STAT406 Project:
Predicting New Covid-19 Cases in U.S. with Multiple Variables

Justin Liang 56007123
Felix Ma 22972558
Yifan Wang 45753621
Jiale Wang 45464484

2021-11-09

Why this project is interesting to us?

Since the beginning of 2020, the Covid-19 virus has spread worldwide and quickly resulted in a pandemic. Until now, Covid-19 still threatens our lives and has dramatically changed the way in which we live, learn, and travel. Fortunately, through the concerted efforts of the world, we now have vaccines to fight the virus. It will be of great insight to study the effects of vaccinations and other relevant data over time and build a model to predict new cases in the United States. This project seems interesting as it may help us understand the effectiveness of the vaccines and predict the future cases.

Introduction

After online research, our group found a complete COVID-19 dataset maintained and updated by *Our world in Data*. This dataset contains daily COVID-19 data for countries around the world and is updated everyday. The dataset has 65 columns that provide important data including country, date, total confirmed cases, new cases, number of ICU patients, number of patients in hospital, COVID-19 test positive rate, number of vaccinations, etc. These variables can be very helpful for us in selecting appropriate predictors and build a model for predictions. Therefore, our eventual goal is to select the variables that best explain the number of new cases in the future and rely on the data in the dataset and fit an appropriate model that can predict future new COVID-19 cases in a given day.

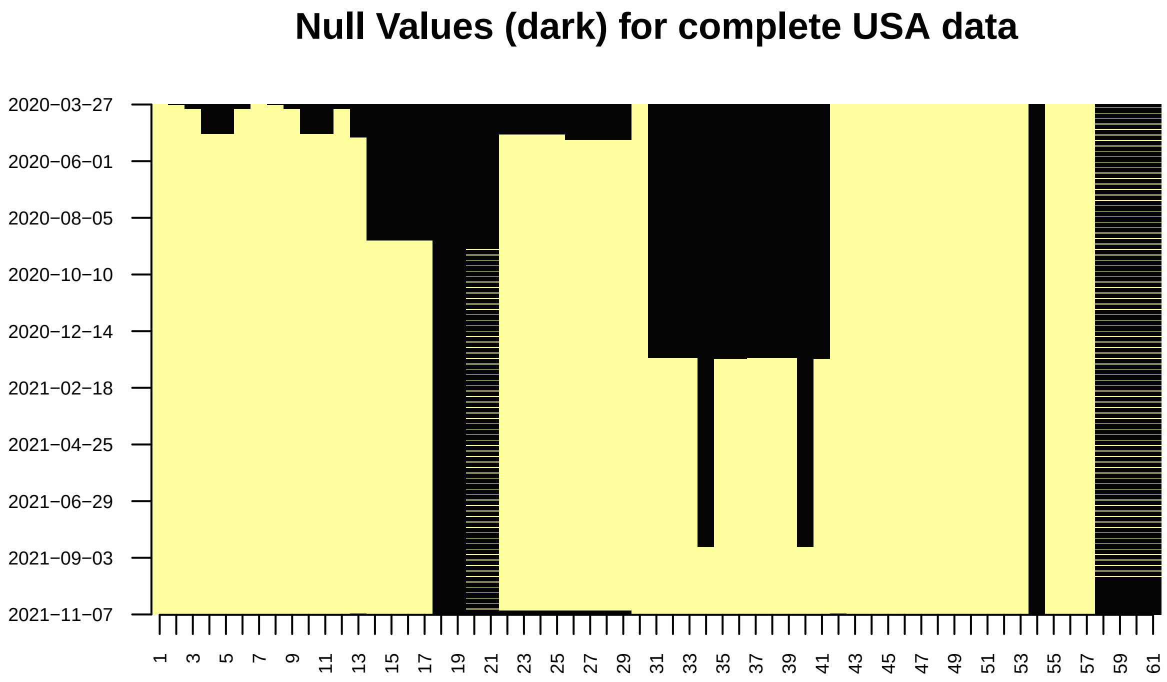
To complete our task, we are going to explore the data first. Daily new confirmed COVID-19 cases is the response variable since it reflects the cases in a given day and fluctuates everyday at the same time which is great for us to see how the factors influences the appearance of new cases. We also restricted the location of our data. Since the virus has been spreading widely in the U.S. over the past year and the U.S. is a pioneer in COVID-19 vaccines, we decided to focus on the U.S. data. As mentioned in the journal named *"Phase I/II study of COVID-19 RNA vaccine BNT162b1 in adults"* from *Nature* published in 2020 August 12th, phases 7 and 14 days before cases are confirmed can be very significant. Therefore, we have decided to use the variables 7 and 14 days before a given day as predictors.

After exploratory data analysis, we will separate our data into training and testing sets and investigate more about the data to find the model with the best prediction power. A discussion about the results and limitations will be shown at the end of the project. We hope to figure out an effective prediction model for future cases in the United States at the end.

Exploratory Data Analysis

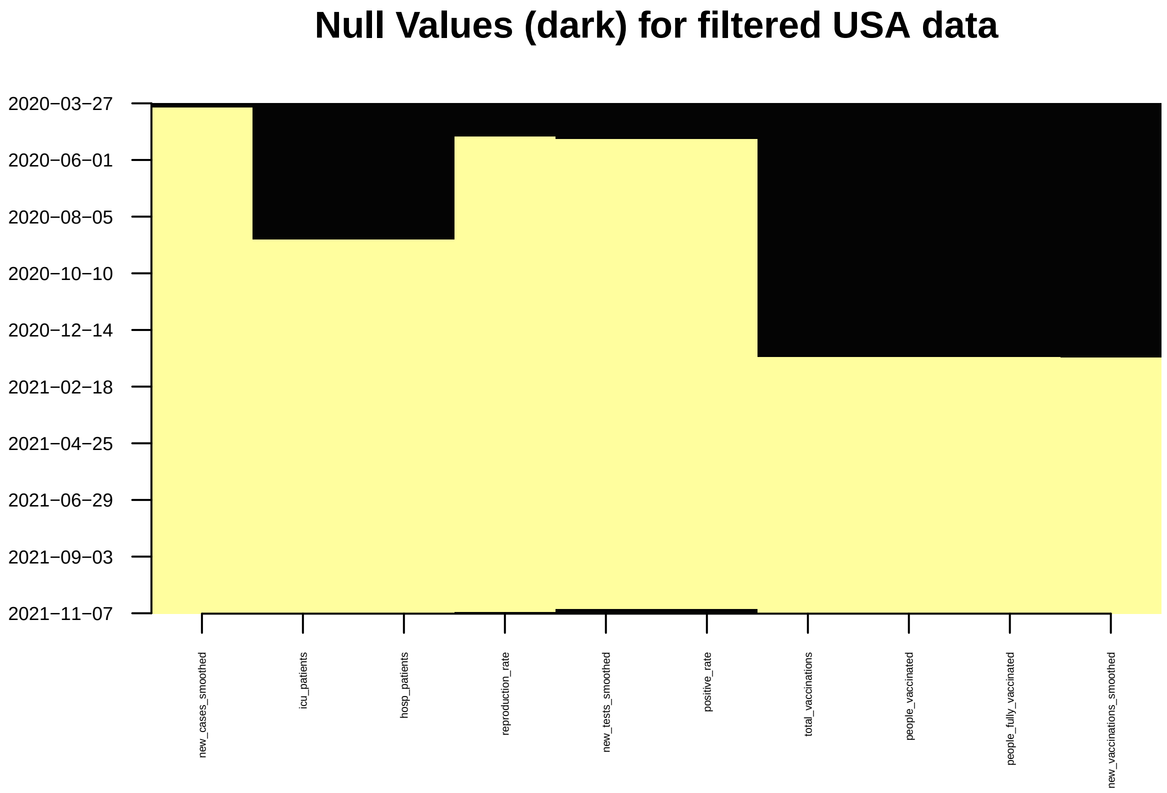
Data Inspection

First, we will inspect the distribution of null values in the data set of United States. Ideally, we would like to use contiguous blocks of data over time that do not contain any null values. We can see that there are many variables that are almost entirely null from the starting date to the end date, these variables will need to be removed. Many variables from index 31 to 41 are null roughly prior to 2021. These variables have names related to vaccinations such as "total_vaccinations" and "new_vaccinations". These predictors could be very useful for our model which means that we would probably need to exclude all data prior to 2021.



Index	Name	Index	Name
1	total_cases	32	people_vaccinated
2	new_cases	33	people_fully_vaccinated
3	new_cases_smoothed	34	total_boosters
4	total_deaths	35	new_vaccinations
5	new_deaths	36	new_vaccinations_smoothed
6	new_deaths_smoothed	37	total_vaccinations_per_hundred
7	total_cases_per_million	38	people_vaccinated_per_hundred
8	new_cases_per_million	39	people_fully_vaccinated_per_hundred
9	new_cases_smoothed_per_million	40	total_boosters_per_hundred
10	total_deaths_per_million	41	new_vaccinations_smoothed_per_million
11	new_deaths_per_million	42	stringency_index
12	new_deaths_smoothed_per_million	43	population
13	reproduction_rate	44	population_density
14	icu_patients	45	median_age
15	icu_patients_per_million	46	aged_65_older
16	hosp_patients	47	aged_70_older
17	hosp_patients_per_million	48	gdp_per_capita
18	weekly_icu_admissions	49	extreme_poverty
19	weekly_icu_admissions_per_million	50	cardiovasc_death_rate
20	weekly_hosp_admissions	51	diabetes_prevalence
21	weekly_hosp_admissions_per_million	52	female_smokers
22	new_tests	53	male_smokers
23	total_tests	54	handwashing_facilities
24	total_tests_per_thousand	55	hospital_beds_per_thousand
25	new_tests_per_thousand	56	life_expectancy
26	new_tests_smoothed	57	human_development_index
27	new_tests_smoothed_per_thousand	58	excess_mortality_cumulative_absolute
28	positive_rate	59	excess_mortality_cumulative
29	tests_per_case	60	excess_mortality
30	tests_units	61	excess_mortality_cumulative_per_million
31	total_vaccinations		

The dataset contains many repeated variables that describe the same kind of measurement but with a different aggregation or smoothing procedure. First, we removed all per_hundred or per_million variables from consideration because we are only analyzing United States. Next, we made a choice between a variable and its smoothed version. We have decided to use smoothed versions of each variable where possible because data is probably updated at irregular intervals. We do not want to model the irregularity in which the data is updated, just the trend. Because we are only analyzing cases for United States, we do not need any of the predictors that only vary among different countries. This includes all the variables from 43 to 57. We also believe that cases related to deaths should not be used because they are not meaningful predictors for new cases. The various vaccination variables carry slightly different meaning, we are not sure which ones to exclude so we have included all reasonable variables. After removing variables with excess null values, repeated variables, constant value variables and variables related to deaths, we end up with the following. We can use all the data from the beginning of 2021 to a few days before November 11 for analysis.

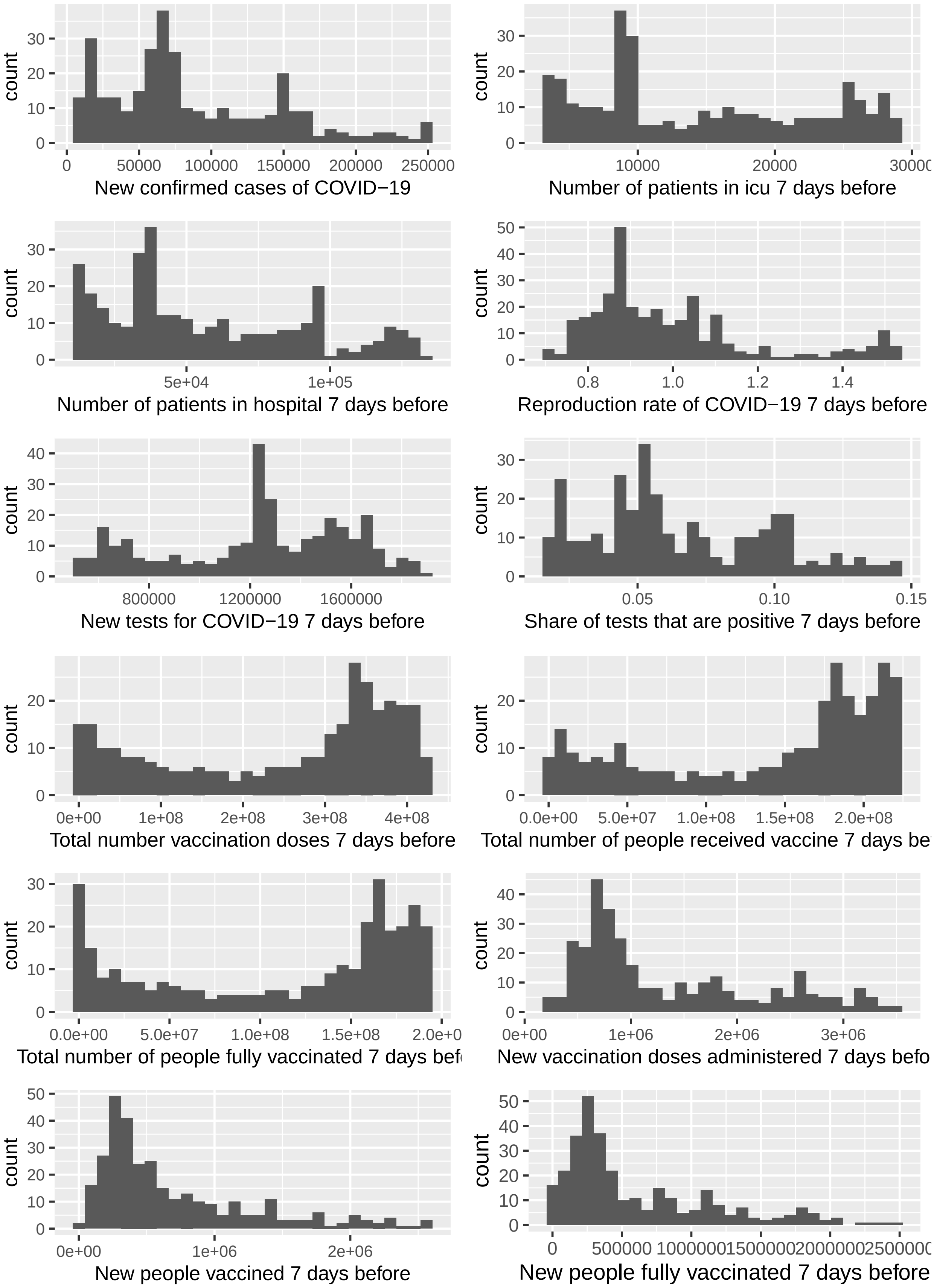


Because we would like to use variables 7 days prior and 14 days prior for prediction, all the predictors will need to be duplicated and aligned properly. Below is a list of all relevant variables after data preparation. A summary of their numerical properties are also listed.

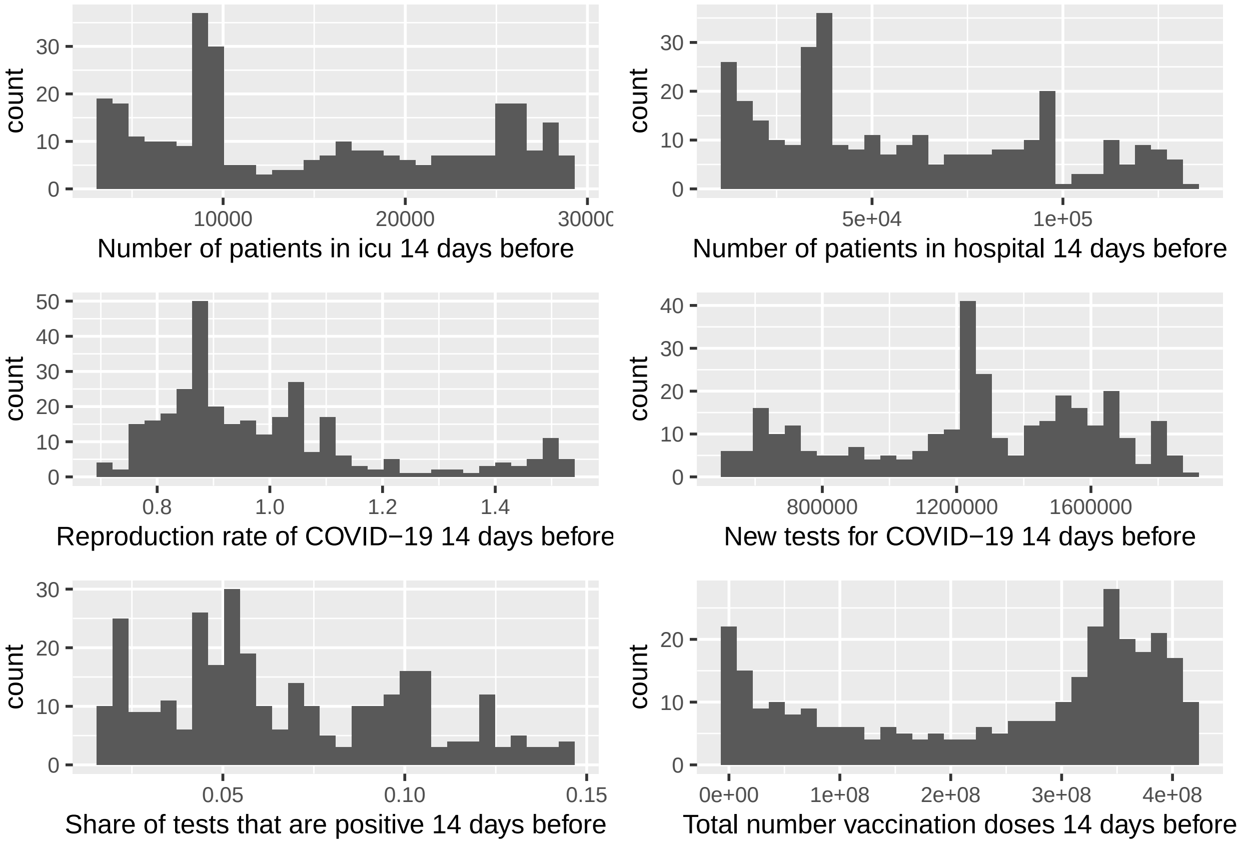
Table of Variable Names	
date	new_cases_smoothed
icu_patients_7	icu_patients_14
hosp_patients_7	hosp_patients_14
reproduction_rate_7	reproduction_rate_14
new_tests_smoothed_7	new_tests_smoothed_14
positive_rate_7	positive_rate_14
total_vaccinations_7	total_vaccinations_14
people_vaccinated_7	people_vaccinated_14
people_fully_vaccinated_7	people_fully_vaccinated_14
new_vaccinations_smoothed_7	new_vaccinations_smoothed_14
new_people_vaccinated_7	new_people_vaccinated_14
new_people_fully_vaccinated_7	new_people_fully_vaccinated_14

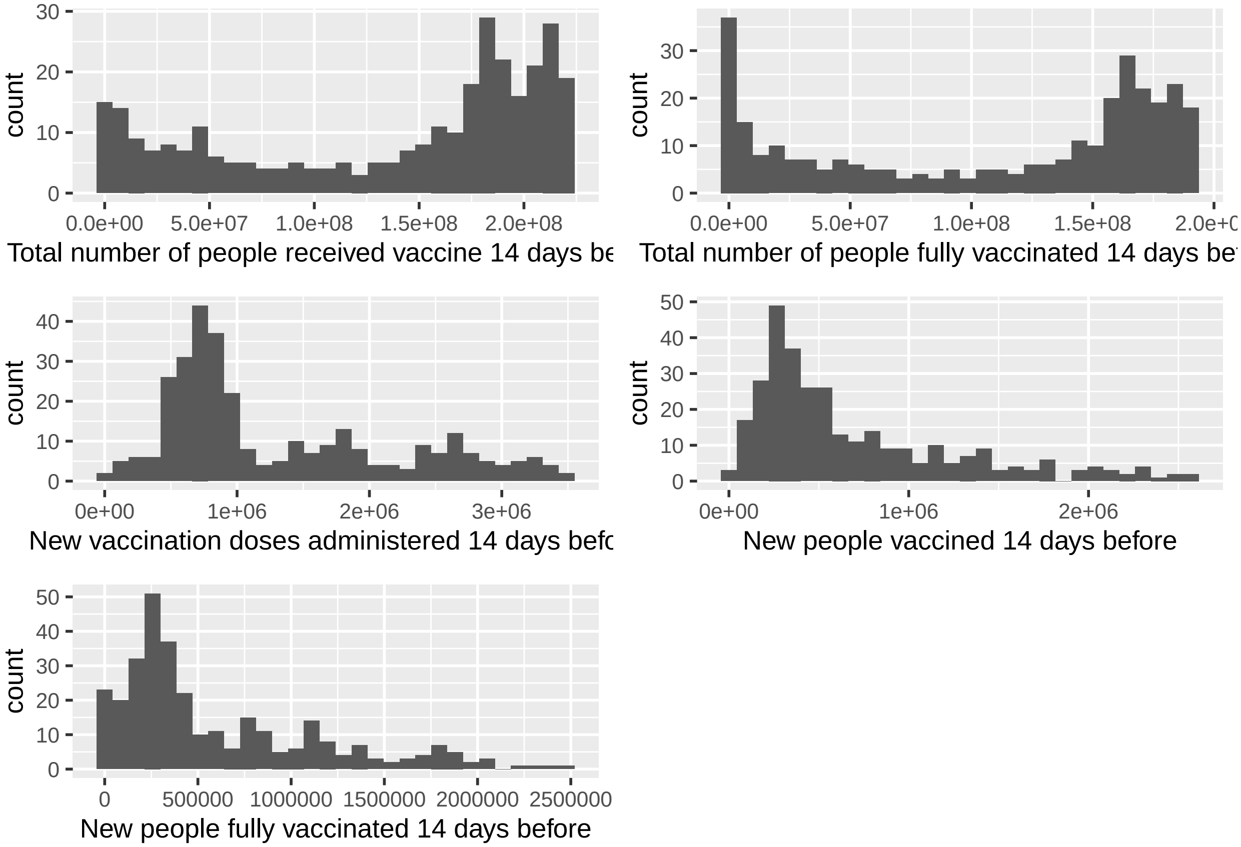
	Min	1st Quartile	Median	Mean	3rd Quartile	Max
new_cases_smoothed	11401	47510	70158	87440	129915	252023
icu_patients_7	3525	8340	12165	14566	22403	28891
hosp_patients_7	12233	32168	45008	56395	84351	133268
reproduction_rate_7	0.710	0.860	0.930	0.992	1.060	1.530
new_tests_smoothed_7	502168	1000966	1259475	1235419	1527400	1878241
positive_rate_7	0.018	0.043	0.056	0.066	0.094	0.145
total_vaccinations_7	1.58e+06	1.10e+08	3.08e+08	2.48e+08	3.62e+08	4.25e+08
people_vaccinated_7	1.54e+06	7.08e+07	1.73e+08	1.40e+08	2.00e+08	2.22e+08
people_fully_vaccinated_7	8.47e+03	3.85e+07	1.42e+08	1.11e+08	1.70e+08	1.92e+08
new_vaccinations_smoothed_7	222279	681781	934360	1338523	1874118	3495281
icu_patients_14	3525	8340	12698	14838	23268	28891
hosp_patients_14	12233	32168	47229	57889	88831	133268
reproduction_rate_14	0.710	0.860	0.930	0.994	1.060	1.530
new_tests_smoothed_14	502168	1000966	1261700	1246225	1544226	1878241
positive_rate_14	0.0180	0.0430	0.0580	0.0675	0.0970	0.1450
total_vaccinations_14	2.46e+04	9.31e+07	2.99e+08	2.38e+08	3.57e+08	4.17e+08
people_vaccinated_14	2.15e+04	5.96e+07	1.68e+08	1.35e+08	1.97e+08	2.20e+08
people_fully_vaccinated_14	3.74e+03	3.17e+07	1.37e+08	1.07e+08	1.68e+08	1.91e+08
new_vaccinations_smoothed_14	4292	665315	904830	1315946	1874118	3495281
new_people_vaccinated_7	12498	293124	479536	702126	941576	2574215
new_people_vaccinated_14	4163	282440	479536	699783	941576	2574215
new_people_fully_vaccinated_7	84	223868	370776	608485	869930	2478463
new_people_fully_vaccinated_14	84	223868	370776	604843	869930	2478463

Univariate Analysis



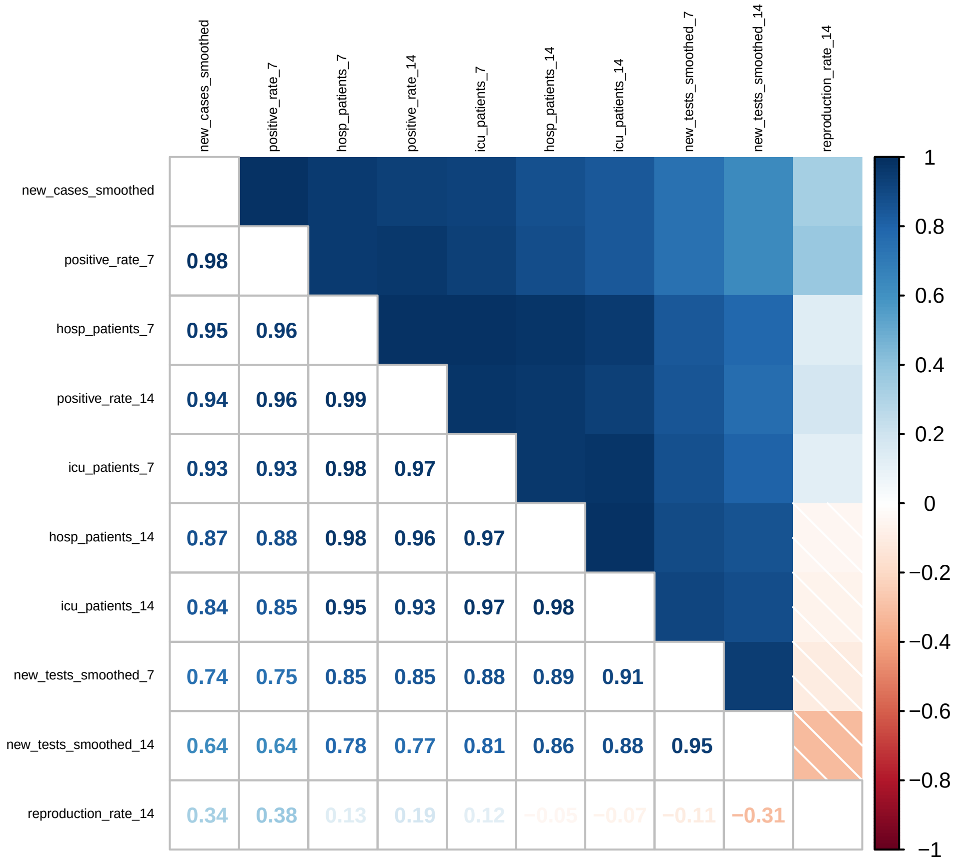
In general, all the univariate plots have rather irregular shapes, we probably cannot assume normality for any of the plots. From the univariate plots we can see that the new COVID-19 cases is quite skewed and not unimodal, there are multiple modes with significantly more count. The number of icu patients has a large mode at around 10000 but is rather uniform throughout. The number of patient in hospital is very irregular with multiple modes. The reproductive rate looks bimodal with a mode at 0.9 and another at 1.5, the data is skewed towards the right. The new tests for COVID-19 has three modes with a largest mode at around 1.2 million. The share of tests that are positive is very irregular. The total number of vaccination doses, total number of people vaccinated, total number of people fully vaccinated graphs all exhibit the same shape with the modes appearing at the tails. New vaccination doses, new people vaccinated and new people full vaccinated are somewhat normally distributed with very heavy right tails.



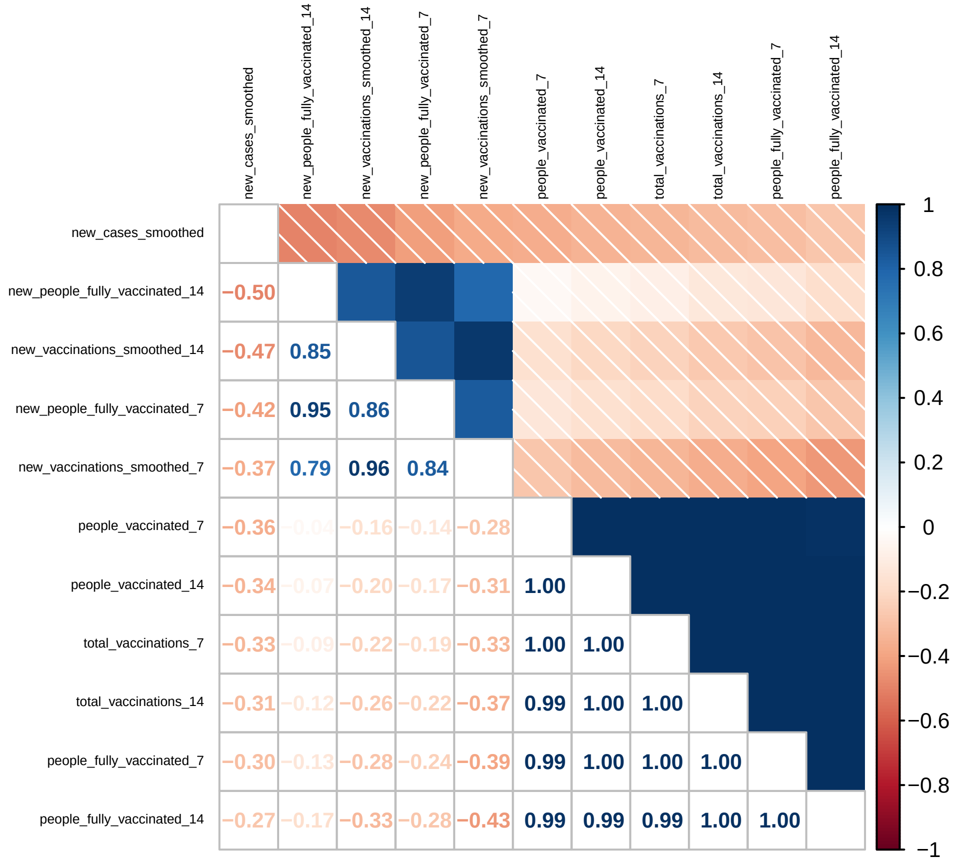

30
20
10
0
count
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
Total number of people received vaccine 14 days be
30
20
10
0
count
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+0
Total number of people fully vaccinated 14 days be
40
30
20
10
0
count
0e+00
1e+06
2e+06
3e+06
New vaccination doses administered 14 days befo
50
40
30
20
10
0
count
0e+00
1e+06
2e+06
New people vaccined 14 days before
50
40
30
20
10
0
count
0
500000
1000000
1500000
2000000
2500000
New people fully vaccinated 14 days before

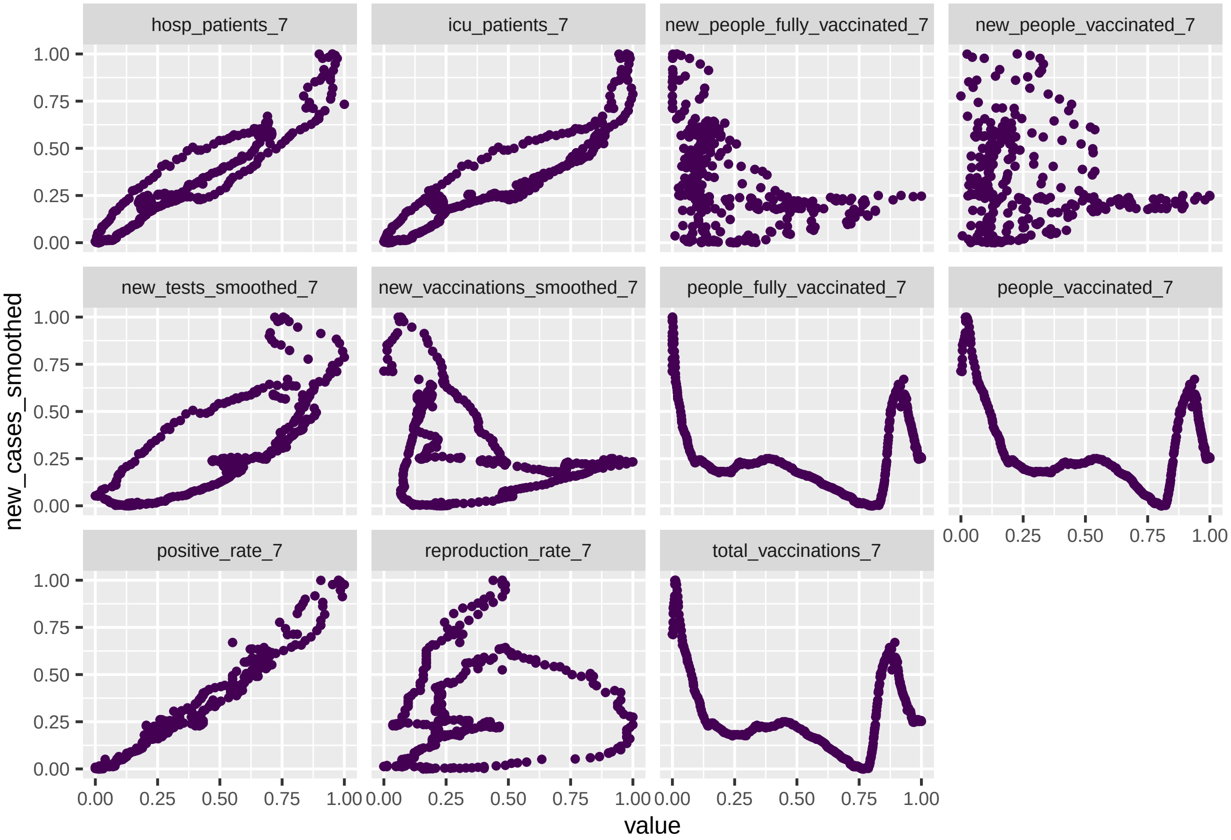
Bivariate Analysis

A correlation Matrix of the variables most positively correlated with New Cases Smoothed and a correlation matrix of the variables most negatively correlated with New Cases Smoothed have been produced. From the most positively correlated variables, positive_rate, icu_patients and hosp_patients are the variables most correlated with new cases. For all the predictors, the 7 day prior predictor has a higher correlation than the 14 day prior predictor. This pattern however is not true for the most negatively correlated predictors. New people fully vaccinated 14 days prior is more negatively correlated than 7 days prior. There is perfect positive correlation between the variables people_vaccinated, total_vaccinations and people_fully_vaccinated. Therefore it will probably be a good idea to remove all except one of the variables due to high colinearity.

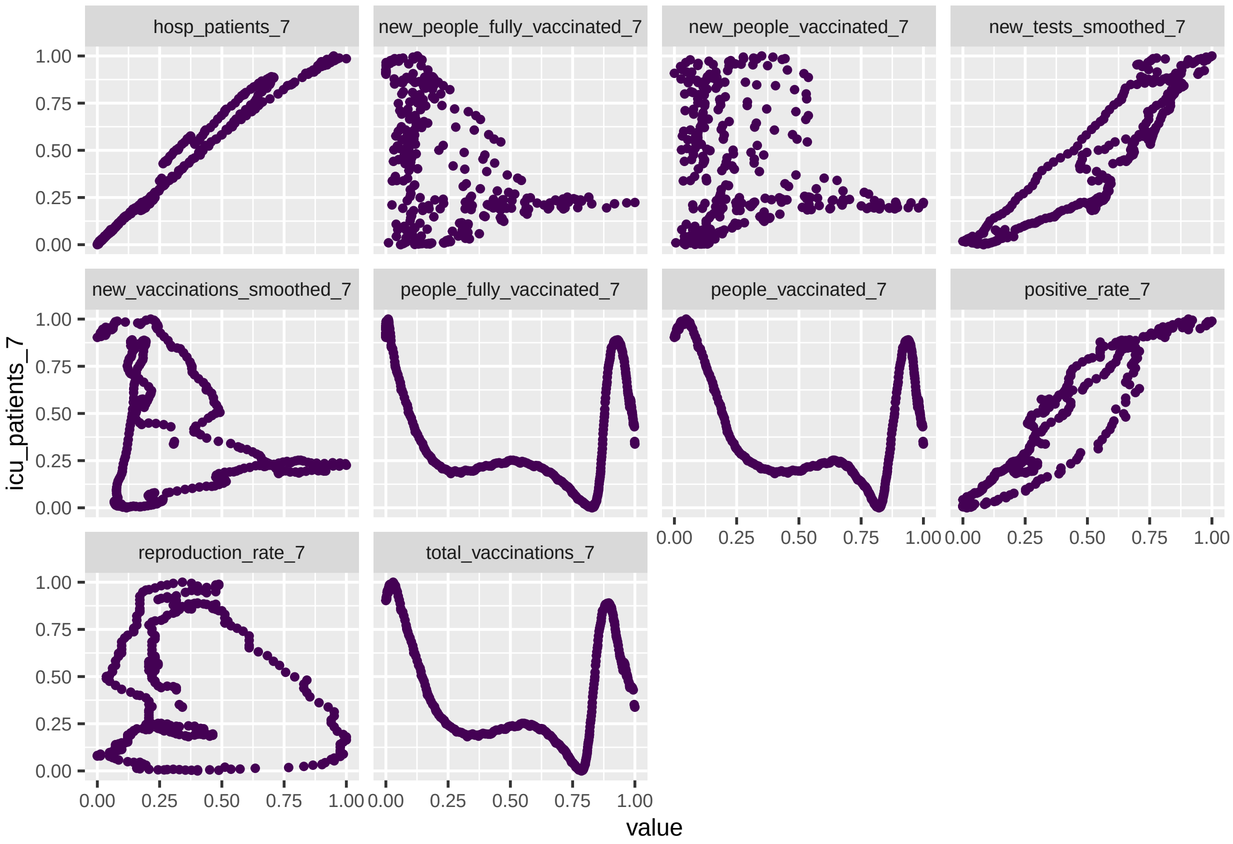


	new_cases_smoothed	positive_rate_7	hosp_patients_7	positive_rate_14	icu_patients_7	hosp_patients_14	icu_patients_14	new_tests_smoothed_7	new_tests_smoothed_14	reproduction_rate_14
new_cases_smoothed										
positive_rate_7	0.98									
hosp_patients_7	0.95	0.96								
positive_rate_14	0.94	0.96	0.99							
icu_patients_7	0.93	0.93	0.98	0.97						
hosp_patients_14	0.87	0.88	0.98	0.96	0.97					
icu_patients_14	0.84	0.85	0.95	0.93	0.97	0.98				
new_tests_smoothed_7	0.74	0.75	0.85	0.85	0.88	0.89	0.91			
new_tests_smoothed_14	0.64	0.64	0.78	0.77	0.81	0.86	0.88	0.95		
reproduction_rate_14	0.34	0.38	0.13	0.19	0.12	−0.05	−0.07	−0.11	−0.31	

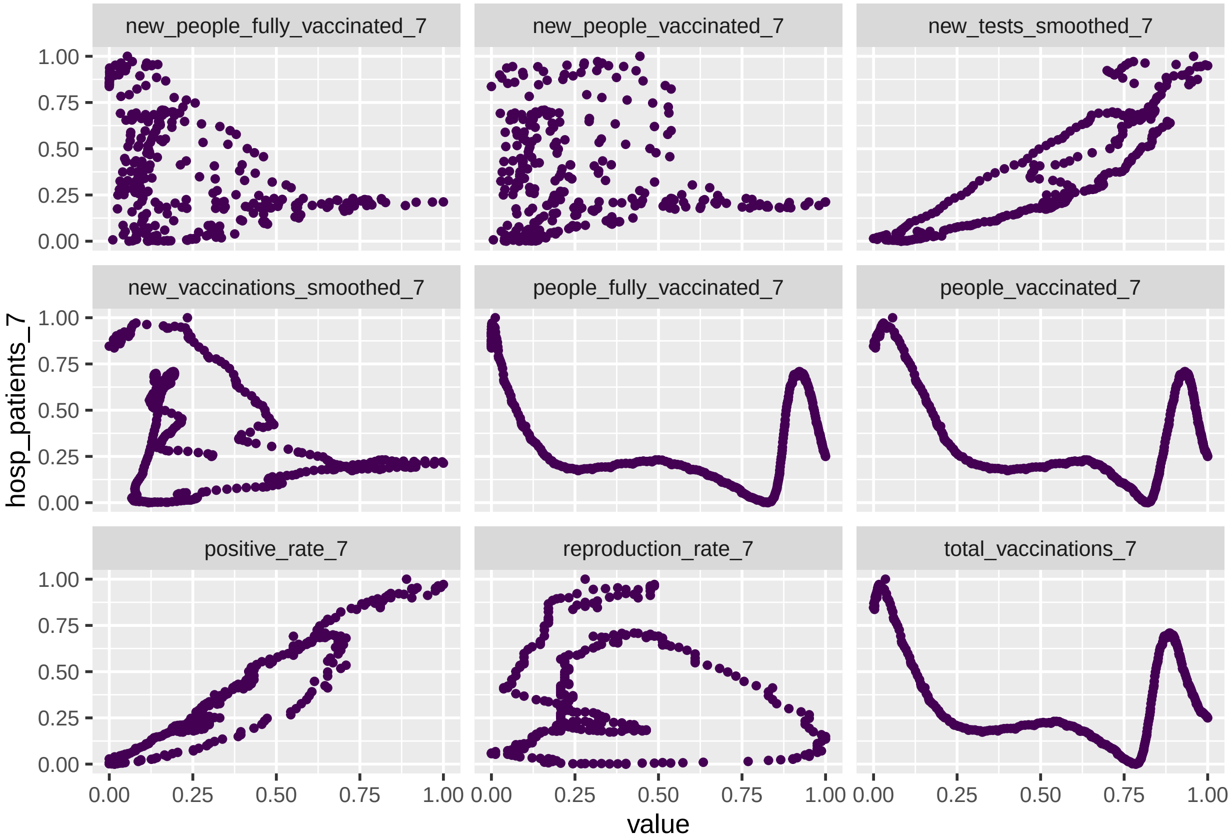

new_cases_smoothed
new_people_fully_vaccinated_14
new_vaccinations_smoothed_14
new_people_fully_vaccinated_7
new_vaccinations_smoothed_7
people_vaccinated_7
people_vaccinated_14
total_vaccinations_7
total_vaccinations_14
people_fully_vaccinated_7
people_fully_vaccinated_14
−0.50
−0.47 0.85
−0.42 0.95 0.86
−0.37 0.79 0.96 0.84
−0.36 −0.04 −0.16 −0.14 −0.28
−0.34 −0.07 −0.20 −0.17 −0.31 1.00
−0.33 −0.09 −0.22 −0.19 −0.33 1.00 1.00
−0.31 −0.12 −0.26 −0.22 −0.37 0.99 1.00 1.00
−0.30 −0.13 −0.28 −0.24 −0.39 0.99 1.00 1.00 1.00
−0.27 −0.17 −0.33 −0.28 −0.43 0.99 0.99 0.99 1.00 1.00
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



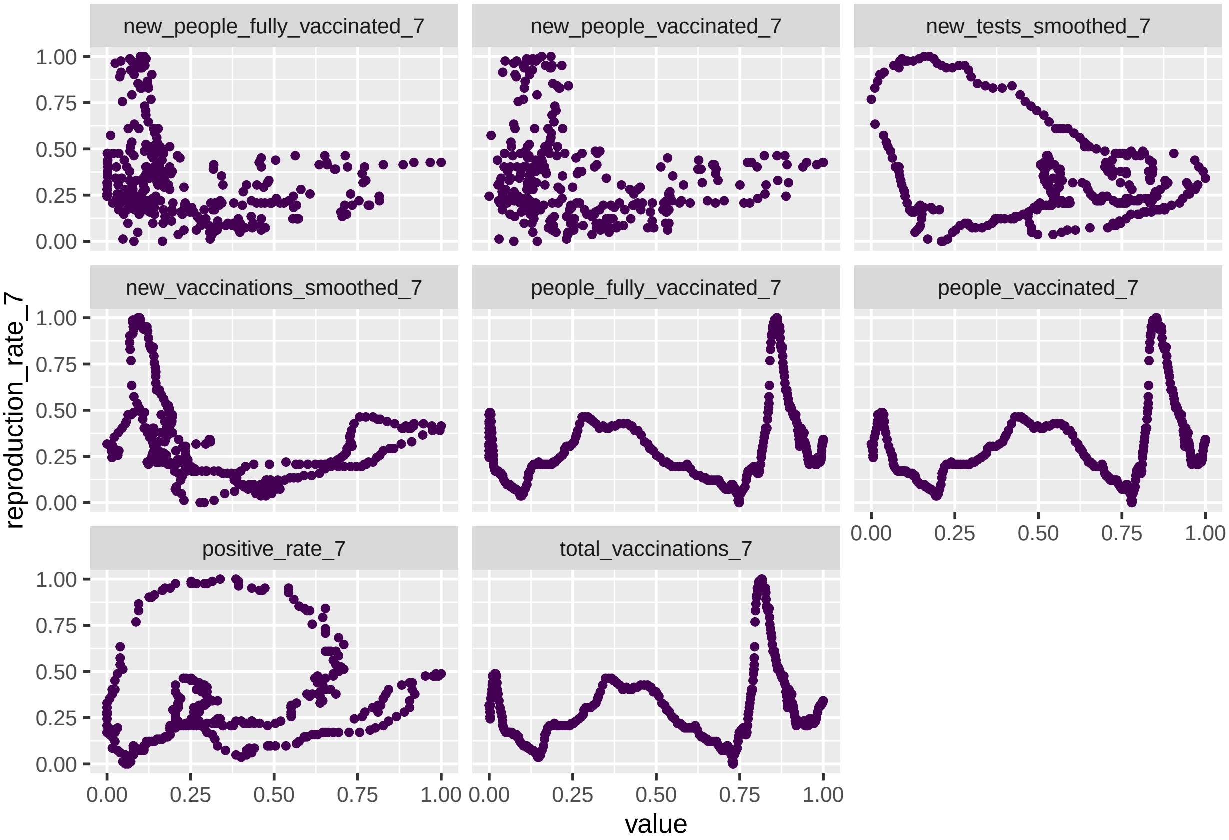
As we can see from the plot, the number of new cases confirmed is positively correlated (highly correlated) with the patients in hospital and number of patients in ICU. Also, the number of new cases confirmed is negatively correlated with number of people vaccinated(fully vaccinated). As the number of people fully vaccinated increases, there is a slight increase int the number of new cases confirmed, but there is a negative correlation between number of new cases confirmed and the number of people fully vaccinated. The total number of vaccinations shows similar trend with people fully vaccinated and people vaccinated. Reproduction rate appears to not have a strong relationship with new cases smoothed. It is interesting to note that as new people vaccinated and fully vaccinated increases, new cases appears to be approaching a value.



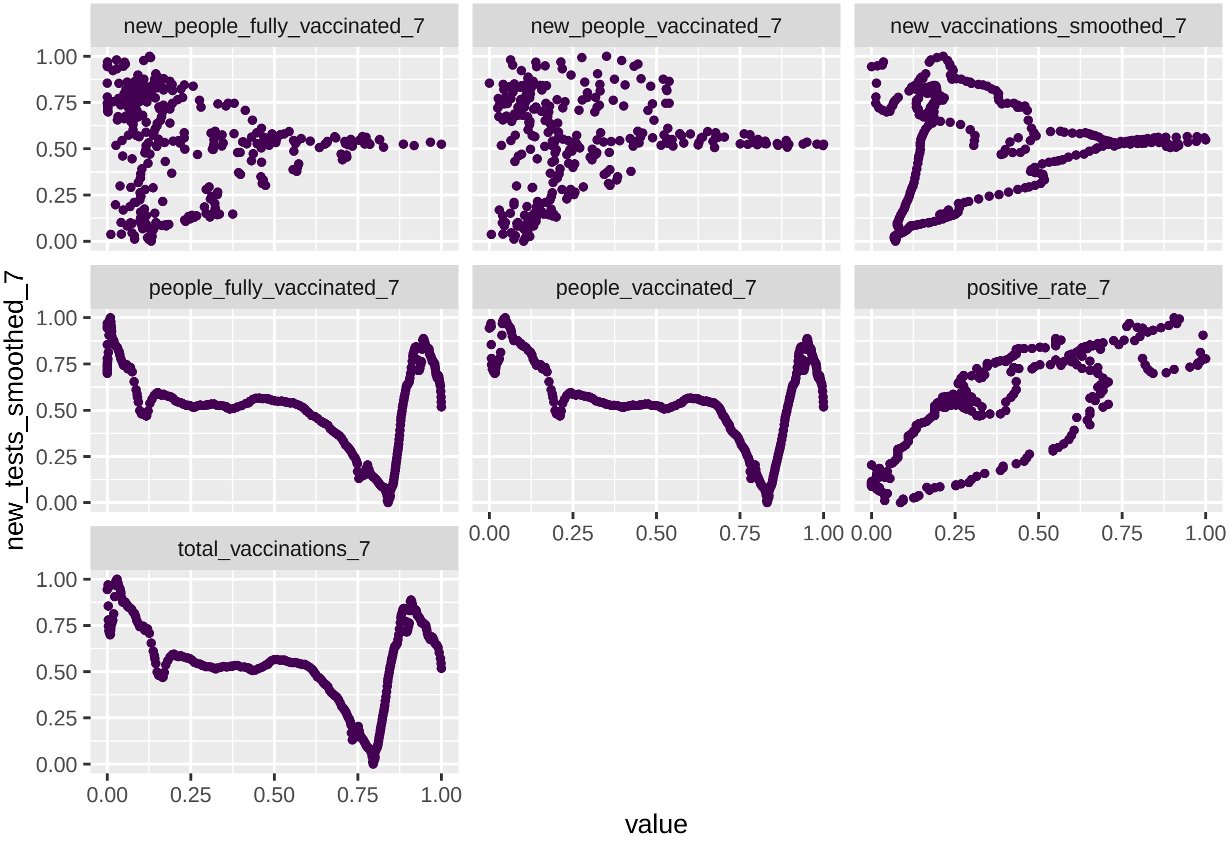
In this plot, the number of icu patients is positively correlated with patients in hospital and the positive rate, also number of icu patients is negatively correlated with number of people get vaccined(fully vaccined).



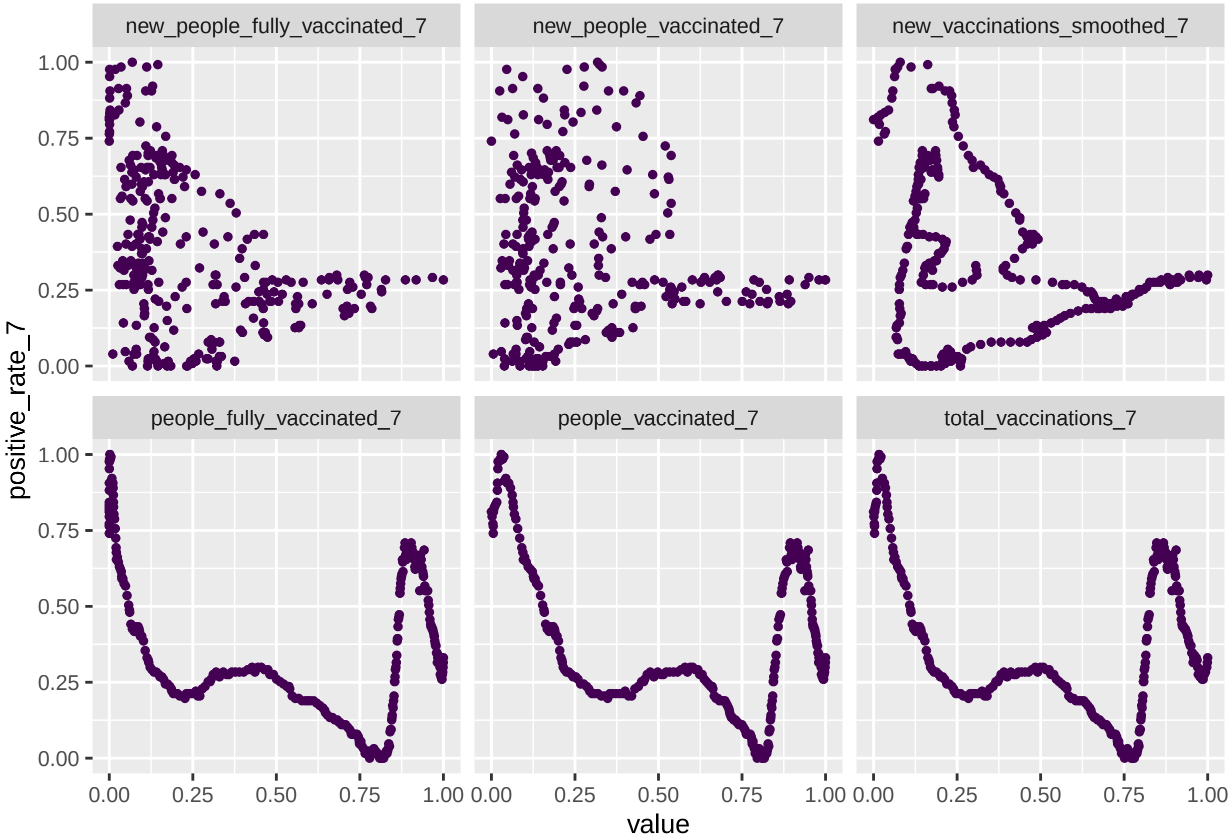
In the plot, the number of patients in hospital is positively correlated with positive rate, also, we can see that there is a slight increase in number of patients in hospital as more people get vaccined, however, the number of patients in hospital is still negatively correlated with number of people get vaccined(fully vaccined).



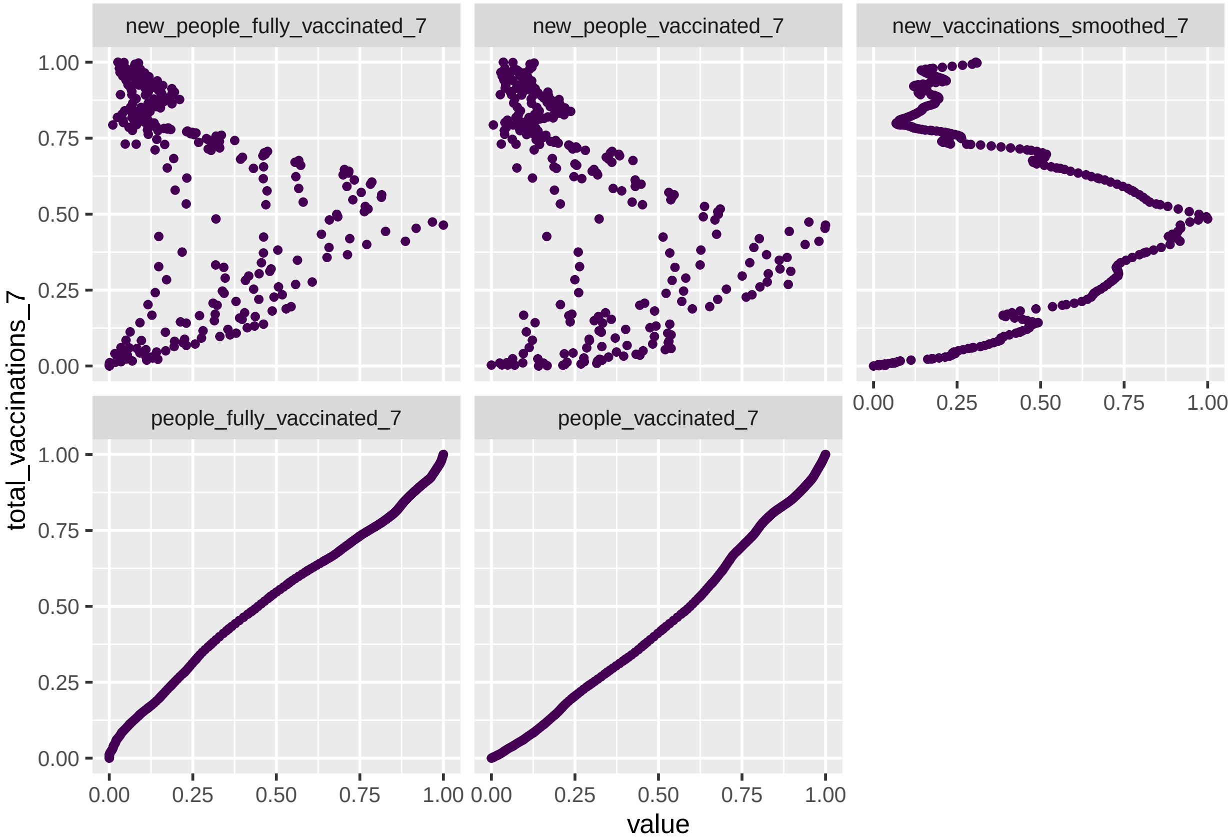
It is obvious that the correlation between reproduction rate and people fully vaccinated have a similar trend with that between reproduction rate and people vaccinated since the first two graphs are similar. There are also similar patterns between reproduction rate and number of people vaccinated and number of total vaccinations.



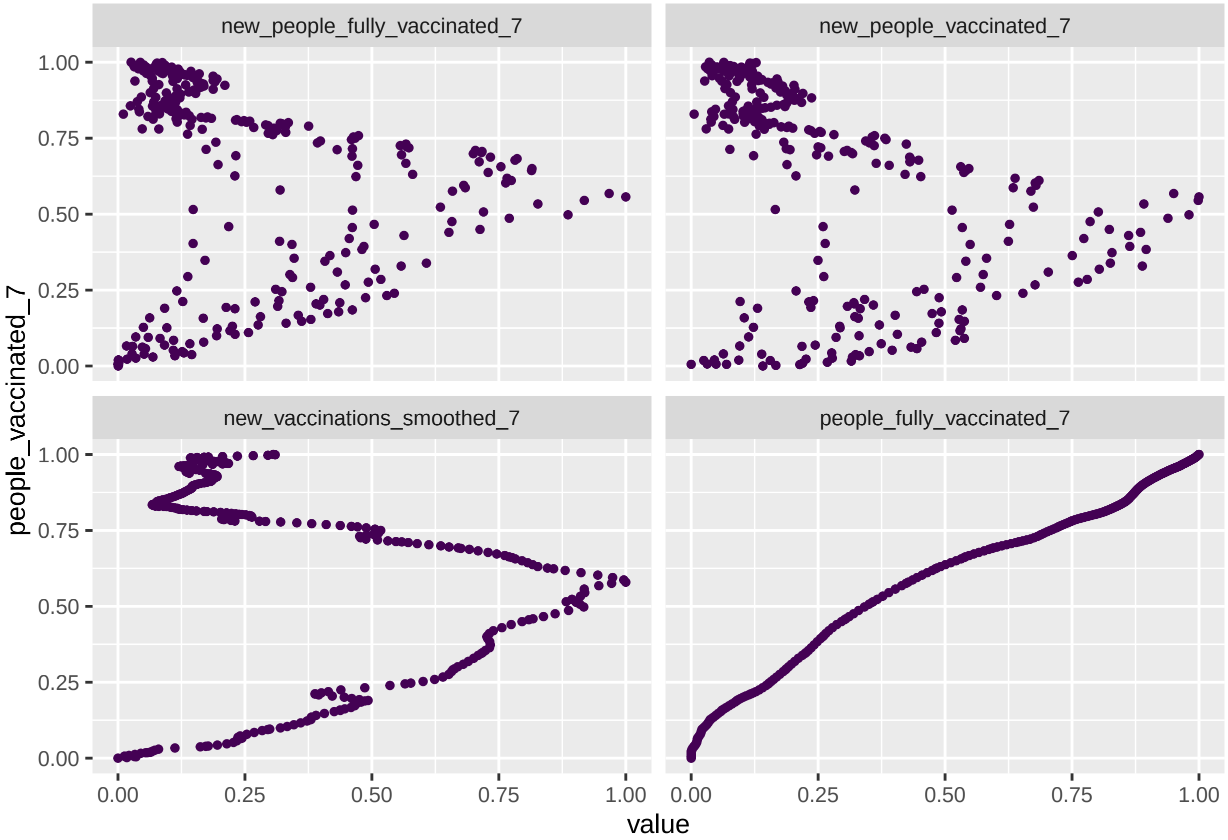
New tests exhibits similar patterns with the variables: new people vaccinated, people vaccinated and total vaccinations.



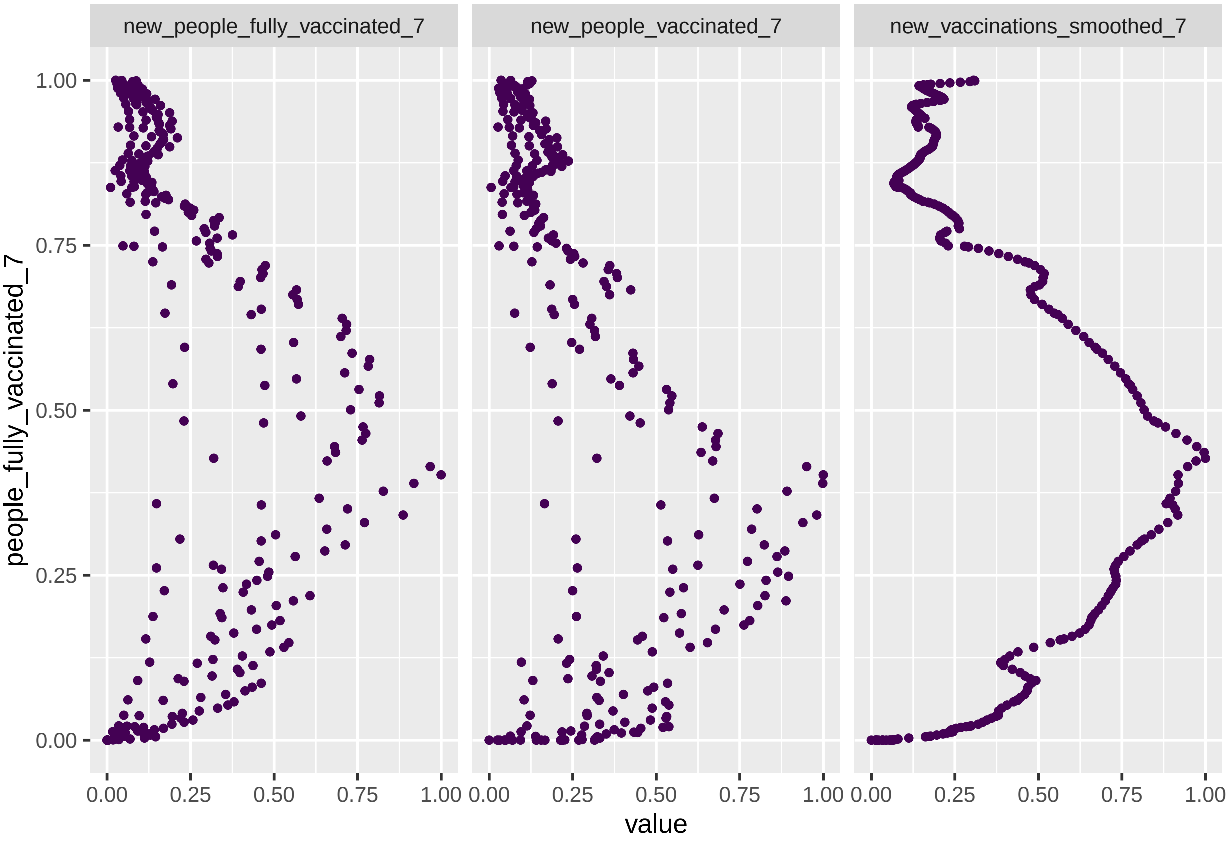
Positive rate exhibits similar patterns with people fully vaccinated, people vaccinated and total vaccinated. It is also interesting that positive rate appears to approach a certain limit as new people vaccinated and fully vaccinated increases.



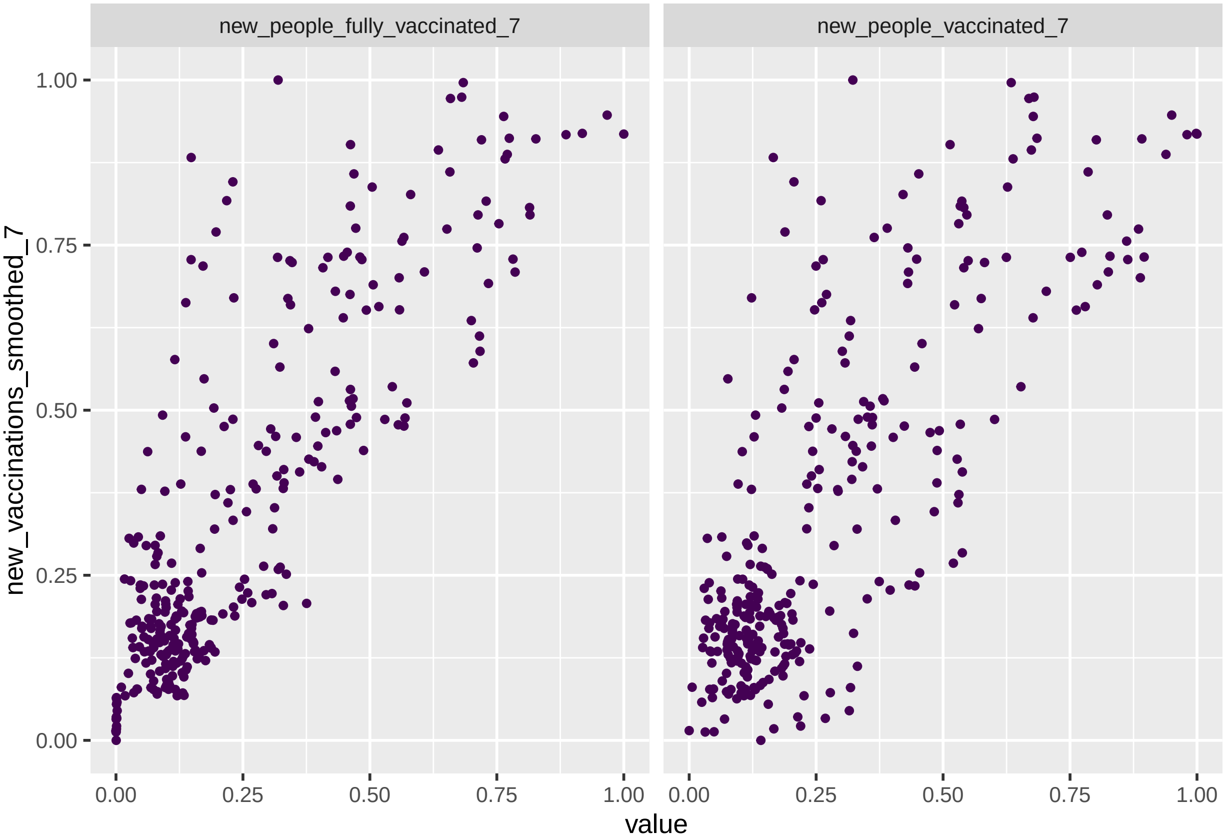
In this plot, we found that the correlation between the factor total vaccinations and the factors people fully vaccinated and people vaccinated seem very close to 1.



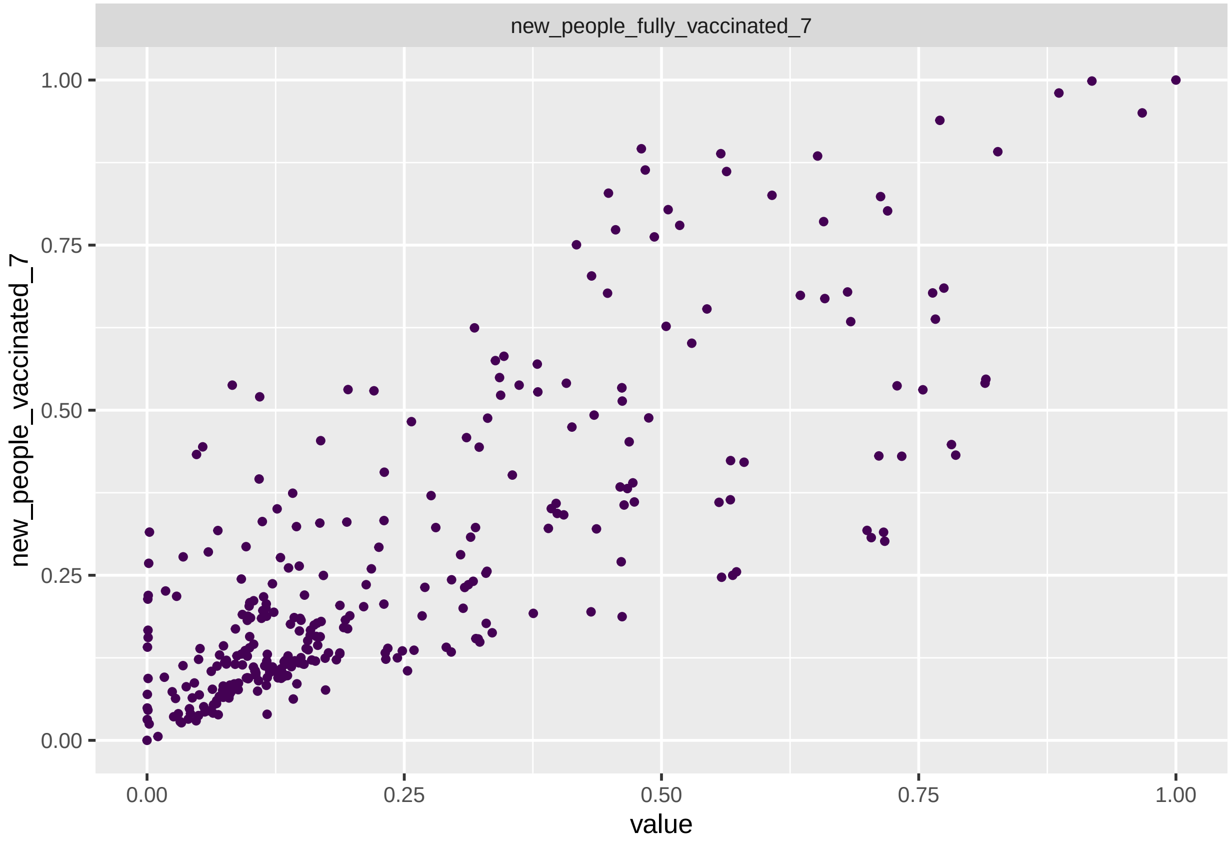
In the plot, people vaccinated and people fully vaccinated are highly correlated. It is also interesting that the new vaccinated graphs first increases then decrease as people vaccinated increases.



All the plots exhibit similar shape, similar to the plots between people vaccinated and the new vaccinated above.

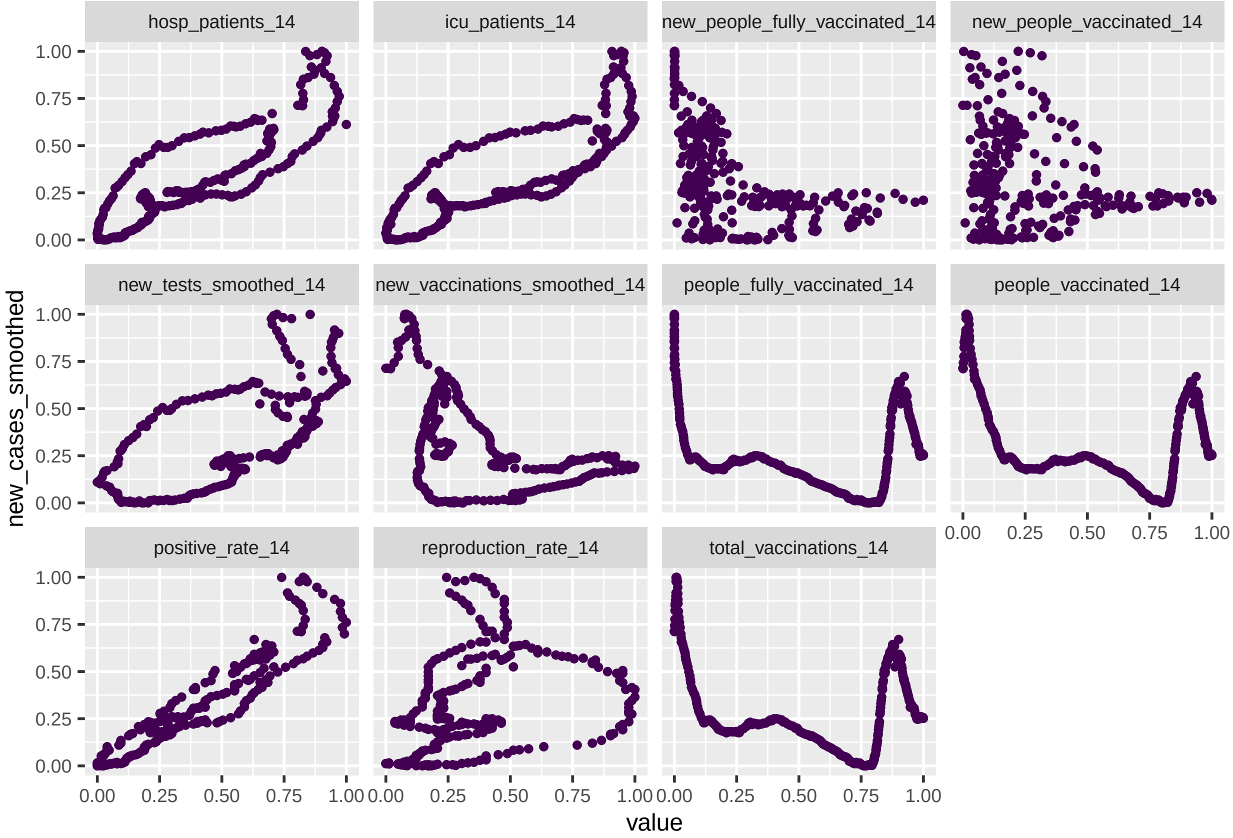


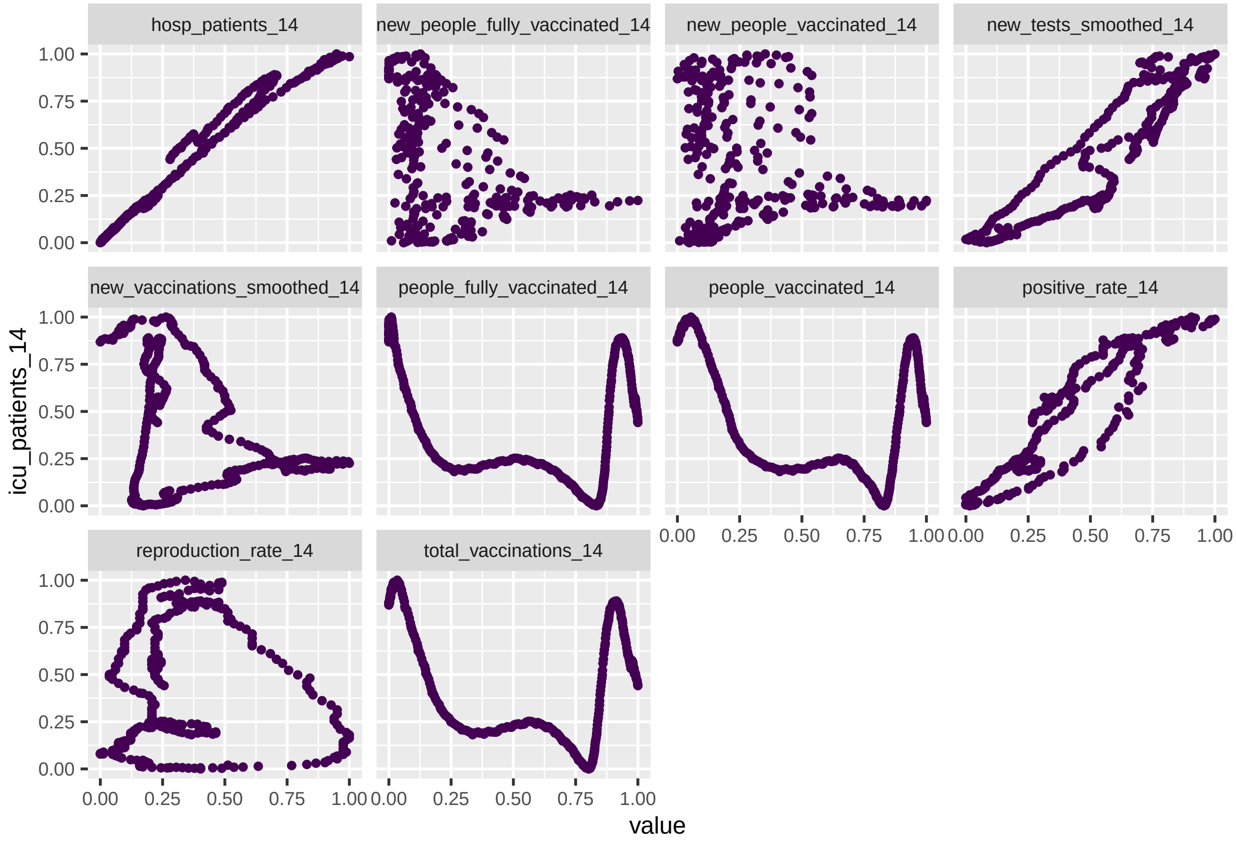
The new COVID-19 vaccinations are positively correlated with increased number of people vaccinated and fully vaccinated.

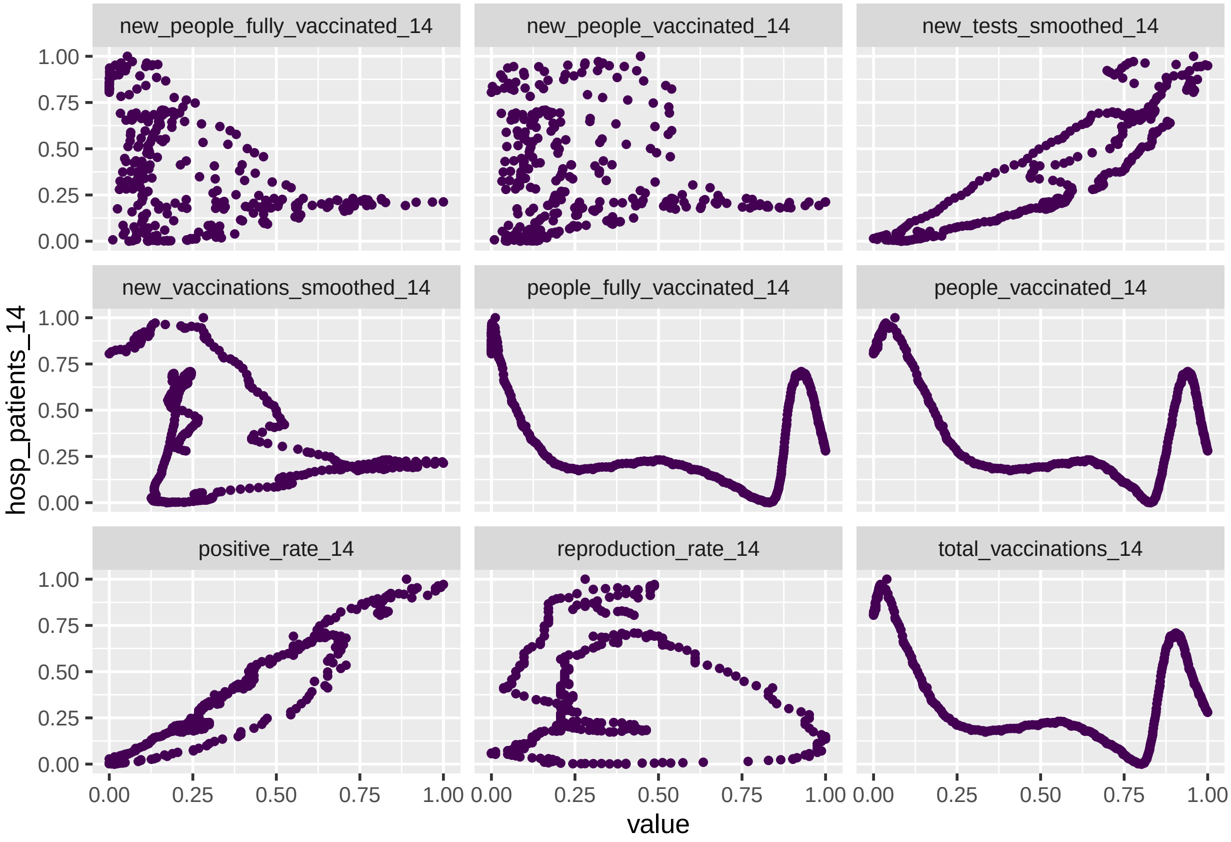


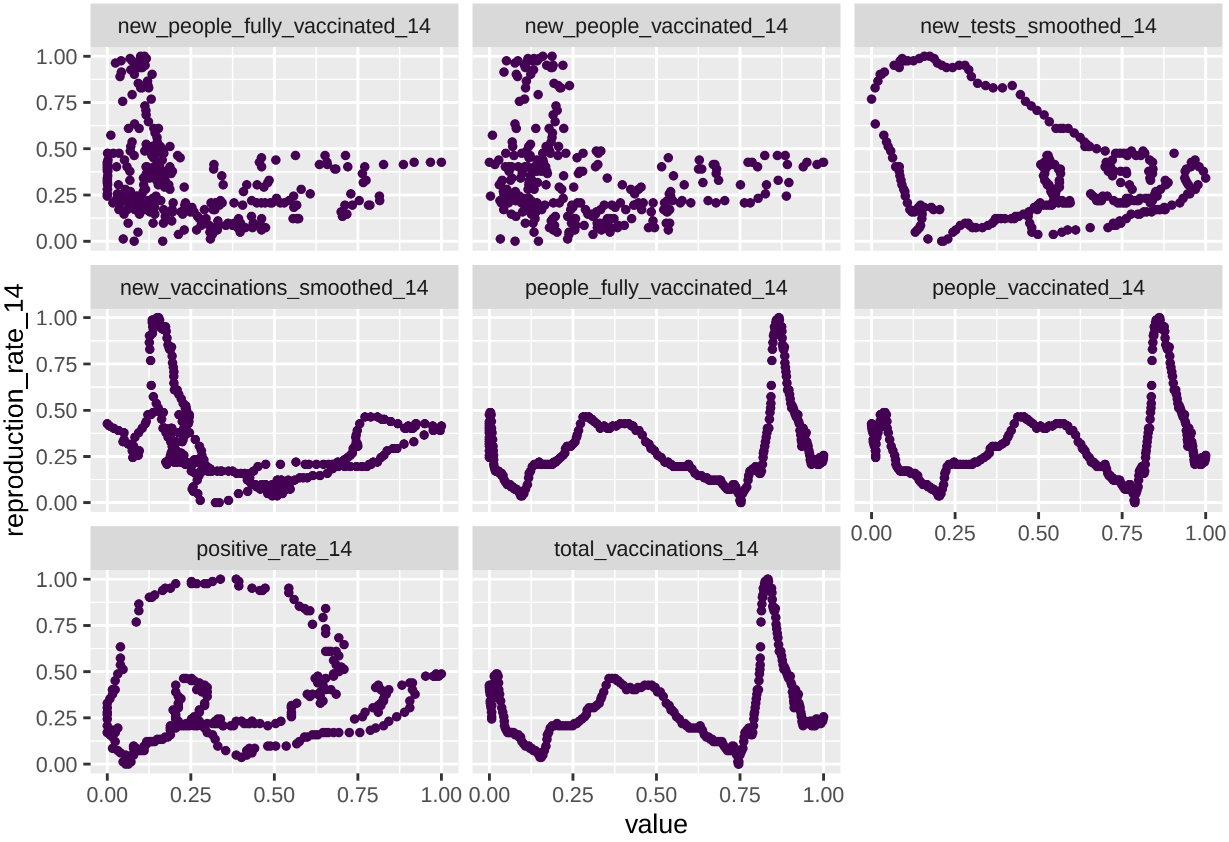
It can be seen that the number of new people vaccinated has a positive relationship with the number of new people fully vaccinated. With increasing number of people vaccinated, there will be more people who are fully vaccinated.

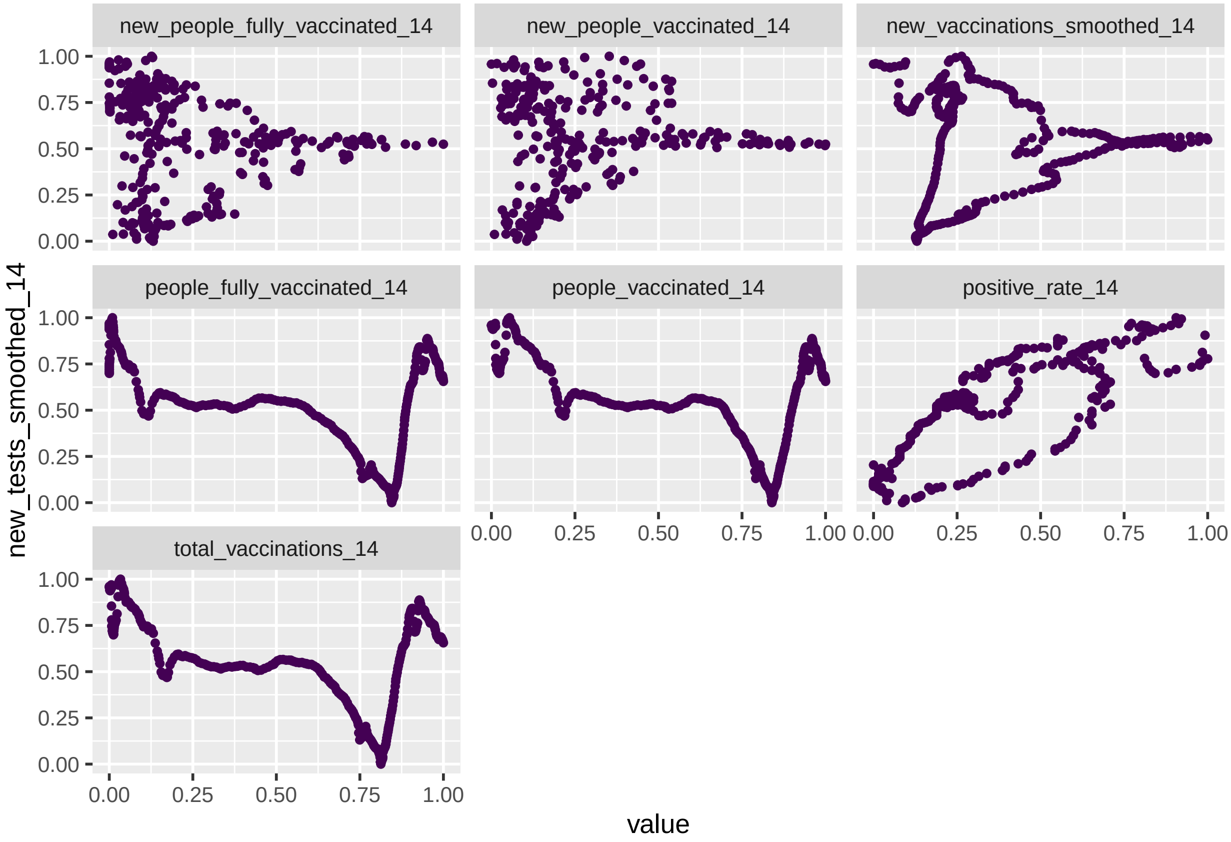
Remarks of the plots below have been omitted because the patterns are similar to the plots above.

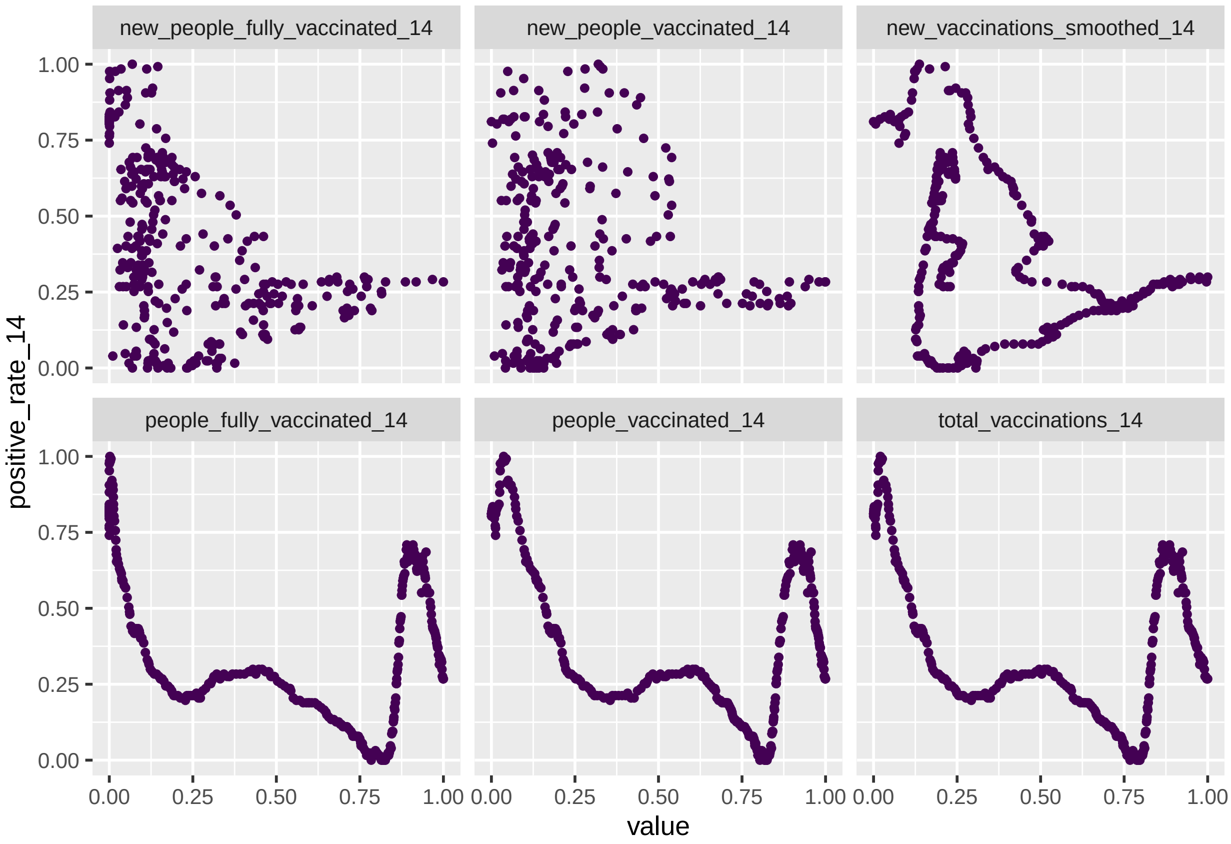


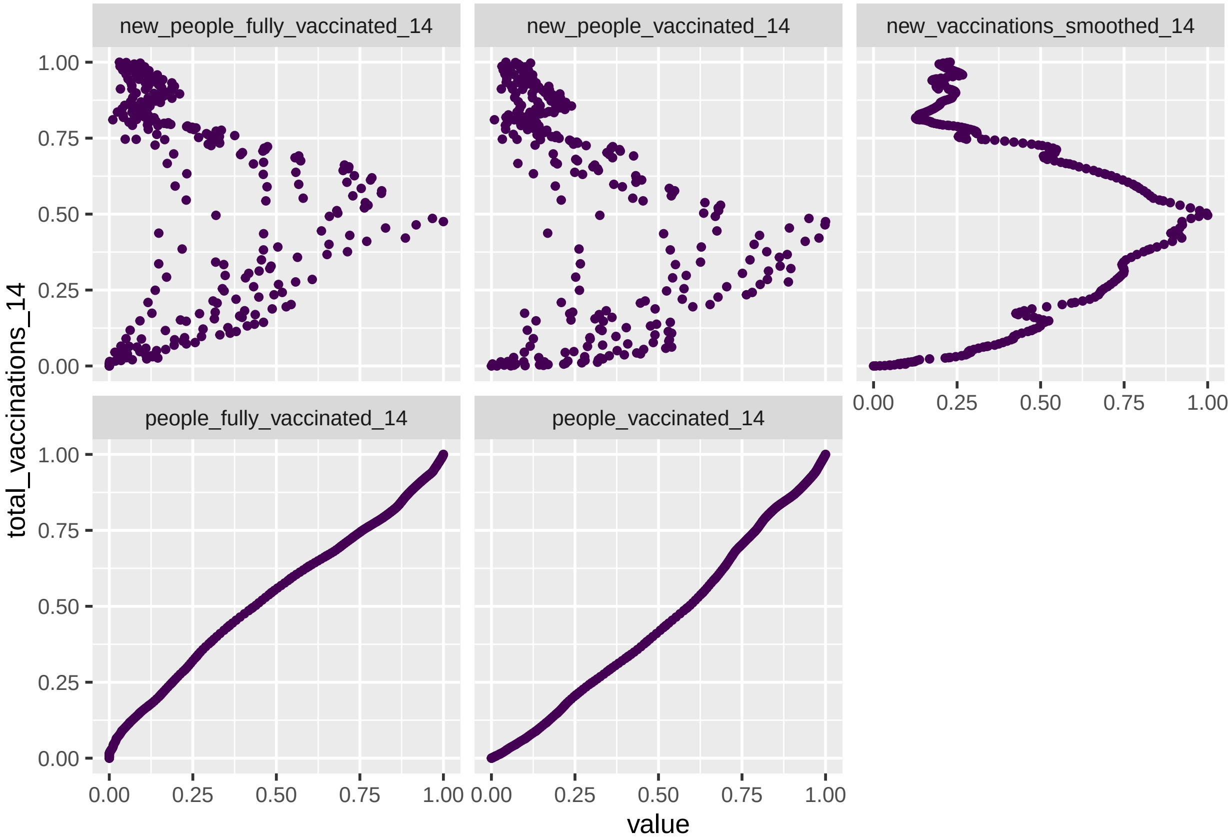

hosp_patients_14
new_people_fully_vaccinated_14
new_people_vaccinated_14
new_tests_smoothed_14
new_vaccinations_smoothed_14
people_fully_vaccinated_14
people_vaccinated_14
positive_rate_14
reproduction_rate_14
total_vaccinations_14
icu_patients_14
value
0.00
0.25
0.50
0.75
1.00

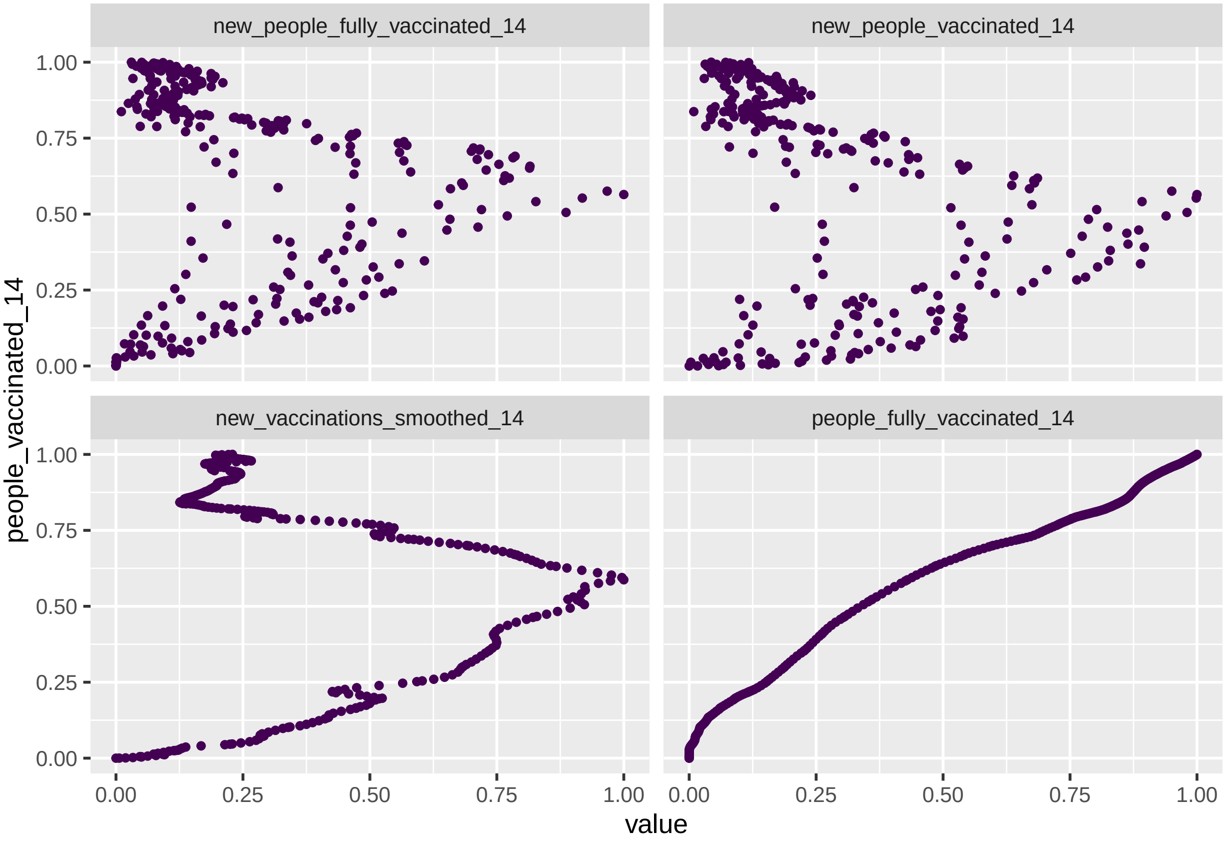

new_people_fully_vaccinated_14
new_people_vaccinated_14
new_tests_smoothed_14
new_vaccinations_smoothed_14
people_fully_vaccinated_14
people_vaccinated_14
positive_rate_14
reproduction_rate_14
total_vaccinations_14
hosp_patients_14
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
value

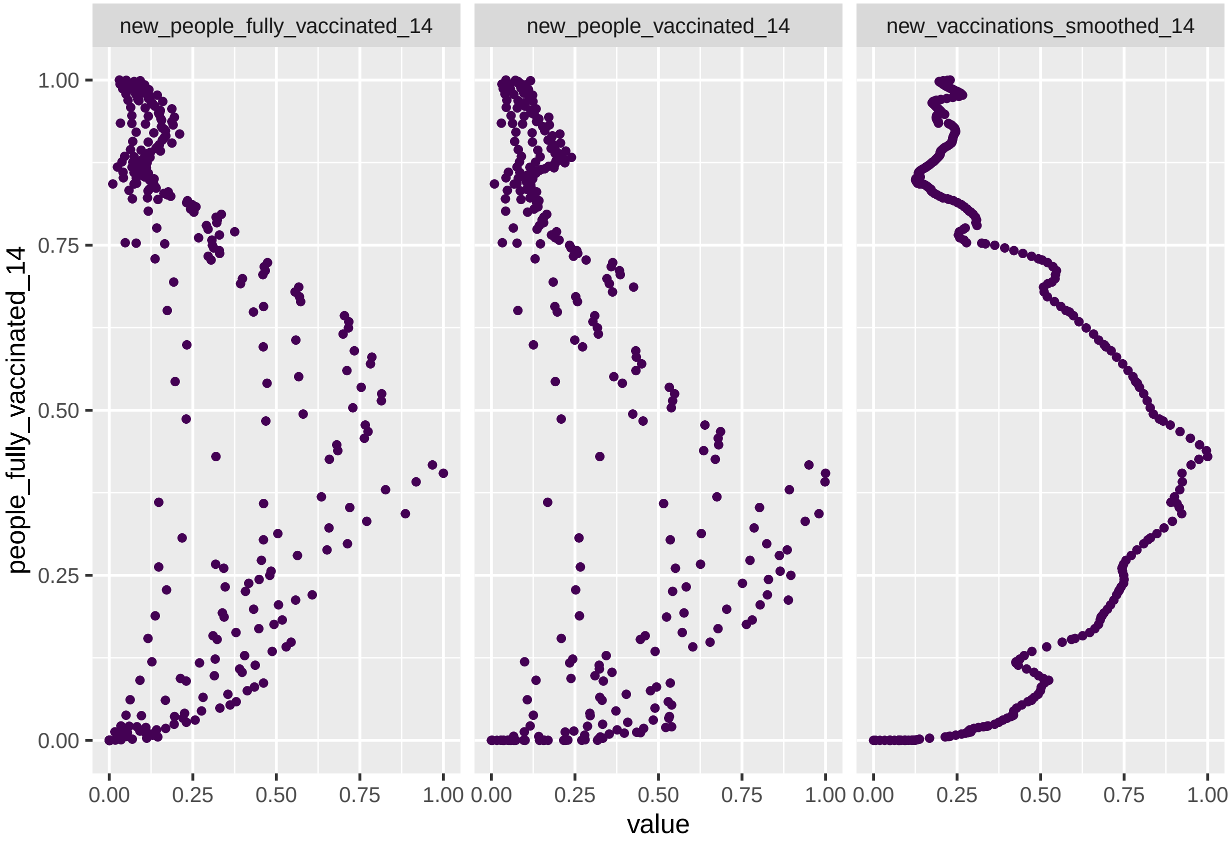

new_people_fully_vaccinated_14
new_people_vaccinated_14
new_tests_smoothed_14
new_vaccinations_smoothed_14
people_fully_vaccinated_14
people_vaccinated_14
positive_rate_14
total_vaccinations_14
reproduction_rate_14
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
value

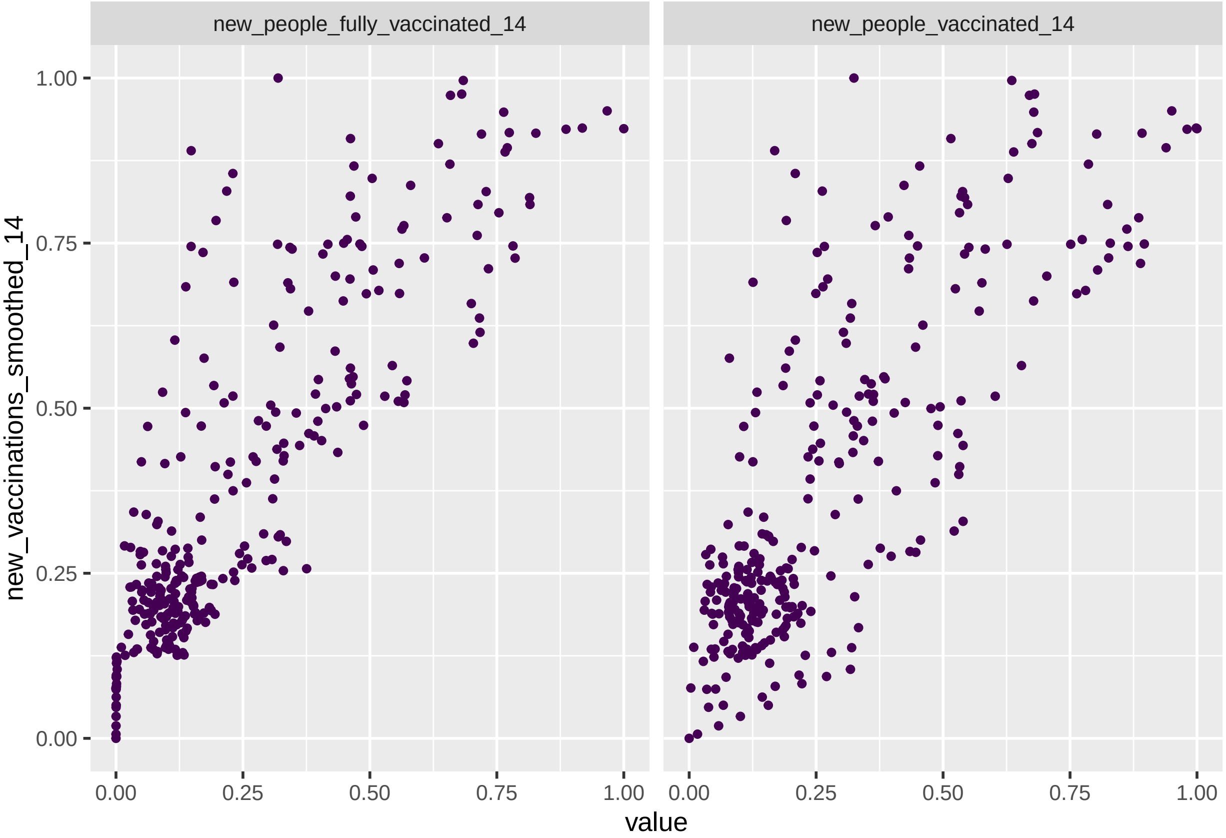

new_people_fully_vaccinated_14
new_people_vaccinated_14
new_vaccinations_smoothed_14
people_fully_vaccinated_14
people_vaccinated_14
positive_rate_14
total_vaccinations_14
new_tests_smoothed_14
value
0.00
0.25
0.50
0.75
1.00

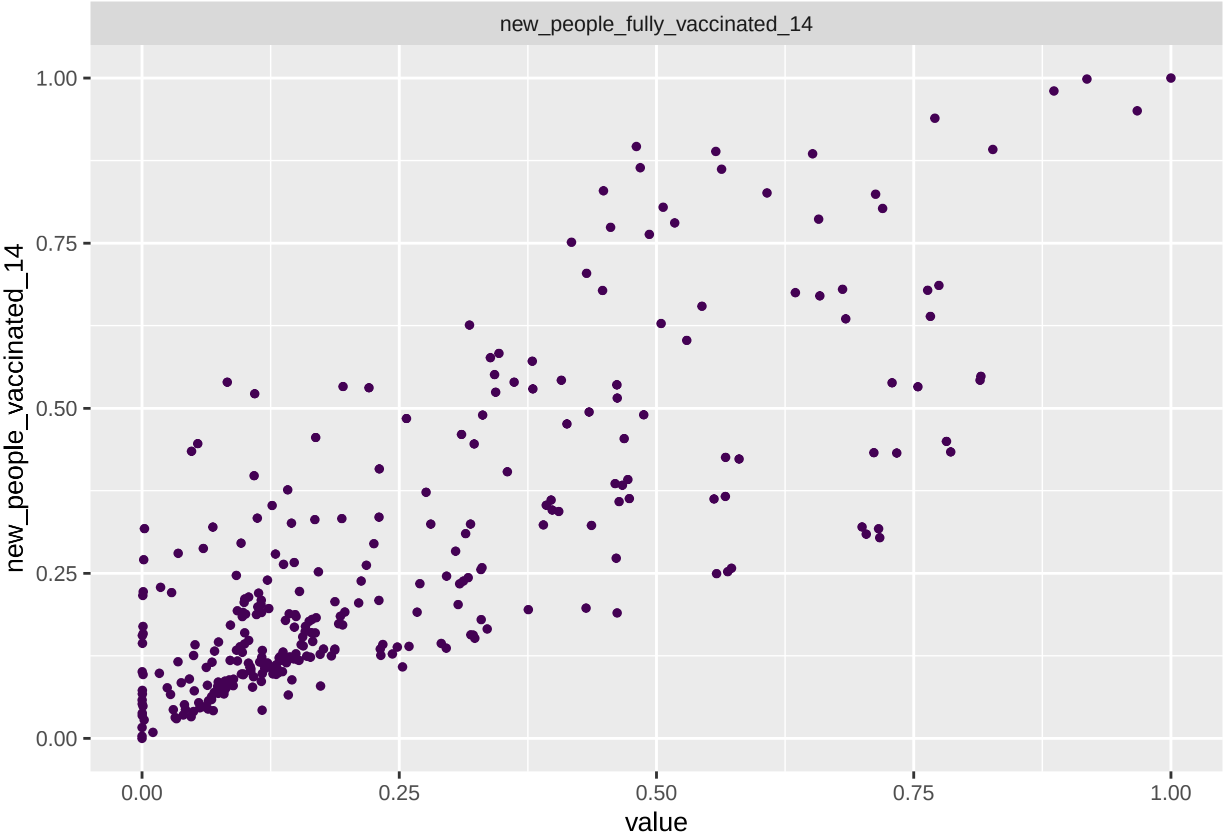

new_people_fully_vaccinated_14
new_people_vaccinated_14
new_vaccinations_smoothed_14
people_fully_vaccinated_14
people_vaccinated_14
total_vaccinations_14
positive_rate_14
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
value


new_people_fully_vaccinated_14
new_people_vaccinated_14
new_vaccinations_smoothed_14
people_fully_vaccinated_14
people_vaccinated_14
total_vaccinations_14
value
0.00
0.25
0.50
0.75
1.00


new_people_fully_vaccinated_14
new_people_vaccinated_14
new_vaccinations_smoothed_14
people_fully_vaccinated_14
people_vaccinated_14
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
value


new_people_fully_vaccinated_14
new_people_vaccinated_14
new_vaccinations_smoothed_14
people_fully_vaccinated_14
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
value


new_people_fully_vaccinated_14
new_people_vaccinated_14
new_vaccinations_smoothed_14
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
value



Closing Remarks

Our exploratory data analysis suggests that multiple variables are highly correlated with each other and we may need to combine or remove some. Nonnormality also indicates that we may wish to transform our target and/or explanatory variables depending on the model we consider.