Contents

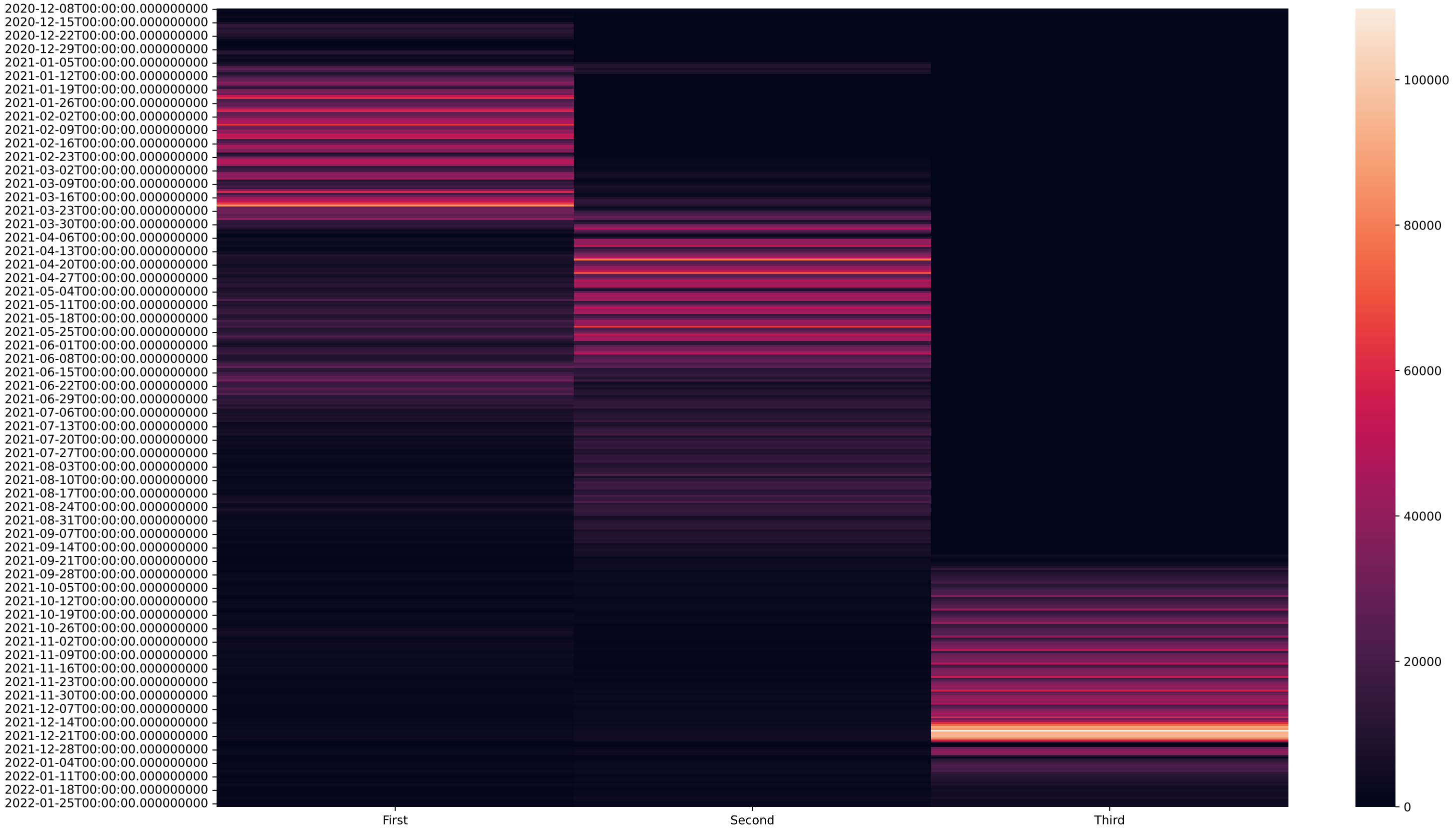
1 Step 4 Machine learning

1.1 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
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 **??**, 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.

1.2 Step 1: Explore the dataset

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

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.

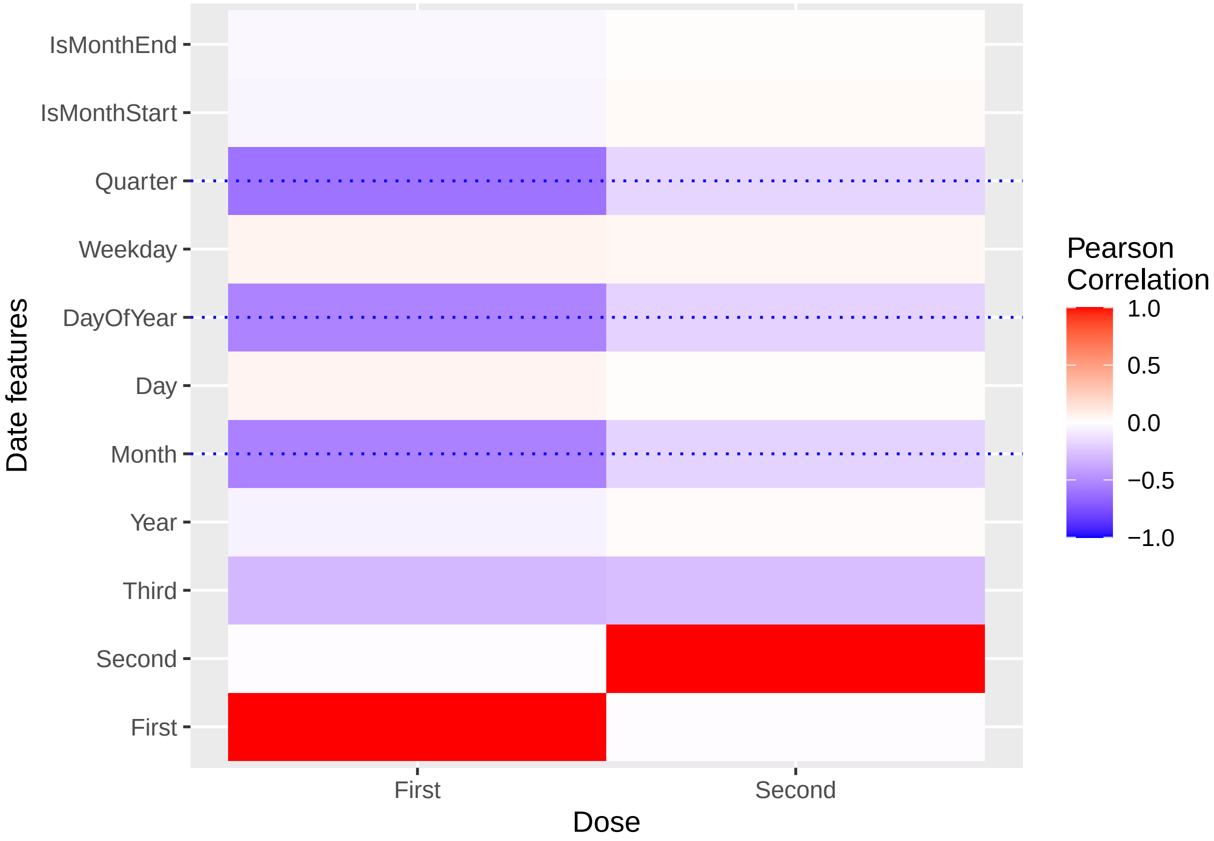
First	double
Second	double
Third	double
Year	double
Month	double
Day	double
DayOfYear	double
Weekday	double
Quarter	double
IsMonthStart	logical
IsMonthEnd	logical

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.

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.

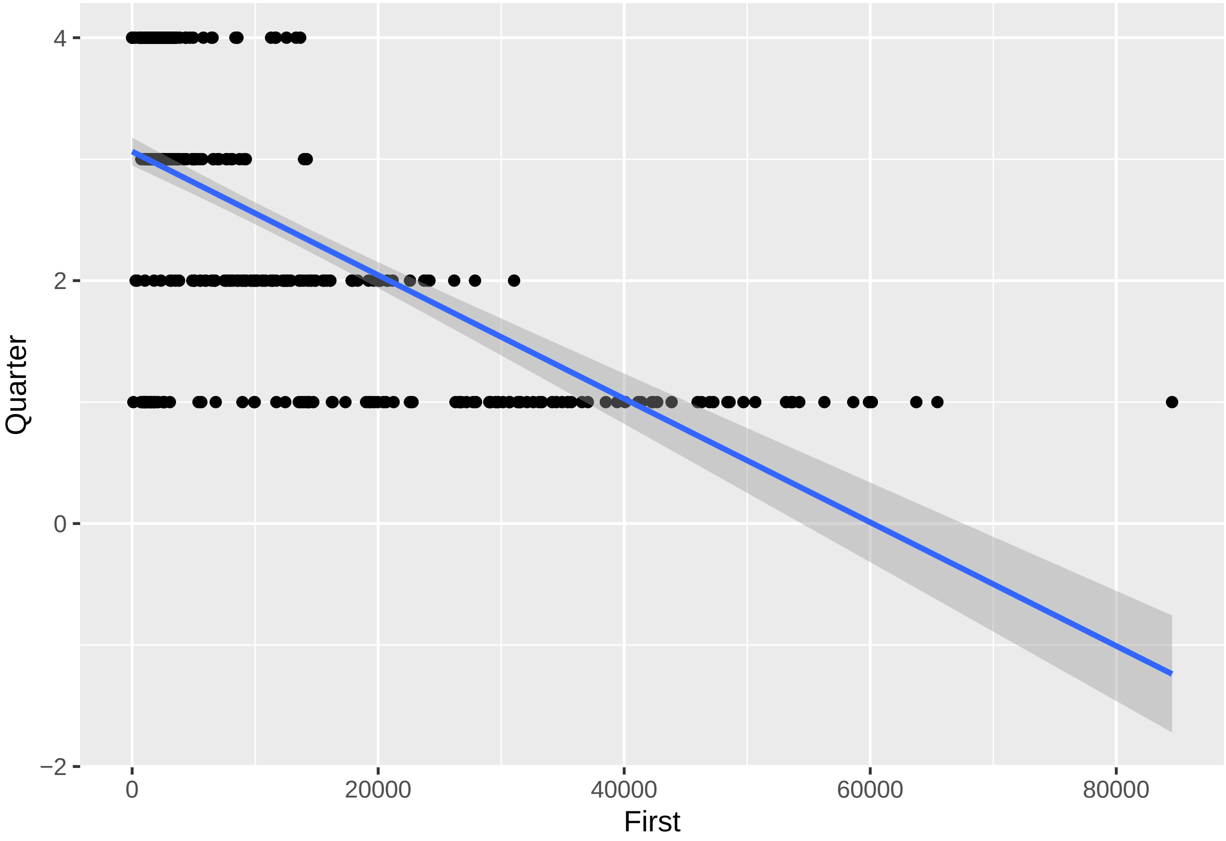


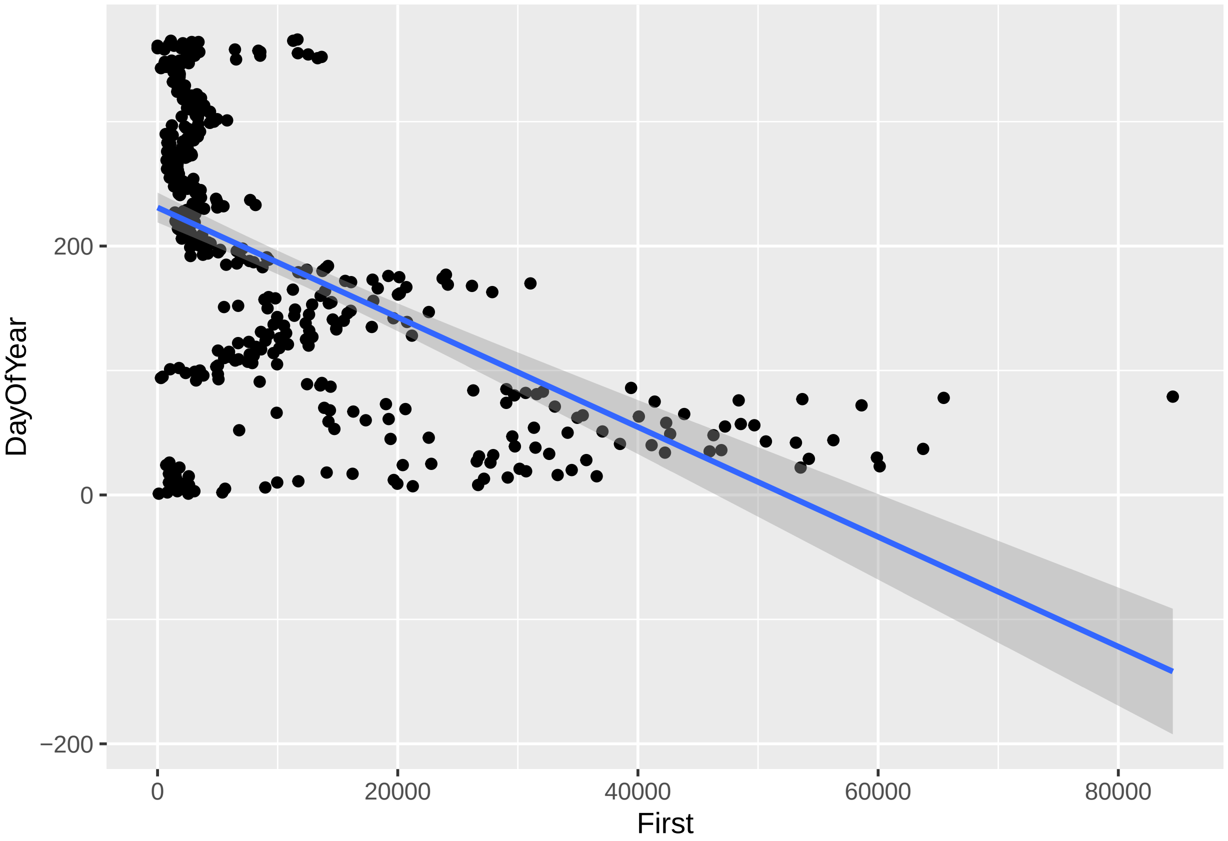
In the table below, we can see the numeric values.

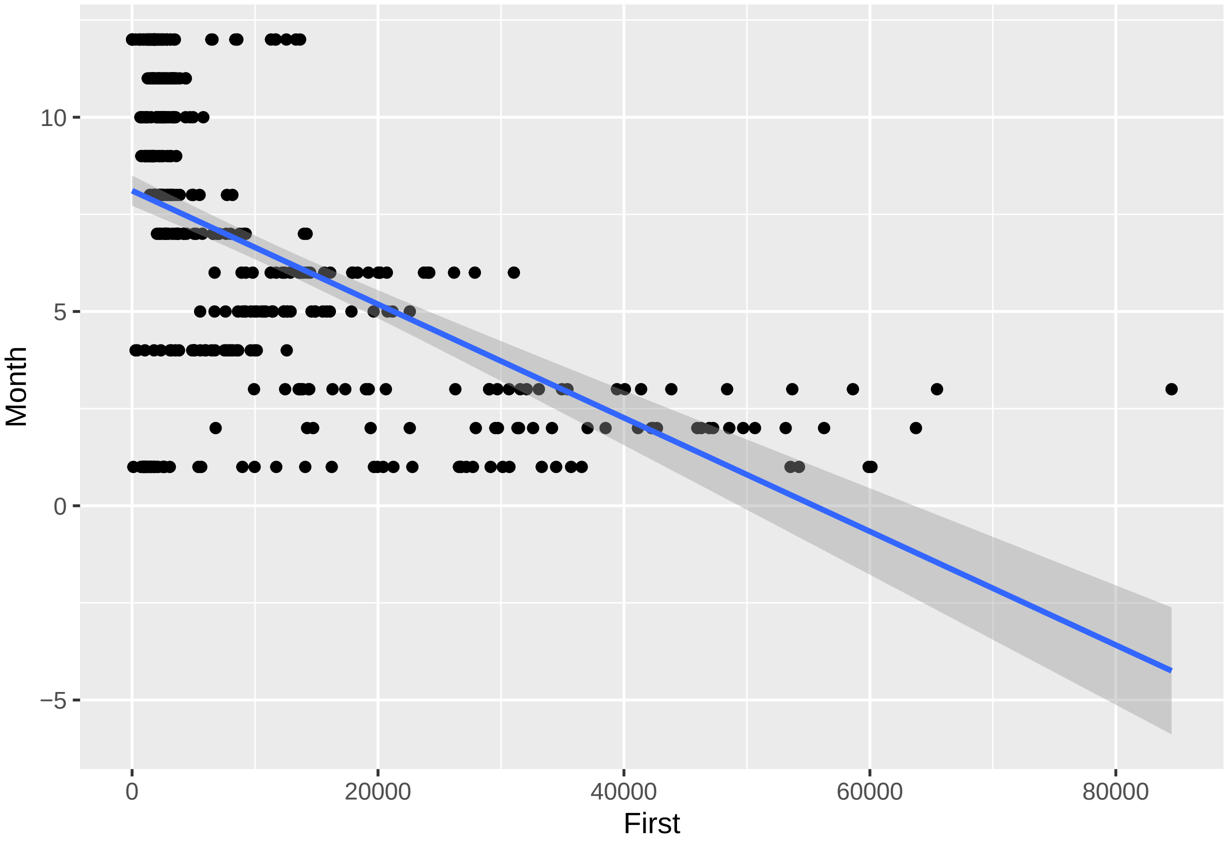
First	Month	-0.5432969
Second	Month	-0.1888013
First	DayOfYear	-0.5343244
Second	DayOfYear	-0.1901012
First	Quarter	-0.6070344
Second	Quarter	-0.1799906

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

- "Quarter",
- "DayOfYear",
- "Month".




DayOfYear
200
0
−200
0
20000
40000
60000
80000
First



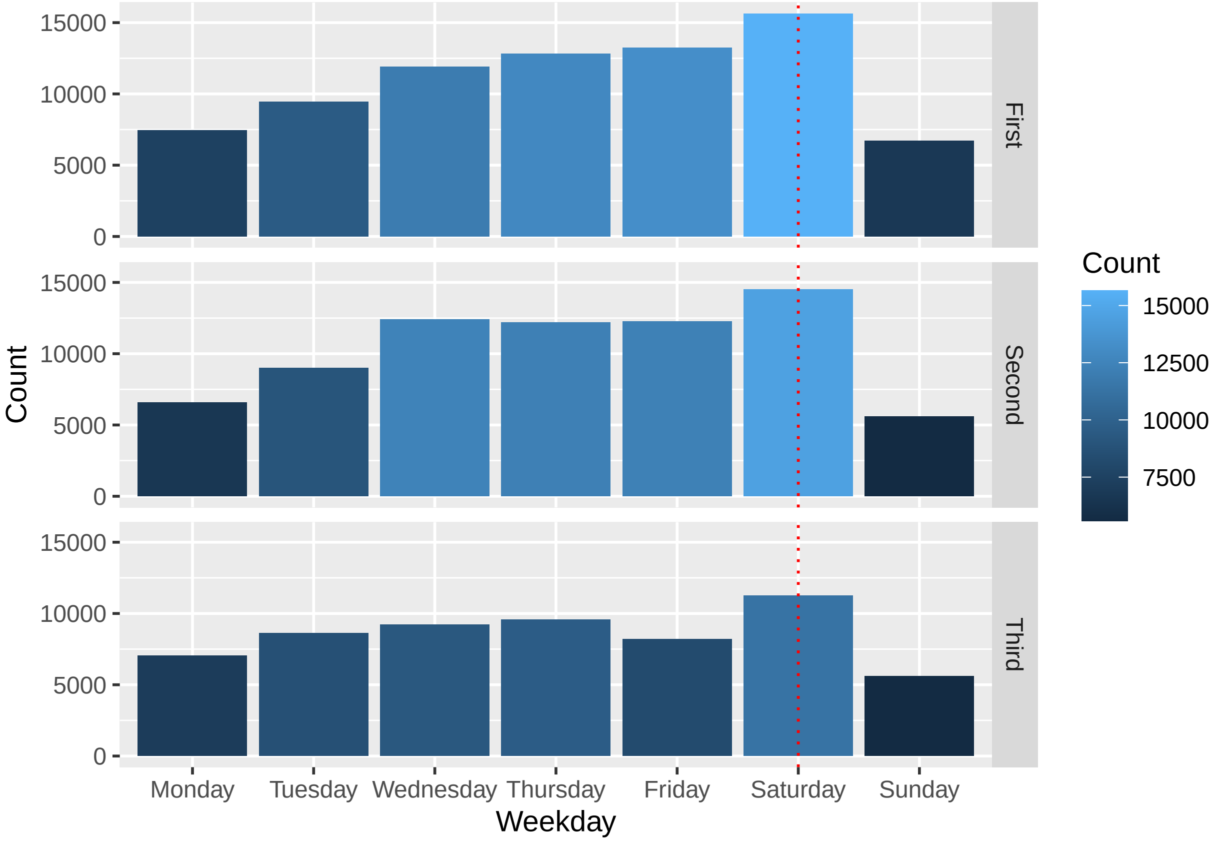
At the same time, the column “Second” doesn’t have strong relationships; but we can use the same columns.

1.2.3 Weekdays

As you remember, I have a question.

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 the side effects go away during the weekend.

1.2.4 Missing values

As we already saw in the previous chapter **??**, 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.

1.3 Step 2: 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.

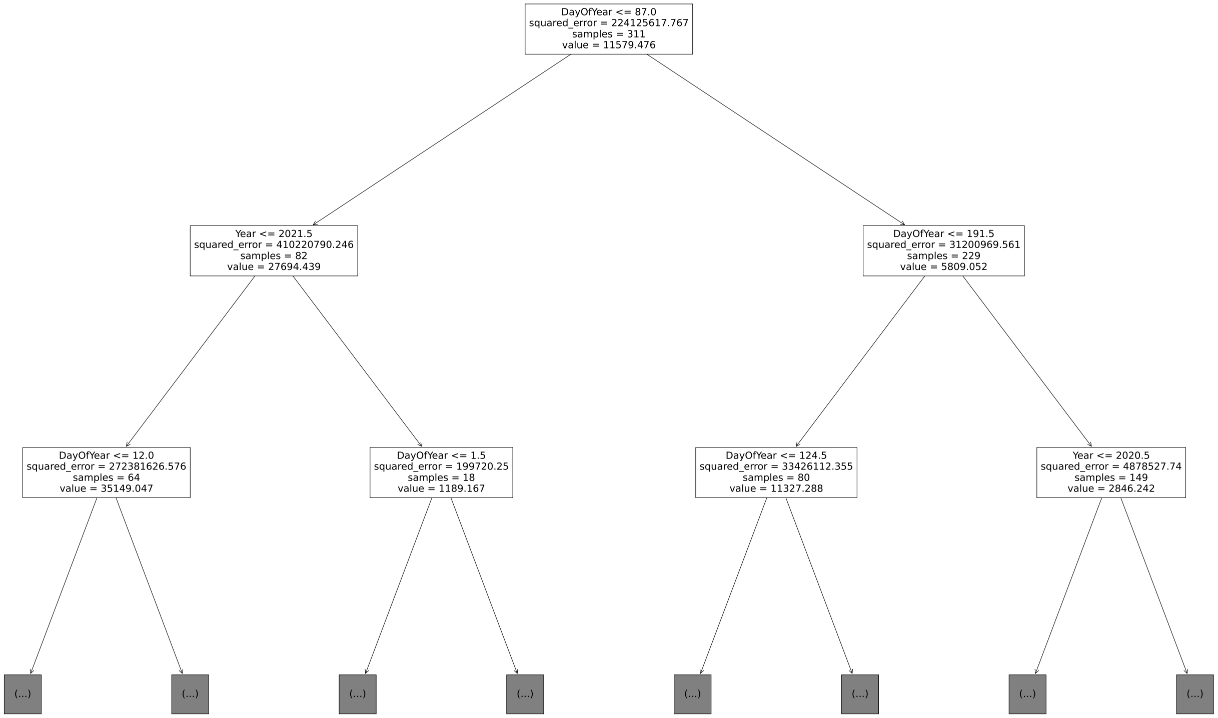
```
y_column = "First"
```

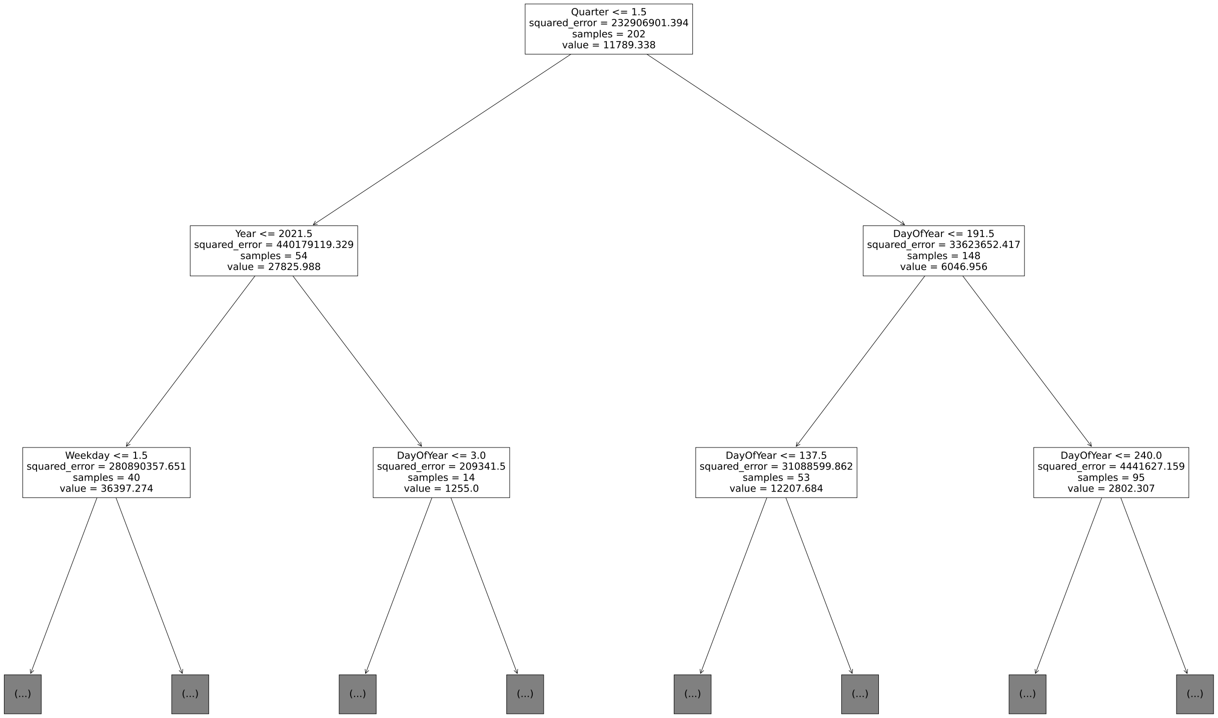
The first column that I will predict is "First".

```
## DecisionTree:  0.719657929335243
```

```
## RandomForest:  0.774580856609961
```

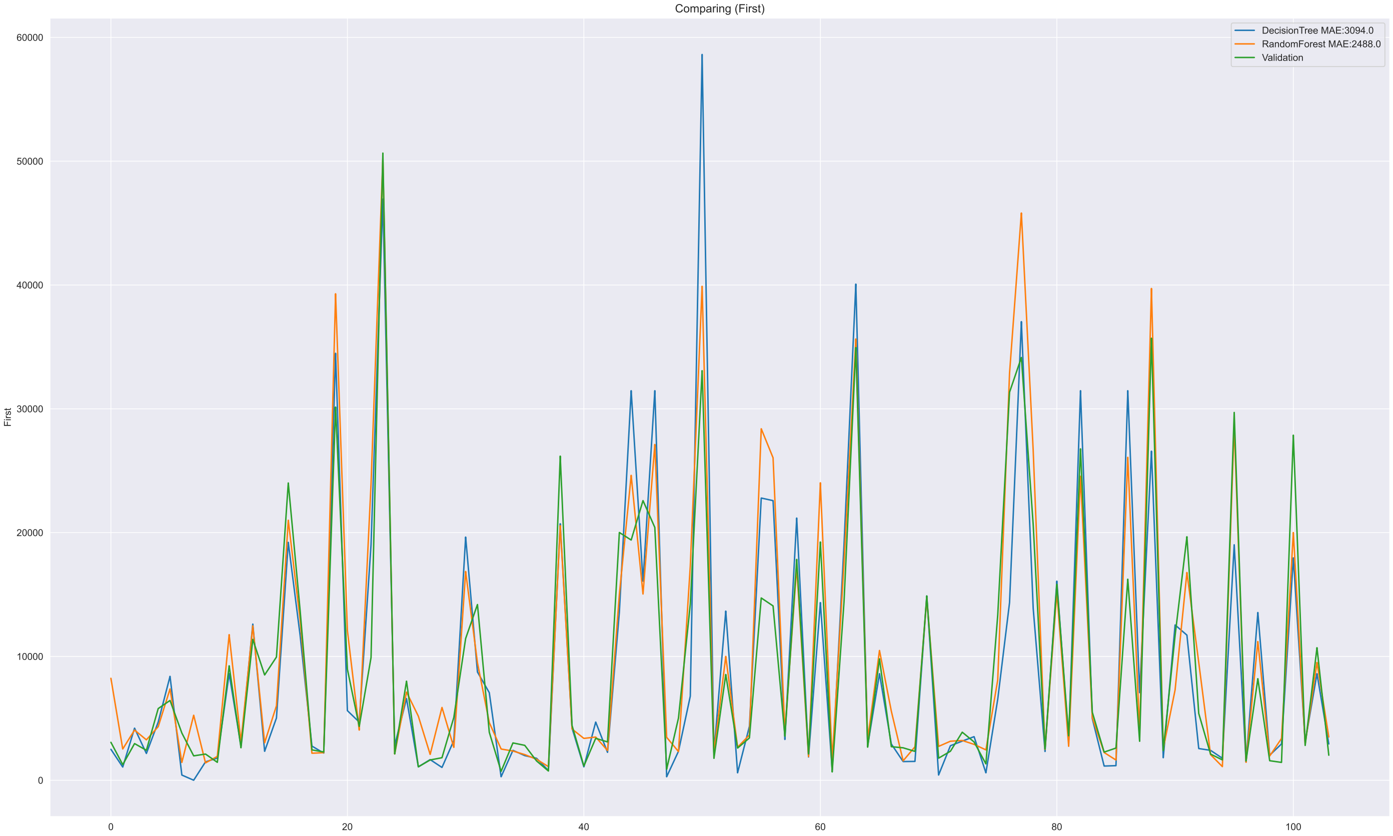
Look at the tree


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
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)



What can we see?

Finally, look at the result.



In my opinion, the result is good.

- The waves were recognized.

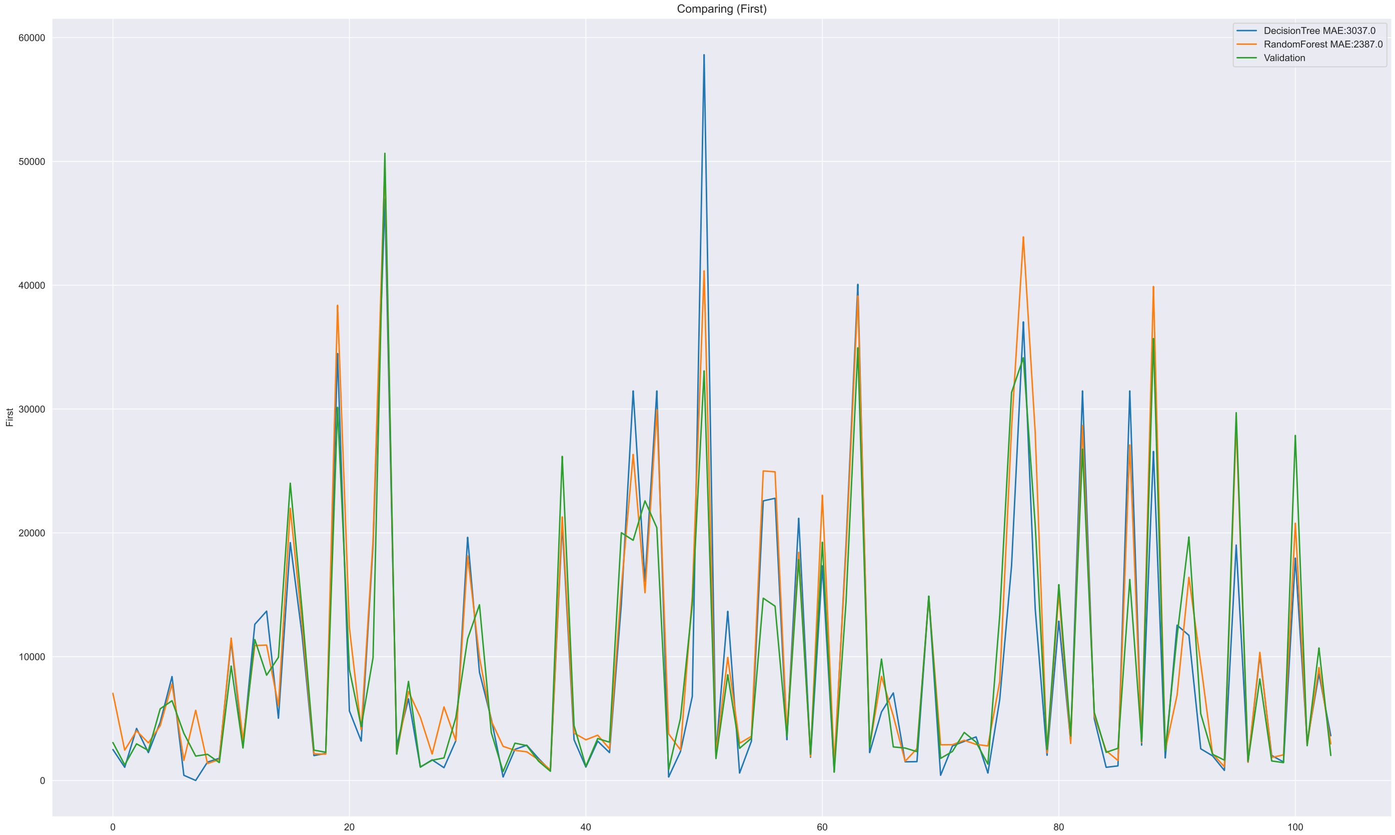
- The extreme values are bigger than in real data.

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.

```
## DecisionTree:  0.7248630326024768
```

```
## RandomForest:  0.7837038702898657
```

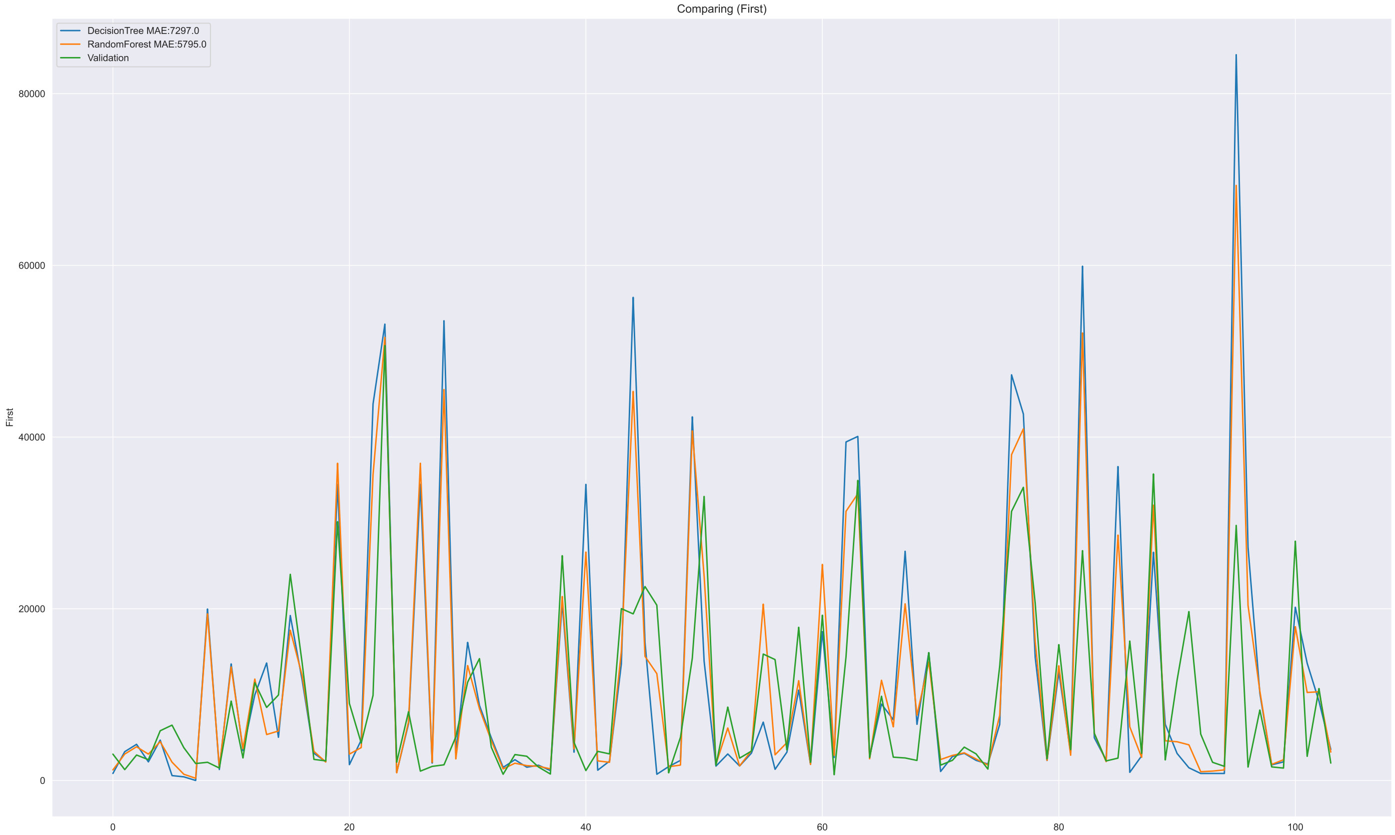


The result is better a little, but extreme values are disappointed.

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

```
## DecisionTree:  0.338881322155111

## RandomForest:  0.47495276411406473
```



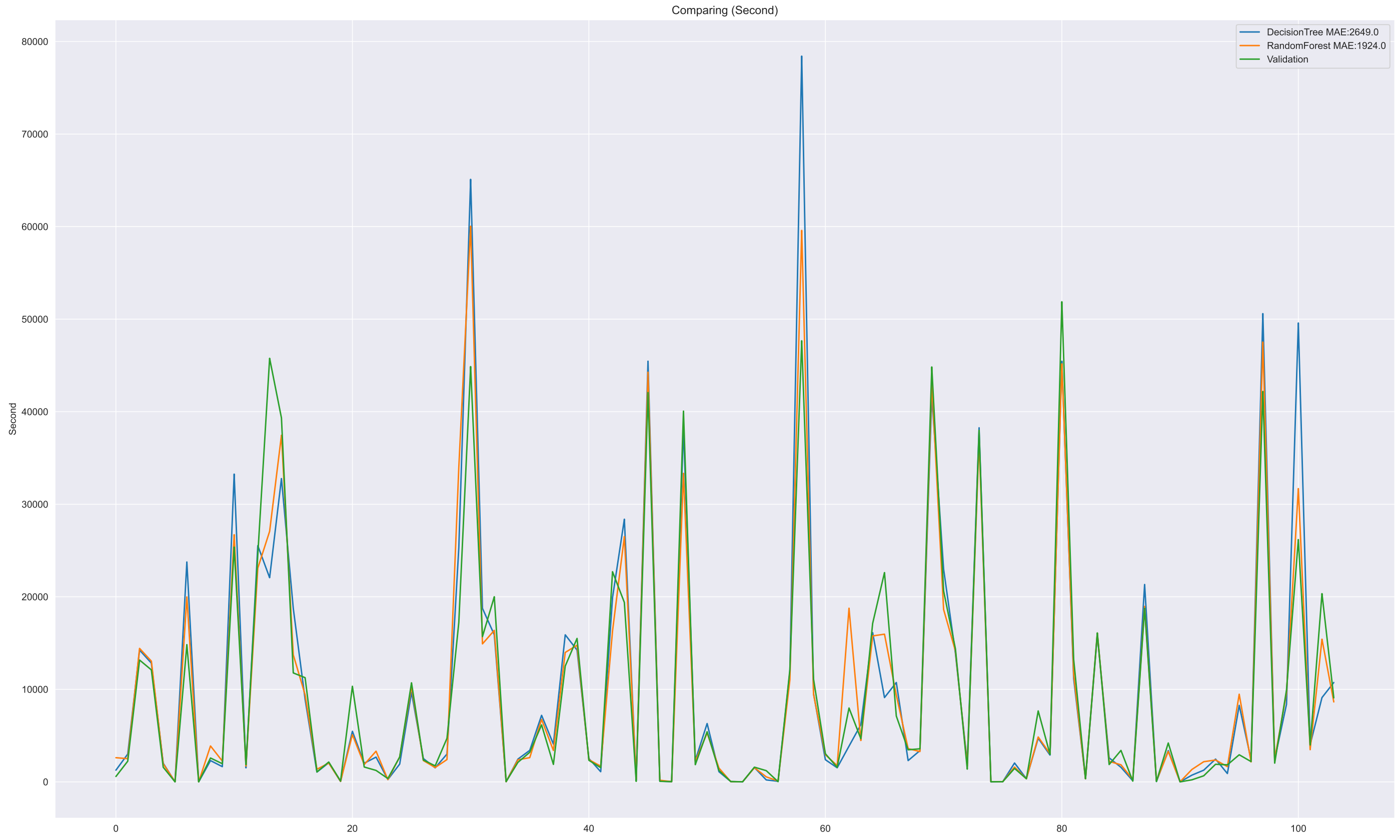
Not so good.

Repeat for the Second

```
y_column = "Second"
```

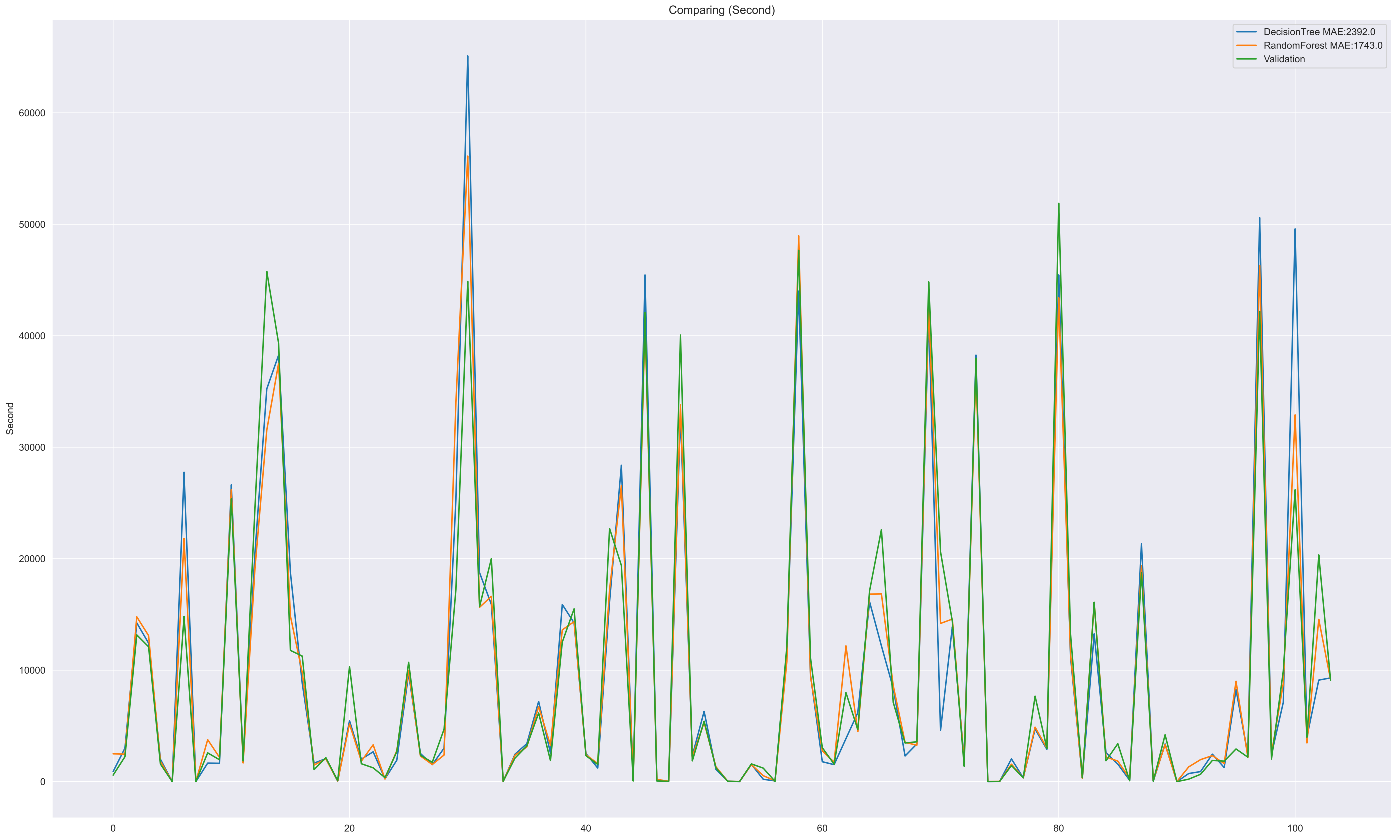
```
## DecisionTree:  0.7445280765098062

## RandomForest:  0.8144473412230163
```



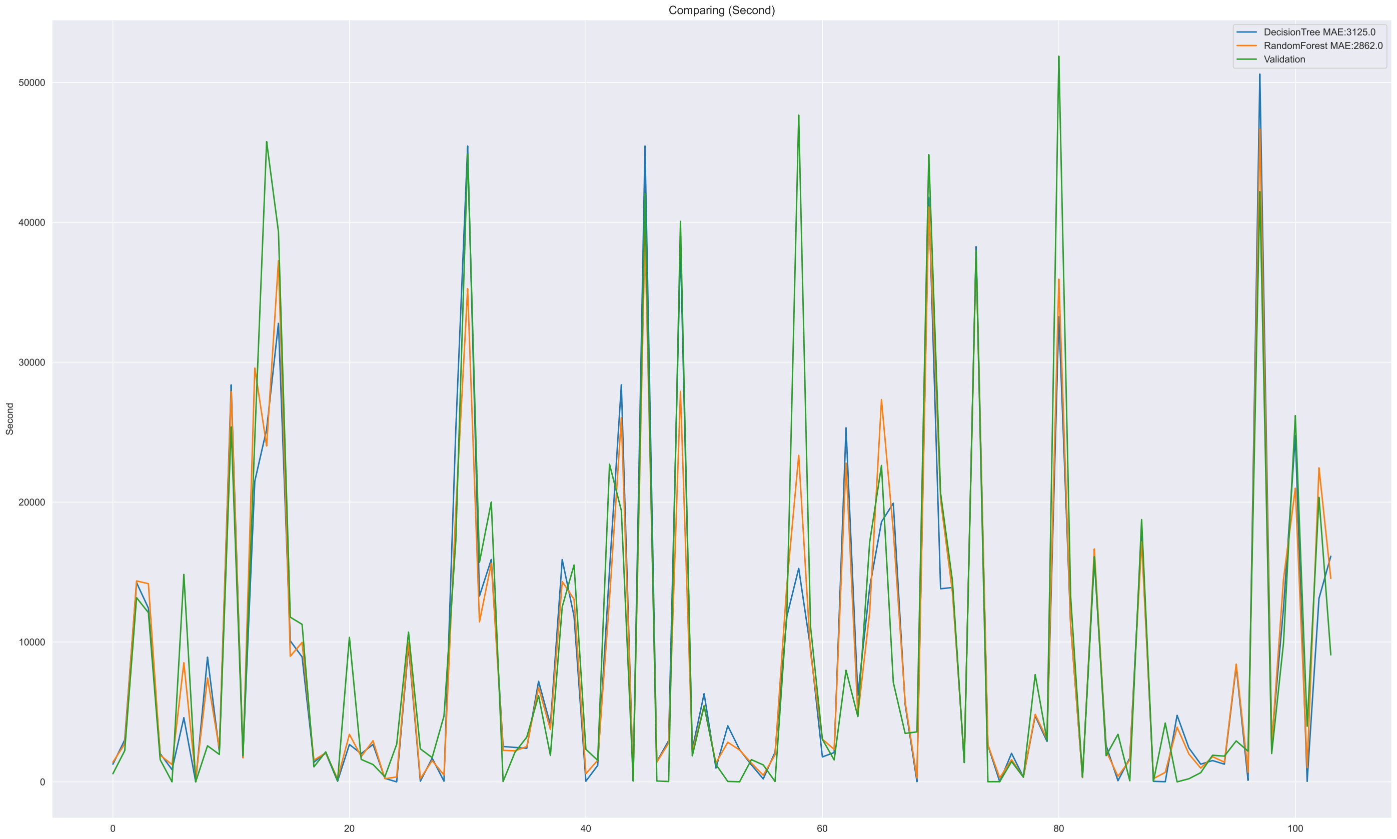
```
## DecisionTree:  0.7692636418171874

## RandomForest:  0.8318561653086721
```



```
## DecisionTree:  0.6985989452917025

## RandomForest:  0.7239113203537064
```

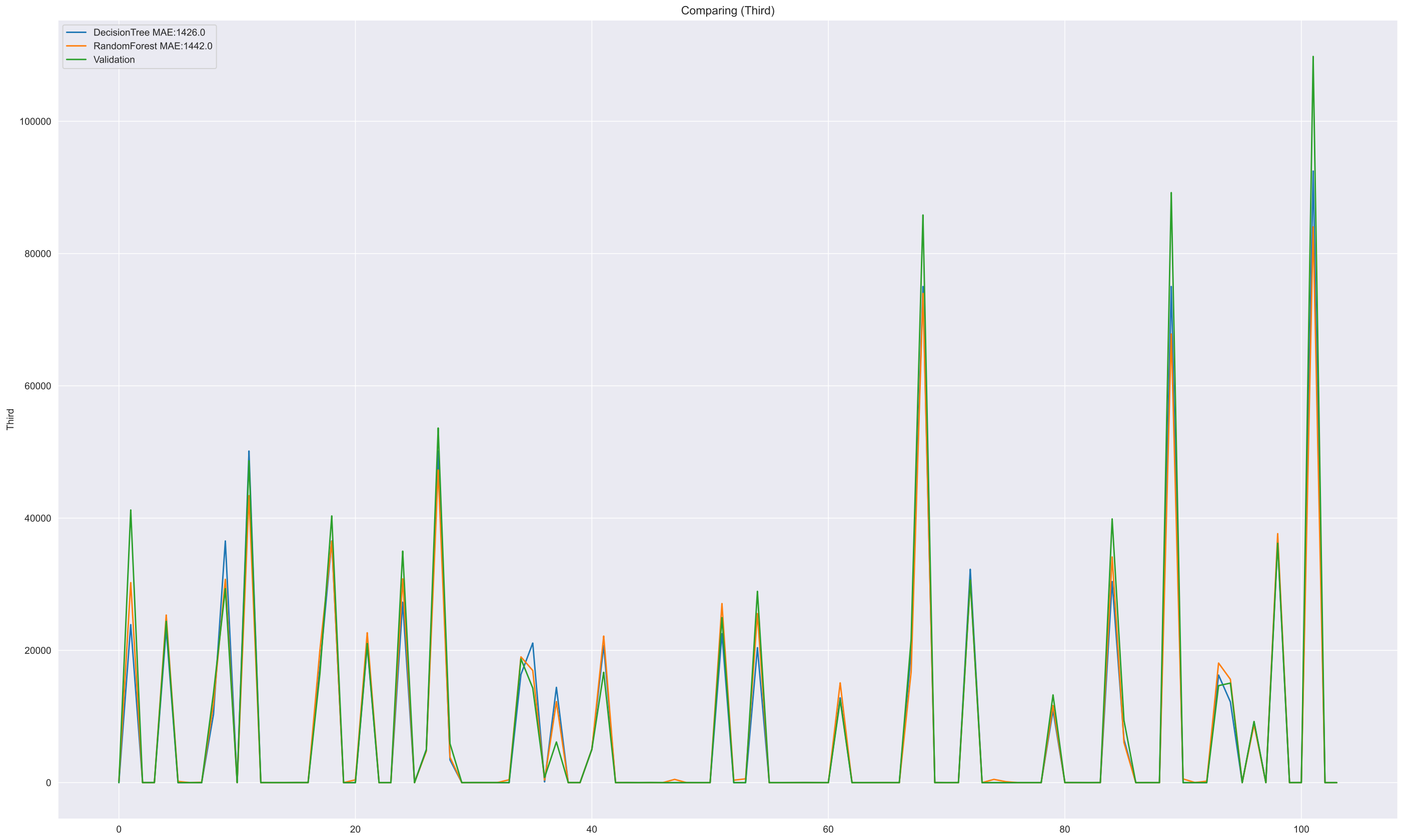


Repeat for Third

```
y_column = "Third"
```

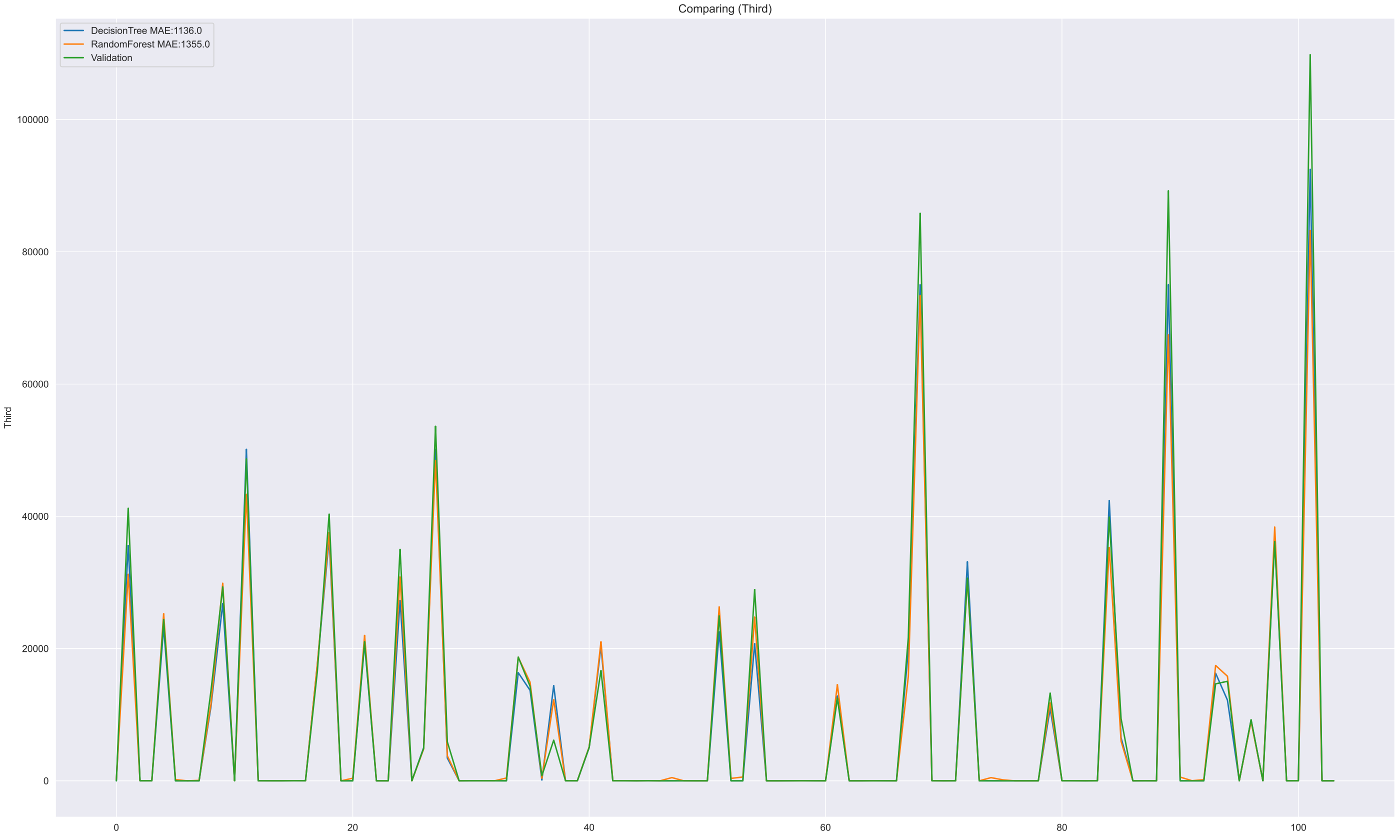
```
## DecisionTree:  0.8323340834065006

## RandomForest:  0.8304365732995669
```



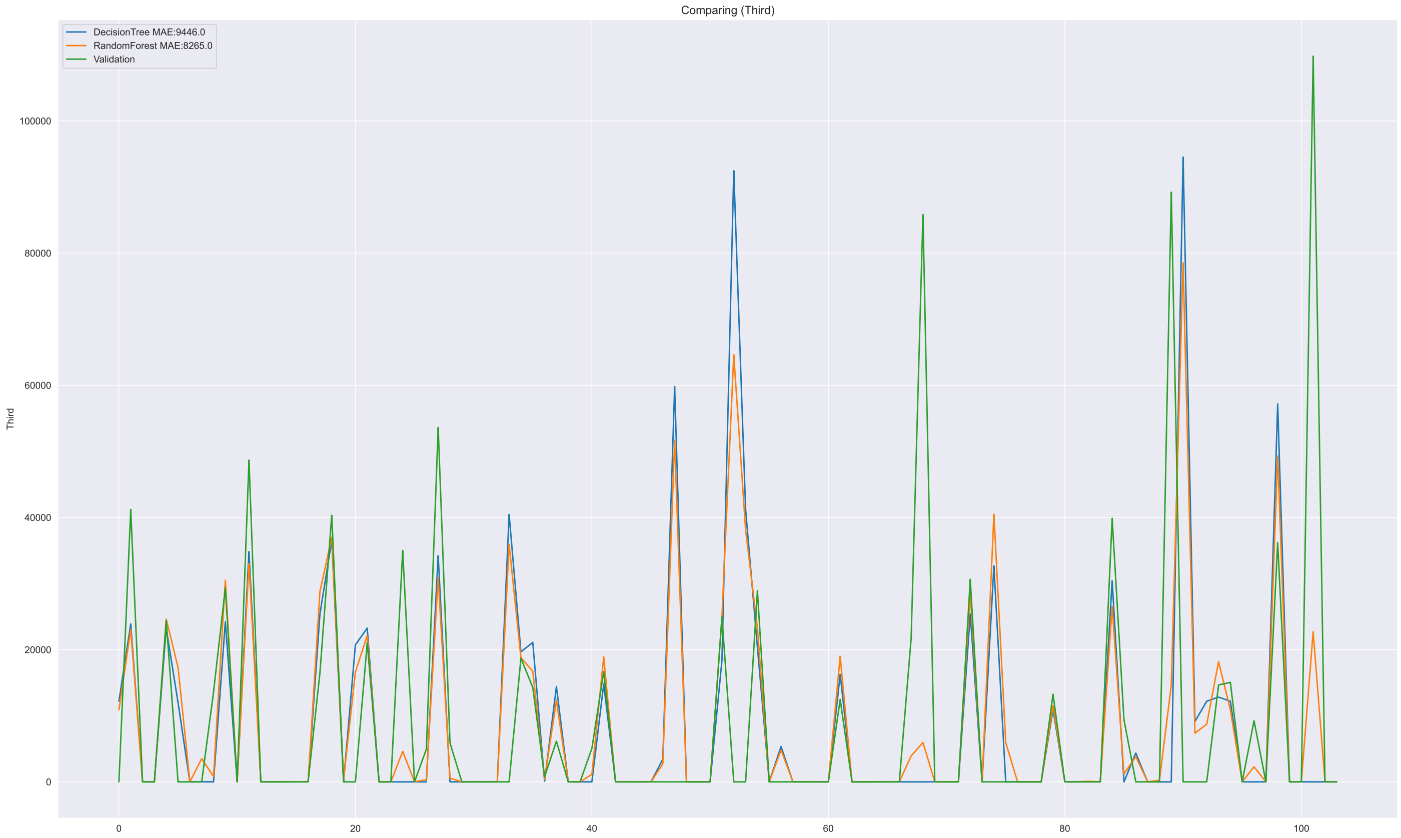
```
## DecisionTree:  0.8664423499345815

## RandomForest:  0.840745098066024
```



```
## DecisionTree:  -0.11041243558502623

## RandomForest:  0.028430358547872348
```



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