WALMART SALES

Presentation by Farah Moataz

Ain Shams Univeristy

INTRODUCTION

Predicting future sales for a company is one of the most important aspects of strategic planning.
In this project, we wanted analyze in depth how internal and external factors of one of the biggest companies in the US can affect their Monthely Sales in the future.

DATA

THE WALMART DATASET CONSISTS OF 21 COLUMNS WITH 9,995 NUMBER OF ROWS.

Row ID => Unique ID for each row.
Order ID => Unique Order ID for each Customer.
Order Date => Order Date of the product.
Ship Date => Shipping Date of the Product.
Ship Mode=> Shipping Mode specified by the Customer.
Customer ID => Unique ID to identify each Customer.
Customer Name => Name of the Customer.
Segment => The segment where the Customer belongs.
Country => Country of residence of the Customer.
City => City of residence of of the Customer.
State => State of residence of the Customer.
Postal Code => Postal Code of every Customer.
Region => Region where the Customer belong.
Product ID => Unique ID of the Product.

Category => Category of the product ordered.
Sub-Category => Sub-Category of the product ordered.
Product Name => Name of the Product
Sales => Sales of the Product.
Quantity => Quantity of the Product.
Discount => Discount provided.
Profit => Profit/Loss incurred.

PLAN OF ACTION

1. **We will build the following Regression models to predict future sales.**

2. **We will perform Time series analysis and gather useful insights.**

MODELLING & MACHINE LEARNING

LINEAR REGRESSION

we got

R^2 Score: 0.63763

DECISION TREE REGESSOR

we got

R^2 Score: 0.76575

RANDOM FOREST REGRESSOR

we got

R^2 Score: 0.8515

The Winner

STEPS

1-select features and target variable

```python
# Select features and target variable
features = df[['Ship Mode', 'Segment', 'Country','Region', 'Category', 'Quantity','Sub-Category','Discount','Profit','Da
target = df['Sales']
```

Python

2- splitting the data

```python
# Convert categorical variables into dummy/indicator variables
features = pd.get_dummies(features)
```

```python
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2, random_state=42)
```

STEPS

3-Random Forest Regression Model

```
from sklearn.ensemble import RandomForestRegressor


RFR=RandomForestRegressor()


RFR.fit(X_train, y_train)
```

```
RandomForestRegressor
RandomForestRegressor()
```

STEPS

4-Model Evaluation

```
    r2_score(y_test,y_pred)
34]

··   0.8515515040022 54
```

TIME SERIES

Objective of time series analysis is to understand how change in time affect the dependent variables and accordingly predict values for future time intervals.

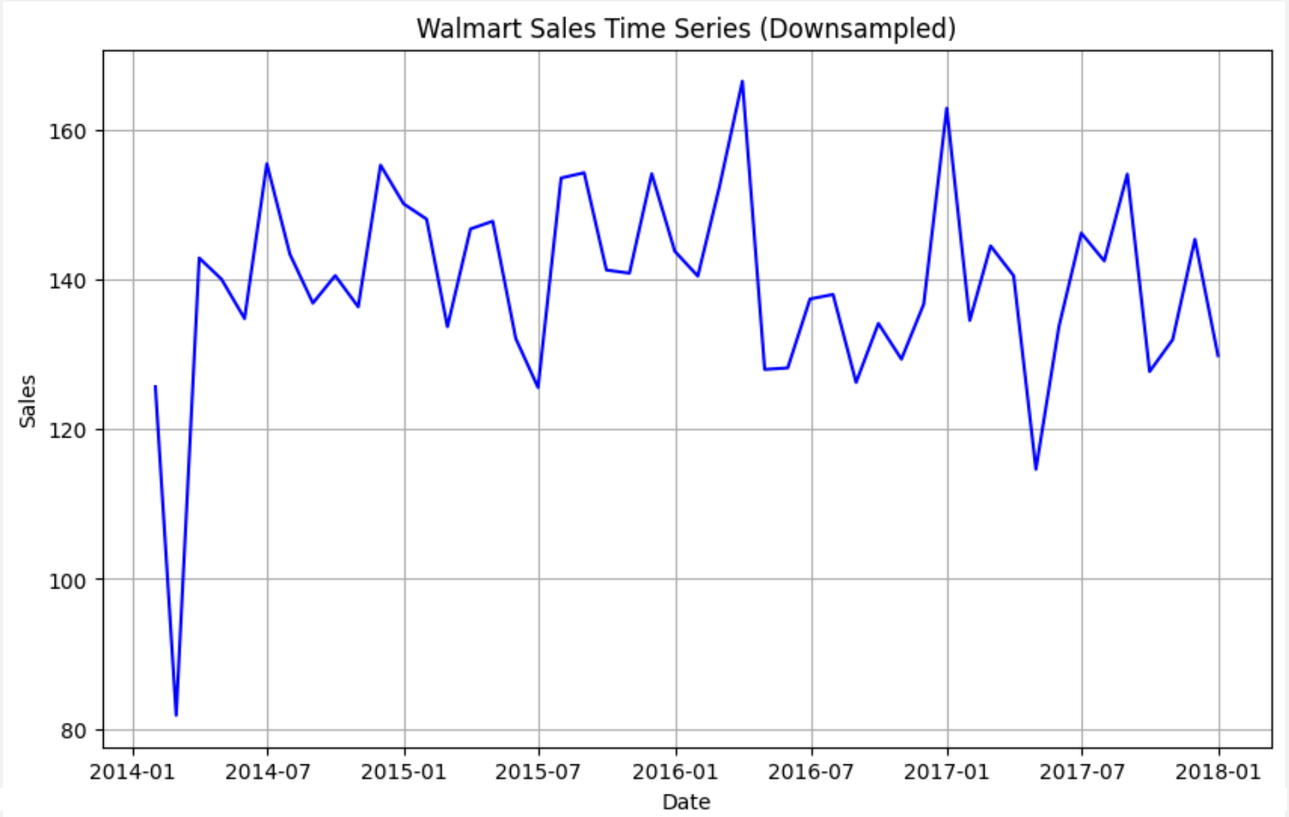
I will discuss the steps and the output of our dataset.

I used ARIMA model

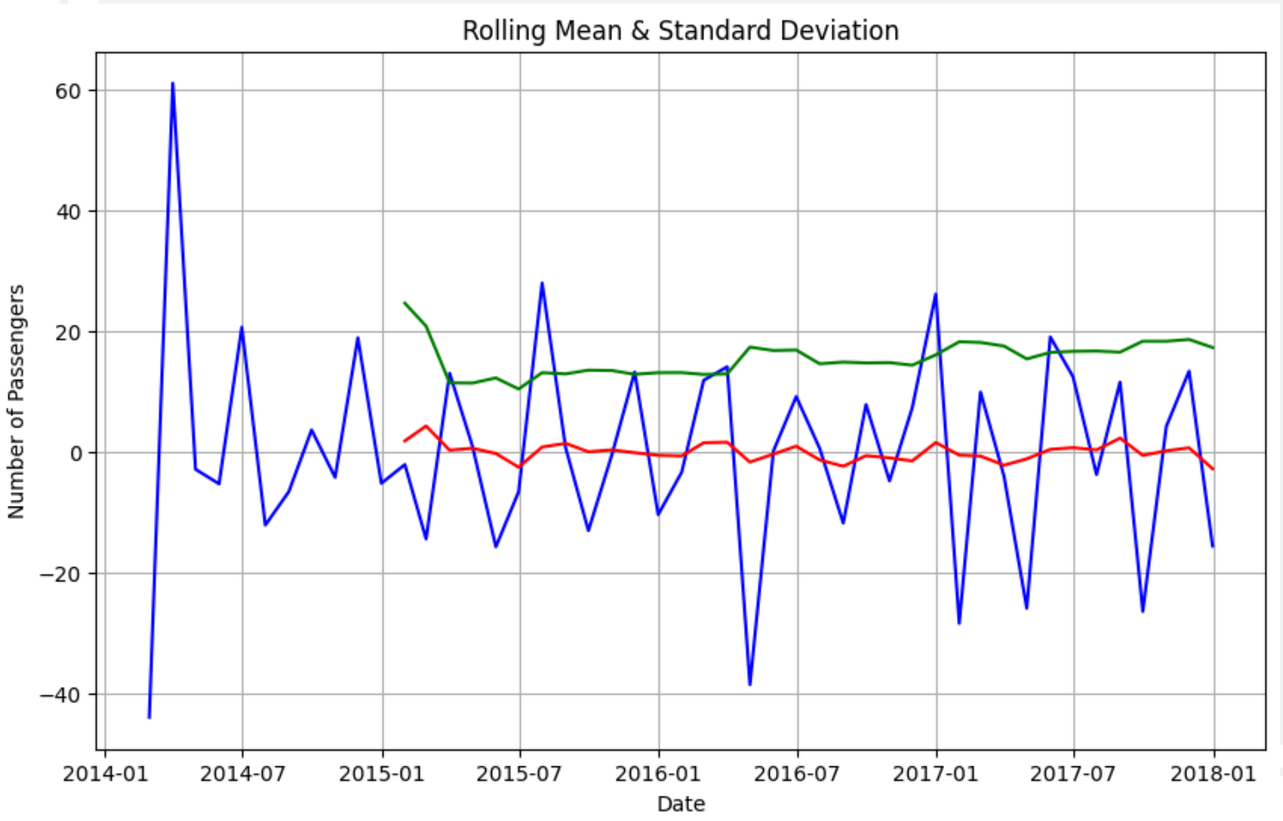
STEPS

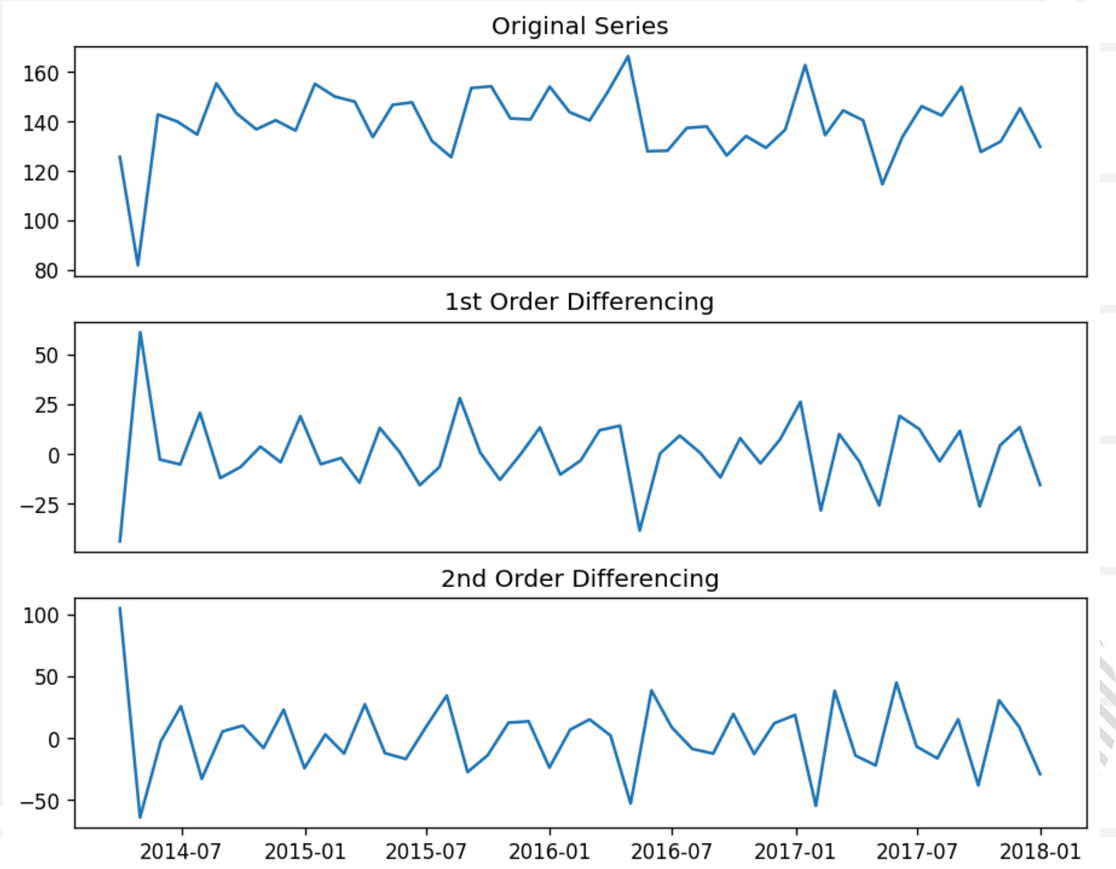
- STEP 1 : VISUALIZE THE TIME SERIES
- STEP 2: STATIONARIZE THE SERIES
- STEP 3: PLOT ACF/PACF CHARTS AND FIND OPTIMAL PARAMETERS
- STEP 4: BUILD THE ARIMA MODEL
- STEP 5: PREDICT

STEP 1 : VISUALIZE THE TIME SERIES

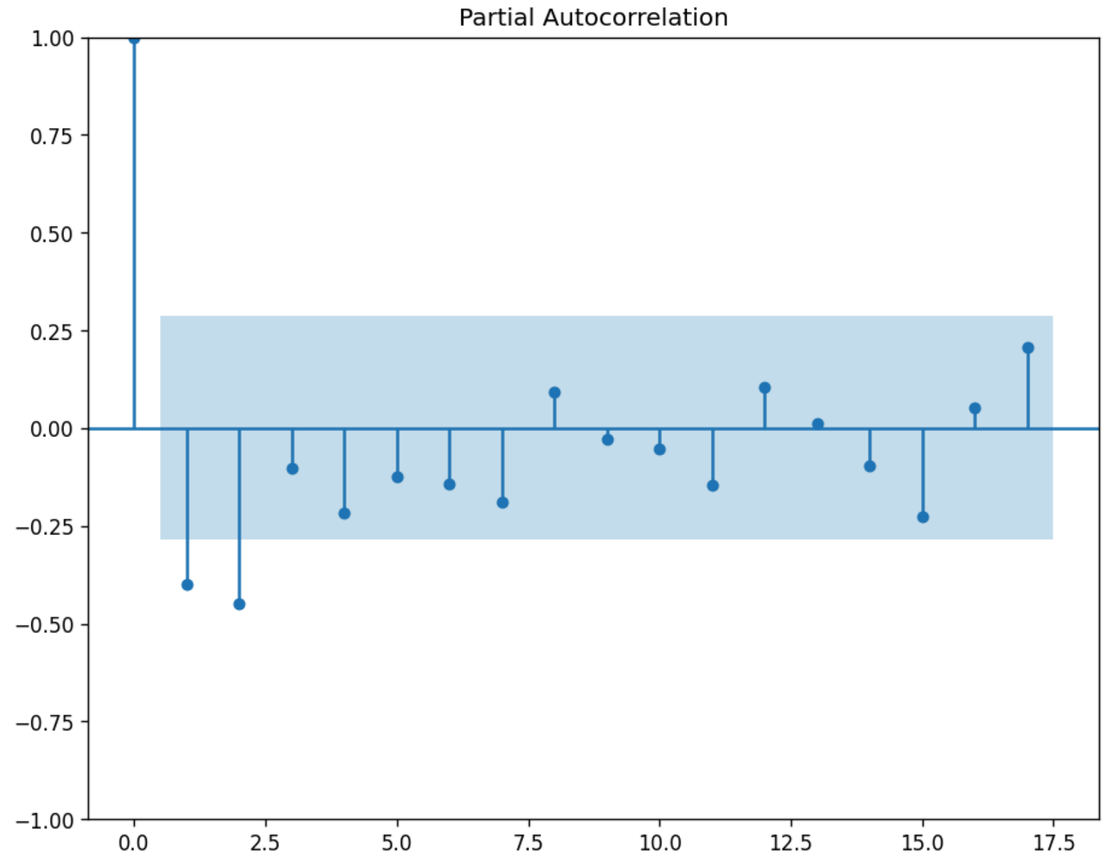


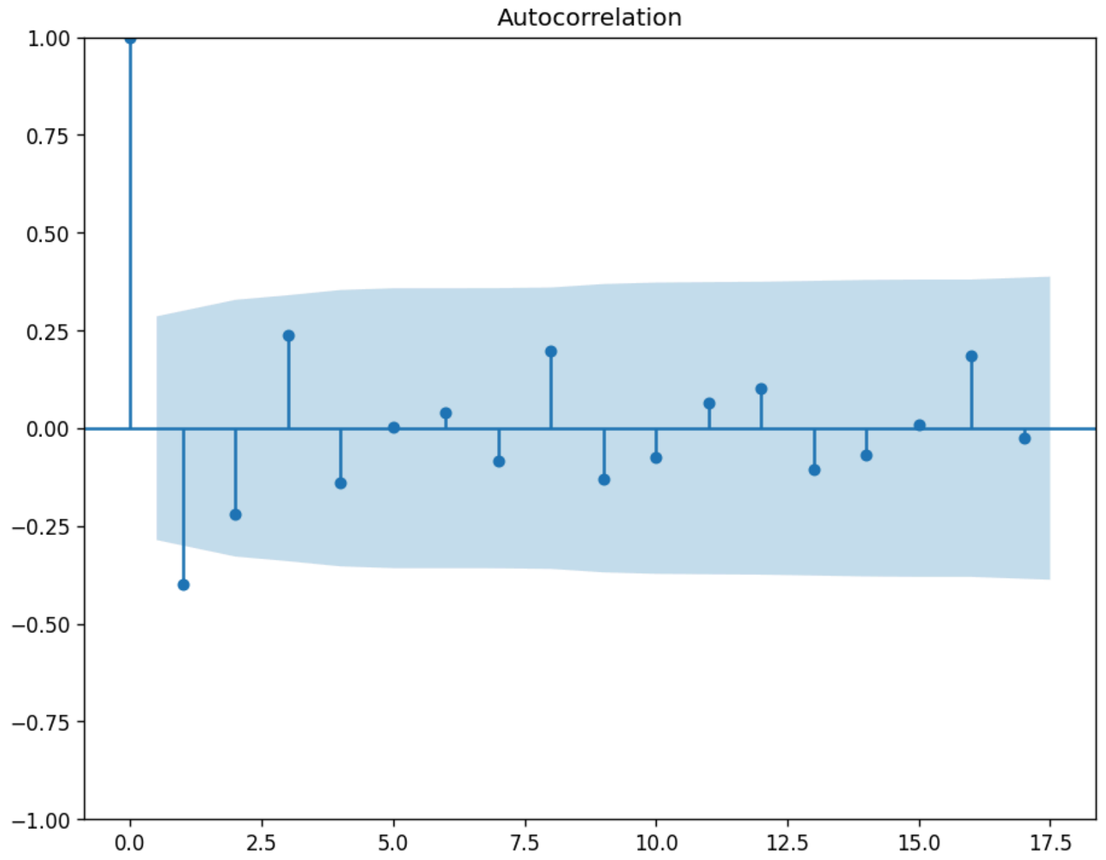
STEP 2: STATIONARIZE THE SERIES



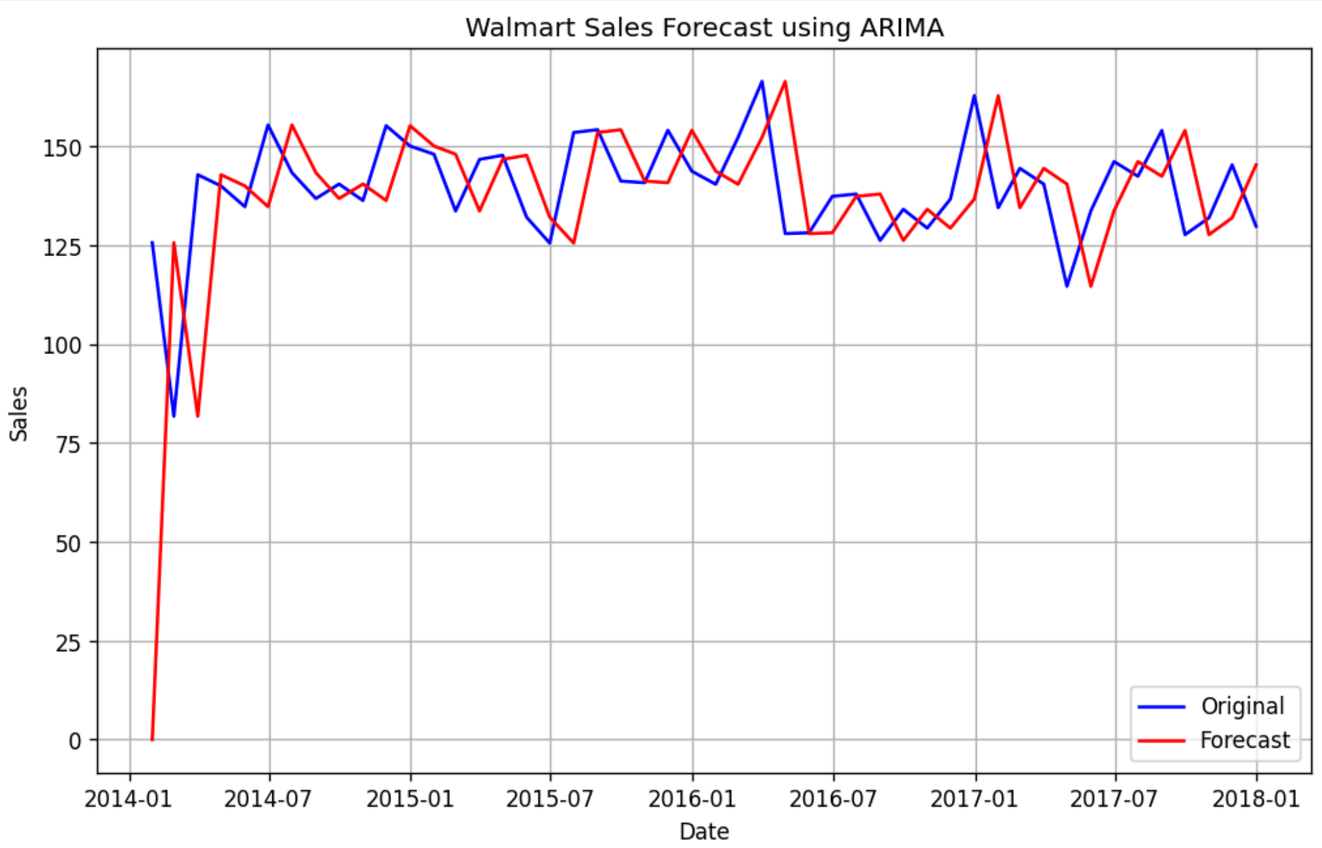


STEP 3: PLOT ACF/PACF CHARTS AND FIND OPTIMAL PARAMETERS

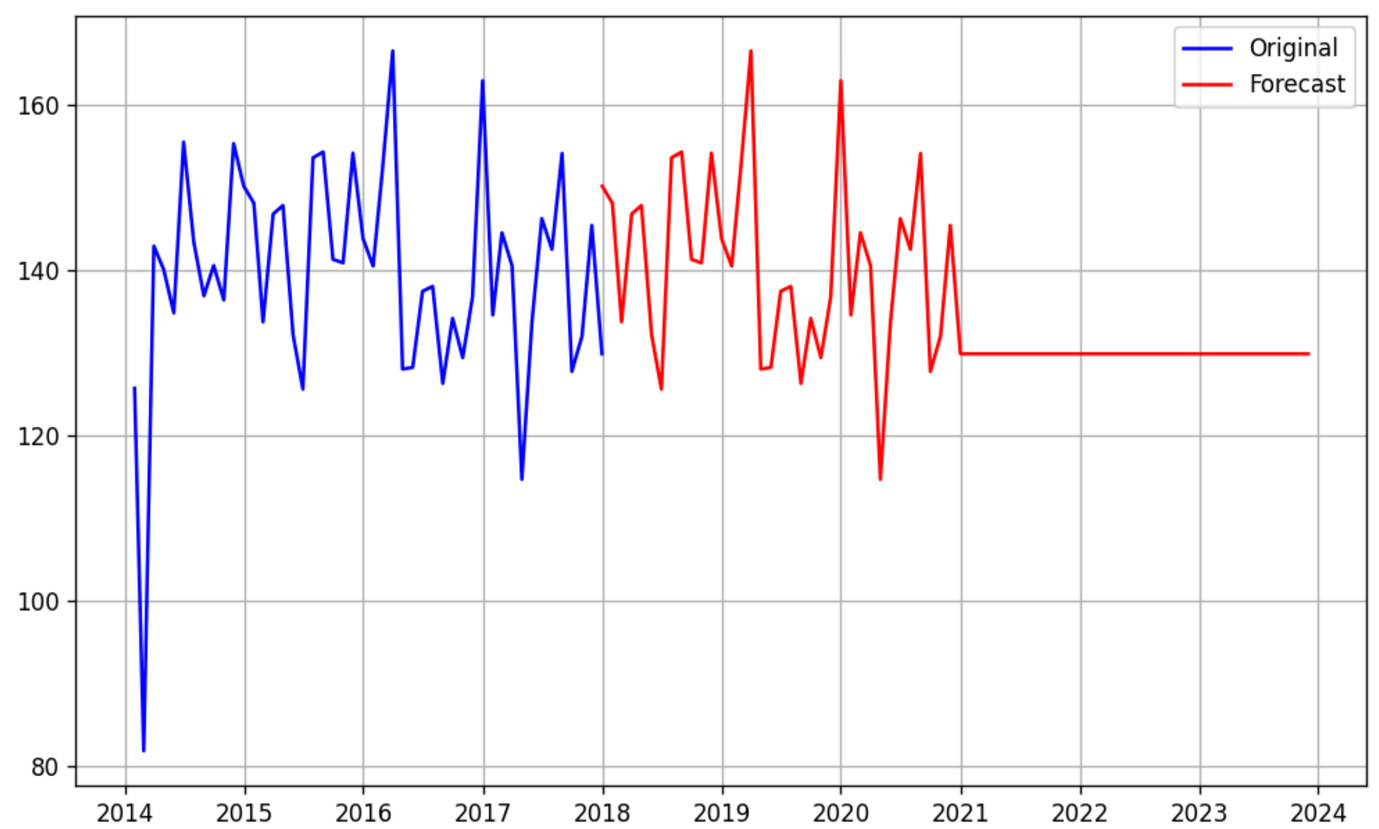




STEP 4: BUILD THE ARIMA MODEL



STEP 5: PREDICT



FOR MORE DETAILS

GitHub