DS 3001
Esha Sharma, exy8eb
Celestine Nguyen, haw3wk
Jennifer Vo, phb8pt
Shruti Bala, aen6ju
Chetu Barot, zrz9ry
Tulsi Patel, cjw6jz

Pre-Analysis Plan

Dataset: https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data
Git repo: https://github.com/CelestineNguyen/DS3001_SalesProject/tree/main

Main/Chosen Question:
How does the type of products sold affect sale rates for stores?

Other Question Ideas:
How can we best predict the future number of sales for items?
How do promotions impact sales across different product families and store types?
What is the effect of holidays and events on store sales?

What is an observation in your study?
Observations in the training set include the different products sold, what product family the item belongs to, the date the data was collected and the number of the store the item is located, and the predicted number of sales for that specific item family.

Are you doing supervised or unsupervised learning? Classification or regression?
We will do supervised learning, specifically we will train a regression model on the data. We will be trying to predict what features in the dataset affect the sales in the dataset. We are mainly trying to predict the sales column and how these values can change.

What models or algorithms do you plan to use in your analysis? How?
We plan to create different regression models using different sets of variables/data and compare which models perform the best effectively giving us insight into what variables are most important to sales data

How will you know if your approach "works"? What does success mean?
We will compare R-scores and RMSE across different models. Success for this project is measured based on how well we can predict the effects of different variables, such as product type, store number, and amount of transactions, on sales. Using an autoregressive model will

help us determine whether our sales forecast is stationary, has a low RMSE, and is statistically significant with the correlation of sales and another variable.

What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?
We could encounter too much data or “noise” where we may need to narrow down the columns we are using. With this, we may encounter models that perform really badly, but hopefully by tweaking our feature matrix X we can mitigate that

Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?
We will be mainly using train.csv which has ‘family’ as a categorical variable that should be one-hot encoded. Additionally, the ‘date’ column is in year-month-day format that should be turned into a numerical variable to be easier utilized. For some of our models, we will be using columns from the other csv’s available where we will need to combine data based on the store identification (‘store nbr’) and time (‘date’).

Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.
We will evaluate our models using R^2 and RMSE. For the model that has the highest R^2 score and lowest RMSE, we will report a table of regression coefficients to gain more insight into which features have the biggest impact when predicting sales on a given day.