Machine Learning: Time Series Retail Sales Forecasting

MOTIVATION AND INTRODUCTION

Time series models are widely used to extract meaning patterns from economic, retail sales, climate, and stock price data, which can catch the seasonal pattern when making predictions for future values. In this project, we aim to employ **time series models (ARIMA, Prophet)** to compare and predict the retail sales of technology and office supplies in the superstore. Another emphasize of this project is to precisely **visualize the seasonal pattern** for prediction.

ACHIEVEMENT FLOW

Technology sales data processing and indexing

↓

Visualizing technology sales time series data

↓

Time series forecasting with ARIMA,model fitting, and accuracy validating

↓

Producing and Visualizing forecasts

↓

Comparing technology sales vs. office supply sales

↓

Forecasting with Prophet for the comparison above

↓

Trends and patterns visualization

TECHNICAL SPECIFICATION

- Python 3
- Jupyter Notebook

DATA PROCESSING

- We used the superstore dataset and a brief overview is provided blow, which includes 9994 observations. A brief overview of
- The dataset contains 19 variables, namely OrderID, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City, State, Postal Code, Region, Product ID, Category, Sub Category, Product Name, Sales, Quantity, Discount, and Profit.
- There are **no missing values**. The order dates are from 2014-01-06 to 2017-12-30. We used the average daily sales value for a month so we used the **first day of each month as the timestamp**.

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	...	Postal Code	Region	Product ID	Category	Sub-Category	Product N
8	CA-2014-115812	2014-06-09	2014-06-14	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	...	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 53 Phone VoIP p
12	CA-2014-115812	2014-06-09	2014-06-14	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	...	90032	West	TEC-PH-10002033	Technology	Phones	Konfte Conference ph Charcoal b
20	CA-2014-143336	2014-08-27	2014-09-01	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	...	94109	West	TEC-PH-10001949	Technology	Phones	Cisco SPA 501 P
27	CA-2016-121755	2016-01-16	2016-01-20	Second Class	EH-13945	Eric Hoffmann	Consumer	United States	Los Angeles	...	90049	West	TEC-AC-10003027	Technology	Accessories	Imation 8GB TravelDrive 2.0 Flash
36	CA-2016-117590	2016-12-08	2016-12-10	First Class	GH-14485	Gene Hale	Corporate	United States	Richardson	...	75080	Central	TEC-PH-10004977	Technology	Phones	GE 3052

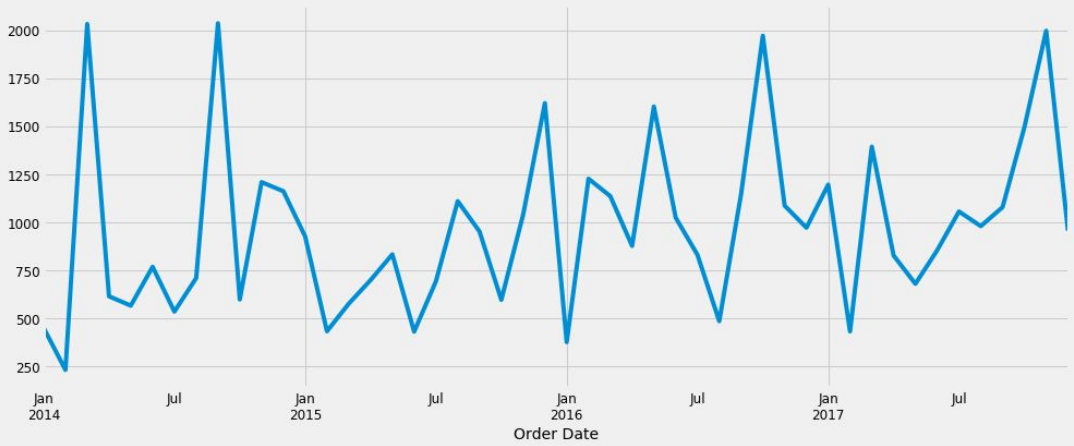
Order Date	Sales
2014-01-06	1147.94
2014-01-09	31.20
2014-01-13	646.74
2014-01-15	149.95
2014-01-16	124.20


```
Order Date
2014-01-01     449.041429
2014-02-01     229.787143
2014-03-01    2031.948375
2014-04-01     613.028933
2014-05-01     564.698588
Freq: MS, Name: Sales, dtype: float64
```

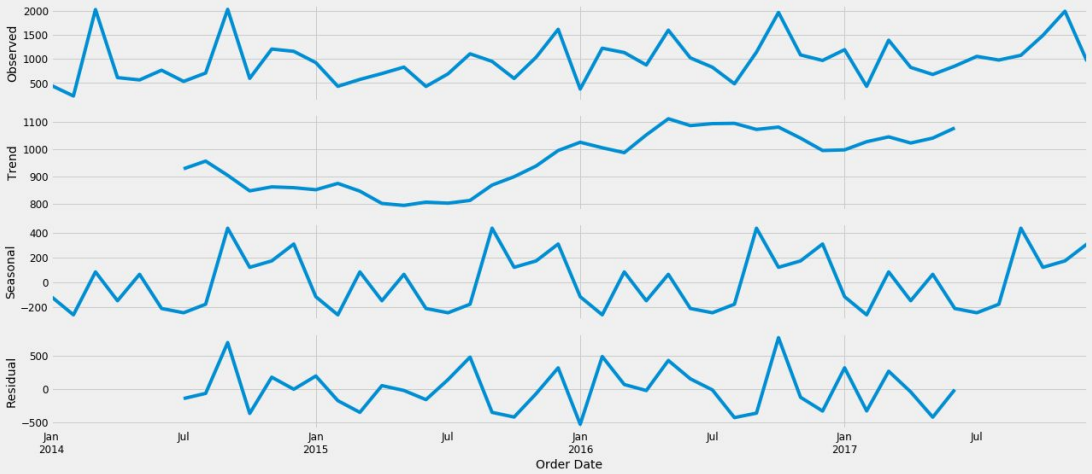
(Technology Sales, indexed with 'Order Data', sorted by 'Order Data')

VISUALIZING TECHNOLOGY SALES

- This time series have **obvious seasonality pattern**. It can be seen in the graph that sales are high at the end of the year and low at the beginning of the year. There are always a couple of low months in the mid of the year.



- We used **time-series decomposition** to decompose our time series into three distinct components: **trend**, **seasonality**, and **noise**, as well as showing the **residual** patterns. We can see that the sales of technology products is not stable with **clear seasonality**.



TIME SERIES FORECASTING WITH ARIMA

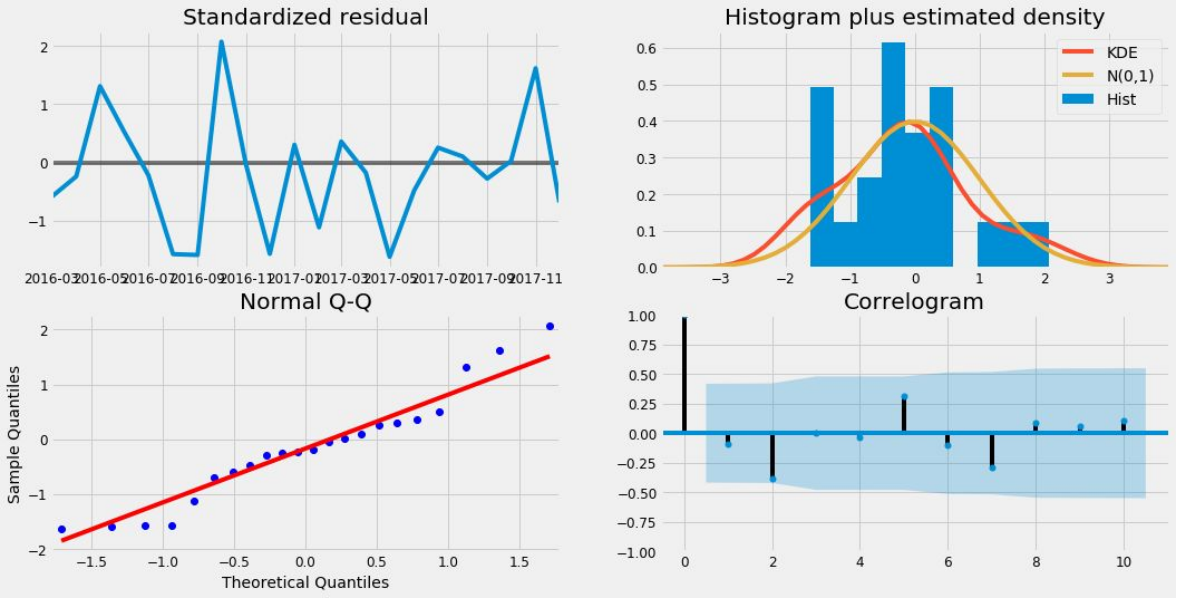
- Model: ARIMA(p,d,q)
  - ❏ p: the number of autoregressive terms
  - ❏ d: the number of nonseasonal differences needed for stationarity
  - ❏ q: the number of lagged forecast errors in the prediction equation
- Used **grid search** to find the set of parameters that yields the best-performance model. With the lowest AIC, we obtained the best option: ARIMA(1, 1, 1) x (1, 1, 0, 12)

```
ARIMA(1, 1, 1)x(1, 1, 0, 12)12 - AIC:343.6037335973577
```

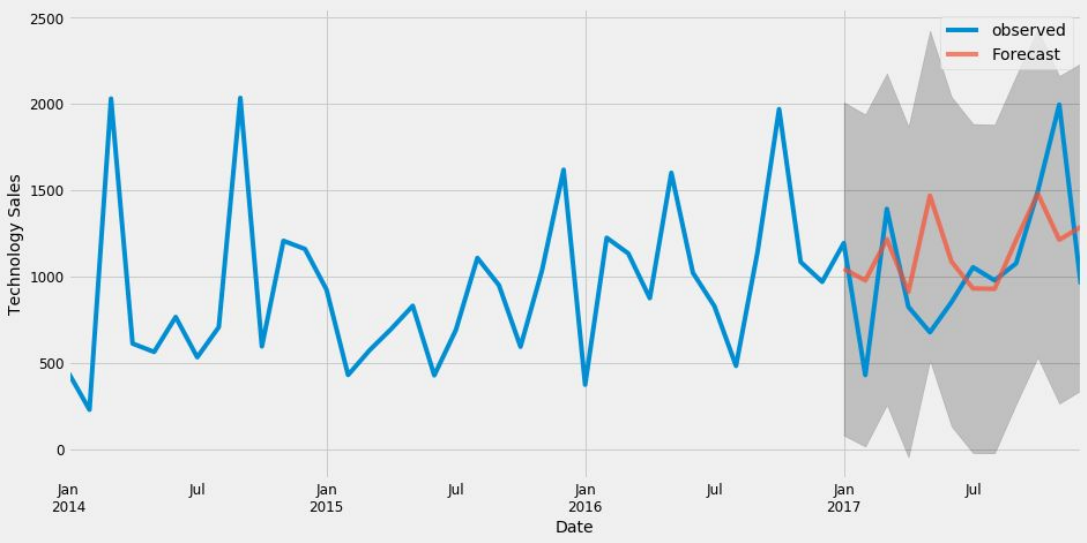
- Model fitting:

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2660	0.249	-1.067	0.286	-0.755	0.223
ma.L1	-1.0001	0.348	-2.870	0.004	-1.683	-0.317
ar.S.L12	-0.5003	0.175	-2.852	0.004	-0.844	-0.157
sigma2	2.243e+05	1.55e-06	1.44e+11	0.000	2.24e+05	2.24e+05

- Model diagnosis: they are mostly satisfied.

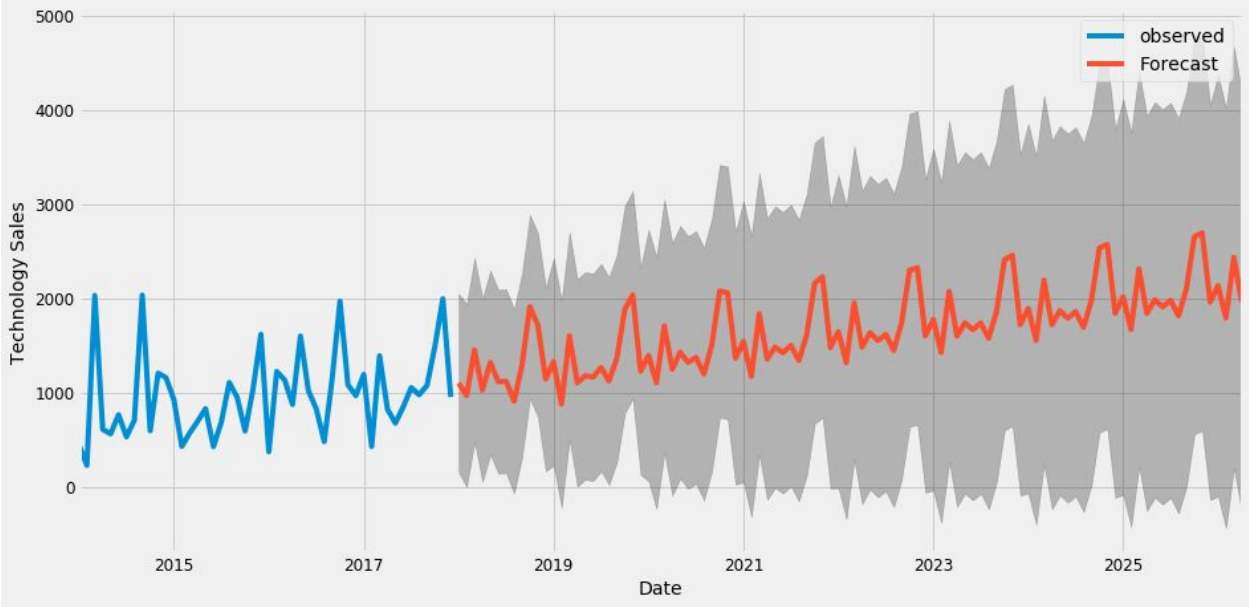


- The line plot shows that the observed values vs predicted values. Our forecasts and observed values both show an upward trend starts from the beginning of the year while the predicted trend is smoother than the observed trend.



- (Root Mean Squared Error) **RMSE: 387.42**. It means that our model could predict the average daily technology sales in the test set **within 387.42 of the real sales.** Our furniture daily sales **range from 230 to 2036**. In my opinion, it is acceptable but need improvement.

PRODUCING AND VISUALIZING FORECASTS

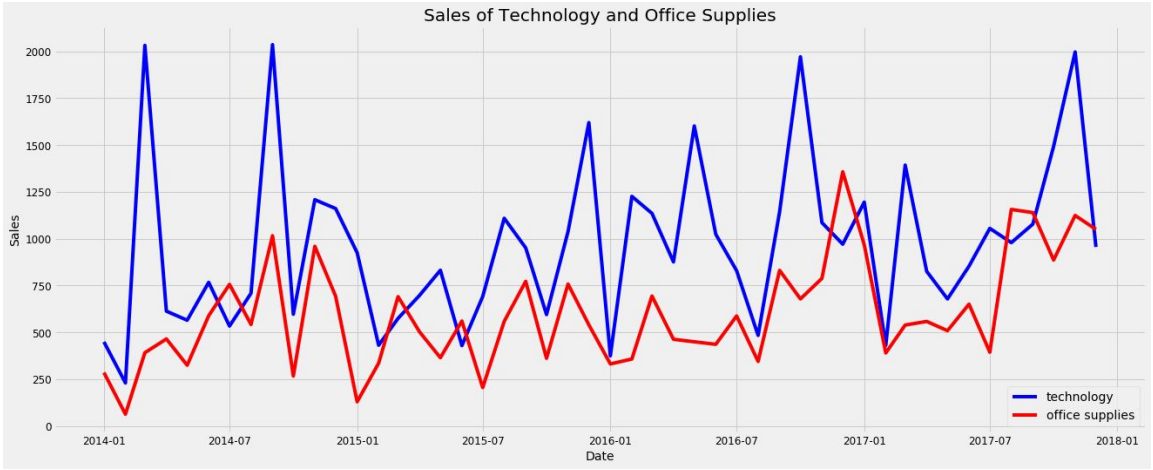


- Our model captured the technology sales seasonality. It is surprising that the sales for technology only have slightly upward trends, which is out of my expectation that technology products have short renewable period. In this case,

we wanted to investigate other categories of products to see whether they had a similar trends. We selected office supplies.

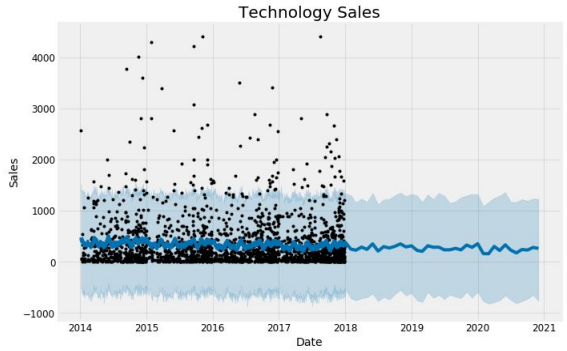
TECHNOLOGY SALES vs. OFFICE SUPPLIES SALES

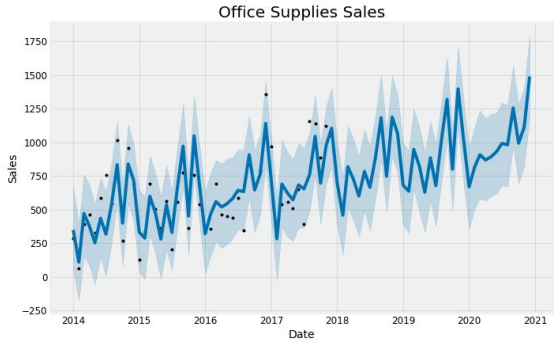
- We can see that technology sales and office supplies sales have similar seasonal patterns. They have mostly higher sales in mid year while have lower sales at the beginning and end of a year except 2017. Overall, the sales of technology are higher than those of office sales except for some occasions.



COMPARING FORECAST WITH PROPHET

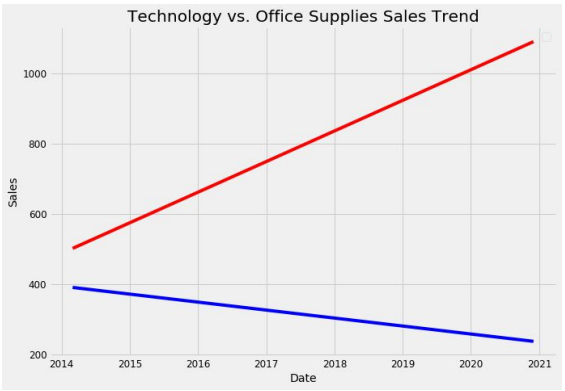
- Prophet: released by facebook, it is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It has excellent performance t with time series that have strong seasonal effects and several seasons of historical data.
- Technology sales have a relatively stable trends with prophet while office supplies sales have an obviously upward trend while with monthly patterns.

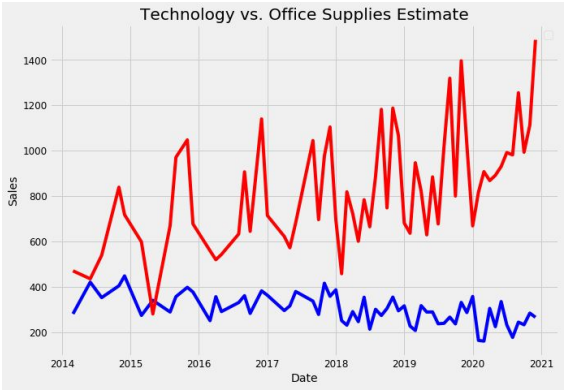




Technology Office Supplies

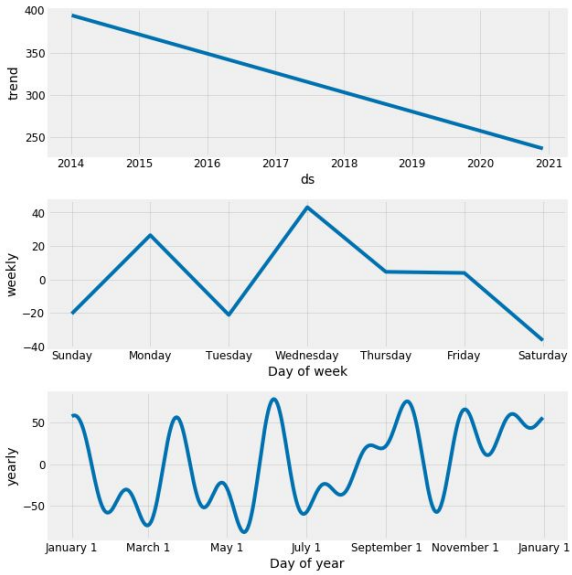
- Joined them together, we found them they have completely opposite trends!

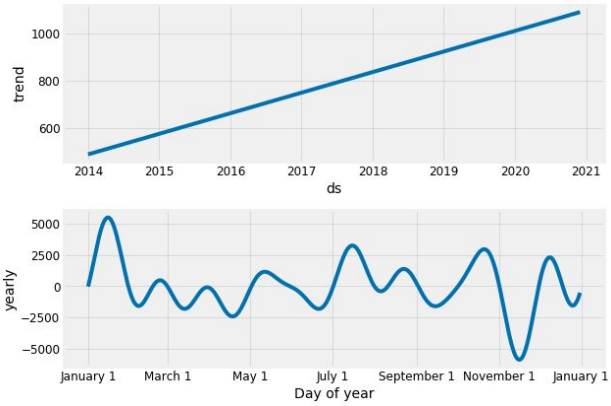




Technology Office Supplies

- Used prophet model to visualize respectively:





Technology Office Supplies

- The sales for technology have been linearly decreasing over time and will be keeping this trend. **The worst month for technology selling is May, and the best months are June and September**. However, the sales for technology shows a very positive increasing trend over time. **The best month for office supplies is February and the worst month is November.**

REFERENCE:

1. https://people.duke.edu/~rnau/411arim.htm
2. https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b
3. https://facebook.github.io/prophet/
4. https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3
5. https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-prophet-in-python-3