**Documentation:**

In order to build a recommendation engine that suggests appropriate products to customers, I was given 7 datasets:

1. **orders** (information on all orders with order IDs, user IDs, order numbers, day of the week orders were placed, time of day orders were placed, and time between consecutive orders)
2. **products** (information on products with product IDs, product names, and IDs of aisles and departments the products were stocked in)
3. **aisles** (aisle IDs and aisle names)
4. **departments** (department IDs and department names)
5. **order products prior** (purchase history of customers with order ID, product ID, what position products were added into the cart, and which products in an order are reorders/purchased in the previous orders)
6. **order products train** (training set with same features as “order products prior” comprised on the most recent orders of a subset of users, new orders)
7. **order products test** (testing set with order ID and product ID, new orders)

I first explored these datasets by reading them into pandas DataFrames and printing data type and count of all features. My goal in this process was to ensure that all important features were in a base 64 data type, and to understand the number of rows per dataset and if any values were missing. From the orders csv file, I noticed that there were only 3,214,874 values for days since the last order in contrast to 3,421,083 values for the rest of the features, but this is due to Null values present for the first order of each user. I contemplated dropping these null values, but as this was still early in the exploratory process, I decided to leave them for now.

To get a holistic perspective, I approached the problem by first hypothesizing what factors could possibly influence customer shopping habits. I then explored any trends in the types of products bought, times of day/week/month when orders are placed, proportion of reorders, and specific products that are most consistently bought.

A picture containing screenshot

Description automatically generatedBy merging the products and prior orders datasets together, and calculating the total reorders for each product, I found that the top 10 most reordered items composed of fruits and vegetables. Specifically, most of these items were organic with bananas and organic bananas being the most popular.

Just to go more in depth, I also explored which products were frequently added into the cart first, and of these, how many were reorders.

A screenshot of a cell phone

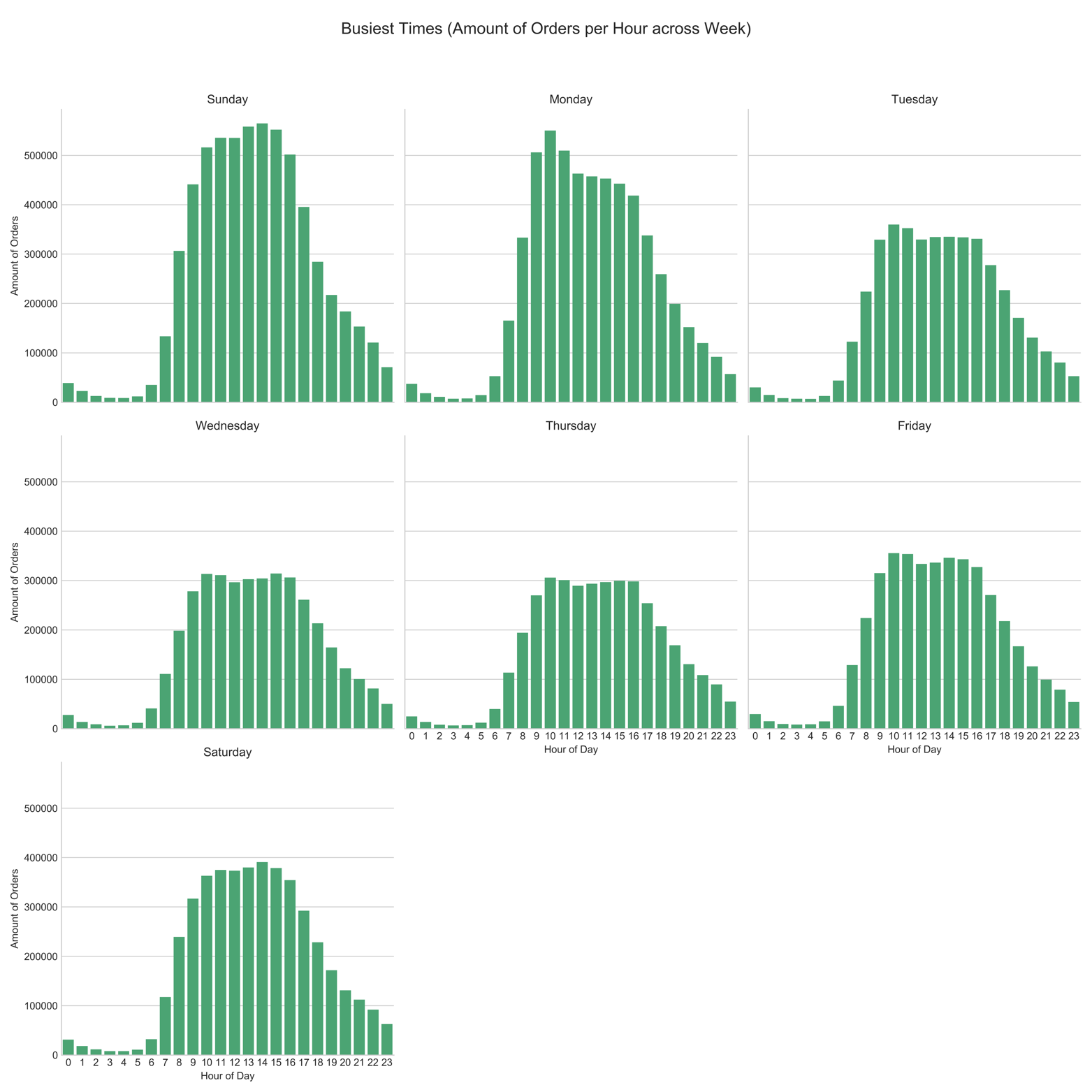
Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

By merging the orders and prior orders datasets, I then explored trends in purchase times. I found that the busiest times were on Sunday afternoon (1-3pm) and Monday morning (9-11am), with a significant decrease during the week.



To understand how many days between consecutive orders / when customers are buying products, I plotted the days since the last order against the amount of orders. Customers are placing orders frequently at the end of the week and at the end of the month.

A screenshot of a cell phone

Description automatically generated

Next, I dove into the reorder ratio in the prior order and train datasets to gauge if there is a similar distribution between the two. 58.97% of all orders in the prior dataset contain reordered products, and 59.89% of all orders in the training set contain reordered products.

Based on this exploratory analysis, I felt that important features to take into consideration when training this model would be order ID, days of the week that orders were placed, hours of the day that orders were placed, a products position in the cart (when products were added to the cart), and whether a product was reordered in the past.

I believed Random Forest Regressor would be an appropriate model to solve this problem as it is able to provide a good balance between precision and over-fitting.

I originally started fitting my model to the orders training set with the features order ID, position in cart, and reorder status, and test on the order test set. However, I realized that this would be difficult as the number of features in the orders train set (n=3) is different from the number of features in the orders test set (n=1).

I then focused on training on the prior orders dataset. I decided to split the prior orders dataset into test and train batches to train and test the model, since this dataset is composed of historical data and could provide for an accurate model. My next step was to then predict product IDs in the orders train dataset, to validate my model. After fitting the model on the training batch of the prior orders dataset, I printed the r squared score of about -0.276. This indicates that this model is very inaccurate at predicting product IDs.

**Conclusions:**

I chose Random Forest due to its ability to prevent overfitting and maintain precision. However, Random Forest Regressor proved to not be a good model to help suggest products to customers. With an r squared value of -0.276, the model is very inaccurate and is not able to find a good correlation between the features to predict products. I believe that tuning of the model is needed, and one possible way I can improve the r squared score for the model is to increase the number of trees. I could also reassess important features for the model’s inputs. A few other analyses to help reassess my model would be to explore the relationships between product ID and its position in the cart and between user ID, product ID, and position in the cart. If these do not work, then a different model would be necessary. Regardless, to implement this model, it would be necessary to continue tracking customer purchase activity and monitor times (day of the week and hour of the day) that customers shop.

It is still possible to translate insights from the exploratory process into suggestions. Organic food is frequently purchased, and most customers shop during Sunday afternoon and Monday morning, and at the end of the week and end of the month. As a starting point, organic food would be a good product suggestion to give to all customers.

**Resources:**

I used a few resources and references while working on this, namely the Titanic Kaggle competition to get a “jumpstart” on the data science process, and documentation on seaborn and sk learn libraries.