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 and profit 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.  From the results tables we created numerous visualizations box plots and line/area charts (shown), time series as well as other visualizations for trends as part of the EDA. 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. |
| Methodology | We selected the Autoregressive Integrated Moving Average (ARIMA) forecasting model, based on Box-Jenkins, for our model to predict sales and profit. 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 business environment of Superstore has a strong tendency to seasonality. |
| Modeling and Results | After the preliminary data preparations and transformations, we decided to drop Discount Rate and Quantity (the latter was brought back later in the analysis), which appear to be stable across sales with little variation based on the descriptive analysis, to have a better, clearer look at Sales and Profit. Shipping Cost has a much larger variation (more outliers) and shipping costs tend to be quite high based on our information. Profit and Shipping Costs are subsets of Sales; therefore, we considered these as a group in order to fully understand profitability trends. Total sales are increasing year-to-year, but average sales are not. We brought back Quantity to examine this more closely. The vast majority of Superstore sales are small-quantity but there are outliers in total sales related to relatively infrequent large quantity orders. This effects average sales despite the fact that most sales are small inside because of the relatively sensitivity of means to outliers. We tried to understand these large quantity outlier sales to determine if they were signal (would produce valuable information to understand consumer behavior and improve profitability) or noise (random and not useful to understand.) Log transformation, followed by correlational analysis of sales and profit and standardization of the scale helped to clarify the fluctuations. It remained clear to us that profit was far less stable than sales. We then calculated rolling averages for Sales, Profit and Shipping Costs and tried to detect “stationary data” (data series that are not actually changing despite the appearance of the time series) using the seasonal and trend decomposition using Loess (STL) on Sales. |
| 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 help with readability as well.  **Question 2: Can we accurately predict sales and profit from 2014-2015 using a model?**  No? |
| Conclusion and Future Work | **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, profits and sales are too unstable to predict, despite robustly increasing sales over time. One way to address this, for example, would be for management in each region administer surveys to consumers in an attempt to determine what other factors may be affecting the decision to shop, what to buy, and other “soft” variables.  We did not have access to a breakdown of online versus brick-and-mortar sales information in this data set. It would be helpful to know this to understand more clearly, what consumers are purchasing in which retail mode. Knowing this might lead to more deliberate strategies for product sales that will reduce shipping costs and encourage more buying – ultimately the most direct, effective way to improve profitability.  **Overall Takeaways:**  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.  In addition to this, [more things] |
|  |  |