Vaccination in the UK

Contents

1 Step 1 Find dataset

- Create a list of metrics for each dataset
- Look at the metrics

Lower Tier Local Authority (LTLA)

```
##  [1] "New people receiving 2nd dose"
##  [2] "New people vaccinated with a booster dose by publish date"
##  [3] "New people vaccinated complete by publish date"
##  [4] "New people fully vaccinated by vaccination date"
##  [5] "New people vaccinated 1st dose by publish date"
##  [6] "New people vaccinated with a first dose by vaccination date"
##  [7] "New people vaccinated 2nd dose by publish date"
##  [8] "New people vaccinated with a second dose by vaccination date"
##  [9] "New people vaccinated with a third dose by publish date"
## [10] "New people vaccinated with a booster dose plus new people vaccinated with a third dose by publish date"
## [11] "New people vaccinated with a booster or third dose by vaccination date"
## [12] "New vaccines given by publish date"
```

Nation

```
##  [1] "New people receiving 1st dose"
##  [2] "New people receiving 2nd dose"
##  [3] "New people vaccinated with a booster dose by publish date"
##  [4] "New people vaccinated complete by publish date"
##  [5] "New people fully vaccinated by vaccination date"
##  [6] "New people vaccinated 1st dose by publish date"
##  [7] "New people vaccinated with a first dose by vaccination date"
##  [8] "New people vaccinated 2nd dose by publish date"
##  [9] "New people vaccinated with a second dose by vaccination date"
## [10] "New people vaccinated with a third dose by publish date"
## [11] "New people vaccinated with a booster dose plus new people vaccinated with a third dose by publish date"
## [12] "New people vaccinated with a booster or third dose by vaccination date"
## [13] "New vaccines given by publish date"
```

So, as we can see, **some metrics are common**. I suggest finding out which metrics are the same for all datasets.

	ltla	msoa	nation	nhsRegion	nhsTrust	overview	region	utla
New people receiving 2nd dose	1	0	1	0	0	0	1	0
New people vaccinated with a booster dose by publish date	1	0	1	0	0	0	1	0
New people vaccinated complete by publish date	1	0	1	0	0	0	1	0
New people fully vaccinated by vaccination date	1	0	1	0	0	0	1	0
New people vaccinated 1st dose by publish date	1	0	1	0	0	0	1	0
New people vaccinated with a first dose by vaccination date	1	0	1	0	0	0	1	0
New people vaccinated 2nd dose by publish date	1	0	1	0	0	0	1	0
New people vaccinated with a second dose by vaccination date	1	0	1	0	0	0	1	0
New people vaccinated with a third dose by publish date	1	0	1	0	0	0	1	0
New people vaccinated with a booster dose plus new people vaccinated with a third dose by publish date	1	0	1	0	0	0	1	0
New people vaccinated with a booster or third dose by vaccination date	1	0	1	0	0	0	1	0
New vaccines given by publish date	1	0	1	0	0	0	1	0
New people receiving 1st dose	0	0	1	0	0	0	1	0

- Add new metrics in a common list
- Build zero-matrix, which dimension is the count of metrics x the count of area types
- Show links

Look at the result

First of all, I am interested in data about the **first jab**. **So, I need to look at the datasets:**

- Build zero-matrix, which dimension is the count of metrics x the count of area types

```
## [1] "Lower Tier Local Authority (LTLA)"
## [1] "Nation"
## [1] "Region"
```

2 Step 2 Ask something

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.
- The city rests every summer when students come back to their homes and works hardly elsewhen.

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? Does the vaccination depend on holidays?

I was vaccinated by

- the first dose on 8 August 2021,
- the second dose on 3 October 2021,
- the booster dose on 8 January 2022.

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: Is there something illogical in data?

3 Step 3 Look at the datasets

3.1 Region

As we can see on the website, Region metrics are available for regions of England. I am interested in the South West and metrics that start with "New":

```
## [1] "areaCode"
## [2] "areaName"
## [3] "areaType"
## [4] "date"
## [5] "newPeopleVaccinatedFirstDoseByVaccinationDate"
## [6] "newPeopleVaccinatedSecondDoseByVaccinationDate"
## [7] "newPeopleVaccinatedThirdInjectionByVaccinationDate"
```

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

areaCode

```
## [1] "E12000009"
```

areaName

```
## [1] "South West"
```

areaType

```
## [1] "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 and columns
- Create long table

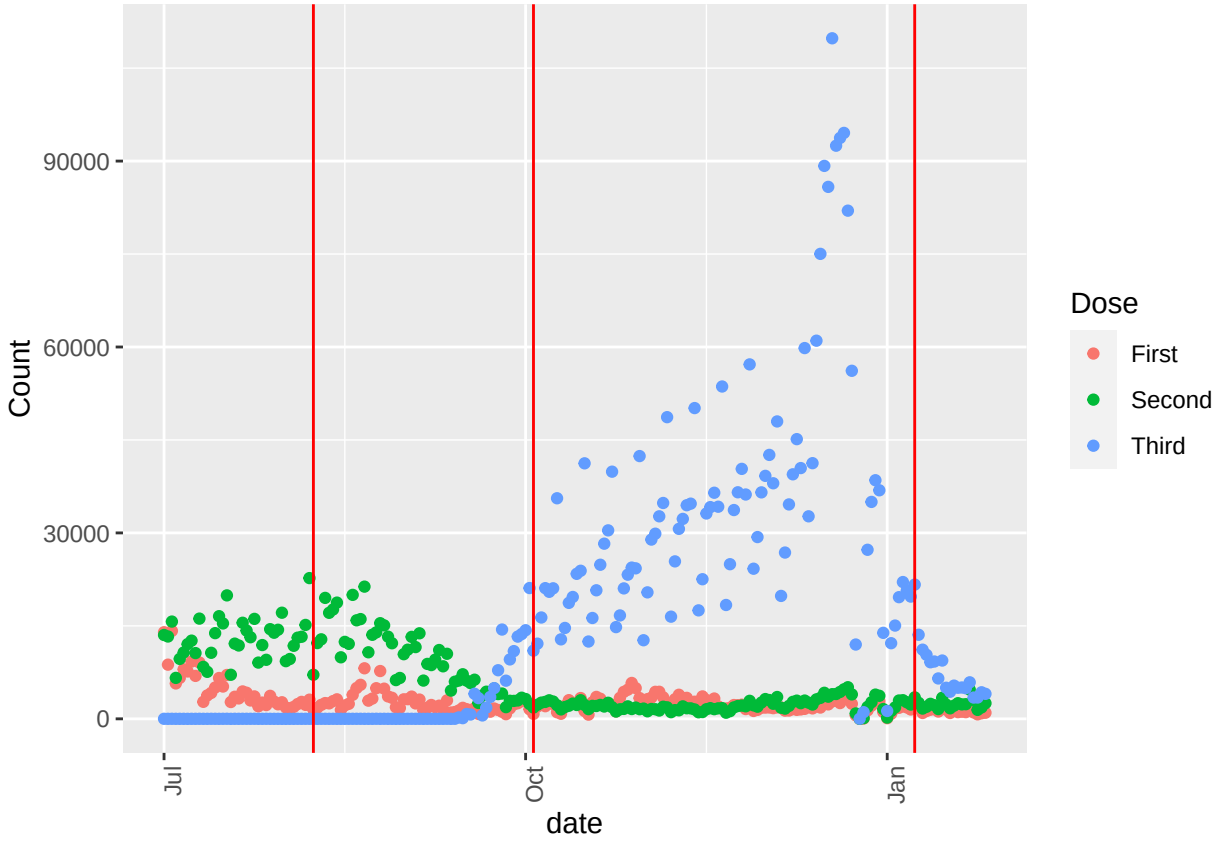
Let's plot something.

areaCode	areaName	areaType	date	First	Second	Third	MonthYear
E12000009	South West	region	2022-01-26	986	2520	4034	1.2022
E12000009	South West	region	2022-01-25	899	1845	4283	1.2022
E12000009	South West	region	2022-01-24	723	1445	3441	1.2022
E12000009	South West	region	2022-01-23	1035	3007	3439	1.2022
E12000009	South West	region	2022-01-22	1822	4709	5896	1.2022
E12000009	South West	region	2022-01-21	1085	2362	4944	1.2022

date	MonthYear	Dose	Count
2022-01-26	1.2022	First	986
2022-01-25	1.2022	First	899
2022-01-24	1.2022	First	723
2022-01-23	1.2022	First	1035
2022-01-22	1.2022	First	1822
2022-01-21	1.2022	First	1085

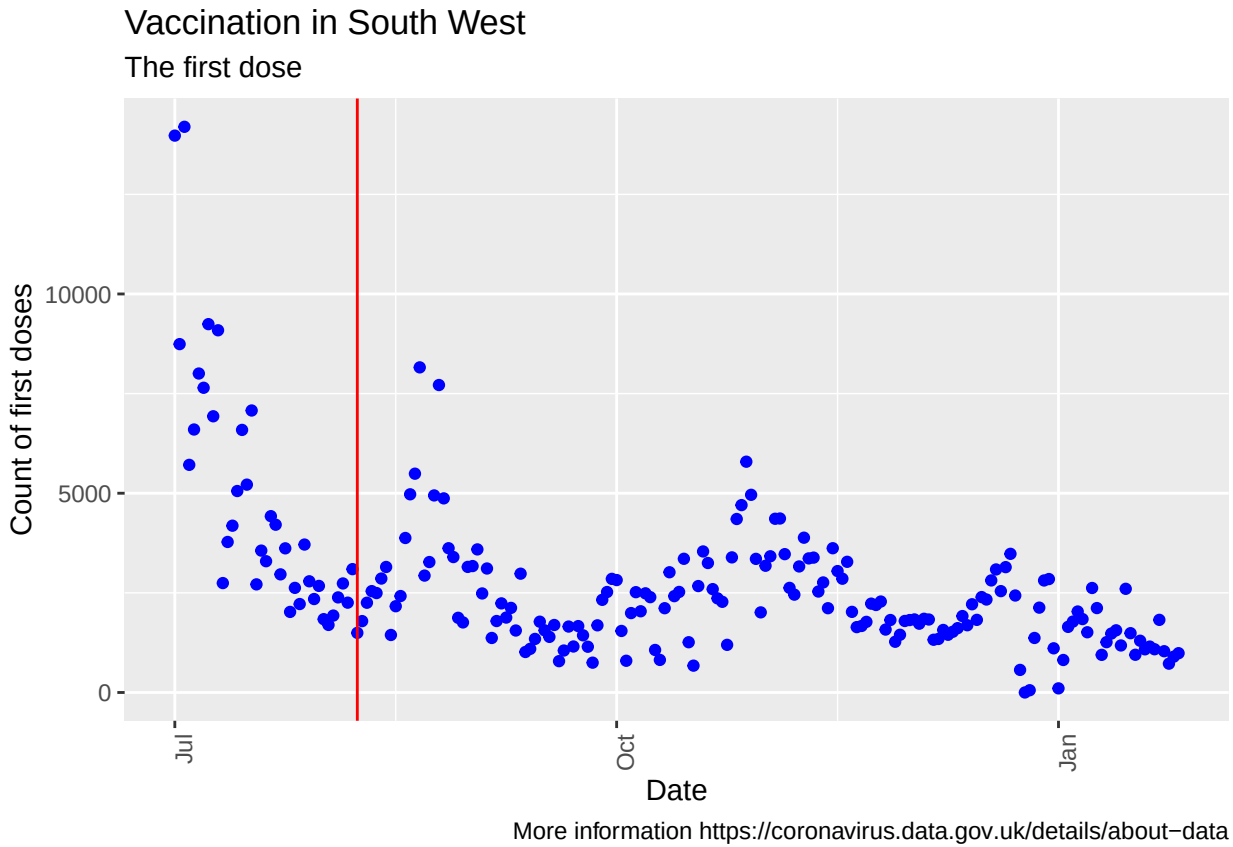
3.1.1 Question 0

Are there dependencies between academic year events and vaccination waves? Does the vaccination depend on holidays?

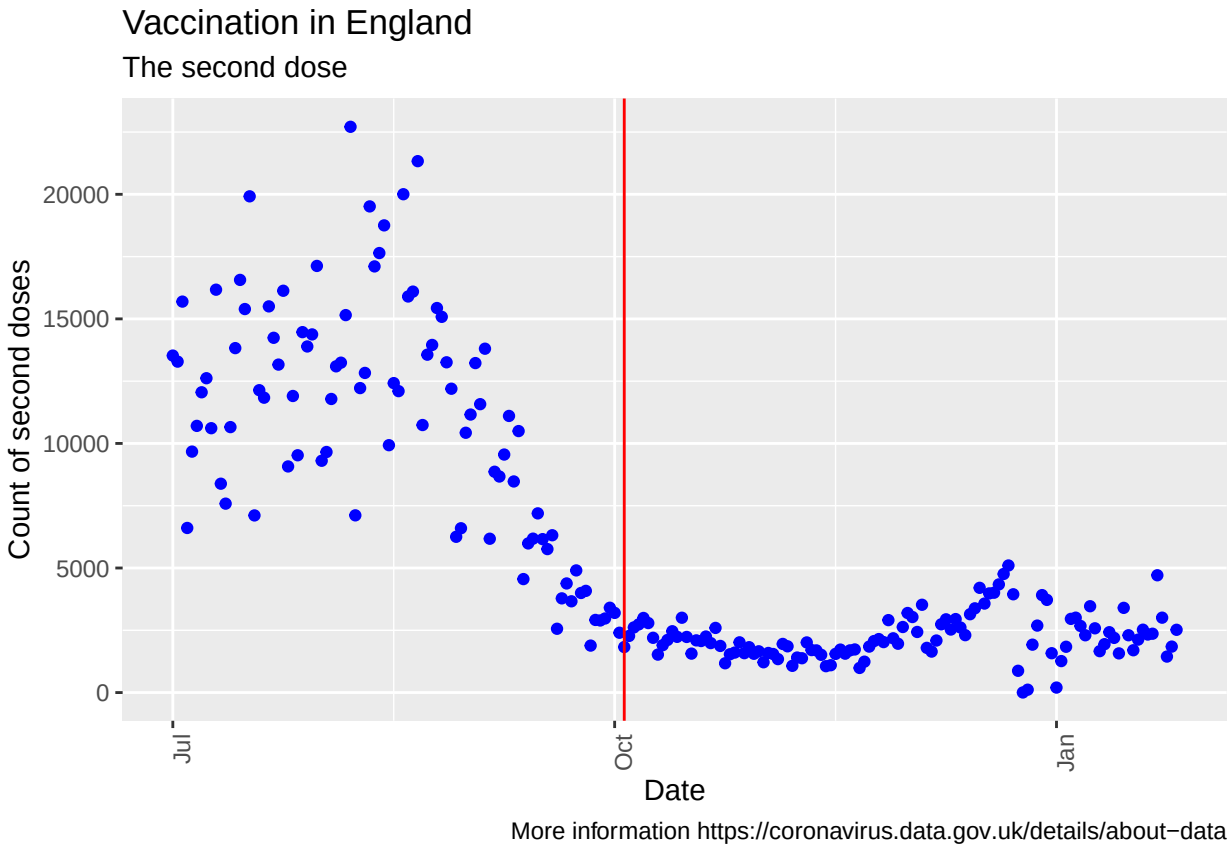


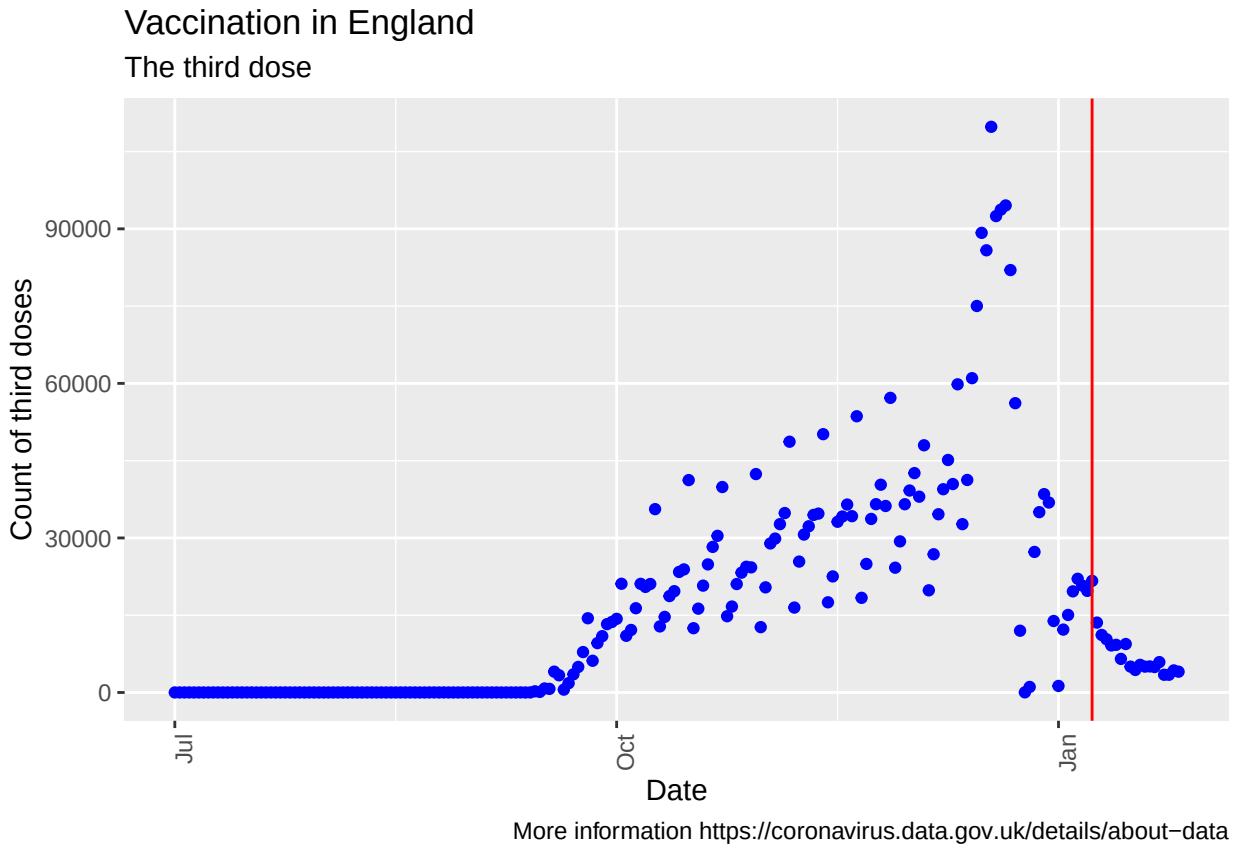
The result is not beautiful because of the active growth of the third jabs count at the end of 2021.

Let's plot them separately.



It is so interesting why the graph is wavy.

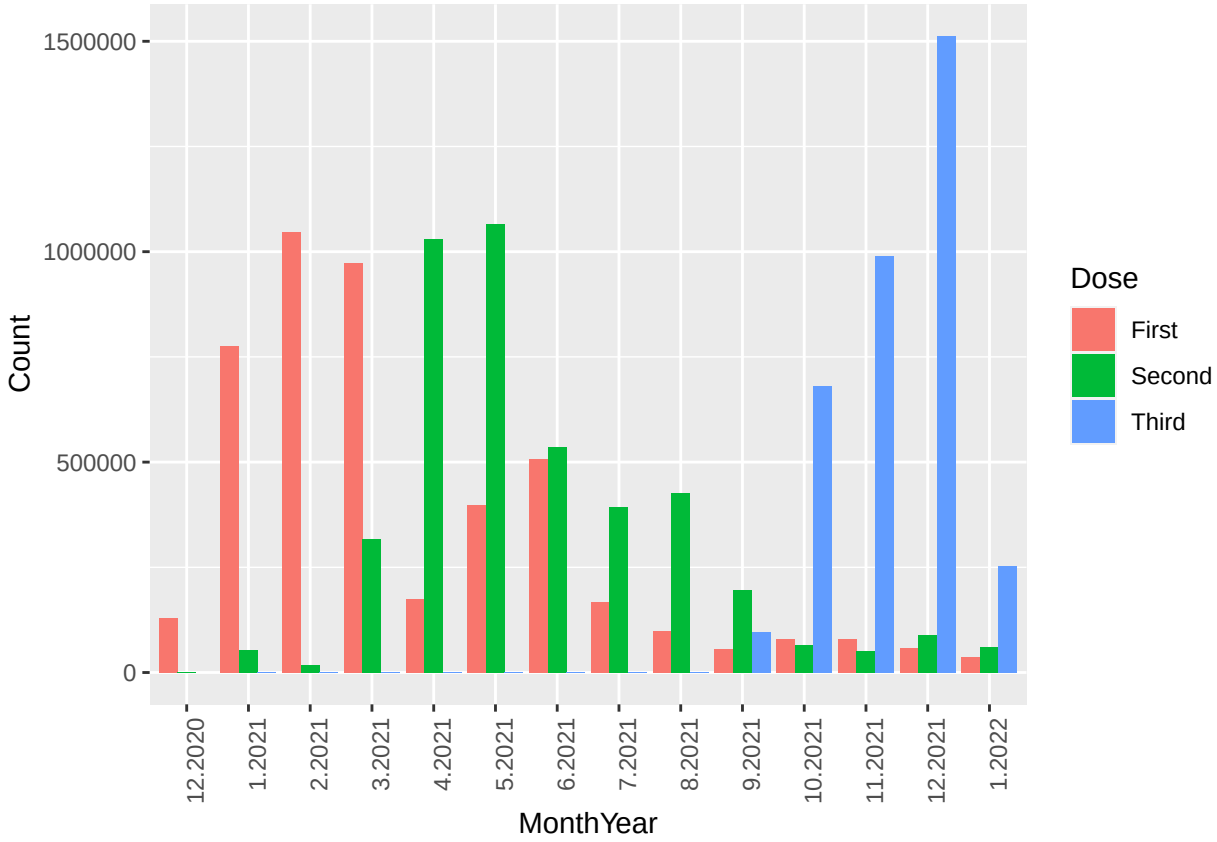

Vaccination in England
The second dose
Count of second doses
20000
15000
10000
5000
0
Jul
Oct
Jan
Date
More information https://coronavirus.data.gov.uk/details/about−data



We can see when the active phase of vaccination by the third dose started.

Let's calculate the date.

```
## Warning: Removed 1 rows containing missing values (geom_col).
```


1500000
1000000
500000
0
Count
12.2020
1.2021
2.2021
3.2021
4.2021
5.2021
6.2021
7.2021
8.2021
9.2021
10.2021
11.2021
12.2021
1.2022
MonthYear
Dose
First
Second
Third

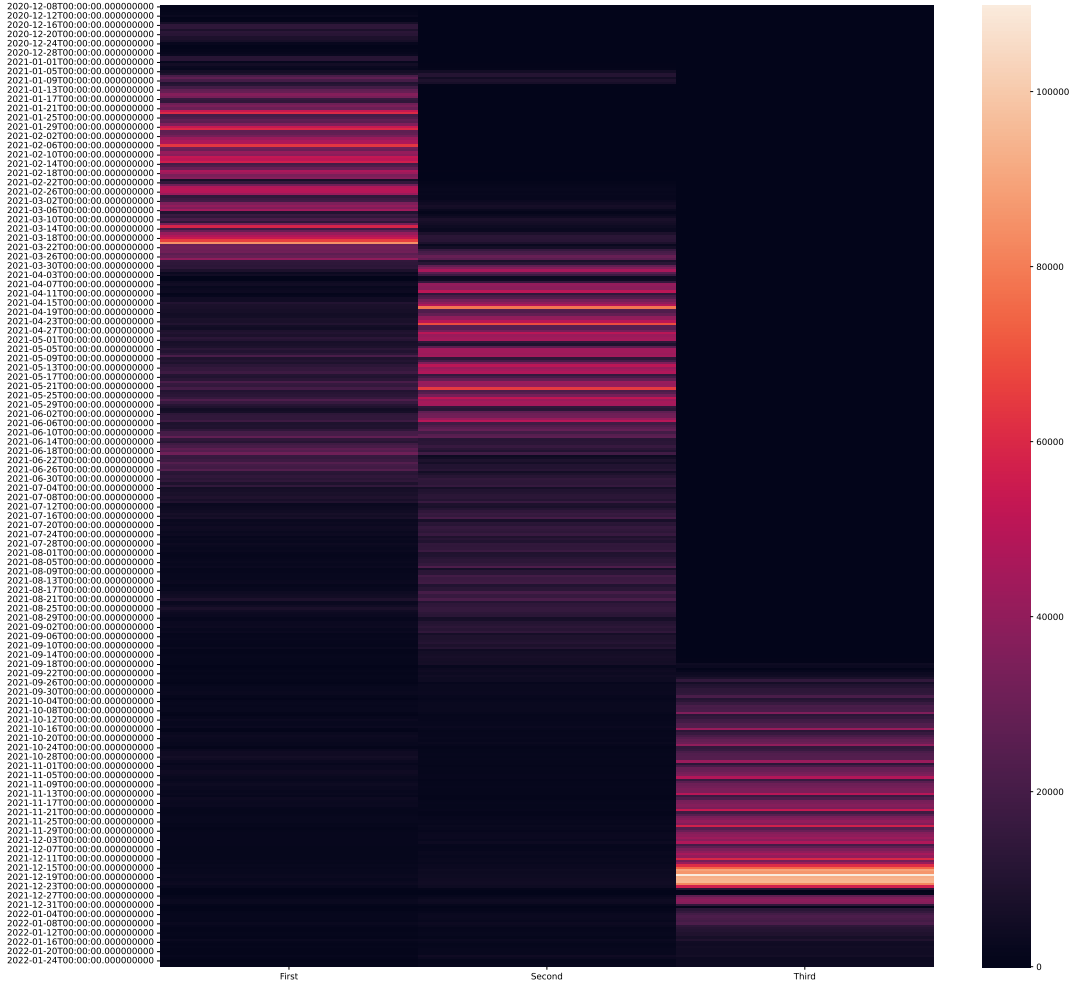
4 Step 4 Machine learning

4.1 Step 0: Look at and Modify the dataset

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

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


2020-12-08T00:00:00.000000000
2020-12-12T00:00:00.000000000
2020-12-16T00:00:00.000000000
2020-12-20T00:00:00.000000000
2020-12-24T00:00:00.000000000
2020-12-28T00:00:00.000000000
2021-01-01T00:00:00.000000000
2021-01-05T00:00:00.000000000
2021-01-09T00:00:00.000000000
2021-01-13T00:00:00.000000000
2021-01-17T00:00:00.000000000
2021-01-21T00:00:00.000000000
2021-01-25T00:00:00.000000000
2021-01-29T00:00:00.000000000
2021-02-02T00:00:00.000000000
2021-02-06T00:00:00.000000000
2021-02-10T00:00:00.000000000
2021-02-14T00:00:00.000000000
2021-02-18T00:00:00.000000000
2021-02-22T00:00:00.000000000
2021-02-26T00:00:00.000000000
2021-03-02T00:00:00.000000000
2021-03-06T00:00:00.000000000
2021-03-10T00:00:00.000000000
2021-03-14T00:00:00.000000000
2021-03-18T00:00:00.000000000
2021-03-22T00:00:00.000000000
2021-03-26T00:00:00.000000000
2021-03-30T00:00:00.000000000
2021-04-03T00:00:00.000000000
2021-04-07T00:00:00.000000000
2021-04-11T00:00:00.000000000
2021-04-15T00:00:00.000000000
2021-04-19T00:00:00.000000000
2021-04-23T00:00:00.000000000
2021-04-27T00:00:00.000000000
2021-05-01T00:00:00.000000000
2021-05-05T00:00:00.000000000
2021-05-09T00:00:00.000000000
2021-05-13T00:00:00.000000000
2021-05-17T00:00:00.000000000
2021-05-21T00:00:00.000000000
2021-05-25T00:00:00.000000000
2021-05-29T00:00:00.000000000
2021-06-02T00:00:00.000000000
2021-06-06T00:00:00.000000000
2021-06-10T00:00:00.000000000
2021-06-14T00:00:00.000000000
2021-06-18T00:00:00.000000000
2021-06-22T00:00:00.000000000
2021-06-26T00:00:00.000000000
2021-06-30T00:00:00.000000000
2021-07-04T00:00:00.000000000
2021-07-08T00:00:00.000000000
2021-07-12T00:00:00.000000000
2021-07-16T00:00:00.000000000
2021-07-20T00:00:00.000000000
2021-07-24T00:00:00.000000000
2021-07-28T00:00:00.000000000
2021-08-01T00:00:00.000000000
2021-08-05T00:00:00.000000000
2021-08-09T00:00:00.000000000
2021-08-13T00:00:00.000000000
2021-08-17T00:00:00.000000000
2021-08-21T00:00:00.000000000
2021-08-25T00:00:00.000000000
2021-08-29T00:00:00.000000000
2021-09-02T00:00:00.000000000
2021-09-06T00:00:00.000000000
2021-09-10T00:00:00.000000000
2021-09-14T00:00:00.000000000
2021-09-18T00:00:00.000000000
2021-09-22T00:00:00.000000000
2021-09-26T00:00:00.000000000
2021-09-30T00:00:00.000000000
2021-10-04T00:00:00.000000000
2021-10-08T00:00:00.000000000
2021-10-12T00:00:00.000000000
2021-10-16T00:00:00.000000000
2021-10-20T00:00:00.000000000
2021-10-24T00:00:00.000000000
2021-10-28T00:00:00.000000000
2021-11-01T00:00:00.000000000
2021-11-05T00:00:00.000000000
2021-11-09T00:00:00.000000000
2021-11-13T00:00:00.000000000
2021-11-17T00:00:00.000000000
2021-11-21T00:00:00.000000000
2021-11-25T00:00:00.000000000
2021-11-29T00:00:00.000000000
2021-12-03T00:00:00.000000000
2021-12-07T00:00:00.000000000
2021-12-11T00:00:00.000000000
2021-12-15T00:00:00.000000000
2021-12-19T00:00:00.000000000
2021-12-23T00:00:00.000000000
2021-12-27T00:00:00.000000000
2021-12-31T00:00:00.000000000
2022-01-04T00:00:00.000000000
2022-01-08T00:00:00.000000000
2022-01-12T00:00:00.000000000
2022-01-16T00:00:00.000000000
2022-01-20T00:00:00.000000000
2022-01-24T00:00:00.000000000
First
Second
Third
100000
80000
60000
40000
20000
0

As we can see, there are waves. So, the count of jabs depends on dates.

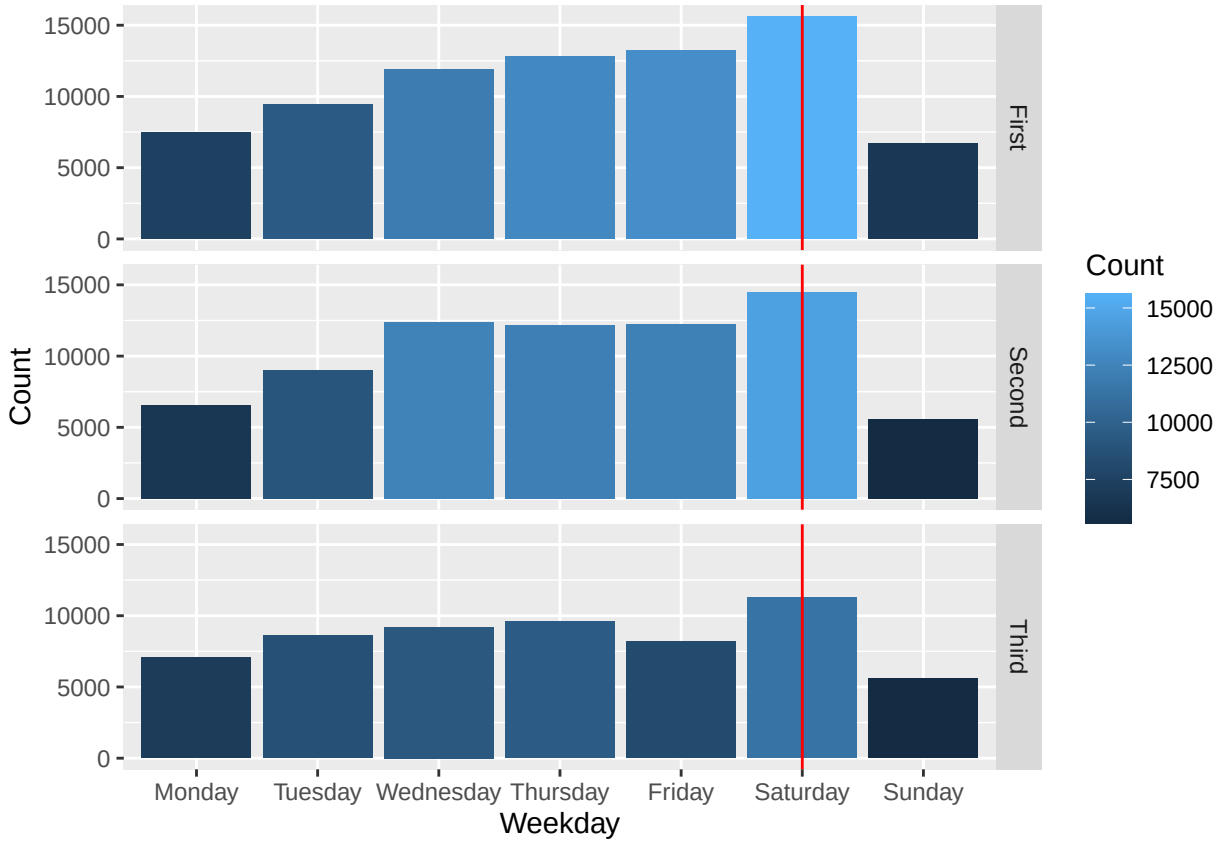
Let's get features: 1) Year 2) Month 3) Day etc.

	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

4.2 Step 1: Explore the dataset

4.2.1 Weekdays

As you remember, I have a question. When do people prefer to get a jab: weekdays or weekends/Saturdays or Sundays?



Let's answer.

So, most of South West's people prefer to get a jab on Saturdays.

4.2.2 Missing values

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.

4.3 Step 2: Split sets, train a Machine Learning Model and Evaluate performance

Define necessary variables

Prepare sets and train models using parameters.

Compare the score with the mean value of the column that we predicted.

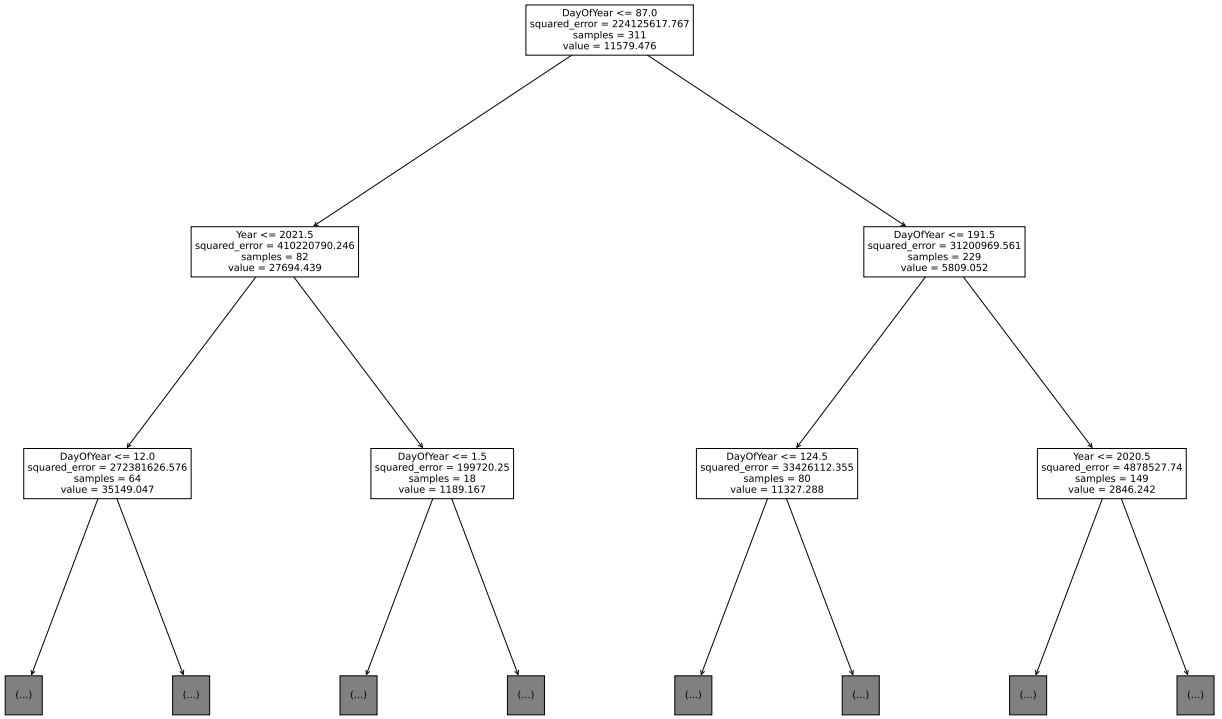
```
## 0.719657929335243
```

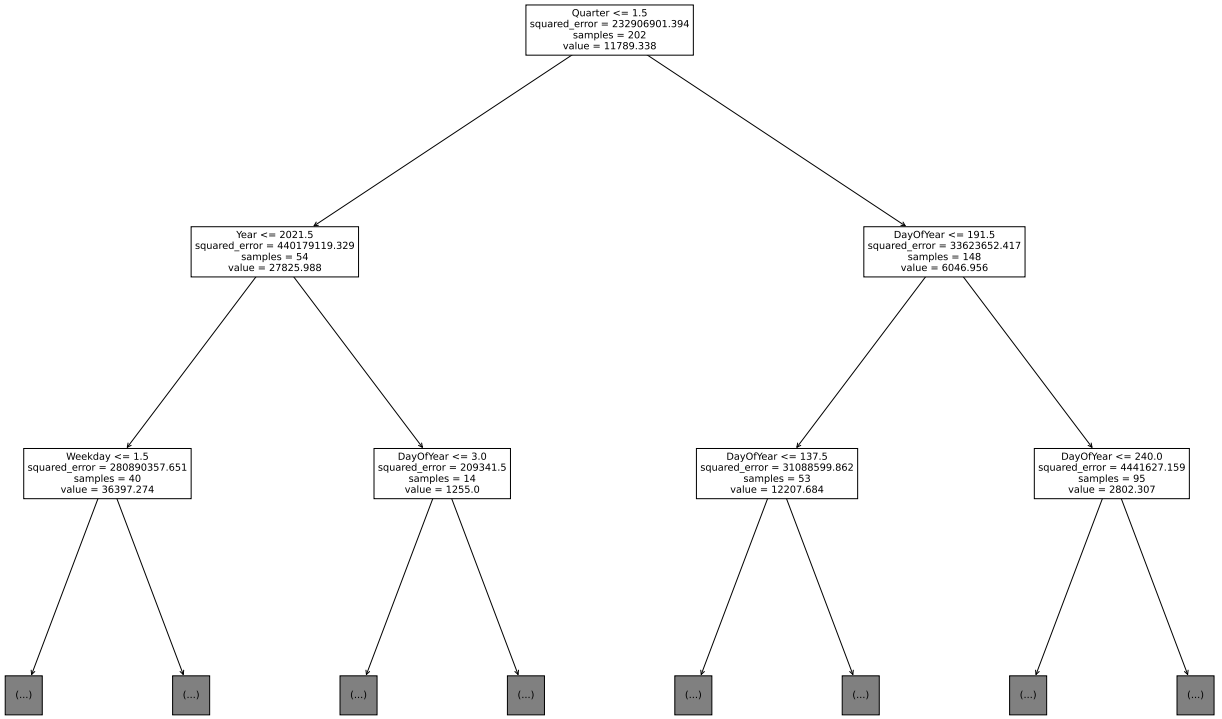
```
## 0.774580856609961
```

```
## 0.7445280765098062
```

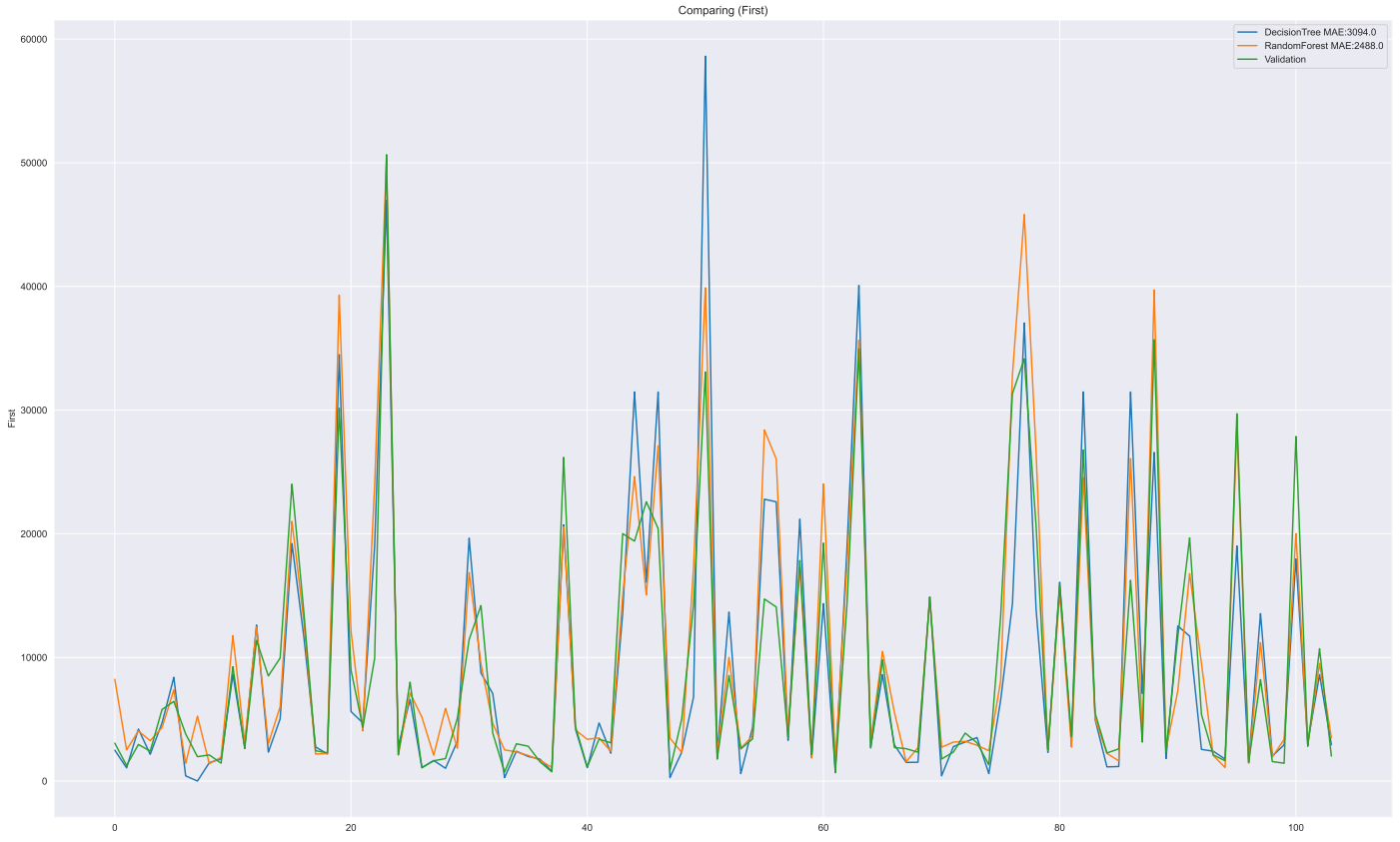
```
## 0.8144473412230163
```

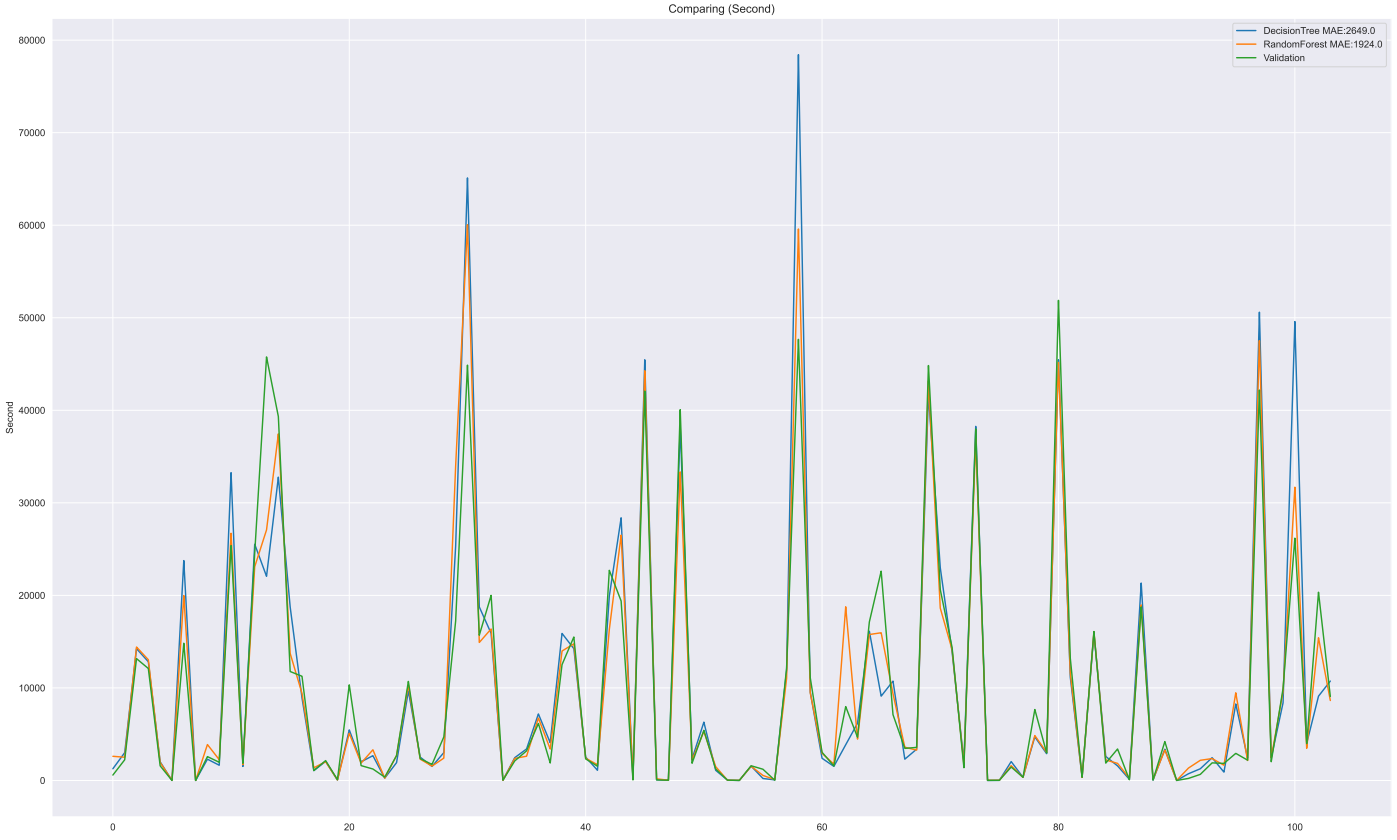
Look at the tree


DayOfYear <= 87.0
squared_error = 224125617.767
samples = 311
value = 11579.476
Year <= 2021.5
squared_error = 410220790.246
samples = 82
value = 27694.439
DayOfYear <= 191.5
squared_error = 31200969.561
samples = 229
value = 5809.052
DayOfYear <= 12.0
squared_error = 272381626.576
samples = 64
value = 35149.047
DayOfYear <= 1.5
squared_error = 199720.25
samples = 18
value = 1189.167
DayOfYear <= 124.5
squared_error = 33426112.355
samples = 80
value = 11327.288
Year <= 2020.5
squared_error = 4878527.74
samples = 149
value = 2846.242
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)


Quarter <= 1.5
squared_error = 232906901.394
samples = 202
value = 11789.338
Year <= 2021.5
squared_error = 440179119.329
samples = 54
value = 27825.988
DayOfYear <= 191.5
squared_error = 33623652.417
samples = 148
value = 6046.956
Weekday <= 1.5
squared_error = 280890357.651
samples = 40
value = 36397.274
DayOfYear <= 3.0
squared_error = 209341.5
samples = 14
value = 1255.0
DayOfYear <= 137.5
squared_error = 31088599.862
samples = 53
value = 12207.684
DayOfYear <= 240.0
squared_error = 4441627.159
samples = 95
value = 2802.307
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

4.4 Step 3: Plot results




Comparing (Second)
DecisionTree MAE:2649.0
RandomForest MAE:1924.0
Validation
Second
80000
70000
60000
50000
40000
30000
20000
10000
0
0
20
40
60
80
100

4.5 Step 4: Improve models by changing the dataset

I am going to work with features.

Define necessary variables

Prepare sets and Train models

Compare the score with the mean value of the column that we predicted.

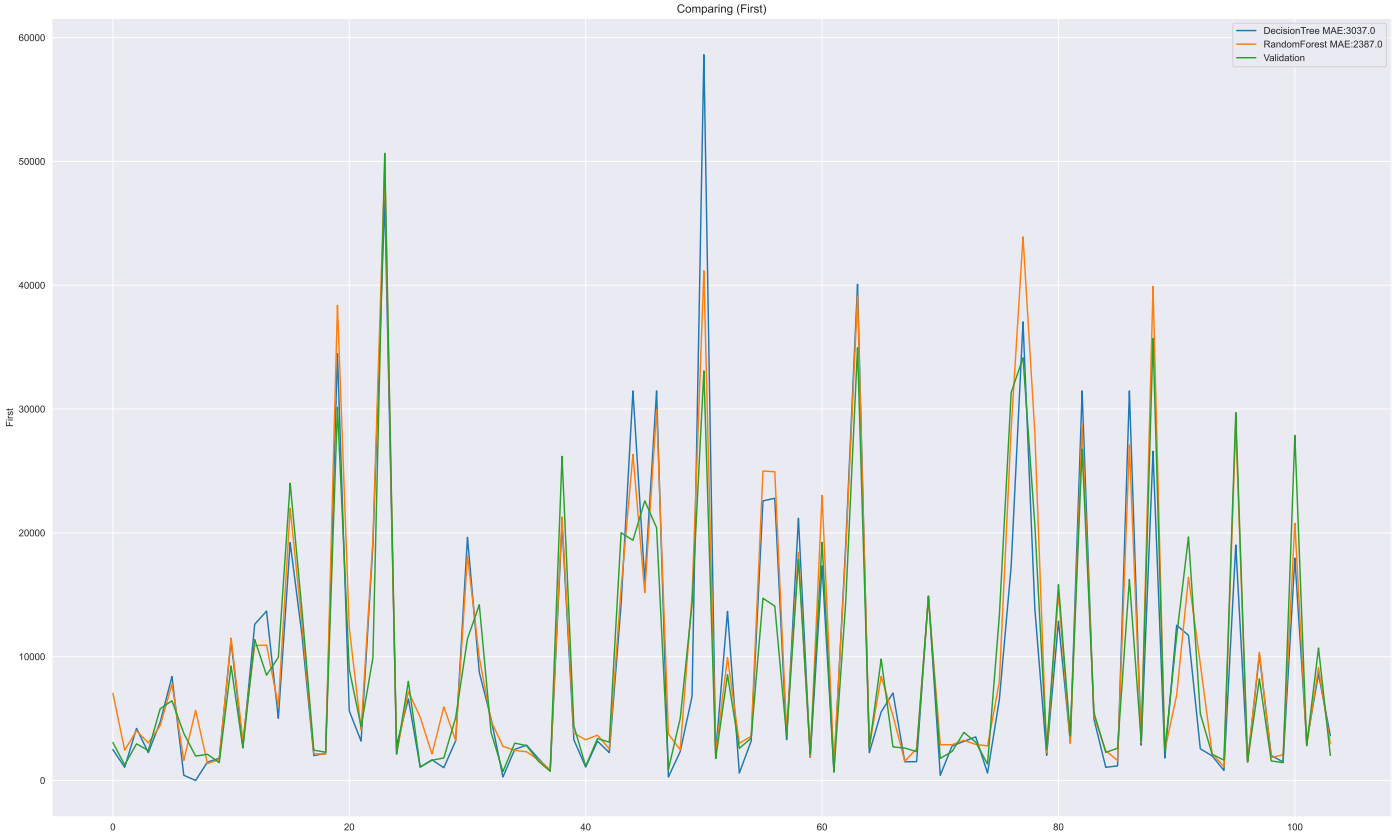
```
## 0.7248630326024768
```

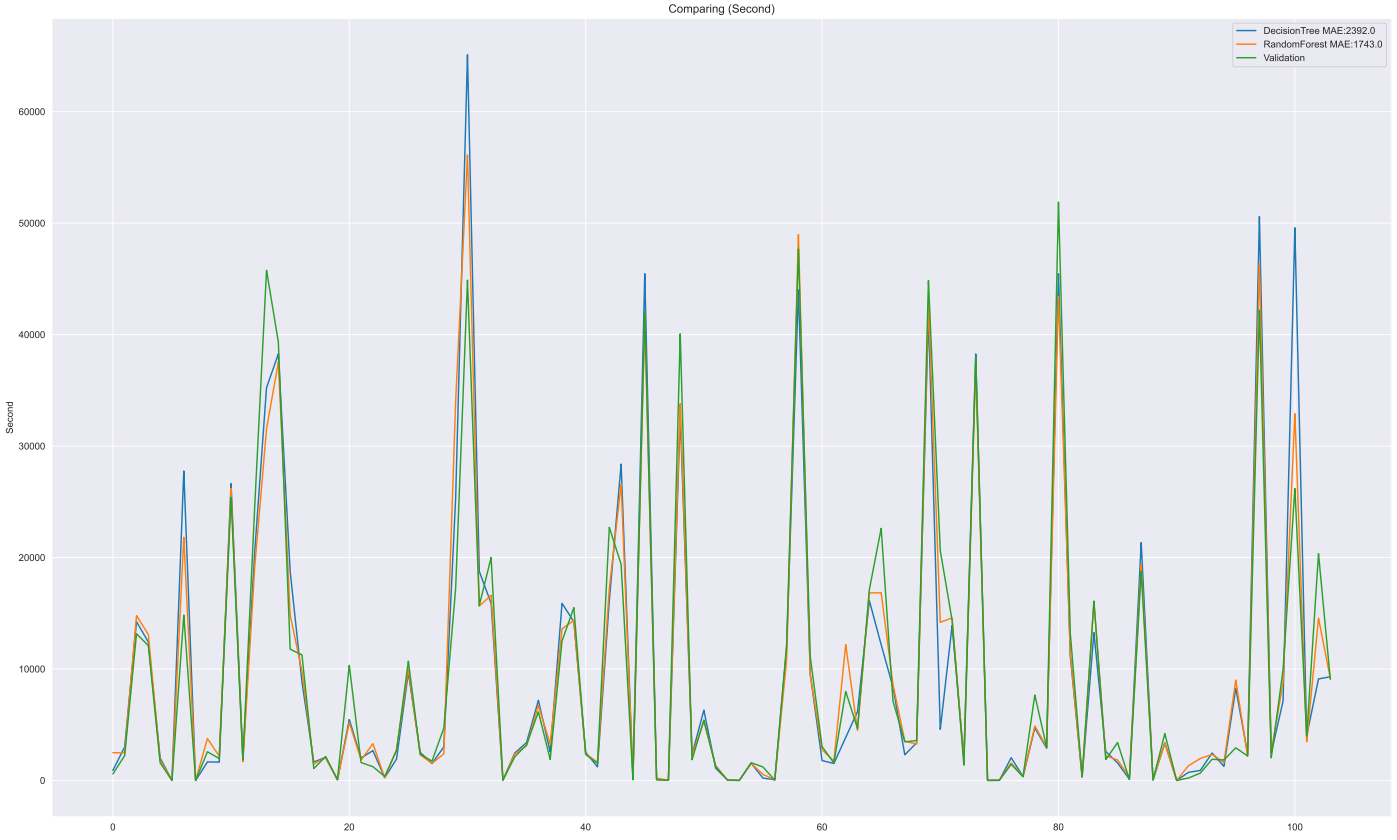
```
## 0.7837038702898657
```

```
## 0.7692636418171874
```

```
## 0.8318561653086721
```

Plot the result.


Comparing (First)
DecisionTree MAE:3037.0
RandomForest MAE:2387.0
Validation
First
60000
50000
40000
30000
20000
10000
0
0
20
40
60
80
100



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