Nancy Melucci, [njm99@drexel.edu](mailto:njm99@drexel.edu)

Sarah Haley, [slh54@drexel.edu](mailto:slh54@drexel.edu)

Zach Carlson, [zc378@drexel.edu](mailto:zc378@drexel.edu)

INFO 659 Final Project Summary

Below is a brief summary of our final project for INFO 659. For more information on the code, dataset, and model please refer to our [Github](https://github.com/zachcarlson/SuperstoreSalesPredictor) repository. The file SuperstoreSalesPredictor.ipynb contains the Python code and figures. The file SuperStoreSalesinR.Rmd contains the R code and figures.

| Outline | Description |
| --- | --- |
| Introduction and Problem Statement | To examine trends in sales and profit and attempt to predict business future trends. We had two questions we were trying to answer with this project:   1. Are sales and profit increasing over time? 2. Can we accurately predict sales for 2014-2015 using a model?   We planned to answer both of these questions in both R and Python using at least ARIMA (see **Methodology**). |
| Data Sources and Preparation | We used a dataset from Kaggle with sales and profit data from the Superstore chain from 2011-2015. The dataset is linked [here](https://www.kaggle.com/jr2ngb/superstore-data). |
| Data Exploration, Visualization, Cleaning and Transformation  Data Exploration, Visualization, Cleaning and Transformation (con.) | The dataset was in comma-separated values form. We used both R and Python based notebook programs to import data and create dataframes for viewing it.  In the EDA stage, we noted that there were many missing data points in Postal Code, and we planned a more broad (global/regional) analysis, so we decided to drop that column.  We created columns to facilitate our analysis by wrangling the dates. These were Year, Year-Month, Abbreviated Month, and Month-Date. If there should be a need for a more granular view of data in a region, we can return to the raw data and devise a way to manage this issue. All Sales/Profits figures are in United States Dollars.  We created “big picture” summary descriptive statistics focused on factors that influence Revenue (Sales, Quantity, Discount Rates, Shipping Rates and Profits) over the four years covered in the dataset. We used these for identification of main revenue factors and hypothesized about influences on these factors.  To see if long-term trends existed in our dataset, we utilized several rolling average windows. For example, 7-day, 30-day, and 360-day.  We created numerous visualizations, box plots and line/area charts in both R and Python. One feature we noted in many of our pictures was an increase in sales, but a much less clear sense of the trends in profits, thus, we also used a logarithmic (log10) transformation on the profit data when plotting both together. However, this transformation was not required in plots where profit and sales were plotted separately (e.g. the Python code). |
| Methodology | We selected the Autoregressive Integrated Moving Average (ARIMA) forecasting model, based on Box-Jenkins, for our model to predict sales and profit and used a step-forward validation technique for our predictions. This is a three-step process of conditioning the data, estimating the model parameters and assessing the strength and utility of the model. As is the case with many other analytic and business project models, this also includes, if needed, a return to step one.  Our ARIMA model prep included the standard four components: Trend, Cycle, Seasonality and Random (“stuff happens”). It is not hard to see that the business environment of Superstore has a strong tendency to seasonality.  We differenced the time series to remove any seasonality but using monthly sales sums data and differencing with a lag of 12 (essentially a yearly difference). We then went through each month of 2014 and predicted the sales one month in advance. After predicting a given month in 2014, we appended the actual data to a list to train the next month. We iterated month by month until we predicted the entire year. |
| Modeling and Results | We considered autocorrelation, autoregression, and moving average terms in our model building. We summarized the sales data by grouping by month and calculating a sum for each month. The frequency used was the start of the month. After accounting for seasonality, by calculating the difference of each month by subtracting the previous year from it (i.e. lag of 12), we optimized the different terms in the ARIMA model: p, d, and q and P, D, Q for seasonality. We ended up with an ARIMA model with the order (2,1,0). Additional configuration could have been done, however our model predicted sufficiently.  Using the step-forward validation technique, we ended up with an RMSE of around $40,000. For context, the monthly total sales data were in the range of $200,000 - $500,000 / month. |
| Evaluation of Results | **Question 1: Are sales and profit increasing over time?**  As we suspected from the start, Sales and Profit are increasing. Visualizing the data using rolling averages helped highlight the long-term trends and seasonality of the data. The 360-day rolling average clearly shows the sales and profit are increasing over time. Plotting the log of the sales and profit data helps with readability as well.  **Question 2: Can we accurately predict sales from 2014-2015 using a model?**  Yes. By utilizing the ARIMA model we were able to accurately predict sales for the year of 2014. The order of our ARIMA model was order=2, 1, 0. We ran the time series through the ARIMA model after calculating the difference on a yearly-basis using monthly sales data and a lag of 12. The RMSE for the year was approximately $46,000, which is good considering the scale of sales is between $100,000 - $500,000. |
| Conclusion and Future Work  Conclusion and Future Work (con.) | **Question 1:**  Sales and profit are increasing over time and rolling average plots are a great tool to visualize long term trends in order to answer this type of question.  **Question 2:**  With our current data cleaning and preparation, sales was able to be predicted.  While the model accurately predicted the sales data, we are far from a state where we can deploy this model in a professional setting.  **Overall Takeaways:**  While our analysis answered both questions sufficiently, our analytic pipeline is nowhere near ready for professional deployment. This is mainly due to context and the reliability of the dataset used. The dataset was posted on Kaggle with **little-to-no documentation**. We had to assume the sales and profits units. While we naturally went with USD, Superstore is a Canadian company, and the values could just as easily be in CAD. Finding a better dataset or contacting the publisher early enough could fix this problem.  Additionally, we used the step-forward validation technique of **assuming we would only predict one month into the future**. This gives us great predictions, however, a team may want predictions made farther into the future. These predictions would obviously be less accurate. If this was the goal, more optimizing would be necessary.  Finally, **the training data only included three years of data.**  Fortunately, these three years were fairly consistent and similar to the fourth so the predictions were sufficient. However, if we want truly accurate predictions that are more robust, more years of data would be essential. |