Stakeholder: Dr. Rania Hodhod - Professor of Computer Science

Project Overview:

You are provided with historical sales data for 45 stores located in different regions - each store contains a number of departments.

- Important Dataset:
    - Sales Dataset
        - This should be the only dataset we need to look at
        - Features: Store (ID), Department (ID), Date, Weekly_Sales (float), Is_holiday (Boolean)
        - Stores
            - Have multiple departments per store
                - Department is dependent
        - Date
            - Starting date of store
            - +7 days gives us next data
            - NOTE: might have to convert dates into two dates to work with weekly sales
                - Start date and end date for the weeks
            - https://towardsdatascience.com/machine-learning-with-datetime-feature-engineering-predicting-healthcare-appointment-no-shows-5e4ca3a85f96
                - Explains date features
        - Is_Holiday
            - Boolean that checks if the week contains a holiday date
            - Holidays Include:
                - Super Bowl
                - Labor Day
                - Thanksgiving
                - Christmas
    - Features Dataset
- Dataset Information:
    - 421570 Instances
    - No NAN values
    - No duplicate instances
    - Dates
        - 143 unique dates
        - Duplicate dates
        - Date Format
            - Day/month/year
    - Only one Numerical feature (Weekly_Sale)
    - Weekly_Sales

        - includes negative numbers
            - 1285 negative sales
            - 73 instances marked as 0 sales
        - Sales is profit
            - We can have 0 profit, positive profit and negative profit.
            - 0 may mean store is closed
            - Negative may mean that we had too many cost
- Dataset Questions:
    - How do I represent dates?
        - Dates are categorical
        - Dates will be split into Day | Month | Year
    - Should I separate months and days from years?

The company also runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas.

The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks.

The goal of this project is to develop a linear regression model for a retail business to enhance marketing strategies.

Data Sources:

1. Use the datasets that include customer demographics, purchase history, and online behavior (consider the attached features, sales, and stores datasets).

Tasks:

1. Clean and preprocess the data, addressing missing values, normalization, and feature extraction.
    - ☑ ~~Get total amount of instances~~
        - 421569 instances
    - ☑ ~~Get total missing value amount (We have no NA variables)~~
        - ☑ ~~Get missing value amount per feature~~
            - No NA variables
            - Weekly sales include negative numbers
                a. 1285 rows
                b. Negative sales may imply losing money?
                c. May be an issue, may or may not remove
    - ☑ ~~Get amount of outlierst~~
        - ☐ Use DBSCAN for outliers
            - May or may not need to do this due to nature of data

- □ Convert dates
    - □ Days is a new feature
    - □ Month is a new feature
    - □ Years is a new feature
- □ Convert Holidays to one hot encoded features
    - □ Use https://stackoverflow.com/questions/45870820/how-to-check-if-today-is-monday-in-python
    - □ SuperBowl
        -
    - □ Labor day
        - First monday of each september
    - □ Thanksgiving
        - 4th thursday of nov
    - □ Christmas
        - Dec 24
- ☑ ~~Convert True and false bools to one label encoding~~
- □ Normalization
        - Might be best to use standardization compared to normalization for store data
            - Better for outliers
        - Store and department ID might have to go through a different process due to not being 'True' ints/floats
- □ Use PCA for feature extraction
    - □ Compare features with most to least correlated (Confusion Matrix)
    - □ Join features based on correlation
    - Things to potentially remove inside Store Data:
        - Store Data - Feature Type:
            - Reasons: No description of what type A,B, or C is.
        - Store Data - Feature Size:
            - Reasons: We do not need to measure the size of a store
    -

2. Conduct Exploratory Data Analysis (EDA) on the provided datasets.
    - □ PCA feature extraction (Demonstrate important features)
    - □
3. Predict the department-wide sales for each store for the following year
    - Prediction input: Store ID, Dates
        - Might have to make a function to loops through stores and departments
    - Prediction Output: Weekly_Sales
    - □ Predict Each stores
    - □ Model yearly Outcome for each store

4. Model the effects of markdowns on holiday weeks
    - □ Label holiday weeks from output data
5. Provide recommended actions based on the insights drawn, with prioritization placed on largest business impact
    - □
6. Use appropriate metrics to evaluate the performance of the models, such as clustering metrics (Silhouette Score)
    - □ https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.silhouette_score.html (Silhouette Score)
    - □
7. Document the methodology, results, and insights gained from the analysis in a comprehensive report.
    - □
8. Include visualizations and discussions on how the segmentation can inform marketing strategies and customer engagement efforts.
    - □
9. Prepare a 5-7 minute presentation summarizing the problem statement, methodology, results, and actionable insights. Emphasize the benefits of using semi-supervised learning for customer segmentation.
    - □