The goal of our project was to investigate the MMR rate across schools in 32 states. We used the Wall Street Journal's (WSJ) US Measles Vaccination Data in our analysis: https://github.com/WSJ/measles-data . The dataset contains data on 32 states and their MMR rates, vaccination exemptions and enrollment count. Our group decided to pursue four algorithms to analyze the data. Kristin pursued decisions trees and random forests. Ben pursued KNN and I pursued linear regression. Each of us coded our own python files with respect to our chosen algorithms. We then combined our codes into one Pyqt5 file for presentation.

The dataset posed a challenge for the linear regression algorithm. The data contained many missing values which had to be imputed. Missing values for Oregon are shown below:

```
Missing Values:
index          0
state          0
year        4681
name           0
type       27174
city       17339
county      5158
district   39009
enroll     12844
mmr            0
overall        0
xrel       34270
xmed       33439
xper       40000
dtype: int64
```

To address the missing values, the means of each column were imputed. I further subset the data to include only numeric variables. I then regressed mmr on enroll, overall, xrel, xmed and xper.

The linear model performed well for several states but performed poorly for others. I suspect the states where the linear model performed poorly are highly imputed. The lack of variation in the data may cause the linear model to fail.

Oregon is a state where the linear model performed well. The coefficient on enroll suggest a one student increase in total enrollment increases the MMR rate by 0.000106percent. The coefficient on the overall MMR rate increases the MMR rate by 0.0983 percent. The coefficient on xmed suggests a one student increase in medical exemption decreases the MMR rate by -

0.146 percent. The signs of the coefficients on xrel and xper do not make sense and suggest poor data and heavy bias for those variables.

```
         Coefficient
enroll    1.068379e-04
overall   9.834023e-01
xrel      5.551115e-17
xmed     -1.466286e-01
xper      1.406847e-01
Mean Absolute Error: 1.6435932013181025
Mean Squared Error: 10.018183429125447
Root Mean Squared Error: 3.1651514069828393
```

Despite probable bias in xrel and xper, the model performs well and displays a score of 93 percent.

```
         Actual  Predicted
36347  97.116844  96.141501
36338  97.202797  97.310538
36165  99.337748  94.200416
36182  98.920863  99.016867
36541  95.238095  96.539395
...          ...        ...
36718  92.950392  93.740052
36840  80.530973  78.571371
36417  96.509240  94.228215
36655  93.989071  95.057575
36390  96.782842  98.233714

[147 rows x 2 columns]
The score of the model:  0.93
```

Montana is a state where the linear model performed poorly. The coefficient on enroll suggests that a one student increase in enrollment decreases the MMR rate by -0.0029 percent. The coefficient on overall, xrel and xper are zero because of high imputation. The coefficient on xmed suggests that a one student increase in medical exemption decreases the MMR rate by -10 percent.

```
         Coefficient
enroll     -0.002925
overall     0.000000
xrel        0.000000
xmed      -10.282346
xper        0.000000
Mean Absolute Error: 32.40150632638283
Mean Squared Error: 1477.8715426690649
Root Mean Squared Error: 38.44309486330497
```

The linear model for Montana does not perform well. It displays large errors and maintains a score of only 2.7 percent.

```
       Actual  Predicted
25140    0.00  68.356510
25085    0.00  68.356510
25067   60.14  67.872564
24926   95.00  68.650577
24665  100.00  68.638877
...       ...        ...
24841   97.33  68.050980
25147    0.00  68.356510
25060   68.97  29.611823
25144    0.00  68.356510
25128    0.00  68.356510

[114 rows x 2 columns]
The score of the model:  0.027
```

Overall the linear model performs well in states that are not heavily imputed and poorly in states that are. Given a state that is not heavily imputed, the linear model is able to predict accurately the MMR rate. The variables that affect the MMR rate the most are the exemption variables: xrel, xmed and xper. These exemption variables significantly decrease the MMR rate.

The linear model would benefit greatly from improvements in data quality. Obtaining data on all 50 states and decreasing the amount of missing data would be a great improvement. I used 5 percent of code from the internet and generated 95 percent of my own.