Vaccination in the UK

Elena Basargina

Contents

1 Introduction

Hello. My name is Elena, and I am a Data Scientist.

When I came to Bristol in May 2021, I decided to be vaccinated because, in my opinion, this is a safer way to live my normal life. So, I started my long research about Covid vaccination. And now I am ready to show you interesting facts.

I live in Bristol. What do I know about Bristol?

- This city is a part of the UK, England, and South West.
- There are two universities.

So, I am interested in data about the UK, England, South West, and Bristol.

Question 0: Are there dependencies between academic year events and vaccination waves?

I was vaccinated by the first, the second, and booster doses on 8 August 2021, 3 October 2021, 8 January 2022, respectively.

Question 1: How many people got their jabs with me?

I got the first and the second jabs on Sunday. There were fewer people in the vaccination centre. When I got the third jab on Saturday, there was a big queue.

Question 2: When do people prefer to get a jab: weekdays or weekends/Saturdays or Sundays?

Question 3: Can I predict vaccination data?

Overall, I need to find datasets to answer my questions.

- Data is about the UK, England, South West, and Bristol.
- Data is about new people who got the jabs by dates.

And I was lucky to find the website "Coronavirus (COVID-19) in the UK".

2   Choosing datasets

Look at the vaccination metrics from the website "Coronavirus (COVID-19) in the UK".

Cumulative people vaccinated booster dose by publish date
Cumulative people vaccinated complete by publish date
Cumulative people fully vaccinated by vaccination date
Cumulative people vaccinated 1st dose by publish date
Cumulative people vaccinated 1st dose by vaccination date

Cumulative people vaccinated 2nd dose by publish date
Cumulative people vaccinated 2nd dose by vaccination date
Cumulative people vaccinated 3rd dose by publish date
Cumulative people vaccinated with a booster dose plus new people vaccinated with a third dose by publish date
Cumulative people vaccinated booster or third dose by vaccination date

Cumulative percentage of people vaccinated with a booster dose by publish date
Cumulative vaccination complete coverage by publish date percentage
Cumulative percentage of people fully vaccinated by vaccination date
Cumulative vaccination 1st dose uptake by publish date percentage
Cumulative percentage of people vaccinated with a first dose by vaccination date

Cumulative vaccination 2nd dose uptake by publish date percentage
Cumulative percentage of second dose vaccination uptake by vaccination date
Cumulative percentage of people vaccinated with a booster dose plus people vaccinated with third dose by publish date
Cumulative percentage of people vaccinated with a booster or third dose by vaccination date
Cumulative vaccines given by publish date

New people receiving 1st dose
New people receiving 2nd dose
New people vaccinated with a booster dose by publish date
New people vaccinated complete by publish date
New people fully vaccinated by vaccination date

New people vaccinated 1st dose by publish date
New people vaccinated with a first dose by vaccination date
New people vaccinated 2nd dose by publish date
New people vaccinated with a second dose by vaccination date
New people vaccinated with a third dose by publish date

New people vaccinated with a booster dose plus new people vaccinated with a third dose by publish date
New people vaccinated with a booster or third dose by vaccination date
New vaccines given by publish date
Vaccinations age demographics breakdown
Vaccination register (NIMS) population by vaccination date

Weekly people vaccinated 1st dose by vaccination date
Weekly people vaccinated 2nd dose by vaccination date

I am going to use metrics

1. which names start with "New",
2. by vaccination date,
3. and are not deprecated.

Which datasets have these metrics?

	ltla	msoa	nation	nhsRegion	nhsTrust	overview	region	utla
cumPeopleVaccinatedBoosterDoseByPublishDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedCompleteByPublishDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedCompleteByVaccinationDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedFirstDoseByPublishDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedFirstDoseByVaccinationDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedSecondDoseByPublishDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedSecondDoseByVaccinationDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedThirdDoseByPublishDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedThirdInjectionByPublishDate	0	0	0	0	0	0	0	0
cumPeopleVaccinatedThirdInjectionByVaccinationDate	0	0	0	0	0	0	0	0
cumVaccinationBoosterDoseUptakeByPublishDatePercentage	0	0	0	0	0	0	0	0
cumVaccinationCompleteCoverageByPublishDatePercentage	0	0	0	0	0	0	0	0
cumVaccinationCompleteCoverageByVaccinationDatePercentage	0	0	0	0	0	0	0	0
cumVaccinationFirstDoseUptakeByPublishDatePercentage	0	0	0	0	0	0	0	0
cumVaccinationFirstDoseUptakeByVaccinationDatePercentage	0	0	0	0	0	0	0	0
cumVaccinationSecondDoseUptakeByPublishDatePercentage	0	0	0	0	0	0	0	0
cumVaccinationSecondDoseUptakeByVaccinationDatePercentage	0	0	0	0	0	0	0	0
cumVaccinationThirdInjectionUptakeByPublishDatePercentage	0	0	0	0	0	0	0	0
cumVaccinationThirdInjectionUptakeByVaccinationDatePercentage	0	0	0	0	0	0	0	0
cumVaccinesGivenByPublishDate	0	0	0	0	0	0	0	0
newPeopleReceivingFirstDose	0	0	0	0	0	0	0	0
newPeopleReceivingSecondDose	0	0	0	0	0	0	0	0
newPeopleVaccinatedBoosterDoseByPublishDate	0	0	1	0	0	0	0	0
newPeopleVaccinatedCompleteByPublishDate	0	0	0	0	0	0	0	0
newPeopleVaccinatedCompleteByVaccinationDate	0	0	0	0	0	0	0	0
newPeopleVaccinatedFirstDoseByPublishDate	1	0	1	0	0	1	0	1
newPeopleVaccinatedFirstDoseByVaccinationDate	1	0	1	0	0	0	1	1
newPeopleVaccinatedSecondDoseByPublishDate	1	0	1	0	0	1	0	1
newPeopleVaccinatedSecondDoseByVaccinationDate	1	0	1	0	0	0	1	1
newPeopleVaccinatedThirdDoseByPublishDate	0	0	1	0	0	0	0	0
newPeopleVaccinatedThirdInjectionByPublishDate	1	0	1	0	0	1	0	1
newPeopleVaccinatedThirdInjectionByVaccinationDate	1	0	1	0	0	0	1	1
newVaccinesGivenByPublishDate	0	0	1	0	0	1	0	0
vaccinationsAgeDemographics	0	0	0	0	0	0	0	0
VaccineRegisterPopulationByVaccinationDate	0	0	0	0	0	0	0	0
weeklyPeopleVaccinatedFirstDoseByVaccinationDate	0	0	0	0	0	0	0	0
weeklyPeopleVaccinatedSecondDoseByVaccinationDate	0	0	0	0	0	0	0	0

Table 1: Vaccination metrics with highlight

So, I need to look at the datasets:

Lower Tier Local Authority (LTLA)
Nation
Region
Upper Tier Local Authority (UTLA)

The dataset "Overview" presents the UK. Which metrics do we have for this dataset?

newPeopleVaccinatedFirstDoseByPublishDate
newPeopleVaccinatedSecondDoseByPublishDate
newPeopleVaccinatedThirdInjectionByPublishDate
newVaccinesGivenByPublishDate

So, we do not have data by vaccination date for the UK. I am going to use metrics from the Table 1 for

1. Bristol (Lower Tier Local Authority (LTLA)),
2. South West (Region)
3. and England (Nation).

That is enough for answering.

3 Exploring datasets

3.1 South West

Look at the dataset for South West.

areaCode
areaName
areaType
date
newPeopleVaccinatedFirstDoseByVaccinationDate
newPeopleVaccinatedSecondDoseByVaccinationDate
newPeopleVaccinatedThirdInjectionByVaccinationDate

We have additional columns. Let's look at them.

- For **areaCode** unique value is E12000009,

- for **areaName** unique value is South West,

- for **areaType** unique value is region.

So, we do not need to look at them in the future because these columns are used for filtering that we have already done on the website.

Let's prepare data for the plotting.

- Rename columns

- Create a long table

date	Dose	Count
2022-01-26	First	986
2022-01-25	First	899
2022-01-24	First	723
2022-01-23	First	1035
2022-01-22	First	1822
2022-01-21	First	1085

Let's plot data.

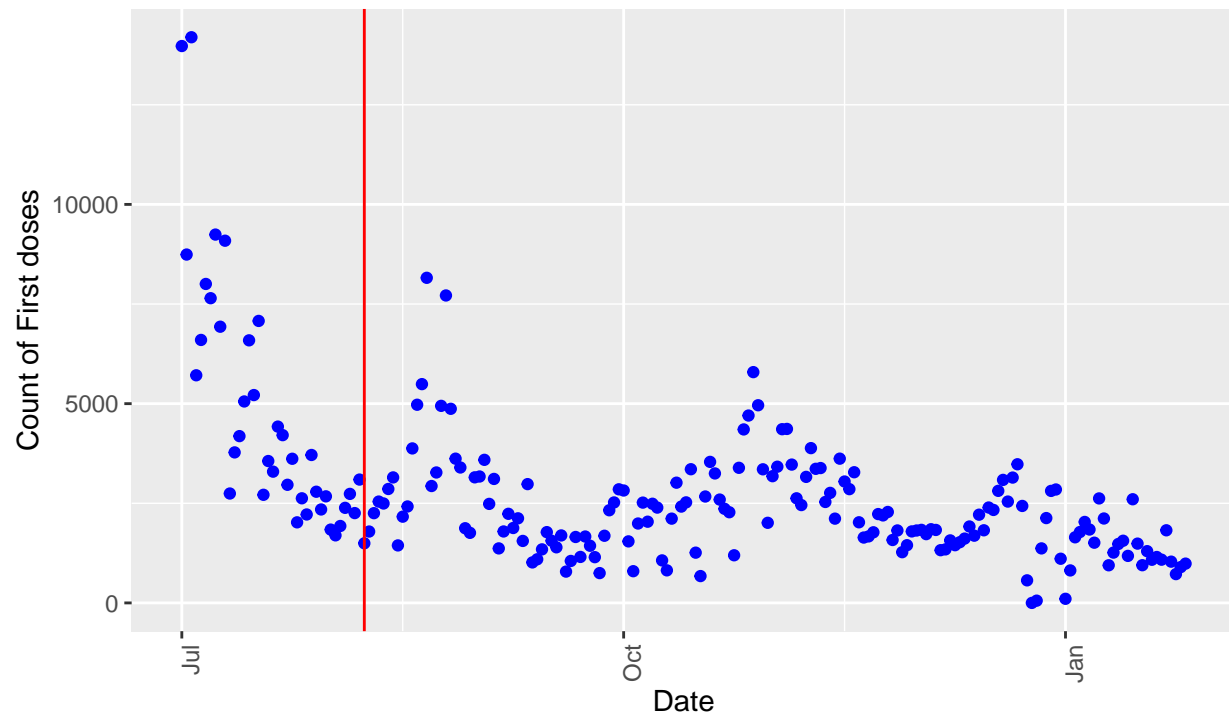As I say in the part 1, I have a question: How many people got their jabs with me?

Let's answer. The red lines are my vaccination dates.
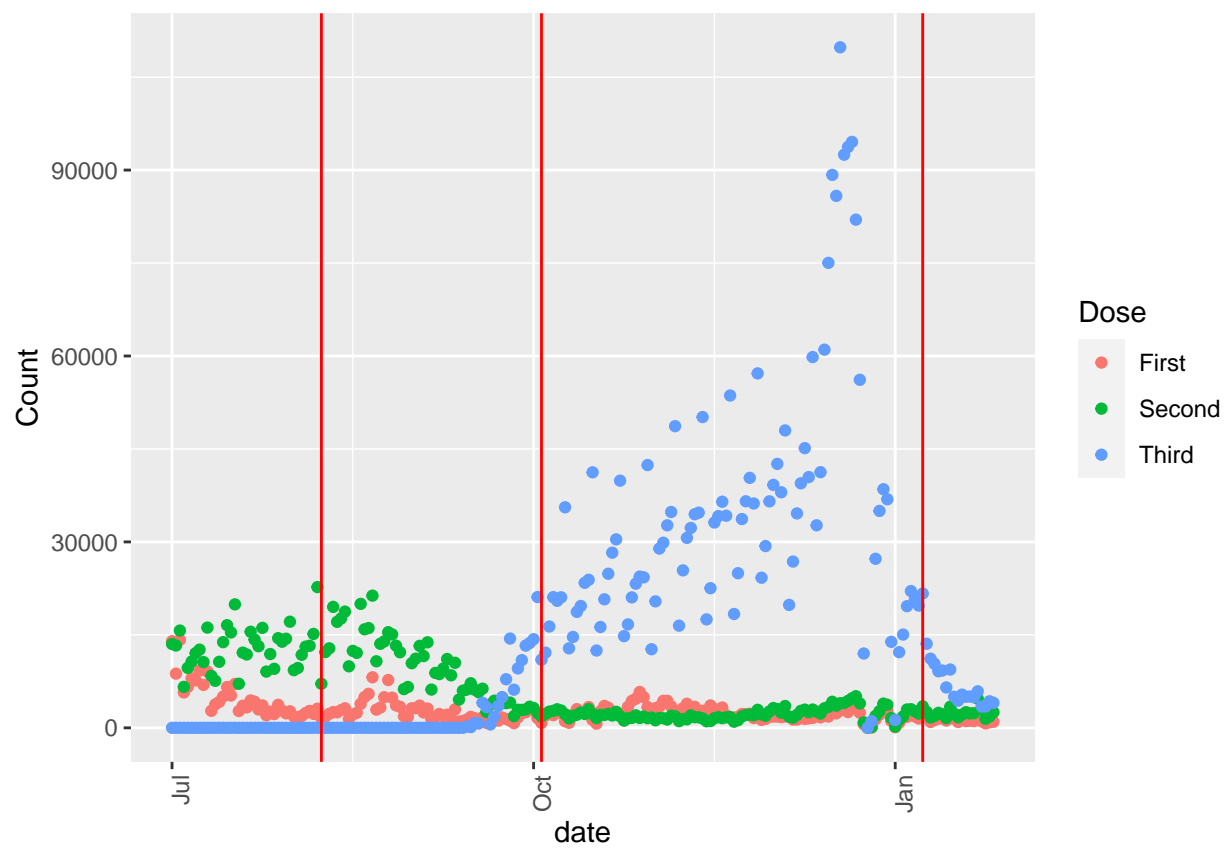
The result is not beautiful because of the active growth of the third jabs count at the end of 2021.

Let's plot them separately.

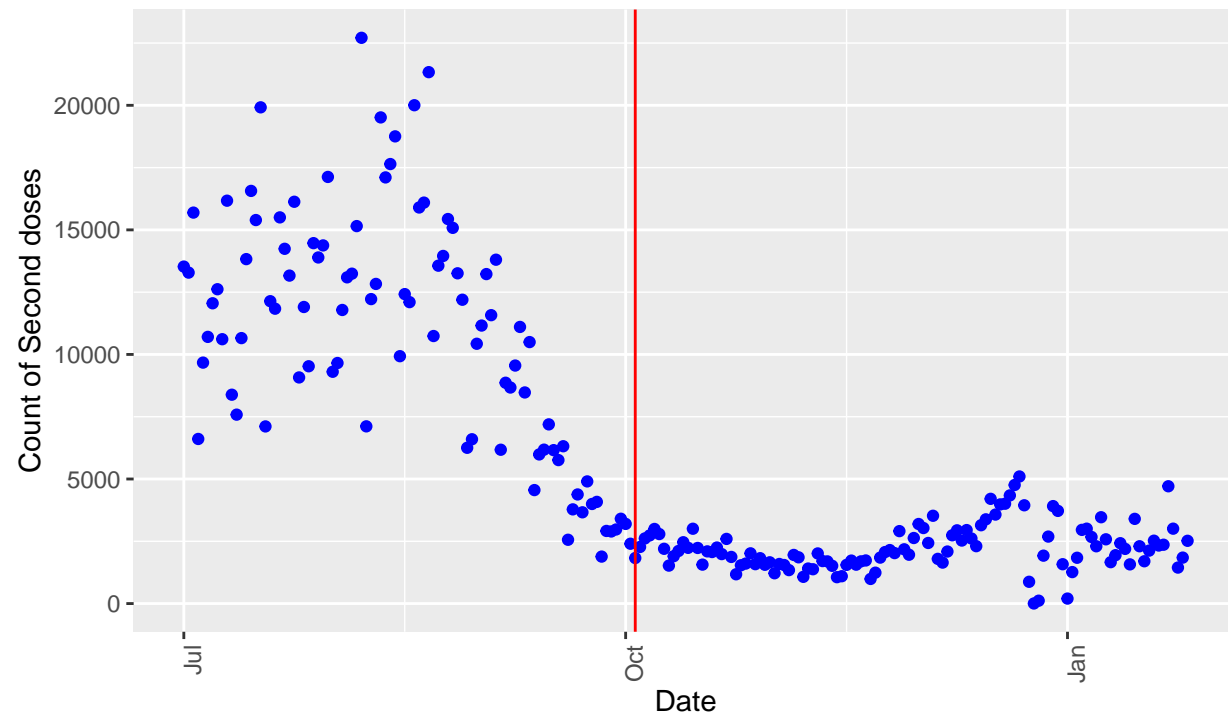## Vaccination in South West
The First dose

More information https://coronavirus.data.gov.uk/details/about−data

It is so interesting why the graph is wavy. 1496 people got their first jabs with me.

Vaccination in South West
The Second dose

More information https://coronavirus.data.gov.uk/details/about−data

1828 people got their second jabs with me.

Vaccination in South West
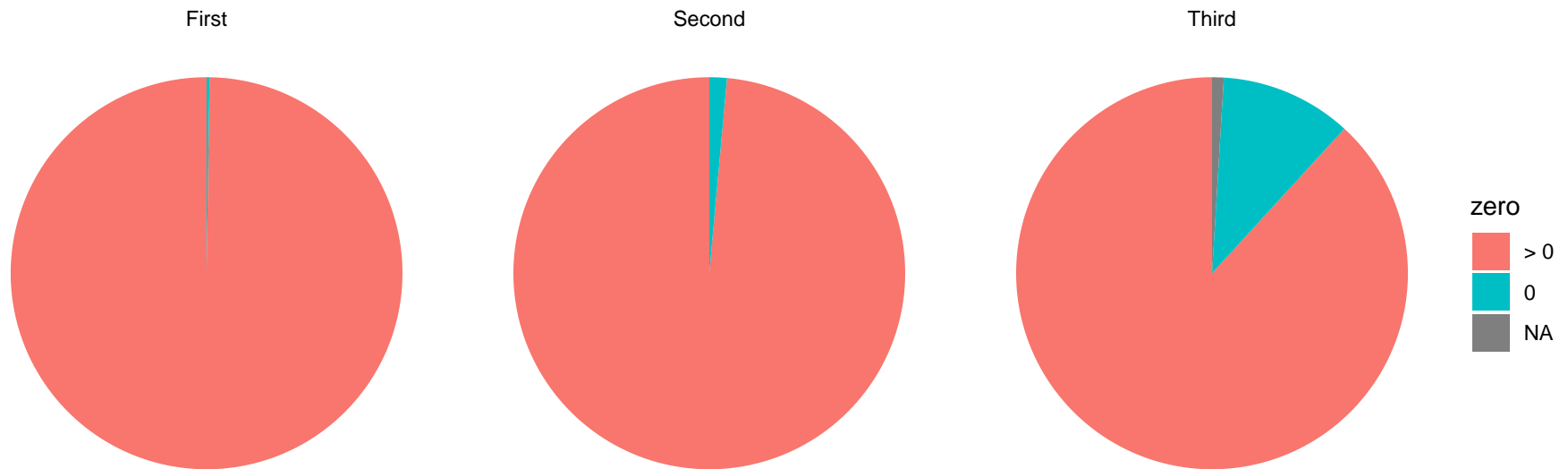The Third dose

More information https://coronavirus.data.gov.uk/details/about−data
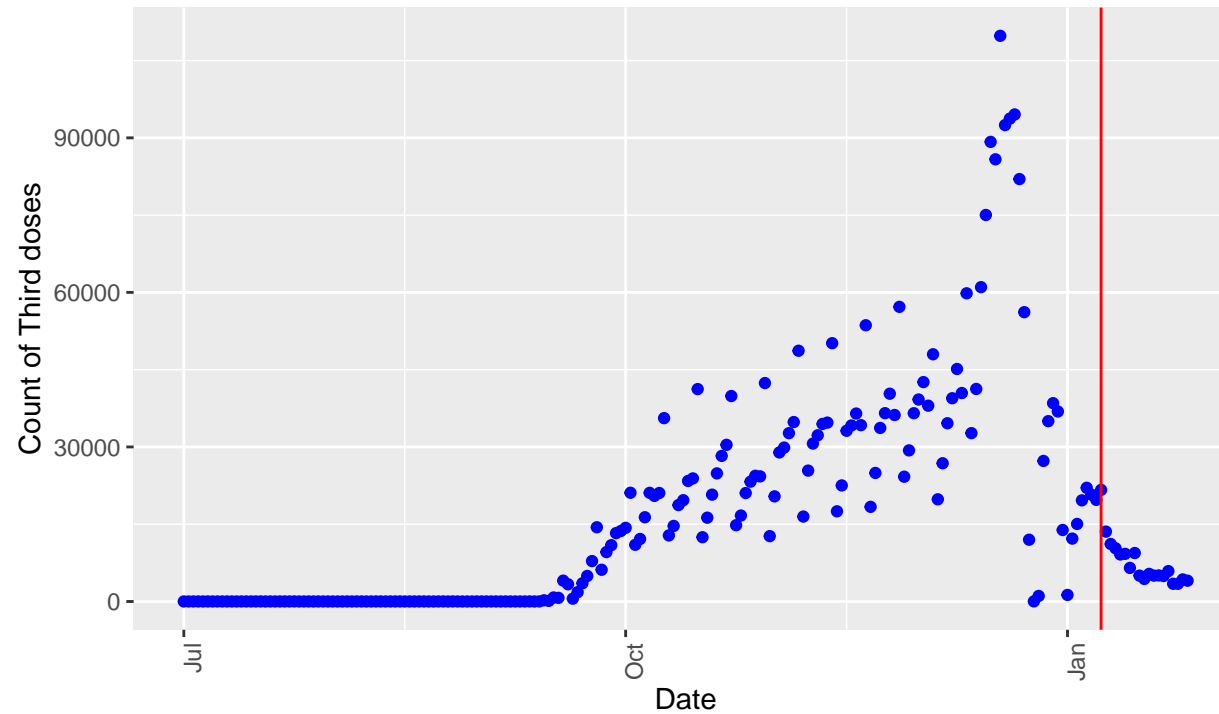
21664 people got their third jabs with me. We can see when the active phase of vaccination by the third dose started.

3.1.1 Zeroes

First   Second   Third

zero
- > 0
- 0
- NA

The column "Third" has more zero values than "First" and "Second; but, I think, it won't influence models' accuracy. Also, we can see missing values for the column"Third"; in our case, missing values mean that nobody got the third jab. I suggest replacing them with zeroes.

Replace missing values.

3.1.2 Data description

Median, percentiles and mean

	mean	median	Q0.25	Q0.75	Q0.9
First	11037.614	4210	2072.0	14213.0	31224.6
Second	10367.352	3998	1659.5	13708.5	28700.6
Third	8589.925	6	2.0	10629.5	33689.0

What can I say?

- Mean and median have a visible difference. So, there are large extreme values.

- For the Third dose, half of the values are below 6. That is not surprising. In the beginning, people needed to get two jabs.

- If we look at "Q0.25", "Q0.75", "Q0.90", we find out that the Third dose's wave caught up with other doses' waves quickly. We already saw this fact on the plots.
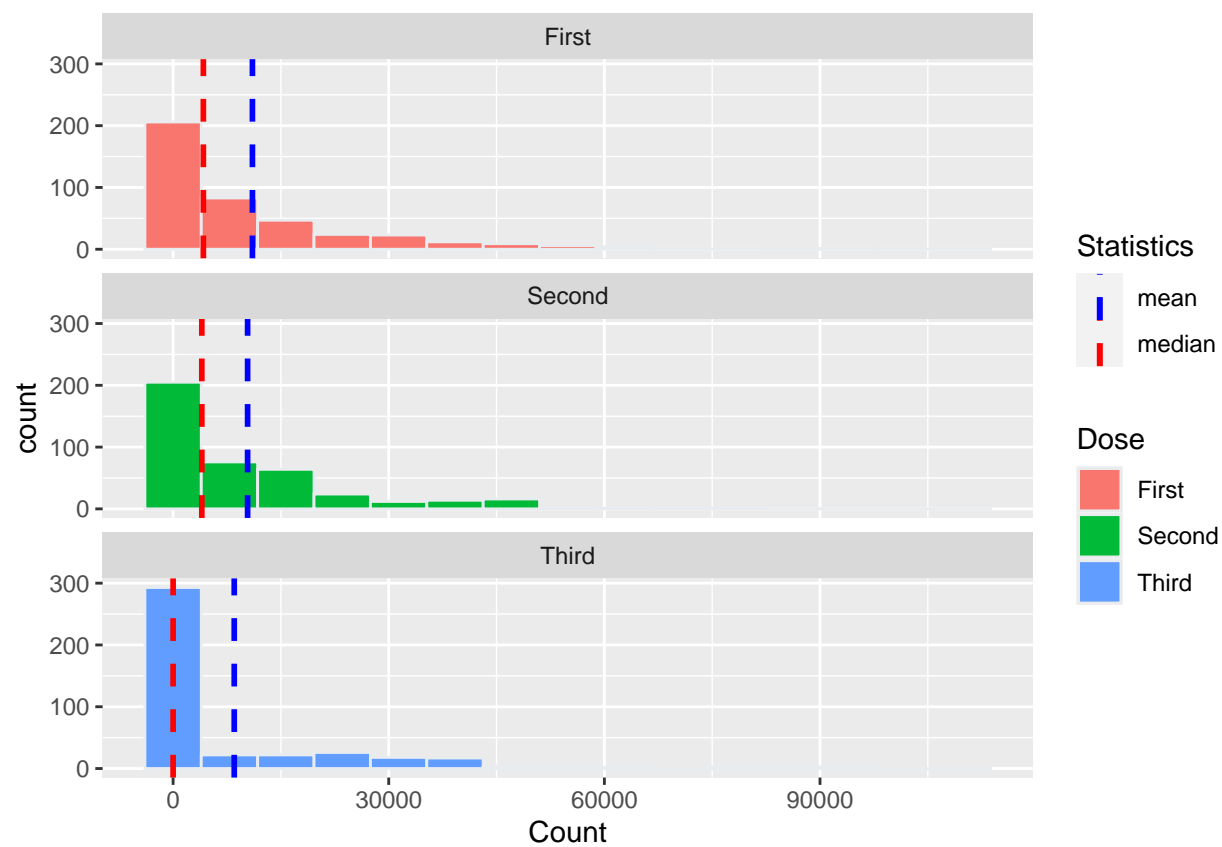
Standard deviation (sd), IQR and range

	sd	range	IQR
First	13971.58	84537	12141.0
Second	13477.22	78425	12049.0
Third	17512.10	109810	10627.5

IQR and standard deviation for each dose are big, consequently, the data spread out. Also, we can see the difference between largest and smallest values in the column "range".

3.2 Bristol

The dataset's columns:

| |
| --- |
| date |
| age |
| newPeopleVaccinatedFirstDoseByVaccinationDate |
| newPeopleVaccinatedSecondDoseByVaccinationDate |
| newPeopleVaccinatedThirdInjectionByVaccinationDate |

Just rename columns and we will move on to answer the questions.

age	date	First	Second	Third
12_15	2022-03-24	6	34	0
16_17	2022-03-24	0	7	23
18_24	2022-03-24	13	15	30
25_29	2022-03-24	3	11	14
30_34	2022-03-24	0	5	13
35_39	2022-03-24	1	3	7

Question 1: How many people got their jabs with me?

I suggest looking at this graph from left to right and from down to up. Also, I put my dates on the graphs as the red lines.
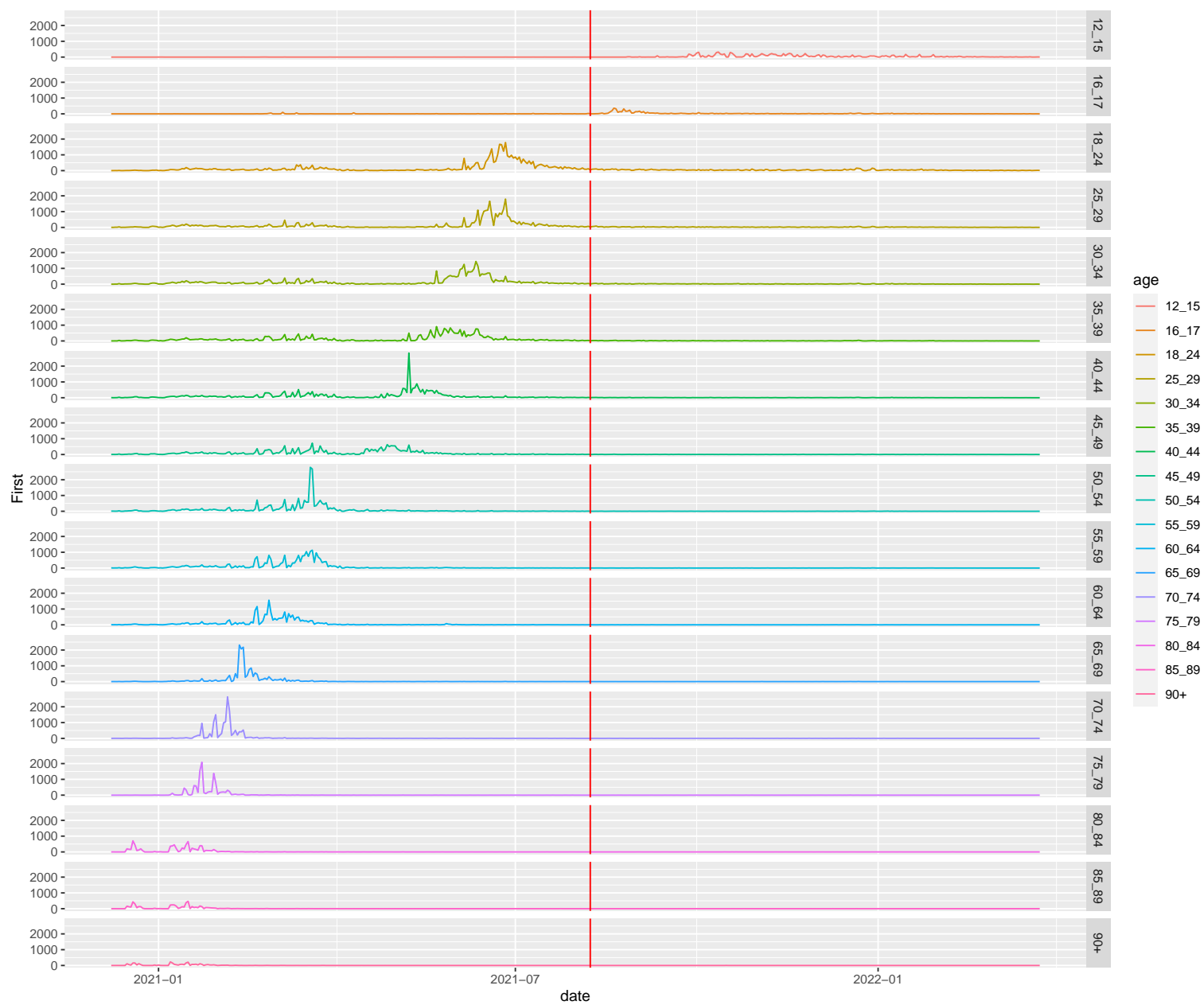
20 people in my age group got their First jabs with me in Bristol.

We can see that the active phases of the vaccination depends on the age group. The first group was people 90 years and older. The last one is the people 12-15 years old. It is not surprising, because, as I know, it was a government strategy.
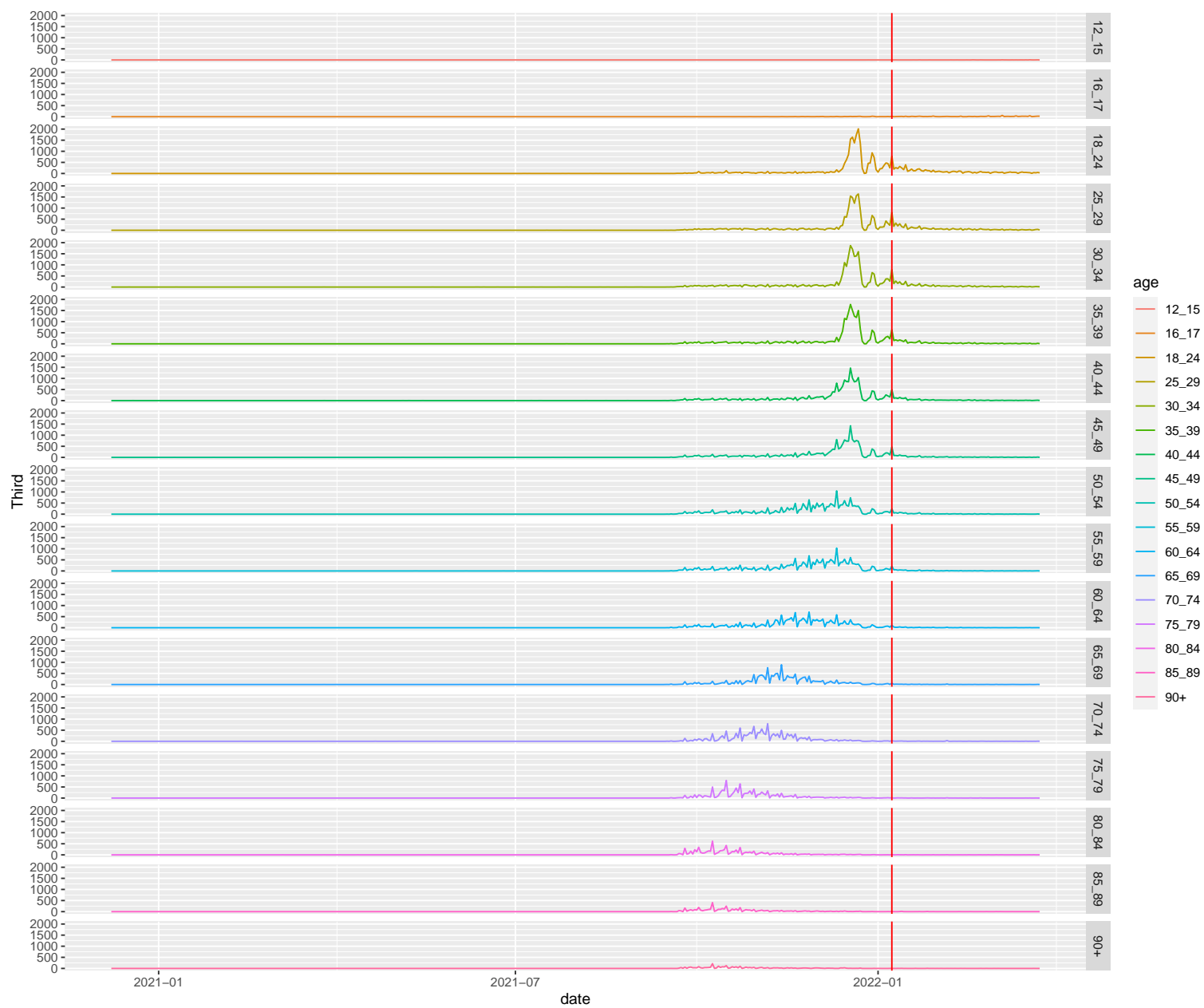
Look at the graph for the second dose.
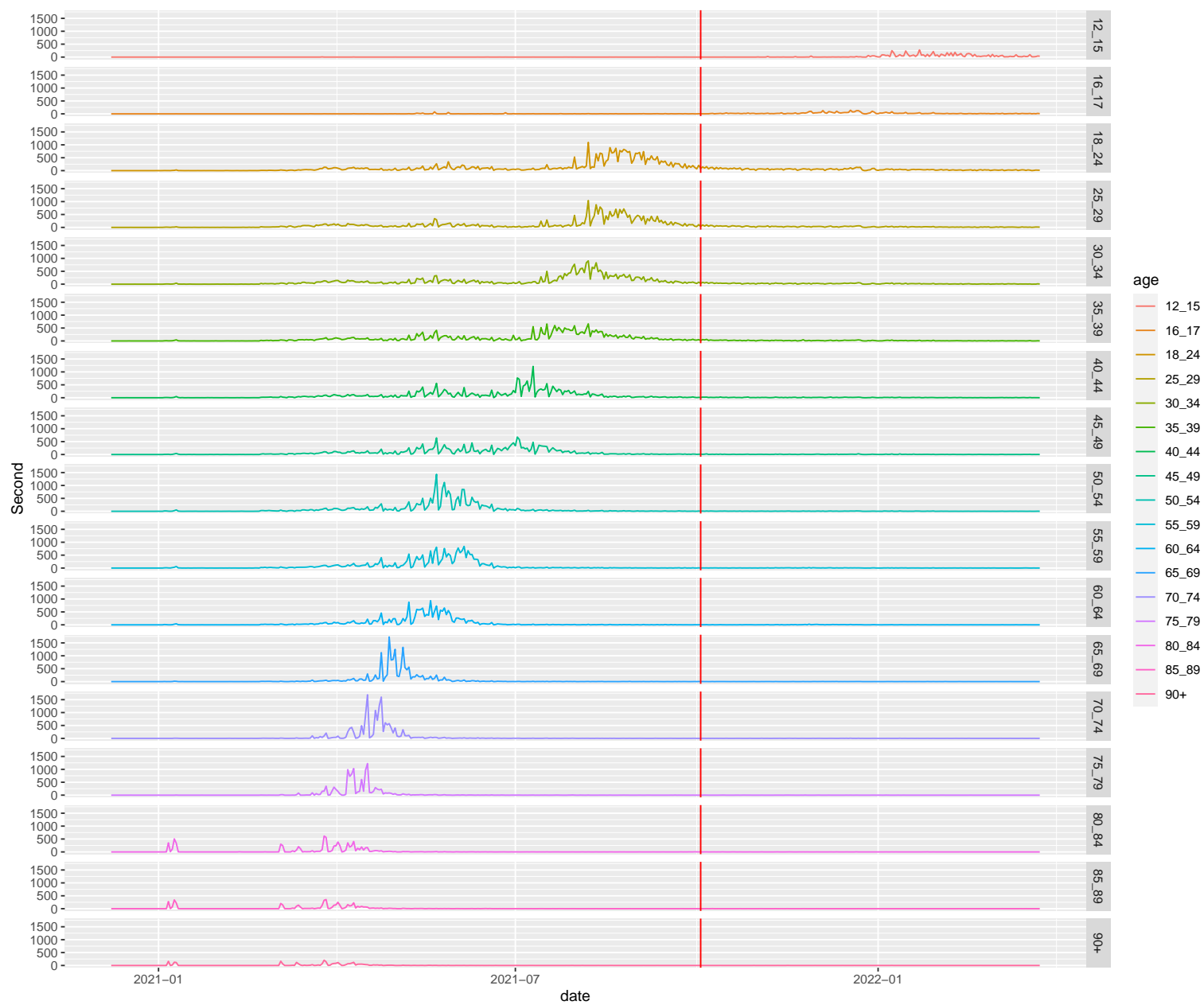
38 people in my age group got their Second jabs with me in Bristol.

The second dose is 2 months after the first. And we can see the waves too.

Look at the graph for the third dose.

804 people in my age group got their Third jabs with me in Bristol.

So, my date of the vaccination was not in the active phase. And I can't imagine how many people were in the vaccination centre at the end of December 2021.

Overall, I got my jabs in the not active phases of the vaccination, but I got the third one with a lot of people. And the count of jabs depends on dates. Keep this fact in mind.

Question 0: Are there dependencies between academic year events and vaccination waves?

There are two large universities in Bristol. I guess that the vaccination waves also depend on the academic year events.

I will be use term-dates of UWE academic years: 2020-2021, 2021-2022, 2022-2023.
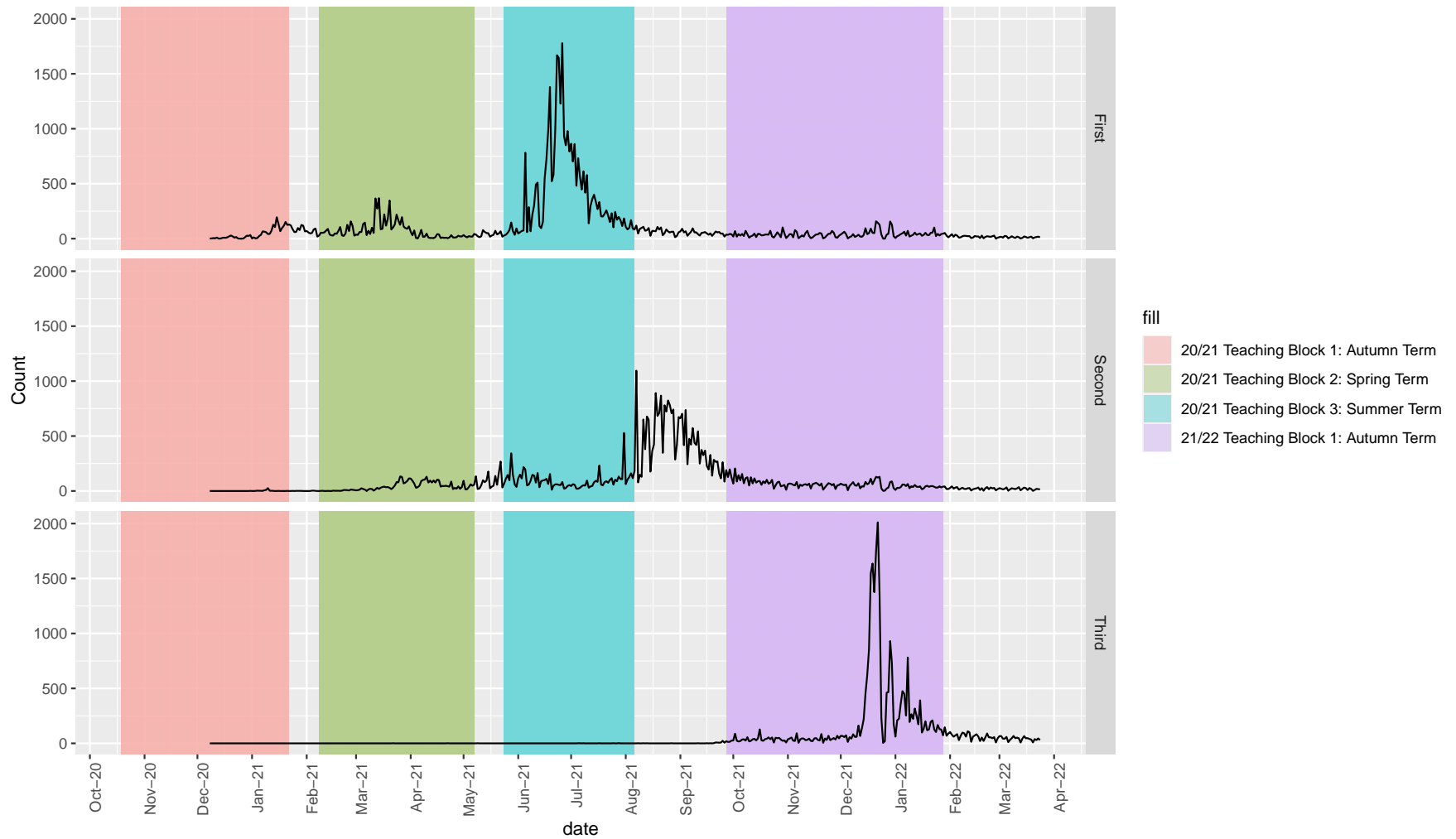
The result is interesting but requires research. As I can see, the active phase of the vaccination for the age group 18-24 was during the Summer term. But why? It may be because students decided to prepare themselves for the new academic year. It may be because of the government or NHS suggestions. It is an open question.

3.3   England

Look at the dataset's columns.

areaCode
areaName
areaType
date
newPeopleVaccinatedFirstDoseByVaccinationDate
newPeopleVaccinatedSecondDoseByVaccinationDate
newPeopleVaccinatedThirdInjectionByVaccinationDate

We are not going to look at the columns areaCode, areaName, areaType, because these columns have one unique value (E92000001 for areaCode, England for areaName, nation for areaType), they are used for filtering on the website.

Rename columns and exclude unnecessary columns.

	Date	First	Second	Third
457	2020-12-08	5370	146	NA
456	2020-12-09	9648	137	2
455	2020-12-10	11888	119	1
454	2020-12-11	12516	82	1
453	2020-12-12	10565	23	3
452	2020-12-13	6134	30	0

I want to be sure that the file has statistics for every day.

The difference between dates is 1. So, we have full statistics by date.

Do we have missing values?

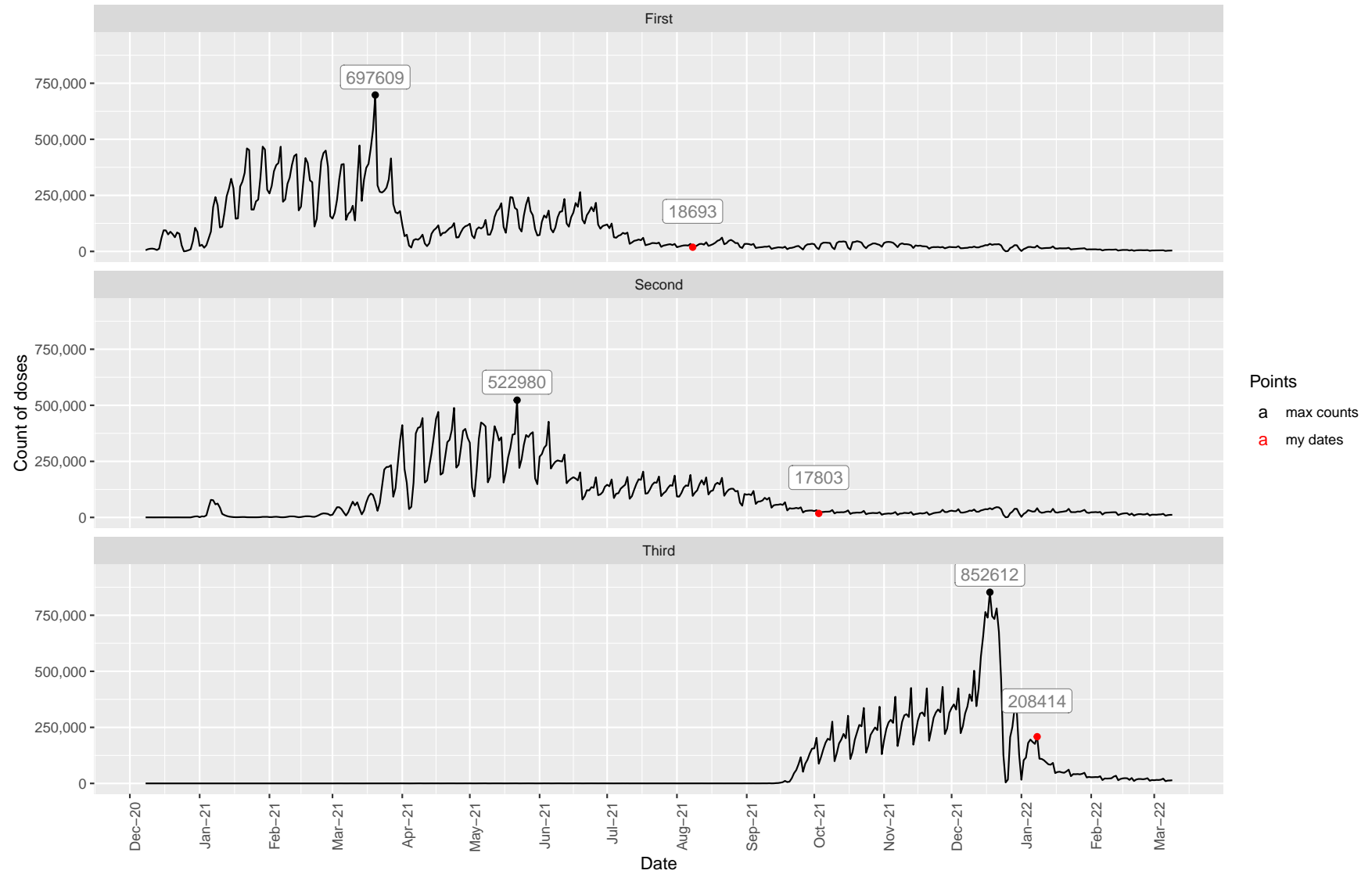|	
Date	0
First	0
Second	0
Third	1
Diffdate	0
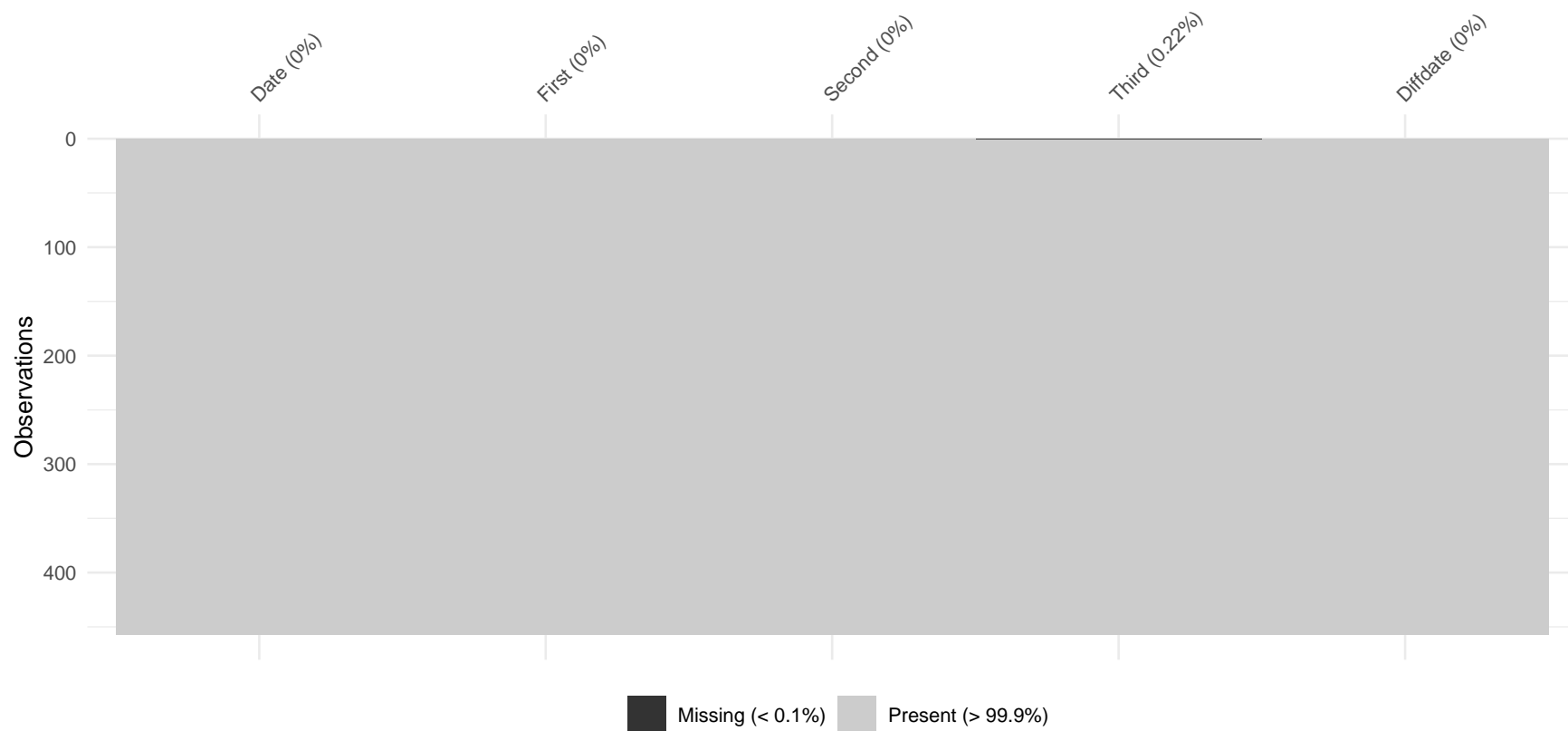
Yes, we have missing values in the column "Third". In this case, missing values and zeroes are equivalent.

Replace by zero.

I have a question.

How many people got their jabs with me?

Vaccination in England

4 Modeling

So, I am curious. Can I predict vaccination data?

4.1 South West

4.1.1 Look at and Modify the dataset

I will work with the South West's vaccination data.

	First	Second	Third
2022-01-26	986	2520	4034
2022-01-25	899	1845	4283
2022-01-24	723	1445	3441
2022-01-23	1035	3007	3439
2022-01-22	1822	4709	5896
2022-01-21	1085	2362	4944
2022-01-20	1152	2330	5058
2022-01-19	1083	2524	5017
2022-01-18	1298	2126	5359
2022-01-17	946	1699	4374
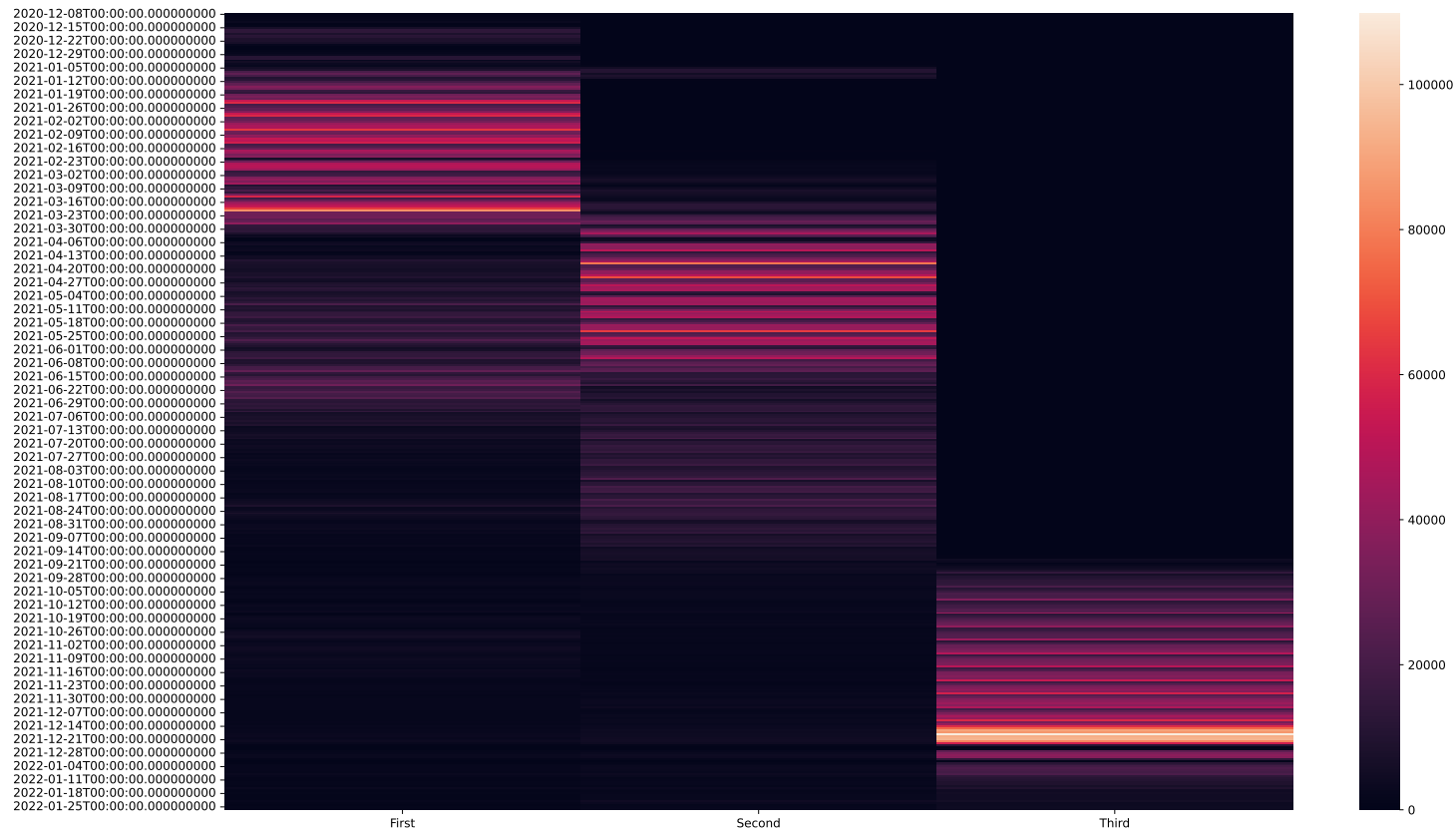
As we discuss earlier 3.1, there are waves. So, the count of jabs depends on dates.

Let's get features: 1) Year 2) Month 3) Day etc.

	First	Second	Third	Year	Month	Day	DayOfYear	Weekday	Quarter	IsMonthStart	IsMonthEnd
2022-01-26	986	2520	4034	2022	1	26	26	2	1	FALSE	FALSE
2022-01-25	899	1845	4283	2022	1	25	25	1	1	FALSE	FALSE
2022-01-24	723	1445	3441	2022	1	24	24	0	1	FALSE	FALSE
2022-01-23	1035	3007	3439	2022	1	23	23	6	1	FALSE	FALSE
2022-01-22	1822	4709	5896	2022	1	22	22	5	1	FALSE	FALSE
2022-01-21	1085	2362	4944	2022	1	21	21	4	1	FALSE	FALSE
2022-01-20	1152	2330	5058	2022	1	20	20	3	1	FALSE	FALSE
2022-01-19	1083	2524	5017	2022	1	19	19	2	1	FALSE	FALSE
2022-01-18	1298	2126	5359	2022	1	18	18	1	1	FALSE	FALSE
2022-01-17	946	1699	4374	2022	1	17	17	0	1	FALSE	FALSE

First of all, I am going to use Regression Machine Learning models:

- Decision Tree

- Random Forest.

What is my plan?

1. Read data

I already did this step.

2. Understand statistics about the data

It will be helpful to choose the right features for better results.

- Work with missing data and categorical variables
- Work with outliers or not completed data.

5. Store prediction target (y) in a Series, selecting multiple features by providing a list of column names inside brackets, define X (subset with features), check the X summary.

6. Choose the library

7. Build and use the model

- What type of model will it be?
- Capture patterns from provided data.
- Predict
- Evaluate = Determine how accurate the model's predictions are

Let's look at the dataset carefully.

4.1.2 Explore the dataset

In the previous chapter 3.1, we already looked at the South West's data. Do we need to know something else? Yes.

4.1.2.1 Data types It is important to know which types of data columns have. Sometimes we don't realise what we see: the string or the number.

First	double
Second	double
Third	double
Year	double
Month	double
Day	double
DayOfYear	double
Weekday	double
Quarter	double
IsMonthStart	logical
IsMonthEnd	logical

The good news is I don't need to convert my variables because they fit into Regression Machine Learning models.

We will move on to correlations.

4.1.2.2   Correlations   What do we need to remember? Correlation does not imply causation. So, the columns that have a strong relationship may show low accuracy in the model.
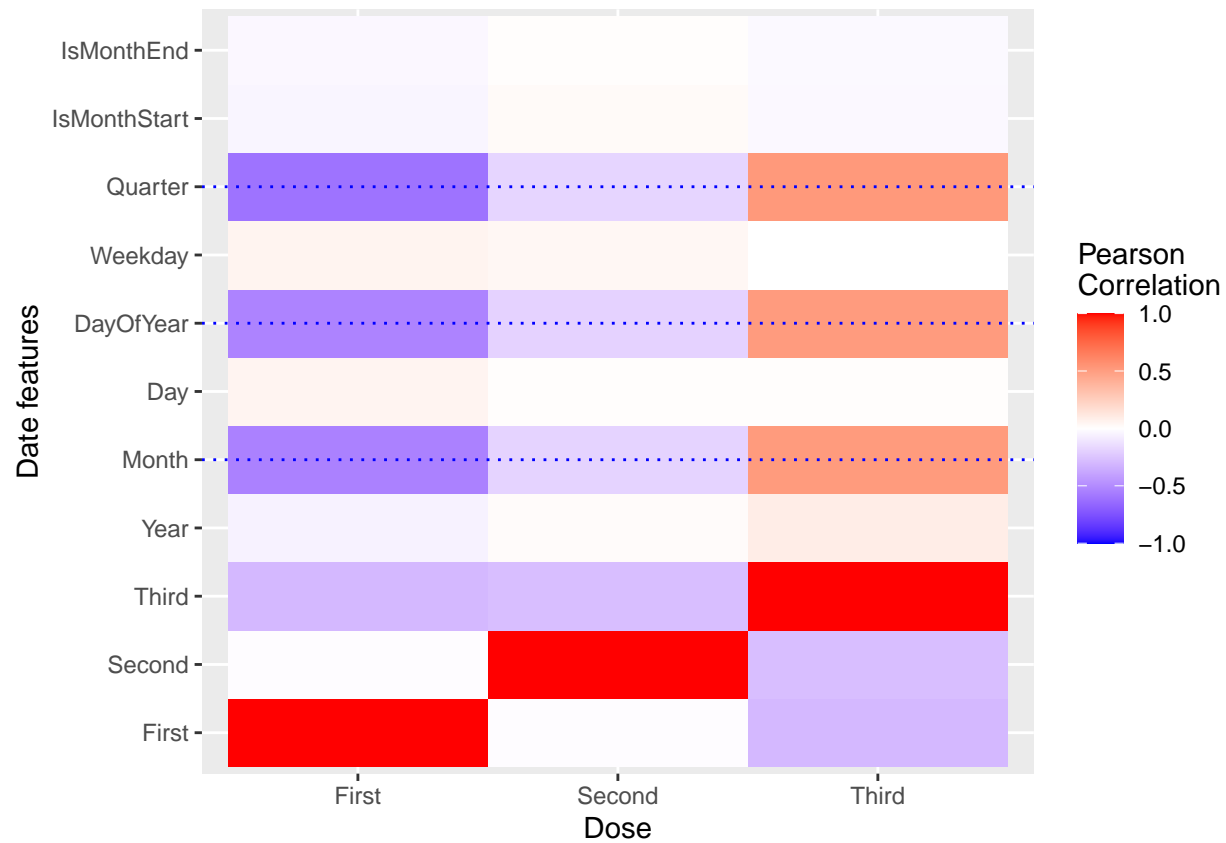
In the table below, we can see the numeric values.

First	Month	-0.5432969
Second	Month	-0.1888013
Third	Month	0.5158655
First	DayOfYear	-0.5343244
Second	DayOfYear	-0.1901012
Third	DayOfYear	0.5133313
First	Quarter	-0.6070344
Second	Quarter	-0.1799906
Third	Quarter	0.5244181

As we can see, the column "First" has a strong relationship with

- "Quarter",

- "DayOfYear",

- "Month".

At the same time, the column "Second" doesn't have strong relationships; but we can use the same columns.

4.1.2.3   Weekdays   As you remember, I have a question. When do people prefer to get a jab: weekdays or weekends/Saturdays or Sundays? It may be helpful to choose the right features.

Let's answer.
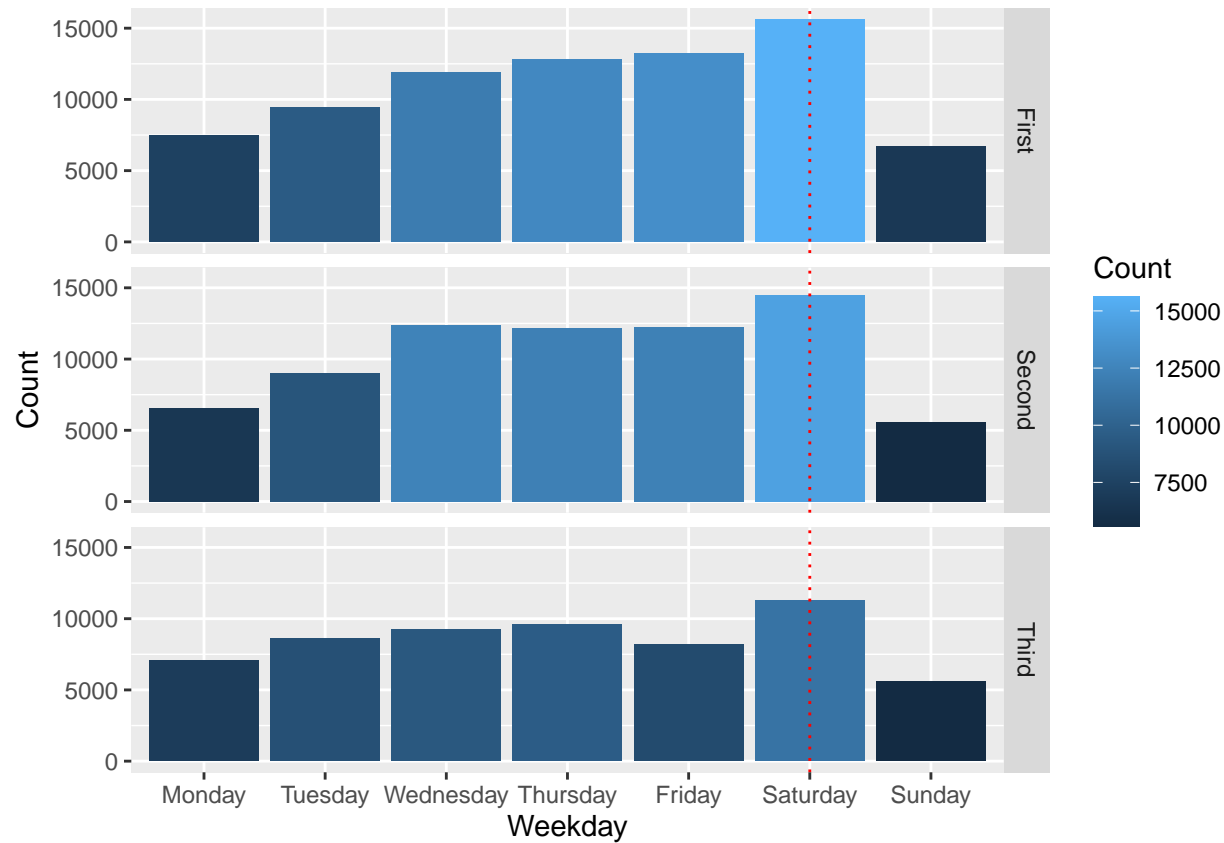
So, most of South West's people prefer to get a jab on Saturdays. That is not illogical because, for example, for me, the side effects go away during the weekend.

4.1.2.4   Missing values   As we already saw in the previous chapter 3.1, the column "Third" has missing values, but we can replace them with zeroes. Do we have the dates when nobody got the jab?

Calculate a count of dates in the dataset.

```
## 415
```

Calculate a count of dates between maximum and minimum dates.

```
## 415
```

There are no missing dates.

So, we have finished the dataset exploring. The next steps are about the models.

4.1.3 Split sets, train a Machine Learning Model and Evaluate performance

Define necessary variables

First of all, I will use all columns that I have.

Year
Month
Day
DayOfYear
Weekday
Quarter
IsMonthStart
IsMonthEnd

Prepare sets and train models using parameters.

The first column that I will predict is "First".

Finally, look at the result.

Comparing (First)
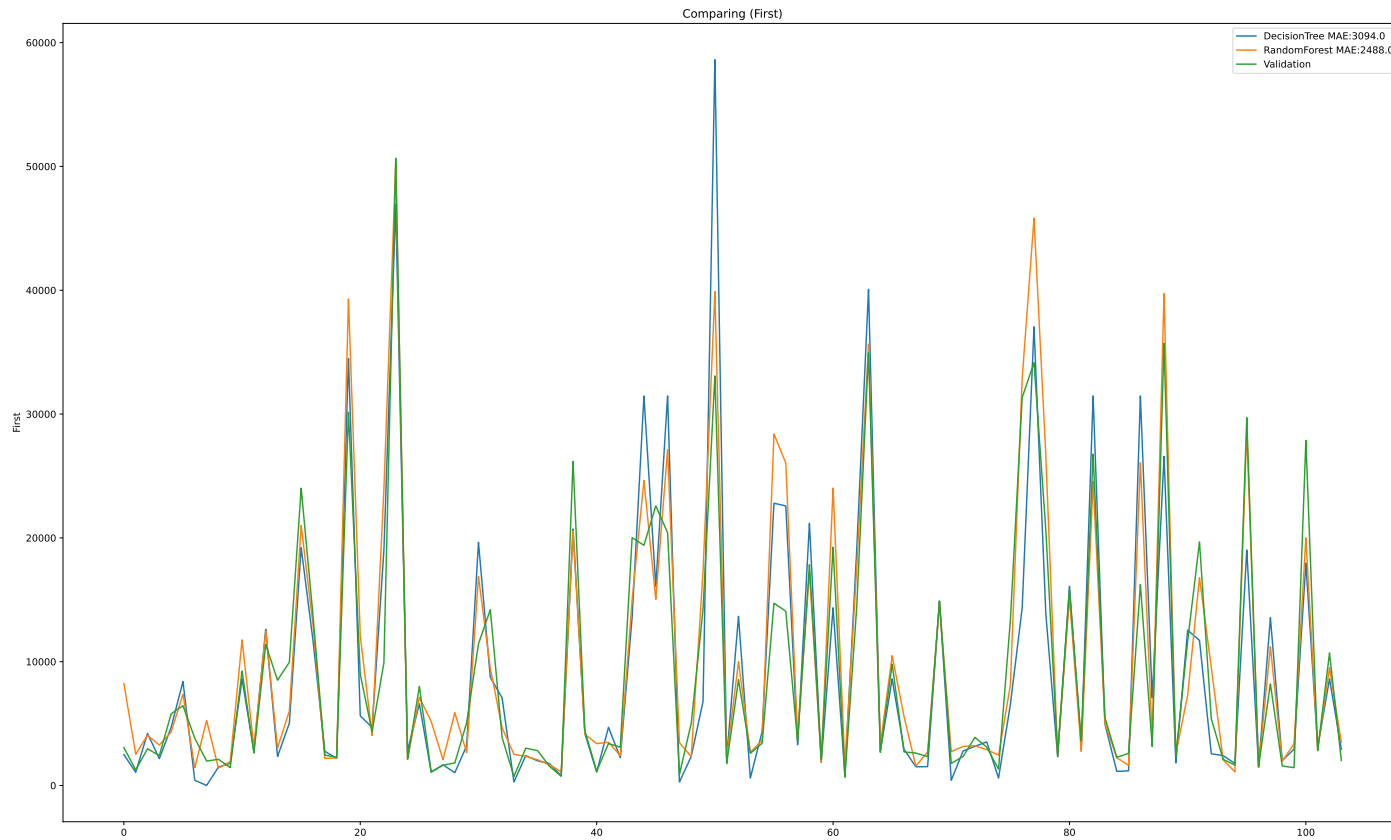
In my opinion, the result is good.

- The waves were recognized.

- The extreme values are bigger than in real data.

Let's work with the columns that I chose during the dataset exploring.

- "Weekday" that we discussed in this chapter influences the wave during the week.

- "Year" is the logical key because of the vaccination steps.

- DayOfYear was chosen because of the dependency on dates.
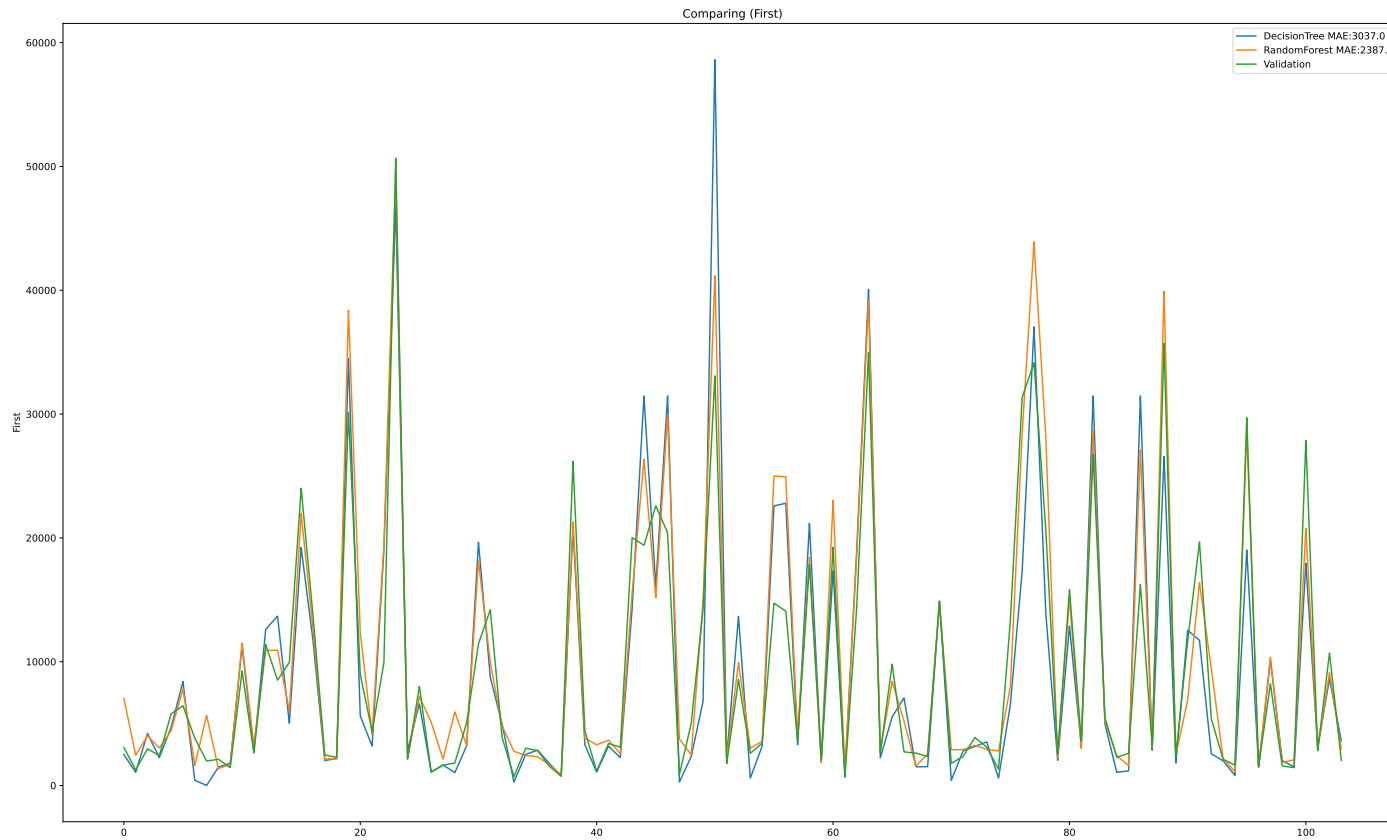
The result is better a little, but extreme values are disappointing.

Also, I suggest checking the model with columns that we discussed during the correlations search.

Comparing (First)
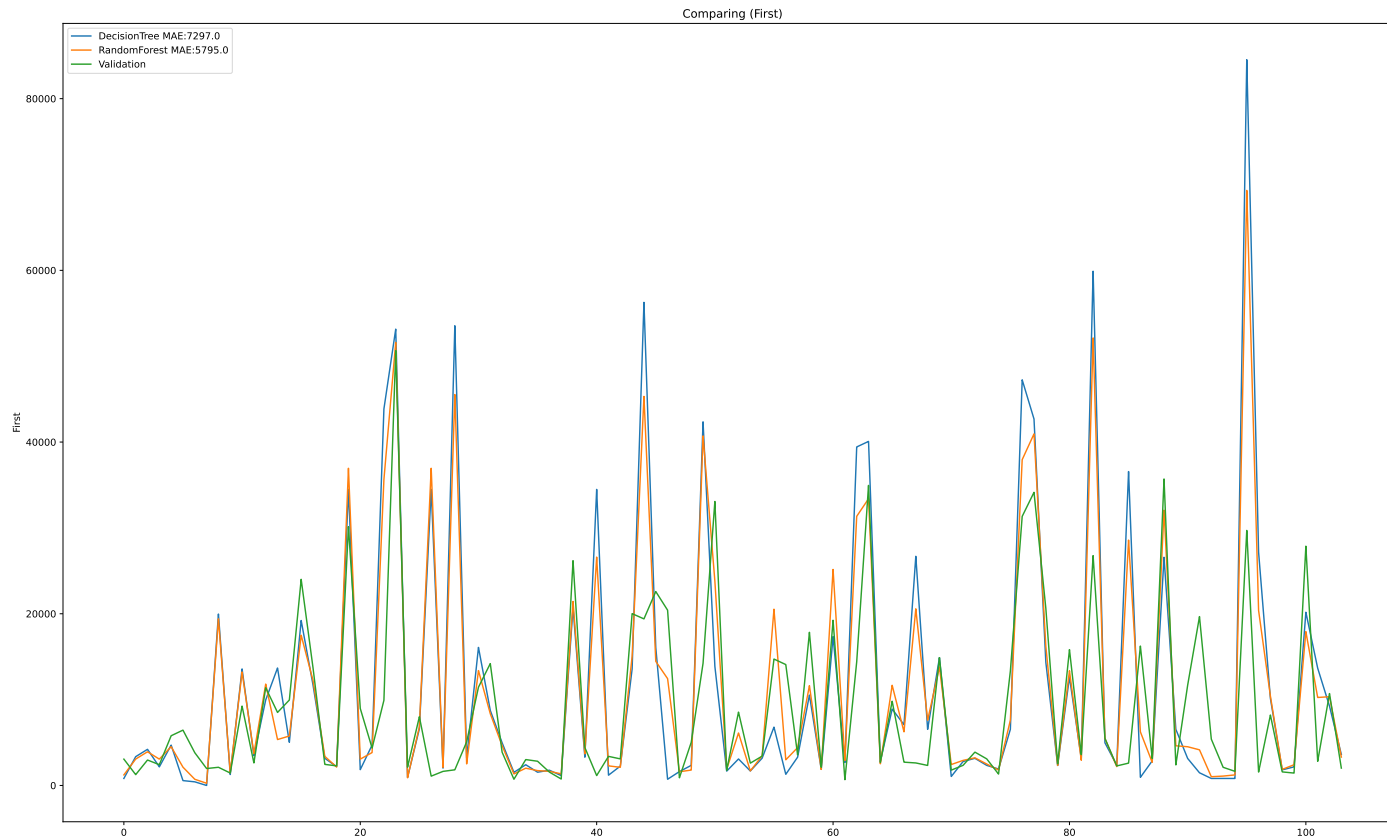
Not so good.

A combination of the following features give us the best result: Weekday, Year, DayOfYear.

We can also look at the models for the Second and Third doses. But now I suggest moving on. Maybe in the future, I will find more fitting models.
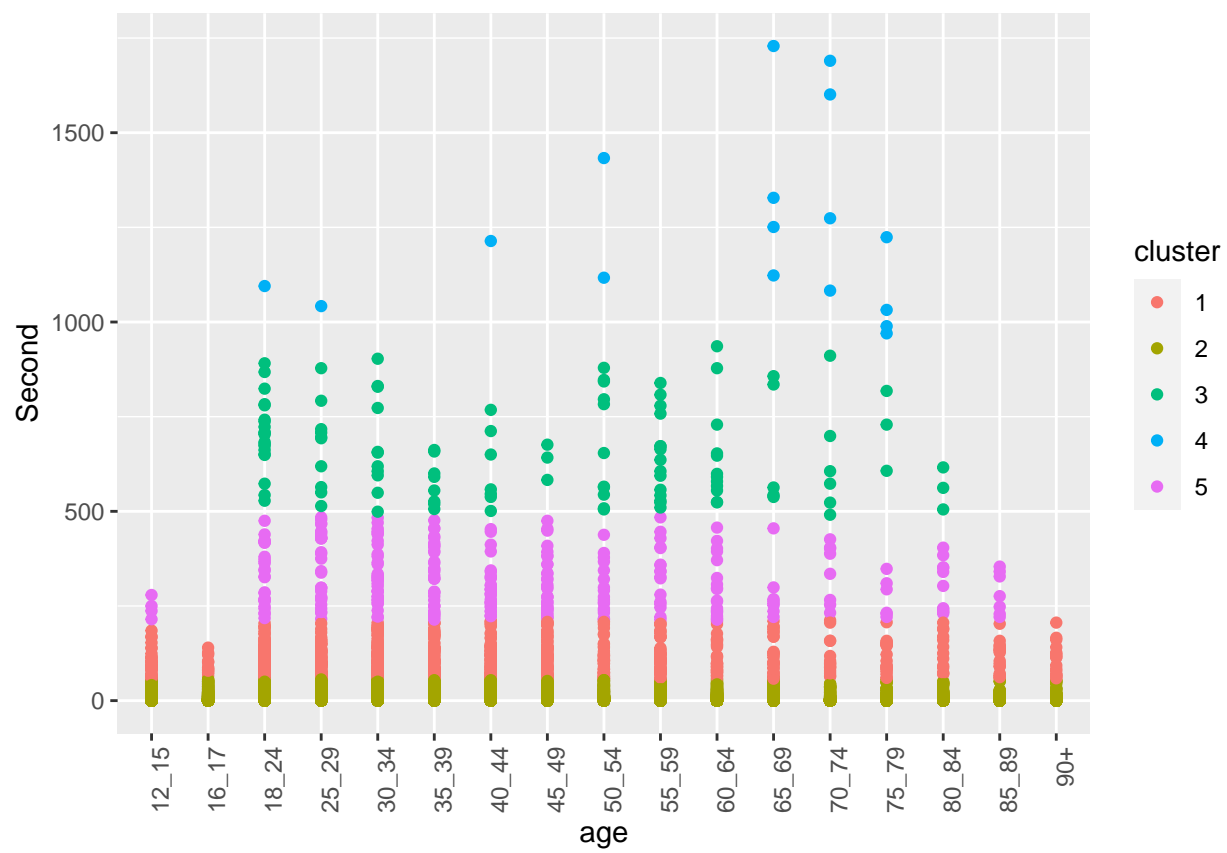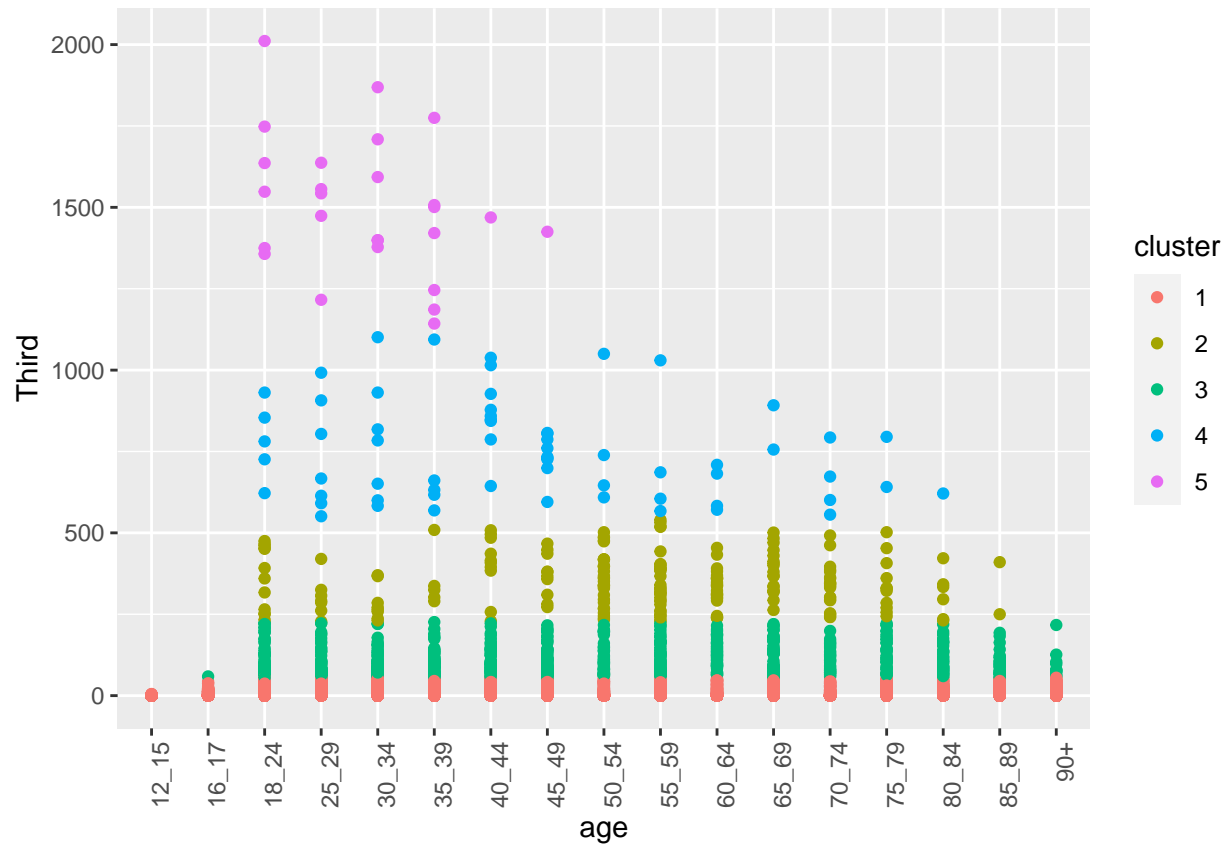
4.2 Bristol

Add the column "Age" for the model using Ordinal Encoding and look at the dataset.

age	date	First	Second	Third	Age
12__15	2022-03-24	6	34	0	1
16__17	2022-03-24	0	7	23	2
18__24	2022-03-24	13	15	30	3
25__29	2022-03-24	3	11	14	4
30__34	2022-03-24	0	5	13	5
35__39	2022-03-24	1	3	7	6

I am going to use the k-means clustering model.
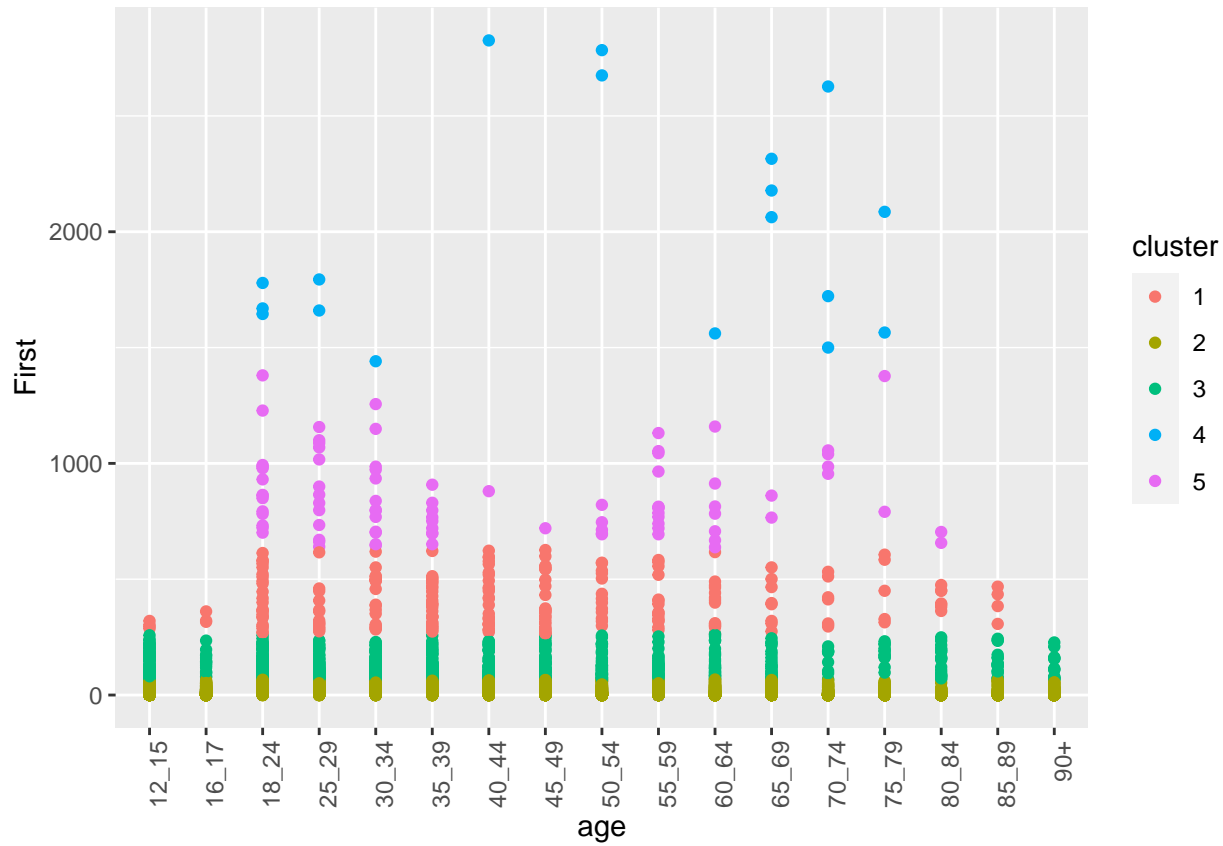
As we can see, sometimes a lot of people in the same age group got their jabs together. This fact may be useful for modeling.

4.3   England

Look at the dataset.

	Date	First	Second	Third	All
457	2020-12-08	5370	146	0	5516
456	2020-12-09	9648	137	2	9787
455	2020-12-10	11888	119	1	12008
454	2020-12-11	12516	82	1	12599
453	2020-12-12	10565	23	3	10591
452	2020-12-13	6134	30	0	6164

Look at the plot for the sum of doses counts.
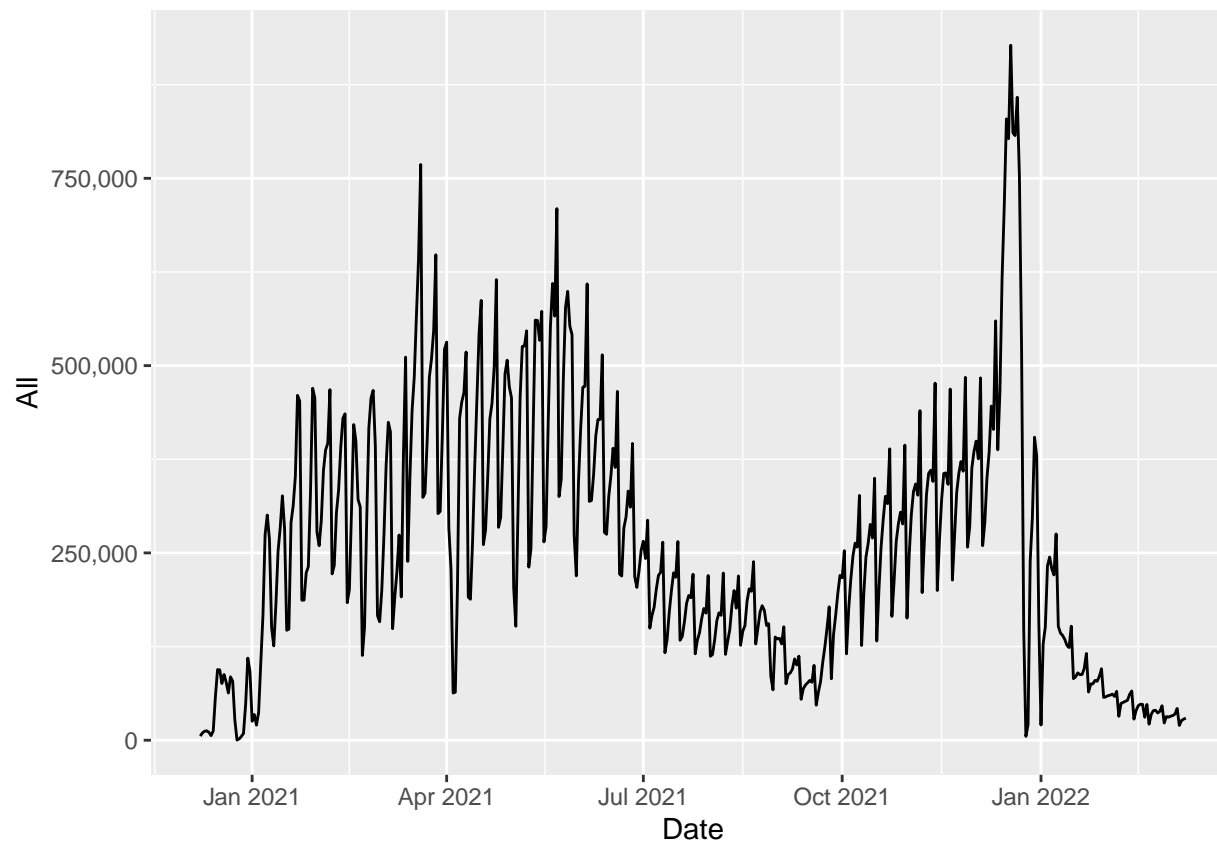
What can we say?

There is a not stationary time series, as the series wanders up and down for long periods.
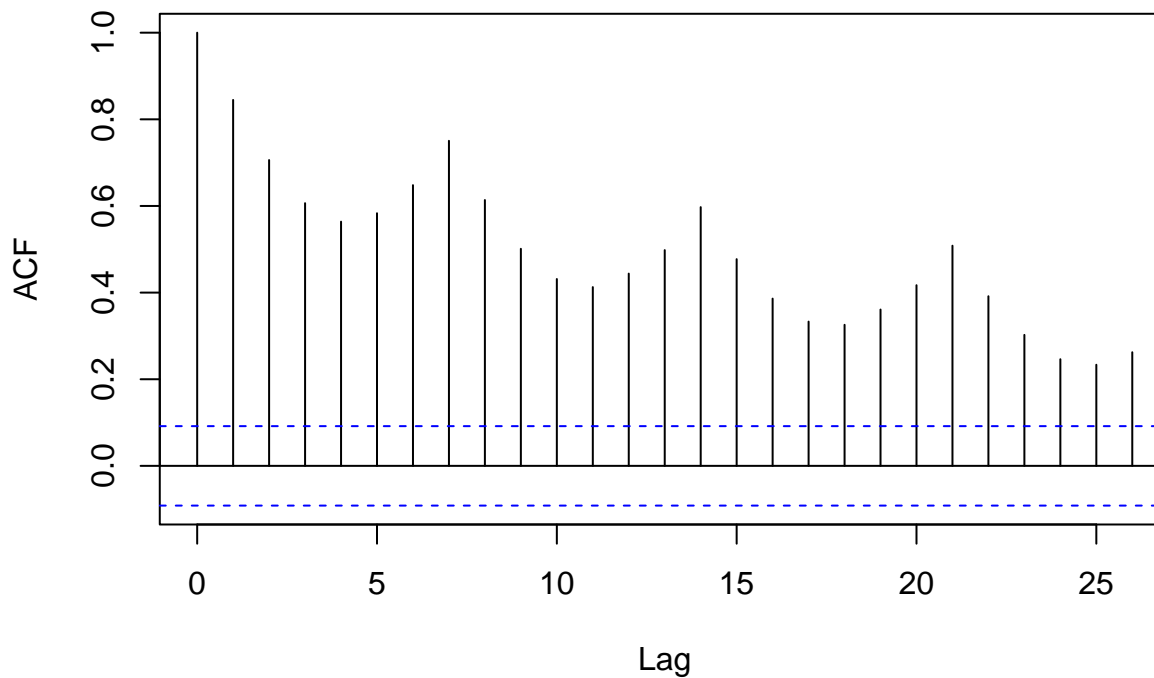
I am going to use an ARIMA model.

ARIMA(5, 1, 3)

- 5 – is the order of Auto-regressive or linear model
- 1 – difference value to make the time series stationary from non-stationary.
- 3 – is the order of Moving Average/ number of lagged values

If the data is stationary, then d=0. So, I was right earlier.

All

As we know, the autocorrelation function (ACF) assesses the correlation between observations in a time series for a set of lags. In an ACF plot, each bar represents the size and direction of the correlation. Bars that extend across the blue line are statistically significant.

So,

- this ACF plot indicates that these time series data are not random.
- the autocorrelations decline slowly.
- When a time series has both a trend and seasonality, the ACF plot displays a mixture of both effects. Notice how you can see the wavy correlations for the seasonal pattern and the slowly diminishing lags of a trend.
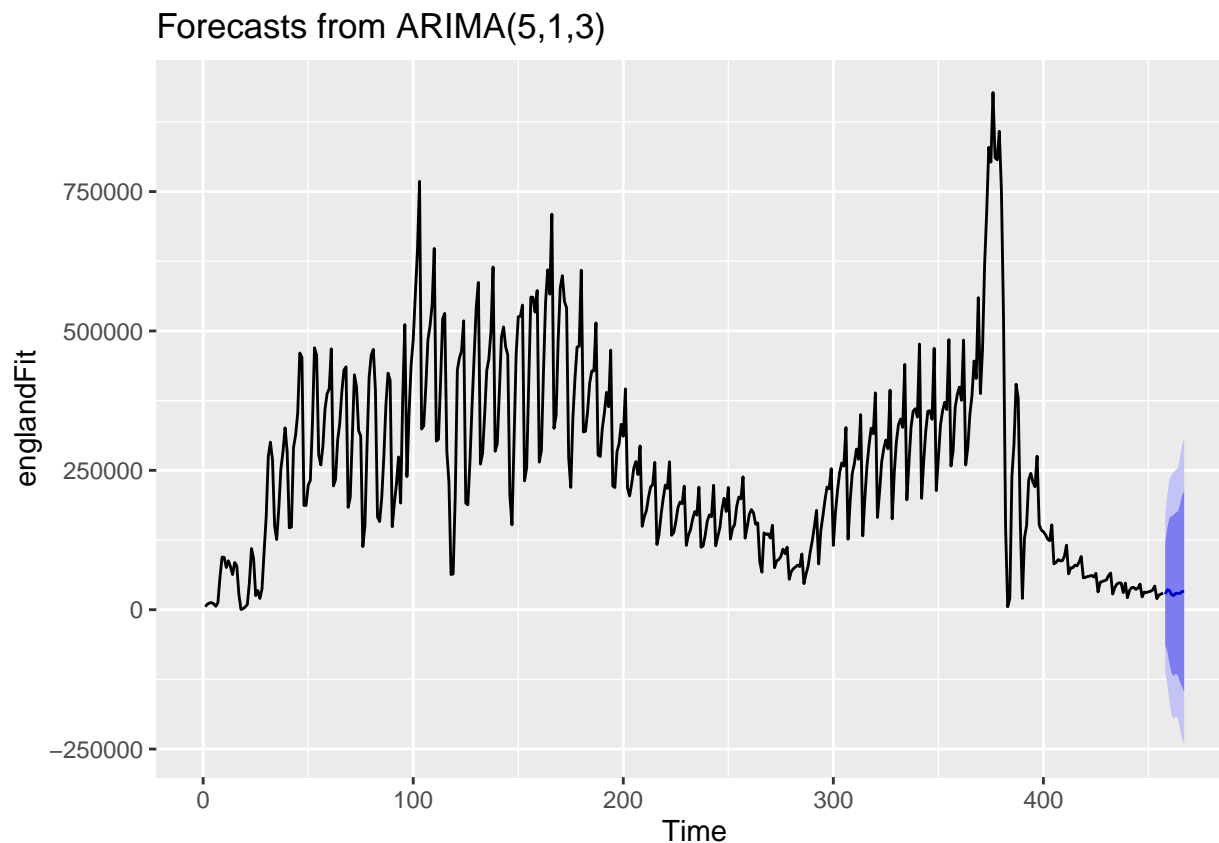
Look at the residuals that tell us a story about the model performance.

Residuals from ARIMA(5,1,3)
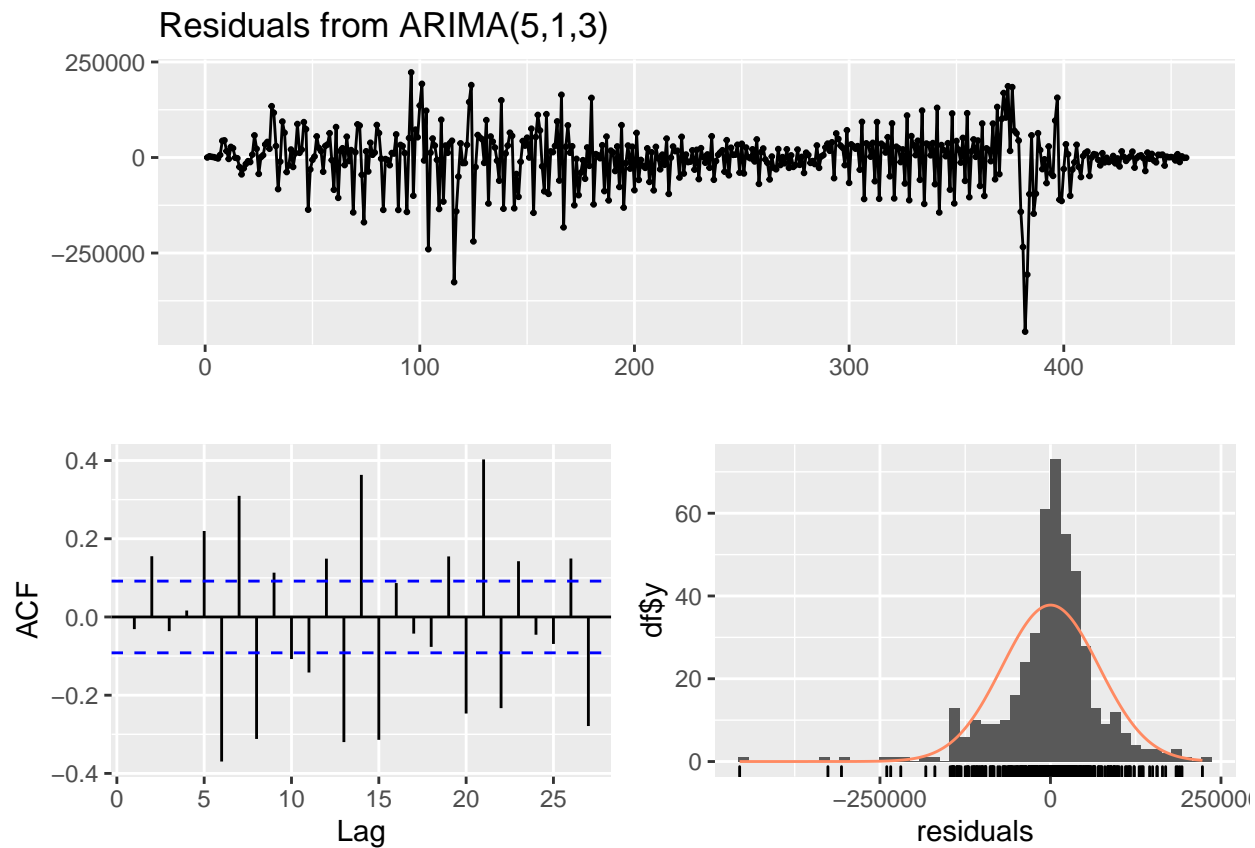
```
## 
##  Ljung-Box test
```

```
##
## data:  Residuals from ARIMA(5,1,3)
## Q* = 209.31, df = 3, p-value < 2.2e-16
##
## Model df: 8.    Total lags used: 11
```

The first plot shows us that the difference between the observations and the corresponding fitted values is not fine.

Forecasts from ARIMA(5,1,3)

I am not impressed because there are negative values. But I am sure that this may be corrected by using not auto.arima. What do I need to think about? Is how to explain to the model that the second dose is two months after the First, and the Third one is three or six months after the Second. Maybe I can find the explanatory variable for that.