Factors Affecting COVID Vaccine Hesitancy in US Counties

Group 3: Abhishek Suragani, Farhaan S. Haque, Huma Ali Meer, Varsha Thebo

ITEC-620: Business Insights/Analytics, American University

Dr. Jay Simon

December 06, 2021

Executive Summary

This project is designed to study and understand the causes behind the incidence of COVID-19 vaccine hesitancy amongst the US population on a county level. It specifically investigates the factors that influence people's decision to avoid the vaccine. The data used for this project is retrieved from several sources such as the United States Census Bureau, Centers for Disease Control and Prevention (CDC), United States Department of Agriculture (USDA), Harvard Dataverse, Household Pulse Survey (HPS), Surgo Ventures, and data repository websites such as livingatlas.com. The project looked at the significance of variables such as race, geographic location of counties, Social Vulnerability Index (SVI), Concern for COVID Vaccine Rollout (CVAC), political affiliation, socioeconomic status, and obesity in explaining the status and variance in COVID-19 Vaccine Hesitancy (VH).

The analysis used for this project consists of both descriptive and predictive methods. To identify the relationship between VH and the selected variables, methods such as correlation analysis and K-means clustering were employed. This process enabled the identification of clusters of counties that are the most similar in terms of VH and other variables of interest. Furthermore, a multiple regression model and a regression tree was created to acquire more conclusive results to assess the direction and the strength of relationship between the dependent and independent variables.

The regression equation used in this report had the lowest Root Mean Squared Error (RMSE) and the highest Multiple R-Squared value, amongst the various regression models that were tested. All the variables used in the final regression equation were statistically significant. The variables with a strong positive relationship with VH are Concern for Vaccine Rollout, Race, and Obesity. On the other hand, Unemployment Rate has a negative impact on Vaccine Hesitancy. Moreover, counties with Democratic majority tend to have lower Vaccine Hesitancy. Counties in the West tend to have lower vaccine hesitancy than counties in Midwest, Northeast and South.

The regression tree used in this report also had the lowest RMSE amongst the regression tree models that were tested. It highlighted those counties with CVAC higher than 60% have higher levels of VH.

Introduction

COVID-19 has an unprecedented impact on the economic and social structures of the world. It has upended the lives of millions across the world and has laid bare the inequalities on every level. While the fight against the pandemic continues, many countries around the world were able to either produce or procure vaccines against this deadly virus. The United States Food and Drug Administration (FDA) issued the authorization for the use of the Pfizer, Moderna and the Janssen vaccines in December 2020 and shortly thereafter, these vaccines were made available and at a minimal to no cost to the US citizens (FDA, 2020). However, despite the dissemination of the scientific data on vaccine safety and efficacy, there are many who are wary of the Covid vaccine and are hesitant to use it.

While resistance to vaccination is not a new phenomenon, it is especially intriguing in these unprecedented times when the world is mired in a global pandemic and access to information is rapid and extensive. It is curious that despite numerous educational campaigns, interactive platforms for the frequently asked questions (FAQs) on government websites and research from well-reputed global health agencies, many in the United States remain hesitant to get the COVID vaccine. This project will explore the multiple variables at play in studying the vaccine hesitancy in the US population on a county level.

The data sources used to craft this report are from the United States Census Bureau, CDC, United States Department of Agriculture, Harvard Dataverse, livingatlas.com and others. While the initial dataset consisted of 3143 US counties, data cleaning resulted in the working data of 3112 counties. The variable that allowed the merging of various datasets is the Federal Information Processing Standards (FIPS). These are a set of five-digit codes that function as unique identifiers for counties or their equivalents in the United States. All data is from 2021 with the exception of Unemployment Rate (2020), Median Household Income (2019) and Diabetes and Obesity (2018).

Description of the Variables

While the final dataset contained 27 variables, the most significant ones are described below:

1. **Vaccine Hesitancy (VH):**

This is an index created by the CDC using the Household Pulse Survey (HPS). HPS is a collaborative effort of federal agencies to collect timely data about the impact of coronavirus pandemic on American households. In trying to assess the unique socioeconomic impact of the pandemic on every household, the set of standardized questions are then aggregated to form an index. Out of the three distinct categories from the survey question for getting the vaccine, the category selected for this variable is "Strongly Hesitant" that includes survey responses indicating that the participants would "definitely not" receive a COVID-19 vaccine when available. This index can take values ranging from 0 to 1 and estimates people's COVID vaccine hesitancy as a continuous variable in terms of percentages (CDC, 2021) (Household Pulse Survey, 2021).

2. **Social Vulnerability Index (SVI):**

This index is also created by the CDC. It summarizes the extent to which a community is socially vulnerable to disasters and is an important metric in assessing the support required in the wake of a public health emergency. SVI is an aggregate of 14 social factors grouped into four themes that include the Socioeconomic Status, Minority Status, Housing Type, etc. SVI takes values between 0 (lowest vulnerability) to 1 (highest vulnerability) (CDC, 2021).

3. **COVID-19 Vaccine Coverage (CVAC):**

This is an index created by Surgo Ventures. It captures supply- and demand- related challenges that may prevent large scale COVID-19 vaccine coverage in U.S. counties, through five specific themes: historic under-vaccination, sociodemographic barriers, resource-constrained healthcare system, healthcare accessibility barriers, and irregular care-seeking behaviors. The CVAC measures the level of concern for a difficult rollout on a range from 0 (lowest concern) to 1 (highest concern) (Surgo Ventures,2021).

4. **Race:**

CDC also provides county level data for the percentage of people belonging to a particular race. The race variables include Hispanic, non-Hispanic White, non-Hispanic Asian, non-Hispanic Black, non-

Hispanic American Indian/Alaska Native, and non-Hispanic Native Hawaiian/Pacific Islander (CDC, 2021).

5. Political Affiliation:

This data is collected from the Harvard Dataverse. It identifies the majority party in a county, i.e. Democrat, Republican or Other. For the sake of the analysis, the values for this variable have been converted into dummy variables (Harvard Dataverse,2020).

6. Region

This variable identifies the region that a particular county is situated in. The regions include South, West, Midwest, and Northeast (US Census,2021).

7. Unemployment Rate 2020 and Median Household Income 2019

These variables are collected from a dataset by the U.S. Department of Agriculture. It represents the county level unemployment rate for 2020 and Median Household Income for 2019 (US Department of Agriculture,2019,2020).

8. Diabetes and Obesity

These two variables were collected from 2018 data available at data repository site livingatlas.com. It represents county level estimates of the percentage of people suffering from diabetes and obesity (Berry, 2018).

Analysis

There were four analytical methods used in this project: Correlation Analysis, Multiple Regression Model, K-Means Clustering and Regression Trees.

Correlation Analysis

This was computed to assess the strength and relationship between multiple variables against VH. The values for a correlation coefficient lie between -1 and 1 wherein a value of 0 indicates no relationship between two variables, a value of 1 indicates a perfectly positive relationship and a value of -1 indicates a negative relationship.

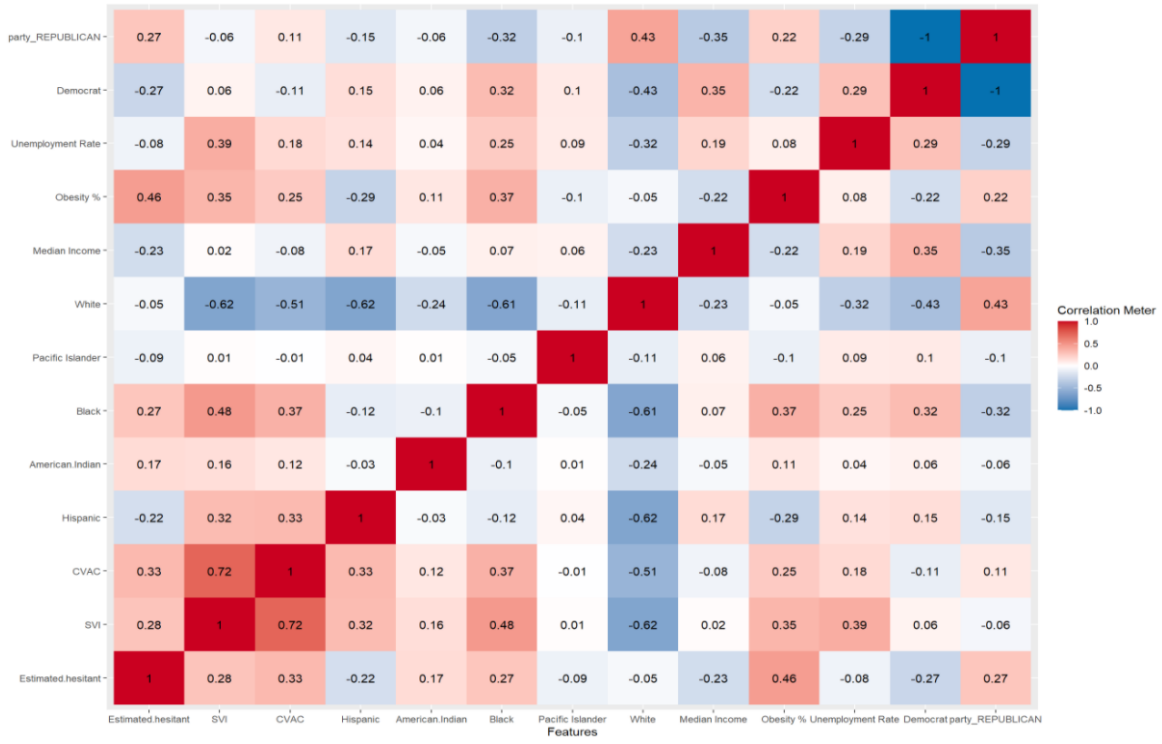


Figure 1

Insights from the correlation plot above are summarized below:

1. SVI was moderately and positively correlated with VH, with a value of 0.28. In other words, the higher the value of SVI, the greater the VH.
2. There was a high positive relationship between obesity and VH with a correlation coefficient of 0.46, thereby revealing that the higher the percentage of obesity, the higher the VH.
3. The relationship between household median income and unemployment rate with VH is negative, albeit a weak correlation with the values of -0.23 and -0.08, respectively.
4. Among races, there was found to be a positive relationship between Native Hawaiian/Pacific Islander population and VH and a negative relationship between Asians and VH.

Additional correlation plots that explore relationship between the independent variables and VH have been included in appendix A.

K-Means Clustering

This method allowed for the identification of existing patterns in the dataset. The characteristics of the vaccine-hesitant population were divided into five clusters, as presented in Figure 2. Clusters with the highest and lowest VH are highlighted in the green and red respectively. The first cluster has the highest rate of VH, second highest value of CVAC, highest SVI, highest diabetic population, and highest Native American Indian population. The same cluster also has the lowest Median Income. On the other hand, the cluster with the lowest VH also has the second-lowest value for SVI and the lowest for value CVAC. The population value for American Indians is negligible and it also has the highest value for median household income. These results are consistent with the relationships obtained from the correlation analysis.

	Vaccine Hesitancy	SVI	CVAC	American Indian	African Americans	White	Diabetes	Democrat	Median Income
1	0.26	0.88	0.77	0.57	0.01	0.31	13.28	0.42	11187.71
2	0.19	0.48	0.49	0.02	0.01	0.75	8.86	0.23	28205.75
3	0.14	0.39	0.25	0.00	0.11	0.68	9.24	0.99	248843.78
4	0.19	0.37	0.39	0.01	0.03	0.89	11.30	0.00	22241.78
5	0.22	0.81	0.82	0.01	0.25	0.58	13.11	0.17	22745.41

Figure 2

A plot visualizing the five clusters (Fig. 13) and the raw R output of the clusters (Fig. 14) is included in Appendix B.

Regression Analysis

The Multiple Linear Regression model shown in figure 3 has a RMSE of 0.0344 and a Multiple R-Squared of 41.2%. All variables included in the model are statistically significant. The model was trained on 60% of the data and tested on 40% of the data.

Figure 3

Independent Variables	Coefficient
CVAC	1.9%
Hispanic	13.1%
African Americans	22.5%
Native American	27.2%
Native Islander	52.1%
White	18.9%
Obesity	0.3%
Midwest	-2.1%
Northeast	-5.0%
South	-0.9%
UR	-0.1%
Democrat	-1.3%

The following important results can be drawn from the regression model:

1. CVAC has a positive impact on VH. A one percent increase in the Level of Concern for Vaccine Rollout in a county, leads to a 1.9% increase in VH.
2. All races have a positive impact on VH. Some races have a higher impact than others. The non-Hispanic Native Hawaiian/Pacific Islander variable has the highest beta coefficient value. It can be interpreted as: a one percent increase in the non-Hispanic Native Hawaiian/Pacific Islander population leads to a 52% increase in VH in a particular county. There are, however, two caveats to this value. Firstly, this variable has the highest standard error value in the model, which means it has outliers. Secondly, it was statistically significant at 95% significance level (or an alpha value of 5%). The rest of the variables in the model were statistically significant at much higher levels, i.e 99.99%. This means that the non-Hispanic Native Hawaiian/Pacific Islander was not as statistically significant as the rest of the variables.
3. A one percent increase in the African American population in a county lead to a 22.5% increase in VH as compared to a 18.5% increase in VH in case of non-Hispanic white population.
4. Obesity, while being very highly and positively correlated with VH with a value of 0.46, has a very low beta coefficient value in the regression equation. A one percent increase in the obese population in a county lead to a 0.3% increase in VH.
5. ‘West’ was used as a base case variable. Therefore, the rest of the region variables can be interpreted as VH is 2.1%, 5% and 0.9% in Midwest, Northeast and South, respectively, as compared to the West.
6. Counties with Democratic majority had 1.3% less VH than counties with Republican majority.

The Figure 4 visualizes these results more clearly:

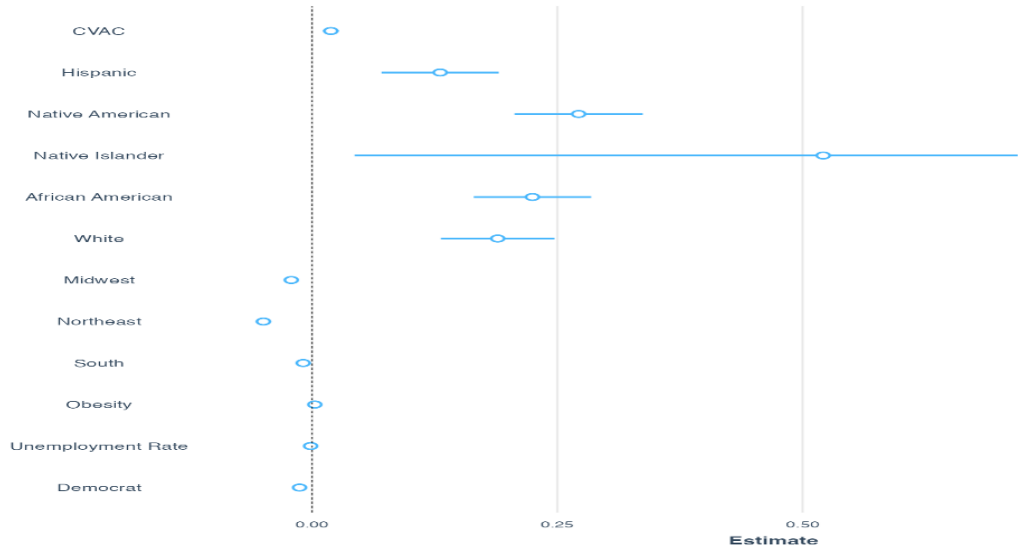


Figure 4

The raw R output (Fig. 15) for the regression model is included in Appendix B.

7. Regression Trees

Figure 5 visualizes the regression tree used for analysis. It has a RMSE value of 0.0355. Both the regression equation and regress tree perform well in minimizing the error. The following insights can be drawn from the regression tree:

1. CVAC divides the regression tree in two portions; One with CVAC greater than or equal to 60% and one with CVAC less than 60%
2. The nodes with darker shades of blue indicate higher VH, most of which lie under the branch of CVAC>0.6
3. The node, where VH is 0.28, the highest in the tree, also has unemployment rate higher than 4.6%, CVAC greater than 19%.

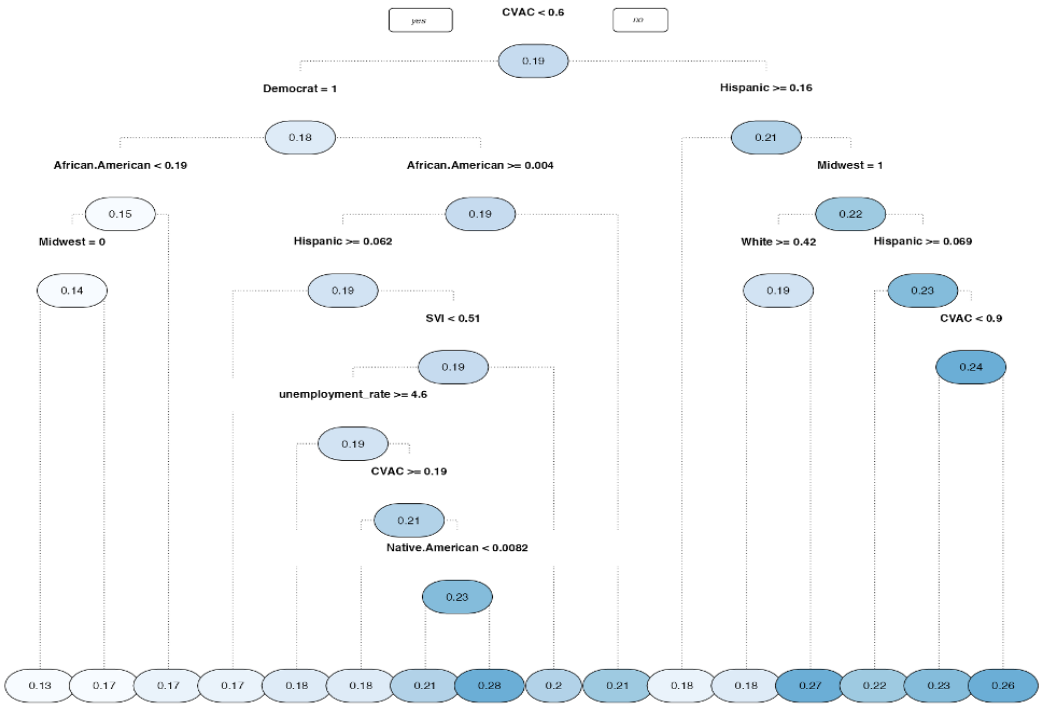


Figure 5

Discussion and Conclusion

The analysis revealed glaring disparities between groups of people who showed hesitancy towards the COVID-19 vaccine compared to people who were less vaccine hesitant. It showed that socioeconomic inequalities, such as unemployment, historical marginalization of communities can compound differences and further alienate those communities in the time of a global crisis. Results from the analysis also revealed that while the nature of disadvantage varies across counties and regions, the minority groups in general, face additional challenges in accessing the vaccine.

While this analysis was made possible by existing data, this project underscores the importance of more updated data on minority communities to identify gaps and offer more effective interventions.

The scope of this analysis could be further strengthened with the availability of data on many facets of the communication campaigns done both for and against the vaccine. Any information on the exposure to mass media could provide unique insights into designing more targeted interventions for the marginalized groups.

References

Agriculture, D. (n.d.). *Unemployment*. USDA ERS - Data Products. Retrieved December 8, 2021, from https://data.ers.usda.gov/reports.aspx?ID=17828.

Agriculture, D. (n.d.). *Vaccine hesitancy for covid-19: State, county, and local estimates*. ASPE. Retrieved December 8, 2021, from https://aspe.hhs.gov/reports/vaccine-hesitancy-covid-19-state-county-local-estimates.

Berry, L. (n.d.). *Diabetes, obesity, and inactivity by US County*. Living Atlas of the World. Retrieved December 8, 2021, from https://livingatlas-dcdev.opendata.arcgis.com/datasets/arcgis-content::diabetes-obesity-and-inactivity-by-us-county/explore?location=45.150673%2C56.951297%2C3.31&showTable=true.

Bureau, U. S. C. (2021, December 1). *Household Pulse Survey Data tables*. Census.gov. Retrieved December 8, 2021, from https://www.census.gov/programs-surveys/household-pulse-survey/data.html.

Center for Disease Control and Prevention. (n.d.). *CDC SVI 2018 documentation - 1/31/2020 please see data ...* cdc.gov. Retrieved December 8, 2021, from https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/pdf/SVI2018Documentation-H.pdf.

Centers for Disease Control and Prevention. (n.d.). *Vaccine hesitancy for covid-19*. Centers for Disease Control and Prevention. Retrieved December 8, 2021, from https://data.cdc.gov/stories/s/Vaccine-Hesitancy-for-COVID-19/cnd2-a6zw.

Commissioner, O. of the. (n.d.). *Covid-19 frequently asked questions*. U.S. Food and Drug Administration. Retrieved December 10, 2021, from https://www.fda.gov/emergency-

preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-frequently-asked-questions.

Mckensey. (n.d.). *Path to the next normal collection - mckinsey & company*. Retrieved December 8, 2021, from https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Navigating%20the%20coronavirus%20crisis%20collected%20works/Path-to-the-next-normal-collection.pdf.

School, H. B. (n.d.). *US Presidential Elections 2020*. Countypres_2000-2020.TAB - U.S. presidential elections. Retrieved December 8, 2021, from https://dataverse.harvard.edu/file.xhtml?fileId=4819117&version=9.0.

Ventures, S. (n.d.). *Surgo U.S. COVID-19 vaccine coverage index*. Surgo Precision For Covid. Retrieved December 8, 2021, from https://vaccine.precisionforcovid.org/.

Appendix A: Correlation Plots

Figure 6

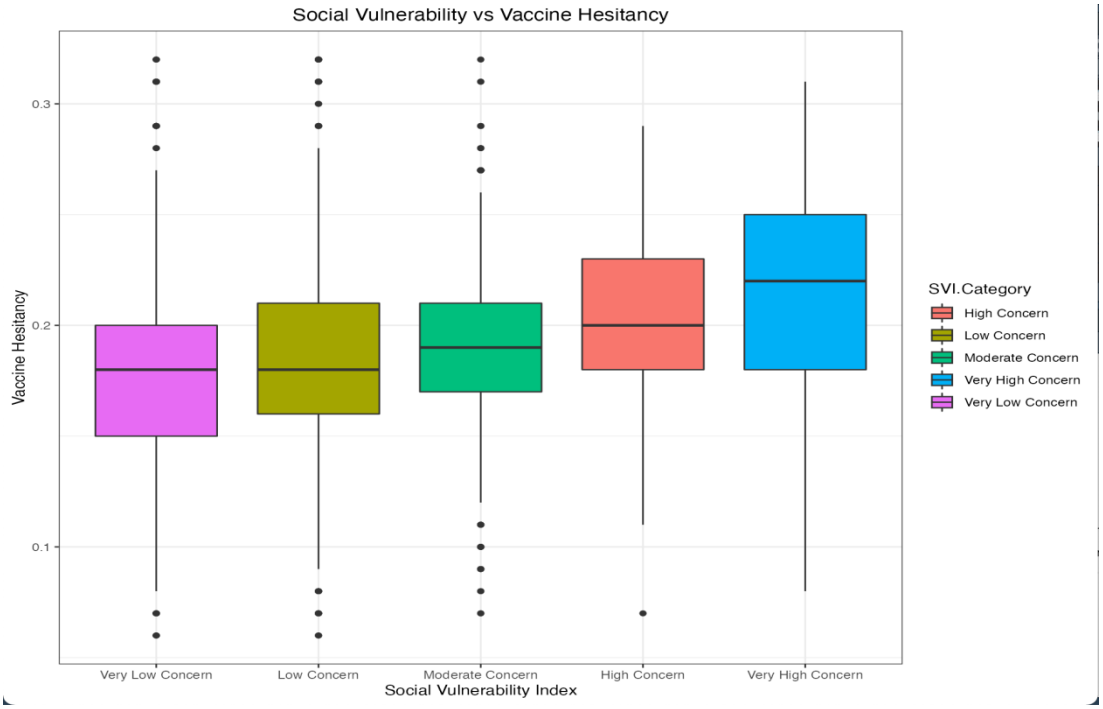


Figure 7

Figure 7 shows the relationship between vaccine hesitancy and political affiliation, with VH on top which is from 0.06 to 0.32. Y-axis represents the votes a particular party received and x axis represents the two political parties. Democrats are in red and Republican in blue. A general trend which can be seen from this facet grid is that VH is high in counties where republicans are in majority.

Figure 8

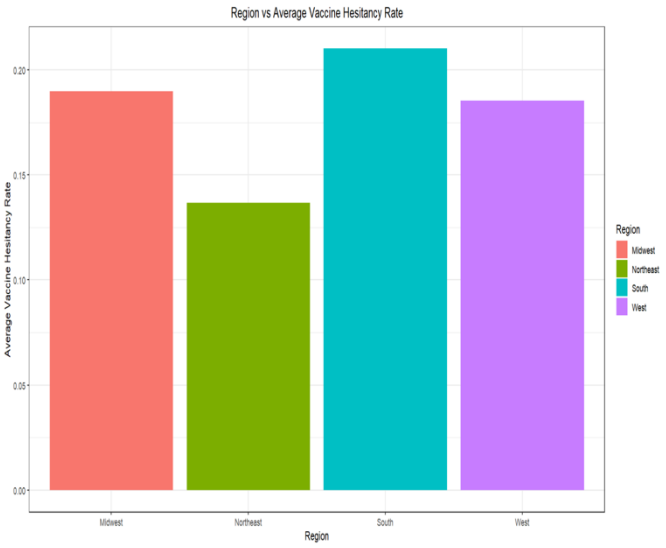


Figure 9

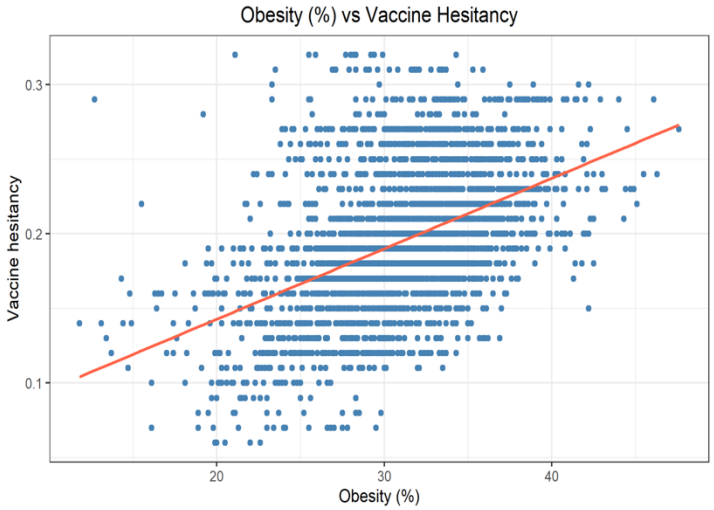


Figure 10

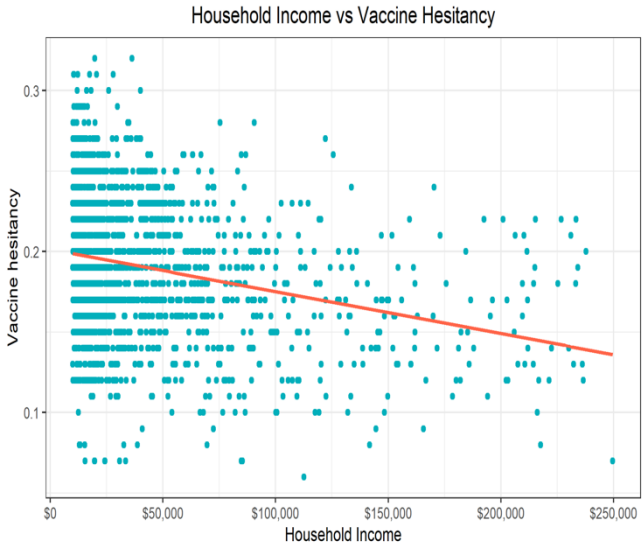
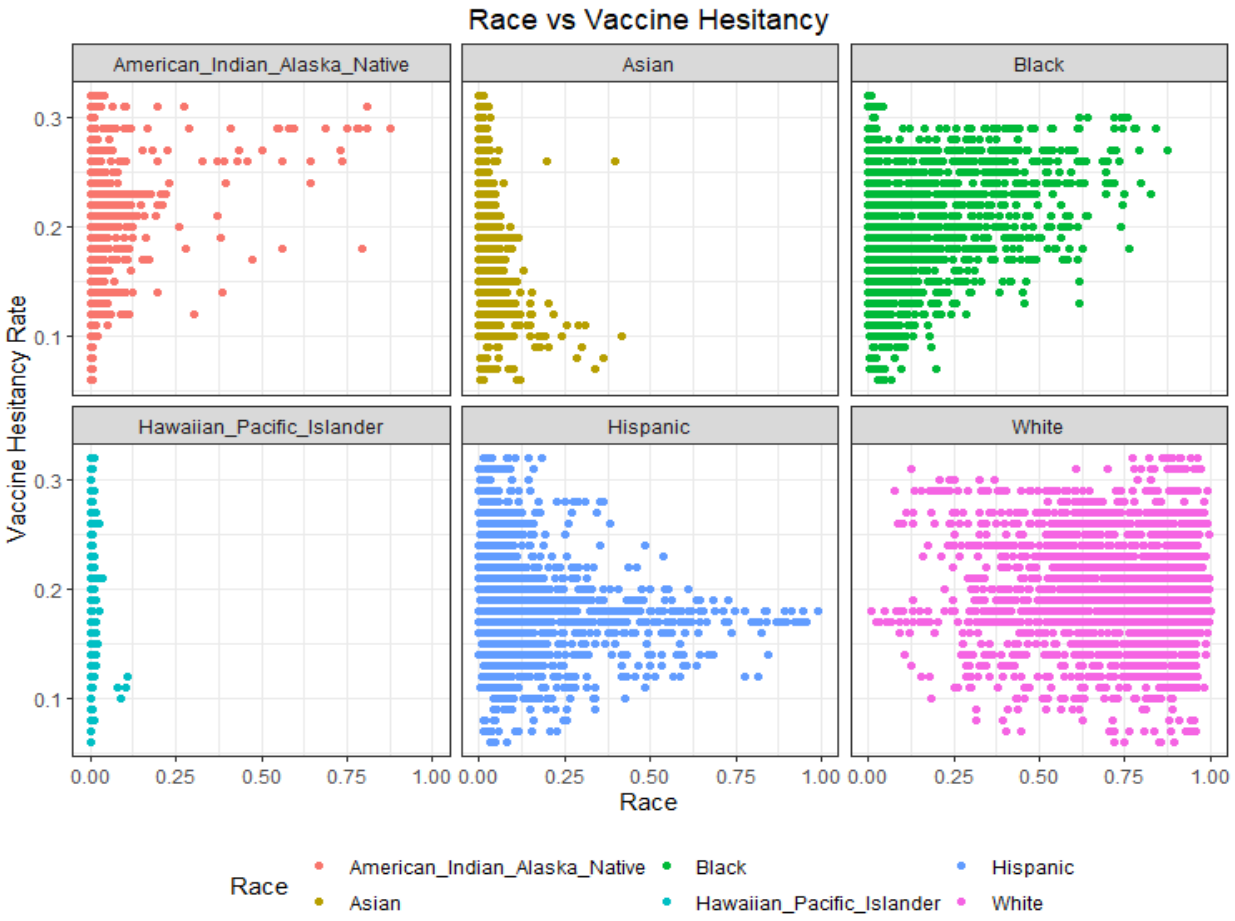


Figure 11

Figure 12



Appendix B

Figure 13: K-means Clustering

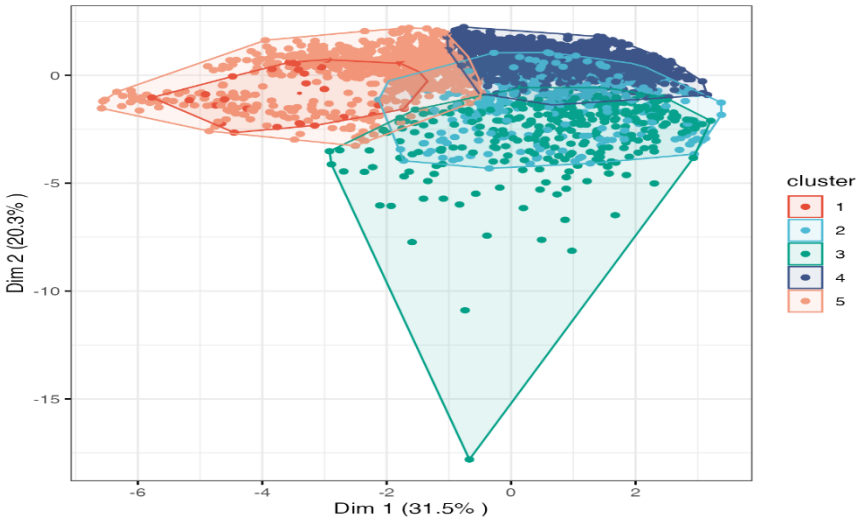


Figure 14: K-Means Clustering R Output

	Estimated.hesitant	SVI	CVAC.level.of.concern.for.vaccination.rollout	Percent.non.Hispanic.American.Indian.Alaska.Native
1	0.2561290	0.8793548	0.7719355	0.572677419
2	0.1888095	0.4756614	0.4877778	0.021874603
3	0.1432595	0.3908228	0.2541772	0.004271835
4	0.1912165	0.3681223	0.3877417	0.007566563
5	0.2235077	0.8076148	0.8152551	0.011061097

	Percent.non.Hispanic.Black	Percent.non.Hispanic.White	Diabetes_Percent	party_DEMOCRAT	Region_West	Median_Household_Income_2019
1	0.011819355	0.3112194	13.277419	4.193548e-01	3.870968e-01	11187.71
2	0.009437831	0.7486254	8.860053	2.301587e-01	1.000000e+00	28205.75
3	0.114835127	0.6796057	9.244620	9.905063e-01	9.493671e-02	248843.78
4	0.028506363	0.8873722	11.300686	-5.225695e-15	3.490522e-15	22241.78
5	0.247165051	0.5768325	13.112500	1.658163e-01	-2.086725e-16	22745.41

Figure 15: Regression Model R Output

MODEL FIT:
$F(12,1854) = 108.2478$, $p = 0.0000$
$R^2 = 0.4120$
$Adj.\ R^2 = 0.4082$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	-0.0569	0.0286	-1.9927	0.0464
CVAC.level.of.concern.for.vaccination.rollout	0.0191	0.0041	4.6995	0.0000
Percent.Hispanic	0.1306	0.0304	4.2896	0.0000
Percent.non.Hispanic.American.Indian.Alaska.Native	0.2717	0.0333	8.1585	0.0000
Percent.non.Hispanic.Black	0.2246	0.0305	7.3580	0.0000
Percent.non.Hispanic.Native.Hawaiian.Pacific.Islander	0.5210	0.2435	2.1395	0.0325
Percent.non.Hispanic.White	0.1892	0.0295	6.4044	0.0000
Region_Midwest	-0.0212	0.0031	-6.8559	0.0000
Region_Northeast	-0.0496	0.0041	-12.1331	0.0000
Region_South	-0.0090	0.0030	-2.9768	0.0030
Obesity_Percent	0.0028	0.0002	11.4300	0.0000
unemployment_rate	-0.0015	0.0004	-3.5618	0.0004
party_DEMOCRAT	-0.0127	0.0029	-4.3312	0.0000