Contents

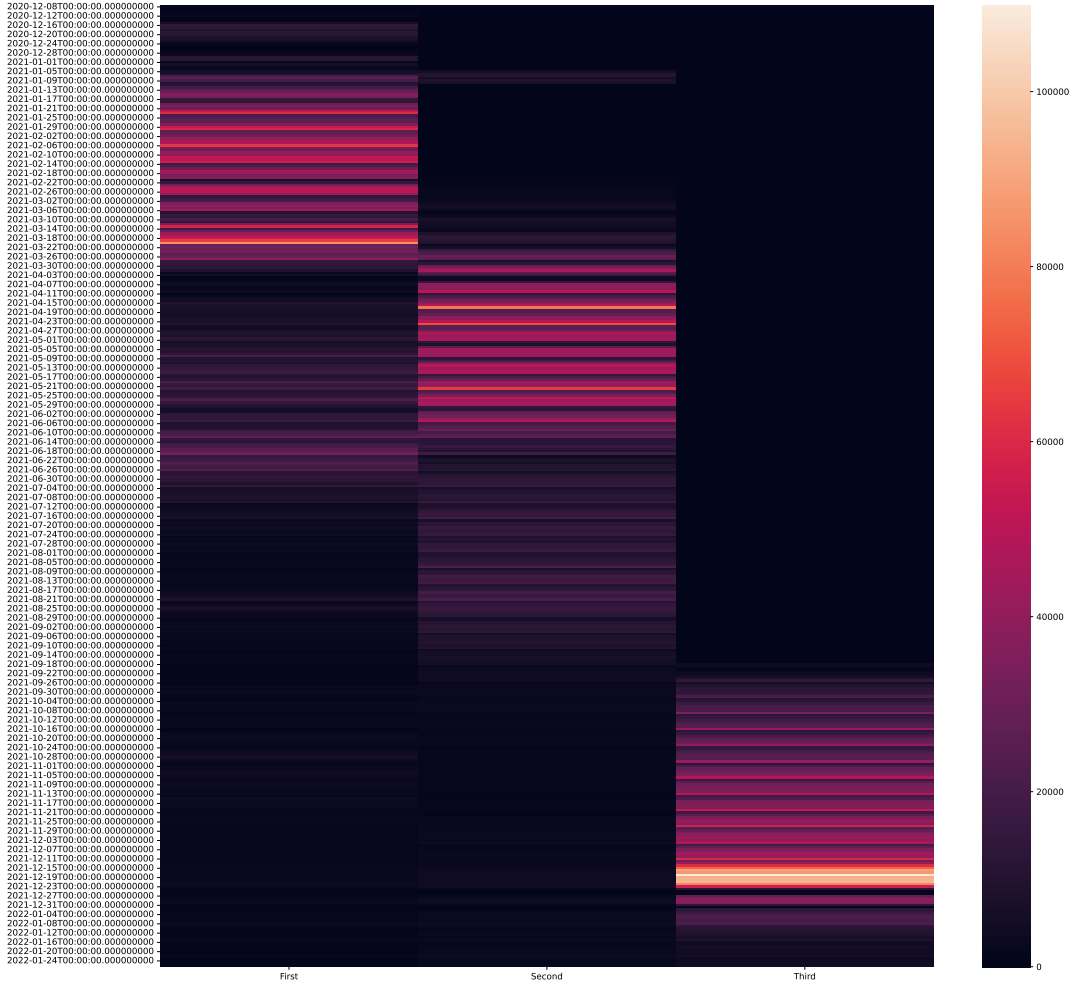
1 Step 4 Machine learning

1.1 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.

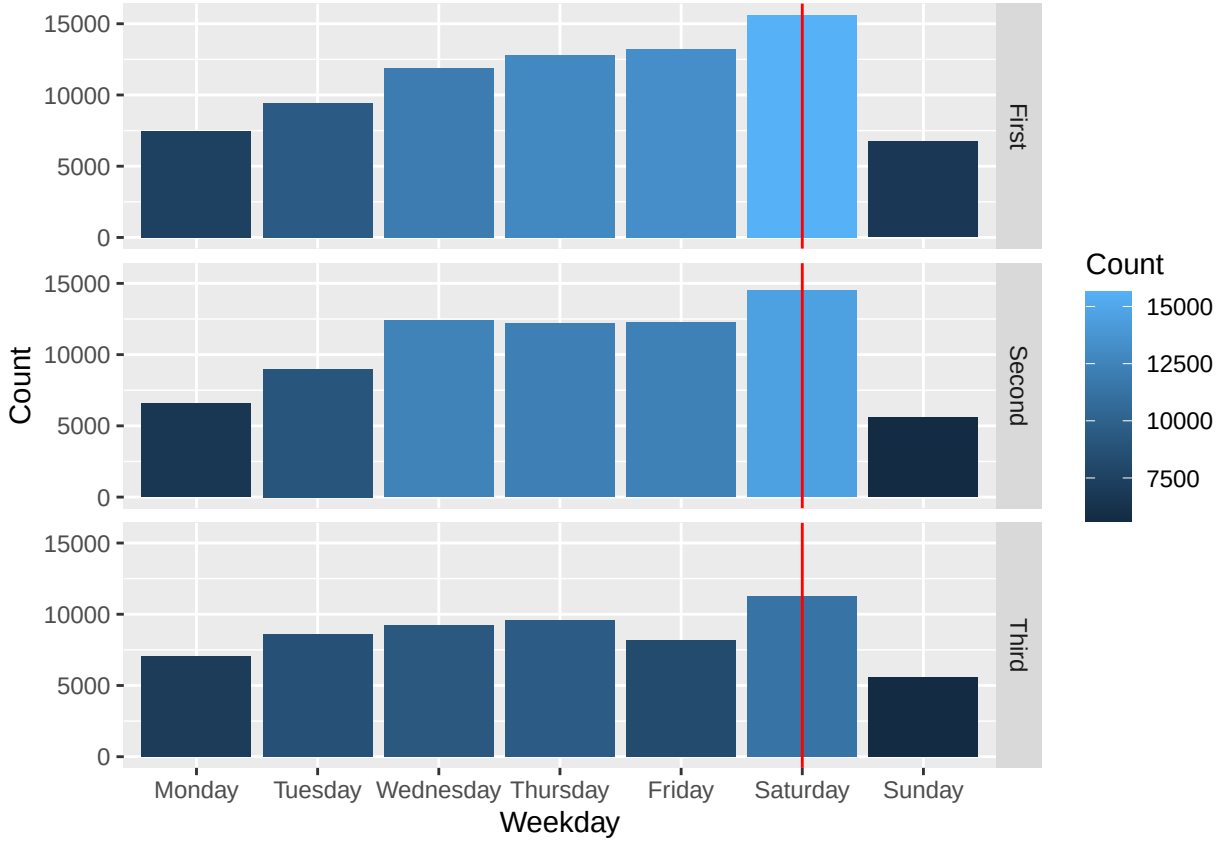
Let's get features: 1) Year 2) Month 3) Day etc.

	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

1.2 Step 1: Explore the dataset

1.2.1 Weekdays

As you remember, I have a question.



Let's answer.

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

1.2.2 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.

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

Define necessary variables

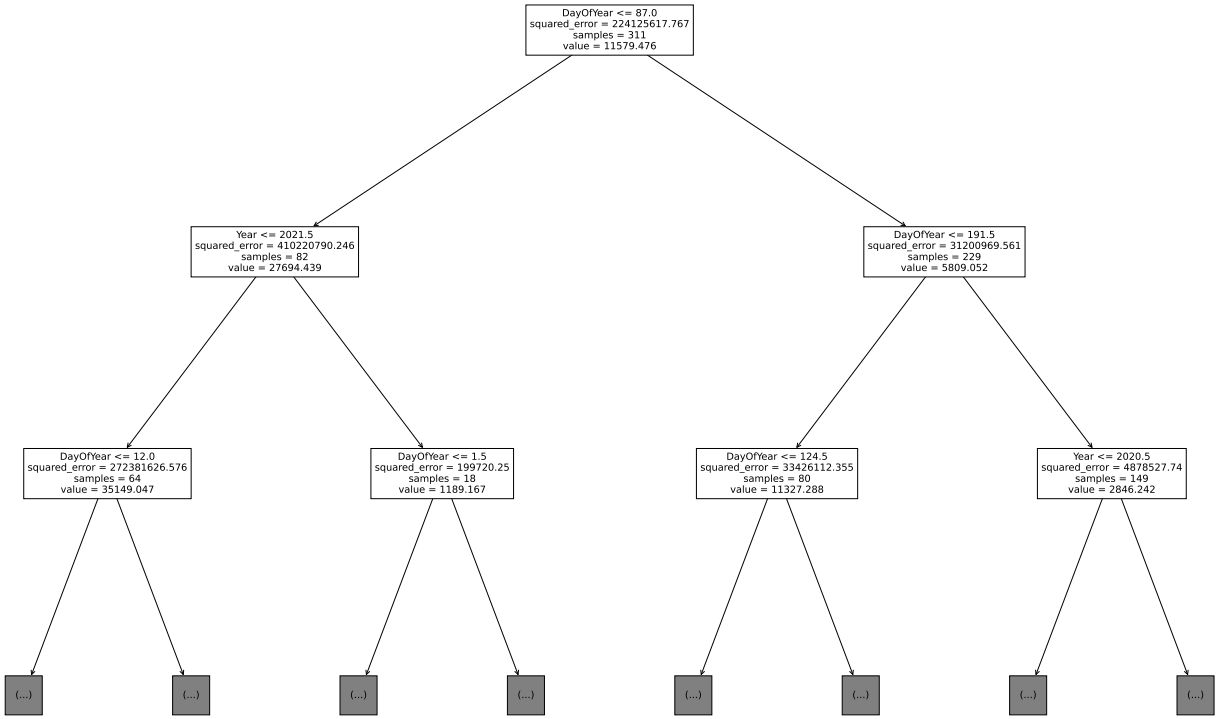
Prepare sets and train models using parameters.

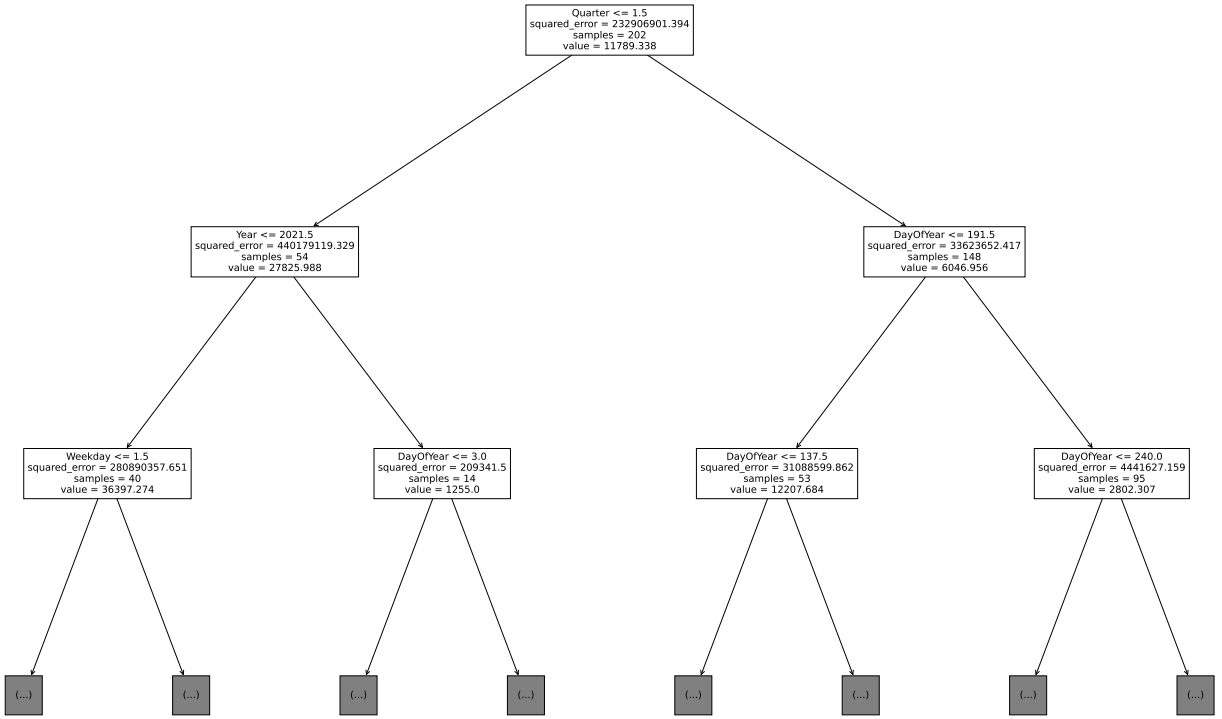
```
y_column = "First"
```

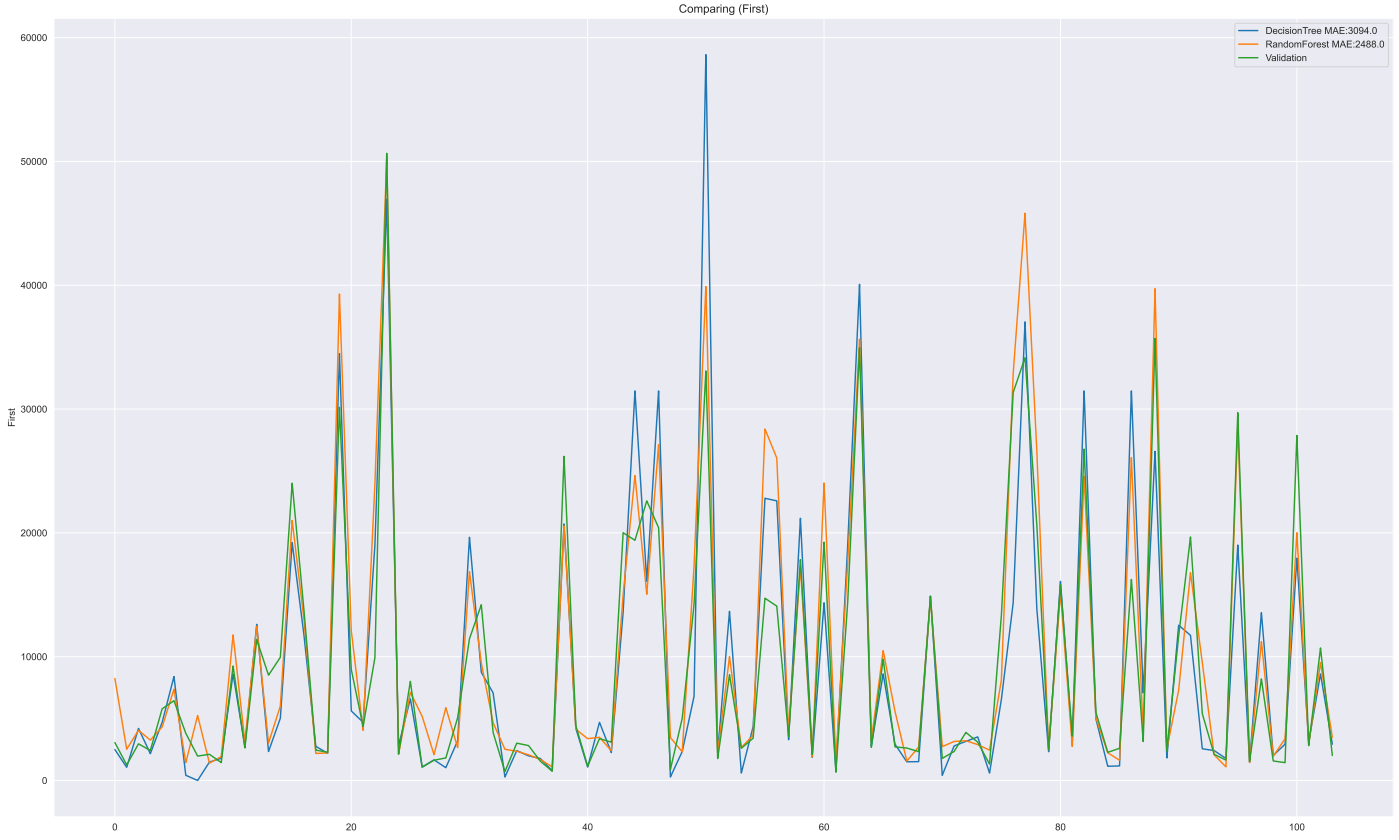
```
## DecisionTree:  0.719657929335243
```

```
## RandomForest:  0.774580856609961
```

Look at the tree

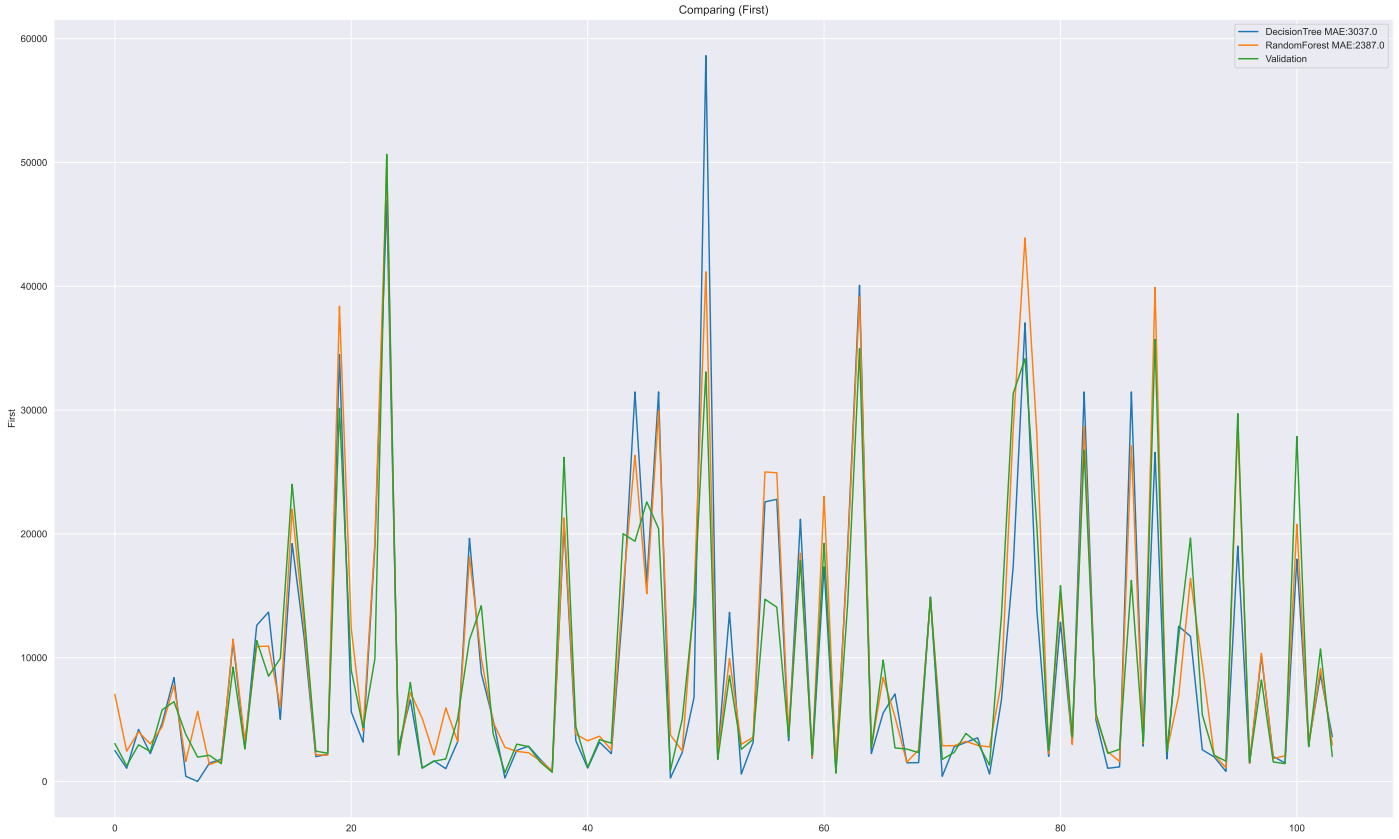

DayOfYear <= 87.0
squared_error = 224125617.767
samples = 311
value = 11579.476
Year <= 2021.5
squared_error = 410220790.246
samples = 82
value = 27694.439
DayOfYear <= 191.5
squared_error = 31200969.561
samples = 229
value = 5809.052
DayOfYear <= 12.0
squared_error = 272381626.576
samples = 64
value = 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
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)


Quarter <= 1.5
squared_error = 232906901.394
samples = 202
value = 11789.338
Year <= 2021.5
squared_error = 4401791 19.329
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
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)



DecisionTree: 0.7248630326024768

RandomForest: 0.7837038702898657

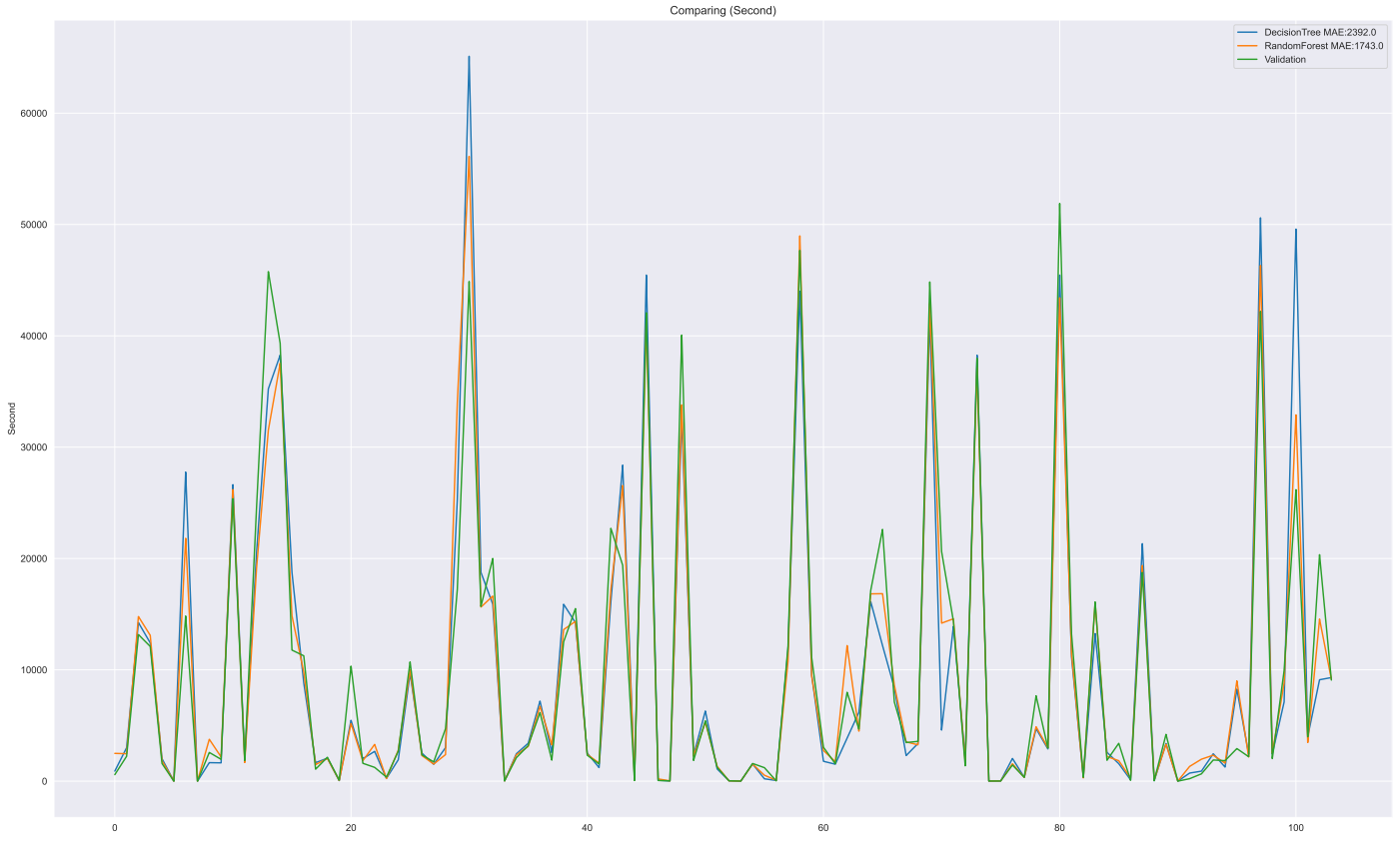


Repeat for the Second

```
y_column = "Second"
```

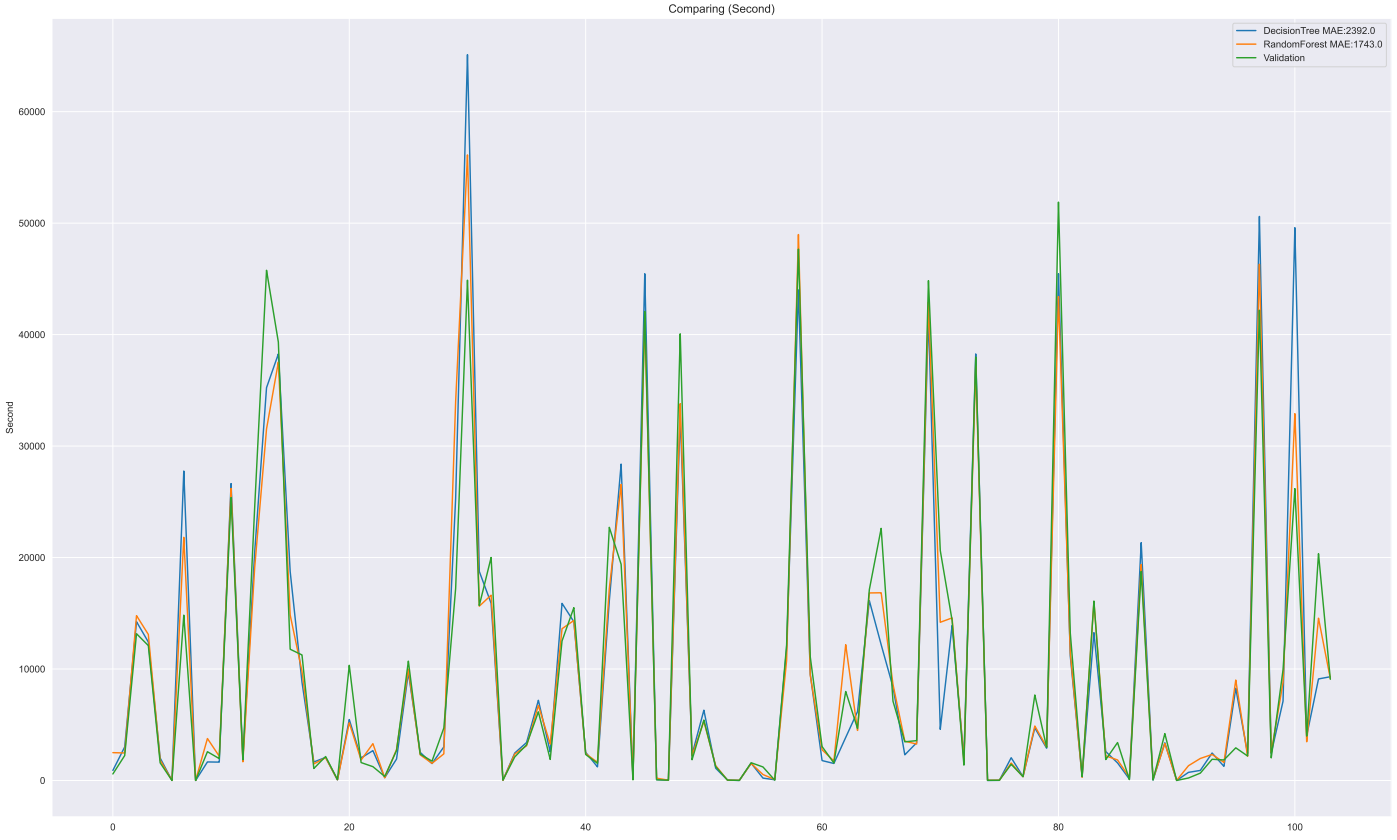
```
## DecisionTree:  0.7692636418171874
```

```
## RandomForest:  0.8318561653086721
```




```
## DecisionTree:  0.7692636418171874

## RandomForest:  0.8318561653086721
```



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

A combination of the following features give us the best result:

- Weekday,
- Year,

- DayOfYear