How does public behavior affect the spread of COVID-19? - Initial report

Quang Vuong

Contents

I am interested in the present dataset because of many reasons. Most importantly, I would like to work in public health in the future, so this project seems like a good starting point. Furthermore, it is shown in the news that there are aspects of public behavior, such as mask-wearing and social distancing, which are more important in controlling the spread of COVID-19 than others, so I wonder if we can quantify this exactly.

0.1 Introduction

We wish to investigate how public behavior affects the spread of COVID-19 within a community. In particular, our goal is to identify which aspects of public behavior, such as visiting restaurants, visiting bars and using public transit, predict case counts. Since many other factors including population demographics, population density and policy decisions also affect the spread of COVID-19, it is best to focus on datasets for a sufficiently small geographical area (like a city or a county) so that the confounding factors unrelated to the study are controlled.

The present dataset, collected from the `{covidcast}` API, contains daily COVID-19 case counts for 88 days from June 5th, 2021 to September 5th, 2021, along with several indicators of public behavior, in Manhanttan, New York. The indicators of public behavior are:

- `distancing` = Percentage of survey respondents reporting that people maintained a distance of at least 6ft (%)
- `public_transit` = Percentage of survey respondents reporting that they used public transit in the last day (%)
- `worked_outside` = Percentage of survey respondents who was indoors (excluding home) in the last day (%)
- `large_events` = Percentage of survey respondents who attended a crowded event in the last day (%)
- `mask_prop` = Percentage of survey respondents who mostly wore a mask outside in the last week (%)
- `other_mask_prop` = Percentage of survey respondents saying that other people mostly wore a mask outside (%)
- `bar_visit` = Number of bar visits per 100000 people (visits per 100000 people)
- `resto_visit` = Number of restaurant visits per 100000 people (visits per 100000 people)

Most of these indicators are obtained from Facebook surveys. Taking into account the issues of survey data such as response bias and subjectivity of answers, it is extremely unlikely that the variables in the dataset track their true respective quantities. Instead, we opt to interpret these indicators as proxies of public behavior, which are potentially useful for prediction.

Now we will summarize our two objectives for this study:

1. Identify which of the above indicators are important in predicting COVID-19 case counts.

2. Identify a model which can be used to predict future COVID-19 case counts.

0.2 Exploratory data analysis

We first examine the descriptive statistics of the response and all predictors. We present the mean and standard deviation of the predictors and case counts in Table 1. The units are as described in the introduction.

Table 1: Descriptive statistics of predictors and case counts.

	Mean	Standard.deviation
distancing	18.43	5.31
bar_visit	55.30	40.10
large_events	19.63	3.51
mask_prop	65.84	7.81
other_mask_prop	40.26	10.25
public_transit	53.55	2.81
resto_visit	101.36	14.48
worked_outside	28.63	2.49
cases	179.94	150.07

It is clear that all variables have very different scales from each other, so models with unstandardized and standardized variables should be examined. However, these statistics do not inform of the distribution of predictors. To gain information about this, we look at boxplots of the predictors and case counts, as in Figure 1. We can see that the distributions of `distancing`, `bar_visit`, `other_mask_prop`, and `cases` are skewed to the right, while the distributions of the others are roughly symmetrical. This concludes our examination of individual variables.

Now, we will look at how predictors are related to case counts and each other. We plot all predictors against the number of cases in Figure 2. `distancing`, `bar_visit` and `large_events` appear to have two clusters with different mean case counts. On the other hand, `mask_prop`, `other_mask_prop`, `resto_visit` and `worked_outside` appear to have two trend lines. `public_transit` also looks like there are two trend lines, but both are quite flat. This strongly suggests that there are two clusters within the data that exhibit different relationships between case counts and public behavior.

To continue the examination of the dataset, we will now look at the covariance matrix of the predictors.

```
##      distancing    large_events other_mask_prop
##       1.0000000      -0.7612822       0.7463194
##  bar_visit  mask_prop
##  1.0000000 -0.6964321
##      distancing    large_events other_mask_prop
##      -0.7612822       1.0000000      -0.6218936
##       bar_visit       mask_prop other_mask_prop
##      -0.6964321       1.0000000       0.6837441
##      distancing    large_events       mask_prop other_mask_prop
##       0.7463194      -0.6218936       0.6837441       1.0000000
## [1] 1
## [1] 1
## [1] 1
```

There is some substantial correlation between some predictors. In particular, `bar_visit`, `mask_prop`, `large_events` and `other_mask_prop` seem correlated with each other. Of these predictors, `large_events` appear to have two clusters with different mean case counts, while the remaining predictors seem to have two trend lines against case counts.

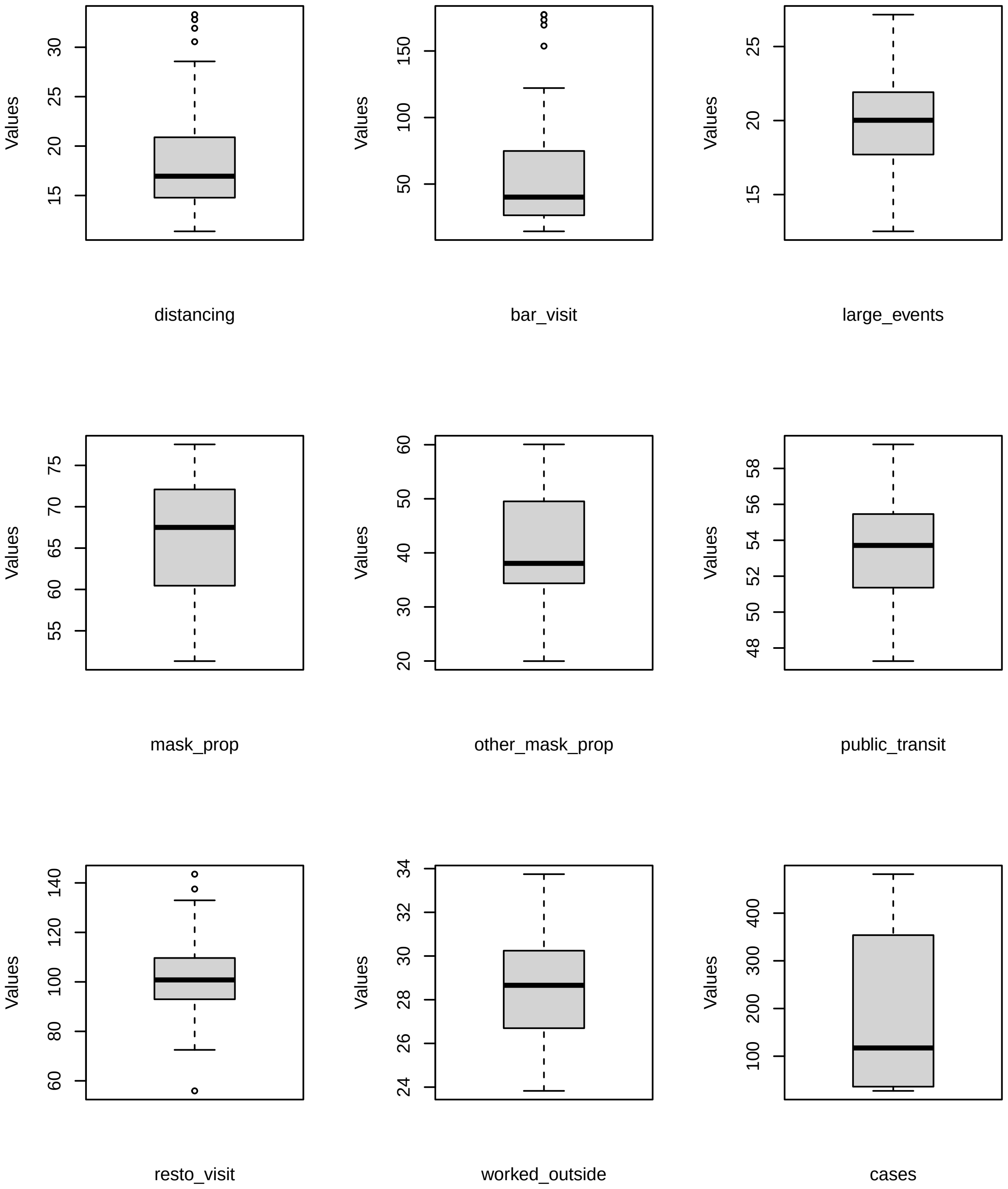

Figure 1: Boxplots of individual predictors and case counts.

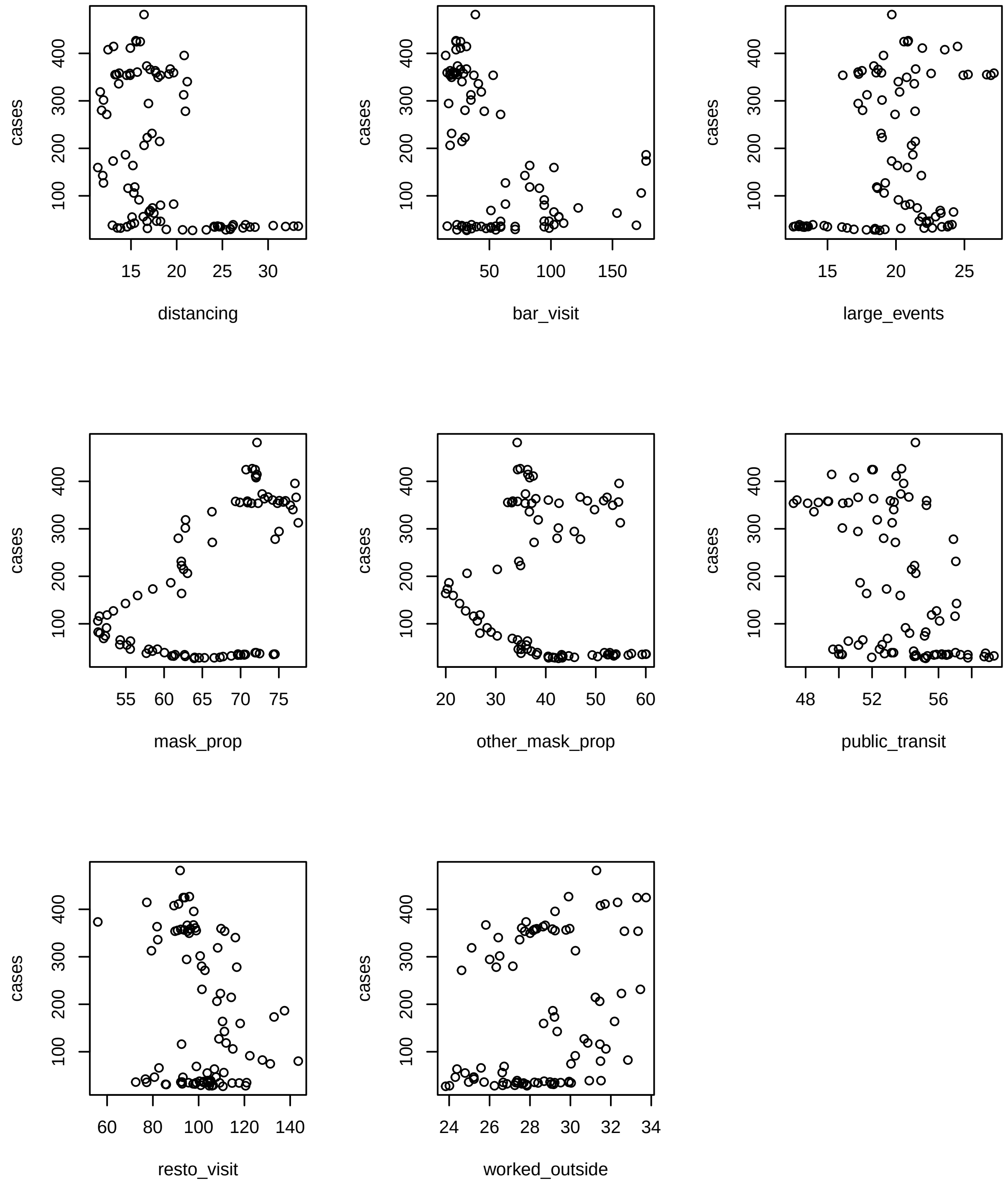

Figure 2: Scatterplots of predictors against case counts.

The most striking observation made so far is that there might be two clusters within the data that shows different relations between case counts and public behavior indicators, so we will now try to identify how the clusters are split. To do this, we attempt to fit a regression tree and look at the split at the root. Please note that we are only using this tree to expedite the exploration rather than seriously considering as a model. The fitted tree is plotted in Figure 3.

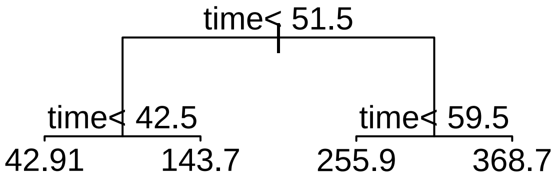

Figure 3: Regression tree to determine clusters in data.

Initial inspection of this tree shows that there is clustering by date. The set of plots in Figure 4 corresponds to the first cluster of the dataset when it is split by date. It is clear that splitting by date reduces the clustering behavior so that trends are now much clearer. For completeness, the plots against cases for the remaining cluster are presented in Figure 5. Interestingly, there does not seem to be trends forming for this cluster.

When checking the correlation matrices within the first cluster, it appears that splitting into clusters has made the correlation between predictors worse. However, the trends are clearer when clusters are made, so we will keep this approach; a regularization method will become useful as a result. The coefficients of `time` in the correlation matrix are large, so it appears that public behavior changes over time, most likely in reaction to the changes in case counts over the months. This is still visible in the correlation matrix of the whole dataset with `time` included, as its coefficients are still relatively high. For this reason, we will analyze the dataset as a whole and in clusters by time.

Now, we will outline potential approaches to the analysis of this dataset. As mentioned above, we will analyze the entire dataset as one, and then repeat when the dataset is divided into clusters or with time as an encoded categorical variable instead. Firstly, univariate linear models of each predictor against case counts will be examined to gain an initial understanding of how each predictor affects case counts, and a full linear model will be presented to see how the predictors fit together. Secondly, a regularization approach will be employed due to the correlation of many predictors. Most likely, LASSO will be used in order to enable the ranking of the impact of predictors, as it matches the goal of the analysis. As interpretation is one of the goals of the analysis, we opt to avoid non-parametric approaches. However, repeating the first two approaches with various transformations of the predictors is a possibility; this enables us to accurately model the dataset without sacrificing too much interpretability. Finally, once the main questions have been answered by the above two methods, we will attempt to see if the clusters can be characterized by the predictors instead of by date, which is a classification problem. We will decide the range of approaches to employ for this last component of the analysis once we have settled whether we want to interpret the characterization of the clusters or not.

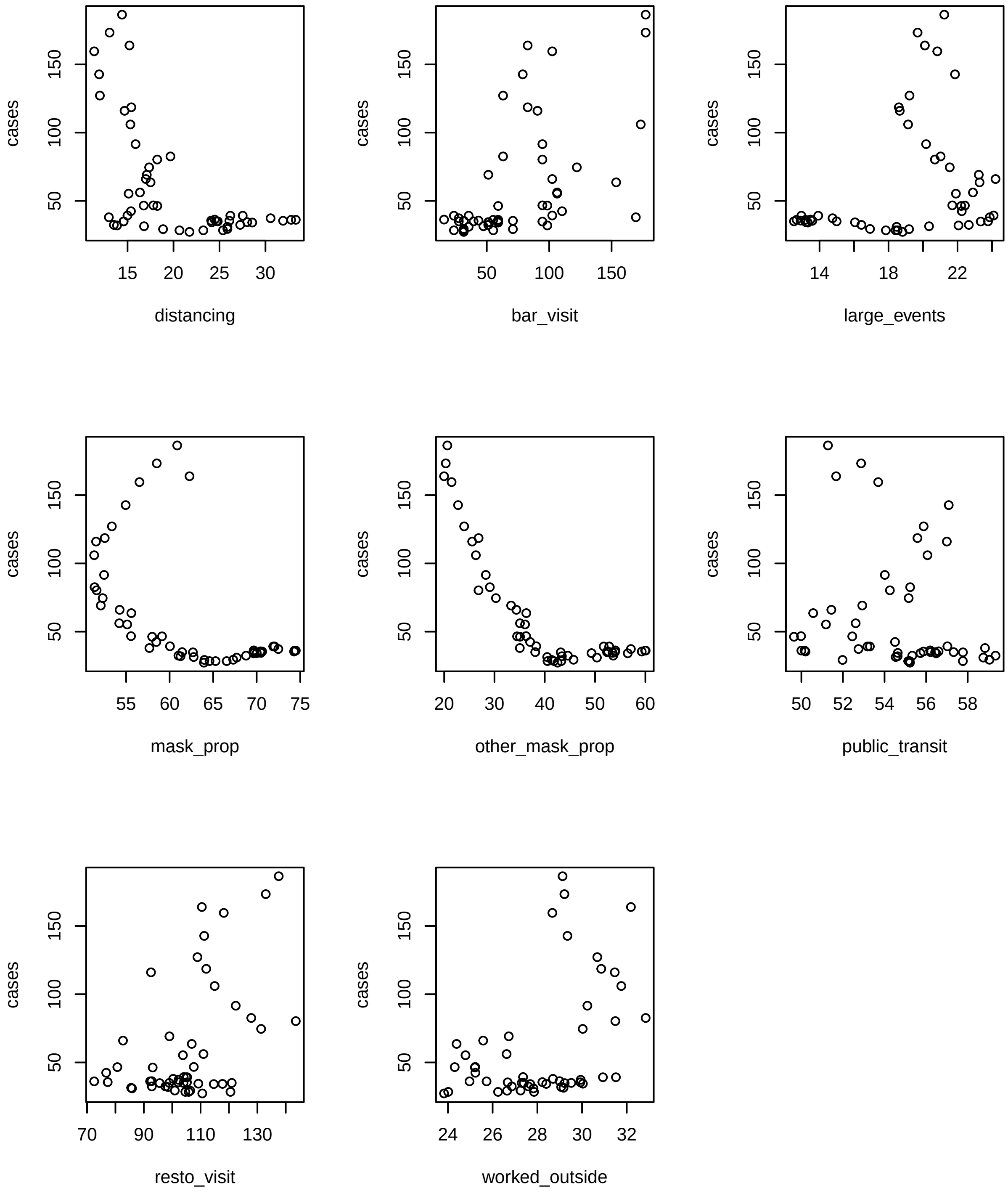

Figure 4: Scatterplots of predictors against case counts on and before July 25th, 2021.

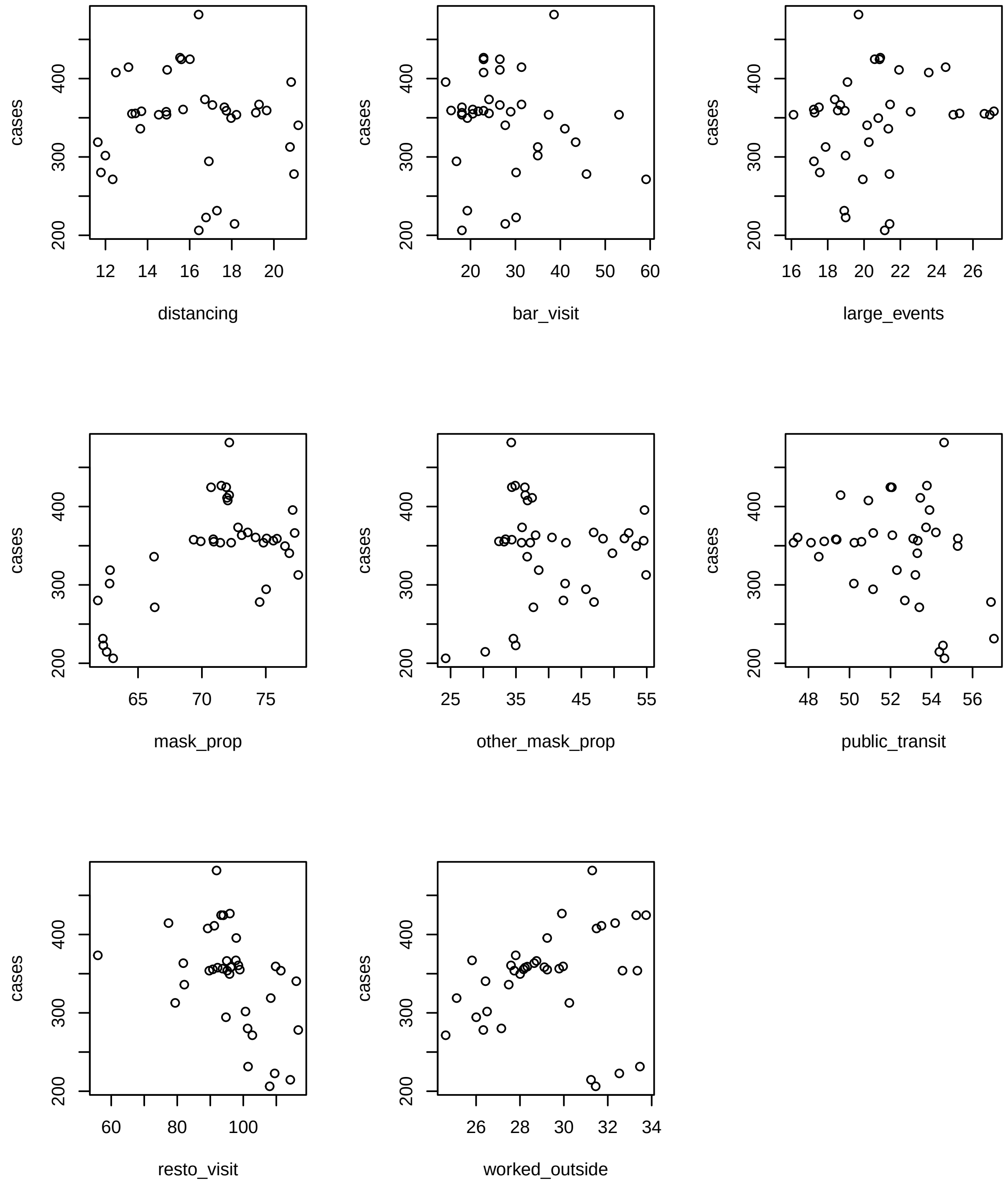

Figure 5: Scatterplots of predictors against case counts after July 25th, 2021.