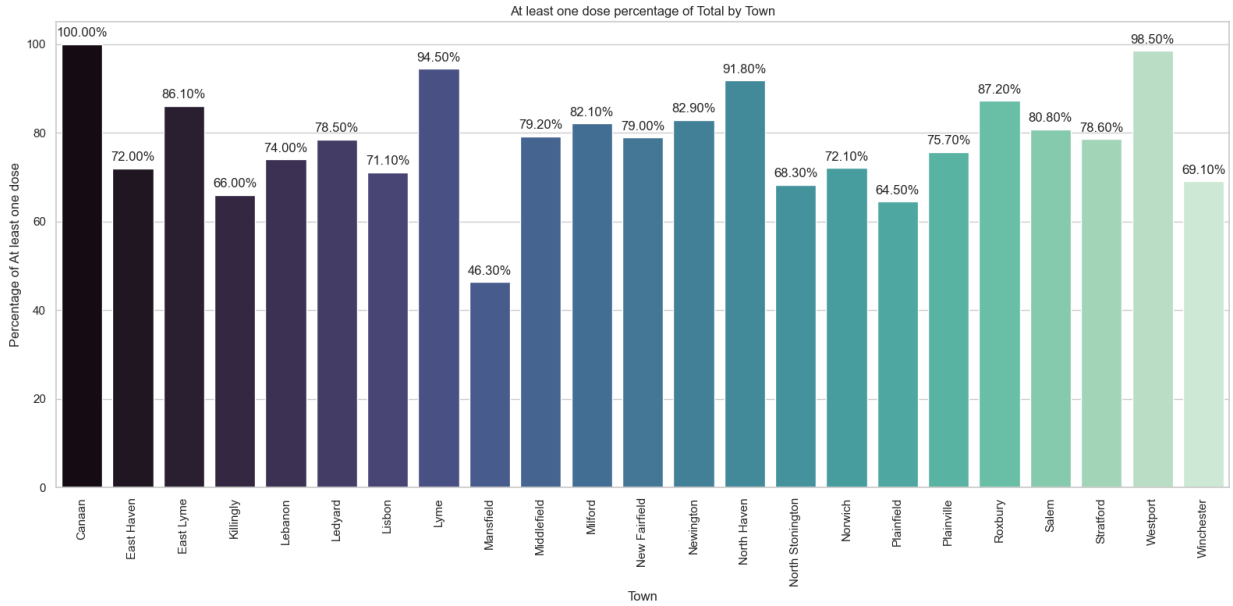

Figure 1: Histogram of Vaccination Rates by Town (Total)

To explore vaccination rates by Town we wanted a visual that wasn't too cluttered, the town names on the x-axis were not readable. We decided to take a sample of 25 towns in Connecticut. Figure 1 above shows the vaccination rates of all individuals within the town that have received at least one dose. Looking at this closely we see that Canaan and Westpoint had the highest rate where individuals had at least one dose of the vaccine, sitting at 100 and 98 percent respectively . On the other hand we see that Mansfield had the lowest rate at around 46 percent. While this is only a sample of 25 towns, at least in our case here in Figure 1, I would say that the rate in which individuals that have gotten at least one dose of the vaccine averages to around 70 percent. Now let's take a look at specific races for these same towns.

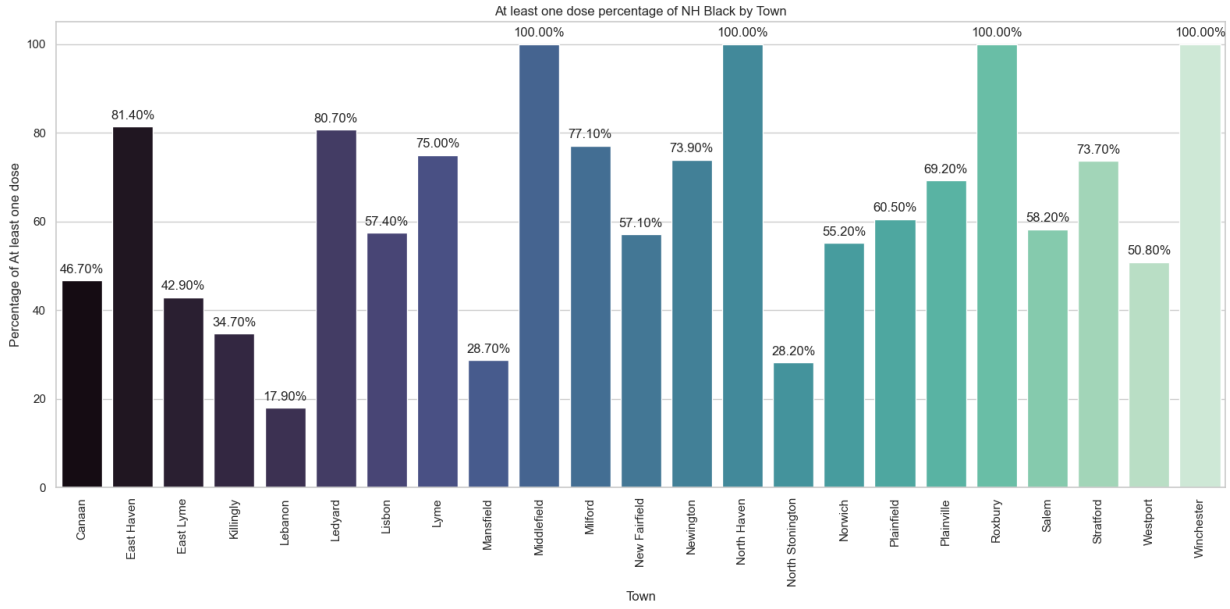

Figure 2: Histogram of Vaccination Rates by Town (Black People)

Now looking at Figure 2 we see a big change was which towns had the highest vaccination rates for at least one dose for Black people compared to all the individuals in the town. The highest percentage towns were Middlefield, North Haven, Roxbury and Winchester. Confusingly from Figure 1 we saw Canaan had 100 percent vaccination rate for a Race/Ethnicity labeled 'Total', from the metadata it does not explain what specifically it stands for but I took it as all individuals in the town but it does not make sense that only 46 percent of black people in Canaan had gotten at least one dose. This may not affect our future modeling and analysis but it was something I noticed while creating these histograms. Now let's compare histograms from the dataset of fully vaccinated individuals.

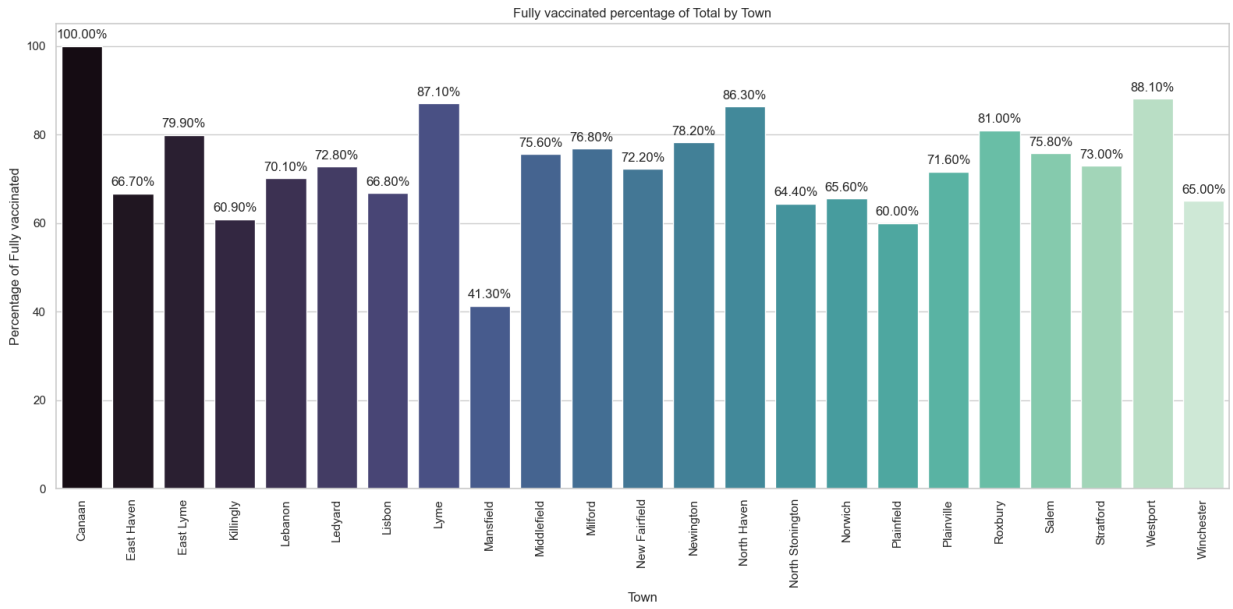

Figure 3: Histogram of Fully Vaccinated Rates by each town (Total)

Now onto Figure 3, we see 100% of Canaan individuals were fully vaccinated, standing to be the only town in our sample to have 100 percent of individuals fully vaccinated. And once again Mansfield was the town with the lowest percentage of 41% of fully vaccinated individuals. Now to look at the fully vaccinated black individuals.

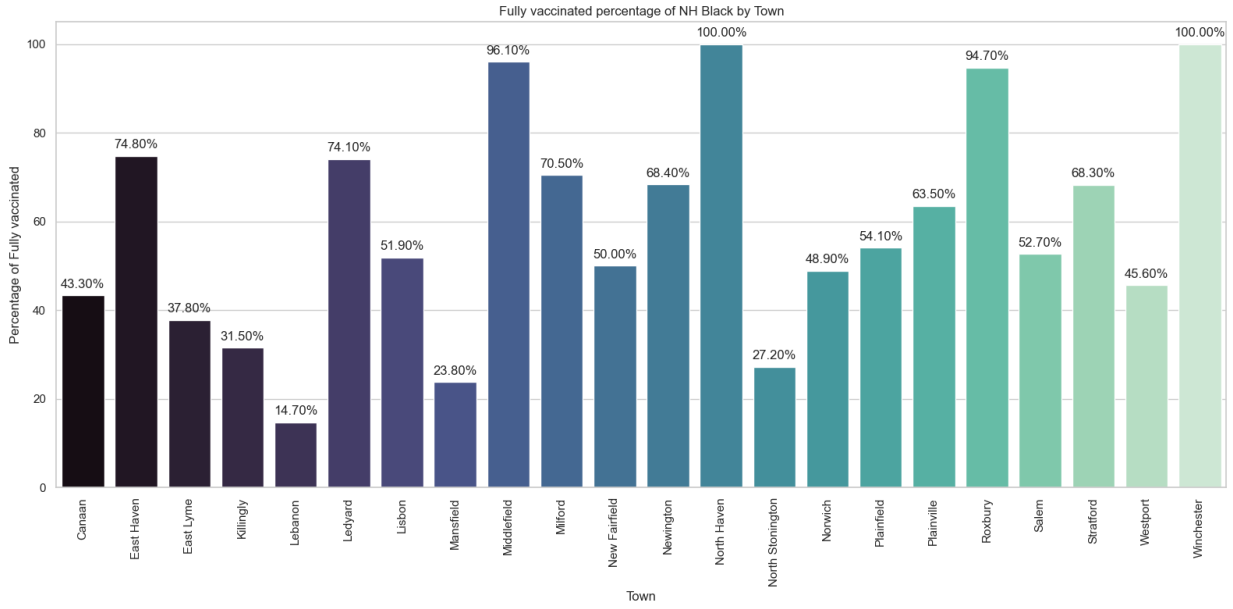

Figure 4: Histogram of Fully Vaccinated Rates by each town (Black People)

Now on Figure 4, we see similar vaccination rates as we did from Figure 2. With only Middlefield and Roxbury dipped below 100 percent which makes sense since the fully vaccinated are a subset of individuals who have received at least one dose.

[Dataset 2: School Attendance by Student Group and District, 2021-2022]

The second dataset focused on the school attendance rates during the 2021-2022 school year was chosen to assess the relations between attendance patterns and vaccination rates. The data was obtained from public school students, PK-12, and stratified by student groups and districts. The source of the dataset was from official school records which are maintained by the Connecticut State department of Education. The dataset includes attendance rates for various different student groups such as homelessness, disabilities, English learned and students by race/ethnicity which is what we will focus on. The other key variables would be district name, attendance rates and student group indicators. Now onto the exploratory data analysis of Dataset 2.

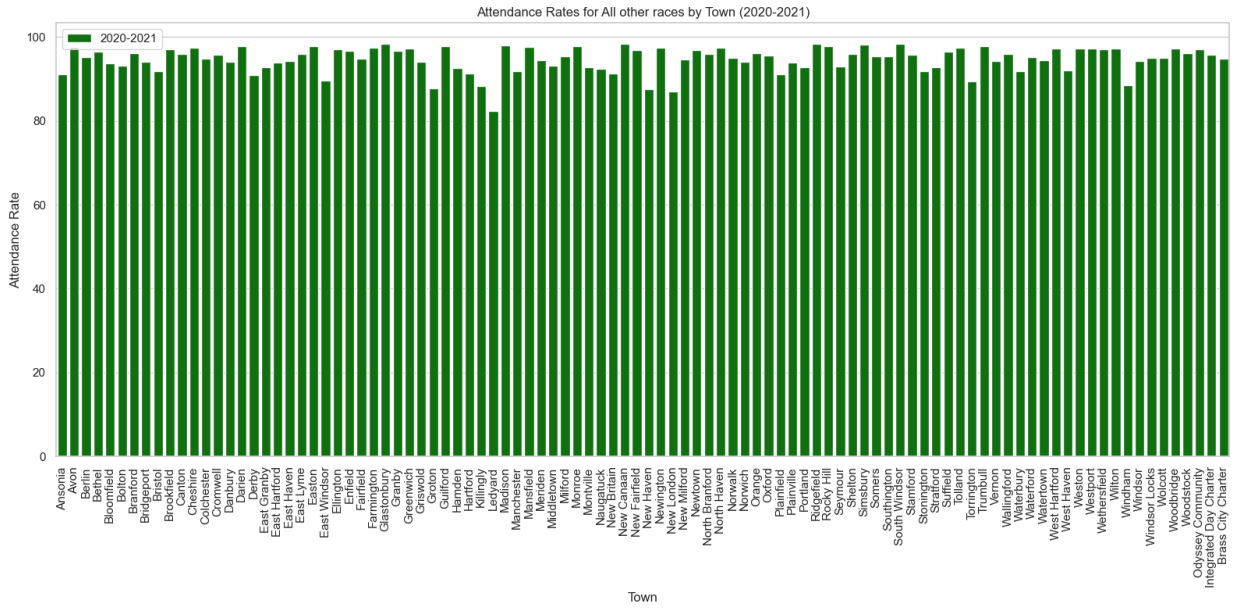

Figure 5: Attendance Rates for 'All other Races' by Town (2020-2021)

From the dataset we wanted to see the attendance rates for towns in a timeframe where the covid vaccine was created and publicly available so we chose the 2020-2021 school year to look closely. While hard to see in Figure 5 all the towns I think it shows how the attendance in the state of Connecticut looks as a whole seemingly like there is a good attendance rate of above or around 85-90 percent. The only towns that seemed to dip below the average were Ledyard. Now let's look at the attendance rates for black people in each town.

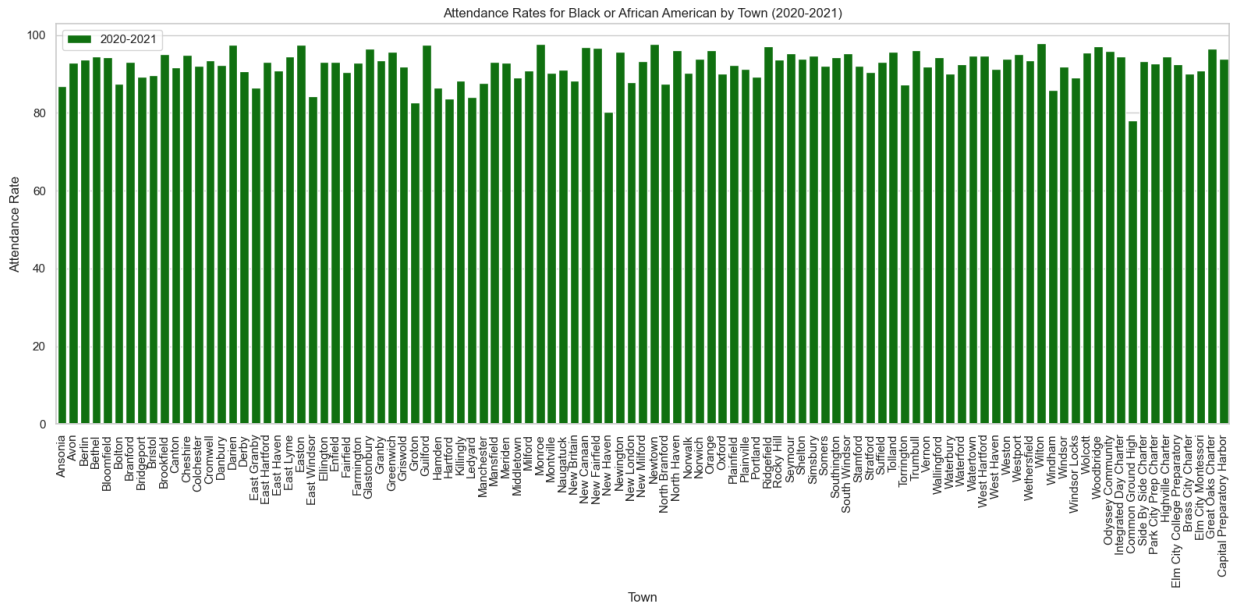

Figure 6: Attendance Rates for 'Black People' by Town (2020-2021)

Now looking at Figure 6 we also see a similar average of attendance for each of the towns we see the lowest rate is at New Fairfield. Along with this we see at the end of the histogram there are some communities in which are not towns in Connecticut and seem like different school districts where they are not named after a town in Connecticut, we will remove these school districts in the future for the sake of keeping the datasets connected by only town names in Connecticut.

Looking more into the datasets we print out some statistics from each. I wanted to first look at the different types of races we had from the vaccination dataset so I could start to think about how we can combine them with the attendance dataset. We saw similarities between each but I decided the ones that did not match with each other would be removed for simplicity. We also looked at each of the towns and school districts and I decided that merging our datasets by the towns would be the easiest and logical thing to do. For this we needed to remove the school districts that were not named after a town, as we noticed in Figure 6.

3. Analysis
    a. Data Transformation and Cleaning took a deeper look into the data and transformed some percentages into the same range (0-100), changed some feature names, and created features as well. Now first off we looked into each dataset and removed null values just to ensure our models won't be missing data. Now specifically for the vaccination dataset I decided to only use the data as it was

updated on “07/20/2022”, we did this because in the attendance dataset the data was last updated on “07/22/2022”. To keep consistent data I choose the closest date each dataset was updated on. I also only wanted percentages from the vaccination data so we kept only data where the feature ‘Data type’ was equal to “Percentage”. Finally we dropped the “Date updated” column. Next onto setting up dataset 2 to prepare merging we first wanted the ‘Category’ column to only contain “Race/Ethnicity” since that is the student group we are focusing on. Next we drop unnecessary columns like ‘District code’, ‘Reporting period’, and ‘Date update’. Now in the ‘District name’ column we had some strings that ended with “School District” so we removed those that did not end with “School District” and then removed “School District” to just get the town names. Now we also had to transform the columns that contained attendance rate. Each of these values were in decimal form so I did some renaming of columns, to keep them consistent, and multiplied all attendance rates by 100 to make them in percentage form. This was done since the vaccination data had them in percentage form already. Finally we just changed the columns ‘District Name’ to ‘Town name’ in order to match our vaccination column. Finally with this we then merged the dataset on the ‘Town name’ column, this also did not include towns that weren’t both in the datasets since I did not want to have a mismatch of data and missing data for some towns. Now that we had a merged data frame we explored the ‘Race/ethnicity’ and ‘Student group’ column in order to see which of the race/ethnicities I could match up together. I then mapped the ‘Race/ethnicity’ by using the values within the ‘Student group’ column and had to drop 2 different ethnicity groups, “NH Asian or Pacific Islander” and “NH American Indian”. Next we split up each of the data frames into 4 different sub data frames based on matching race/ ethnicities.
Once we had created 4 different sub data frames I wanted to feature engineer a No vaccine status. To do this I created a function which took in a dataframe and copied a selected row then changed the ‘Vaccination status’ to “No Vaccine” and then had to modify my ‘Value’ column which was the vaccination percentage. We got this percentage by subtracting the percentage from “At least one dose” row since this percentage is the largest of vaccinated people, since it is a subset.
After this process our 3 sub data frames were ready. Each of the towns within the data frame have all 3 vaccination statuses.

4. Model Development and Application Model
    a. Now for our models I choose to do Linear Regression , Random Forest Regressor, Decision Tree Regressor , XGBoosting Regressor and finally Support Vector Regressor. Now to validate my models since we are doing regression we used the Mean Squared Error, Mean Absolute Percentage Error and R-Squared values. Each of these models were run on the different vaccination status and all the 4 sub data frames.

Linear Regression

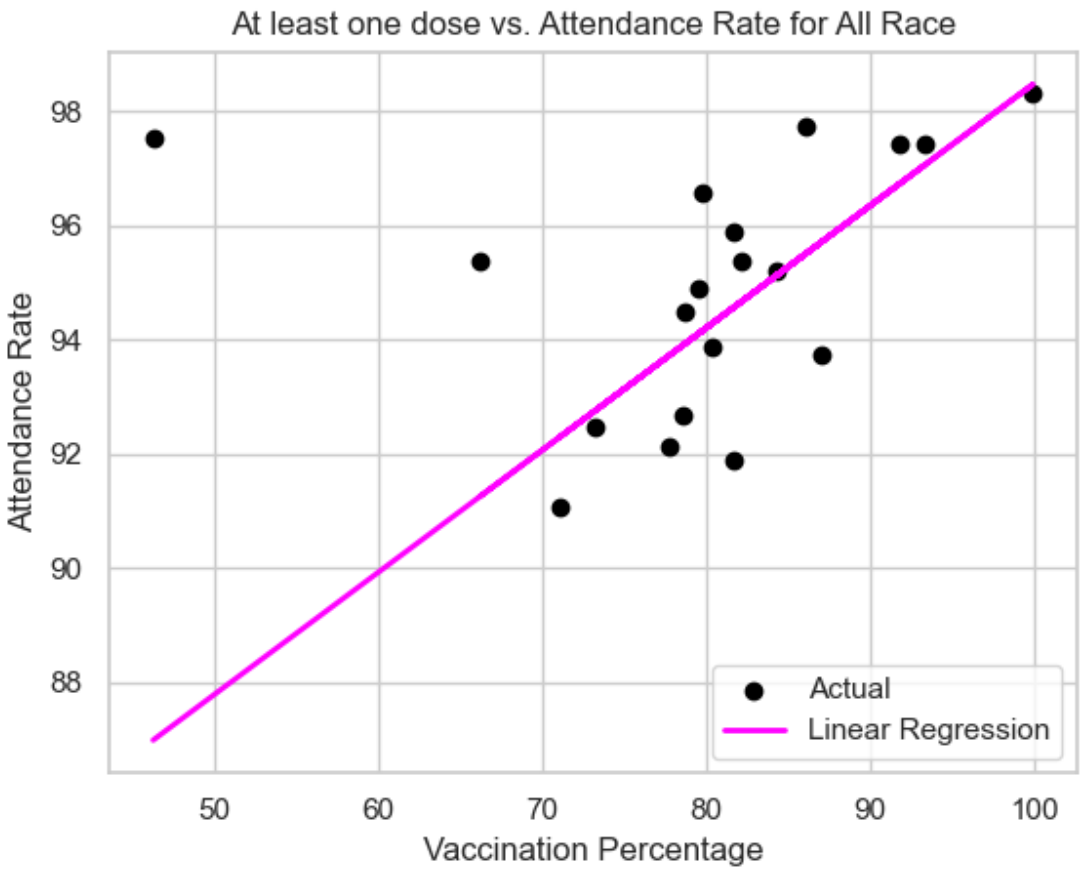

Figure 7: Linear Regression (One Dose) (All Race)

We see there is at least a positive correlation between vaccination and attendance percentages. Looking into our values to look at the performance of our model we see an MSE of 8.435 and MAPE of 1.84, while having an error of 8 and percentage error of 1.8 the model seemed to have a good fit but in terms of variability of attendance rate was -0.8 which is not good since that mean we are explaining a negative amount of the data. Showing us this model does not explain the variance in attendance rate.

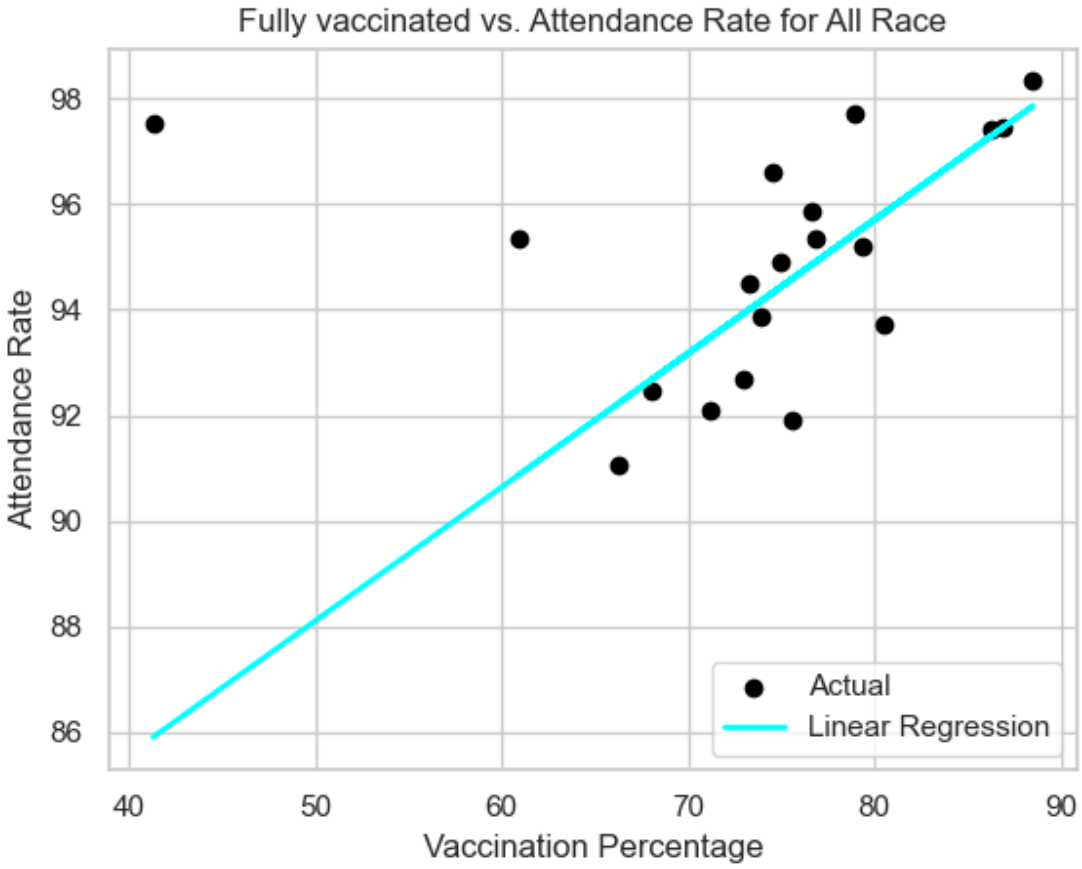

Figure 8: Linear Regression (Fully Dosed) (All Race)

Now in Figure 7, we are showing the linear regression model of fully vaccinated individuals of all races. We got a MSE of 9.7 and an MAPE of 1.8 which again were both quite low values which is good that the model did a nice job at predicting the attendance rate. But once again we got a negative R-Squared value of -1.1 which shows that linear regression is not a good model to predict the vaccination rate vs. attendance rate of all race fully vaccinated individuals.

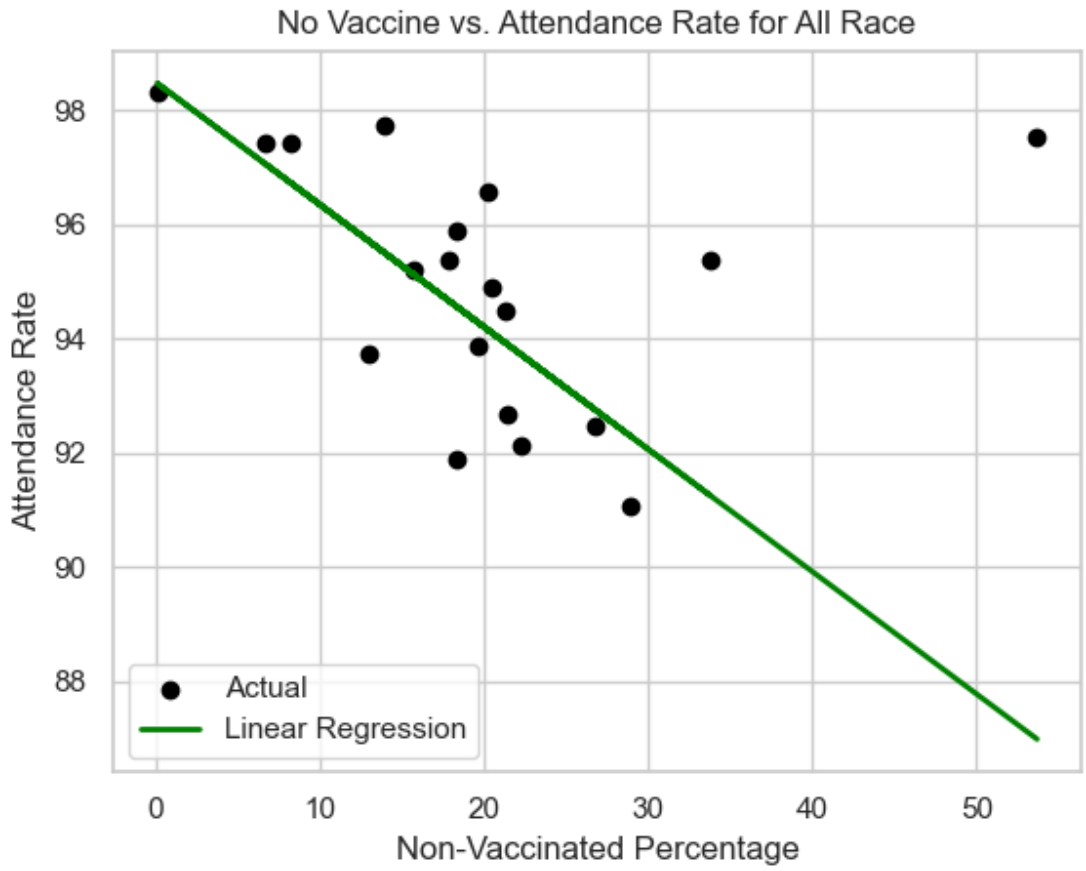

Figure 9: Linear Regression (No Dose) (All Race)

From Figure 9, we have a linear regression model based on the parameters of all races and not vaccinated at all. From this there is a negative correlation which does not make sense from our hypothesis that the higher the vaccination rate the higher the attendance rate would be, but in this mode we see that a low vaccination rate could also lead to a higher attendance rate. In this we had an MSE of 8.4 and MAPE of 1.8, which again was similar to the linear regression models we did on one dose and fully dosed. We once again see a negative R-Squared value showing us this model does not explain the variance in attendance rate.

Below is a table of MSE, MAPE of the combination of vaccination status and the different races.

	All Races (MSE)	All Races (MAPE)	Black (MSE)	Black (MAPE)	White (MSE)	White (MAPE)	Hispanic (MSE)	Hispanic (MAPE)
Fully vaccinated	9.73	1.83	7.06	2.42	2.27	1.19	12.06	2.99
At least one dose	8.44	1.84	7.2	2.45	2.31	1.2	12.32	2.99
No Vaccine	8.44	1.84	7.2	2.45	2.31	1.2	12.32	2.99

Figure 10: Results of Linear Regression

From this table in Figure 10, we can see that the race that had the lowest values of MSE and MAPE were white people. To further this observation let's look at the R-Squared values

```
R-SQUARE SCORE (Fully vaccinated, All Race): -1.08
R-SQUARE SCORE (Fully vaccinated, Black): 0.11
R-SQUARE SCORE (Fully vaccinated, White): 0.56
R-SQUARE SCORE (Fully vaccinated, Hispanic): 0.10
R-SQUARE SCORE (At least one dose, All Race): -0.80
R-SQUARE SCORE (At least one dose, Black): 0.09
R-SQUARE SCORE (At least one dose, White): 0.55
R-SQUARE SCORE (At least one dose, Hispanic): 0.08
R-SQUARE SCORE (No Vaccine, All Race): -0.80
R-SQUARE SCORE (No Vaccine, Black): 0.09
R-SQUARE SCORE (No Vaccine, White): 0.55
R-SQUARE SCORE (No Vaccine, Hispanic): 0.08
```

Figure 11: R-Square values of Linear Regression

Looking at figure 11 the only R-Squared values that explained 50 percent variance were all vaccination status for white people. While not a great amount of variance we can see that the linear regression model explained some variance of attendance values.

Now I wanted to run Linear Regression on the merge dataset without looking at separating by vaccination status and race / ethnicity and came up with a much better result.

```
Evaluation Metrics for Linear Regression:
--------------------------------------------------
Mean Squared Error: 1.36
Mean Absolute Percentage Error: 0.01
R-squared: 0.75
```

Figure 12: Linear Regression on merged dataframe

We see that from Figure 12 we had a much better Mean Squared Error, Mean Absolute Percentage Error and also even a R-Squared value of 0.75 which means 75% of the variance is explained from all columns in the merged dataset.

Random Forest Regression

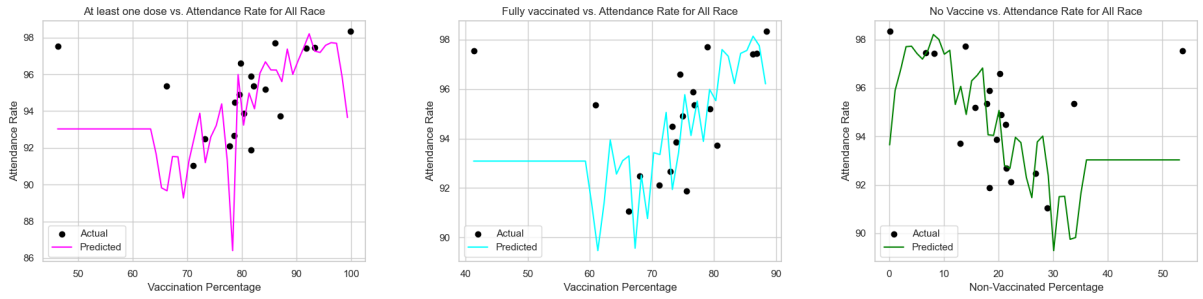

Figure 13: Random Forest Regression on all race

In order to save some space I have combined all vaccination status for each race into an image. For this Figure 13, we have in one dose, full dose and no dose of all races run through a random forest regressor. We see that the graphs are similar to that of the Linear Regression but let's see if there are any differences in MSE, MAPE and R-Squared values.

	All Races (MSE)	All Races (MAPE)	Black (MSE)	Black (MAPE)	White (MSE)	White (MAPE)	Hispanic (MSE)	Hispanic (MAPE)
Fully vaccinated	8.47	2.52	13.24	3.41	2.21	1.16	8.15	2.47
At least one dose	10.31	2.47	13.38	3.31	2.31	1.21	11.32	2.73
No Vaccine	10.39	2.52	13.09	3.26	2.31	1.22	11.49	2.76

Figure 14: Random Forest Regression Results (MSE) and (MAPE)

From Figure 14, we see a similar trend in the Random Forest Regression result table as the Linear Regression table, we have a higher MSE all around and then same MAPE values for each of the races and vaccination status. Next let's look at the R-Squared Values

```
R-SQUARE SCORE (Fully vaccinated, All Race): -0.81
R-SQUARE SCORE (Fully vaccinated, Black): -0.67
R-SQUARE SCORE (Fully vaccinated, White): 0.57
R-SQUARE SCORE (Fully vaccinated, Hispanic): 0.39
R-SQUARE SCORE (At least one dose, All Race): -1.20
R-SQUARE SCORE (At least one dose, Black): -0.69
R-SQUARE SCORE (At least one dose, White): 0.55
R-SQUARE SCORE (At least one dose, Hispanic): 0.16
R-SQUARE SCORE (No Vaccine, All Race): -1.22
R-SQUARE SCORE (No Vaccine, Black): -0.65
R-SQUARE SCORE (No Vaccine, White): 0.55
R-SQUARE SCORE (No Vaccine, Hispanic): 0.15
```

Figure 15: R-Square values of Random Forest

Now once again Figure 15 we see that some races like all races and black, the random forest regression model did not explain any of the variance but once white had around 50 percent of variance explained. Now since we didn't get the greatest results I decided to run the model on just the merged data frame again.

```
Evaluation Metrics for Random Forest Regressor:
--------------------------------------------------
Mean Squared Error: 0.01
Mean Absolute Percentage Error: 0.0
R-squared: 1.0
```

Figure 16: Random Forest Regression on merged dataframe

Now looking at the results of this we see a R-Squared value of 1, which means that the model has explained 100% of the variance and the Mean Squared Error and mean Absolute Percentage Error values of 0.01 and 0.0 shows very accurate predictions. While a perfect fit may seem desirable it could indicate overfitting. We should see how standardizing the data could affect the mode.

```
Evaluation Metrics for Random Forest Regressor w/ StandardScaler:
--------------------------------------------------
Mean Squared Error: 0.01
Mean Absolute Percentage Error: 0.0
R-squared: 1.0
```

Figure 17: Random Forest Regression with Standard Scaling

Even with a standardscaler on the dataset we got the exact same results. It seems like the random forest regression model was a good predictor of attendance rate. This makes sense since random forest models are types of ensemble learning methods.

Decision Tree Regressor

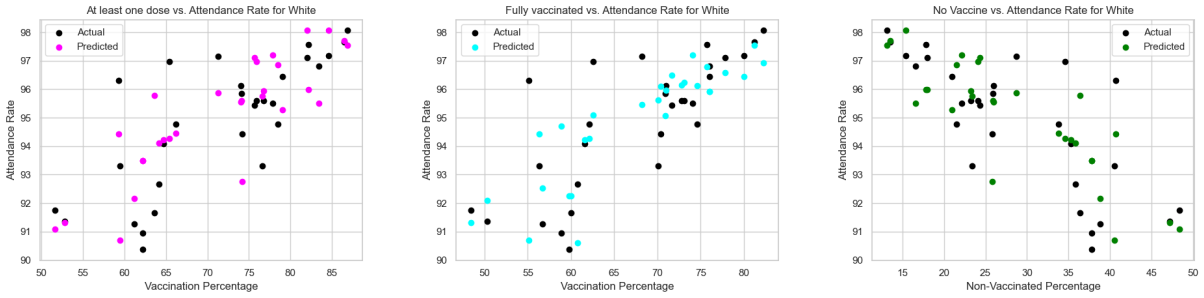

Figure 18: Decision Tree Regressor on All White

From the past 2 models we have seen how ‘White’ has had some success for explained variance so this time instead of showing all races Figure 18 shows a scatter plot of the actual and predicted values from the decision tree regressor model. The black dots are the actual data points from the ‘white people’ data set while the magenta, cyan and green points show the predicted ones. The colors are meant to separate each of the doses of vaccine so we are not confused. Looking at each of the scatterplots we see that the predicted points are quite close to the actual values. So hopefully we see a low MSE and MAPE for ‘white’.

	All Races (MSE)	All Races (MAPE)	Black (MSE)	Black (MAPE)	White (MSE)	White (MAPE)	Hispanic (MSE)	Hispanic (MAPE)
Fully vaccinated	12.56	2.99	21.05	4.3	2.86	1.34	10.55	2.68
At least one dose	21.38	3.32	22.98	4.19	2.92	1.48	16.49	3.1
No Vaccine	21.42	3.37	21.94	4.08	2.93	1.48	16.49	3.1

Figure 19: Decision Tree Regressor Results

Analyzing the table in Figure 19, we see that our prediction of a lower MSE and MAPE value for ‘white’ while the other races have higher values. Looking into the MSE values for each vaccination status we see that the fully vaccinated individuals generally have a lower MSE compared to those who have received either one dose or no dose, indicating a better predictive performance from this subset. Doing the same thing with MAPE we see a similar pattern, fully vaccinated individuals generally have lower MAPE, suggesting better accuracy in predicting attendance rates.

```
R-SQUARE SCORE (Fully vaccinated, All Race): -1.68
R-SQUARE SCORE (Fully vaccinated, Black): -1.65
R-SQUARE SCORE (Fully vaccinated, White): 0.44
R-SQUARE SCORE (Fully vaccinated, Hispanic): 0.21
R-SQUARE SCORE (At least one dose, All Race): -3.57
R-SQUARE SCORE (At least one dose, Black): -1.90
R-SQUARE SCORE (At least one dose, White): 0.43
R-SQUARE SCORE (At least one dose, Hispanic): -0.23
R-SQUARE SCORE (No Vaccine, All Race): -3.58
R-SQUARE SCORE (No Vaccine, Black): -1.77
R-SQUARE SCORE (No Vaccine, White): 0.43
R-SQUARE SCORE (No Vaccine, Hispanic): -0.23
```

Figure 20: R-Squared Values for Decision Tree Regressor

In Figure 20, we have the R-Squared values from the decision tree regressor. We see that once again the ‘white’ race indicates that the model is explaining some of the variance but worse than past models. Also looking at the vaccine status of one dose and no dose, we see all negative values, aside from ‘white’, meaning the model is also not explaining much of the variance of attendance rates in these 2 vaccination status.

```
Evaluation Metrics for Decision Tree Regresssor:
-------------------------------------------------
Mean Squared Error: 0.01
Mean Absolute Percentage Error: 0.0
R-squared: 1.0
```

Figure 21: Decision Tree Regression on merged dataframe

Now running the Decision Tree Regressor Model on the original data frame we got the same results as the Random Forest Regressor Model. With these results it would make sense since random forest is a method where the model is made up of decision trees. With these results we can understand why the results of the random forest model did so well, decision trees did great as well. Now since a perfect R-Squared value raises concern about overfitting we will do some hyperparameter tuning using GridSearchCV.

```
Evaluation Metrics for Decision Tree Regressor w/ Hyperparameter tuning---------------
Best Parameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5}
Tuned Model Metrics:
Mean Squared Error: 0.01
Mean Absolute Percentage Error: 0.0
R-squared: 1.0
```

Figure 22: Decision Tree Regression with hyperparameter tuning

Now after successfully performing hyperparameter tuning on the decision tree regressor using GridSearchCV, found the best hyperparameters are a max_depth of 20, min_sample_leaf of 1 and min_samples_split of 5. With the tuning we resulted with the same MSE, MAPE, and R-Squared values.

XGBoost Regressor

	All Races (MSE)	All Races (MAPE)	Black (MSE)	Black (MAPE)	White (MSE)	White (MAPE)	Hispanic (MSE)	Hispanic (MAPE)
Fully vaccinated	14.99	3.26	19.38	4.14	3.01	1.42	19.46	3.8
At least one dose	18.99	3.11	15.54	3.35	2.24	1.25	17.54	3.45
No Vaccine	8.35	2.47	18.51	3.94	1.98	1.23	16.37	3.11

Figure 23: XGBoost Regressor Results

From Figure 23, we have the results of using XGBoost regression, it seemed like once again the ‘white’ race has the lowest values compared to the different other races, with the ‘black’ race performing the worst, especially with the fully vaccinated status.

```
R-SQUARE SCORE (Fully vaccinated,All Race): -2.20
R-SQUARE SCORE (Fully vaccinated,Black): -1.44
R-SQUARE SCORE (Fully vaccinated,White): 0.41
R-SQUARE SCORE (Fully vaccinated,Hispanic): -0.45
R-SQUARE SCORE (At least one dose,All Race): -3.06
R-SQUARE SCORE (At least one dose,Black): -0.96
R-SQUARE SCORE (At least one dose,White): 0.56
R-SQUARE SCORE (At least one dose,Hispanic): -0.31
R-SQUARE SCORE (No Vaccine,All Race): -0.78
R-SQUARE SCORE (No Vaccine,Black): -1.33
R-SQUARE SCORE (No Vaccine,White): 0.62
R-SQUARE SCORE (No Vaccine,Hispanic): -0.22
```

Figure 24: R-Squared Values of XGBoost Regression

In Figure 24, the model 'white' race explains some sort of variance of the attendance rate. Aside from the 'white' race all other R-Squared values were negative no matter the vaccination status, showing that XGBoost is not a good model to explain the variance in attendance rates as whole, except for maybe the 'white' race.

```
Evaluation Metrics for XGBoost Regressor:
--------------------------------------------------
Mean Squared Error: 0.02
Mean Absolute Percentage Error: 0.07
R-squared: 1.0
```

Figure 25: XGBoost Regression with merged dataframe

We once again run the XGBoost model onto the original merged data frame and get the results as shown in Figure 25. We see a similar result in terms of the R-Squared value comparing it to the Random Forest and Decision Tree models but we see that the MSE and MAPE have increased a little bit. But that aside we still see an R-Squared value of 1 which suggests that there is potential overfitting, so we will conduct another hyperparameter tuning process for the model.

```
Best Parameters: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200}

Evaluation Metrics for Tuned XGBoost Regressor:
--------------------------------------------------
Mean Squared Error (Tuned): 0.02
Mean Absolute Percentage Error (Tuned): 0.08
R-squared (Tuned): 1.0
```

Figure 26: XGBoost Regression w/ Hyperparameter Tuning Result

Now in Figure 26, we see the best parameters are a learning_rate of 0.1, max_depth of 7 and number of estimators is 200. With the best parameters we have the same MSE and R-Squared value and a different MAPE of 0.08 (compared to a 0.07). We still have a R-Squared value of 1 so we did create a perfect fit to the data.

Support Vector Regression

	All Races (MSE)	All Races (MAPE)	Black (MSE)	Black (MAPE)	White (MSE)	White (MAPE)	Hispanic (MSE)	Hispanic (MAPE)
Fully vaccinated	9.04	1.83	7.72	2.49	2.34	1.18	12.12	2.98
At least one dose	7.42	1.8	7.74	2.51	2.4	1.21	11.99	2.85
No Vaccine	7.42	1.8	7.74	2.51	2.4	1.21	11.99	2.85

Figure 27: Support Vector Regression Results

We see in Figure 27, grouping the Fully vaccinated group the MAPE for ‘all races’ and ‘white’ are relatively low, suggesting good accuracy of the model, and ‘hispanic’ having the highest values. Onto the one vaccination status we see that within one dose and no dose we have the same pattern in which ‘all races’ and ‘white’ have the lower MAPE values suggesting better accuracy compared to ‘hispanic’

```
R-SQUARE SCORE (Fully vaccinated, All Race): -0.93
R-SQUARE SCORE (Fully vaccinated, Black): 0.03
R-SQUARE SCORE (Fully vaccinated, White): 0.54
R-SQUARE SCORE (Fully vaccinated, Hispanic): 0.10
R-SQUARE SCORE (At least one dose, All Race): -0.59
R-SQUARE SCORE (At least one dose, Black): 0.02
R-SQUARE SCORE (At least one dose, White): 0.53
R-SQUARE SCORE (At least one dose, Hispanic): 0.11
R-SQUARE SCORE (No Vaccine, All Race): -0.59
R-SQUARE SCORE (No Vaccine, Black): 0.02
R-SQUARE SCORE (No Vaccine, White): 0.53
R-SQUARE SCORE (No Vaccine, Hispanic): 0.11
```

Figure 28: R-Squared Values for Support Vector Regression

Interestingly looking at Figure 28, we see positive values for all the races except for ‘All Race’. Each racial group has a different R-Squared value which tells us that the mode performs differently for various racial groups within each vaccination status category. We should try to understand the factors that contribute to these variants in order to improve the model, which is hard to do with the limited data we have.

```
Evaluation Metrics for SVR Regressor:
-------------------------------------------------
Mean Squared Error: 1.42
Mean Absolute Percentage Error: 1.0
R-squared: 0.74
```

Figure 29: SVR on merged dataframe

Looking at Figure 29, we have the evaluation metrics of SVR on our original merged data frame we have Mean Square Error of 1.42, meaning that the squared difference between predicted and actual attendance rates is 1.42. The Mean Absolute Percentage Error is 1 meaning on average the model predictions deviate by 1 percent from the actual values. The R-Squared value of 0.74 suggests that the model explains 75% of the variance in the attendance rate. ‘

In this next section I just wanted to show the graphs / plots created for each of the models above.

Linear Regression

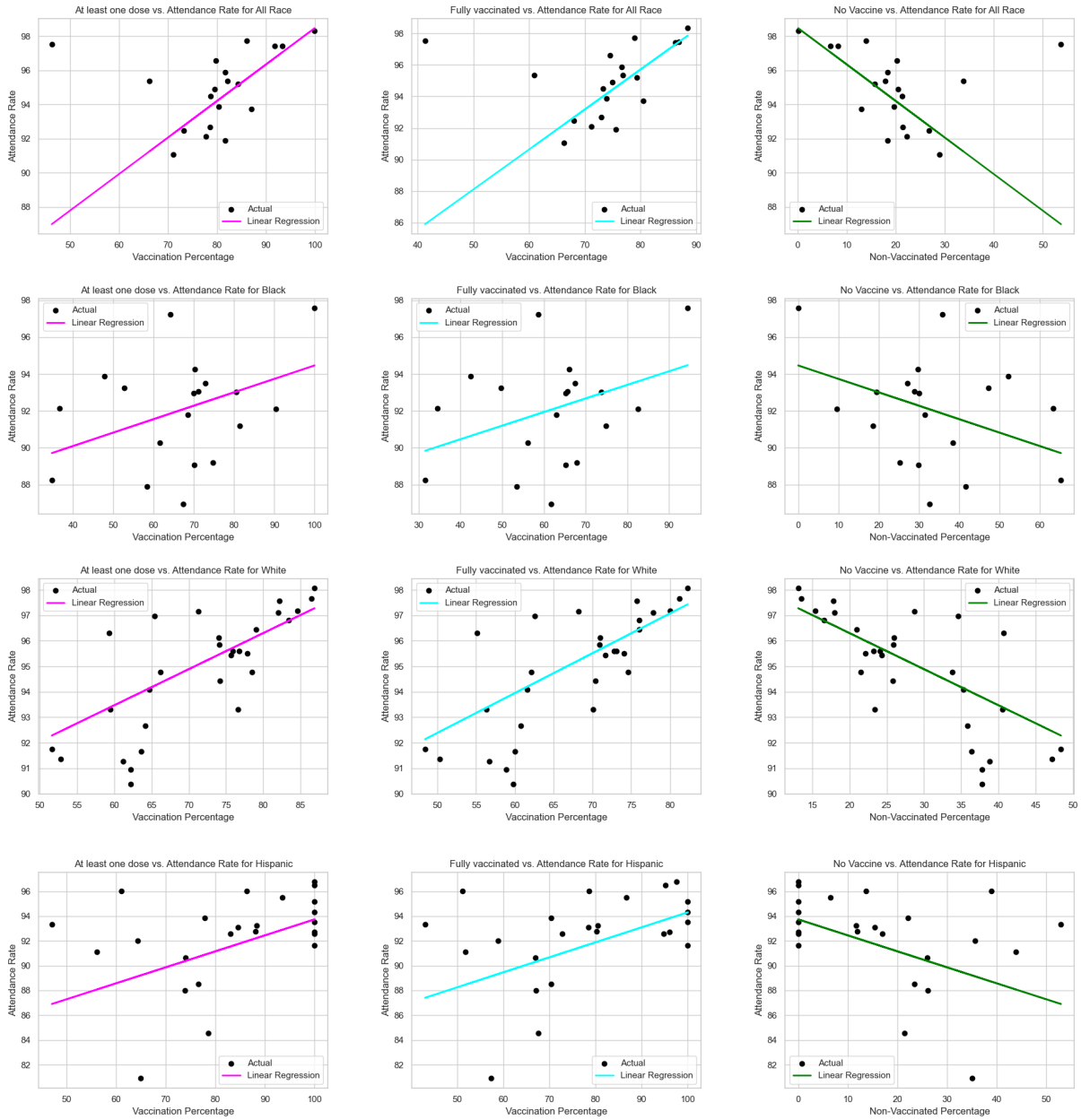

Figure 30: Linear Regression Model Plots

Each of the rows shows the linear regression of each race while each column is the vaccination status. This shows a positive correlation between vaccination rates and attendance rates for at least one dose and fully dosed and a negative correlation between non vaccinated percentage and attendance rate.

Random Forest Regressor

Figure 31: Random Forest Regressor Plots

Each of the rows shows the random forest regressor of each race while each column is the vaccination status. We see that for All Race (Row 1) and White (Row 2) plots seem to follow in line with the actual points plotted.

Decision Tree Regressor

Figure 32: Decision Tree Regressor Plots

Each of the rows shows the scatter plots of actual vs. predicted values of the Decision Tree Regressor of each race while each column is the vaccination status. In 'Black' (Row 2) we can see how all the colored points (predicted) are not close to the actual points (black). While 'All Race' (Row 1) and 'White' (Row 3) the predicted points are very close to the actual points.

XGBoost

Figure 33: XGBoost Regressor Plots

Each of the rows shows the scatter plots of actual vs. predicted values of the XGBoost Regressor of each race while each column is the vaccination status. We see that “All Race” (Row 1) and “White” (Row 3) also have the predicted points close to the actual points while the ‘Black” (Row 2) and “Hispanic” (Row 4) predicted points are farther away from the actual points.

Support Vector Regressor

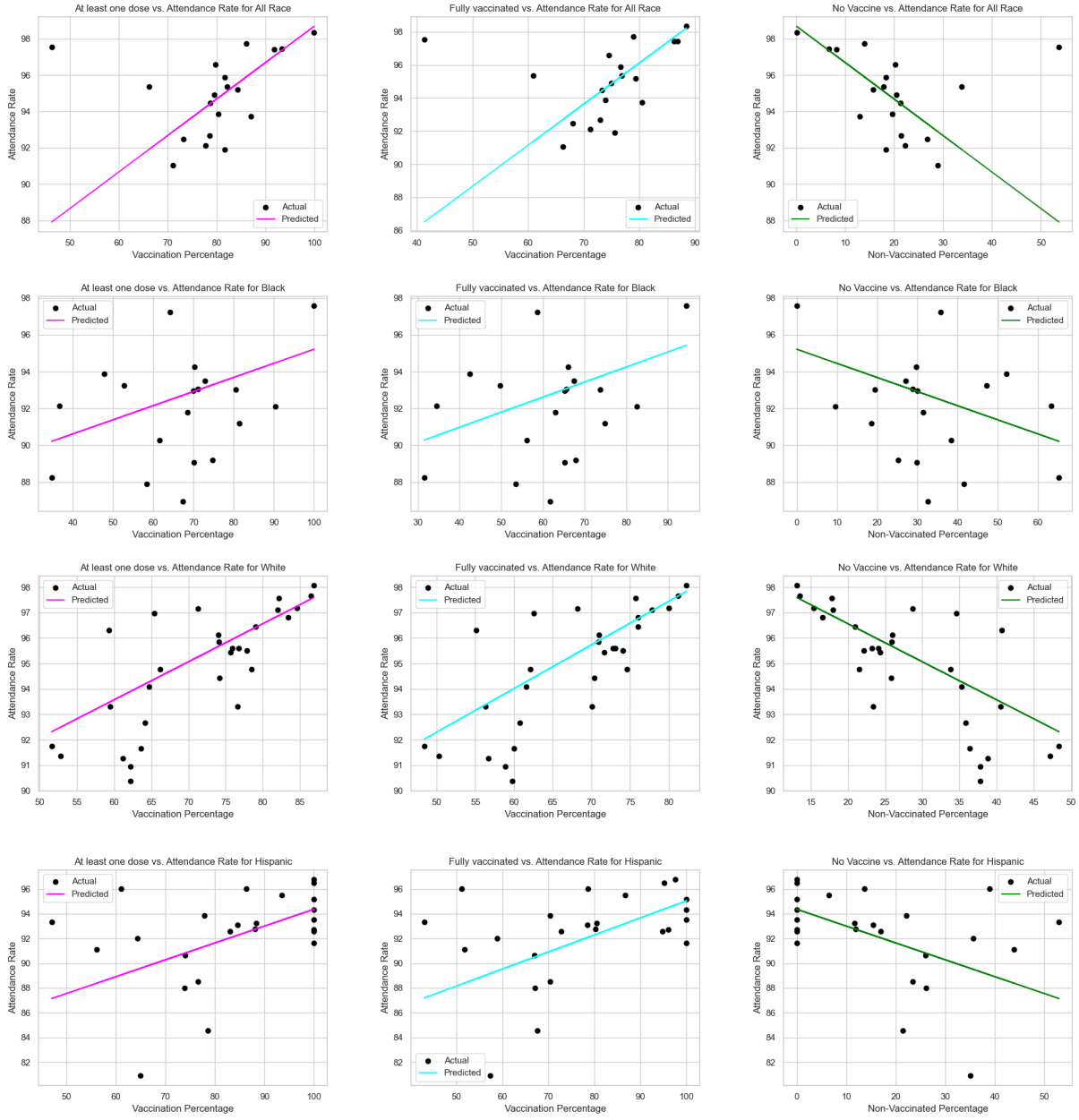

Figure 34: SVR Plots

Each of the rows shows a scatter plot of actual data points and then the Support Vector Regressor Line of each race while each column is the vaccination status. It seems like “All Races” and “Whites” points and lines seem to be close together and a good model while the “Black” and “Hispanic” points are not near the line.

5. Conclusions and Discussion
    a. In this project, our goal was to explore the relationship between vaccination status and percentages, demographic factors, and attendance rates in school. We employed 5 machine learning models, including Linear Regression, Random Forest, Decision Tree Regressor, XGBoost Regressor, and Support Vector Regression, to predict attendance rates based on our 2 datasets.
    For model evaluation we saw that the Random Forest Regressor, Decision Tree Regressor and XGBoost Regression models achieved a perfect R-Squared score of 1.0 on just the merged datasets. These models demonstrated remarkable abilities to capture the patterns in the data and make accurate predictions. The Linear Regression and Support Vector Regression model performed well as well with an R-Squared value of 0.75 and 0.74 respectively.
    Just from this knowledge about our 3 perfect models we can see why these performed perfectly with the data. For Random Forest and Decision Tree we saw the connection between them due to the fact that Random Forest is an ensemble learning method which is made up of multiple Decision Trees. If a Decision Tree has a perfect fit then multiple Decision Trees would also create a perfect fit. Now while XGBoost is also an ensemble method it doesn't use Decision Trees like Random Forest, it starts off with a Decision Tree as the 'learner', in other terms it also bases it's model on a Decision Tree.
    Our key findings that the status of vaccination had an impact. Our analysis relieved that fully vaccinated individuals showed a negative R-Squared score for all demographic groups, indicating challenges in predicting attendance solely based on vaccination status, the results for at least one dose or no vaccine at all were more promising than fully vaccinated. The second finding was that there are demographic variations, the impact of vaccination status on attendance rates were different across all racial / ethnic groups, but with some models showing a positive R-Squared scores for white individuals may show that there is a stronger predictive ability for this demographic compared to others.
    Throughout the project, we make progress in terms of data exploration, analysis and model selection. Initially exploring the data involved understanding the distribution of attendance rates by towns and then exploring potential correlation with vaccination status. Then after the exploration of data we used models to make predictions based on our chosen features. With using 3 tree-based models it proved very effective but the linear regression and SVR added a different look and perspective on the data.
    Next steps in this would be to explore additional features to transform existing features that may enhance the models performance. We also may look for future Connecticut datasets dealing with attendance rates and see if our models would

see how accurate the models would be at predicting the attendance rates in said future.

In conclusion, the project explored correlations between vaccination status, vaccination percentage, demographics, and attendance rates in school. We used machine learning models that showed strong predictive capabilities and could offer insights for educators or policymakers. As with any project with limited time and resources there always can be more techniques or data to get a better understanding of factors that influence attendance rates in school.

6. Poster
    a. Within Zip File
    b. Github Link