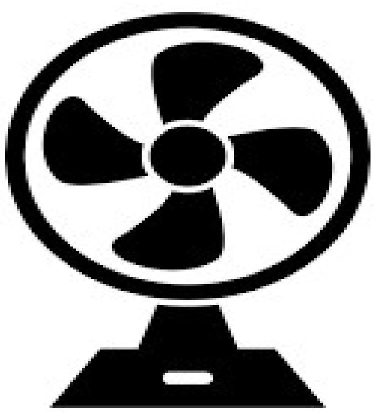
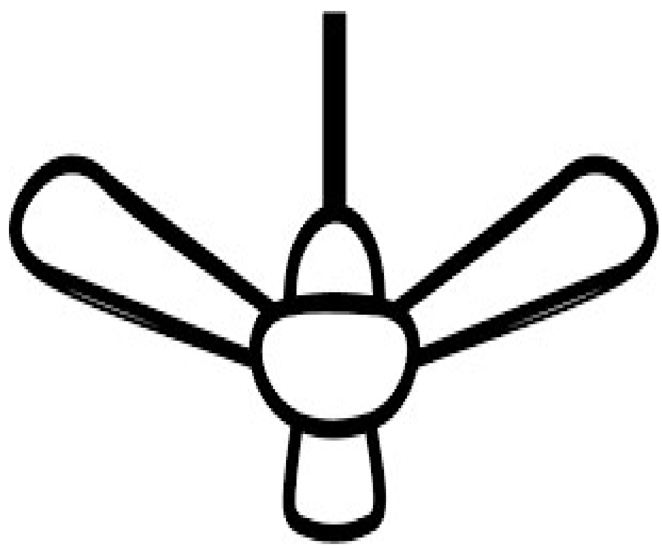
# TEAM 278





# INDEX

# Introduction:

### Industry:

A company manufactures electrical fans and sells them to consumers all-over India through enormous dealers in four regions: North, South, East, and West. Each region has its warehouse, and dealers place orders in their respective region's warehouses. The dealer places orders to the region-specific warehouse,which is fulfilled by the inventory available when receiving the order. The company wants to improve sales forecasting to maintain the demand and supply ratio.

### Task:

The task is to improve forecasting using advanced time series and machine learning algorithms. Sales forecast for the month of June-2021 has to be madeusing the train data available from the time period of April-18 to May-21.

### Approach:

We have used data visualization tools to generate insights to support our forecasting models. We have developed various time series forecasting modelsand evaluated them on a validation set for model selection. We have applied models ranging from the simple data imputation regression models to the complex ones like LSTM, ARIMA, SARIMA, SES, HLT, FB Prophet to minimize the error (MAPE).

# Data Description and Scoring:

There are a total of four features in the dataset, and they are as follows:

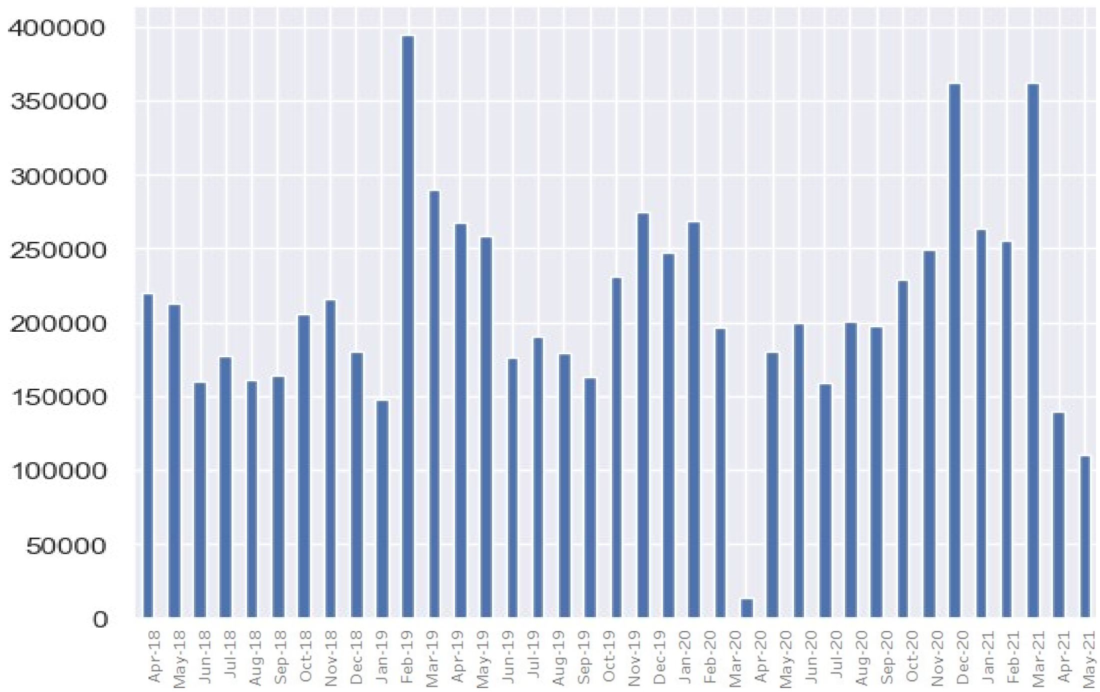
- Warehouse ID: This column contains numbers 1,2,3,4, which indicates four different warehouses
- Region: This has the direction to which the warehouse supplies fans.
- SKU ID: Each id is unique and corresponds to different warehouses.
- Apr-18 to May-21: Each month in this period has the number of fans sold.

The scoring is based on Mean Absolute Percentage Error (MAPE), which is defined as follows:

MAPE =
ABS (Actual Sales - Forecasted Sales) / Actual Sales, if Actual Sales > 0
0%, if Actual Sales and Forecasted Sales are both 0
100%, if Actual Sales = 0 and Forecasted Sales < > 0)

# Data Visualization:

1. **Total Monthly Sales:**

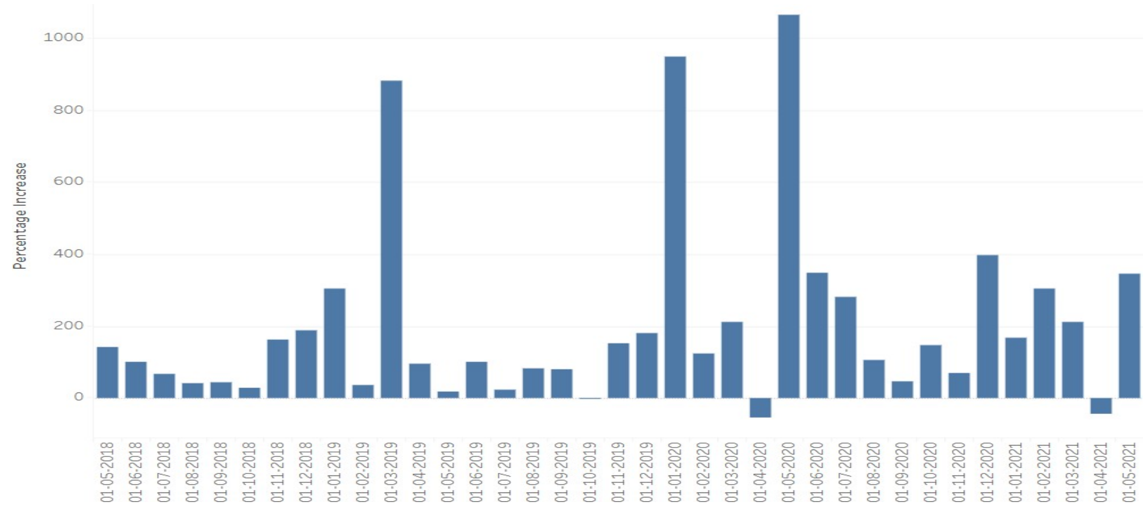



**Observations:**

- There is a dip in sales in April 2020, which indicates the pandemic lockdown.
- April and May 2021 have also been impacted by the second wave ofthe pandemic lockdown, which is indicated by the dip in sales.

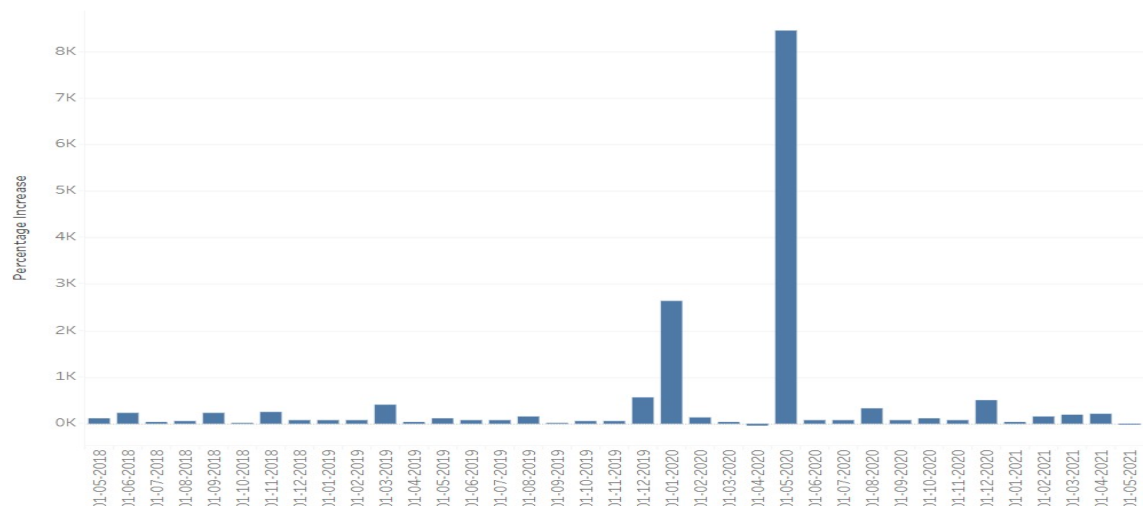
**2. Region-wise Sales Rolling Rate of Change:**

a. North Region:



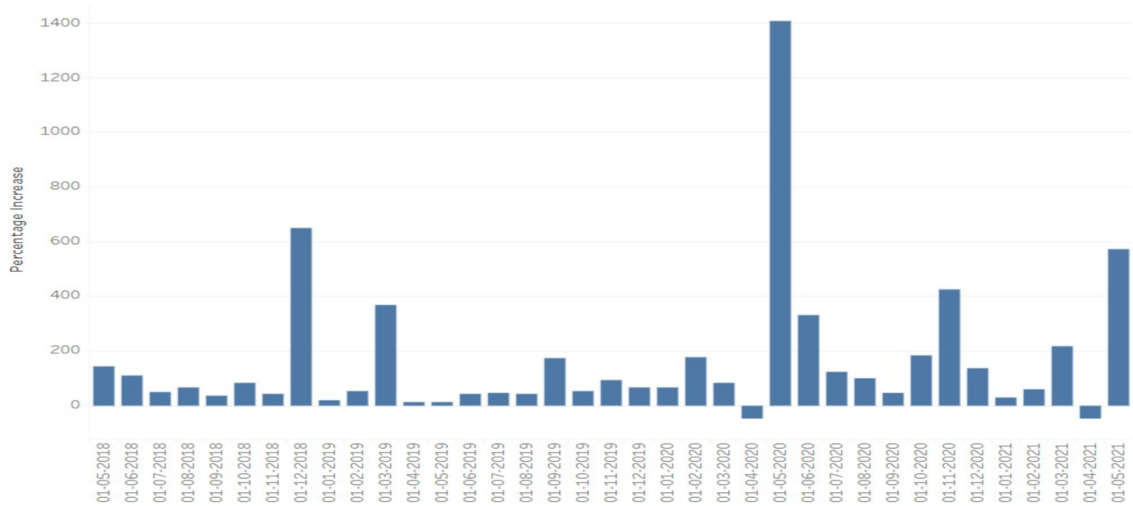

- The sales have decreased in April 2020 and April 2021 due to COVID-19 pandemic and strict nationwide lockdown.

b. East Region:



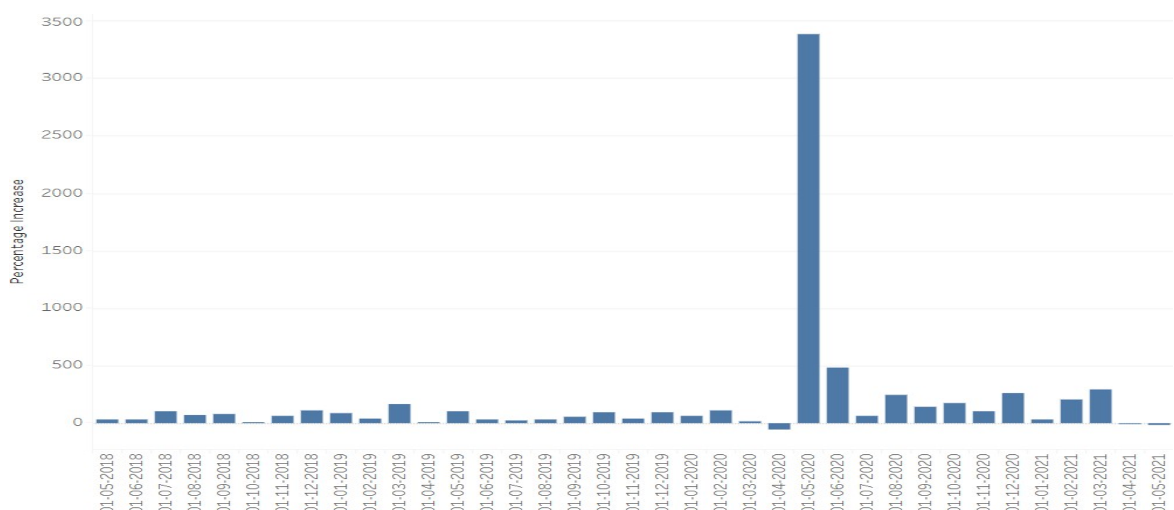

- Due to COVID-19 pandemic and nationwide Lockdown, nearly zero sales could be seen in March-2020, April-2020, and April-2021.

c. West Region:



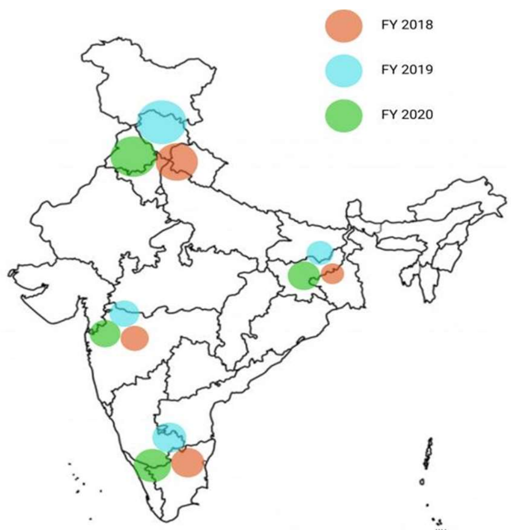

- The Sales have decreased in April 2020 and April 2021 due to COVID-19 pandemic and strict nationwide lockdown.

d. South Region:




- The Sales have decreased in April 2020 and April 2021 due to COVID-19 pandemic and strict nationwide lockdown.

### 3. Geo-visualization:

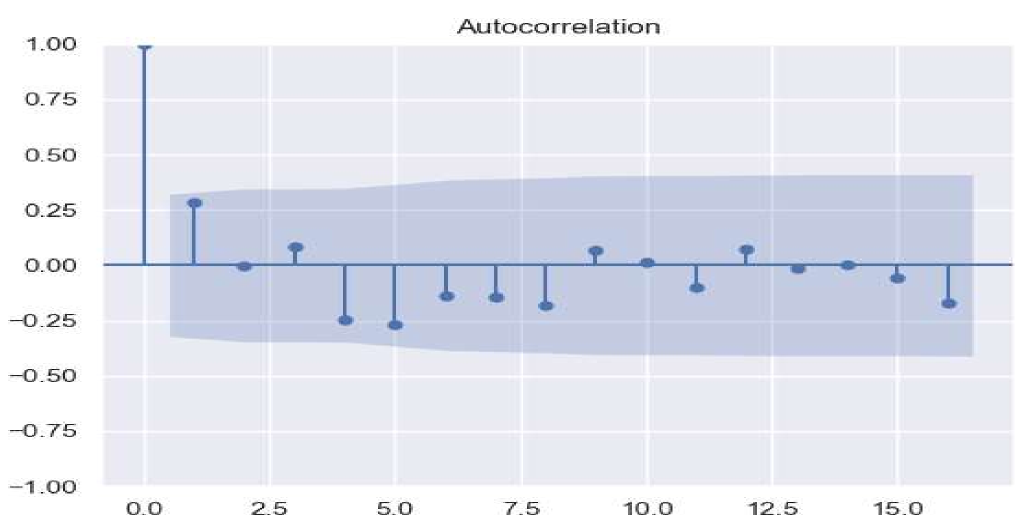



Total Sales by Region and Financial Year

## Exploratory Data Analytics:

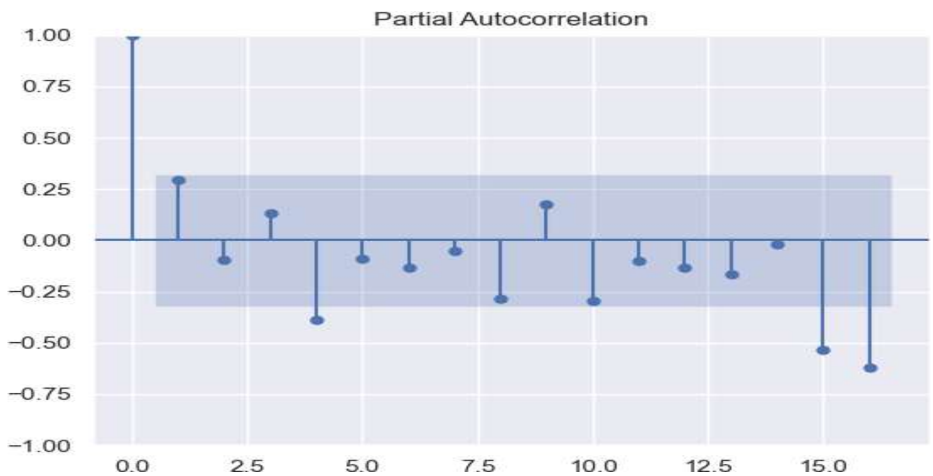
### 1. Autocorrelation Plot:

An autocorrelation plot is designed to show whether the time series elements arepositively correlated, negatively correlated, or independent of each other.



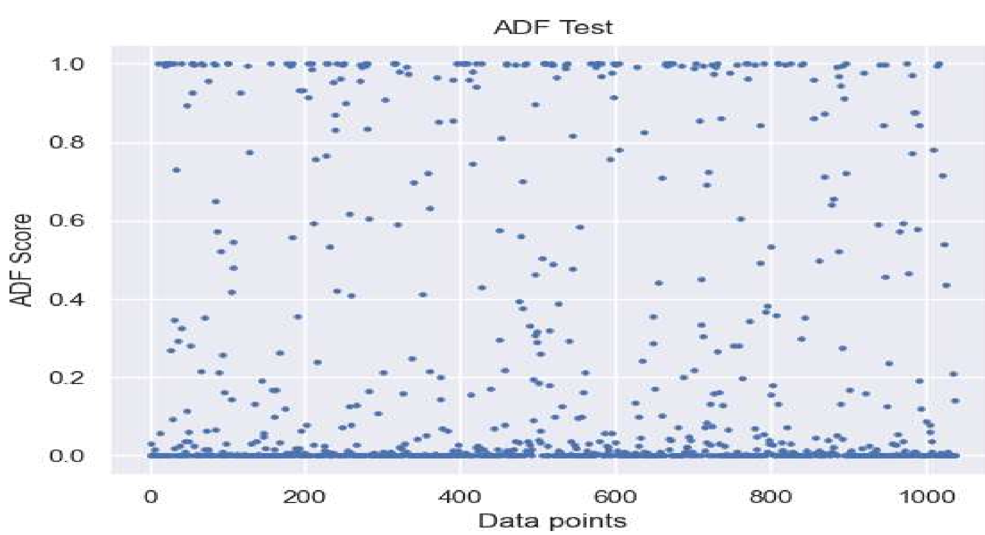
**2. Partial Autocorrelation Plot:**

A partial autocorrelation summarizes the relationship between an observation ina time series with observations at prior time steps with the relationships of intervening observations removed.

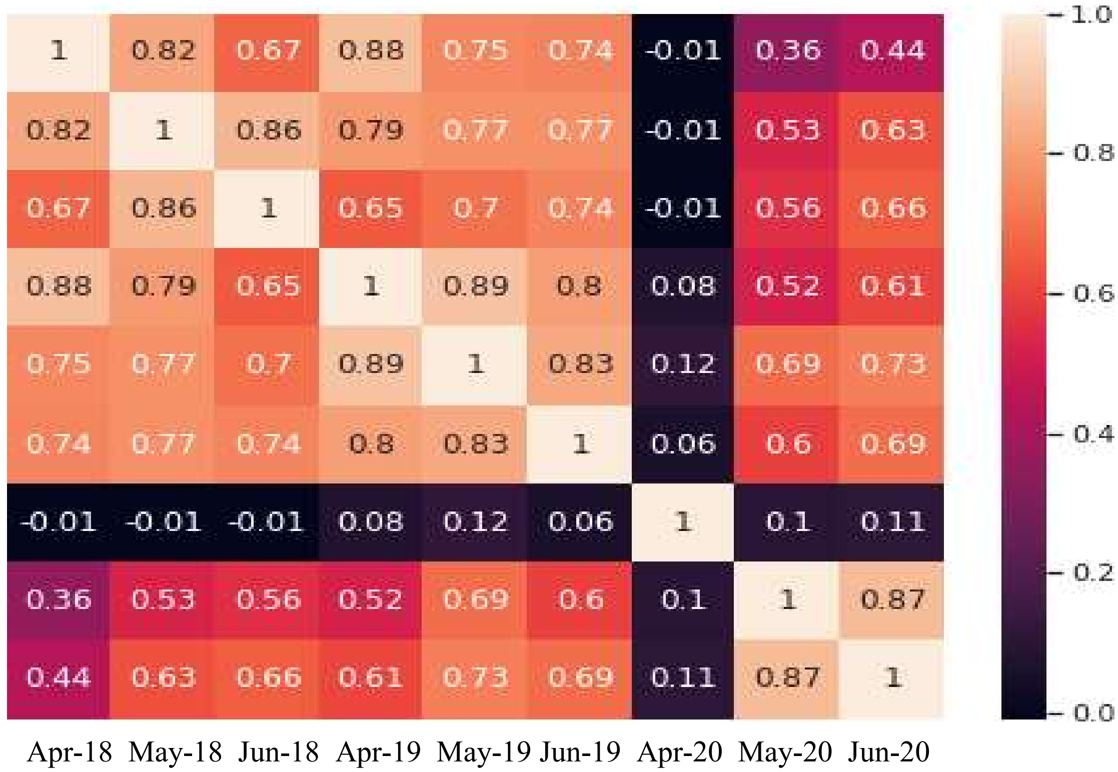



**3. Augmented Dickey-Fuller (ADF) Test:**

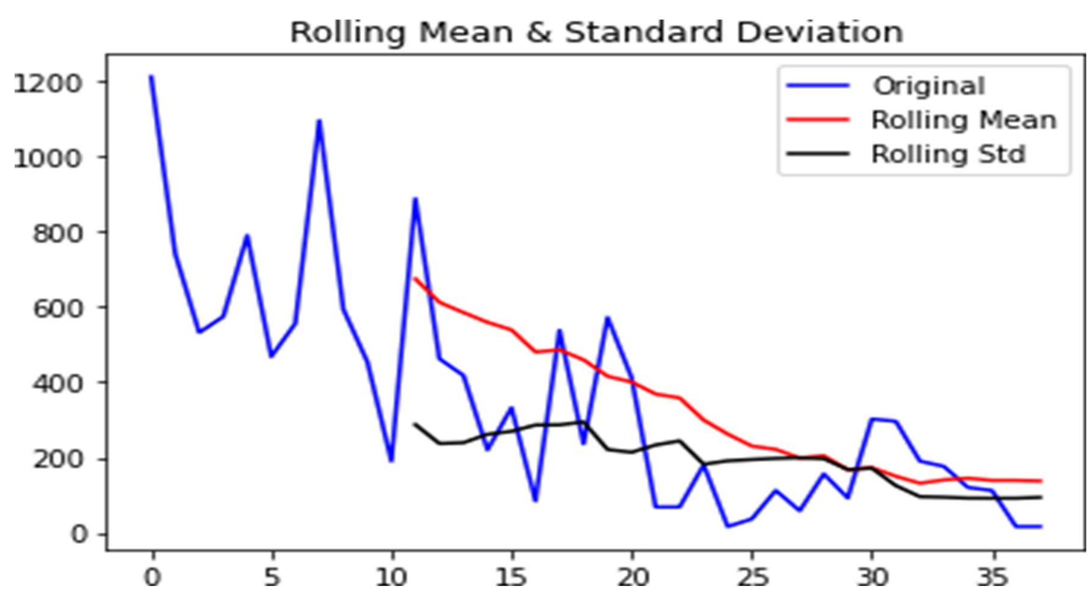
An augmented Dickey-Fuller test (ADF) on each row of the dataset shows that some of the data (about 20%) are highly varying (ADF p-value > 0.8), and about25-30% of data is way more stationary (constant sales for most of the months except the last months) than expected, which can also be seen from the following examples.



### 4. Correlation Heatmap:




### 5. Rolling Mean and Rolling Standard Deviation:



# Models:

## 1. Machine Learning Models:

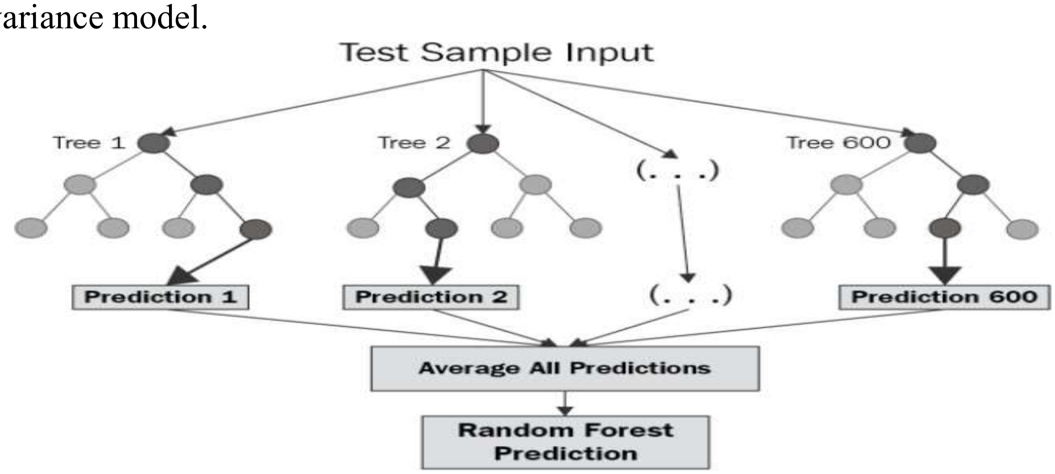
In Machine Learning, we use various algorithms to allow machines to learn the relationships within the data provided and make predictions based on patterns orrules identified from the dataset. So, regression is a machine learning technique where the model predicts the output as a continuous numerical value.

Regression analysis is often used in finance, investing, and others and finds out the relationship between a single dependent variable (target variable) dependenton several independent ones.

### a. Random Forest Regression:

A Random Forest is an ensemble technique that uses several decision trees anda method called Bootstrap and Aggregation, also known as bagging, to solve both regression and classification problems. The main idea is to combine multiple decision trees to determine the outcome.

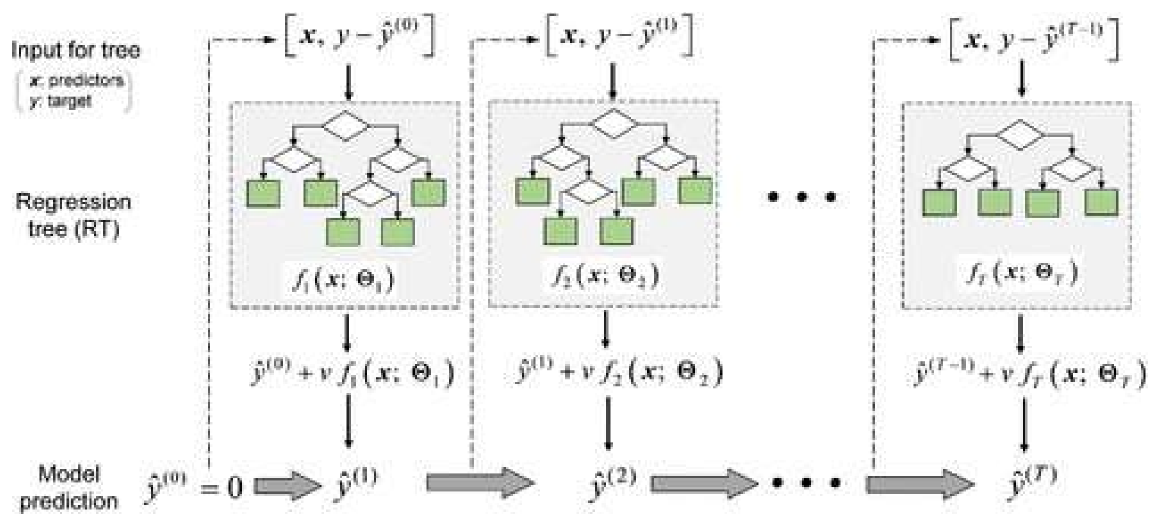
Random Forest is a good model when it comes to high performance when compared to interpretation. It is parallelizable, which means that the data can besplit into multiple machines to run, resulting in faster computation time. It is excellent with high dimensionality quick prediction and handles unbalanced data. Using this, we can create a low bias and moderate variance model.



**b. XG Boost Regressor:**

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. XGBoost is an efficient implementation of gradient boosting that can be used for predictive regression modeling. It is designed to be both computationally efficient (e.g., fast to execute) and highly effective, perhaps more effective than other open-source implementations.

It is an approach where new models are created that predict the errors or residuals ofprior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.




## 2. Statistical Models:

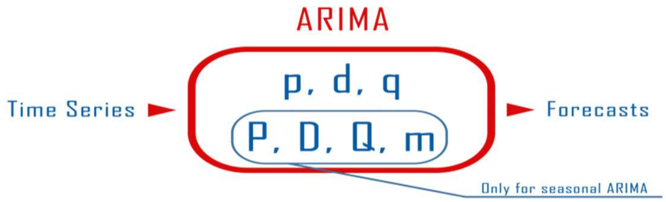
Statistical forecasting implies using statistics based on historical data to projectwhat could happen in the future.

**a. Autoregressive Integrated Moving Average (ARIMA):**

Autoregressive Integrated Moving Average or ARIMA utilizes data that varieswith time, uses statistical analysis, and predicts future trends. ARIMA model utilizes the lagged moving average, which smoothens the time-varying data. This model assumes that the future trend is based on past values.

The ARIMA model has parameters of p, d, and q, which are standard notations. ‘p’ representsthe number of lag observations in the model or the lag order. ‘d’ denotes the degree of differencing or the number of times the observations are different. ‘q’ denotes the order of the moving average.

**Shorthand Notation:**




**Formula:**

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_0 Y_0 + \epsilon_t$$

$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + .. + \beta_0 Y_0 + \epsilon_{t-1}$$

**b. Simple Exponential Smoothing (SES):**

It is a type of time series forecasting that does not require seasonality or trendbut univariate data. It takes only a single parameter, ‘ɑ’ alpha, called the smoothing factor. It mainly sets a value between 0 and 1. For a considerable value, it indicates that the model pays primary attention to only recent past observations, and for smaller values, its history is taken into consideration.

**Formula:**

$$F_{t+1} = \alpha D_t + (1 - \alpha) F_t$$

**c. Seasonal Autoregressive Integrated Moving Average (SARIMA):**

Seasonal Autoregressive Integrated Moving Average or SARIMA varies from the ARIMA model in only one step- it utilizes the seasonal trends. This featureis significant in any time series model as repeated seasonal effects are possible.The parameters are the same as ARIMA with an extra parameter ‘m’. It represents the number of time steps for a single seasonal period.

It also takes into account seasonal differencing. It means the difference of values of the current month and its value at the previous season. When plottingand observing the data, seasonal differencing is hugely significant as it helps usunderstand the repeated trend.

**Shorthand Notation:**

$$\text{ARIMA}\ (p, d, q) \times (P, D, Q)S$$

**Formula:**

$$\Phi(B^S)\phi(B)(x_t - \mu) = \Theta(B^S)\theta(B)w_t$$

**d. Holt Linear Trend:**

Holt’s linear trend method is also known as Double exponential smoothing. Holt extended simple exponential smoothing to allow forecasting of data with atrend. It is nothing more than exponential smoothing applied to both levels (theaverage value in the series) and trends. We can use additive and multiplicative seasonality in Holt’s linear trend method. It takes two parameters, one for the overall smoothing and the other for the trend smoothing equation.

**Formula:**

$$\begin{aligned} \text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t + hb_t \\ \text{Level equation} \quad & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ \text{Trend equation} \quad & b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \end{aligned}$$

**3. FB Prophet:**

The Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends fit with yearly, weekly, and daily seasonality andholiday effects. It works best with time series with strong seasonal effects and several seasons of historical data. Prophet is robust to missing data, and trend shifts and typically handles outliers well. The input to Prophet is always a data frame with two columns: ds (datestamp) and y is the measurement we wish to forecast (number of products in our case).

**Formula:**

$$\mathbf{y(t) = g(t) + s(t) + h(t) + e(t)}$$

- g(t) refers to trend (changes over a long period of time)
- s(t) refers to seasonality (periodic or short-term changes)
- h(t) refers to the effects of holidays on the forecast
- e(t) refers to the unconditional changes specific to a business, person, or circumstance. It is also called the error term.
- y(t) is the forecast.

**4. Deep Learning Models:**

Deep learning is a subset of machine learning meant to improve the overallaccuracy of the basic algorithms. These models help us to achieve high accuracy and perhaps exceed human-level performance. These models are trained by inputting extensive data and are adjusted to improve overall accuracy.

**a. Long Short-Term Memory (LSTM):**

Long Short-Term Memory Network or LSTM is a Recurrent Neural Network that allows continuous data or data to continue. This model helps recognize audio, videos, or any other flowing data. A significant feature of this model is that it eliminates the vanishing gradient problem. The significance of the modelin this problem is that a future prediction is required from the continuous data.

There are several hyperparameters in LSTM. The number of nodes and hidden layers is a crucial parameter to tune. It represents the number of layers between the input and output layers. One must be careful not to increase it too much as it might overfit.

The Learning rate is another crucial hyperparameter. It represents how quickly the model updates its parameters for better accuracy.

## Final Model:

- The methods mentioned above were tried in our initial approach. The likes ofRandom Forest, XG boost regressor, ARIMA, SES, SARIMA, HOLT linear trend and FB prophet turned out to be much more complex when compared toour final approach.

- To arrive at our final model, we first compared sales of consecutive months andsaw very strong correlations between them. This further led us to explore the heat map and we saw significant correlations upto the previous two months.

- We saw a very reduced MAPE when just predicting this month's sales taking the previous month's sales as reference when compared to the different complexmodels mentioned above.

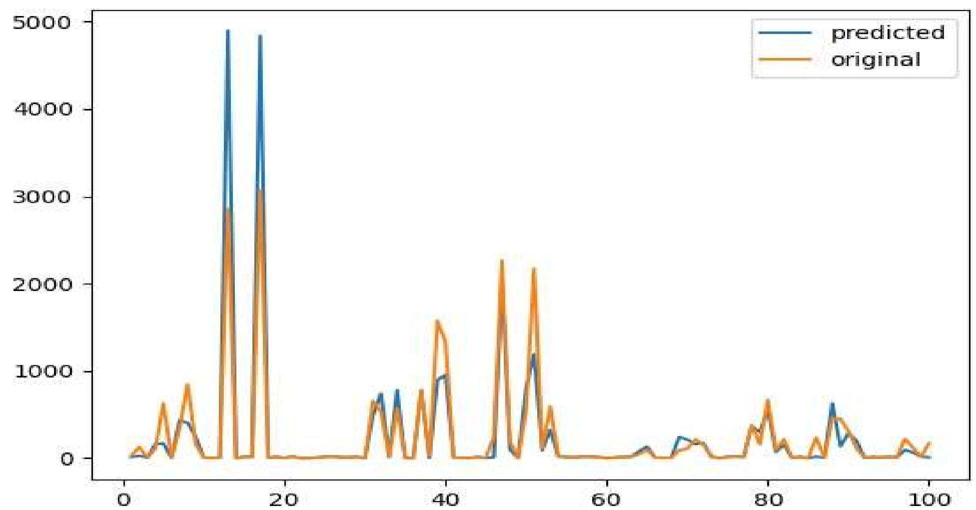
- Now the question is should we predict this month's sales (June's data) as the most recent month's sales (May's sales)? Average of last two months sales ((May + April) / 2)? April's sales? Or any other metric?

- We also had the knowledge of the fact that the MAPE favours under-forecast rather than over-forecast. This is caused by the fact that the percentage error can't exceed 100% for forecasts that are too low, while there is no upper limitfor forecasts which are too high.

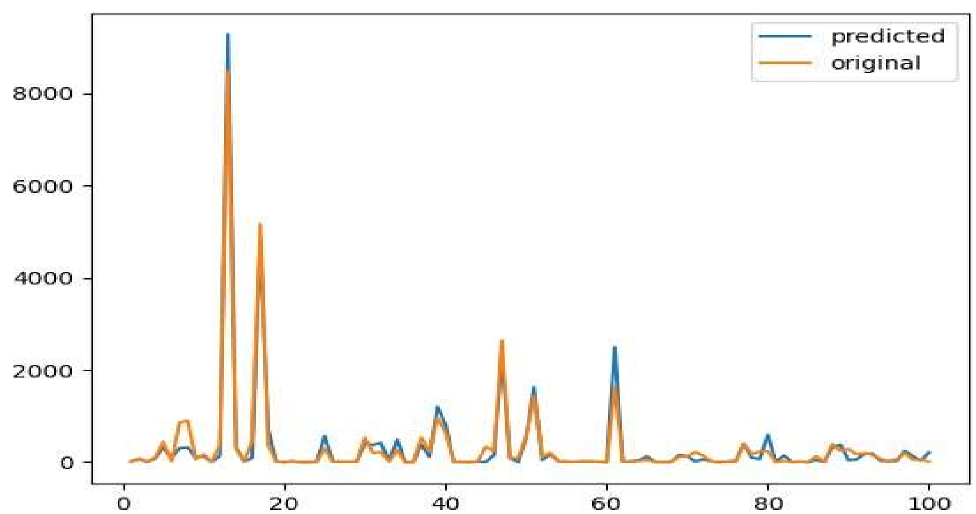
- So to minimise our prediction we went with the following simplistic approach:Predicted June's sales = Minimum (April's sales, May's Sales)

- This boosted our performance and we got an even less MAPE with this approach. For May 2021 we got MAPE = 46.55 %

**For 2018**




**For 2019**

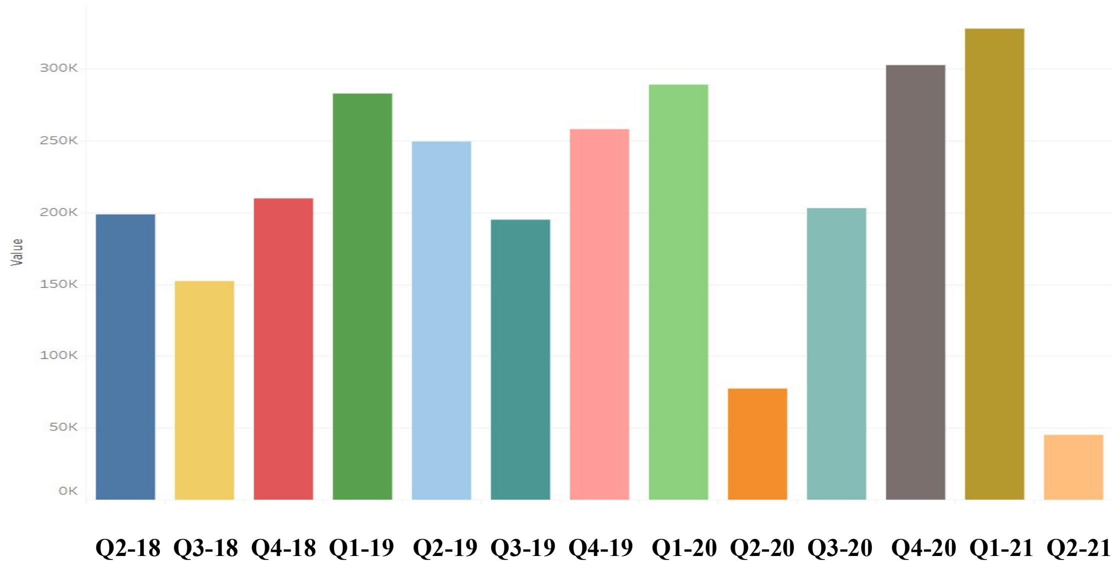


# ANNEXURE

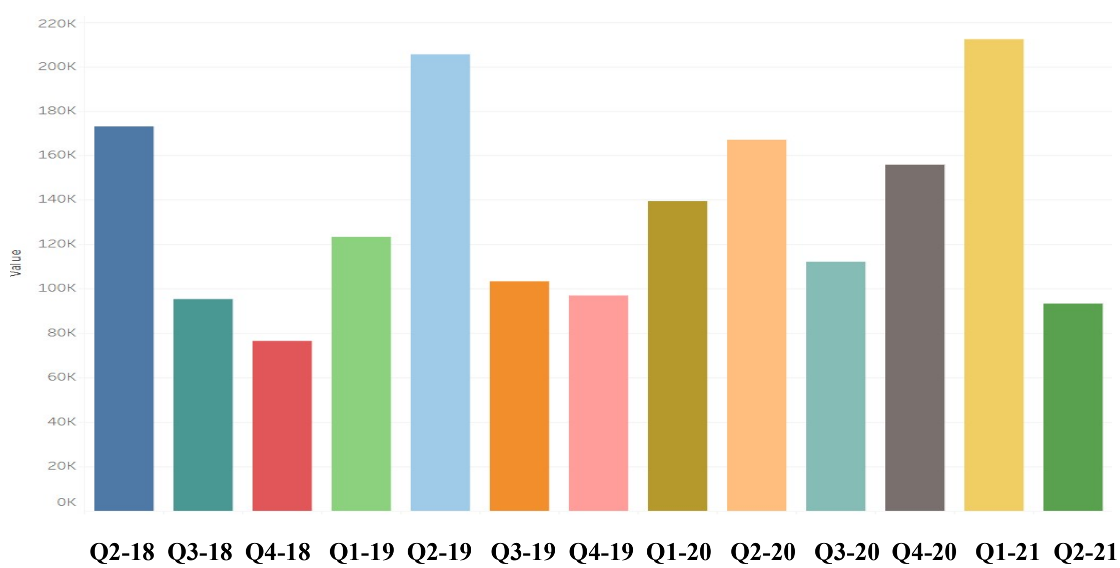
## A. Further Graphical Analysis:

### 1. Region-wise Quarterly Sales

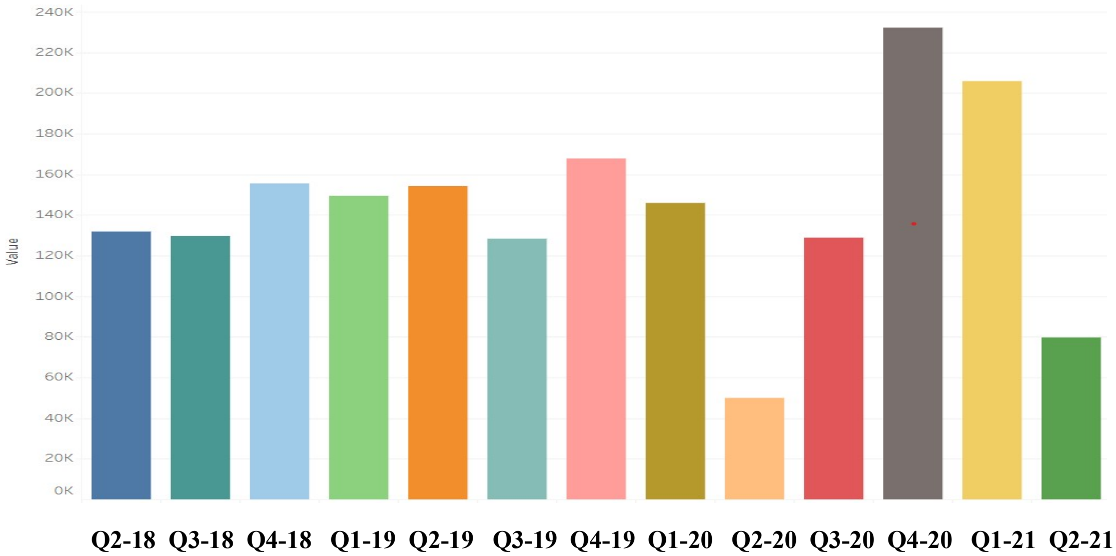
a. North Region



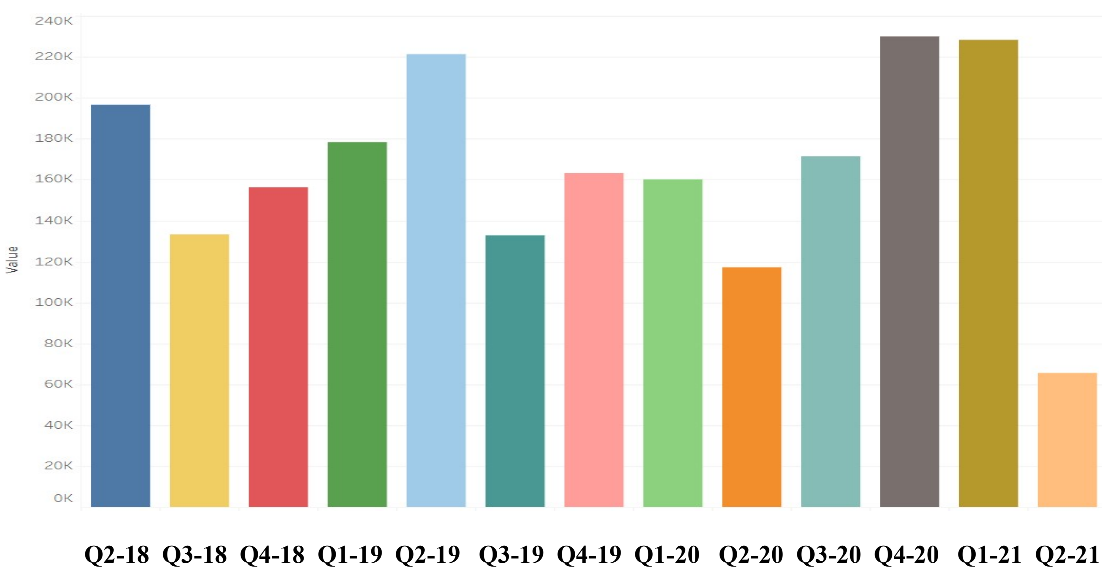

b. East Region:



c. West Region:




d. South Region:



# References:


1. Machine Learning Techniques for Sales Forecasting
2. TIme Series Forecasting with Deep Learning
3. XGBoost Documentation
4. XGBoost Regressor Implementation
5. Stats Model Documentation
6. Stats Model Implementation
7. FB Prophet Documentation
8. FB Prophet Implementation
9. LSTM Colah's Blog
10. LSTM Implementation
11. LSTM performs better on small data