Analysis of COVID-19 vaccination disparity with socioeconomic & demographic data

Akash Sonowal (20114003); Prince Singh (20114014)

IME672 Project **Fall, 2021**

1. INTRODUCTION

In this work, our objective is to understand the racial disparity rate in COVID-19 Vaccination in the US with the help of socioeconomic privilege and political ideology data. As outlined in the work done by Agarwal et. al., it is very important to address these issues at all levels, as all these disparities if persisted can take additional lives which is very disastrous and would lead to unrest in the world. So, to study the relevant factors that affect our outcome of disparity in covid-19 vaccination rate, we have employed various machine learning algorithms that help us better predict the future with the most important features and also studying the direct correlations of variables with our output variable, COVID-19 disparity rate. However, there are some constraints that we are bound to while conducting this study- among them are we have data limited to only a few counties of the US.

In our work firstly, we have done the preliminary analysis (**exploratory data analysis**) of all the variables, dataset preparation (**normalization, outlier removal, numerical encoding of categorical variables, dropping less important variables**). Then we have fitted linear and non-linear models to our dataset for predicting our outcome variable. We have also performed hyperparameter tuning wherever necessary and reported the three best models (details of other models are provided in the Appendix). Finally, we have plotted **SHAP (SHapley Additive exPlanations)** to understand the final effect of different variables in our model output.

2. ANALYSIS

2.1 Data Description:

Our dataset has 756 entries and 18 variables with 1 dependent variable (CvdVax_DisparityY) and 17 independent variables featuring information about demography, socioeconomic status etc.

2.2 Exploratory Data Analysis:

To understand our data better, we did the plotting of each variable with our output variable. For numerical variables, we have plotted scatter plots, histograms (with KDE plots), box plots to understand the feature characteristics like linearity, normality of data, presence of outliers. For categorical variables, we have plotted violin plots to understand their impact on the output variable distribution.

Note: Because of constraints in the page limits we have only mentioned the plots of two numerical variables (**IT_WholeRate** and **MediaInc_Disparity**) out of 16 variables given in our dataset. And for categorical variables, only the plot for the **State** variable is shown. For other variables, we have mentioned them in the Appendix.

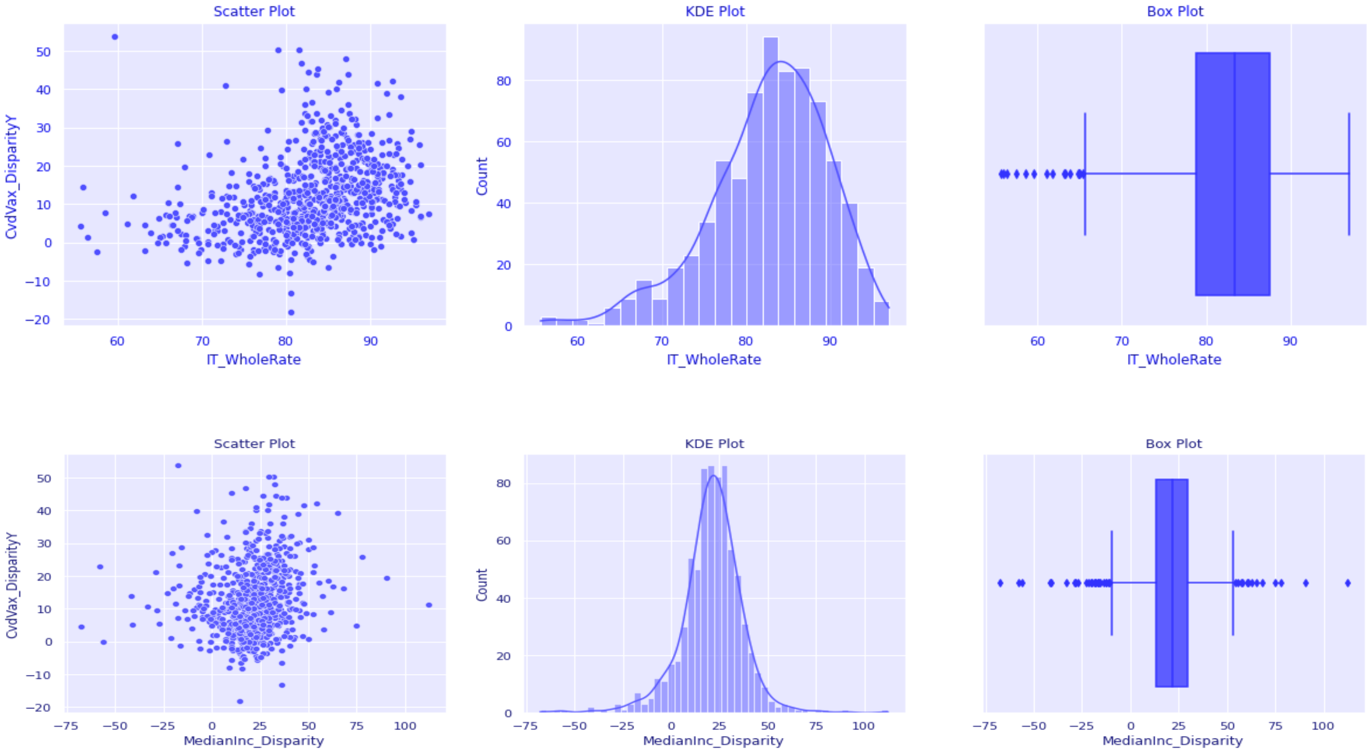


Fig.1 Exploratory Data Analysis of IT_WholeRate and MedianInc_Disparity

In Fig. 1, we see that **IT_WholeRate** is positively correlated with the output variable, it is left-skewed. Likewise, the **MedianInc_Disparity** is **Leptokurtic** (high kurtosis) distribution. Similar inferences can be made for all other numeric variables.

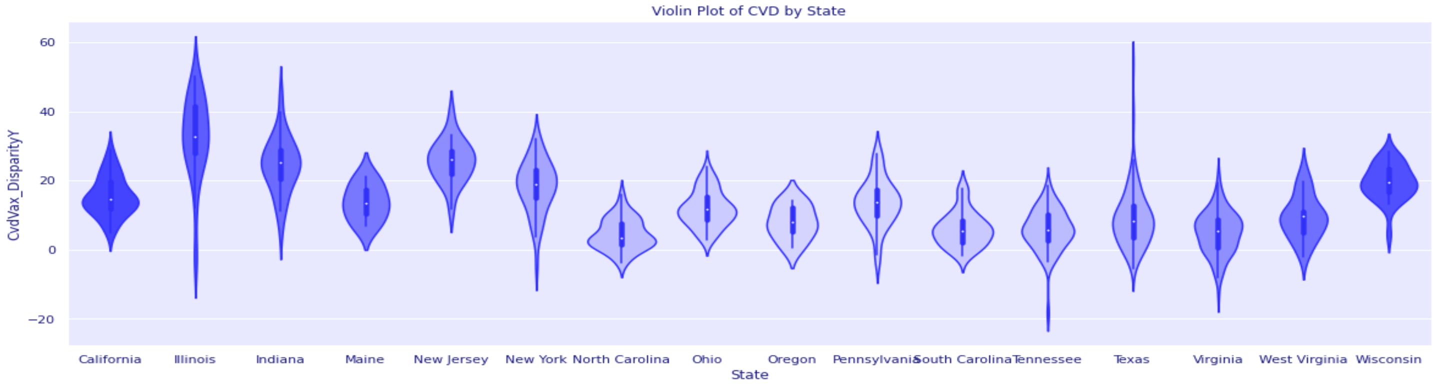


Fig.2 Exploratory Data Analysis of Categorical variable State

In Fig. 2, we can infer the effect of different state locations on the output variable. The most important observations are: Illinois state has the most variation in the output variable indicating that mostly all counties in Illinois have different covid-19 vaccination disparity rates.

3. DATA PREPARATION

In this work, our main task is to predict the output variable which is continuous. As known from the standard pieces of literature on the Internet, behind every model there are many assumptions. So, in our regression problem, we have tried to fulfil all assumptions of a linear regression model. Although we don't limit ourselves just to the linear regression model, it is a heuristic that works well for other models as well.

Below are some of the assumptions and data structures we have tried to achieve in our dataset.

3.1 Linear Assumption: The relationship between input variables and output variables is linear. In our dataset, almost all variables had linear relationships with output variables except a few binary and categorical variables.

3.2 Remove Multicollinearity: When there is multicollinearity in our data features, we cannot determine the effect of individual variables in our output variable leaving us with an inefficient model specification.

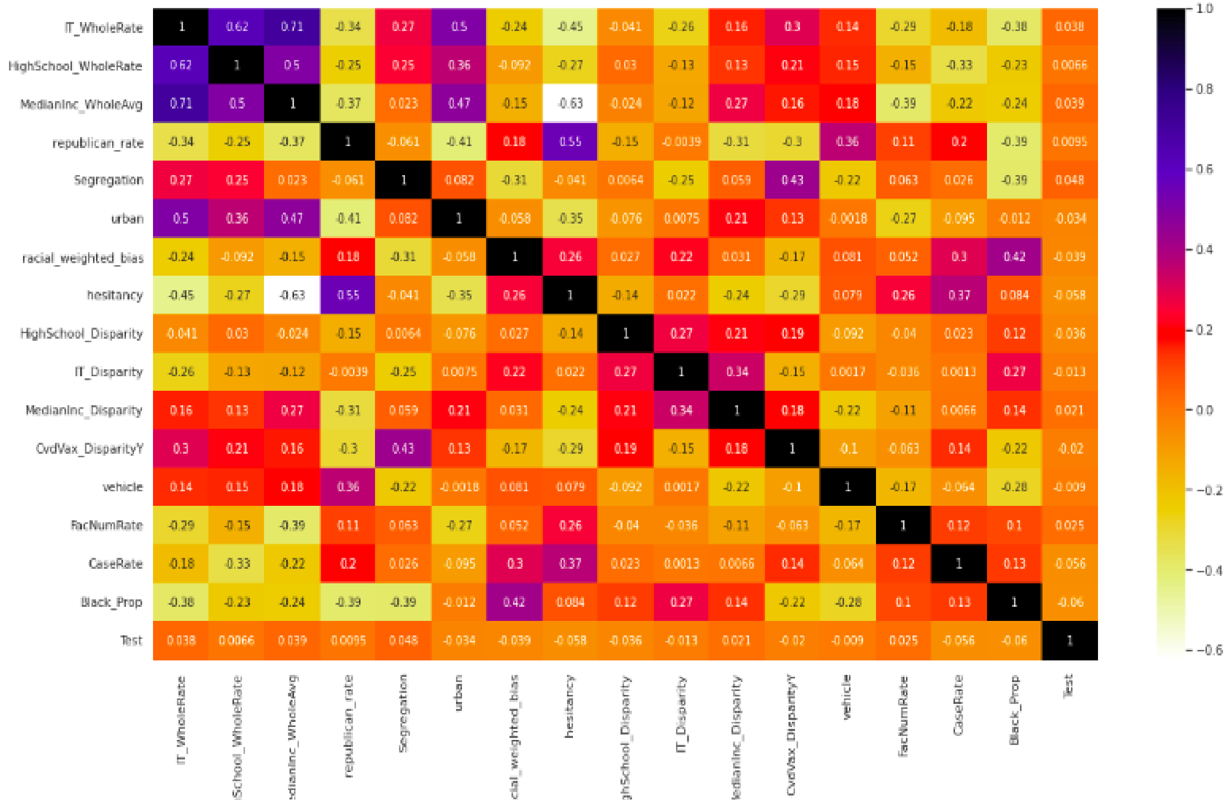


Fig.3 Correlation plot between all numeric continuous variables.

In our dataset, we have determined multicollinearity from the correlation matrix and removed one among variables that form highly correlated pairs. To be specific we have dropped HighSchool_WholeRate and MedianInc _WholeAvg as they have a high correlation with IT_WholeRate because of Pearson correlation values greater than 0.5 i.e., 0.62 and 0.71. This was further verified with VIF scores calculated later which had values greater than 10.

3.3 Gaussian Distribution: Our models will make more reliable predictions if our input and output variables have a Gaussian distribution. In practice, we have techniques like log transform, square-root and box-cox transform which normalise our data. However, in our dataset, we achieved the best transformation with the yeo-johnson transformation (an alternate version of box-cox power transform).

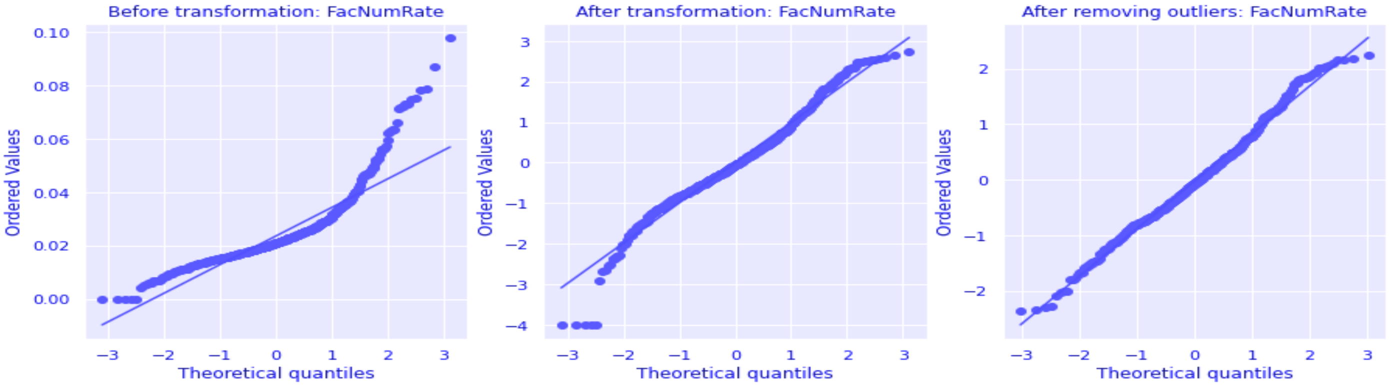


Fig. 4 Q-Q plot while applying Yeo-johnson transformation to our variable FactNumRate

3.4 Remove noise: Linear regression assumes that your input and output variables are not noisy. The model fitting is largely affected by the presence of outliers. As seen from Fig. 4, outliers get removed with transformations but sometimes it resides. So, to remove all outliers we have used the interquartile range (IQR) method to discern outliers.

Outlier removal with upper limit = 75th percentile + 1.5*IQR
Outlier removal with lower limit = 25th percentile - 1.5*IQR

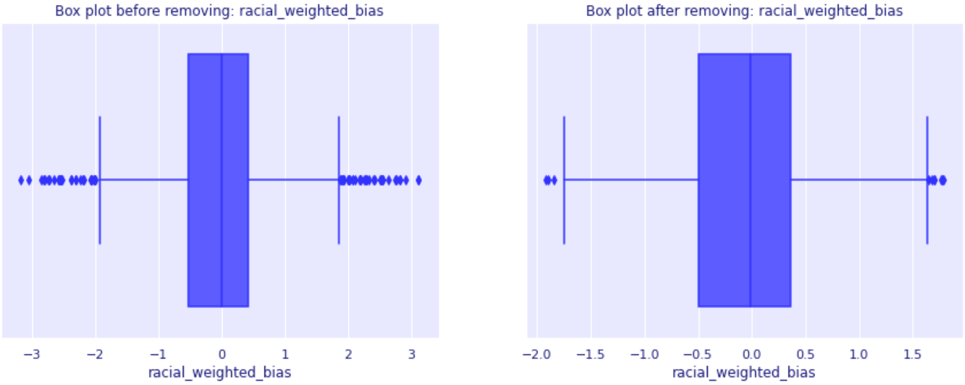


Fig. 5 Box plots with and without outliers in racial weighted bias variable.

3.5 One Hot Encoding: As our model can take values only in number, so we have to convert our categorical variables into binary variables.

3.6 Rescaling inputs: All variables need to have values on a similar scale so that they have equal impact in terms of magnitude on the output variable for linear regression to make more reliable predictions. We have employed a **standard scalar** method to standardize our data.

Please note that we have rescaled only numerical variables and not binary variables.

3.6 Train and Test split: We have split our dataset into train and test based on the test column attribute values given in our dataset.

4. MODELS

We have applied various models like Linear Regression with modifications like Robust Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, other regression methods like Support Vector Regression, Random Forest Regression and also complex non-linear methods like Neural Network.

To evaluate our models, we have considered out-of-sample R^2 on test data. Moreover, for a sanity check, we have also calculated values of MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) in our test data.

Furthermore, we have also performed hyperparameter tuning for models like Ridge, Lasso, SVR, Random Forest to find the best models. To perform this, we have used 10-Fold cross-validation with grid search to optimize the hyperparameters.

5. RESULTS

After hyperparameter tuning in our models, we find the best models with the following specifications:

Ridge regression: alpha = 0.76
Lasso regression: alpha = 0.001
Elastic Net regression: alpha = 0 and ratio = 0.4
Random Forest regression: n_estimators = 1000
Support vector regression: C = 1, kernel = 'rbf'

Although these hyperparameters gave good results in their own variant, we have only selected models with good final R^2 scores in different variants (i.e., we have chosen only one linear regression model among its variants). The results of the models with the out of sample R^2 and other metrics are mentioned in the table below.

MODEL	MAE	MSE	RMSE	R^2
Linear Regression	0.093158	0.014786	0.121597	0.612138
Robust Regression	0.093878	0.015197	0.123275	0.601364
Ridge Regression	0.084072	0.013023	0.114116	0.658394
Lasso Regression	0.085159	0.013262	0.115161	0.652112
Elastic Net Regression	0.085159	0.013262	0.115161	0.652112
Artificial Neural Network	0.098548	0.016739	0.129378	0.560914
Random Forest Regression	0.102353	0.017758	0.133257	0.534186
Support Vector Regression	0.075872	0.010248	0.0101231	0.731185

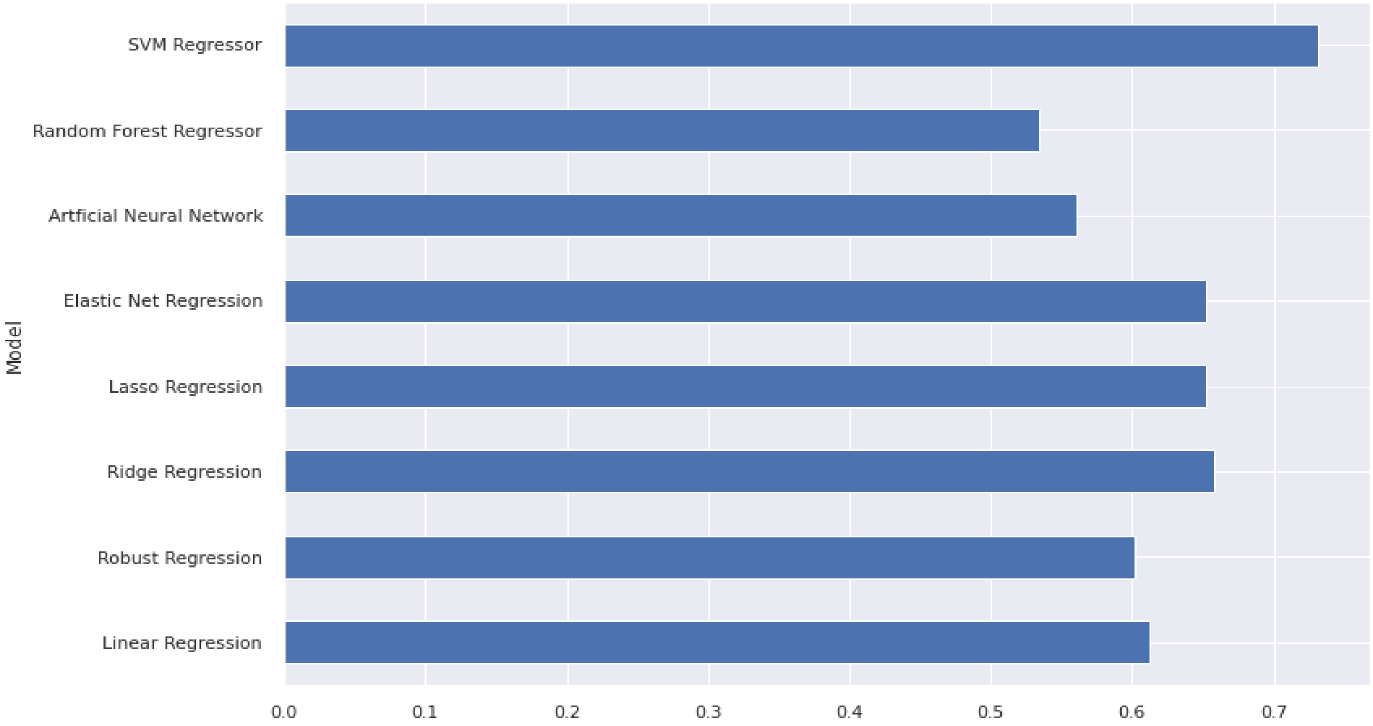


Fig. 6 Comparison of different models based on R^2 scores.

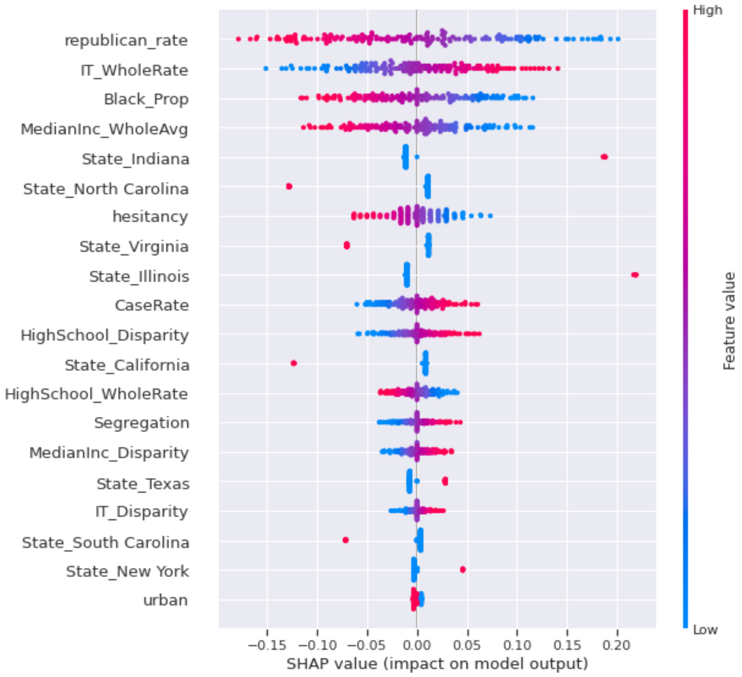
As it is apparent from the table above, the Support Vector Regressor gave the best results with an R^2 value of 0.731185 among others like Ordinary Least Squares Regression, Random Forest, shrinkage methods like LASSO, Ridge Regression and complex nonlinear methods like neural networks.

The reason for good performance in terms of curve fitting of SVR can be attributed to the nature of linearly separable data we are given and also to the fact how SVR is fitted according to the **epsilon insensitivity method which doesn't allow overfitting and thus give a better R^2 in the test data.**

5.1 Model Interpretation of the three models chosen:

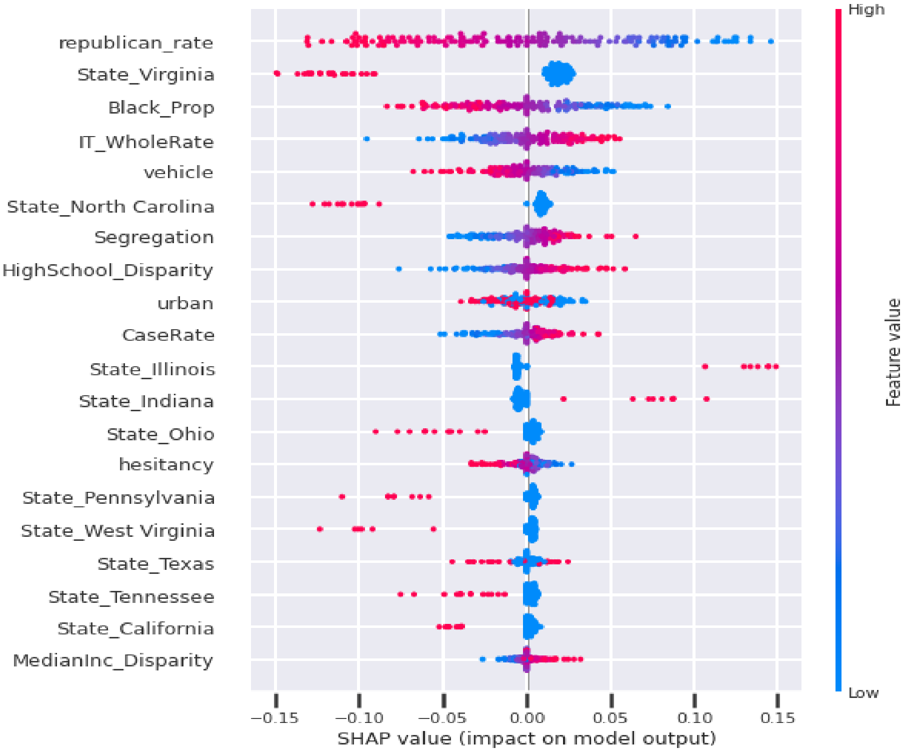
To interpret the models, we have given a summary plot: SHAP (SHapley Additive exPlanations), a game-theoretic approach to explain the output of any machine learning model that combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value. The colour represents the value of the feature from low to high. Overlapping points are jittered in the y-axis direction, so we get a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance. In the summary plot, we see the first indications of the relationship between the value of a feature and the impact on the prediction.

Ordinary Least Squares Regression:



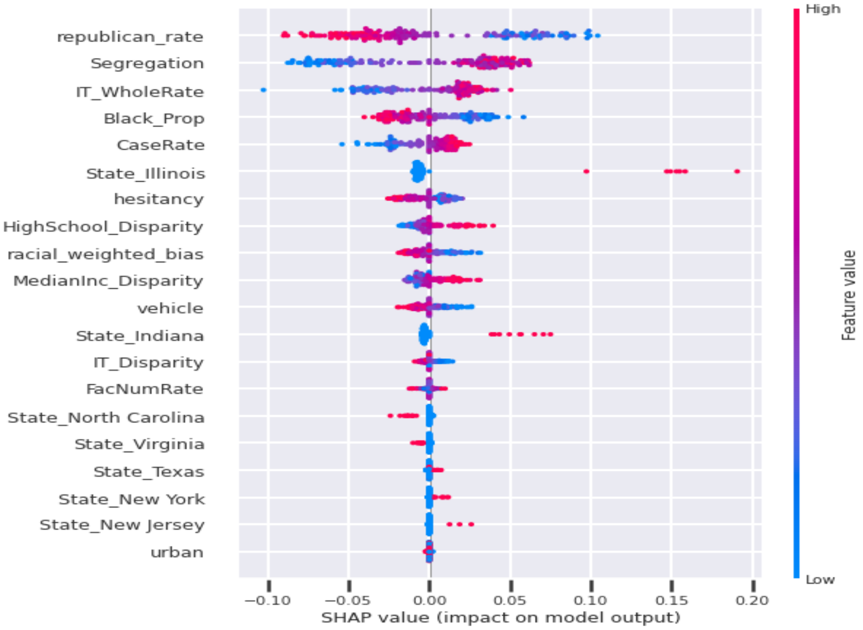
In the above plot, we can see that the most important feature is the republican rate which is negatively correlated with disparity. Furthermore, IT_WholeRate is positively correlated with disparity, Black_Prop and MedianInc_WholeAvg are negatively correlated with disparity. Similarly, for other variables, we can make the related observation from the above plot.

Support Vector Regressor:



Similar to OLS, we have the most important variable as the republican rate which has a negative correlation with disparity. Furthermore, State_Virginia and Black_Prop are a negative correlation with disparity and IT_WholeRate is positively correlated with disparity.

Random Forest Regressor:



Here, again, the republican rate is the most important feature which is negatively correlated with disparity. Segregation, IT_WholeRate are positively correlated with disparity and Black_Prop is negatively correlated with disparity

6. CONCLUSION

In this work, the final regression coefficients showed similar nature in terms of negative or positive affecting variables which is reported in the tables below. Table 1 is by authors and Table 2 is from our OLS summary. It is to be noted that some variables that are dropped (based on VIF score) by authors like hesitancy are not dropped by us as they didn't give a high VIF score in our dataset.

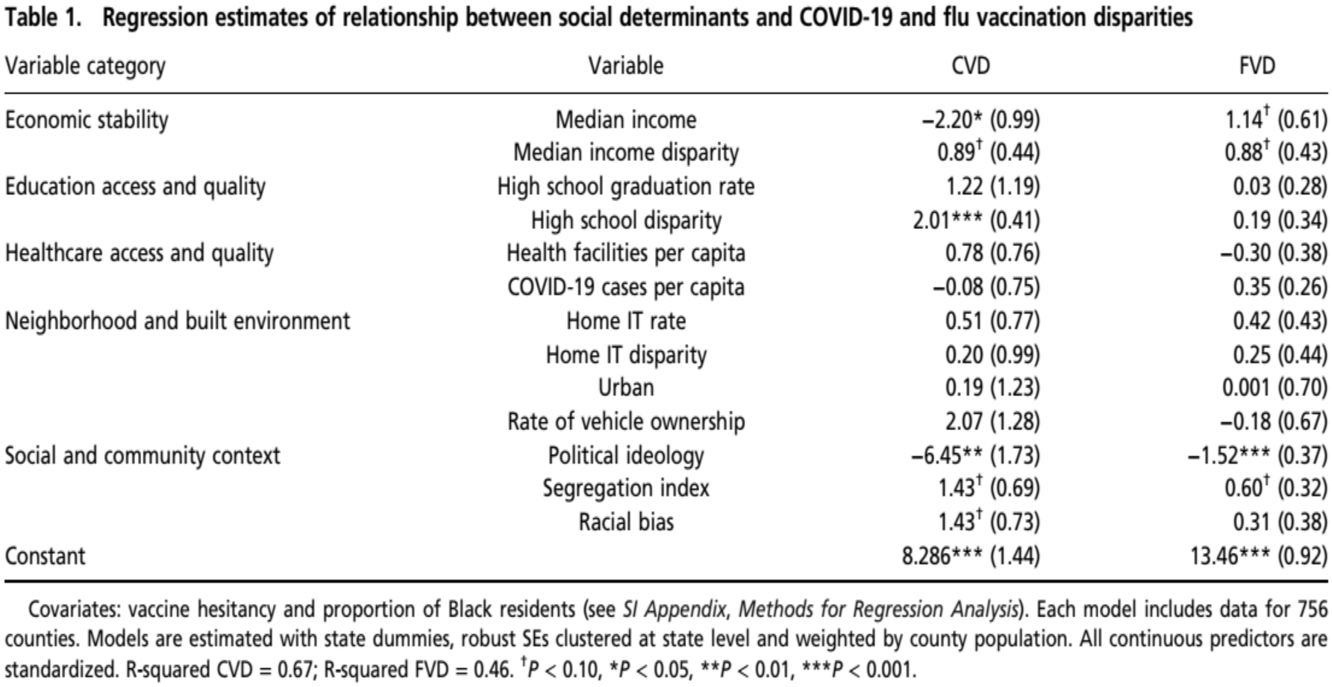


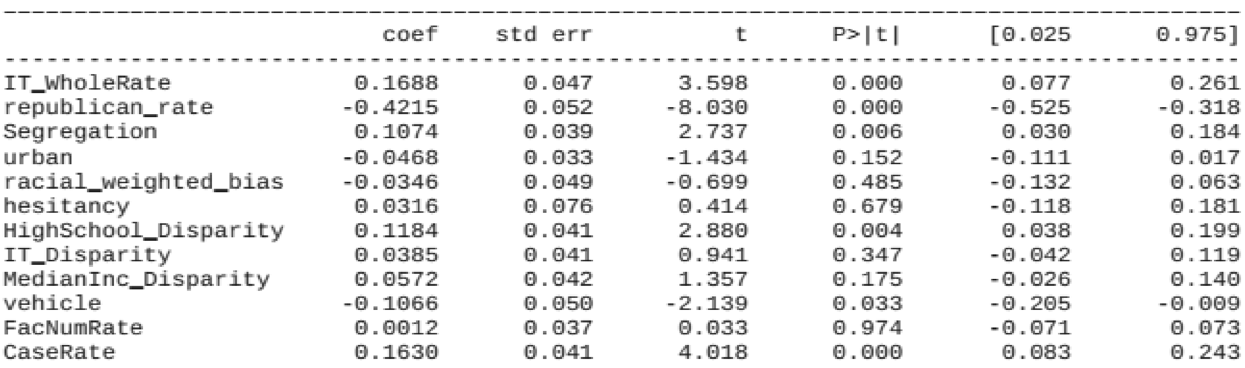
Table 1. Regression estimates of relationship between social determinants and COVID-19 and flu vaccination disparities

Variable category	Variable	CVD	FVD
Economic stability	Median income	−2.20* (0.99)	$1.14^{\dagger}$ (0.61)
	Median income disparity	$0.89^{\dagger}$ (0.44)	$0.88^{\dagger}$ (0.43)
Education access and quality	High school graduation rate	1.22 (1.19)	0.03 (0.28)
	High school disparity	2.01*** (0.41)	0.19 (0.34)
Healthcare access and quality	Health facilities per capita	0.78 (0.76)	−0.30 (0.38)
	COVID-19 cases per capita	−0.08 (0.75)	0.35 (0.26)
Neighborhood and built environment	Home IT rate	0.51 (0.77)	0.42 (0.43)
	Home IT disparity	0.20 (0.99)	0.25 (0.44)
	Urban	0.19 (1.23)	0.001 (0.70)
	Rate of vehicle ownership	2.07 (1.28)	−0.18 (0.67)
Social and community context	Political ideology	−6.45** (1.73)	−1.52*** (0.37)
	Segregation index	$1.43^{\dagger}$ (0.69)	$0.60^{\dagger}$ (0.32)
	Racial bias	$1.43^{\dagger}$ (0.73)	0.31 (0.38)
Constant		8.286*** (1.44)	13.46*** (0.92)

Covariates: vaccine hesitancy and proportion of Black residents (see *SI Appendix, Methods for Regression Analysis*). Each model includes data for 756 counties. Models are estimated with state dummies, robust SEs clustered at state level and weighted by county population. All continuous predictors are standardized. R-squared CVD = 0.67; R-squared FVD = 0.46. $^{\dagger}P < 0.10$, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

Note: In the original paper, the authors have grouped each variable in categories mentioned in Table above so, to compare our values with their work, we have considered the following variables to be equivalent:

IT_WholeRate = Home_ITrate,
Repulican_rate = Political ideology
HighScool_Disparity = High school disparity
IT_Disparity = Home IT disparity
MedianInc_Disparity = Median Income disparity
Racial_weighted_bias = Racial bias
Vehicle = Rate of vehicle ownership
FactNumRate = Health facility per capita

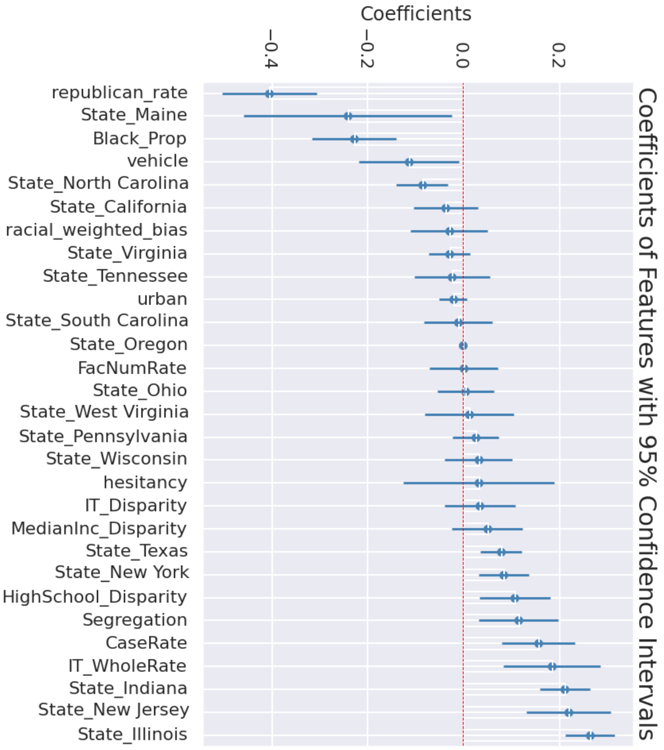


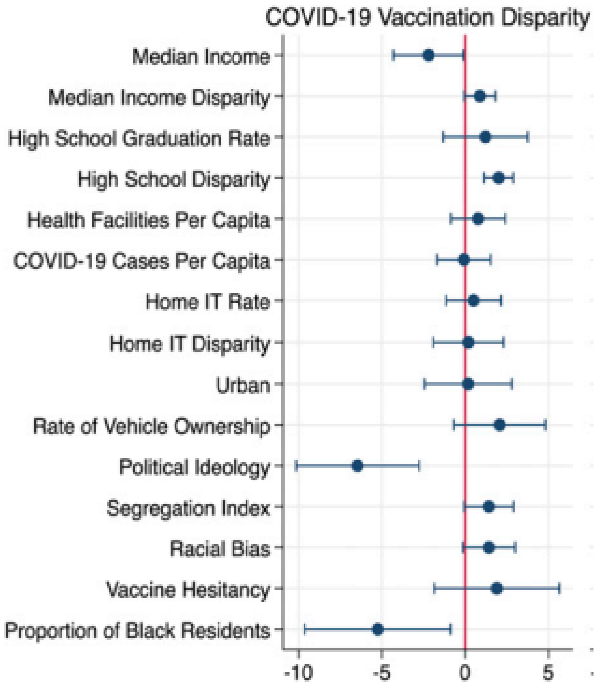
```
=======================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
IT_WholeRate              0.1688      0.047      3.598      0.000       0.077       0.261
republican_rate          -0.4215      0.052     -8.030      0.000      -0.525      -0.318
Segregation               0.1074      0.039      2.737      0.006       0.030       0.184
urban                    -0.0468      0.033     -1.434      0.152      -0.111       0.017
racial_weighted_bias     -0.0346      0.049     -0.699      0.485      -0.132       0.063
hesitancy                 0.0316      0.076      0.414      0.679      -0.118       0.181
HighSchool_Disparity      0.1184      0.041      2.880      0.004       0.038       0.199
IT_Disparity              0.0385      0.041      0.941      0.347      -0.042       0.119
MedianInc_Disparity       0.0572      0.042      1.357      0.175      -0.026       0.140
vehicle                  -0.1066      0.050     -2.139      0.033      -0.205      -0.009
FacNumRate                0.0012      0.037      0.033      0.974      -0.071       0.073
CaseRate                  0.1630      0.041      4.018      0.000       0.083       0.243
```

We can find that the coefficient in most of the variables follows the same nature in terms of negative or positive effects. However, they differ in the values pertaining to the reasons that our project dataset has fewer data points as compared to the original dataset used by the authors. Also, we have removed some variables like “**HighSchool_WholeRate”** and “**MedianInc_WholeAvg”** due to the high correlation factor. We have also used **yeo-Johnson** transformation to achieve Normal distribution of each variable which is not used in the actual result.

In the original work, the authors have achieved R^2 of 0.67 with OLS and we achieved R^2 0.612 with OLS, R^2 0.658 with Ridge, R^2 0.652 with Lasso but we achieved the best R^2 0.731 with SVR.

Below is the comparison of the 95% confidence interval between ours and the authors.

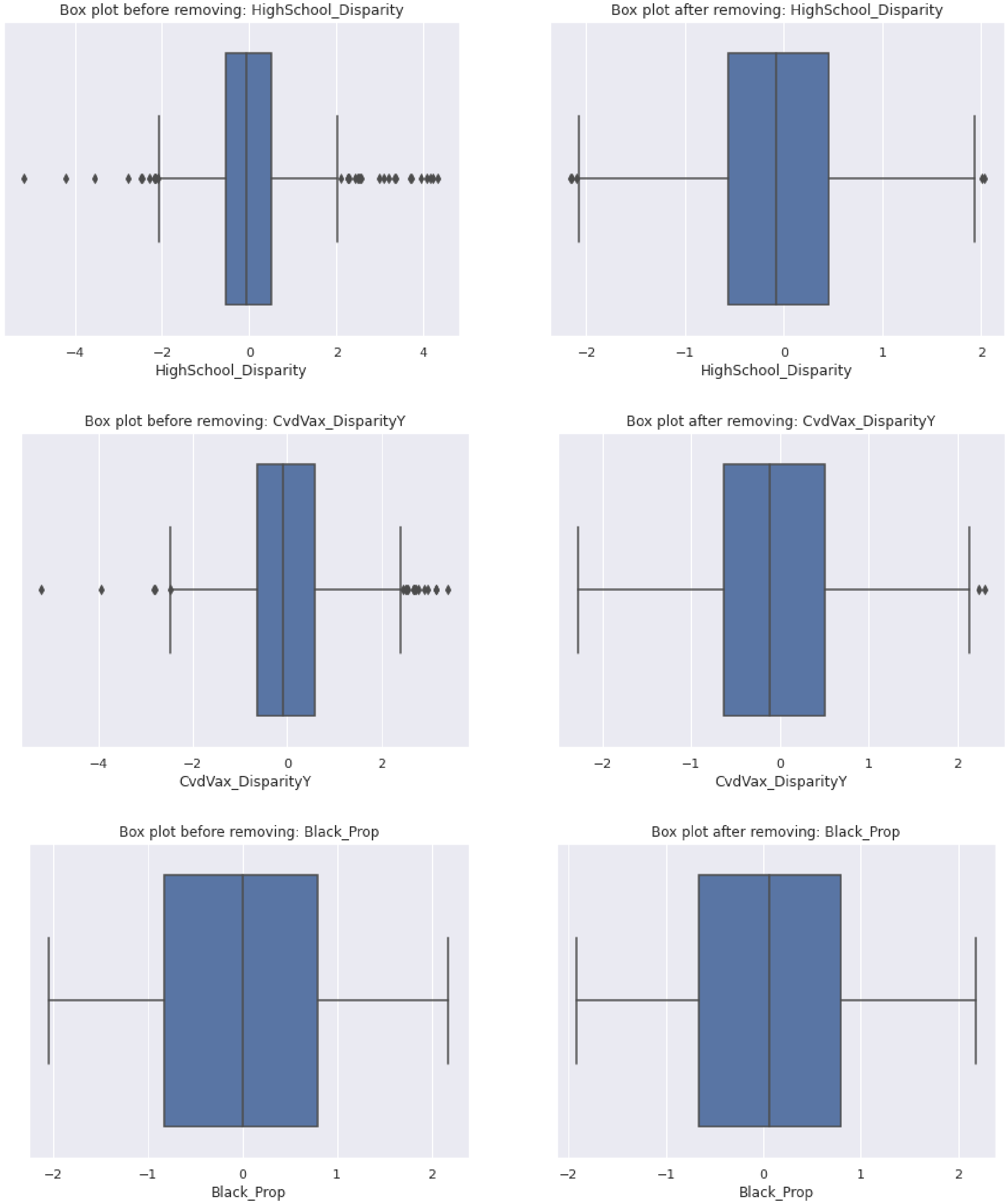


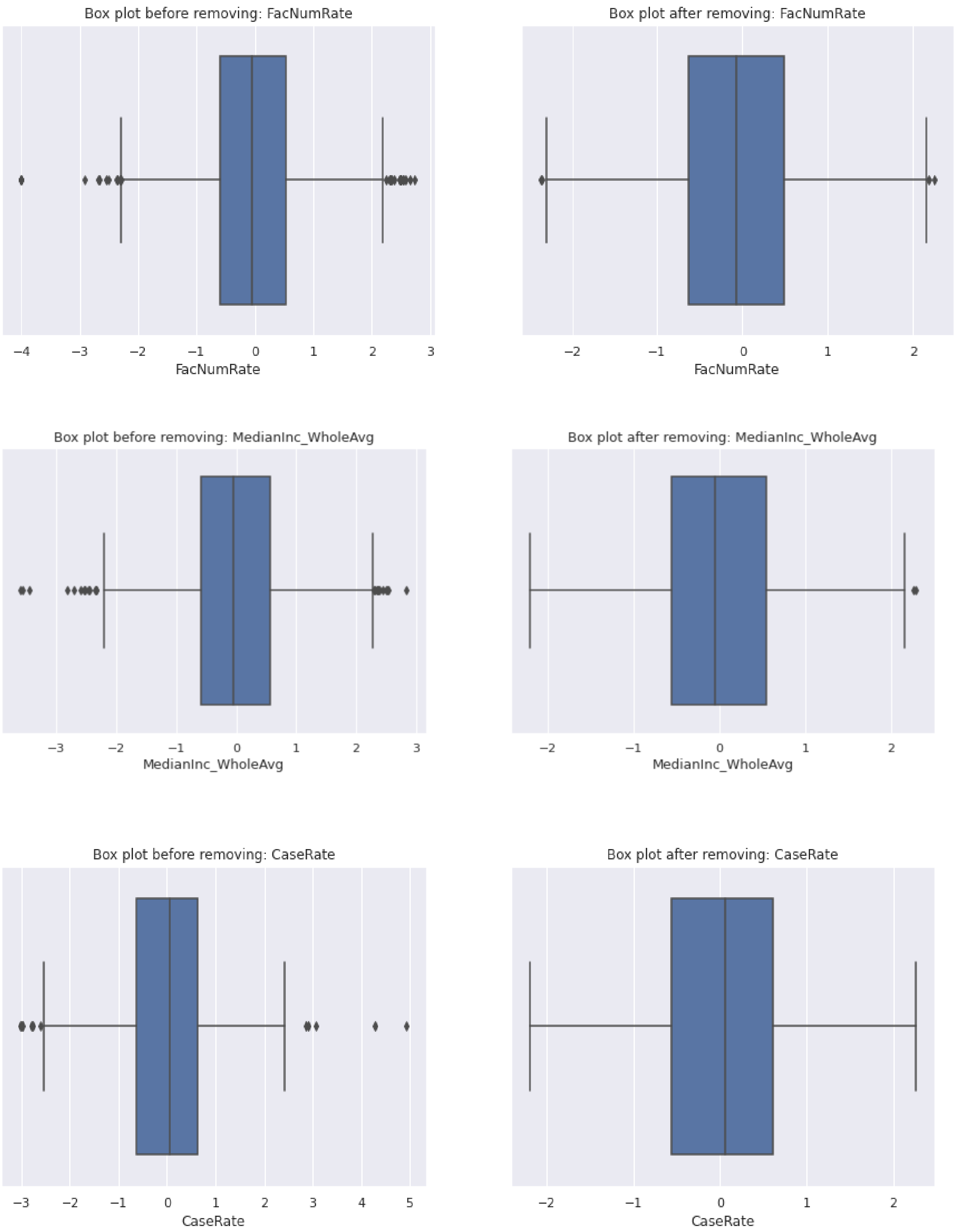


APPENDIX:

Section 3- Data Preparation

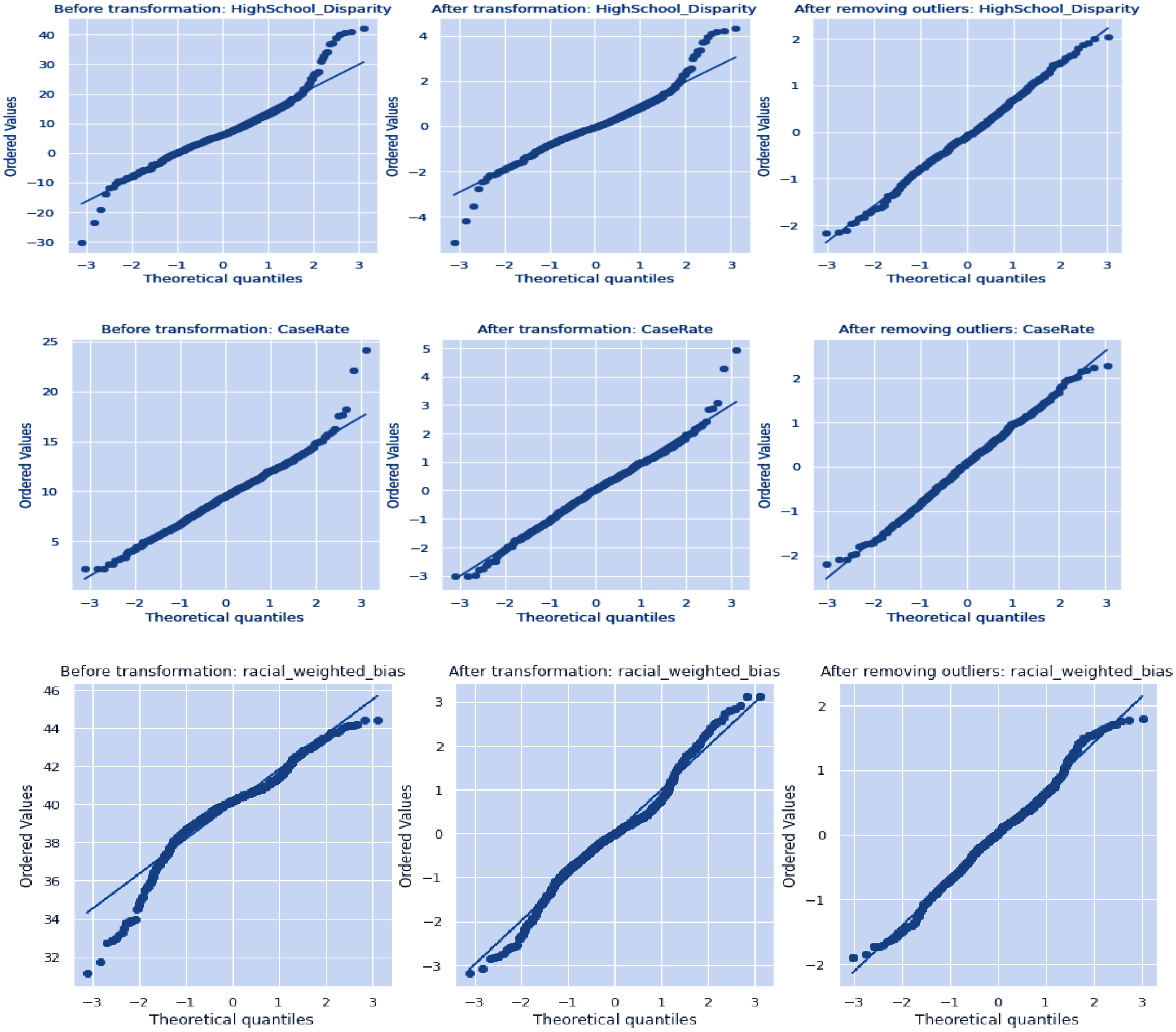
Box plot of Remaining variables before outlier removal and after outlier removal

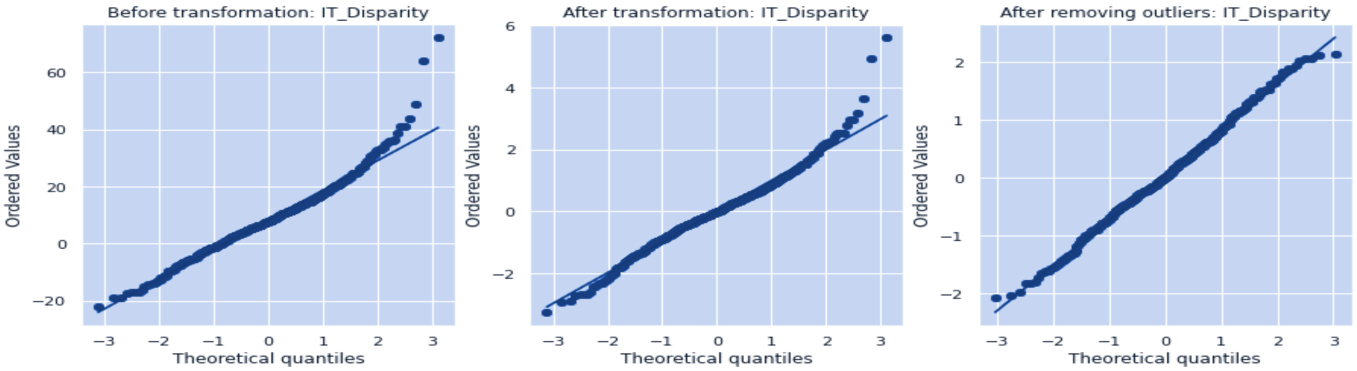



Box plot before removing: FacNumRate
−4
−3
−2
−1
0
1
2
3
FacNumRate
Box plot after removing: FacNumRate
−2
−1
0
1
2
FacNumRate
Box plot before removing: MedianInc_WholeAvg
−3
−2
−1
0
1
2
3
MedianInc_WholeAvg
Box plot after removing: MedianInc_WholeAvg
−2
−1
0
1
2
MedianInc_WholeAvg
Box plot before removing: CaseRate
−3
−2
−1
0
1
2
3
4
5
CaseRate
Box plot after removing: CaseRate
−2
−1
0
1
2
CaseRate

<u>Section 3:</u> Data Preparation

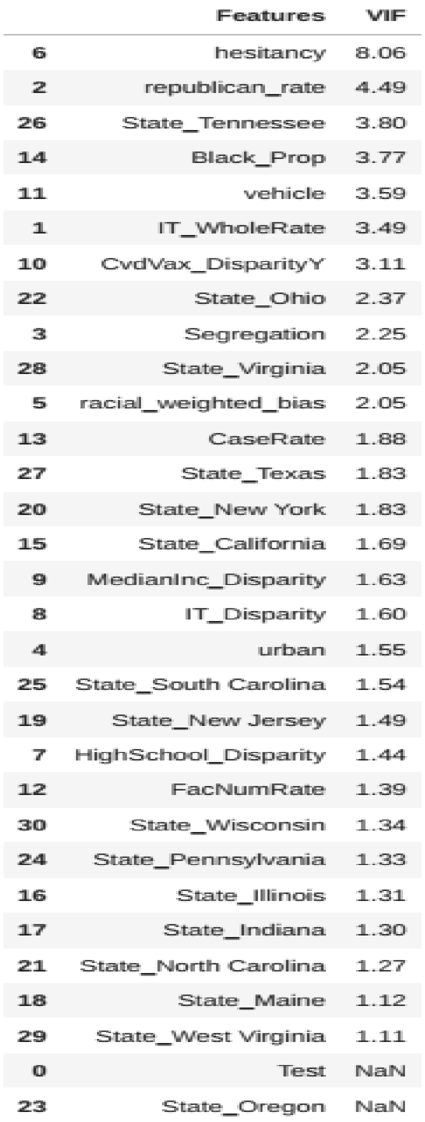
Transform each variable with yeo-Johnson transformation below is Q-Q plot.





SECTION 2 Data Preparation:

We calculate the VIF factor and find all factor is less than 10 hence multicollinearity is not of significant value



	Features	VIF
6	hesitancy	8.06
2	republican_rate	4.49
26	State_Tennessee	3.80
14	Black_Prop	3.77
11	vehicle	3.59
1	IT_WholeRate	3.49
10	CvdVax_DisparityY	3.11
22	State_Ohio	2.37
3	Segregation	2.25
28	State_Virginia	2.05
5	racial_weighted_bias	2.05
13	CaseRate	1.88
27	State_Texas	1.83
20	State_New York	1.83
15	State_California	1.69
9	MedianInc_Disparity	1.63
8	IT_Disparity	1.60
4	urban	1.55
25	State_South Carolina	1.54
19	State_New Jersey	1.49
7	HighSchool_Disparity	1.44
12	FacNumRate	1.39
30	State_Wisconsin	1.34
24	State_Pennsylvania	1.33
16	State_Illinois	1.31
17	State_Indiana	1.30
21	State_North Carolina	1.27
18	State_Maine	1.12
29	State_West Virginia	1.11
0	Test	NaN
23	State_Oregon	NaN