Final Project Proposal

By: Ismet Okan Celik, Rishab Katteri, Swapnali Patki

Problem Statement: We will be using a time-series related data set to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer.

More specifically, we will try to build an adequate ARIMA model that more accurately predicts the unit sales for thousands of items sold at different Favorita stores. Using test and training data, and with a couple of predictions, we will see what features are important or what features we should engineer in order to maximize profit and to predict when it will happen when Corporación Favorita reaches a certain sales point.

Description of Dataset: The provided data from this kaggle competition will be used for this project. The data can be found through the following link:
https://www.kaggle.com/c/store-sales-time-series-forecasting

It mainly contains the transaction records, any holiday events, store locations with the training and testing dataset.The data might need a bit of cleaning and preprocessing. We look forward to gathering insight on foresight and see what techniques we can use at our disposal from class in a constructive and innovative way.

Implementation Plan:

March week 1: Data Preprocessing- Following this idea, we will strive to come up with new ways to modify our dataset and find the best way to structure our data so that we may derive the most useful information from our analysis each week.
March week 3: Program Design- We will try to realize a practical and novel way of interpreting this information and see if we can implement something that best fits the data according to what we have gone over in class.
April week 1 to week 3: Program Implementation- Within this timespan, we reckon to complete the coding and debugging. We will finalize any revisions we have in this game plan and see what we might have to change from our original problem statement.
April week 4: Testing and Final Documentation

Team Members & Task Allocation:

Ismet: Pre-processing the data and handling the training/testing process through different learning systems to best predict certain goals for the company
Rishab: Building the ARIMA model and a focus on the time series analysis with added residual analysis
Swapnali: Using multiple linear regression and feature engineering to identify the best and most feasible predictors while checking its statistical significance, while also testing how effective this regression holds for out-of-sample data using cross-validation.