Final Projects (65 points)
Introduction

The final project of the course entails analyzing inventory data of 132 stores of an international coffee shop chain with more than 1000 stores worldwide and $400M revenue. The purpose of the analysis is to use the sales prediction and inventory alignment to improve sales and reduce waste, resulting in a more efficient operation.

In a team of three, in section one, one member mainly focus on analyzing the data of three specified stores (StoreID=18, 117, 332) and two persons mainly focuses on the data of the entire 132 stores. Single store results and corporate-wide results should be communicated between team members to help them with their individual analysis and ensure consistency. But you will be graded as one team. In section two, all team members work together.

Data Description

Click to Download: Link
"https://drive.google.com/file/d/1PLppvwAQ-9sRYAZGZt54sjBo5r8uBDDR/view?usp=sharing"

- **StoreID:** ID of store
- **BusinessDate:** Date of record
- **PLU:** ID of inventory
- **Description:** Name of product
- **ItemType:** Type of product
- **CategoryLvl1Desc:** Main level of product's category
- **CategoryLvl2Desc:** The 2nd level of product's category
- **CategoryLvl3Desc:** The 3rd level of product's category
- **ReceivedQuantity:** The amount stores received from the distributor. They typically receive perishable items every 2 or 3 days based on the customer's feedback. But it could be every day or more than 3 days for some products.
- **SoldQuantity:** Quantity sold on a particular day.
- **EndQuantity:** Quantity at the end of the day – inventory condition. (Please note if they throw away expiring items, they record *zero* at the end of the day and they order them again to have them the next day.)
- **LatestOrder:** The number of items they requested.
- **StockedOut:** They record when a customer asks for an out-of-stock item. Yet, the cashier may not always record all the stocked out. But if they do, we are sure of the case.
- **GroupID:** Not Applicable
- **MissedSales:** Not sure what these data signify.

Section 1: Data Visualization (Coding and chapter one of the report) 30 points
Deliverables: Codes and chapter 1 of the report. Tabulate the team member names and specify each person's role in the beginning of the report.

Description:
Part A: Analyze the data of individual stores and draw insights based on the following:

1. Provide the box plots and statistics of 27 products, inventory patterns, stock out patterns and missed sales (note: do not rely on the missed sales data in the data set. See part 4 below for more information).
2. Show graphs of best seller and worst seller products of top 25% and bottom 25% and provide your insight into data. (The average selling price of a product is 3$). Go beyond graphs and just analyzing data. Think about "So What" when writing your report!
3. Show graphs of best and worst products based on their inventory management - Top 25% and bottom 25% and provide your insight into data (the average cost of a product is 0.5$) Identify where/when the store gets rid of the unpurchased products. Go beyond graphs and just analyzing data. Think about "So What" when writing your report!
4. Identify stock outs and estimate the loss of sales per year per product. Assume when we are out of stock, we conservatively lose 75% of the average of sales in the previous 4 weeks on the same weekday. You are welcome to make other reasonable assumptions. Clearly explain your assumption.
5. Show graphically how the product sales and inventory waste change. Investigate
    a. Impact of day of the week on sales and stocks (7 days)
    b. Monthly changes and patterns (for the duration of the data)
    c. Impact of weather condition based on two factors:
        i. Temperature
        ii. Weather condition (sunny, raining, cloudy, etc.)
        iii. Find the weather data from publicly available sources and add them as features to your data set. Weather has shown to have a huge impact on sales.
6. Investigate whether drive thru feature causes certain products to sell better or worse.
7. Investigate the impact of weekday/weekends and National Holidays by adding extra features.
8. Based on the store data, identify the stocking patterns across multiple stores. Are they provided with new products every day or restocking happens less frequently based on your insights?
9. Draw conclusions and suggest a recommendation to optimize the stocking. We will have a deeper dive into it in section 3.

Part B: Analyze the data of stores and draw insights based on the information provided above. Keep in mind that:

1. Stores have a variety of products that may not be found in the designated stores.
2. Seasonal changes based on the weather feature cannot be linked to weather because of the diversity of the locations and unavailability of zip codes but still you can identify the patterns based on the months of the year and US daily average temperature.

3. Individuals working on specified store data and all store data should communicate their findings to help each other draw impactful conclusions.
4. Analyze the data of all stores and identify best and worst stores based on top 25% and bottom 25% and provide your insight into data.
5. Based on the store data, identify the stocking patterns across multiple stores. Are they provided with new products every day or restocking happens less frequently based on your insights?
6. Draw conclusions and suggest a recommendation to optimize the stocking. We will have a deeper dive into it in section 3.

Note that we are working on real data. You may need to clean data, or estimate missing data based on averages or using other reasonable assumptions.

StoreID	Zip Code	Drive Thru	Address
18	91101	No	Pasadena, CA 91101
117	91105	No	Pasadena, CA 91105
332	92122	Yes	San Diego, CA 92122

Suggestions to optimize stocking:

- Managing an optimized stocking for the best-sellers is essential to maintain and grow sales (that when consumers ask the products are always available) meantime reducing waste. We notice that the top 25% of best-sellers are also the top missed sales and high waste. This is due to the fact the stores want to maintain high availability; however, the pattern of consumer demands is hard to predict.
- Sales are changing along with many factors, such as weekends, holidays, month-by-month, and weather. These factors can be carefully collected, then used through a predictive model to predict the quantity for the next week. When any factors (such as weather) change, re-input the features and predict again. Through the store manager's professional experience with an AI-driven prediction (multiple predictions) model(s), hopefully, we can optimize the inventory.
- Driving through is an important feature. We observe store 332 with a drive-through feature has two-fold to three-fold more sales than stores 18 and 117 without this feature. This feature is obvious for a coffee store to attract consumers for a cup of coffee or treat. We also notice store 332 has better inventory management overall. When a store has a drive-through, sales are more robust and stable, it is easier to predict the inventory to stock up.

Section 2: Prediction (Coding and chapter two of the report) 35 points
Deliverables: Codes and chapter 2 of the report.
The purpose of this section is to come up with a predictor for **sales of each product** based on which we can optimize the restocking and inventory management of stores 18, 117, 332. [(optional)You can try finding a model for inventory prediction]

1. Produce synthetic data using Generative models (GANs) to get accurate predictions even with little historical data where necessary.
2. Iterate through different combinations of features to identify the optimal features and remove potential correlated features (if any) for your predictions. Add weather, weekdays, holidays and temperature data to your features.
3. (optional) Start with a quick linear regression to get a sense of data. Linear regression may not result in a great prediction.
4. Use ensemble models. Develop the following models and compare the accuracies by comparing the confusion matrix, true positive rate, true negative rate, precision, and F1-score.
    a. Random Forest
    b. Gradient Boosting Machine
    c. XGBoost

    Perform a quick sensitivity analysis on the parameters of the model and try to finetune the default values where you see an opportunity to improve the model. Improving the machine learning models is where a data scientists will shine in their career and ahead of the game from others.
5. Document and highlight model improvements. Extra credit will be considered for team's effort on improvements. (5 points)
6. As we are dealing with time-series data, we would like to compare the results of the previous models with the following deep methods:
    a. CNN
    b. LSTM
    c. Transformer
7. Based on the results from the models above, we would like to predict sales based on the following scheme:
    a. Weather forecast i) 1 day ahead ii) 3 days ahead (shipping from corporate warehouse) iii)10 days ahead (distributer order to the manufacturer)
    b. Week of the day
    c. What models provide the best prediction for the sales forecast 1 day from today, 3 days from today and 10 days from today.
8. Use 80% of data for training and 20% of data for testing. Compare the model accuracy for training and test data sets.
9. For individual store data, teams 1-10 focus on the first 13 products and groups 11-25 focus on the second 14 products. Eliminate the products that are not common between your three stores.
10. Find the best models across all stores (the model with best predictions, i.e., the lowest error) Apply the individual store model to 10 other stores. Discuss the accuracy and

where to improve. Note if the models heavily depend on the weather data, you may need to remove that feature from your data set for predictions of the remaining stores. Average US daily temperature could be a great substitute.

11. All team members should collaborate on all sections of this section regardless of their role in Section 1.

Extra Credits:

1. Create a dashboard to deploy the model (see below). **10 points**

Explainerdashboard: Link

It is a powerful dashboard for quick deployment of machine learning models. Please review the document in the link.

- The dashboard example works well for categorical variables and classification purposes with just a couple of lines of code.
- Extra credit will be given to the teams that deploy the ML model for the final project. It would be a great opportunity for you to showcase your coding abilities to employers in the future by showing the dashboard on your homepage.
- Feel free to incorporate any other dashboards that you are familiar with provided they are embedded in your codes and are open source. Please do not use proprietary codes and licensed templates that are not available for free to public.
- Discuss the important charts of the dashboard that help you understand the models better and make better judgments.

Coding and Report requirements.

1. Data should be saved on Google Drive and loaded to Google Colab.
2. Codes should be developed professionally with proper documentation including notes, assumptions and variable definitions for your teammates and others to easily understand and follow.
3. The report should be a Google Doc file prepared in collaboration between team members. Tell your story, provide insights, and organize your charts to support your thoughts.
4. Provide an executive summary at the beginning of each chapter and final conclusions at the end.

Final Deliverables on the deadline (Dec 2023 - TBD) in a zip file
1. Link to the code on Colab and exported python code
2. Link to the Google Doc report and the exported file in MS Word format (.docx)
3. PDF of the final Report
4. (Optional) Link to the deployed Dashboard, the code on Colab (if different from 1) and exported python code