Time Series Forecasting on Interns Datasheet

Report:

Swastik Saxena (ms2004101011@iiti.ac.in)

Saurav Kumar (ms2004101009@iiti.ac.in)

Step 1: Importing necessary libraries

We have used following libraries for our work

- Pandas : For reading and saving dataset and also to perform different operations on data like grouping, apply(), map(), info(), isna() etc.
- Seaborn : This is used for making visualization. We have created different types of graphs for visualization purpose like countplot, histogram, linegraph, barplot etc.
- Datetime : Since we are dealing with date and time so we need to convert the required date into correct format before processing further. For parsing and converting date format, this library is used.
- Sklearn : This library provides many important function for different models and different evaluation metrics. We have used RMSE for evaluating our model which is provided by sklearn.meterics.
- Statsmodel : This module provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.

Step 2: Data Cleaning:

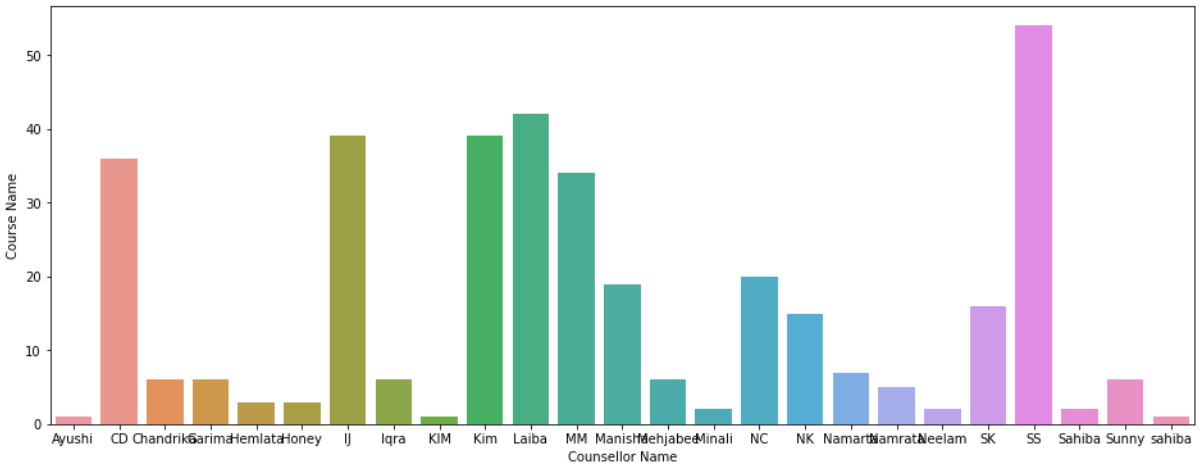
- The data for student consists of many null values for course names.
- To deal with these values, we have used to methods. First we have experimented by dropping all the null rows from the dataset but that will drop 40% of data so other methods we have follow is that we have filled the null values of course name with the values of previous rows. Since it can

be a possibility that due to same course name, the name has not been mentioned so we have carried our experiment with second approach.

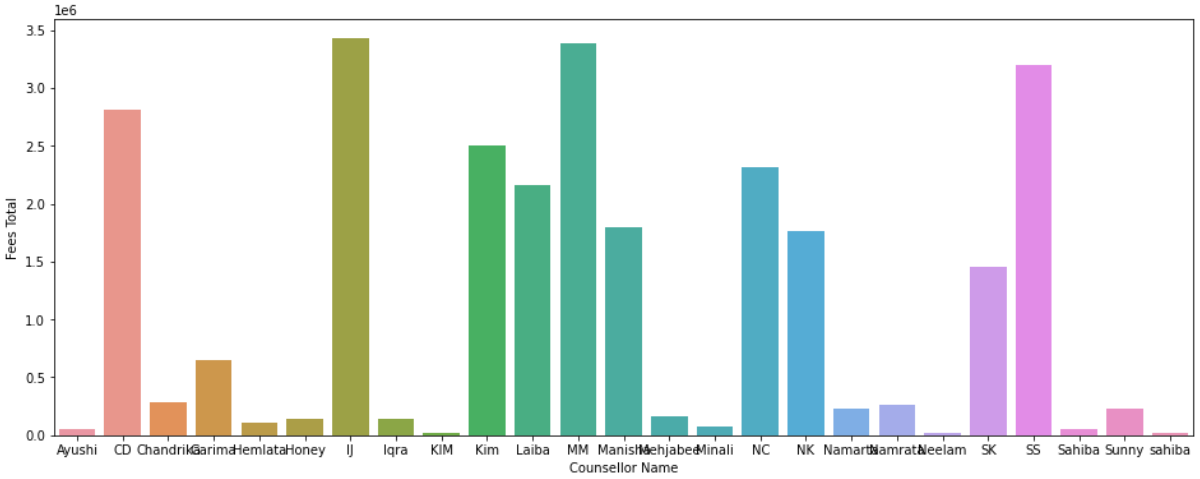
- The Fees Total, Fees Received and Fees pending all have object datatypes but the fees should be numerical. These columns have some string entries of string types like “Fees Forwarded”, so we have removed these rows as they are small in numbers.
- We then changed the data type to int64 for fees columns.
- We have changed the index of dataset from default to “Month” column as we are dealing with time so it will make sense to make prediction on date and time.
- The type of Month dataset is object so we have parsed using datetime library to make inferences on date.
- We have also changed the columns name of all the sheets to same name as different datasets have different names for columns.
- We have used following columns name : Counsellor Name, Course Name, Fees Total, Fees Received, Fees Pending.

Step 3: Feature Enginnering:

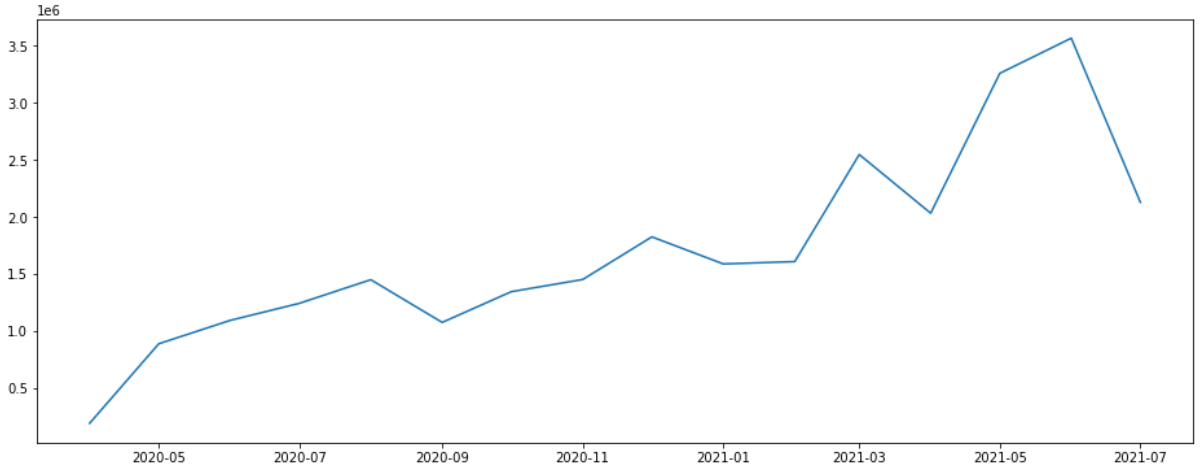
- In this step, we try to do data analysis and tried to get some insights of data.
- We have made different visualization on the basis on number of course each counsellor taking. The graph below shows different number of course taken by each counsellor.



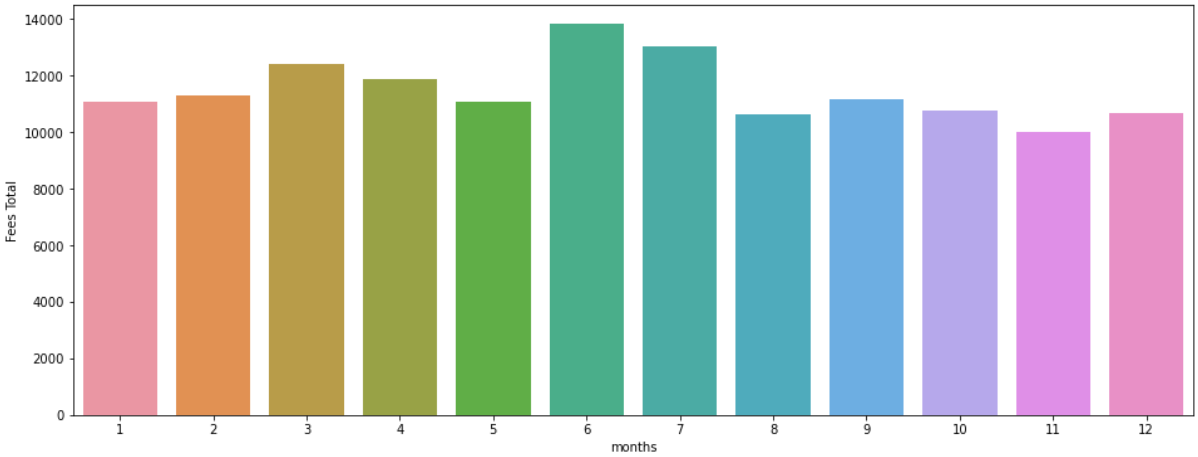
- The following graph shows the total fees generated by each counsellor.



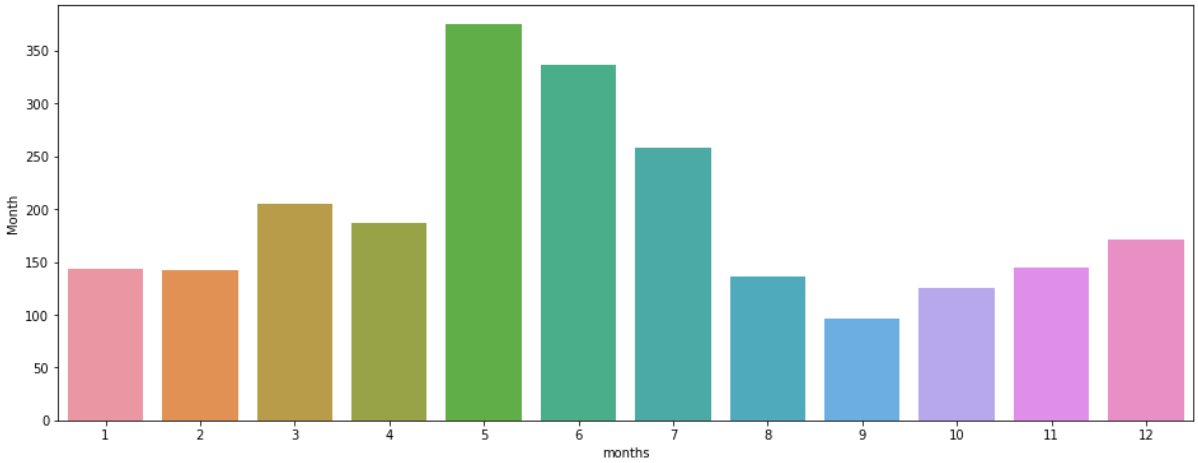
- The following line graphs shows the total fees generated month wise.



- The following graphs shows the mean of total fees collected monthly wise.



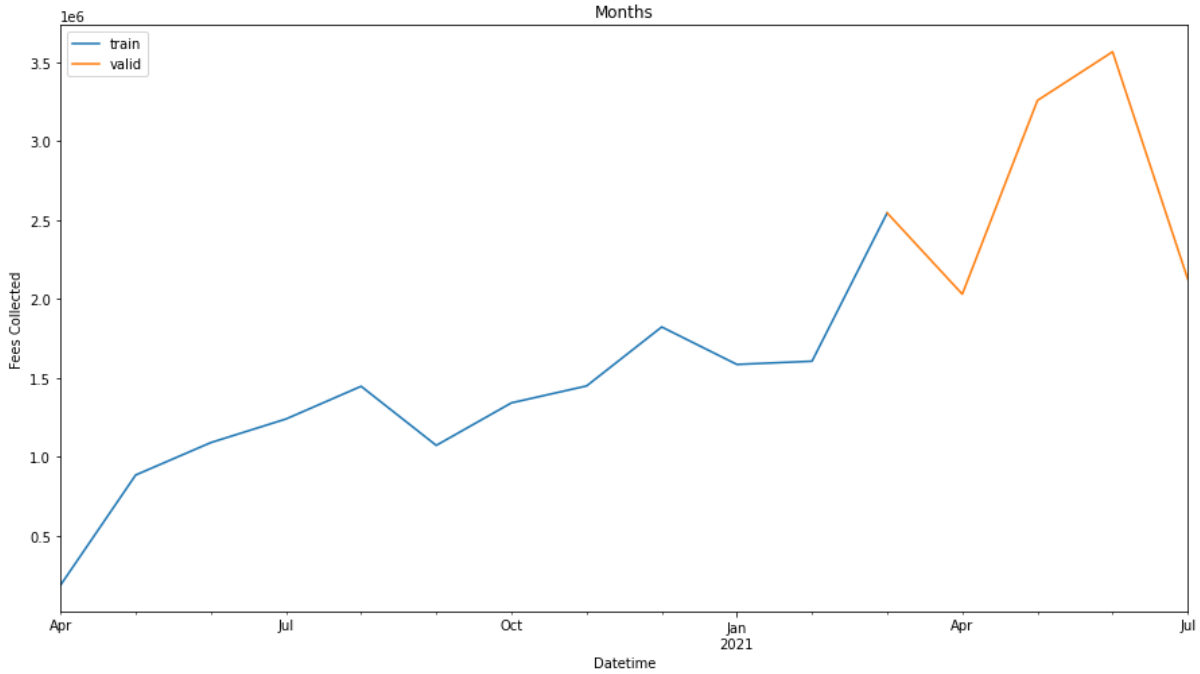
- We can make inference that due to not availability of data for months before April in 2020 and month after July in 2021, the total fees for June and July is high as we have data of both years for these months but not for others.
- After doing visualization on fees collected, we have done the same with number of sales per month.
- The following graphs shows the number of sales for each month.



- After doing EDA, we go on defining model for forecasting on the total fees collected.

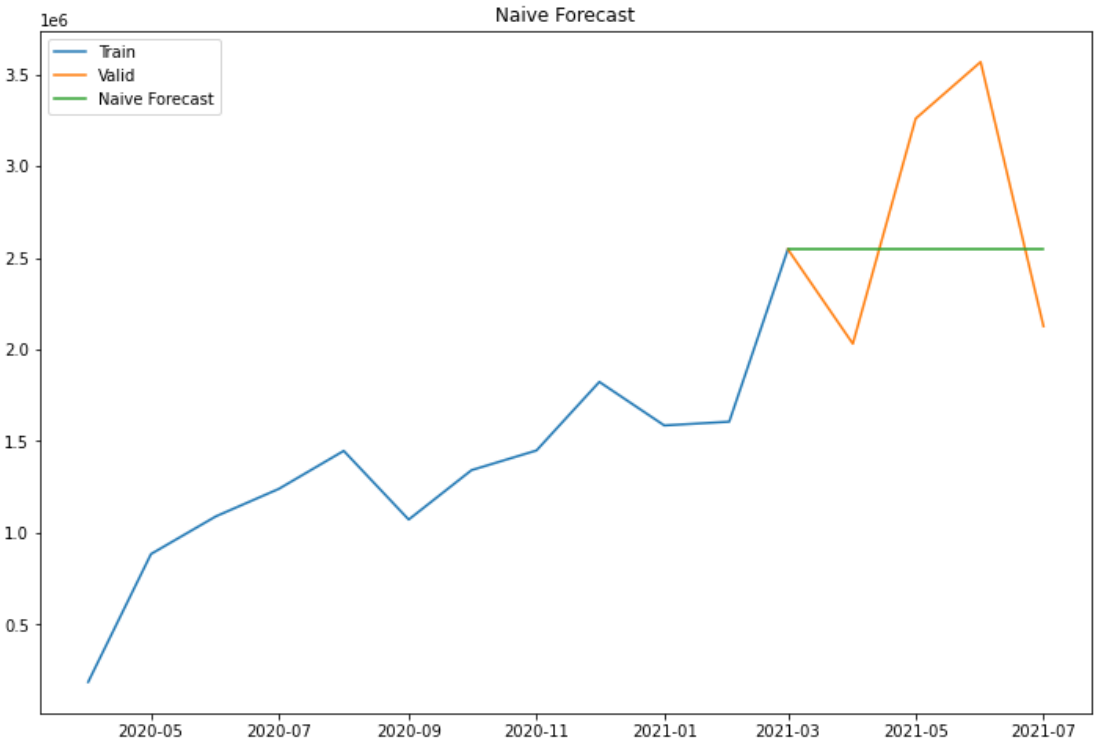
Step 4: Defining Data Models and Evaluating.

- We have defined a new dataset for forecasting, which have months as index and total fees collected month wise.
- Datetime
- 2020-04-01 187000
- 2020-05-01 884935
- 2020-06-01 1090372
- 2020-07-01 1239974
- 2020-08-01 1447287
- 2020-09-01 1072710
- 2020-10-01 1342199
- 2020-11-01 1449374
- 2020-12-01 1823178
- 2021-01-01 1585790
- 2021-02-01 1606208
- 2021-03-01 2546000
- 2021-04-01 2031171
- 2021-05-01 3258768
- 2021-06-01 3567337
- 2021-07-01 2126604
- Name: Fees Total, dtype: int64
- We have split our dataset into 70% and 30% for training and validation respectively.
- Train=pd.DataFrame(x3.loc['2020-04-01':'2021-03-01'])
- valid=pd.DataFrame(x3.loc['2021-03-01':'2021-07-01'])



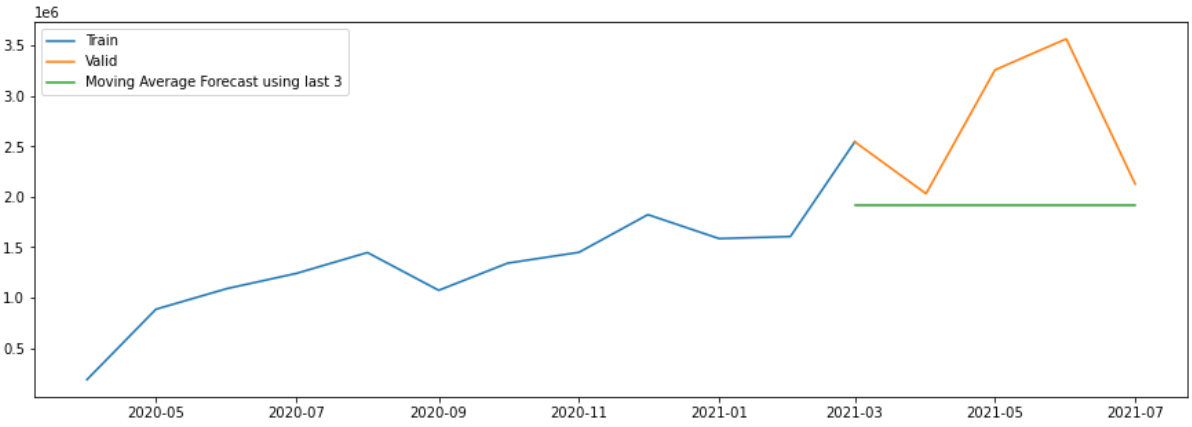
Models:

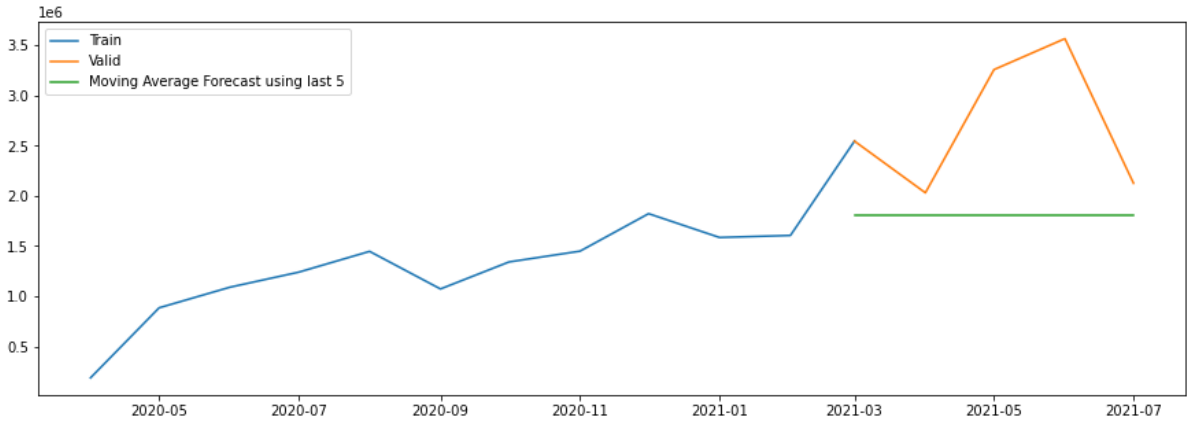
1. **Naïve model:** In this model the forecasting values of new data is same as the value of last known data.
    a. For example, If we have last month fees in training data as 1000, then for all the coming months the same fees will be forecasted.
    b. Following graph shown this model

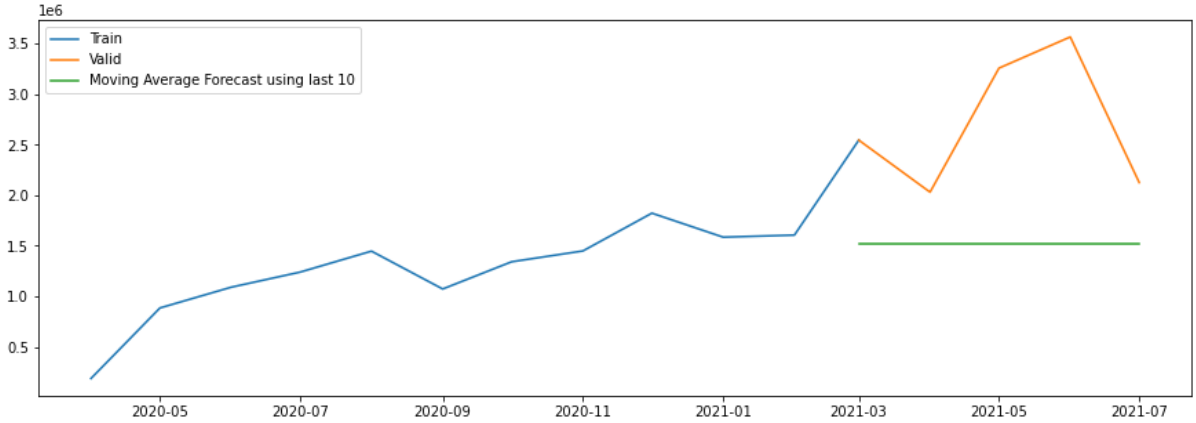


c. The rmse value for this model is 631206.6846049716

2. **Moving average**: In this method, we take average of some last observations and forecasting in done on that basis.
   a. For eg, we have done experiment by taking last 3, 5 and 10 observations.
   b. Following figures shows the prediction on validation dataset.





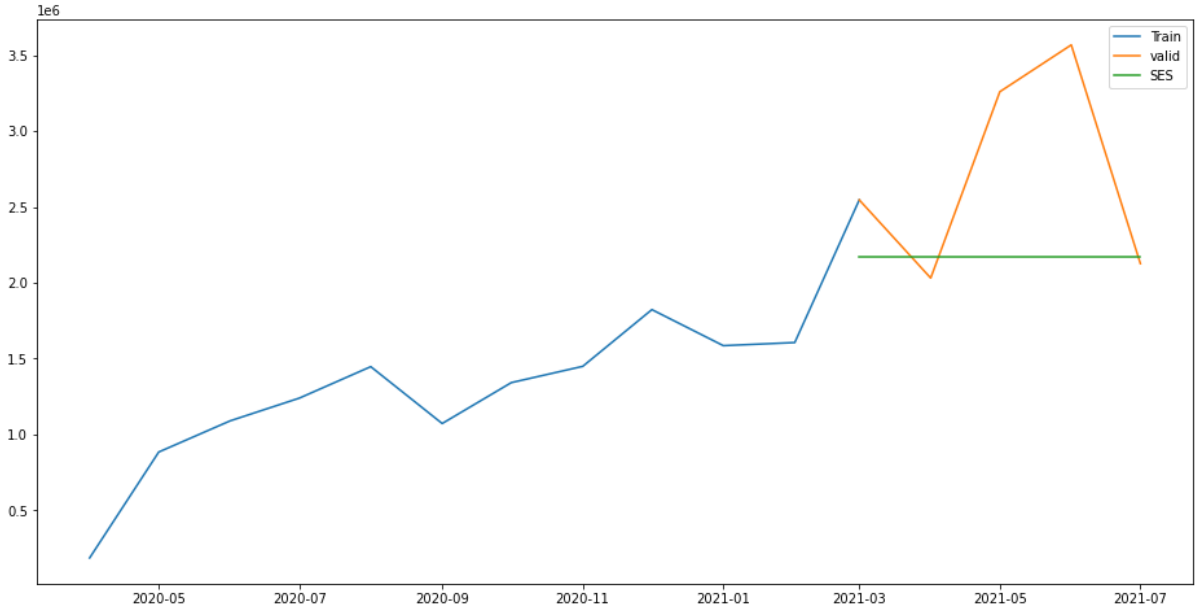


RMSE value for this model is 1333654.8724224868

3. Simple Exponential Smoothing:

In Exponential Smoothing, the far observations get less weightage and recent observations got more weightage.

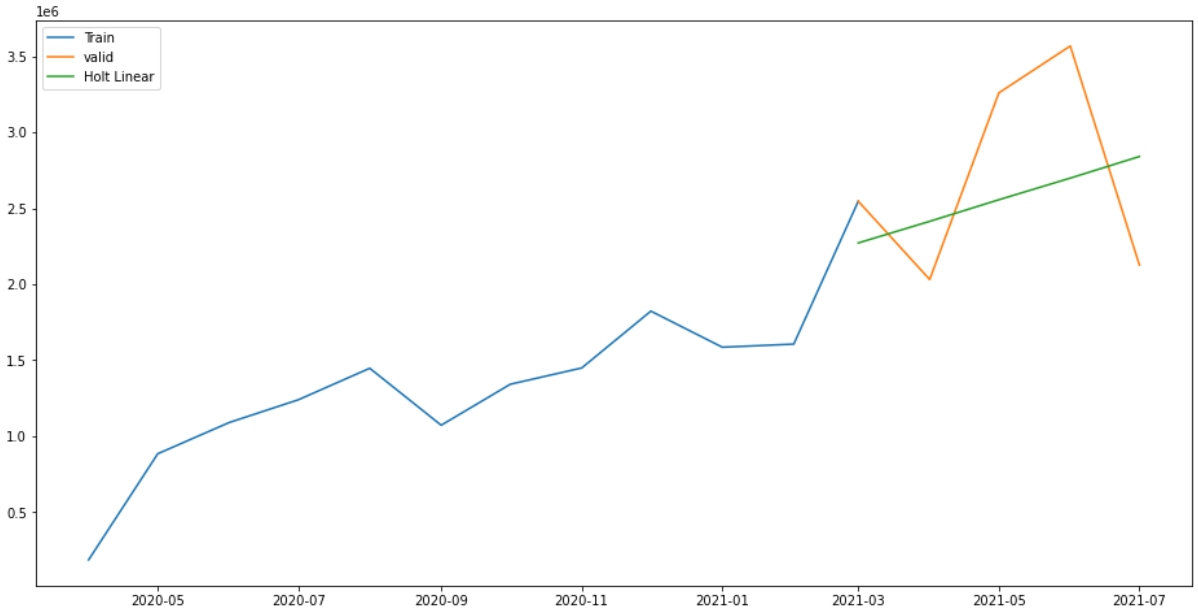
RMSE value for this model is 812052



4.Holts Linear Trend Model

In this method, we take trend of data also into consideration with simple exponential smoothing.

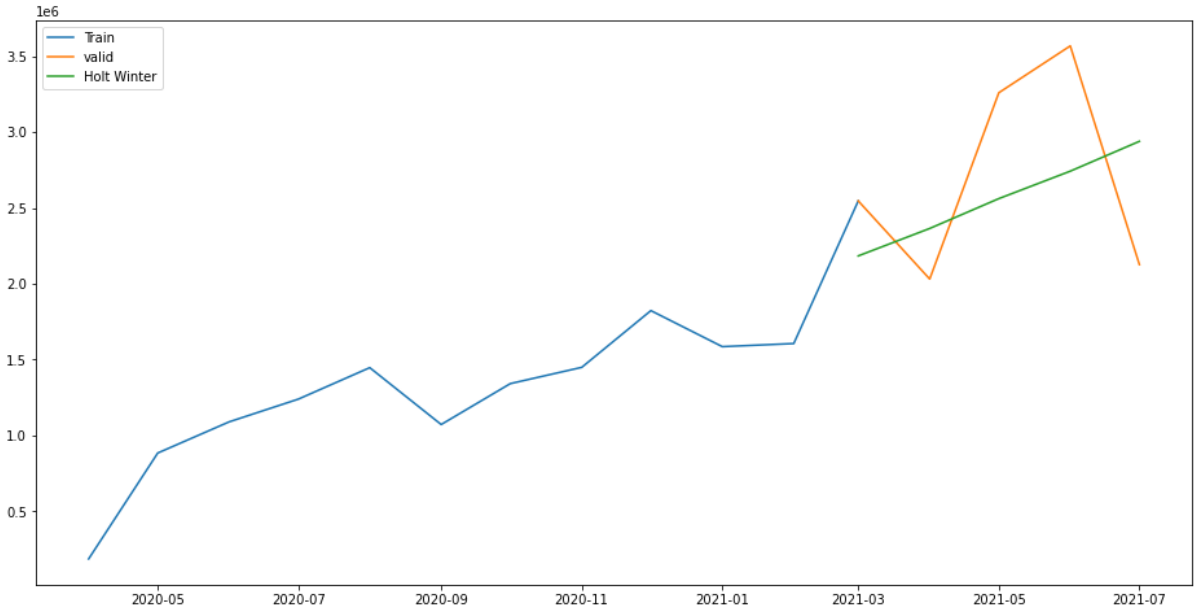
RMSE for this model is 629378.



5. Holt's Winter Model

The previous methods doesn’t take seasonality into account. This method takes trend and seasonality both to predict future fees.

RMSE of this model is 643377



These are the different models we have applied to predict future fees collected. We got good RMSE value for Holts Winter Model i. 643377. It can be further improved by collecting more data since we have only fees collected for 15 months only.

We have done forecasting on number of sales for future months also using the same methods.