Example Manuscript Template for a Data Analysis Project

Andreas Handel

2021-10-29

#Loading required packages  
library(dplyr) #for data processing

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(here) #to set paths

## here() starts at C:/Users/Priyanka/Desktop/new/Priyanka\_G-MADA-project

library(tidyverse) #all required data manipulation packages

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.3 v stringr 1.4.0  
## v tidyr 1.1.4 v forcats 0.5.1  
## v readr 2.0.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# 1 Summary/Abstract

In past pandemics, vulnerable populations faced greater disease burden and decreased testing and treatment access.1 As coronavirus disease 2019 (COVID-19) spreads in the USA, concern is growing that even the early stages of this pandemic have disproportionately impacted vulnerable communities.2–4 However, the relationship between social vulnerability and ethnicity remains unknown.

# 2 KEY QUESTIONS

What is already known about this subject?

COVID-19 has disproportionally affected racial/ethnic minority groups.

What will this study add or answer?

The present analysis will attempt to estimate the percent of the population in each US region that may be heistant to get a vaccine.

Relationship between ethnicity and social vulnerability index

How might this impact clinical practice?

Identify and address barriers to COVID-19 vaccination. Continued monitoring of vaccination coverage by social vulnerability metrics is critical for developing tailored, local vaccine administration and outreach efforts to reduce COVID-19 vaccination inequities.

## 2.1 General Background Information

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Coronavirus disease 2019 (COVID-19) was declared a pandemic in March 2020. COVID-19 vaccine is the most sustainable option to manage the current pandemic. However, vaccine hesitancy by even a small subset of the population can undermine the success of this strategy. The Social Vulnerability Index (SVI) is a tool that uses census data to identify and map places where a community may have more difficulty preventing human suffering and financial loss in a disaster.The SVI assesses the extent that 15 known vulnerabilities (indicators) are present within a community and categorizes them into four themes: socioeconomic status, household composition and disability, minority status and language minority and housing type and transportation. Indicators like poverty and transportation can highlight places where people may have difficulty accessing COVID-19 testing, treatment and vaccination

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

The dataset used for this analysis is publicly available on CDC website. <https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw>

The SVI index at the county level using the Public Use Microdata Sample (PUMS). The data was collected by utilizing survey question. The SVI index is categorised into

1. Very Low (0.0-0.19);
2. Low (0.20-0.39);
3. Moderate (0.40-0.59);
4. High (0.60-0.79);
5. Very High (0.80-1.0).

CVAC Level Of Concern 1) Very Low (0.0-0.19); 2) Low (0.20-0.39); 3) Moderate (0.40-0.59); 4) High (0.60-0.79); 5) Very High (0.80-1.0).

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

The present analysis will attempt to estimate the percent of the population in each county that may be vaccine hesitant

Relationship between ethnicity and vaccine hesitancy

# 3 Future analysis

I plan to present a summary of social vulnerability index based on race and ethnicity. I will plot some graphs (box-plot, scatter plot). I am also planning to do some regression analysis too

# 4 Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

## 4.1 Data aquisition

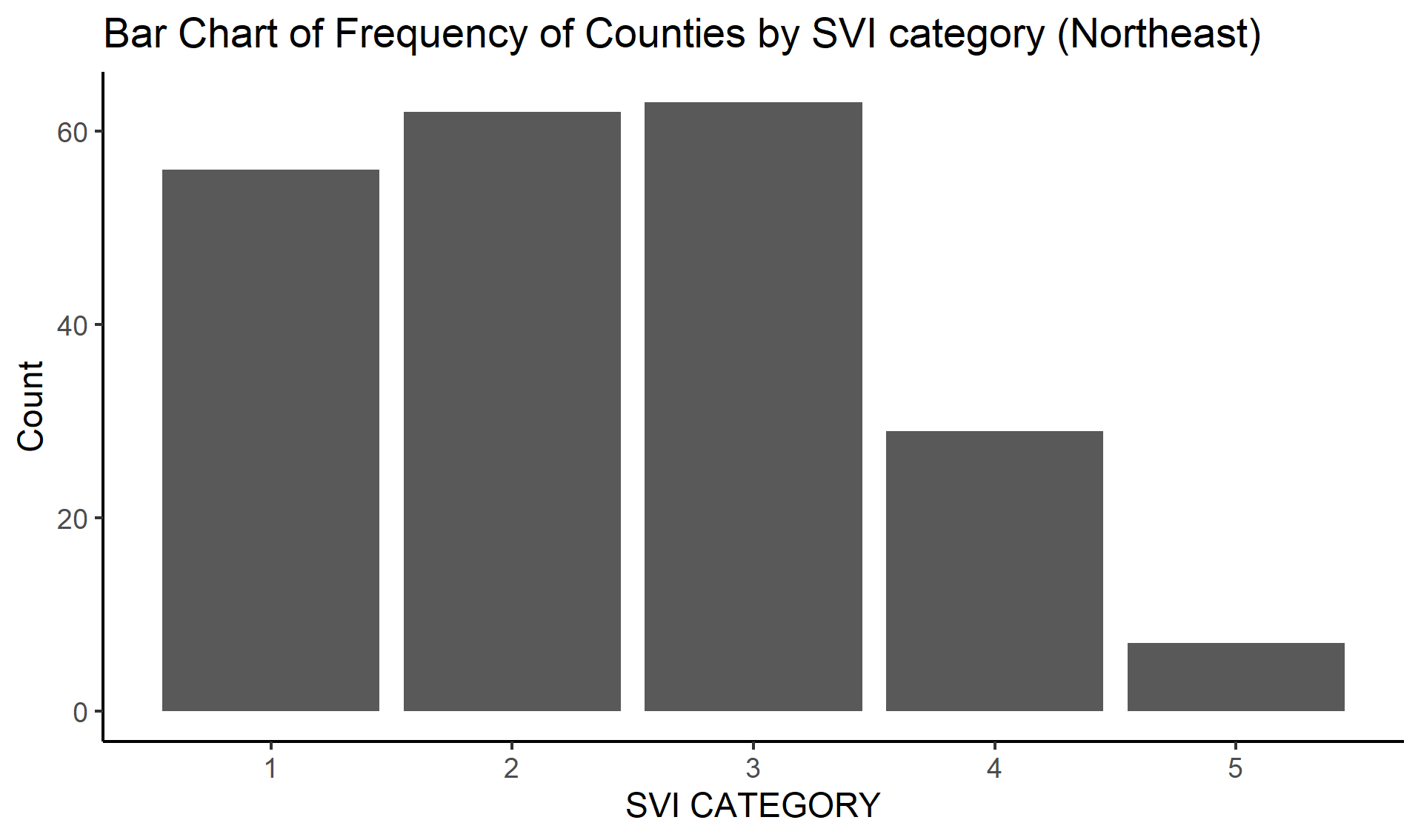
*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 4.2 Data import and cleaning

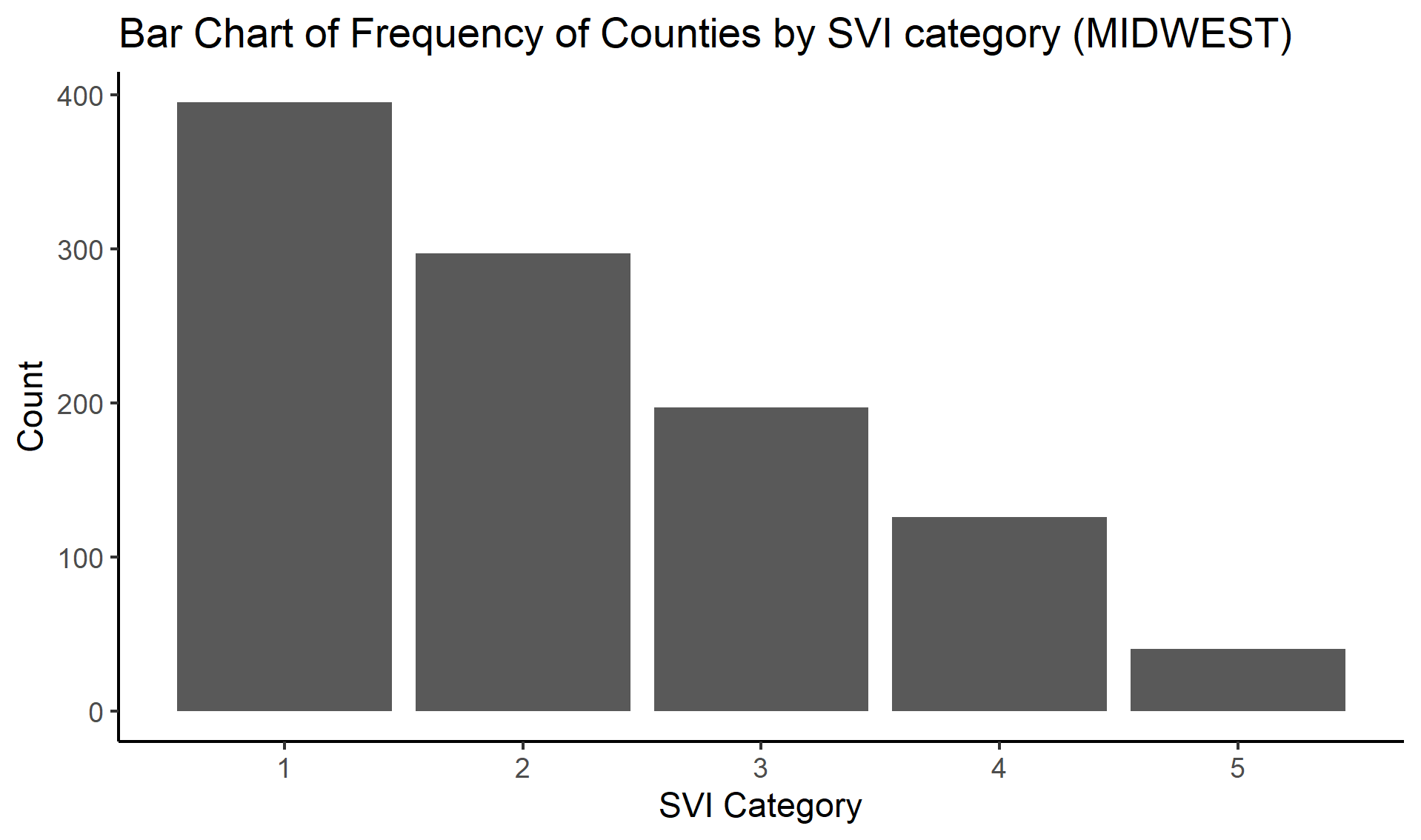
## 4.3 Exploratory analysis

##BAR Charts

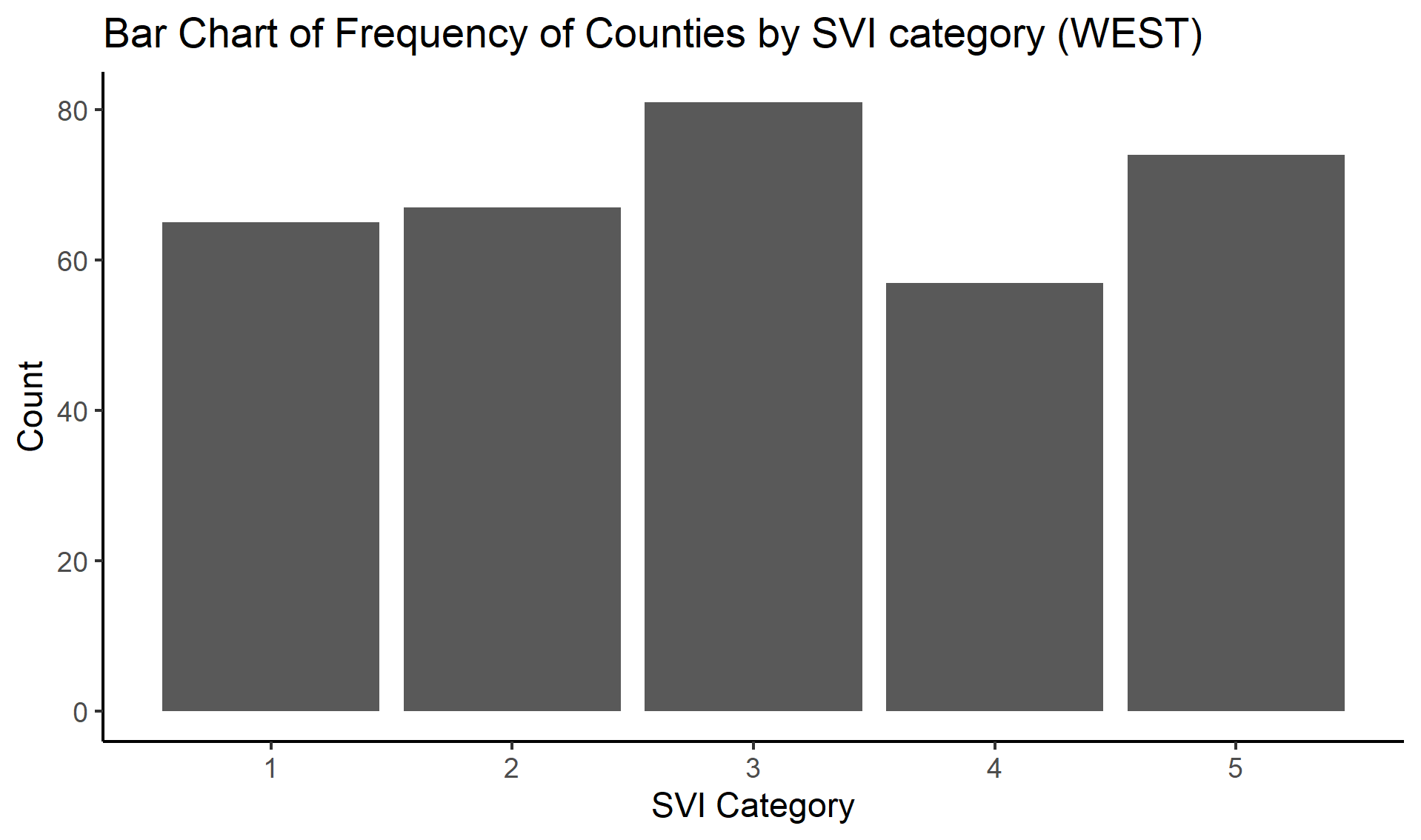
##Bar Chart of Frequency of Northeast states by SVI category



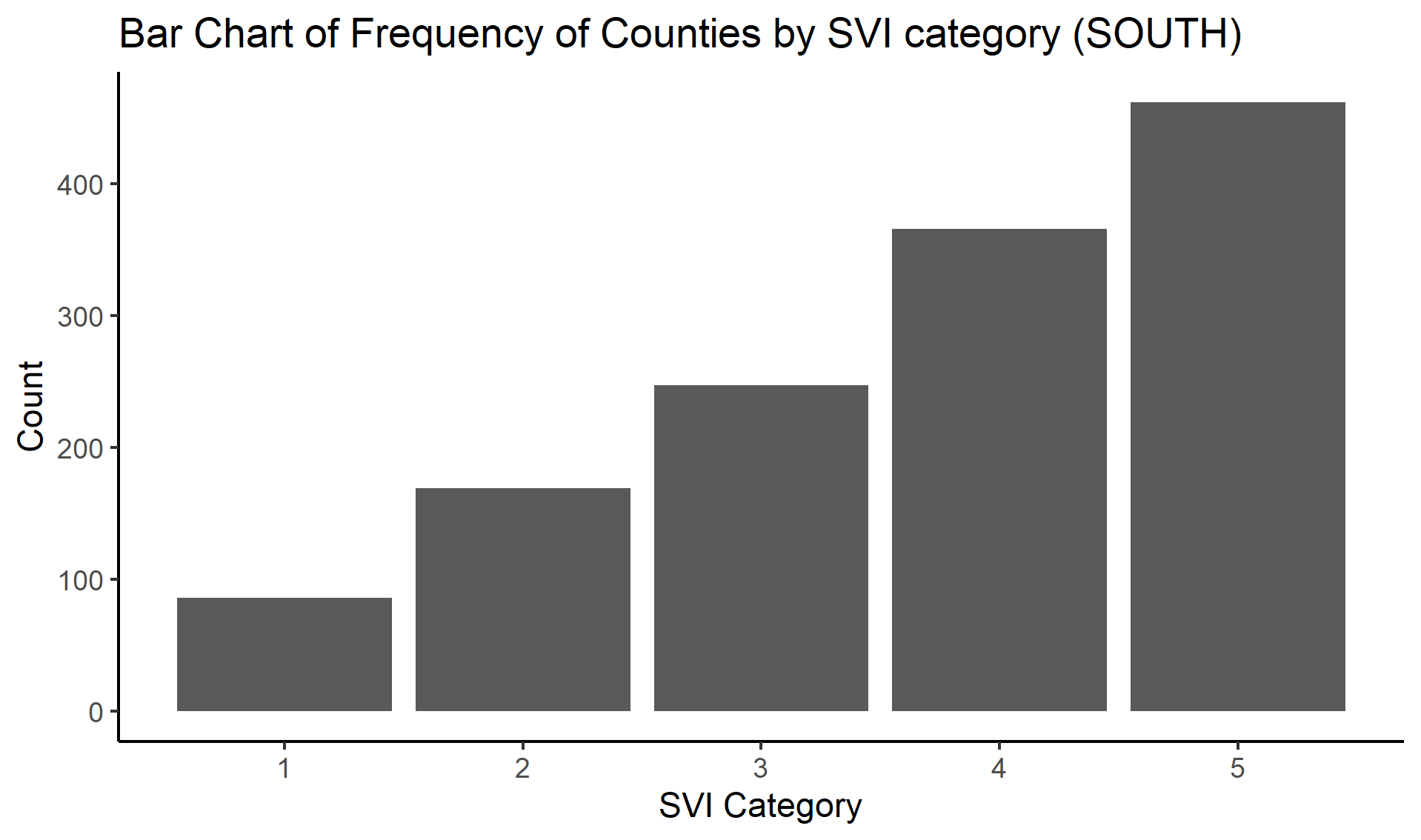
##Bar Chart of Frequency of Midwest states by SVI category



##Bar Chart of Frequency of West states by SVI category

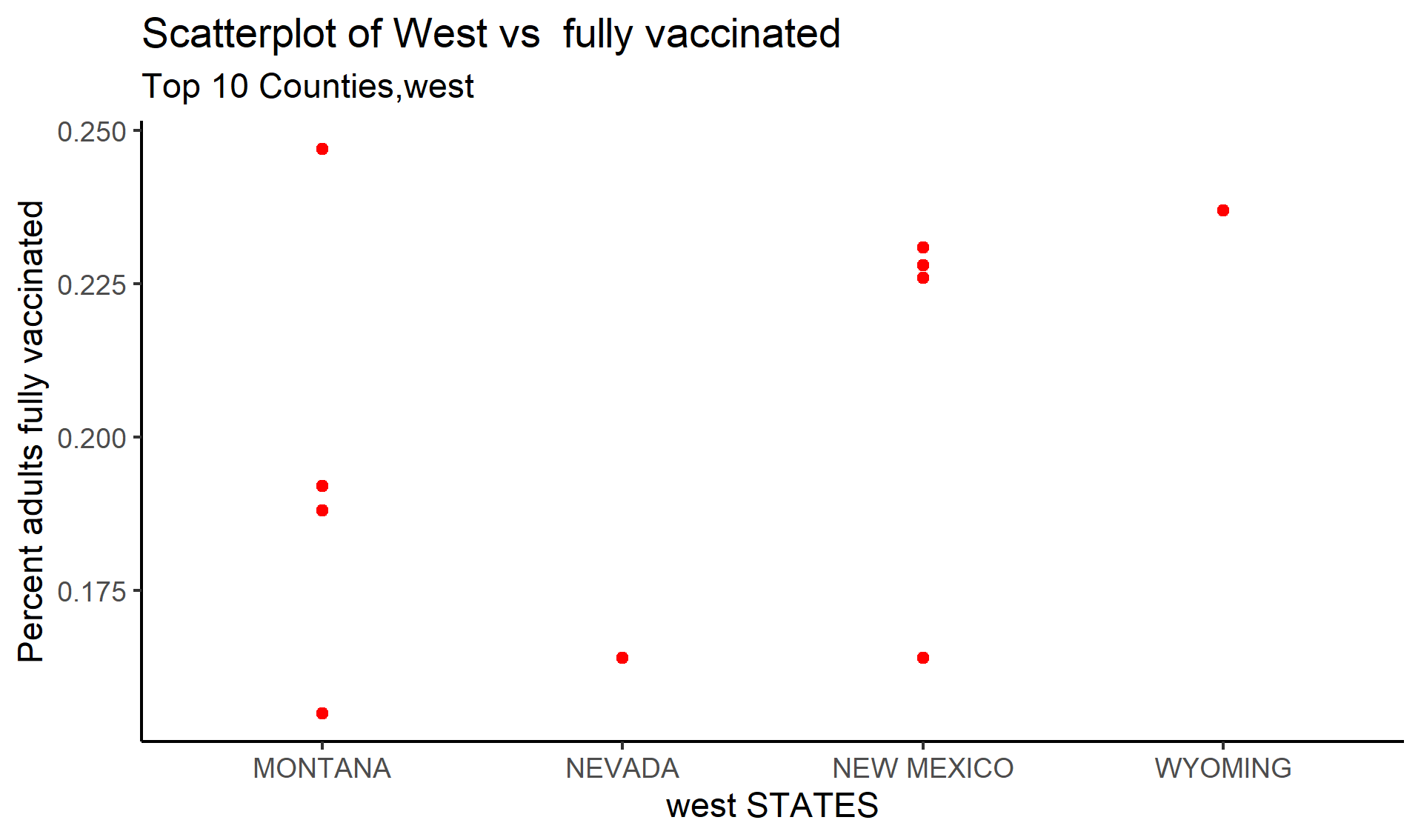


##Bar Chart of Frequency of South states by SVI category

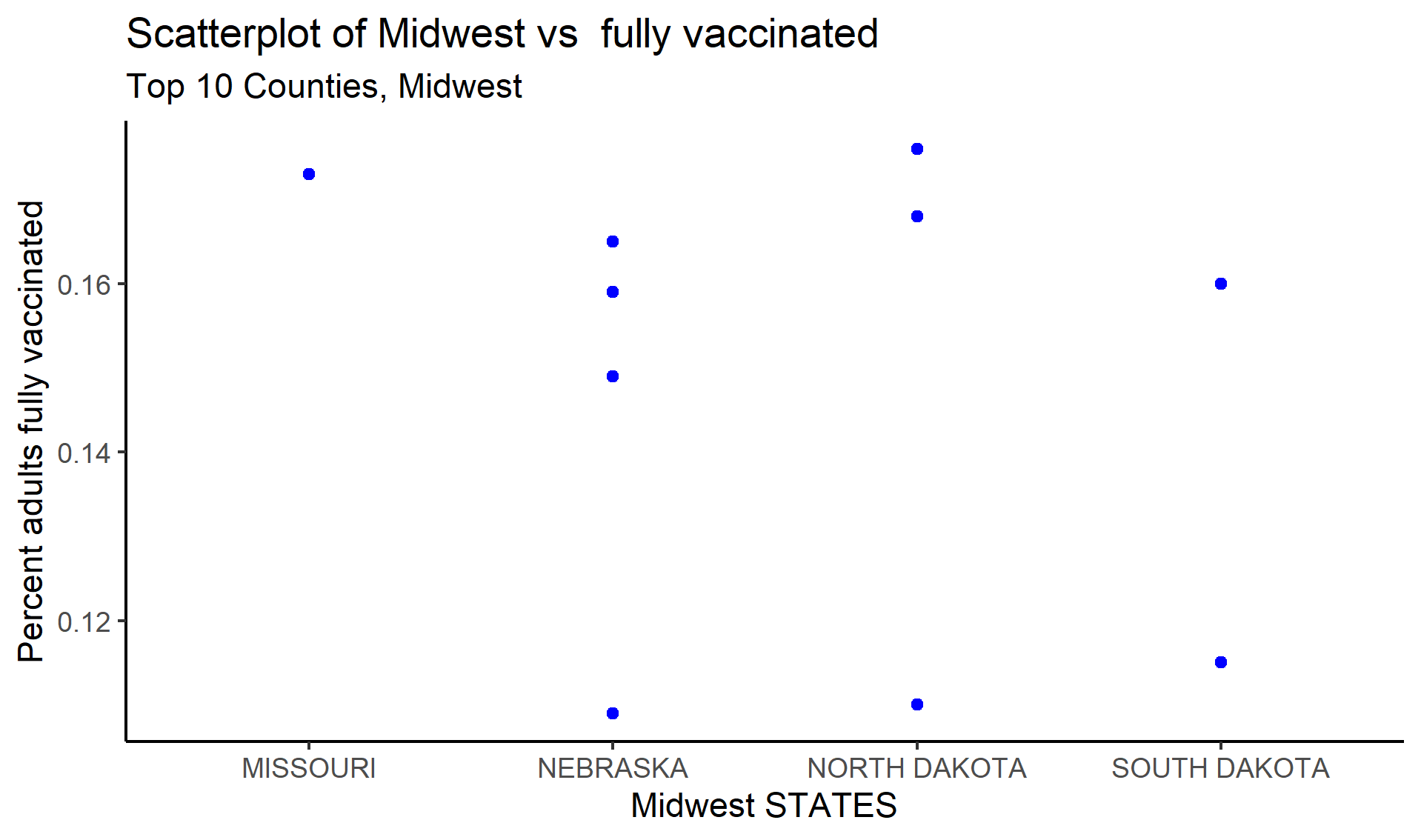


##SCatter plot between STATE and percent fully vaccinated (Top 10 counties)

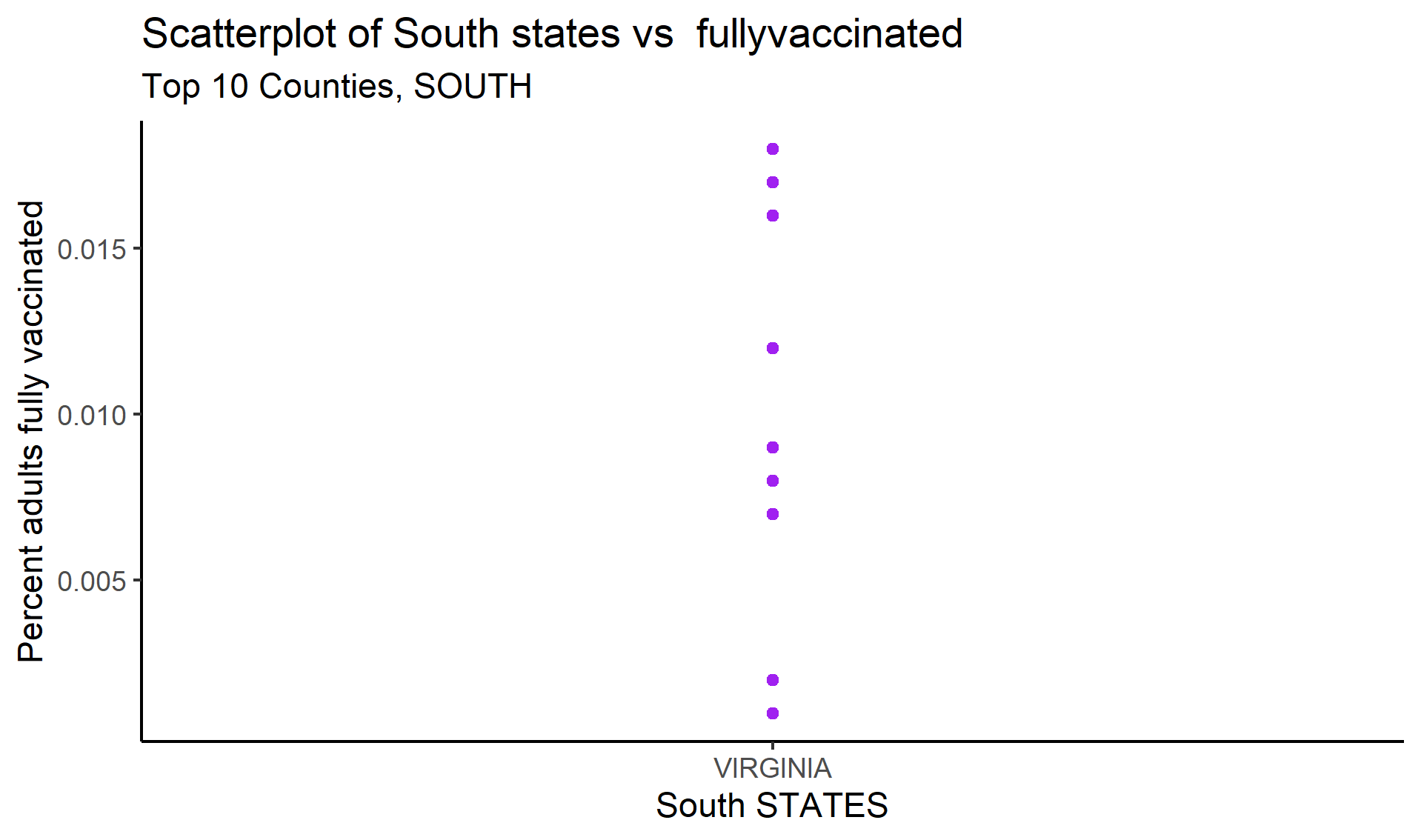
## 4.4 TOP 10 counties in west states v percent fully vaccinated



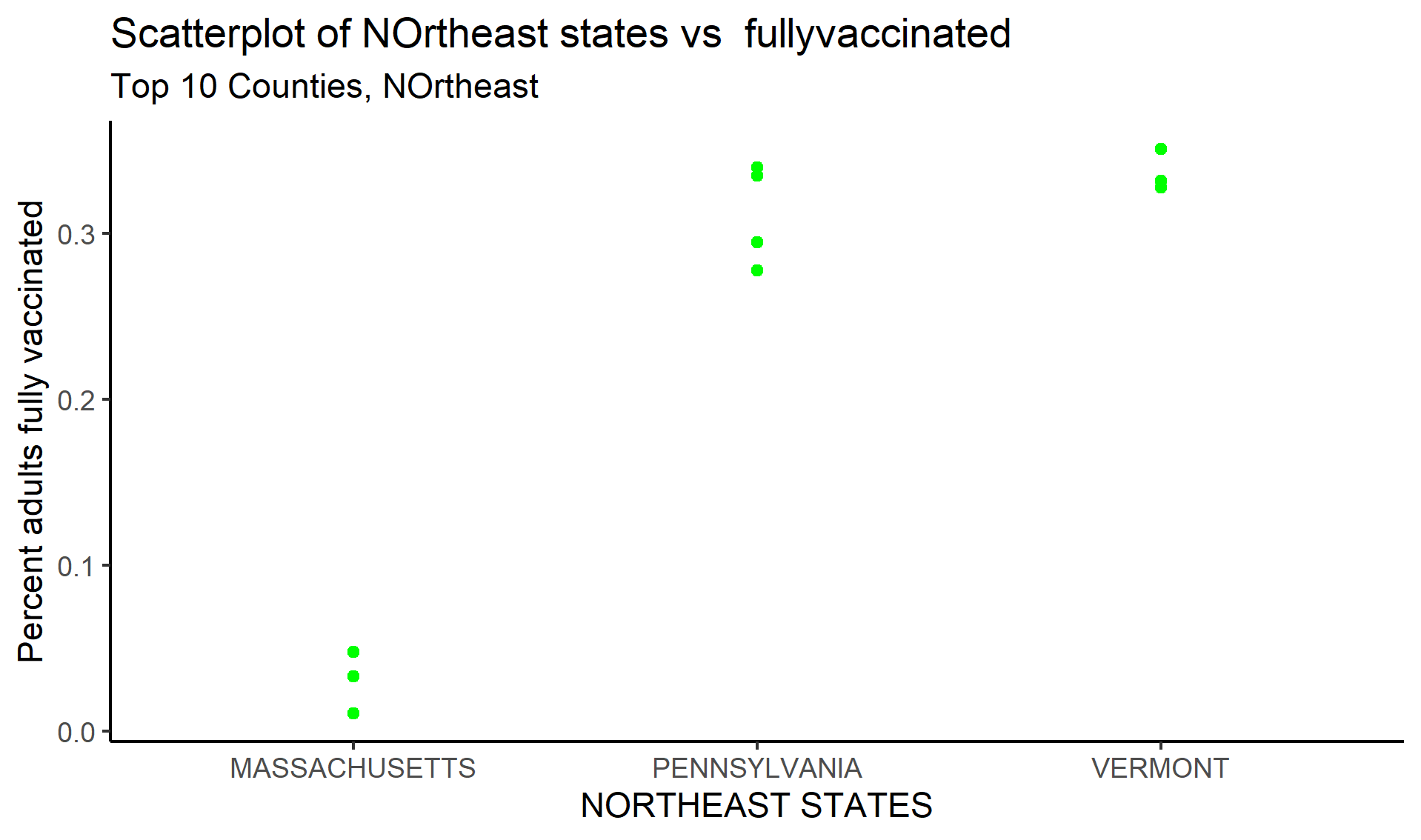
## 4.5 TOP 10 counties in Midwest states v percent fully vaccinated



## 4.6 TOP 10 counties in South states v percent fully vaccinated

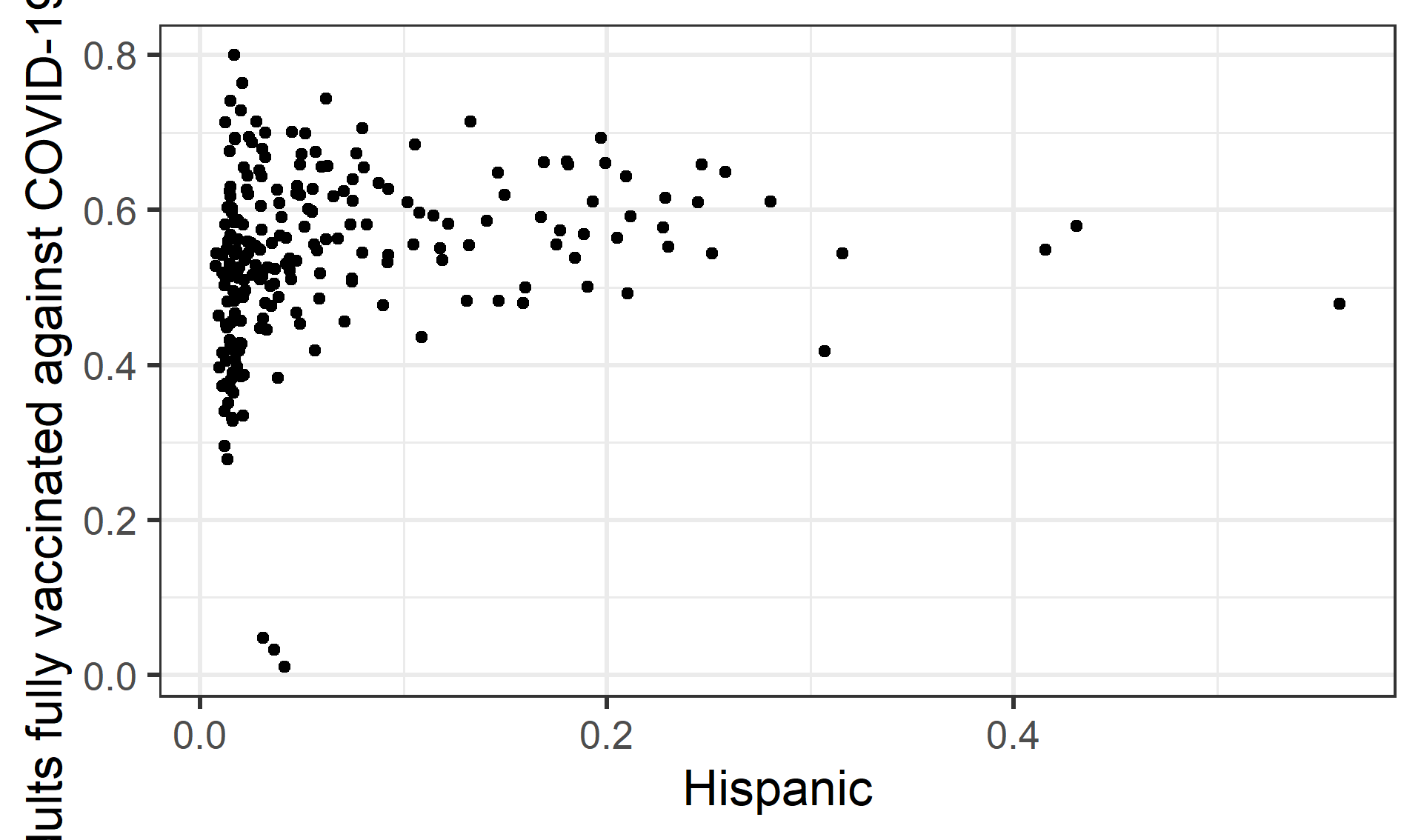


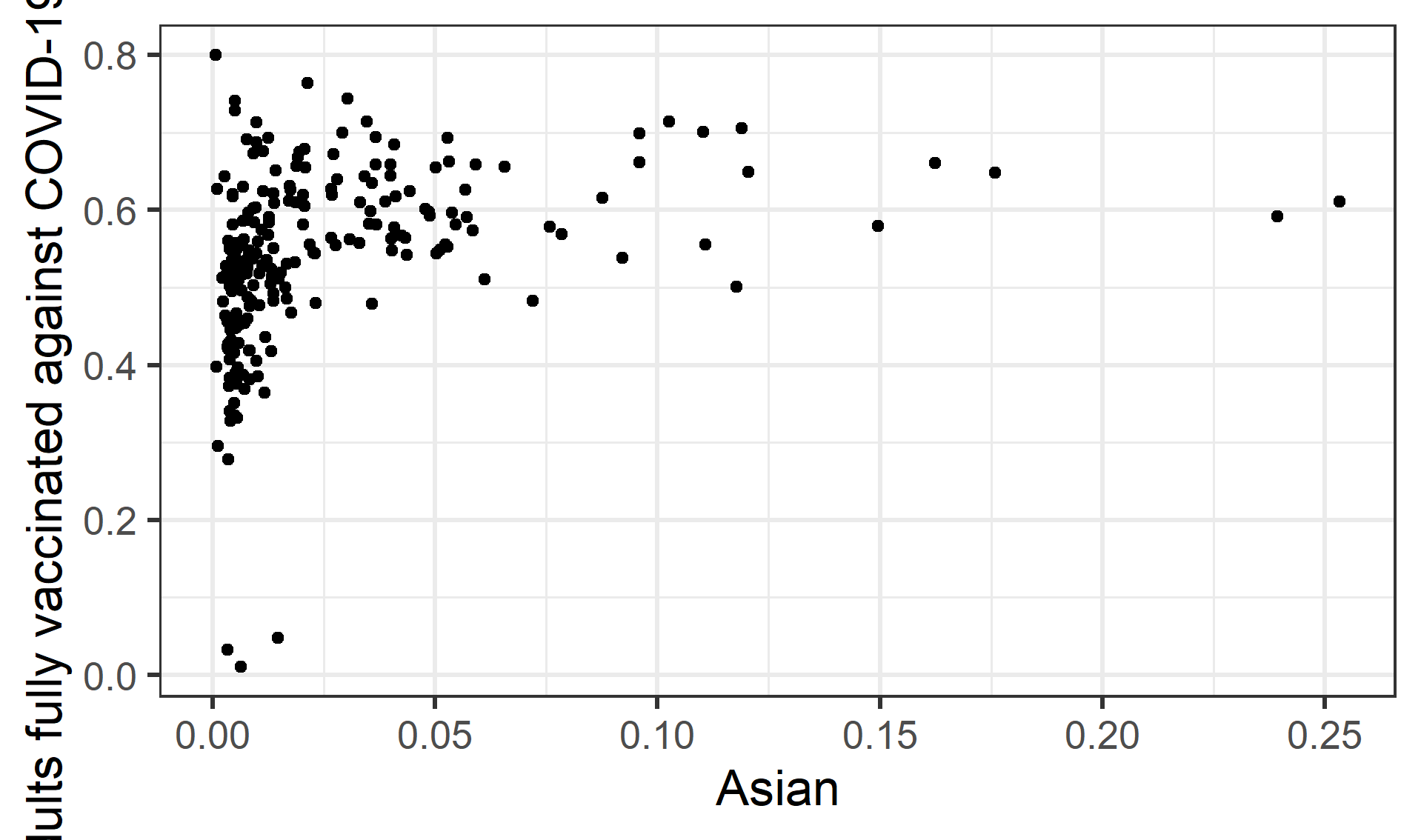
## 4.7 TOP 10 counties in NOrtheast states v percent fully vaccinated

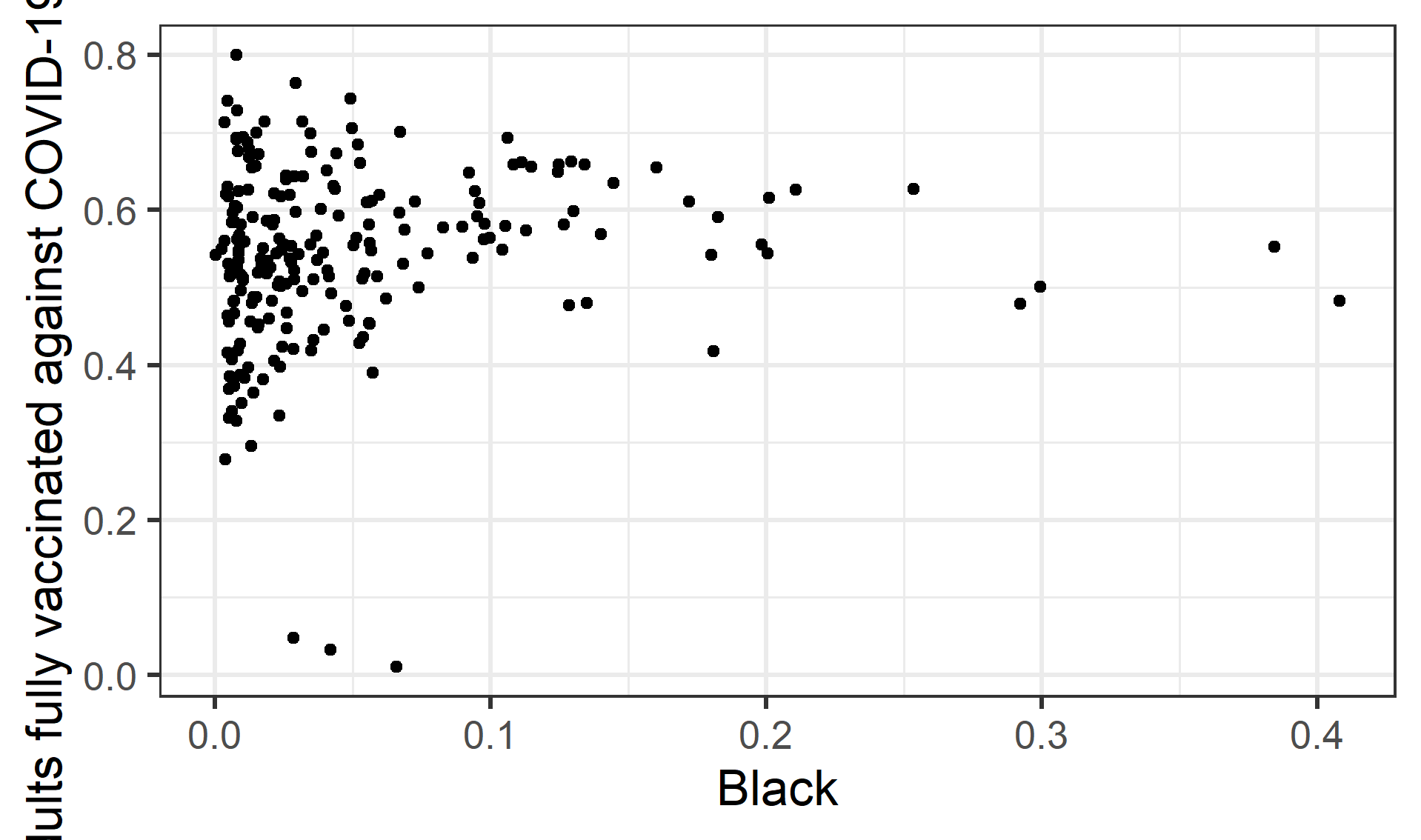


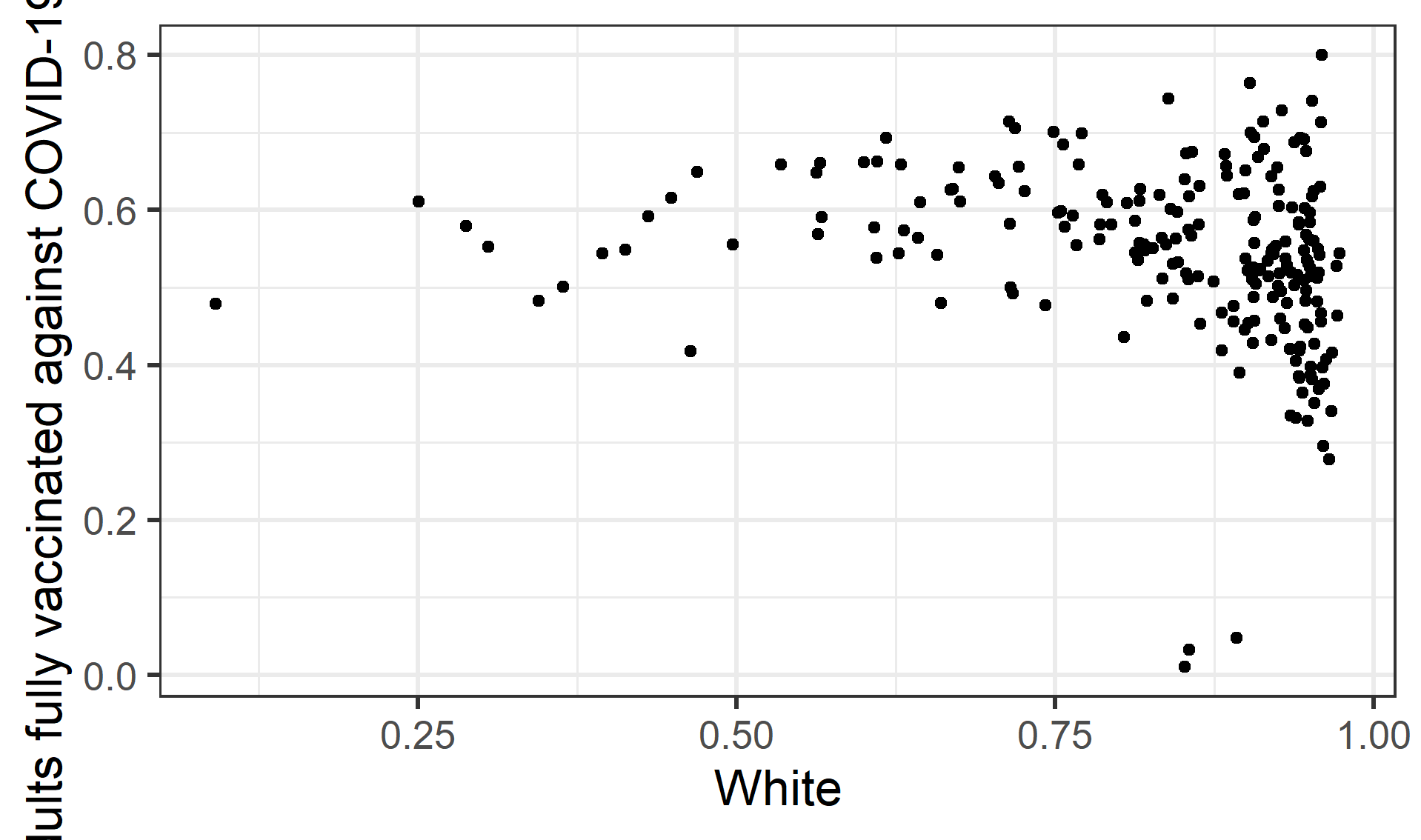
## 4.8 Comparing ethinicity in percent fully vaccincated in all US regions

##Northeast Vs ethinicity

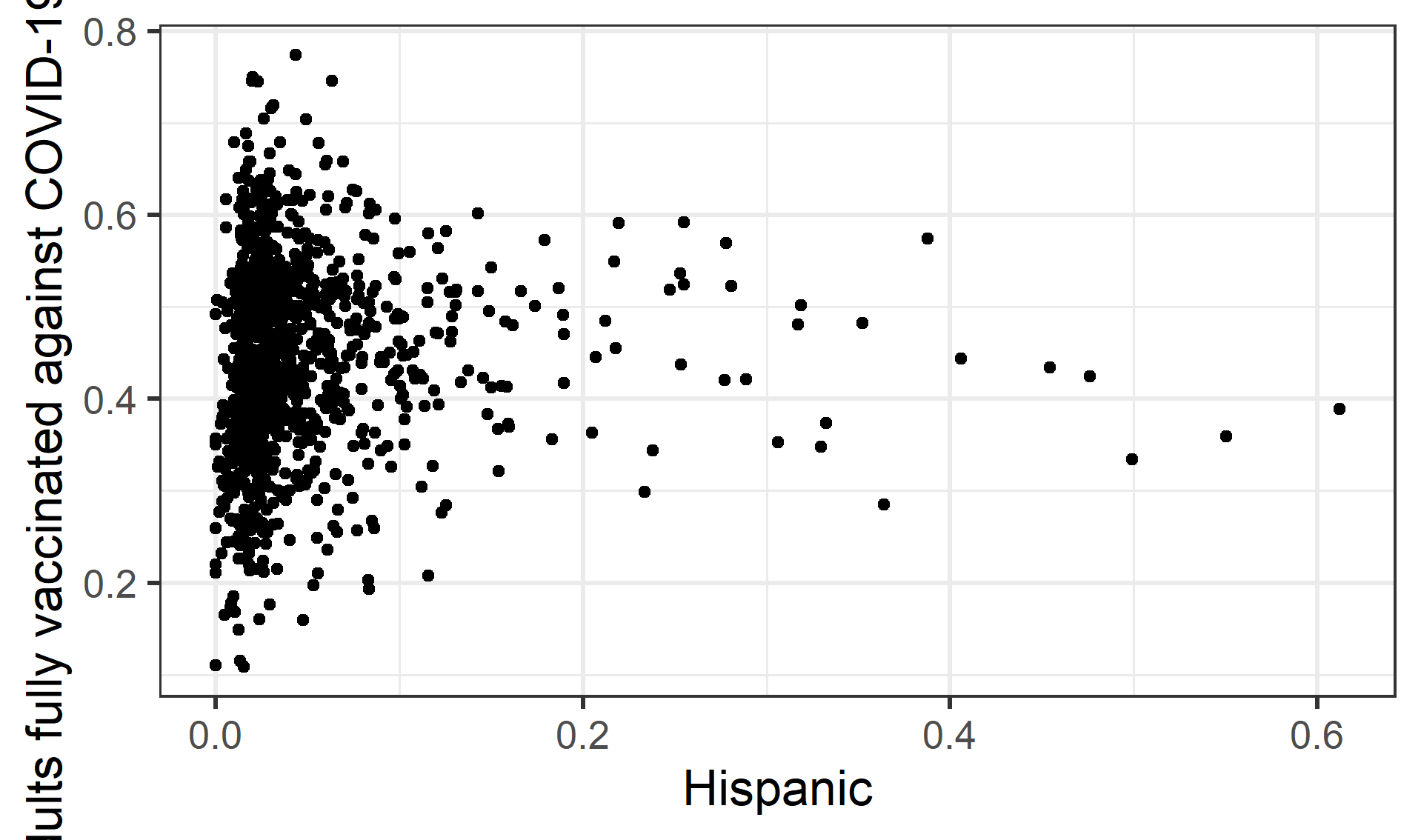


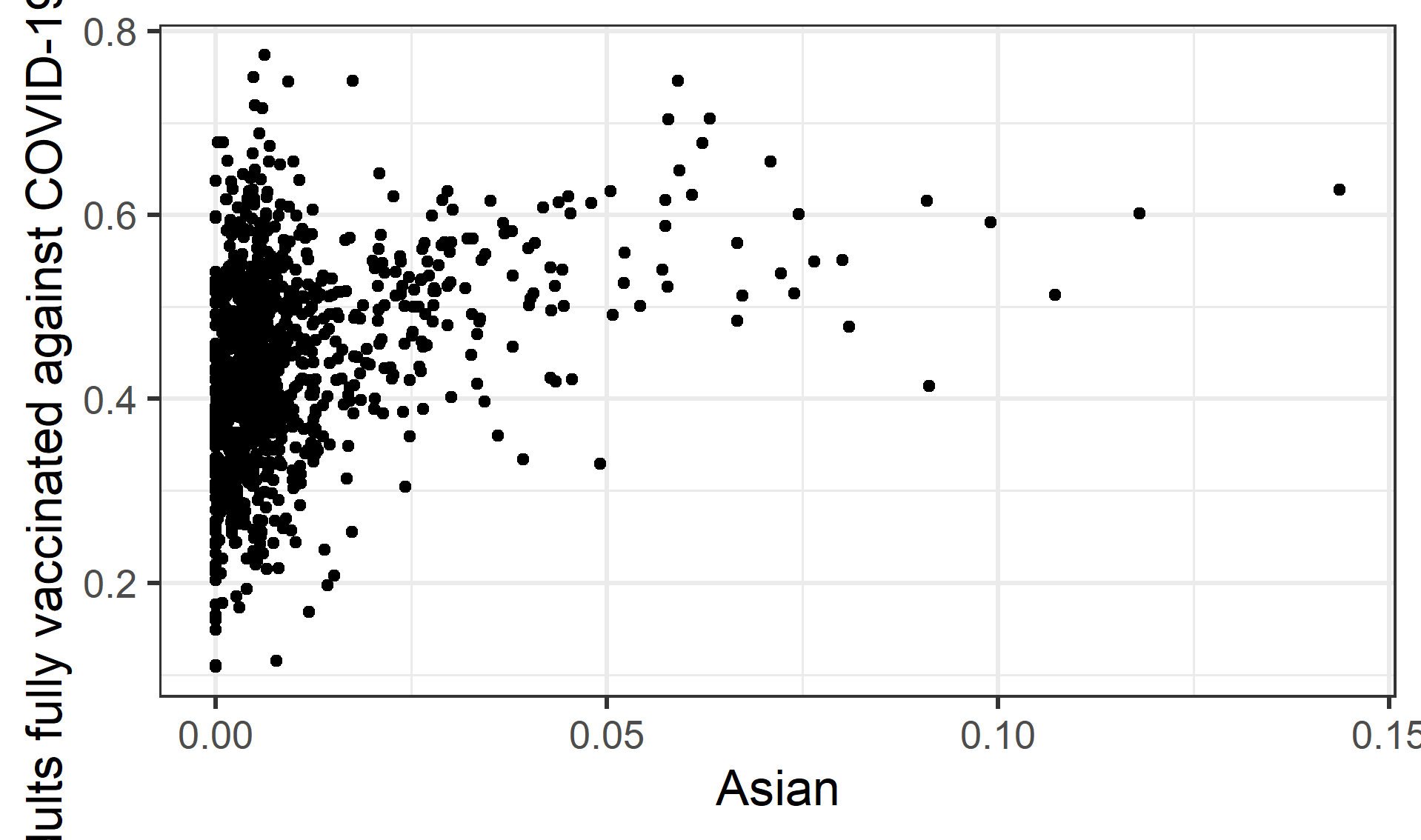


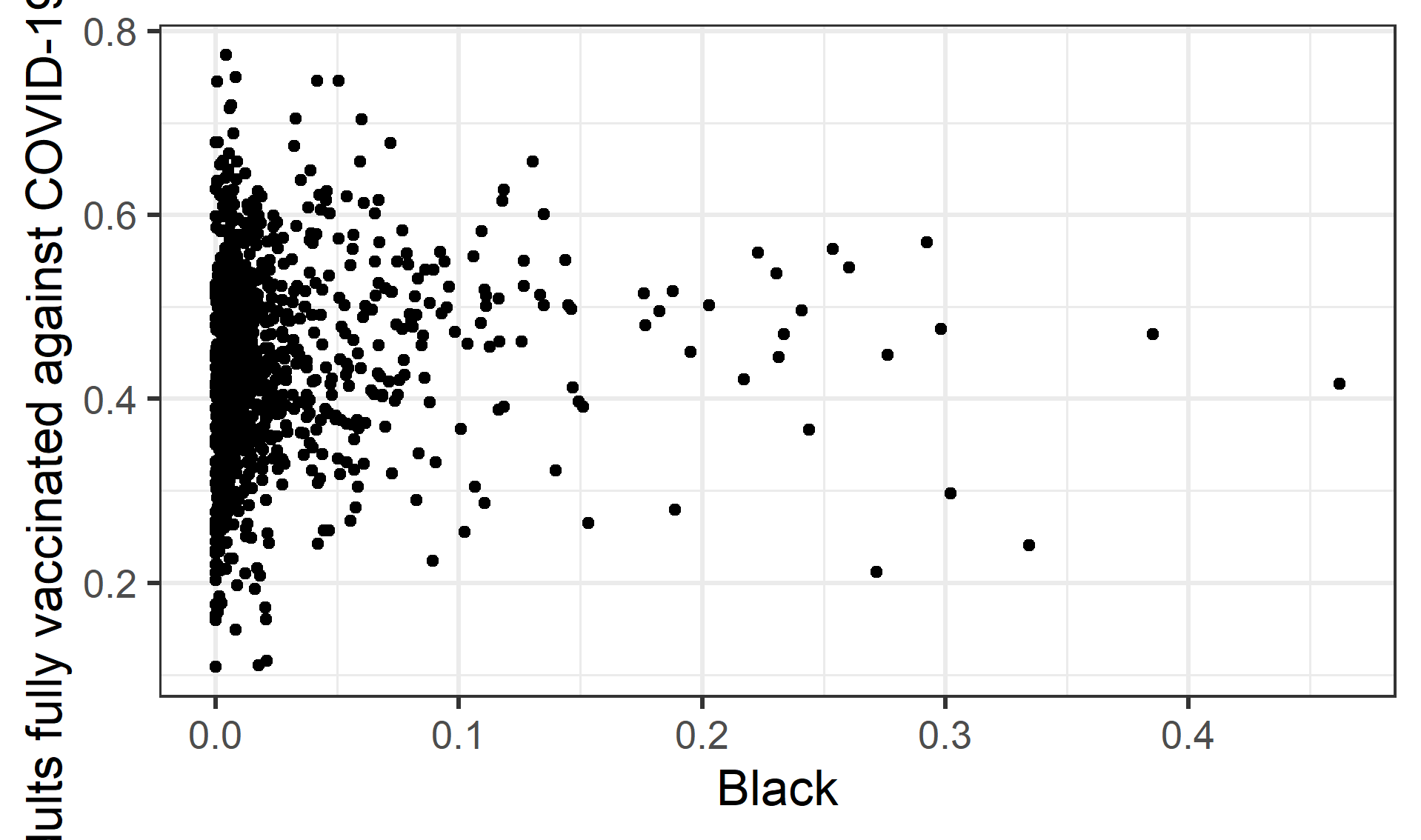


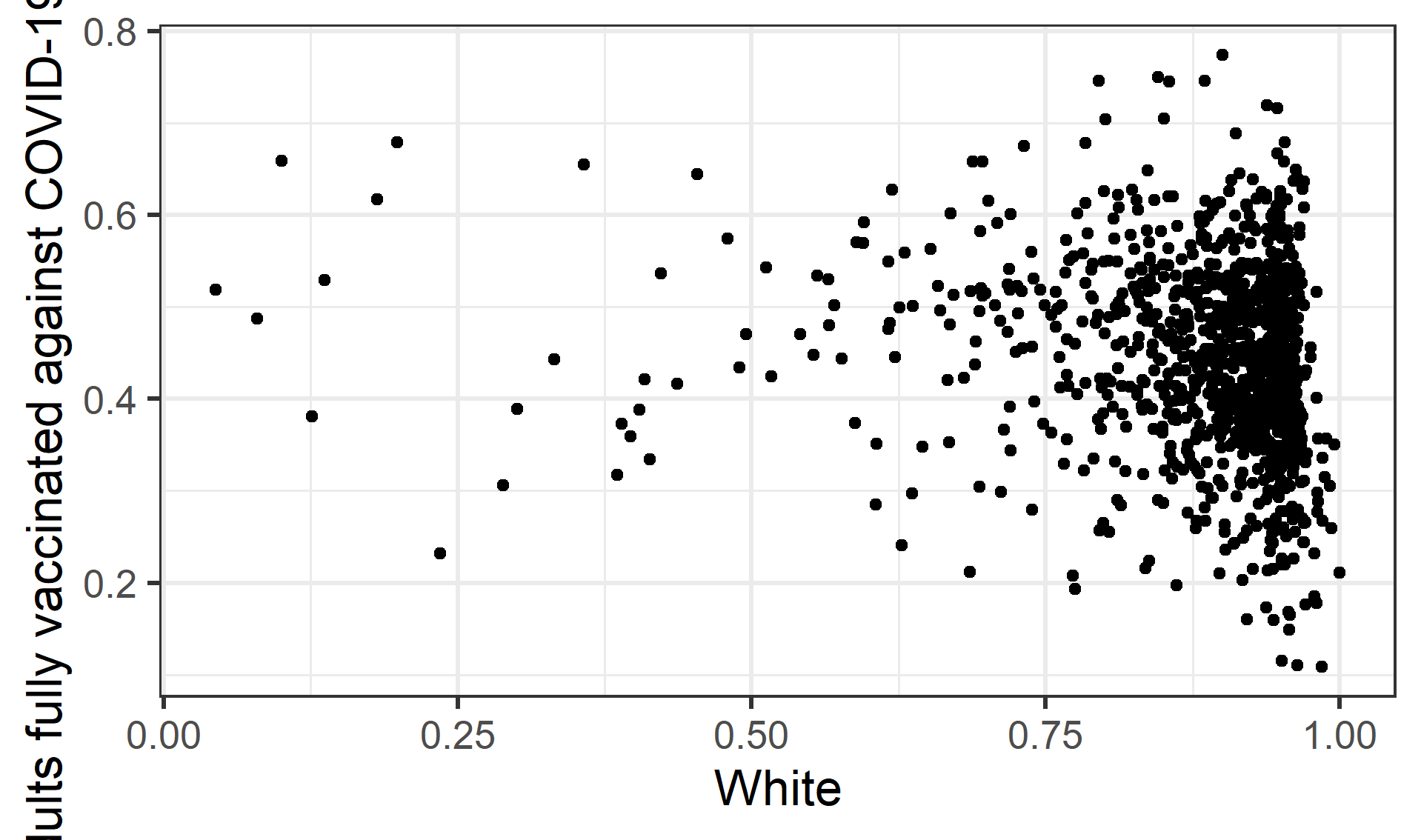


##Midwest Vs ethinicity

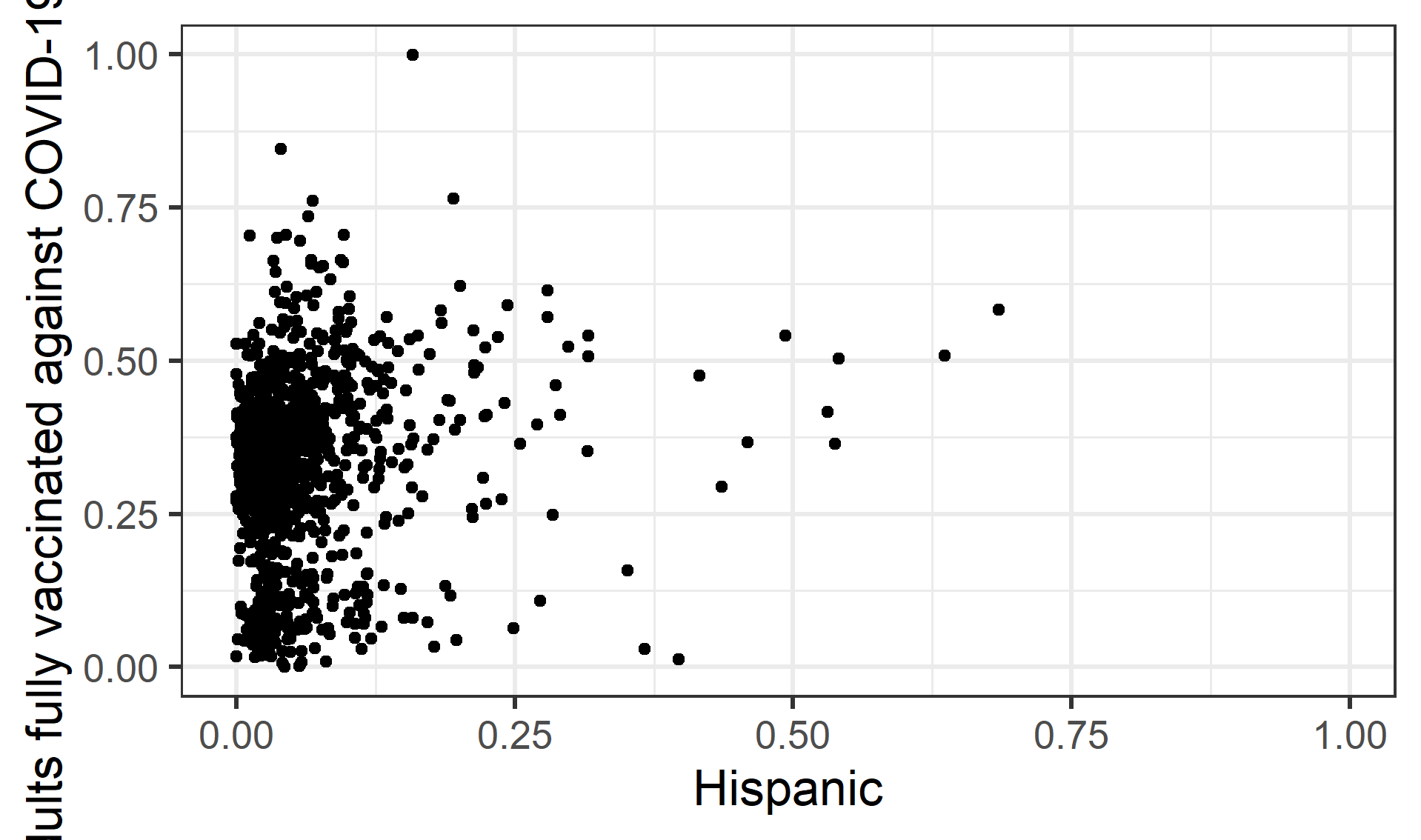


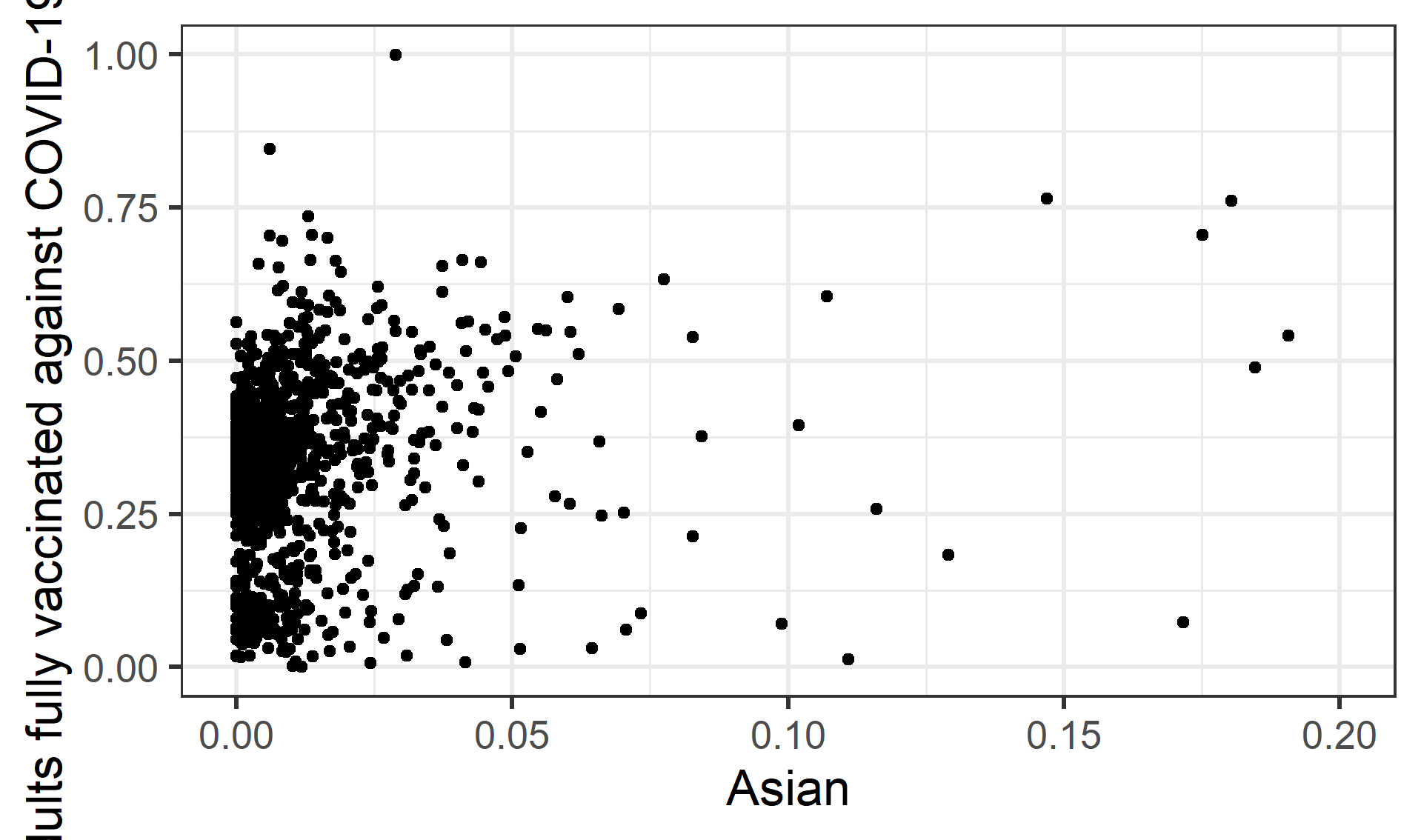


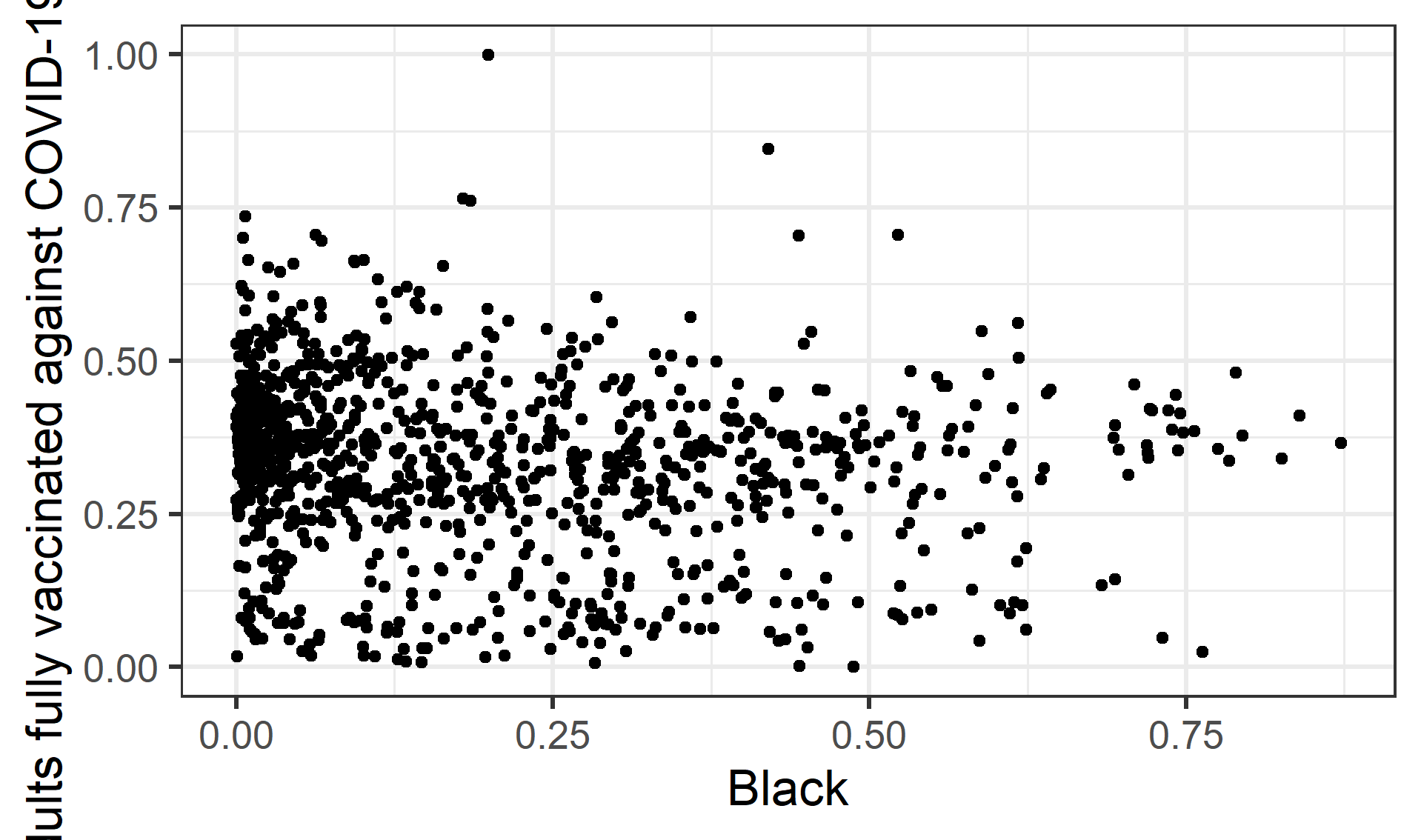


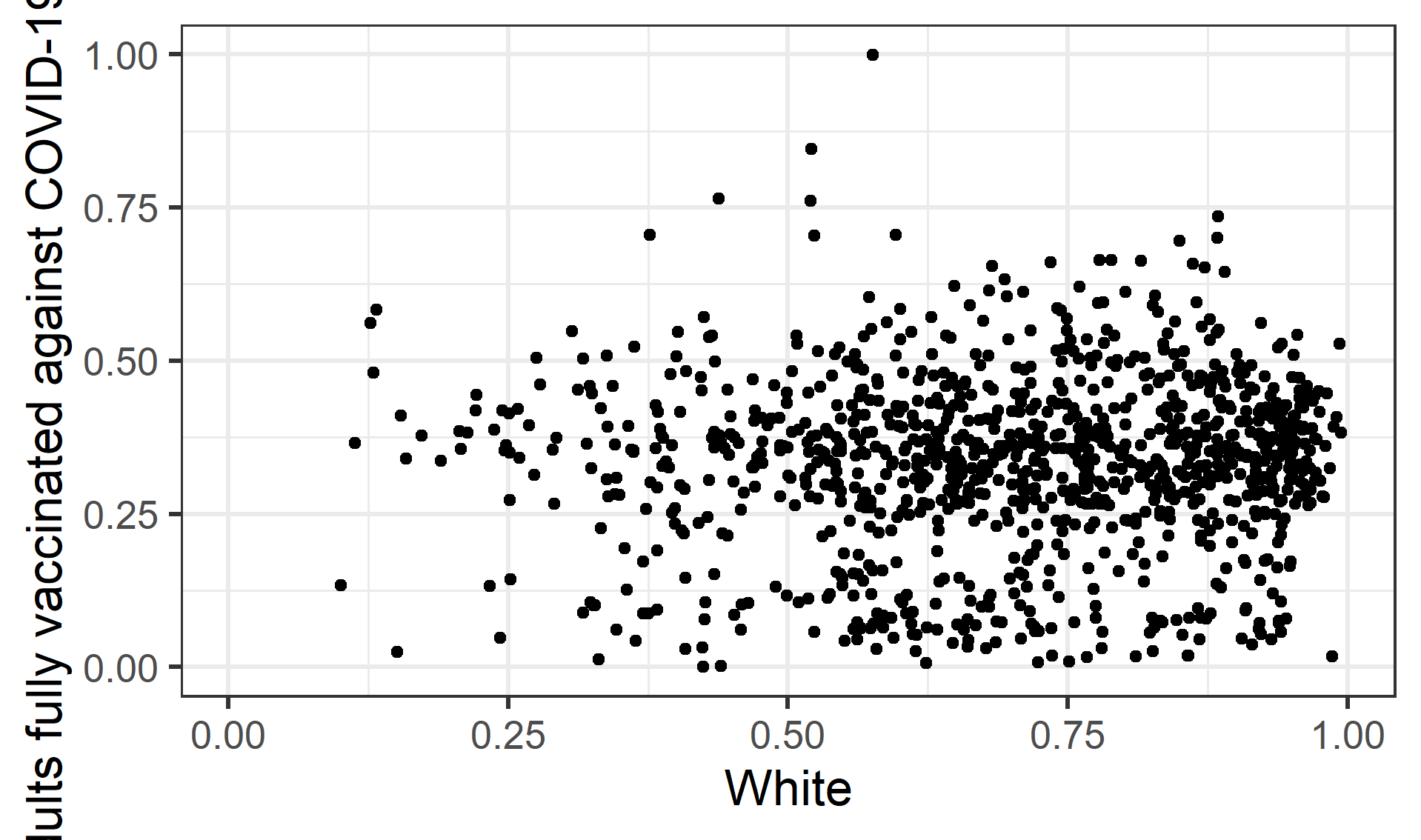


##South states Vs ethinicity

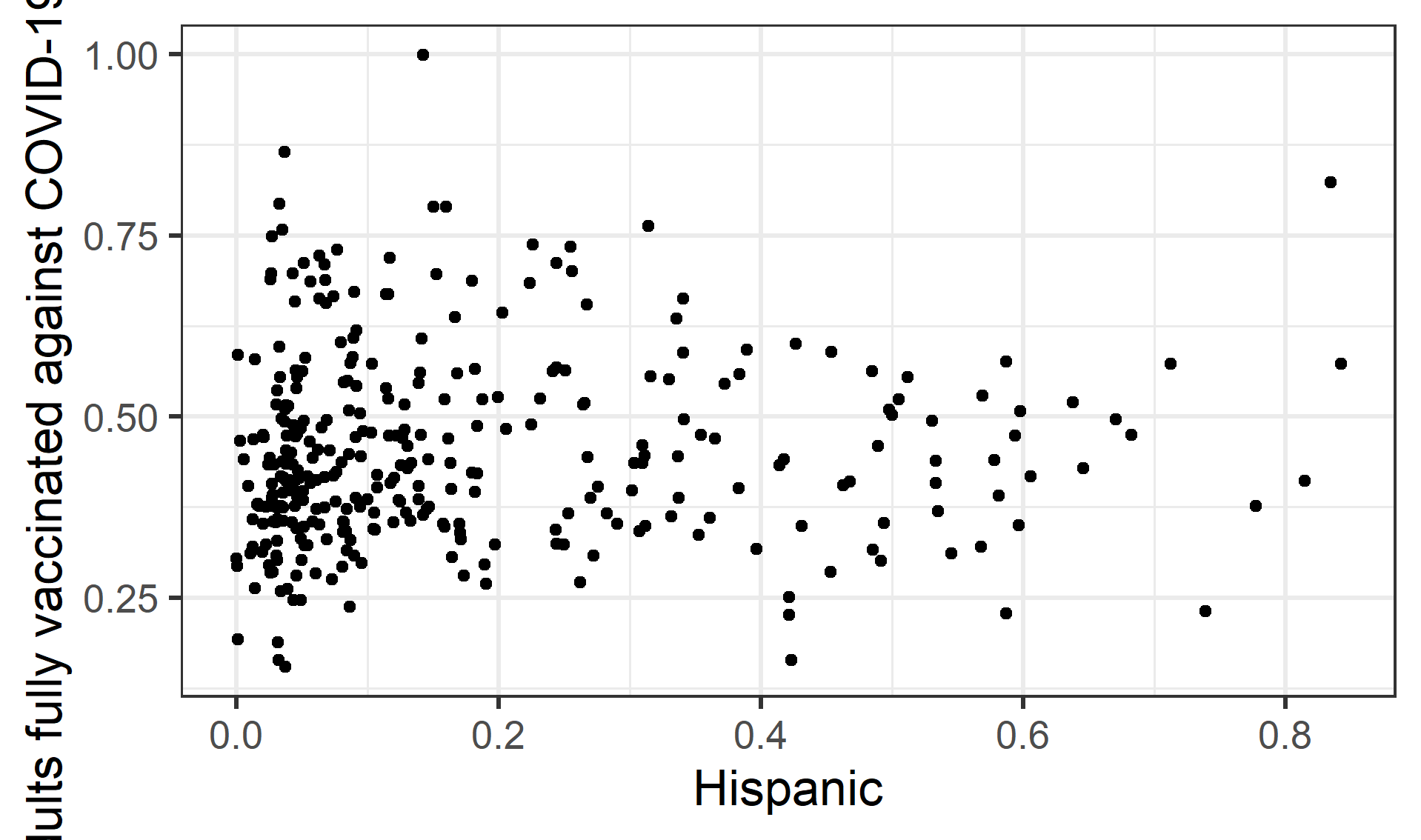


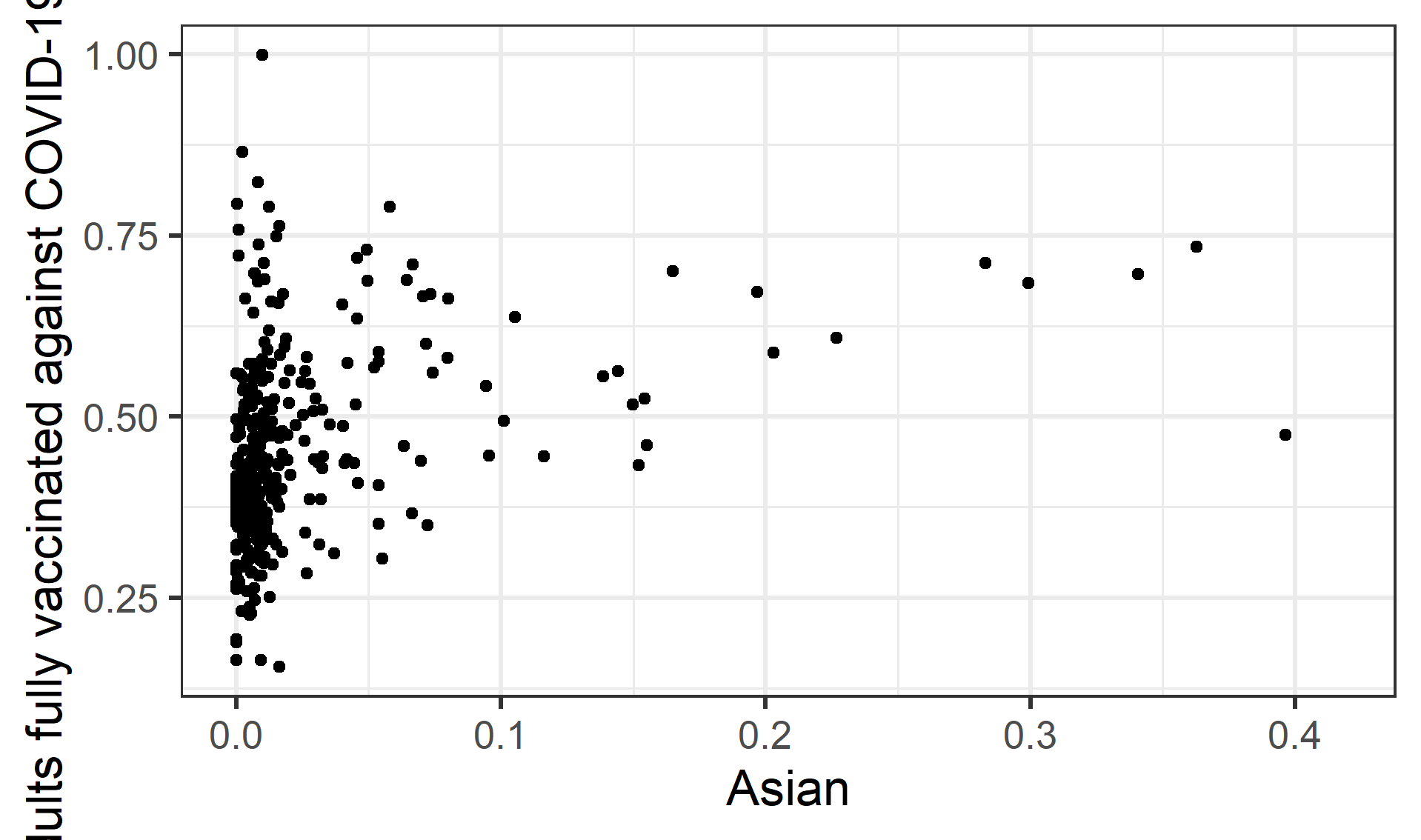


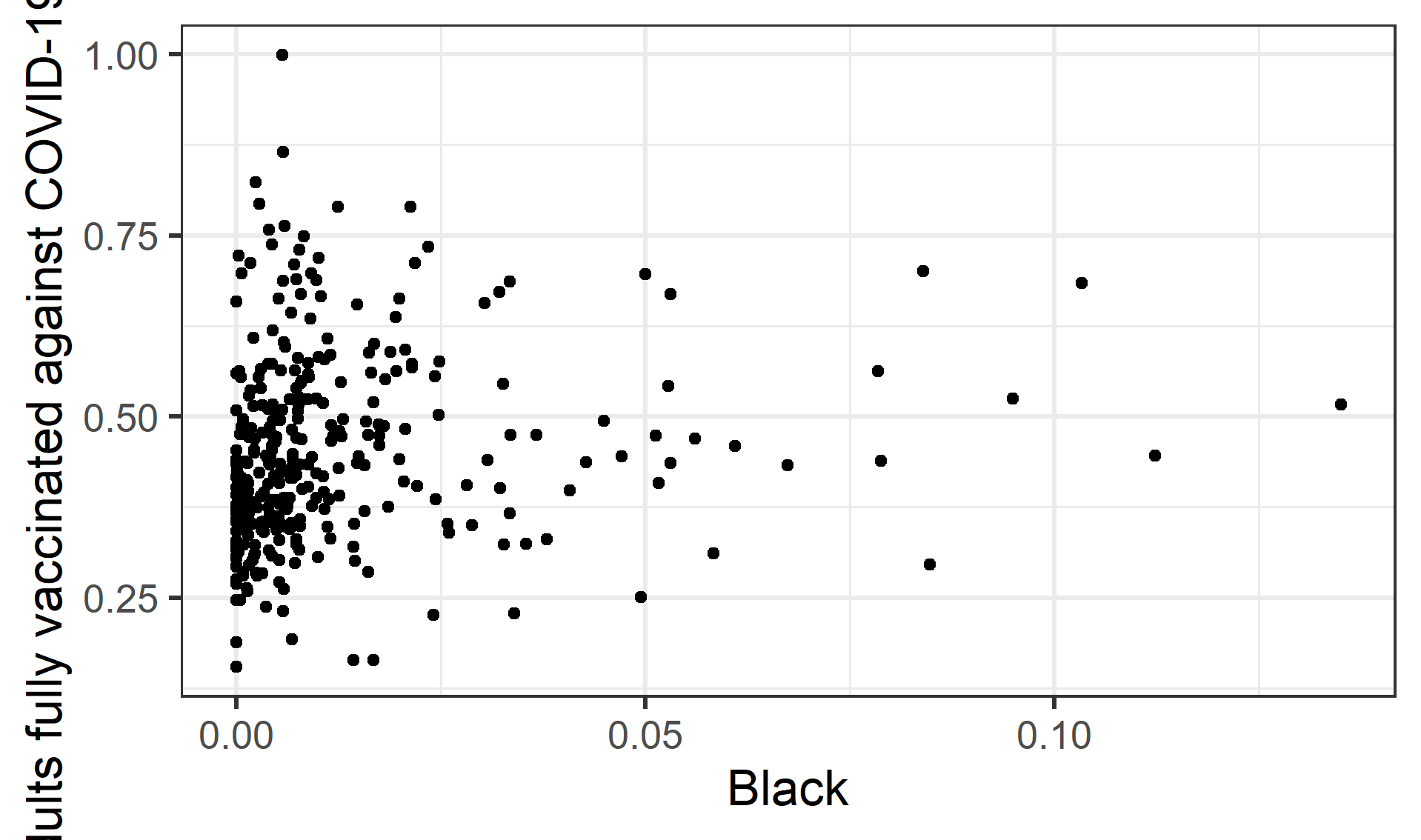


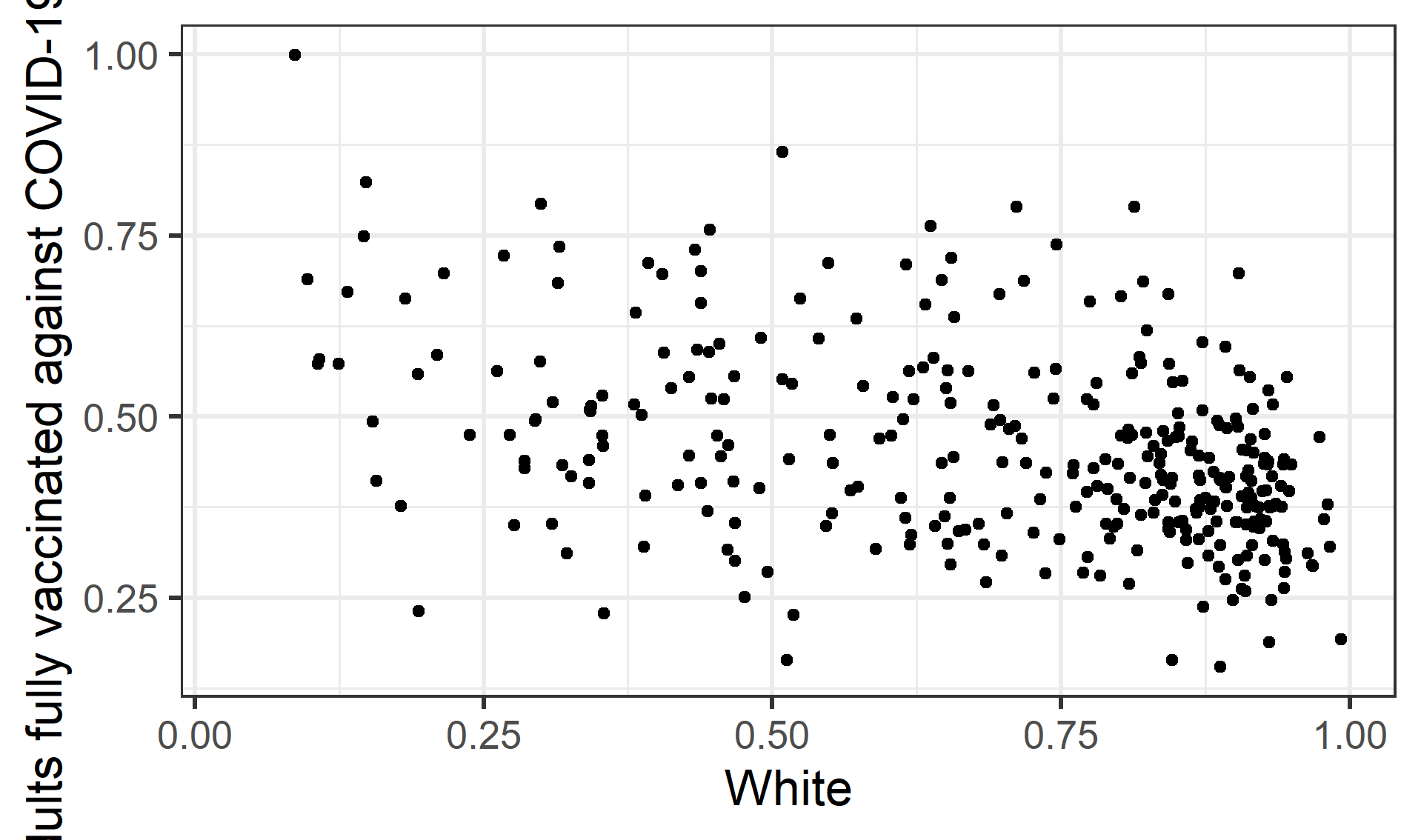


##West states Vs ethinicity







 ## Full analysis

# 5 fit linear model

# 6 Create a linear model of Percent fully vaccinated and race in all regions

Table 6.1: Linear model fit table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.7767846 | 0.0370430 | 20.969811 | 0.0000000 |
| White | -0.3913793 | 0.0408063 | -9.591156 | 0.0000000 |
| Asian | 0.4969402 | 0.1480797 | 3.355898 | 0.0008848 |
| Black | -0.8114566 | 0.4155598 | -1.952683 | 0.0517153 |
| Hispanic | -0.3163635 | 0.0489379 | -6.464595 | 0.0000000 |

# 7 Discussion

## 7.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 7.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 7.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*, 1314–1315. <https://doi.org/10.1126/science.aaa6146>