Main objective of the analysis that also specifies whether your model will be focused on a specific type of Time Series, Survival Anlaysis, or Deep Learning and the benefits that your analysis brings to the business or stakeholders of this data.

Main objective is to explore Superstore sales data to identify the most value-generating category and model their sales with different variants of Time Series models, to potentially gather any insights to help us better understand recent market trends.

Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.

Given data set "Sample - Superstore" on 9994 records of daily sales ordered during 2011-14 at Superstore, with 21 feature measurements covering the following:

1) Order details (e.g. discount, quantity, segment, category, sub-category, product name)
2) Customer info (e.g. customer name, country, region, state, city, postal code)
3) Time components (e.g. order date, ship date)
4) Sales metrics (sales, profit)

Brief summary of data exploration and actions taken for data cleaning or feature engineering.

1) Simplify Time Series:
   - target variable: Sales
   - grouping variable: Category
   - time component: Order Date

	Category
Office Supplies	6026
Furniture	2121
Technology	1847

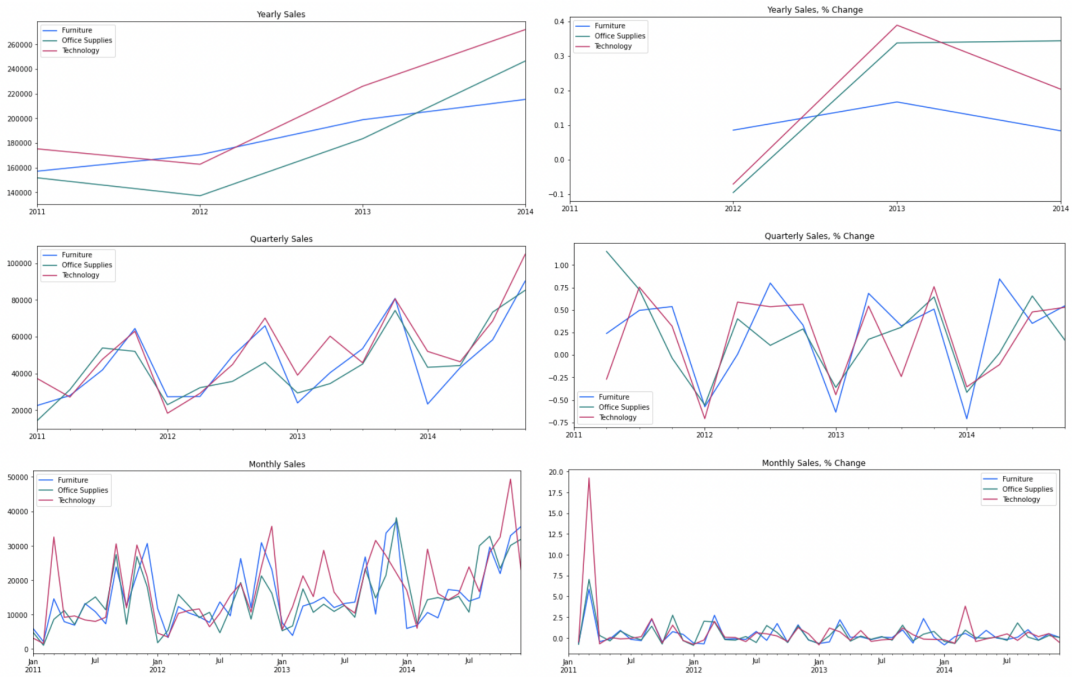
	Sales
Category	
Office Supplies	719,047.03
Furniture	741,999.80
Technology	836,154.03

Note that Technology items make up the lowest share in order counts (18%), but have contributed the most across all categories in terms of Sales (36%).

2) Data Cleaning:
   - Fill in dates with no orders to get a complete date range as index
   - Impute null values with zeros for dates where there were no sales on certain categories
   - Confirm all records have no null values otherwise

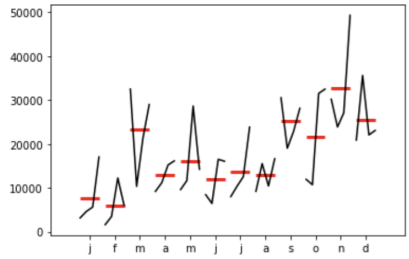
3) <u>Feature Engineering & Visualizations:</u>

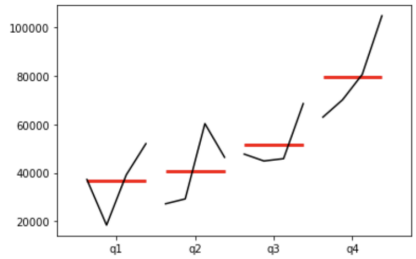
- Resampling daily records to monthly, quarterly and yearly frequencies
- Create run-sequence plots for preliminary insights, and check for trends
- Calculate and plot out % change, cumulative sum, rolling mean & standard deviation
- Create seasonal plots to check for seasonality
- Create autocorrelation & partial autocorrelation plots to check for autocorrelation
- Divide time series of focus into chunks, and calculate statistics to check for heteroscedasticity



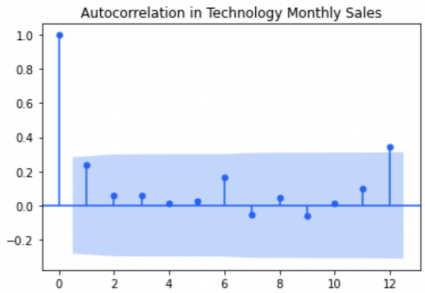
Note that Technology sales have remained the highest almost at all times across categories, and had its largest growth in 2013 after which the growth rate slowed down.

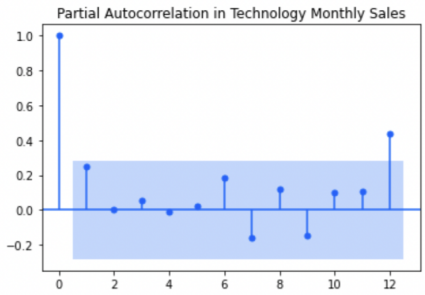
Note also that on a high level there's growing trend year over year, and there seems to be a seasonal sales spike in the later part of each year potentially related to holiday promotions, which can be further confirmed by the seasonal plots below:



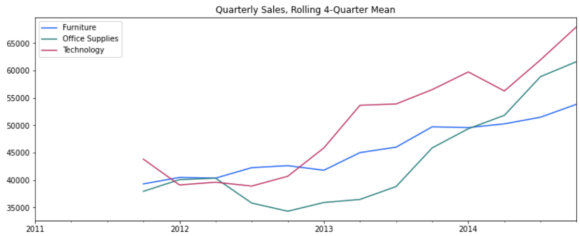


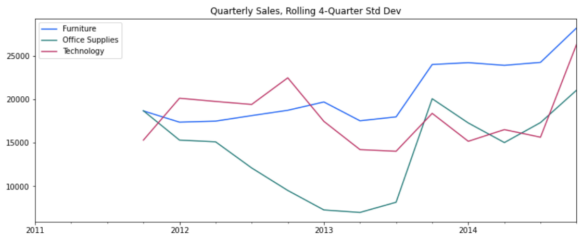
Otherwise, the monthly trend fluctuates in a relatively random fashion, and the sales in a given month does not seem to depend noticeably on previous month(s). That is, there may not be significant autocorrelation in our data. The ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots below confirm the same.





Lastly, the rolling 4-quarter views below shows also the growing trend. And especially for our focused category (Technology), the data seems to be varying within a relatively stable range over time. However, the chunk statistics suggest that the variance may not be considered constant. In other words, our monthly Technology Sales data presents heteroscedasticity.





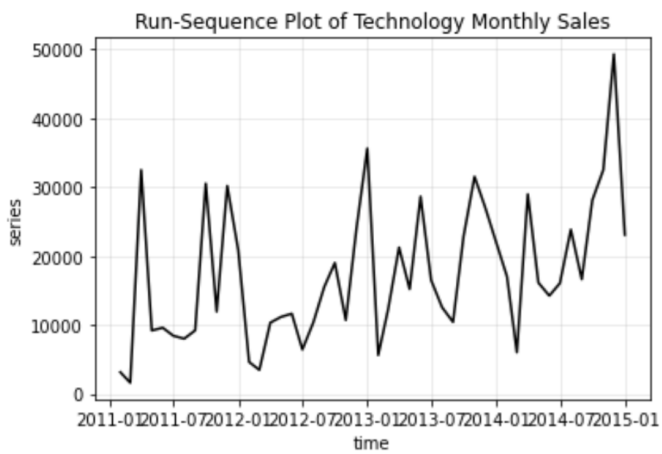
Chunk	Mean	Variance
1	32,247	2.5e+07
2	55,392	5.8e+07
3	23,829	2.9e+07
4	57,562	1.6e+08
5	49,746	1.1e+08
6	63,285	3e+08
7	49,250	7.9e+06
8	86,766	3.3e+08

4) Data Exploration Summary

- Highest value-generating category: Technology
- Time series of focus for modeling: Monthly Technology Sales
- Time series is identified as non-stationary:
    1) Trend: yes - growing at yearly level
    2) Seasonality: yes - at yearly cadence, lasts for 4 months towards end of year
    3) Autocorrelation: minimal dependence on past values, differencing may not be needed
    4) Heteroscedasticity: yes, a log transformation will be needed to remove it
- Multiplicative model is best suited: magnitude of Seasonality & Residuals fluctuate with Trend

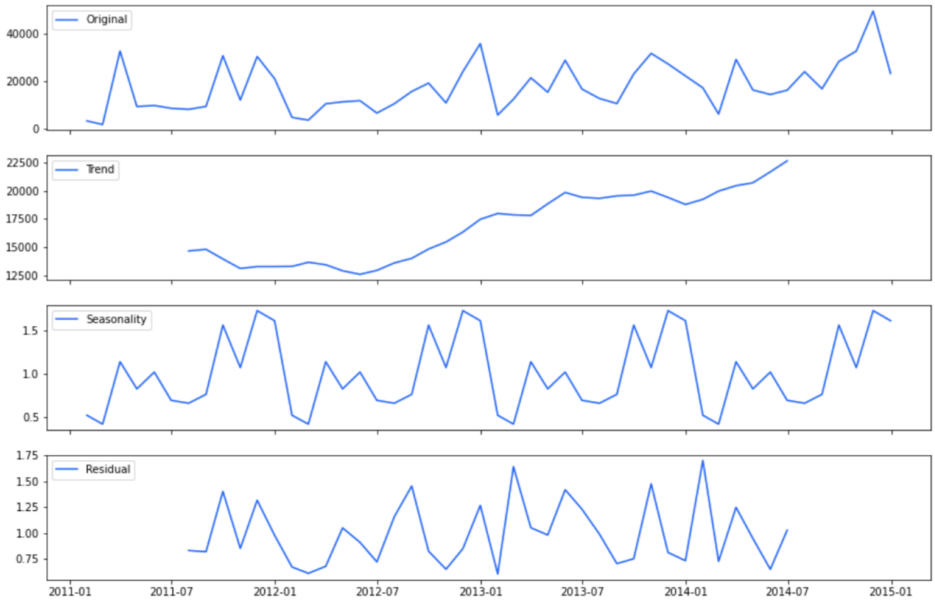
Summary of training at least three variations of the Time Series, Survival Analysis, or Deep Learning model you selected. For example, you can use different models or different hyperparameters.

As discussed above, we'll be exploring different Time Series models on the Technology monthly sales data as shown below. Historical data from Jan'2011 through Jun'2014 will be used as training set, while the last 6 data points from Jul'2014 through Dec'2014 will be our test set for model validation purpose.

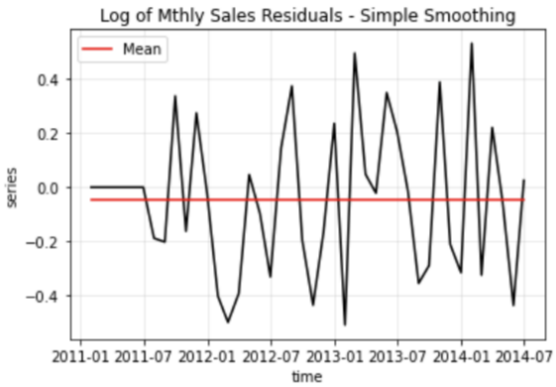


1) Smoothed Model

The Smoothed Model first decomposes our training set into Trend, Seasonality, and Residual as below, assuming a multiplicative relationship among the 3 terms.

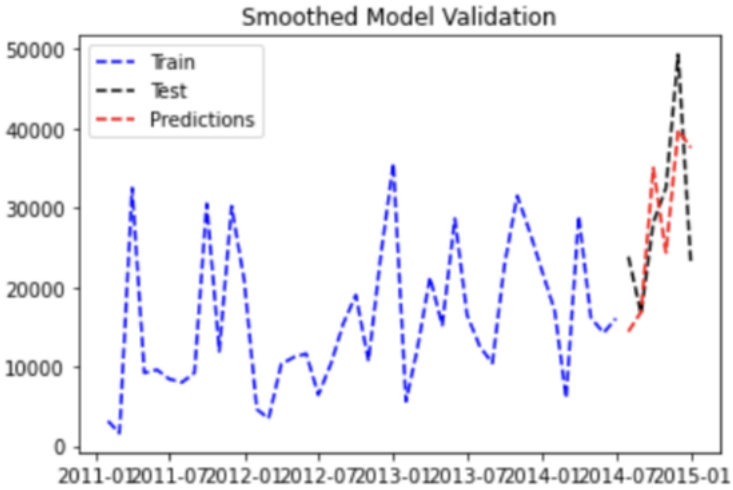


Note that the Residuals, despite a wide fluctuation, fluctuate around a constant mean. We hence need to perform the log transformation to bring down the magnitude of fluctuation (i.e. remove heteroscedasticity), before we can apply a Simple Smoothing, as shown below:



With Double Exponential Smoothing on Trend, and Triple Exponential Smoothing on Seasonality, we're then able to extrapolate 6 months further and resemble those 3 parts to get the predictions on the test set: Trend * Seasonality * exp(log of Residuals).

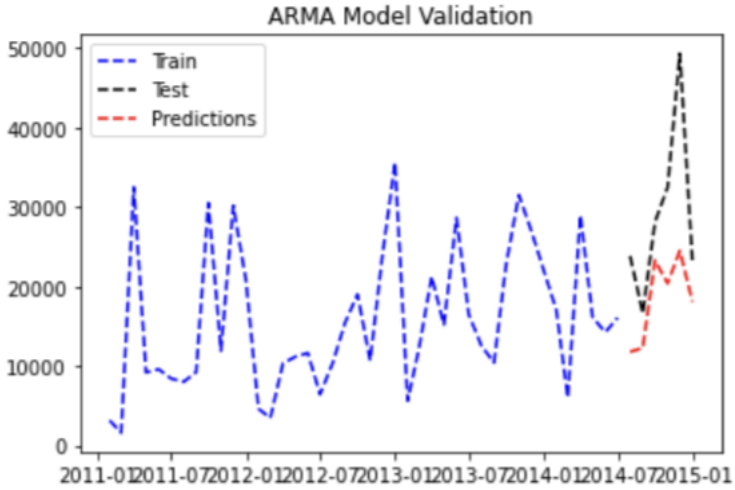
The validation graph below suggests that our Smoothed Model does a descent job forecasting on our out-of-time sample, with MSE of 0.8e8.



2) ARMA Model

The ARMA Model may be not as well-suited in our case, due to 1) non-stationarity; and 2) minimal dependence on past values and past forecast errors. However, we could try to capture the growing trend as well as the yearly seasonality effect by fitting a ARMA(p=12, q=0) model.

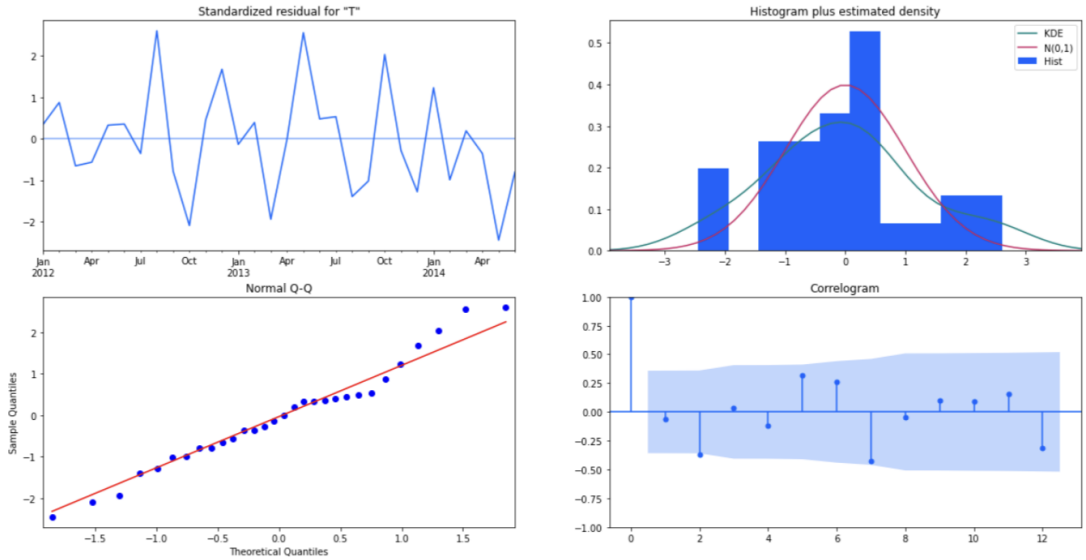
As expected, the validation graph below suggests that our ARMA Model may be able to vaguely capture the seasonal pattern by looking back 12 months prior, but it fails to extrapolate the growing trend that's more recognizable on the yearly level. It has a MSE of 1.6e8.



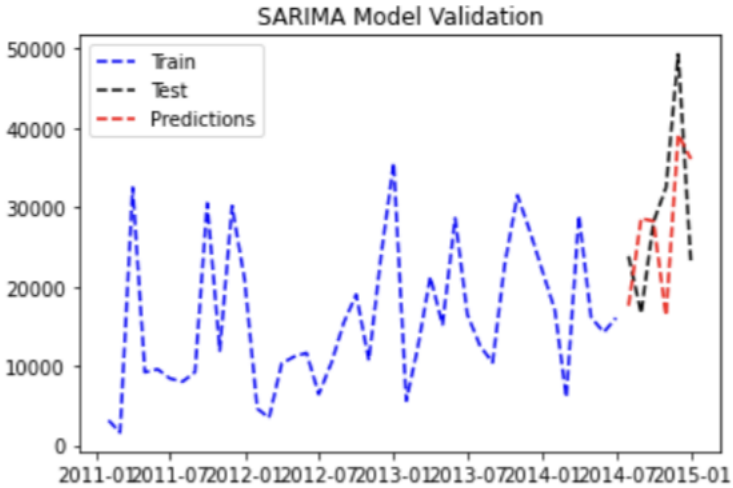
3) SARIMA Model

The SARIMA Model could accommodate our needs on capturing year-over-year trend as well as yearly seasonality, with hyperparameters set to be SARIMA(p=12, d=0, q=0)(P=0, D=1, Q=0, S=12).

Based on the plot diagnostics below, our Residuals seem to be centered around zero, and approximately following a Normal distribution - except for the last few data points from first half of 2014. Combined with our run-sequence plot, there might have been a calendar event (potentially the series of hurricanes in early 2014) that affected the timing of our Sales within that year.

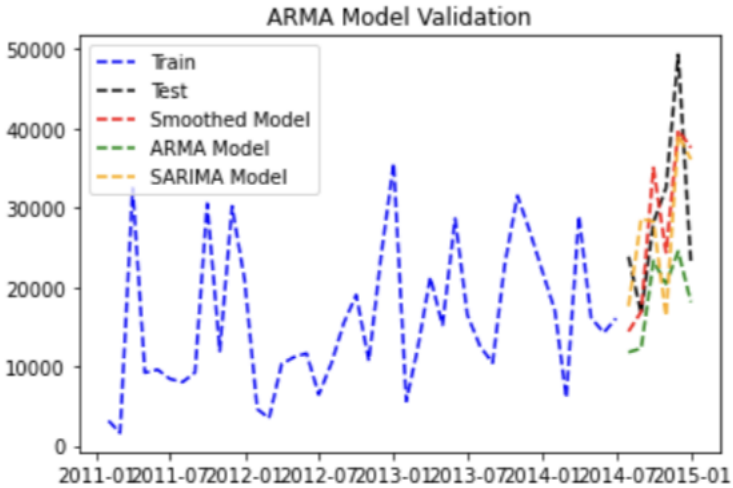


The validation graph below indicates that our SARIMA Model have successfully overcome the critical drawbacks of our ARMA Model and provides a comparable forecasting results as our Smoothed Model on the test set, with MSE of 1.2e8.



A paragraph explaining which of your models you recommend as a final model that best fits your needs in terms of accuracy or explainability.

Plotting the model validation results on the same graph below, we can easily tell that the Smoothed Model is the winner. The comparison of their error rates on the test set suggests the same, with our Smoothed Model having the lowest MSE across the three attempts.



The Smoothed Model also provides the visualization of the decomposed parts (Trend, Seasonality, Residuals) to allow for better interpretation as well as empirical adjustments (e.g. business process changes, industry rule changes, and other out-of-control calendar events) for extrapolating further into the future to plan and steer the business more effectively.

The Smoothed Model is therefore recommended as the final model that best fits our needs.

Summary Key Findings and Insights, which walks your reader through the main findings of your modeling exercise.

1) Technology is considered the highest sales-generating category at Superstore in 2011-14. With 18% share in count of orders, it makes up 36% of the sales total. Its sales have been steadily growing in those years, with a minor dip in 2012 followed by a huge spike in growth rate going into 2013, which then slowed down a bit in 2014.

2) Despite the year-over-year growing trend, the golden period for sales promotion is the last 4 months of every year where seasonal spikes occur in a multiplicative manner.

3) Certain calendar events (like hurricanes) may distort the sales trend, but depending on the nature of the event it could be more of a timing shift within the year rather than a permanent loss or incremental gain.

Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model or adding specific data features to achieve a better model.

1) Perform similar analysis with different groupings to identify other key drivers, whether it's product-specific, consumer-specific, region-specific, deal-specific etc.

2) Gather sales data beyond 2014 to help understand the market trend today and look into the future.