FINAL PROJECT PRESENTATION

- Abdelrahman Abdelnasser
- Khalid Osama
- Mahmoud Ahmed
- Taha Mahmoud

SALES FORCASTING

WHAT & WHY

- This project focuses on sales forecasting for a retail superstore—a critical task for making informed decisions in inventory management, marketing, and operational planning. Using historical daily sales data, our goal is to build a model that predicts future sales trends and provides actionable insights to help the company reduce waste, improve efficiency, and better meet customer demand.
- We chose this problem because it represents a common and impactful business challenge in retail, and solving it offers both practical value and a chance to apply advanced time series forecasting methods in a real-world context.

WHY THIS DATASET

- We selected this dataset because it offers a rich and comprehensive set of historical data spanning approximately four years, which is a strong foundation for building reliable forecasting models. The dataset not only includes daily sales figures but also provides detailed information on products, regions, cities, quantities, discounts, and profit margins. This level of granularity allows us to explore various factors that influence sales trends and build a model that captures the complexity of real-world retail dynamics. Its depth and time range make it a well-suited choice for our forecasting objectives.

RESULTS OF EDA

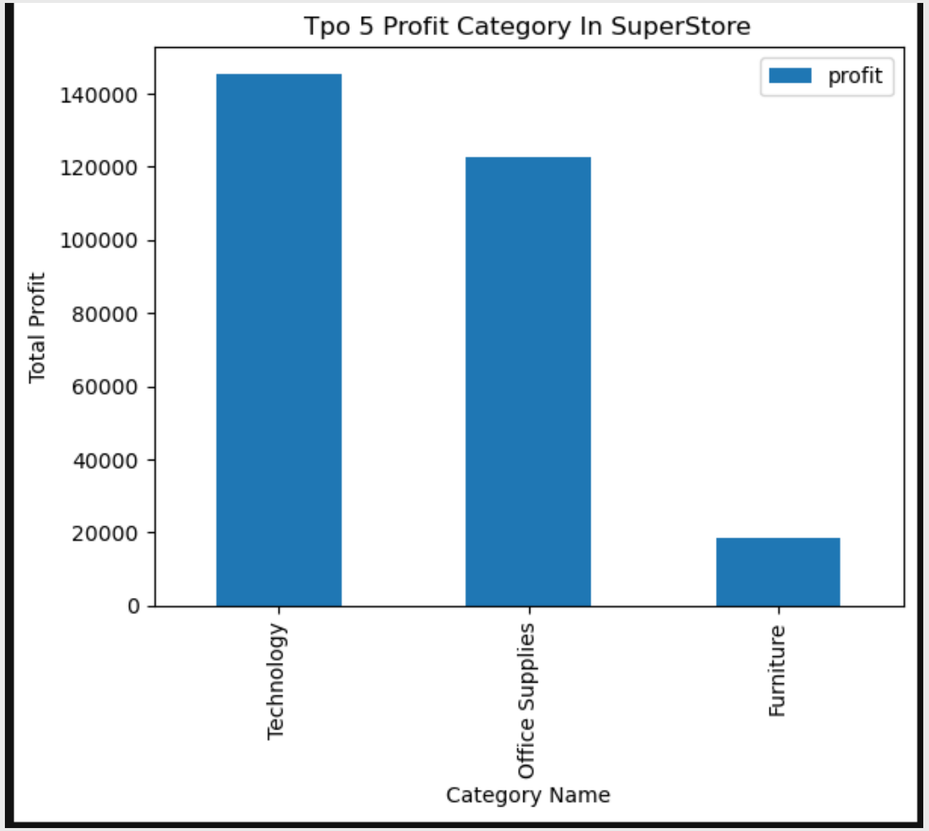
key findings include:

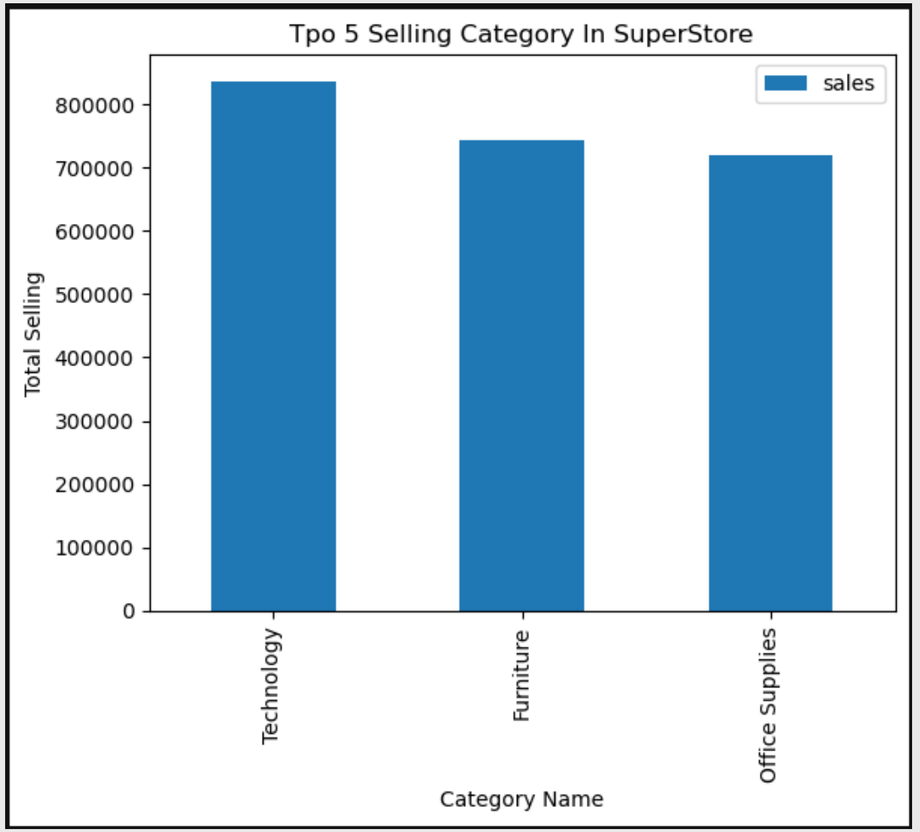
- the top 5 highest sales and most profitable products, categories and sub_catgories.

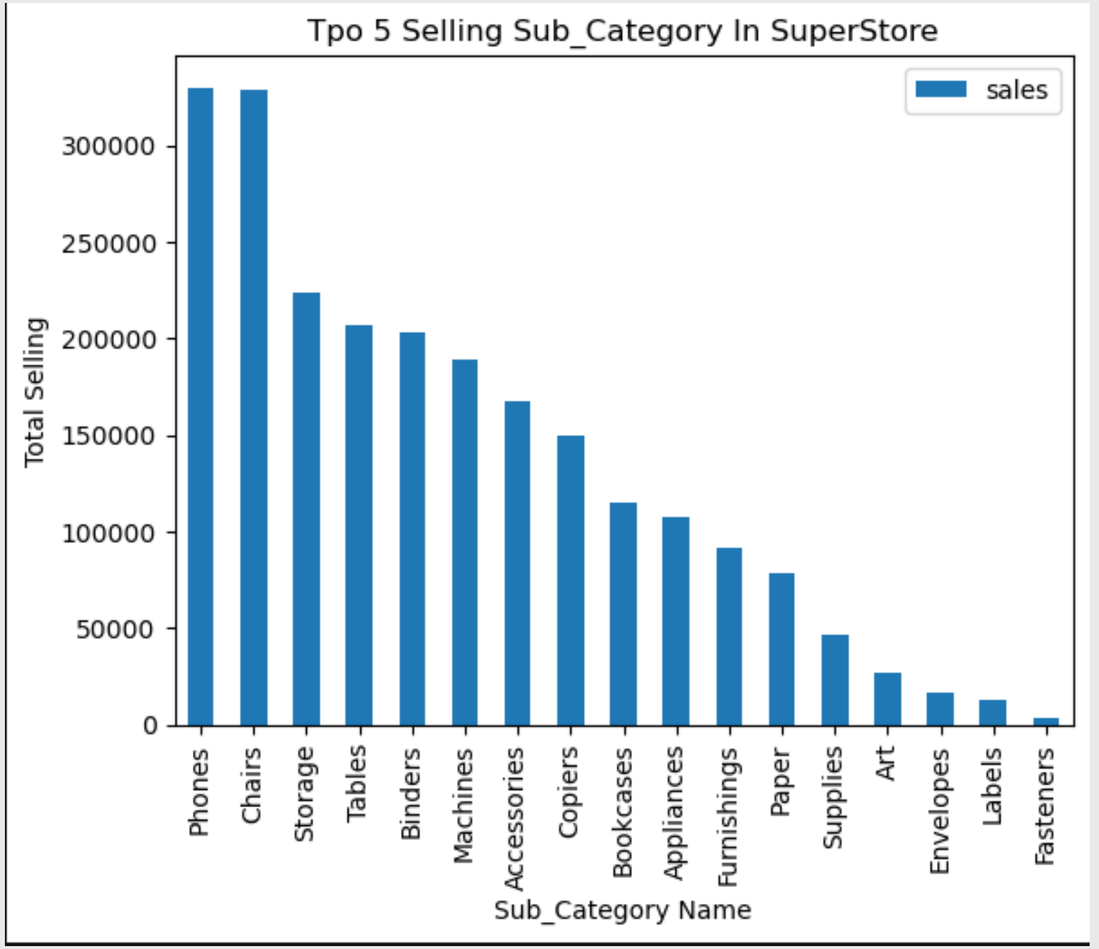
- Top-Selling Products:
- Canon Camera
- Electric Plastic Comb
- Cisco TelePresence System
- Chair
- ElectricBinding System

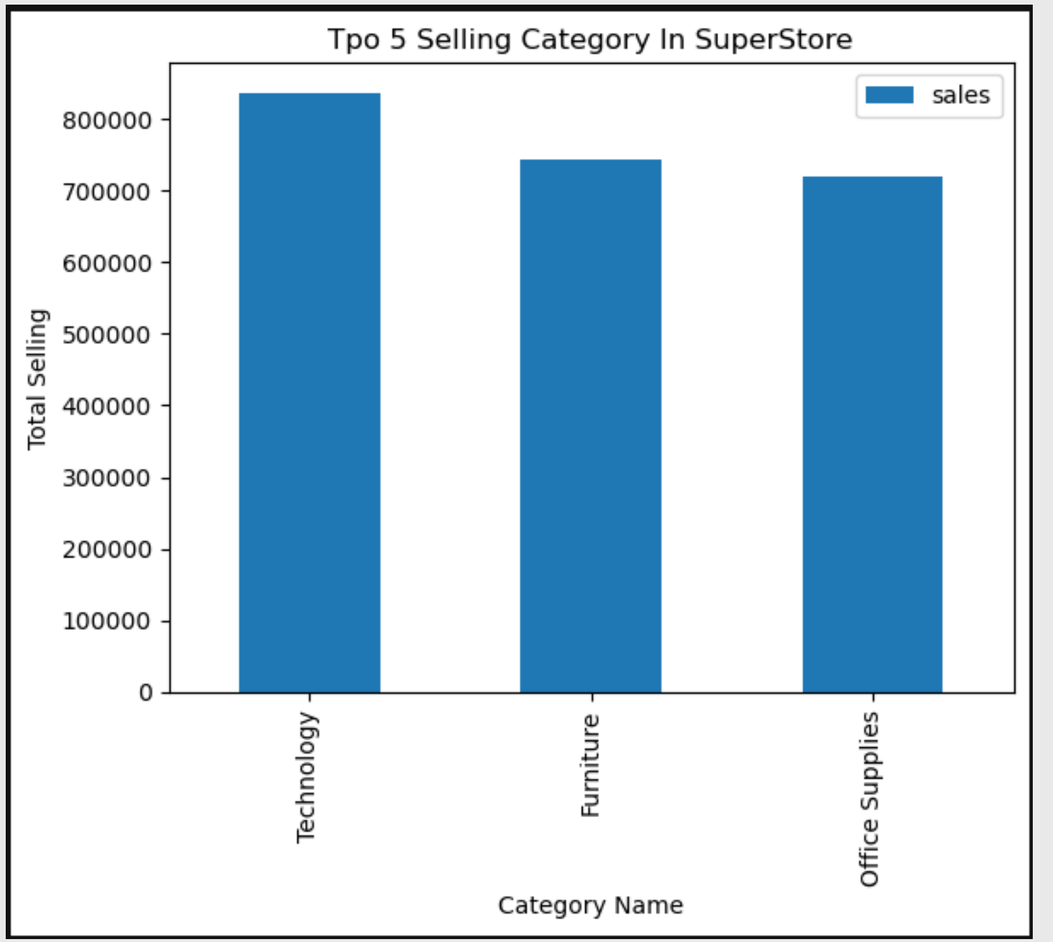
- Top-Profitable Products:
- Canon Camera
- ElectricPlastic Comb
- PackardLaser Jet
- Canon Personal Laser Copier
- HPDesignJet 24-Color Printer

- Top-Profitable Categories :
- Technology
- Office Supplies
- Furniture


Tpo 5 Profit Category In SuperStore
profit
Total Profit
0
20000
40000
60000
80000
100000
120000
140000
Technology
Office Supplies
Furniture
Category Name

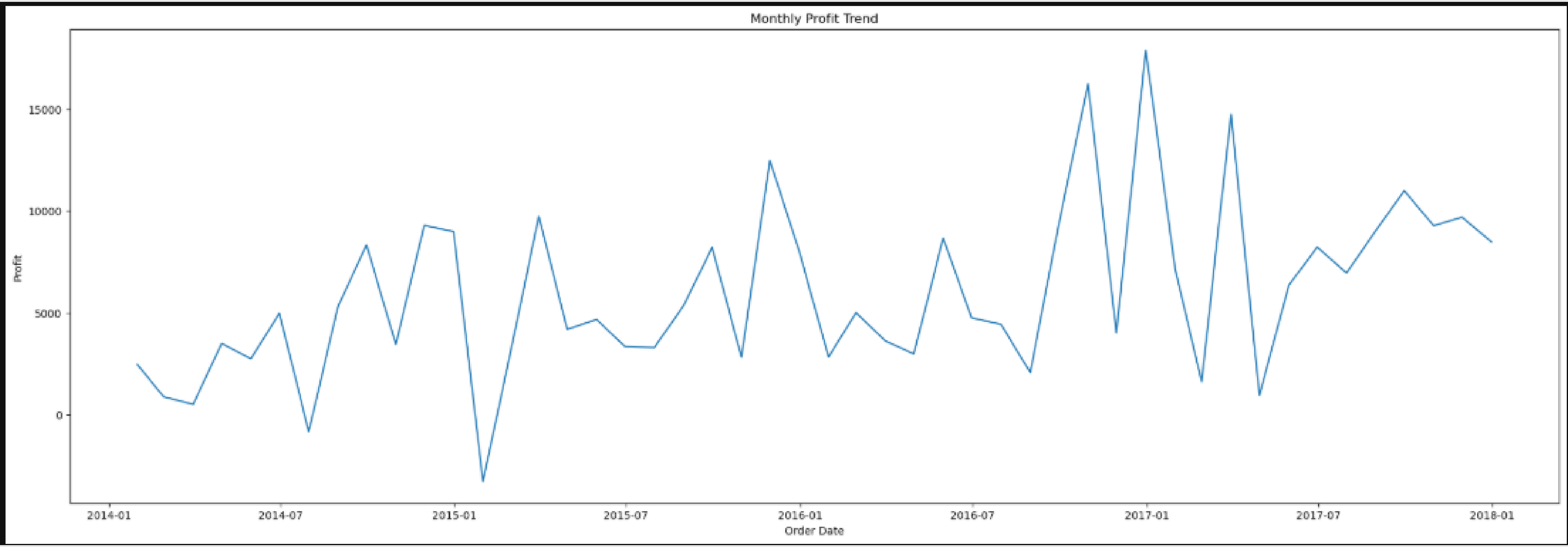

Tpo 5 Selling Category In SuperStore
sales
Total Selling
0
100000
200000
300000
400000
500000
600000
700000
800000
Technology
Furniture
Office Supplies
Category Name

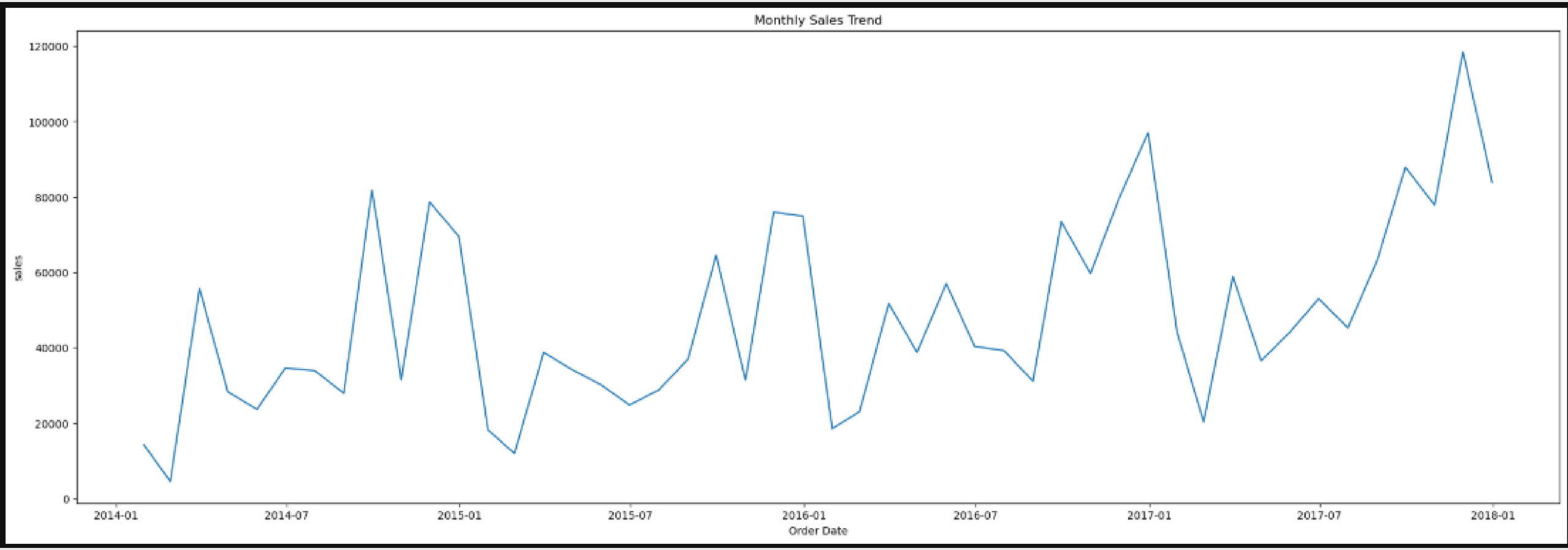

Tpo 5 Selling Sub_Category In SuperStore
sales
Total Selling
300000
250000
200000
150000
100000
50000
0
Phones
Chairs
Storage
Tables
Binders
Machines
Accessories
Copiers
Bookcases
Appliances
Furnishings
Paper
Supplies
Art
Envelopes
Labels
Fasteners
Sub_Category Name

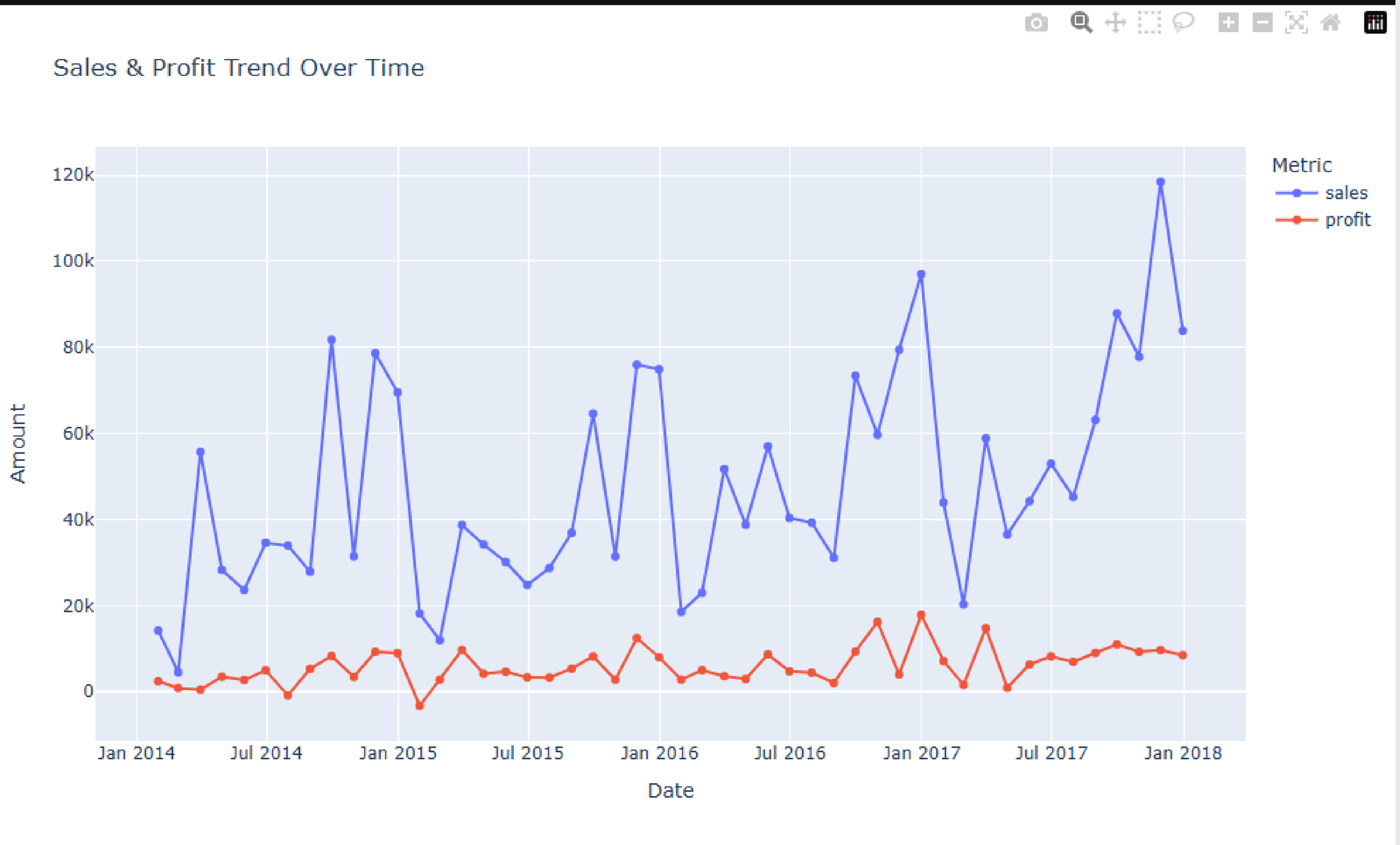

Tpo 5 Selling Category In SuperStore
sales
Total Selling
800000
700000
600000
500000
400000
300000
200000
100000
0
Technology
Furniture
Office Supplies
Category Name

RESULTS OF EDA CONT

- Some of the Top categories and subcategories are performing better at sales than at profit wise indicating a very clear problem with pricing and discounts.
- The same pattern is repeated for the products This issue causes a decrease in profit margin increase in unprofitability.
- The top selling and the top profitable are not the same which proves the existence of this problem furthermore and hence problems with pricing and discounts.
- For the sales trends weekly monthly quarterly and yearly across the data set shows a seasonal linear pattern of increasing and decreasing in seasonality and an overall general increase over time.

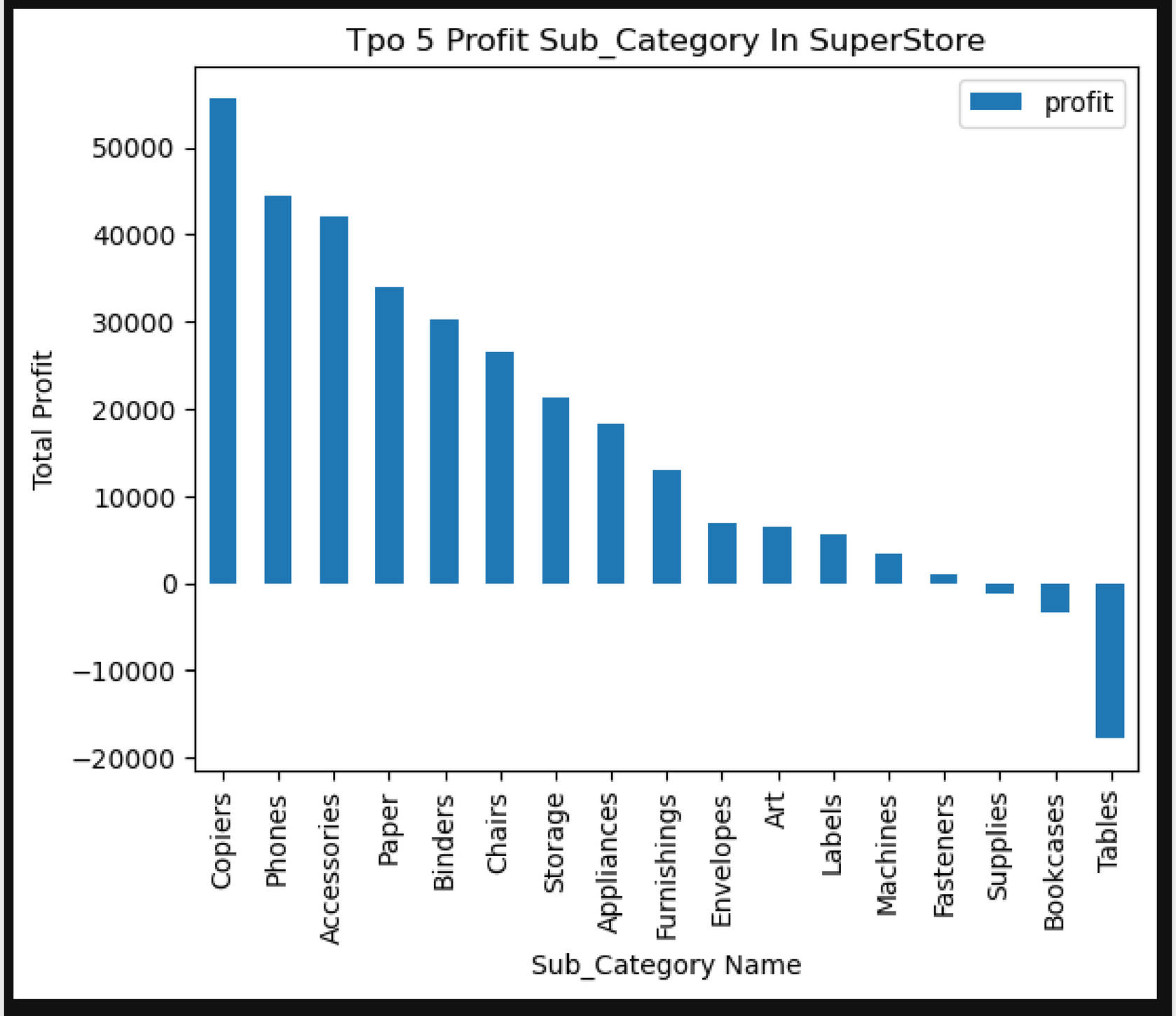

Monthly Profit Trend
15000
10000
5000
0
Profit
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Order Date

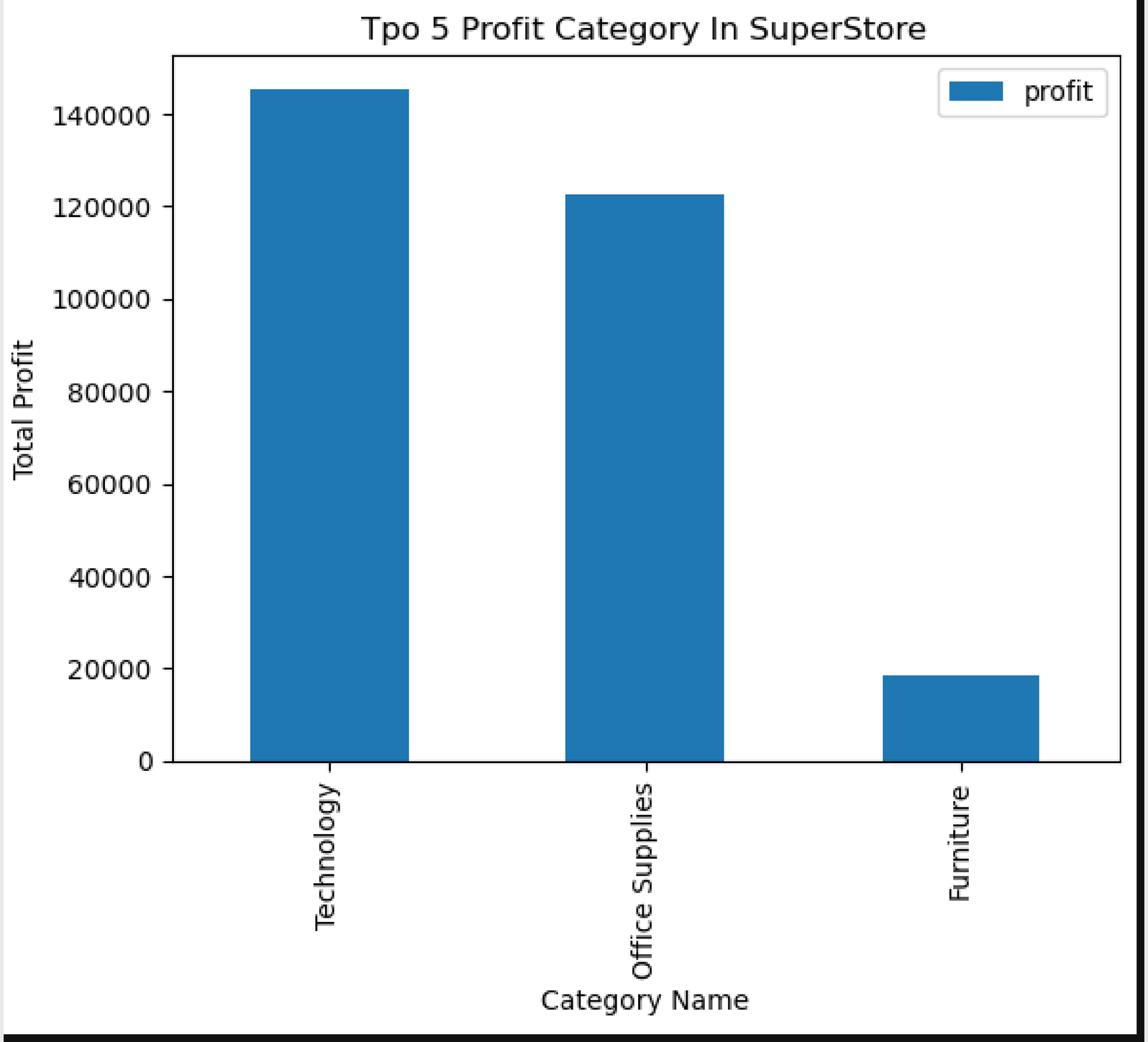

Monthly Sales Trend
120000
100000
80000
60000
40000
20000
0
sales
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Order Date

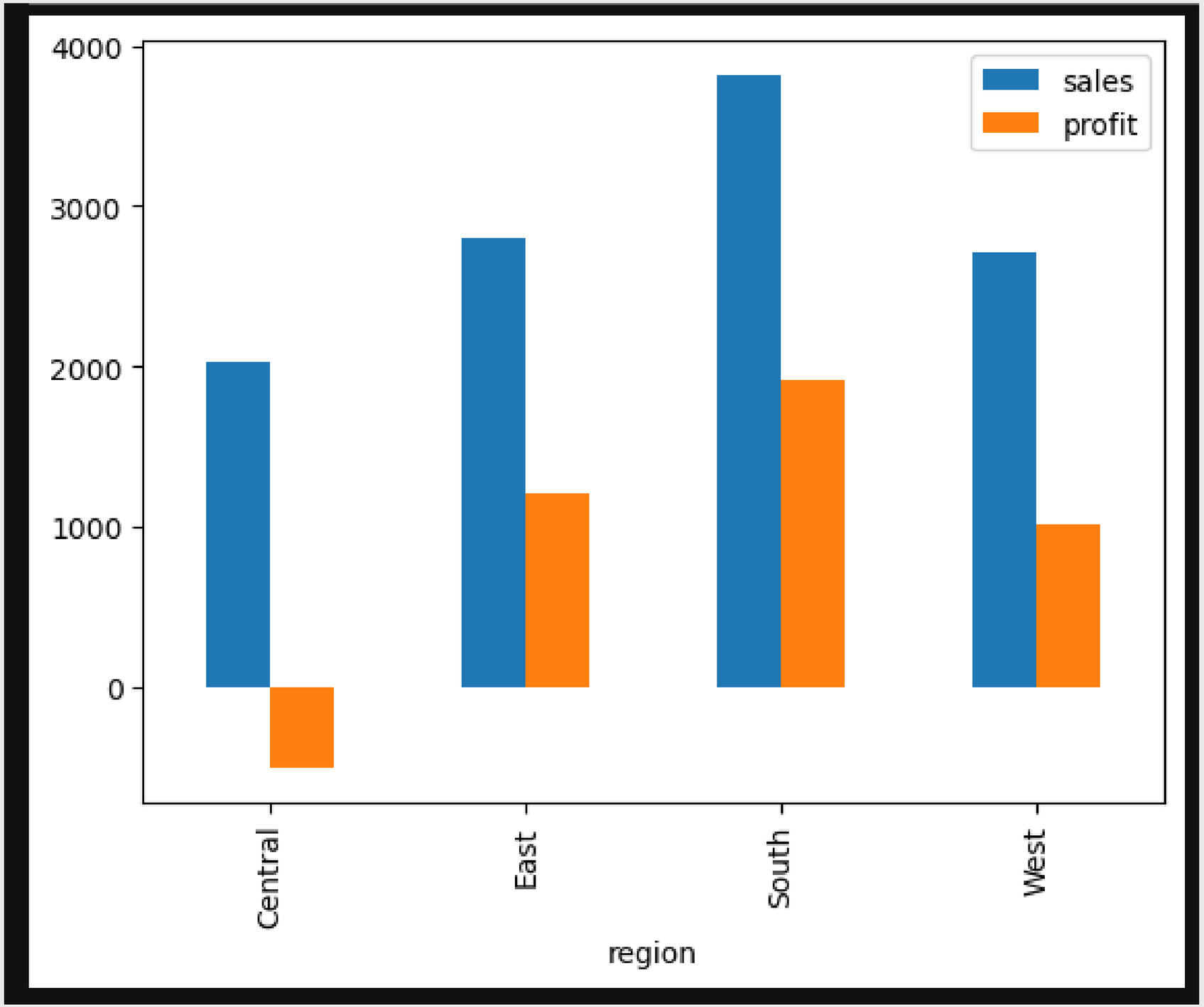

Sales & Profit Trend Over Time
Metric
sales
profit
120k
100k
80k
60k
40k
20k
0
Amount
Jan 2014
Jul 2014
Jan 2015
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018
Date

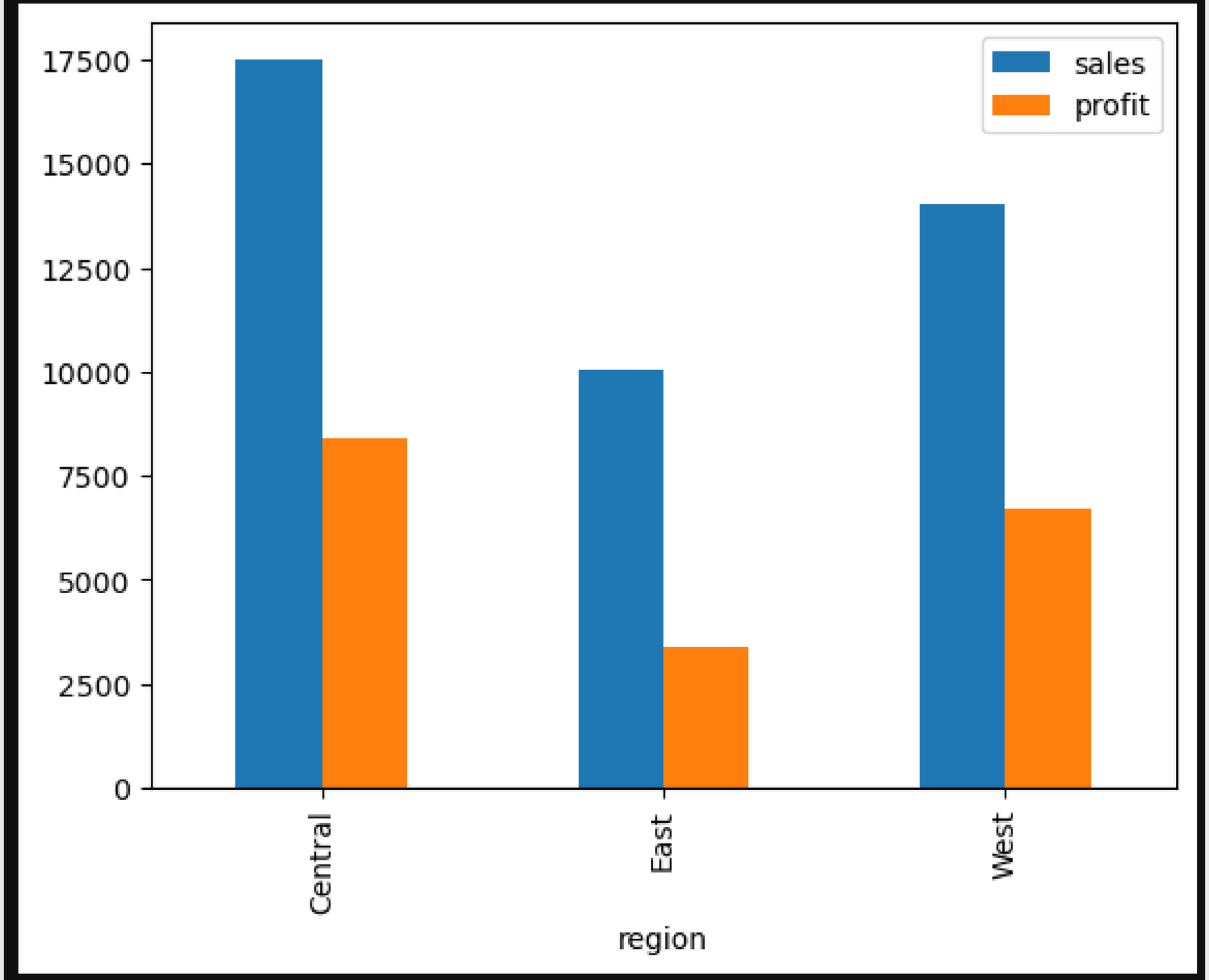
RESULTS OF EDA CONT

- The Unprofitability issue was also visible here but we were also able to conclude that The West and East regions are the top selling and the top profitable in comparison to the other two region which recommends that we focus on them more to generate more revenue and profit. The same pattern continues over to the cities in each region.
- When analyzing the discount percentage in accordance to sales and profit we find that higher discounts do generate a lot more sales but it's the complete opposite for profit as it decreases or it can come as a loss
- In the same way we find the highest sales as well as profitability present when the discount was from zero all the way to 20%
- The company should focus more on the technology category as it provides more sales as well as profitability proven by the fact that it has the highest profitability and sales out of all the category in addition to having the two highest selling and profitable products


Tpo 5 Profit Sub_Category In SuperStore
profit
50000
40000
30000
20000
10000
0
−10000
−20000
Total Profit
Copiers
Phones
Accessories
Paper
Binders
Chairs
Storage
Appliances
Furnishings
Envelopes
Art
Labels
Machines
Fasteners
Supplies
Bookcases
Tables
Sub_Category Name

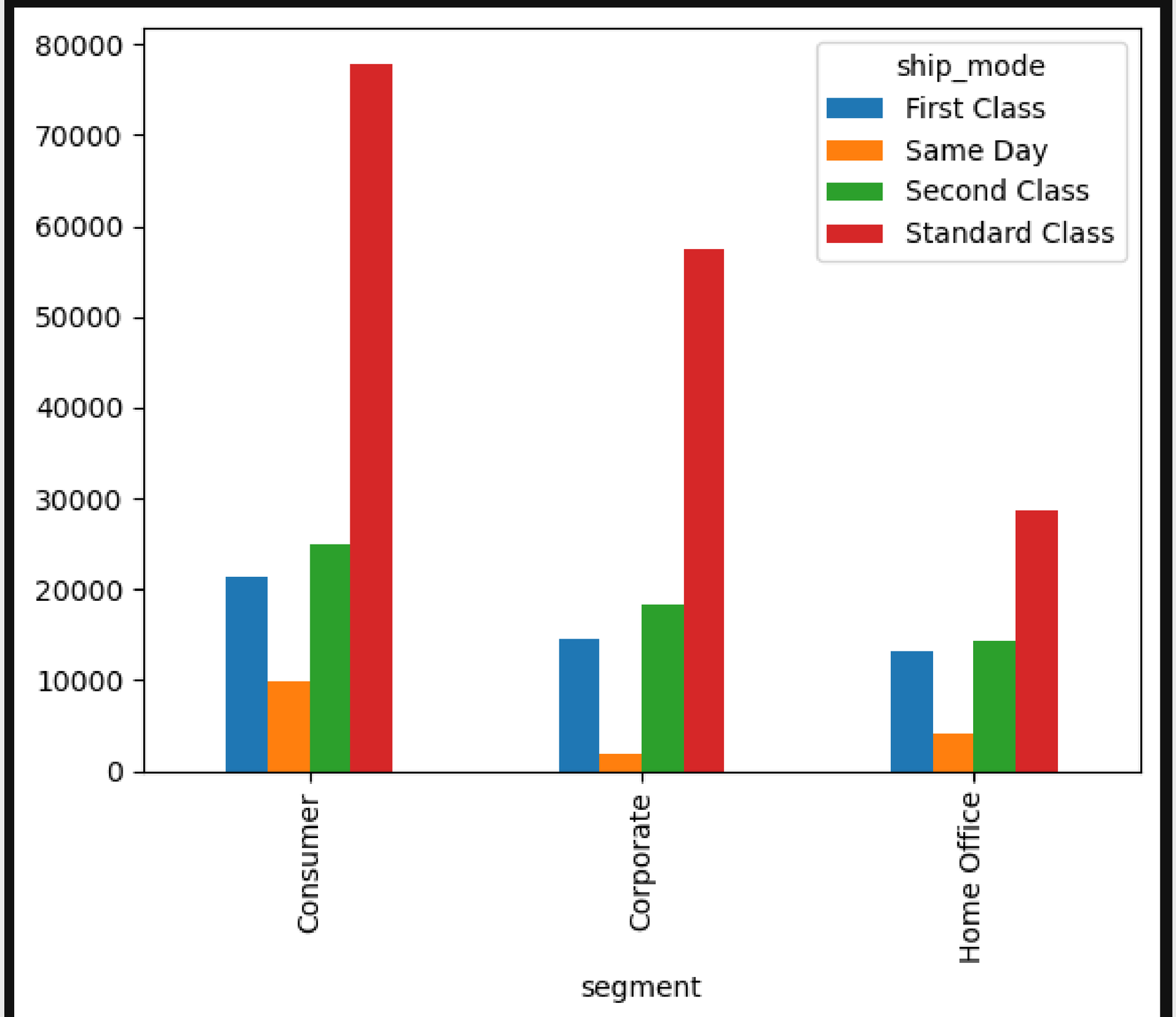

Tpo 5 Profit Category In SuperStore
profit
140000
120000
100000
80000
60000
40000
20000
0
Total Profit
Technology
Office Supplies
Furniture
Category Name


4000
3000
2000
1000
0
sales
profit
Central
East
South
West
region


sales
profit
17500
15000
12500
10000
7500
5000
2500
0
Central
East
West
region

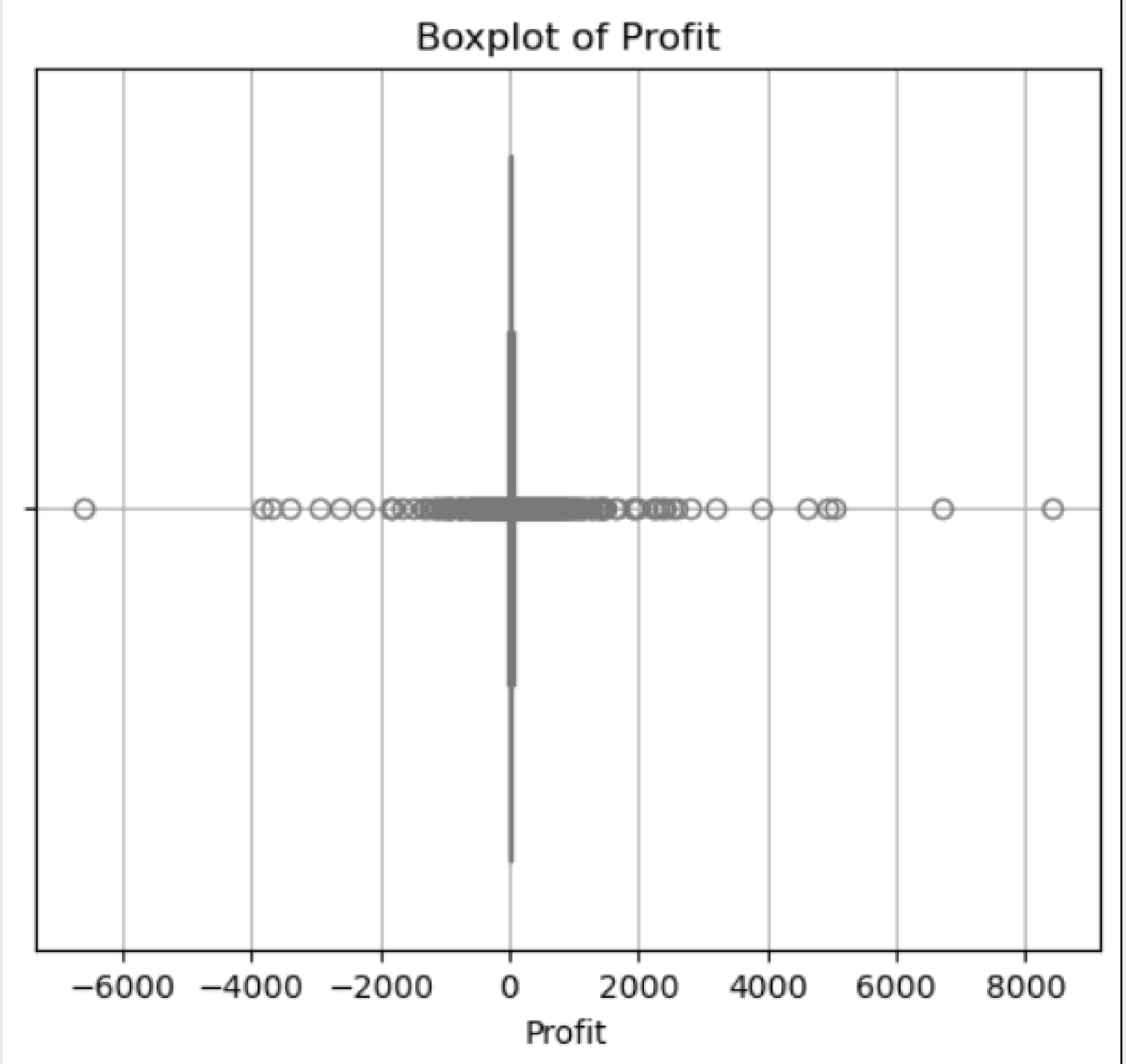
RESULTS OF EDA CONT

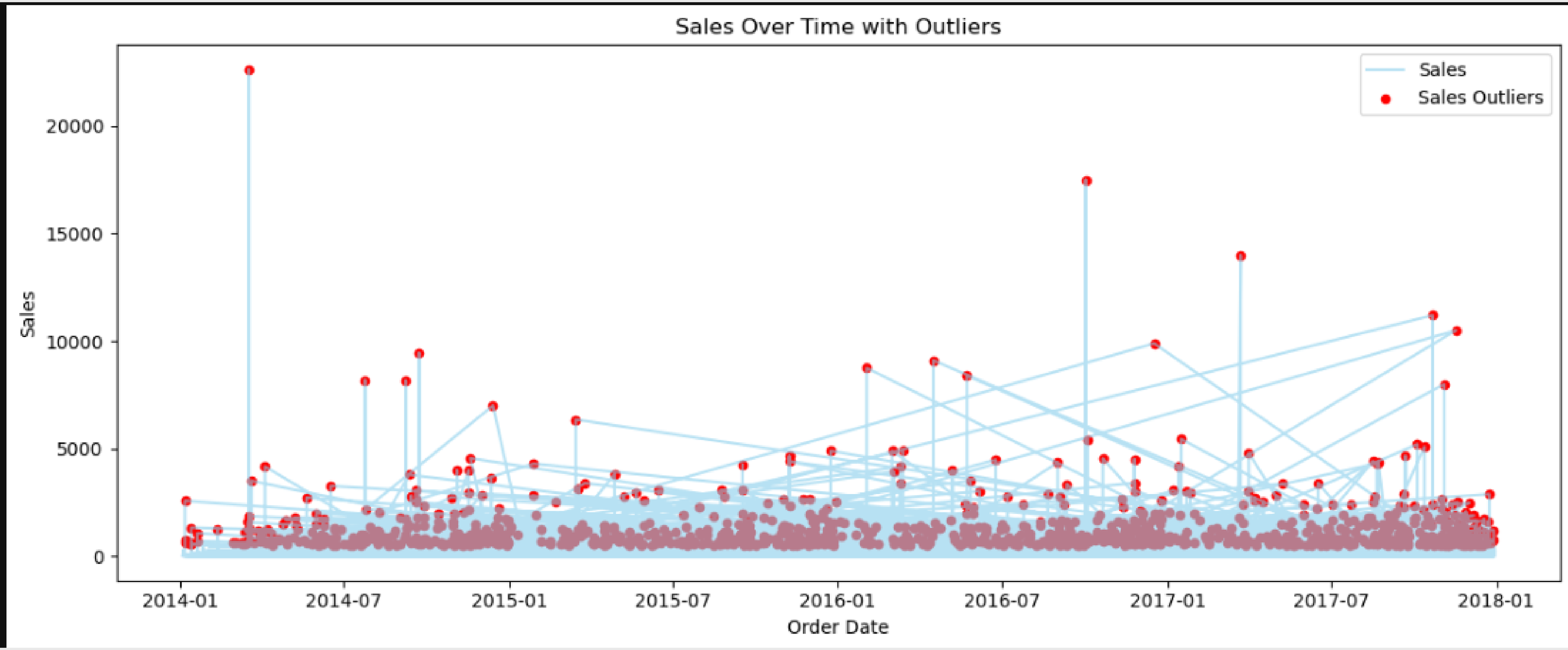
- In the customer side We find that consumers corporate then home office are the rankings for the most orders.
- In addition to that the consumer as well as the standard class shipping mode are the ones that are driving the most amount of sales followed by corporate and second class shipment then finally home offics.

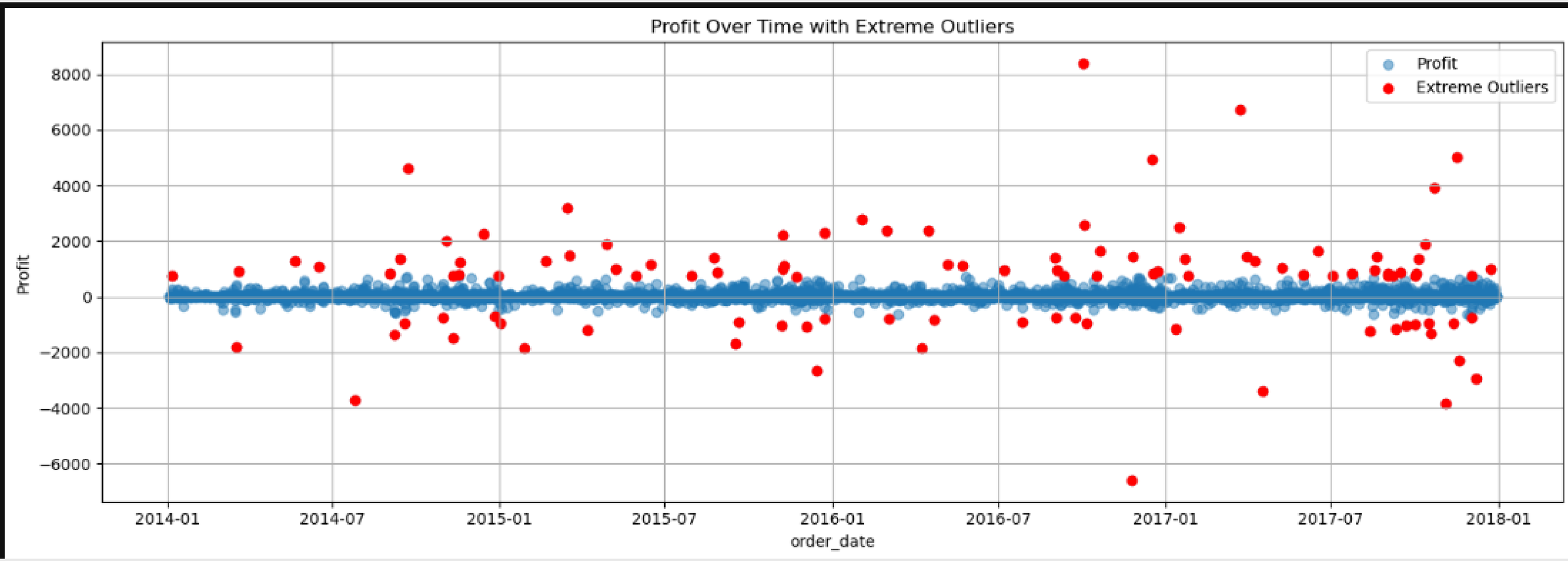

80000
70000
60000
50000
40000
30000
20000
10000
0
ship_mode
First Class
Same Day
Second Class
Standard Class
Consumer
Corporate
Home Office
segment

ANALYSIS AND OUTLIERS DETECTION.

- Through the analysis process we were capable of visually Proving all of the previous findings that we got through the exploratory data analysis.
- We also provided summarizing visualizations spanning over the sales and the profit when it comes to the products the categories and the subcategories in addition to the customer, region/state/city level findings
- and lastly Through plotting and visualizations we were capable of identifying all of the outliers in the data set and marking them as normal outliers as the data itself is normally messy, It's also worth pointing that there Was extreme outliers that we were also able to detect and visualize so that we have a better control over the forecasting process as they could the accuracy of the model.


Boxplot of Profit
−6000
−4000
−2000
0
2000
4000
6000
8000
Profit


Sales Over Time with Outliers
Sales
Sales Outliers
20000
15000
10000
5000
0
Sales
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Order Date


Profit Over Time with Extreme Outliers
Profit
Extreme Outliers
8000
6000
4000
2000
0
−2000
−4000
−6000
Profit
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
order_date

MODEL SELECTION.

- this Is a very crucial part of being able to correctly forecast and predict future trends and ultimately leading to stronger business decisions that are derived by accurate forecasting.
- the following are the models of choice:
- XGBOOST
- ARIME/SARIME
- PROPHET
- for the following reasons: .....

MODEL TYPE

- Based on the datasets as well as our goals to be able to accurately predict the sales into the future in combination with the data set containing historical data that spanned from the beginning of 2014 all the way till the end of 2017 we conclude that the This is a time series forecasting problem.
- Now there are different types of models out there that could be used but our choice is not 100 percent accurate as it depends on the characteristics and the quantity of the available historical data.

WHY PROPHET?!

- Prophet is an open-source time series forecasting model developed by Facebook, designed to handle business time series data with strong seasonal effects and historical trends. It is an additive model that decomposes time series into trend, seasonality, and holiday effects, making it especially useful for data with repeating patterns over time. One of Prophet's standout features is its simplicity and ease of use – it requires minimal data preprocessing and is highly beginner-friendly, allowing users to generate accurate forecasts with just a few lines of code. What distinguishes Prophet from other models is its built-in ability to automatically detect and model yearly, weekly, and even daily seasonality, along with handling missing data and outliers gracefully. It also provides intuitive tools to incorporate holidays, special events, or custom seasonality directly into the forecast. In our project, Prophet was an ideal fit because of its strong performance with time-based data, its flexibility in handling real-world business scenarios, and its interpretability, which makes it easy to explain results to non-technical stakeholders.

PROPHET TIME >>

- To start working with profit you need to first of all understand that we need to prepare the data so that it only includes two columns and those two columns need to be renamed as ‘ds’ and ‘y’
- The two columns are going to represent the dates as ‘ds’ and The value that we desire to predict which is the sales as ‘y’
- In our case we needed to do some changes first because the dates that we have Were recorded per order meaning that we could have multiple dates if in a single day we received multiple orders and this is not acceptable because we need to make sure that the type of date data that we feed into the model is Fit in a way to show direct and continuous dates across the data set, And on the other side should be the sum of the sales for each date/day.
- This specific model is special in the way that it can identify dates and time, Especially across varying data that has outliers and strong seasonal patterns.

- Our work will be separated into multiple sequential steps: .....

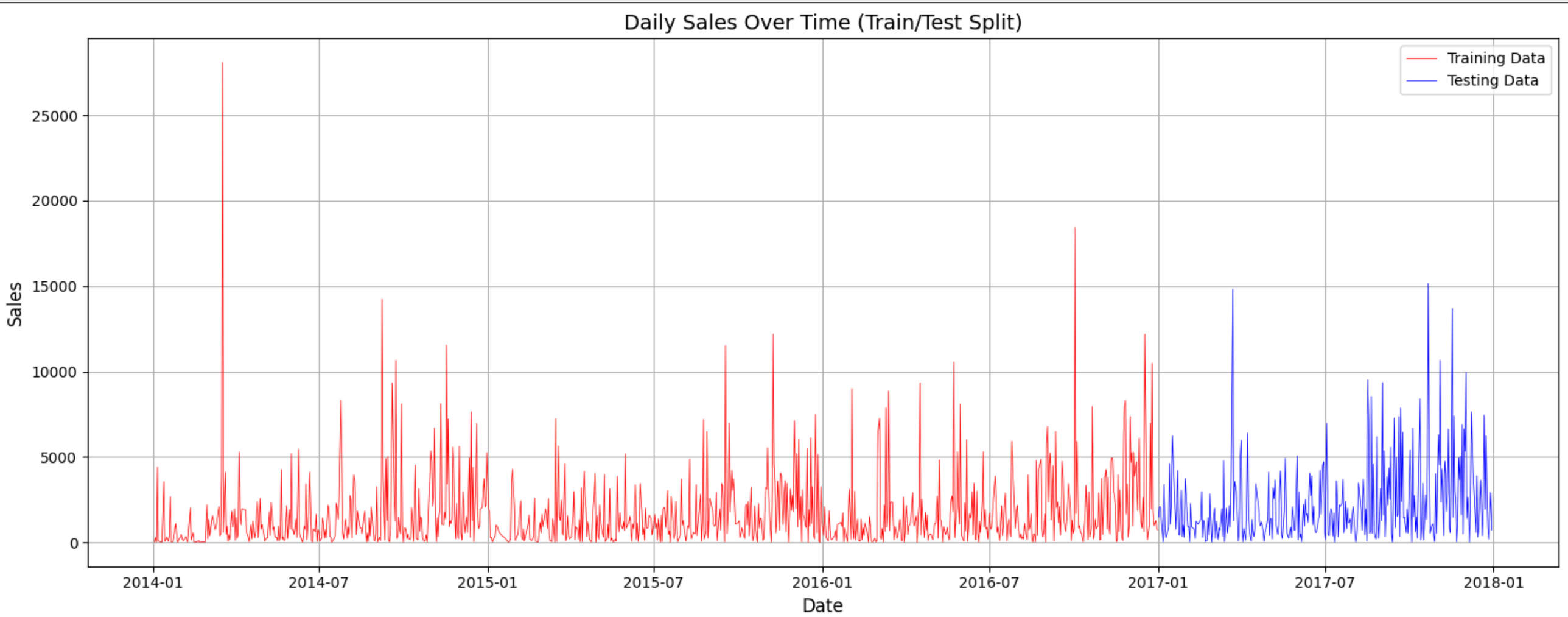
STEPS?!

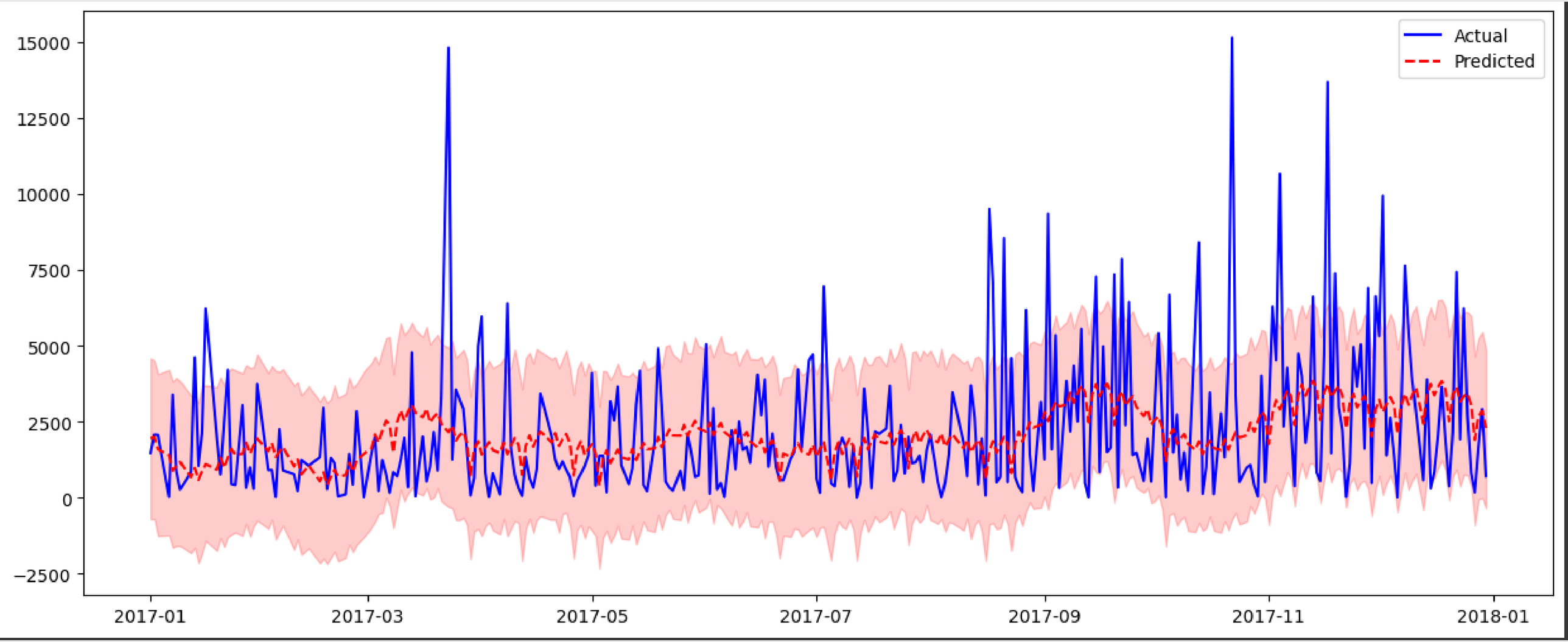
- The steps will be as follows:
- First we need to develop a baseline model without any extra regressors or any extra parameters nor arguments to serve as a baseline that we could compare to all of the other models that we will create.
- In order for that to happen all we have to do is to train test split the data, And And then start fetting the model, After that we proceed with forecasting on the test dates and visualize the results to see how accurate the baseline model is when it comes to its forecasts in comparison to the actual testing data.
- The then is extract the components of the model which are basically all of the patterns that the model is capable of picking on through time and they vary for the years the months or even the days of the year.
- And lastly we use the evaluation metrics to measure how good our models prediction are.

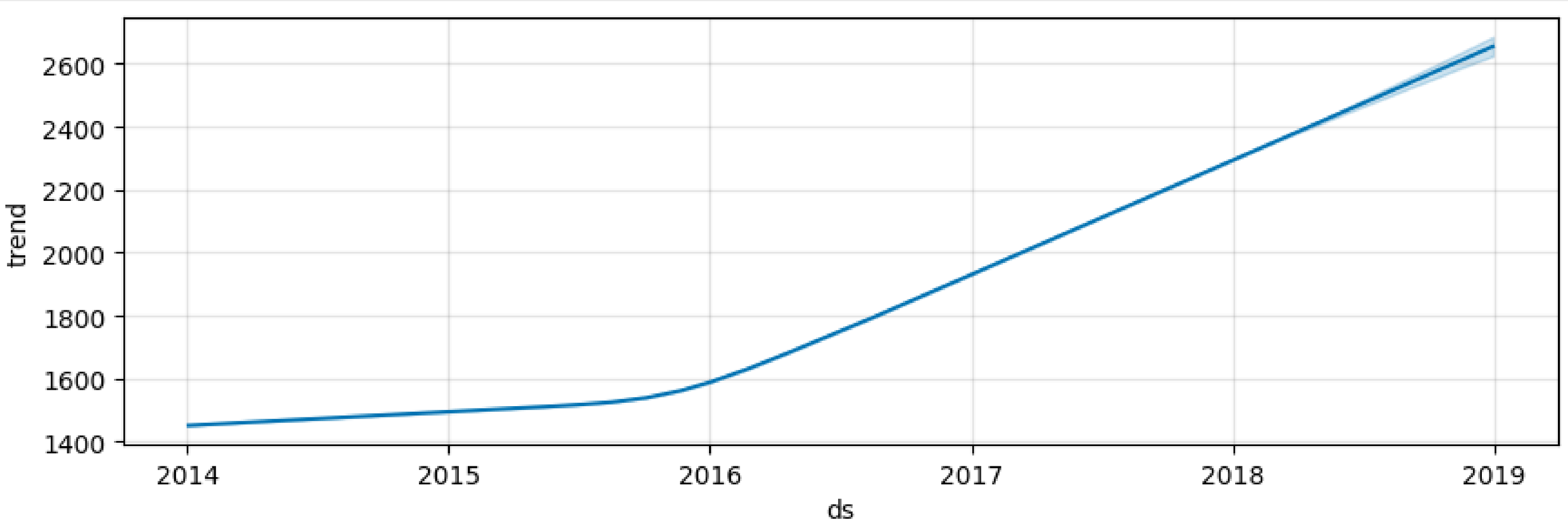
EVALUATION METRICS?!

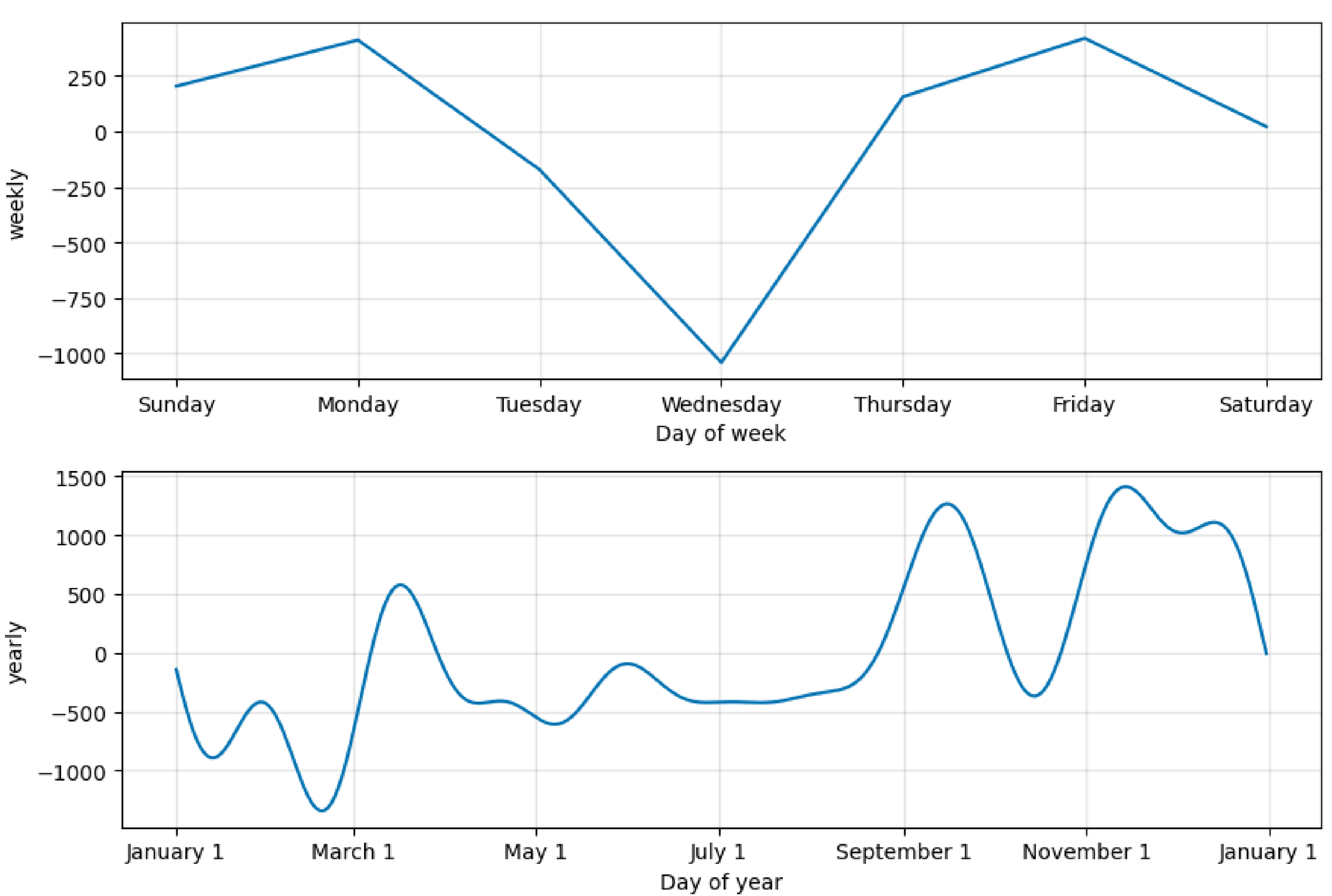
- The steps will be as follows:
- We picked four evaluation metrics for our model:
- MAE
- RMSE
- MAPE
- SMAPE

- NEXT :
- We visualize the components of the model which are basically the patterns and the trends that the model has picked up on and they could vary from yearly to monthly or even days of the year.
- the visualizations of our work.


Daily Sales Over Time (Train/Test Split)
Training Data
Testing Data
Sales
25000
20000
15000
10000
5000
0
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Date

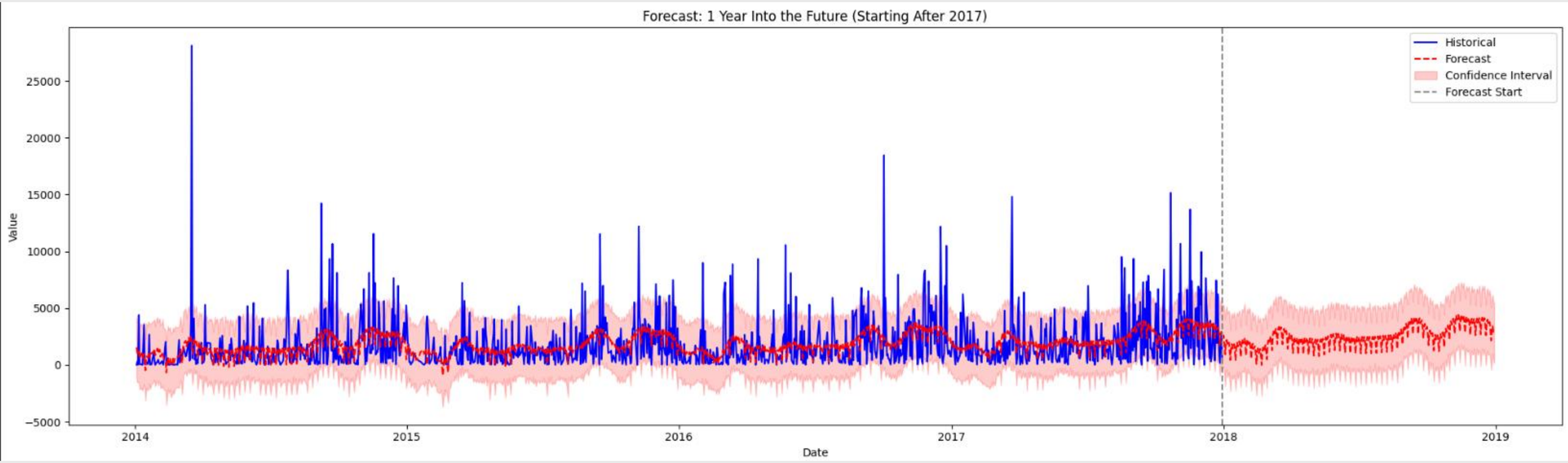

Actual
Predicted
15000
12500
10000
7500
5000
2500
0
−2500
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01


2600
2400
2200
2000
1800
1600
1400
trend
2014
2015
2016
2017
2018
2019
ds


weekly
250
0
−250
−500
−750
−1000
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week
yearly
1500
1000
500
0
−500
−1000
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year

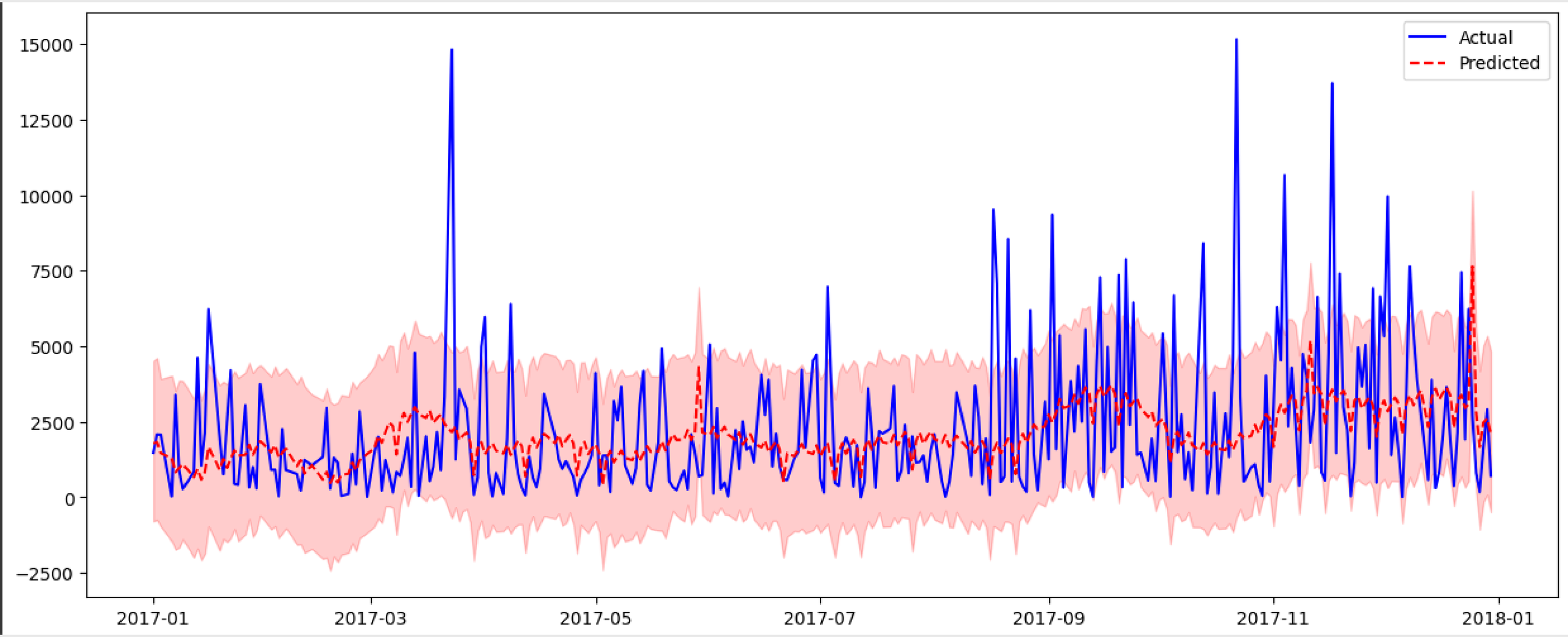
STEPS?!

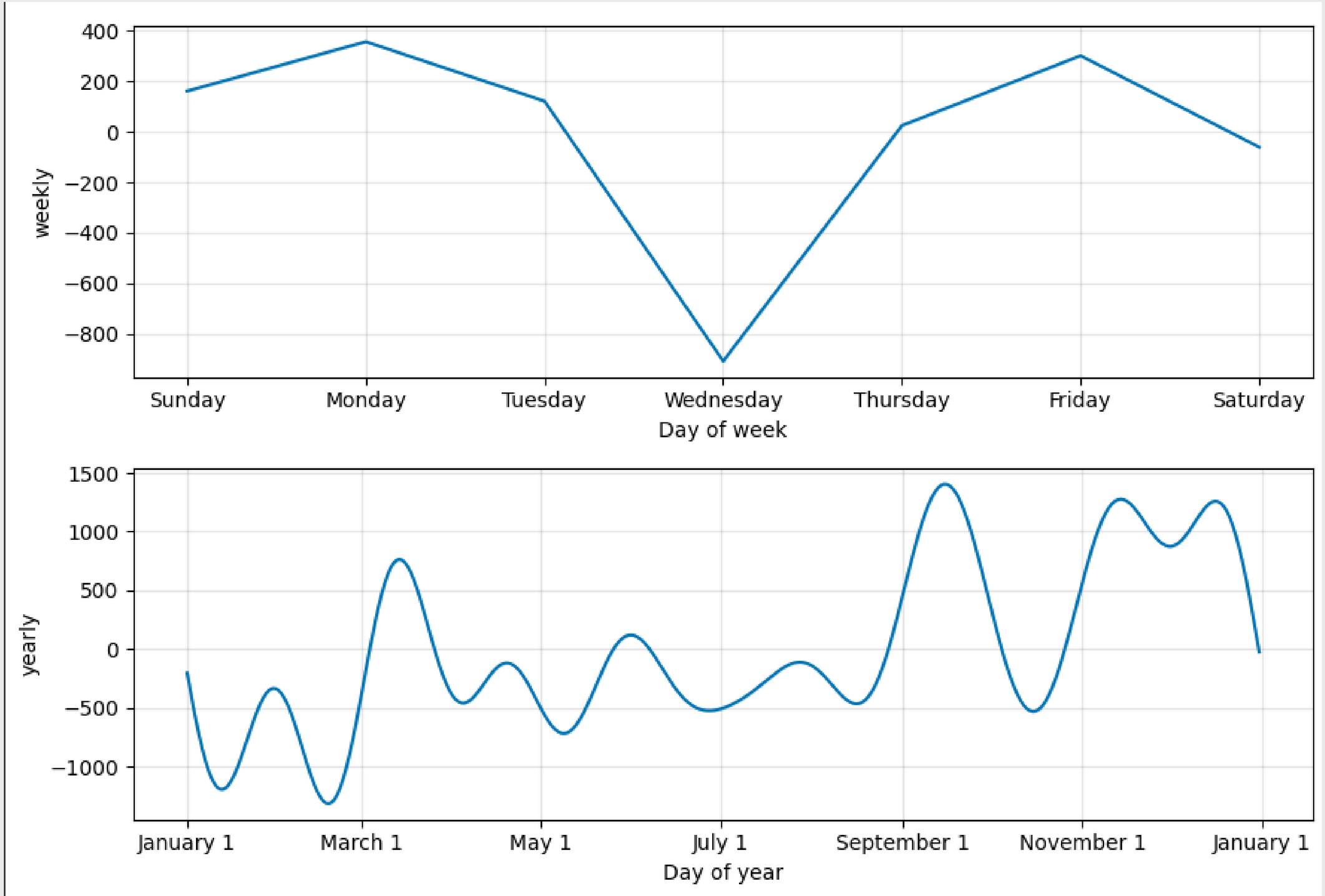
- The steps will be as follows:
- next we forecast into the future which is one of the strengths of the prophet Model, We do that by training the model on the entire data set from the beginning of 2014 all the way till the end of the 2017 basically we're going to include both splits of the data the train as well as the tests then we will specify the next year as our target for forecasting and then visualize the results.

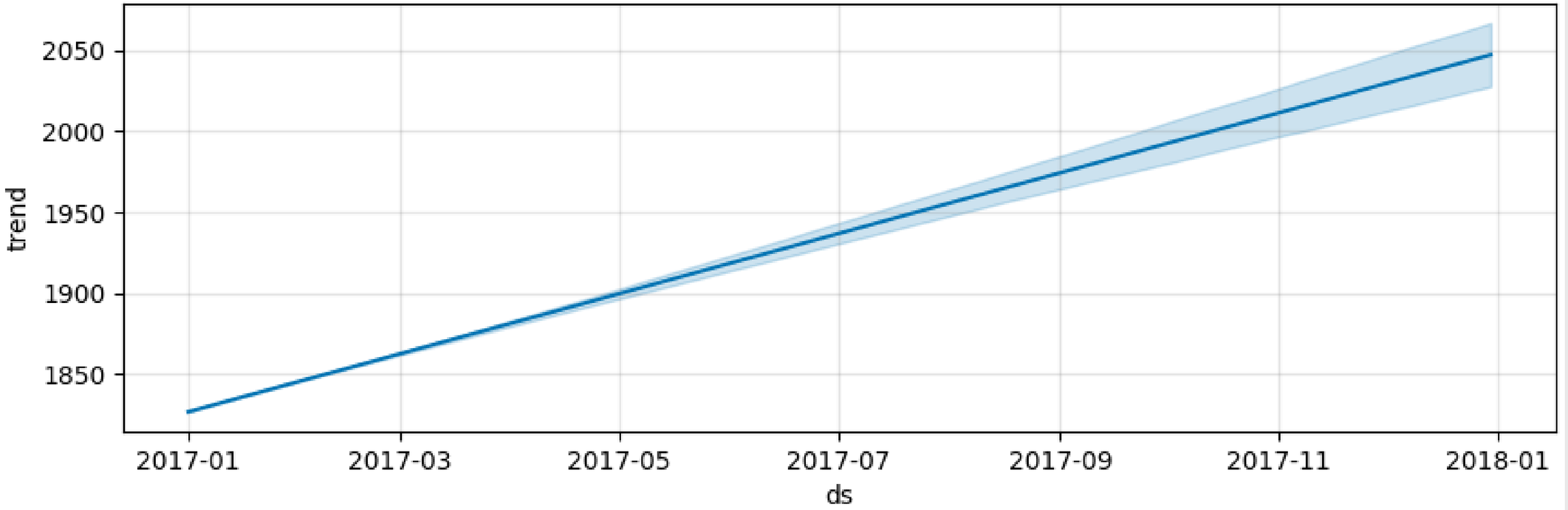

Forecast: 1 Year Into the Future (Starting After 2017)
Historical
Forecast
Confidence Interval
Forecast Start
Value
25000
20000
15000
10000
5000
0
−5000
2014
2015
2016
2017
2018
2019
Date

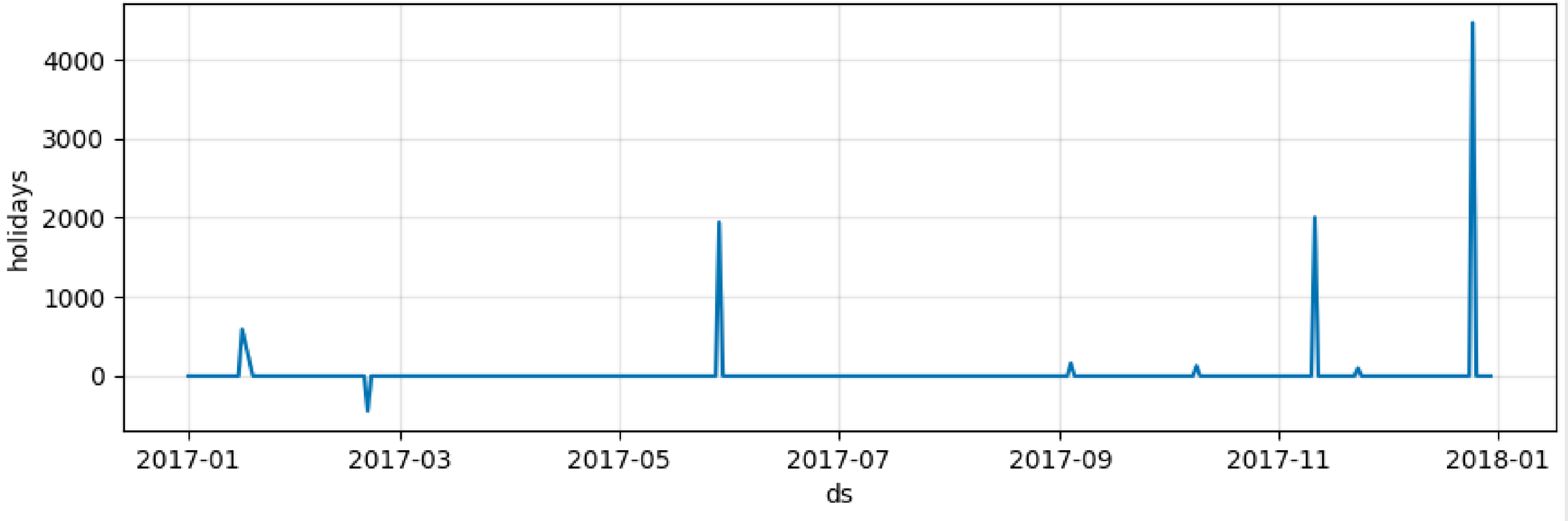
STEPS?!

- The steps will be as follows:
- The next step is to add the regressor of the holidays because as we all know sales of stores are heavily influenced by many external factors aka Regressors such as holidays discount Announcements and so on.
- So we're going to refit the model once again with the The training section of the data only then we're going to predict into the future using the test set visualizing the results but this time it's going to be unique because we're going to add the regressor holiday and see how the model is going to pick up on the trends and how accurate is it's forecasting going to become.


Actual
Predicted
15000
12500
10000
7500
5000
2500
0
−2500
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01


400
200
0
−200
−400
−600
−800
weekly
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week
1500
1000
500
0
−500
−1000
yearly
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year


2050
2000
1950
1900
1850
trend
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
ds


4000
3000
2000
1000
0
holidays
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
ds

MORE IMPROVMENTS?!

- The steps will be as follows:
- If we wanted to get more accurate results we could do that simply by working on a different level of data what do I mean by that?
- It basically means that instead of working at the level of the daily sales we could work at the level of the daily sales for each and every single subcategory in the superstore in order for us to be able to isolate any outsider trends from different subcategories that may overlap together in addition to getting a more accurate predictions into the future.
- however this was halted by the time and Capability constraint as training 19 models for every single subcategory is impossible If we abide by the time constraint.

WHAT IS XGBOOST?

XGBOOST (EXTREME GRADIENT BOOSTING) IS A POWERFUL MACHINE LEARNING ALGORITHM BASED ON DECISION TREES.

IT'S KNOWN FOR ITS SPEED, ACCURACY, AND PERFORMANCE, ESPECIALLY IN STRUCTURED/TABULAR DATA LIKE SPREADSHEETS

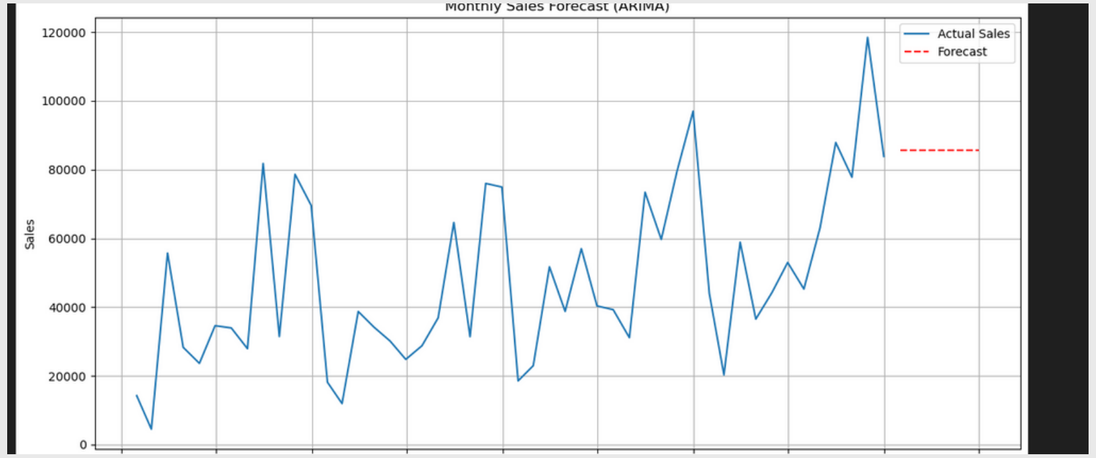
- **IT USES HIGHLY EFFICIENT INTERNAL CODE (OPTIMIZED CODE)**
- **The code that runs in the backgrOUND (WHICH YOU DON'T DIRECTLY SEE) IS WRITTEN IN FAST PROGRAMMING LANGUAGES LIKE C++ INSTEAD OF PYTHON.**
- **THIS MAKES IT MUCH FASTER THAN OTHER SIMILAR ALGORITHMS WRITTEN IN STANDARD WAYS.**
- **IT TAKES ADVANTAGE OF TECHNIQUES SUCH AS:**
- **PARALLEL PROCESSING – RUNNING MULTIPLE OPERATIONS AT THE SAME TIME.**
- **MEMORY OPTIMIZATION – USING LESS RAM.**
- **CACHE AWARENESS – USING THE DEVICE'S MEMORY SMARTLY TO BOOST PERFORMAN**

🧩 MY CONTRIBUTION: PREDICTIVE MODELING USING XGBOOST

- 📌 1. Profit Prediction using XGBoost Regressor
- In this part, I built a regression model to predict profit using the XGBoost algorithm. The model used several features such as:
- Quantity
- Discount
- Shipping method (One-hot encoded)
- Region (One-hot encoded)
- Product category and sub-category
- Shipping time
- Day of week and month of the order
- ☑ Implementation Steps:
- Prepared the data by selecting relevant features (X) and setting the target variable as profit (y).
- Split the data into training and testing sets using train_test_split.
- Trained the model using XGBRegressor.
- Evaluated the model using metrics such as Mean Squared Error (MSE) and R^2 score.
- 🎯 Results:
- MSE = 6413.00
- R^2 Score = 0.87
- This means the model explains 87% of the variance in profit, which is considered a strong predictive performance.

📌 2. SALES FORECASTING USING ARIMA (TIME SERIES ANALYSIS)

- I also applied Time Series Forecasting using the ARIMA model to predict future sales based on historical order dates.
- ☑ Implementation Steps:
- Aggregated the sales data by Order Date.
- Performed stationarity checks using the Augmented Dickey-Fuller (ADF) test.
- Chose ARIMA(0, 1, 1) as the best-fit model based on statistical analysis.
- Trained the model and forecasted sales for the next 30 days.
- Plotted the actual vs. predicted values to visualize future trends.



DEPLOYMENT-DASHBOARD

THANK YOU