# Data Glacier Group Project: Retail Forecasting

**Modeling Report**

## Team Member Details:

Kierra Dangerfield
kierradachelle@yahoo.com
United States of America
Freelance
Specialization: Data Science

## Problem Description:

The large company which is into beverages business in Australia. They sell their products through various super-markets and also engage into heavy promotions throughout the year. Their demand is also influenced by various factors like holiday, seasonality. They needed a forecast of each of the products at item level every week in weekly buckets.

## Github Repo Link:

https://github.com/KierraDangerfield/Data-Glacier/tree/main/Week_13

## Data Overview

**Product:** Product id
**date:** date of sales
**Sales** **:** sales amount
**Price Discount (%):** percentage of discount applied to sales
**In-Store Promo:** whether or not if there was in store promotion
**Catalogue Promo:** whether or not if there was catalog promotion
**Store End Promo:** whether or not if there was an end date to a promotional period
**Google_Mobility:** anonymized measurement of people's movements surrounding the products
**Covid_Flag:** whether or not if there was Covid to worry about on that day > during the lockdown or not
**V_DAY:** whether or not if it was valentine's Day
**EASTER:** whether or not if it was Easter
**CHRISTMAS:** whether or not if it was Christmas > no = 0, yes = 1

## Data Cleaning & Transformation

The dataset only has 1218 entries. There are no null values or duplicate values.

- I transformed the date column from object to datetime.
- I changed the “Price Discount (%)” column from string to float and renamed the column to “Discount”

I added 3 columns:
**DayOfWeek:** which day the date was on
**Quarter:** Quarter of the year the date of sales is in
**WeekNum:** The week number of the year
**Year**: year of date
**Month**: month of the date

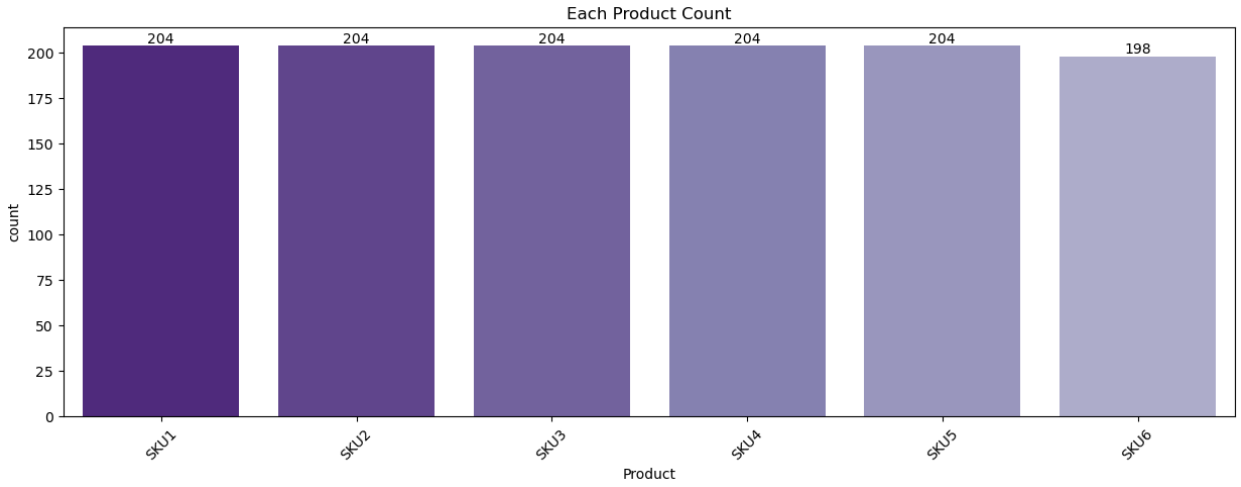
There are outliers in the sales column. Because I am doing a forecasting model, I did not remove the outliers. Most of the

outliers come from quarter 3 and 4. If the outliers are affecting the model, I might have to remove them from the dataset or replace the values with mean, median, or mode values.

I am looking at experimenting with Linear Regression, tree-based models, and other models as well as some type of regularization/standardization.

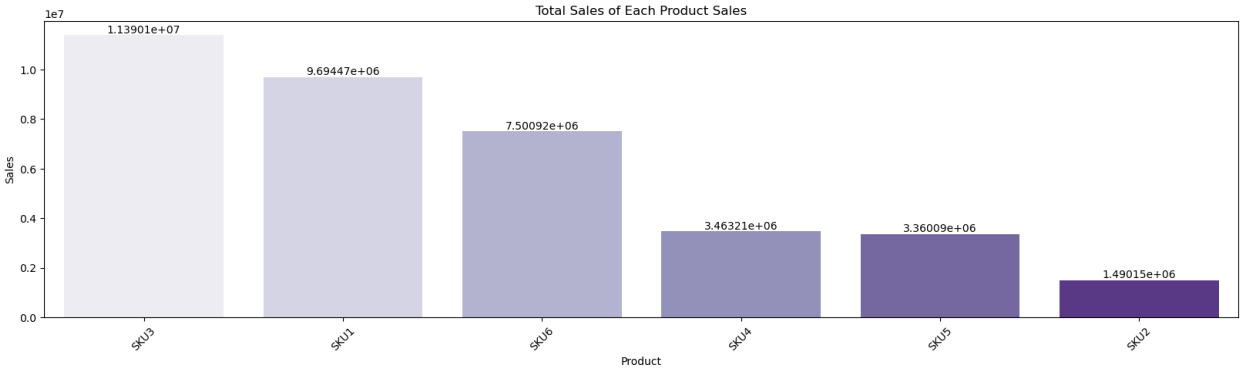
## EDA

### 1. Each Product Count



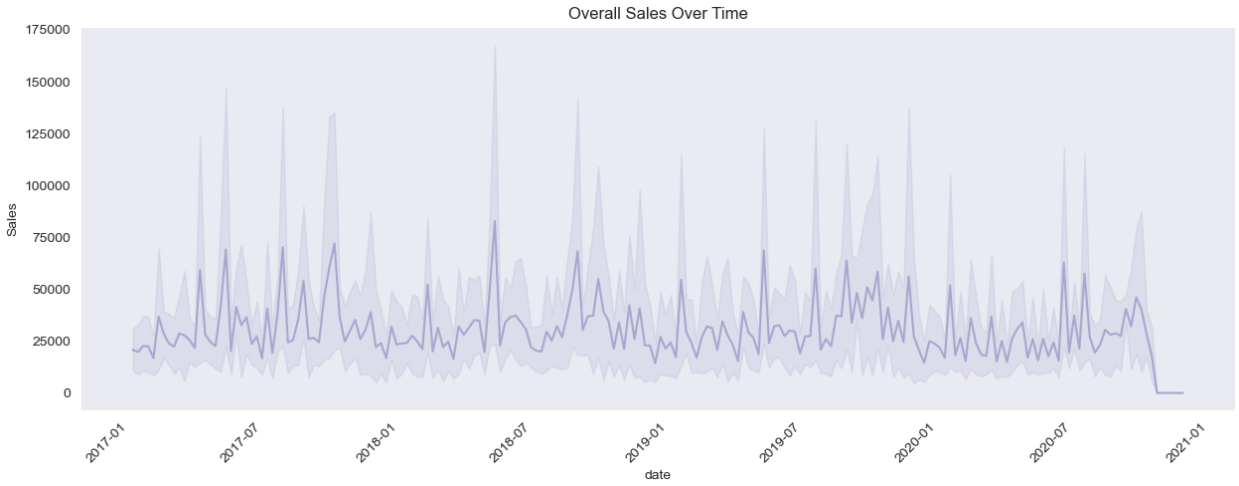

Product 6 has slightly less amount of products sold, but it is still around the same amount of beverages sold.

## 2.Total Sales of Each Product



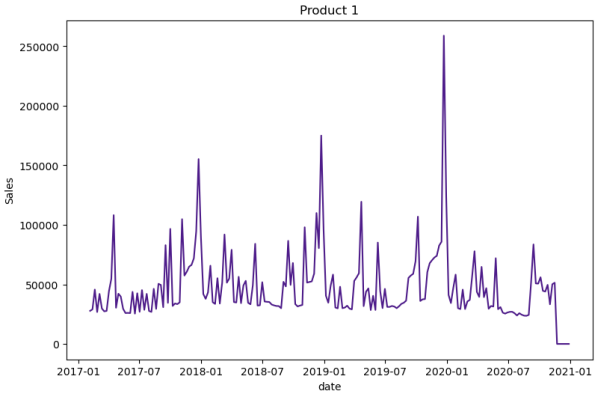

Product 3 has the most amount in total sales, and product 2 has the least amount in total sales. Product 3 could have a higher price compared to the others.

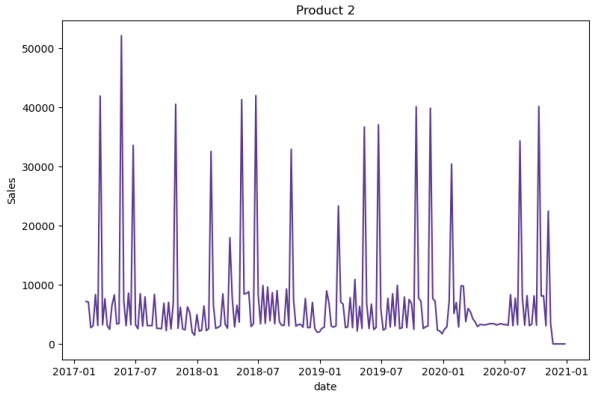
## 3.Overall Sales Over Time

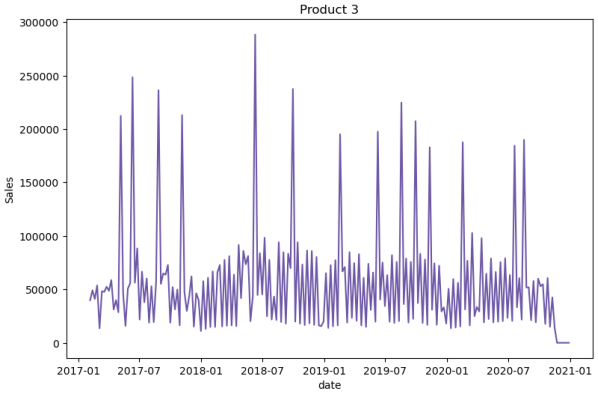


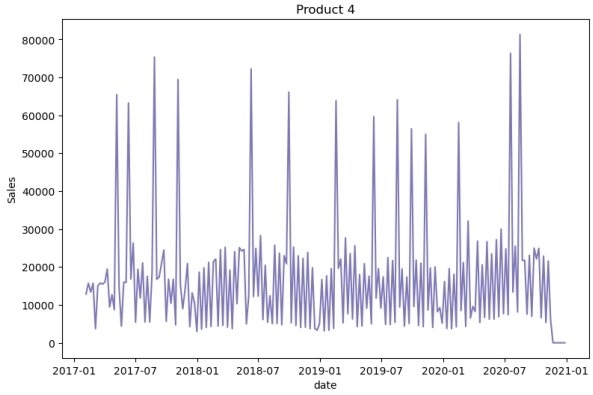

The chart above shows the overall sales from January 2017 to January 2021. You can see the highpoints in some areas.

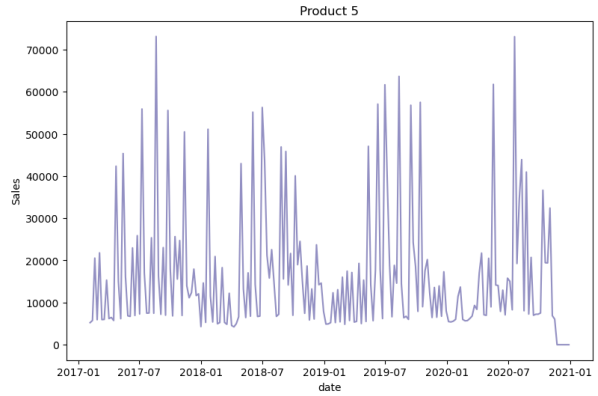
# 4. Sales Based Each Product

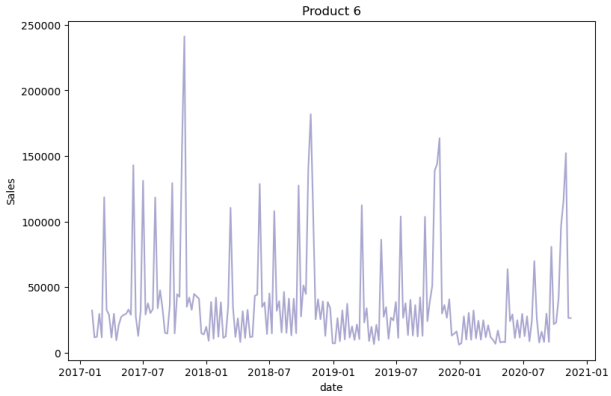


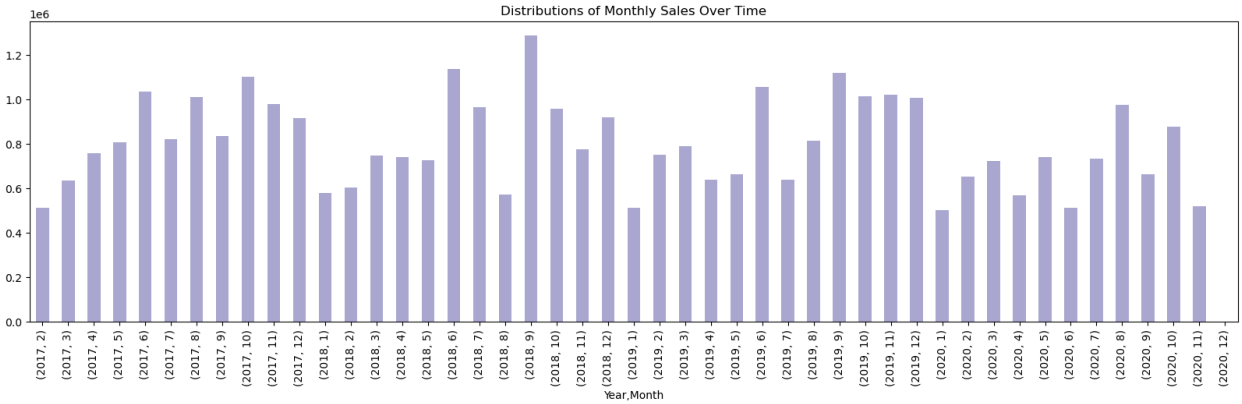




The above line plots show the sales over time for each product. For each product we can see when the product has high points and low points of sales.

## 5. Distributions of Monthly Sales Over Time

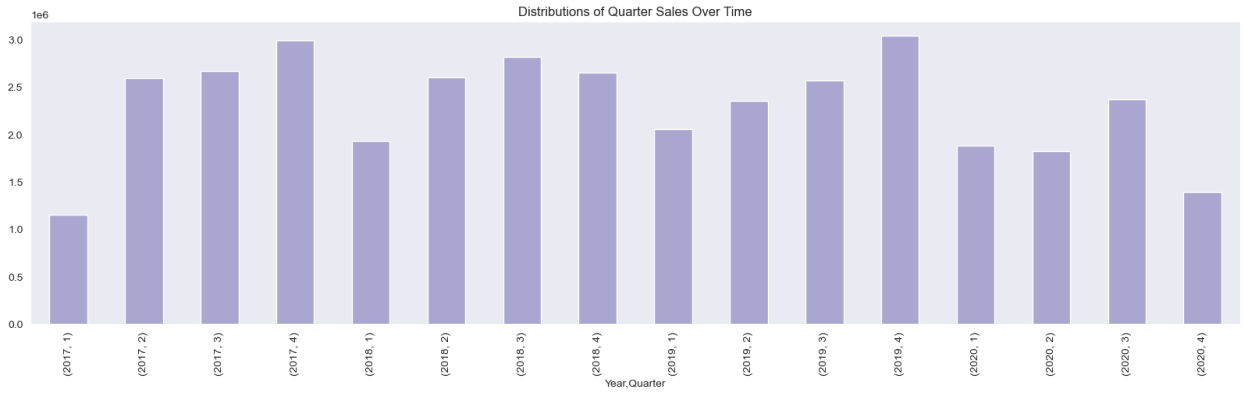



The above chart shows the distribution of sales over time. Overtime, the later months of the year typically have higher sales.

## 6. Distributions of Monthly Sales Based On Product

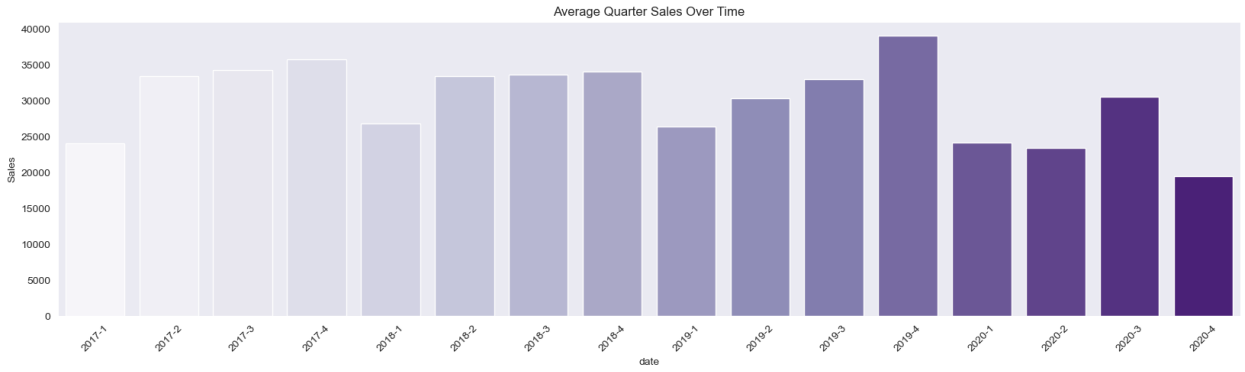

The above shows the distributions of monthly sales based on each product. Based on the error bars, there is a lot of uncertainty, especially with product 1 in month 12.

## 7. Distributions of Quarter Sales Over Time



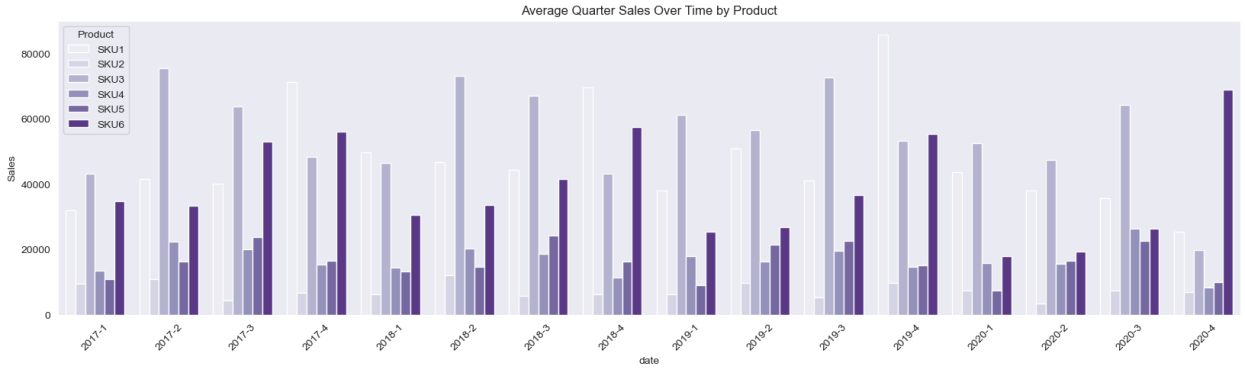

The first quarter had the lowest sales of the year except for quarter 1 in 2020. In 2019 there was a steady, linear increase.

## 8. Average Quarter Sales Over Time




The above shows the average quarter sales. It seems that 2020 has been a slower year for the company. In 2019 they had consistent growth with each quarter.

## 9. Average Quarter Sales Over Time by Product

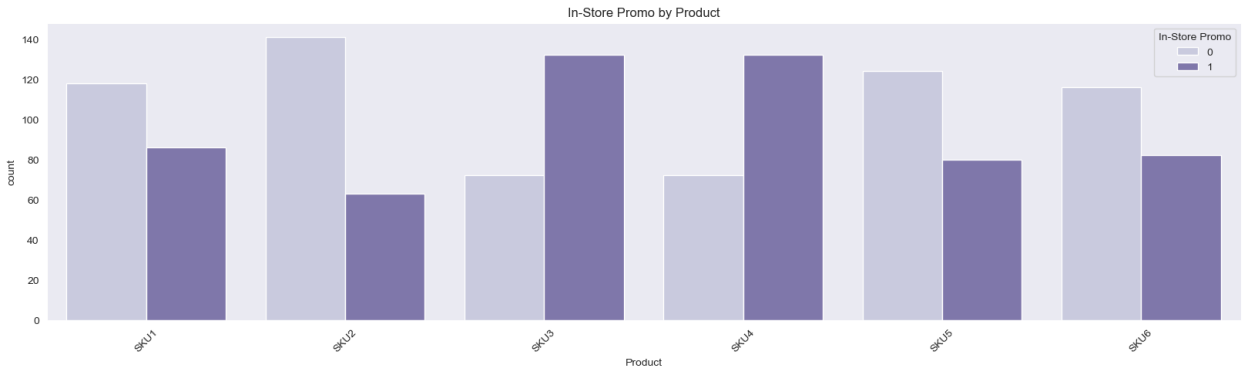



The above chart shows the average quarter sales over time by product. Product 2 has the lowest sales. The above chart shows the average quarter sales over time by product. There is a significant difference between Product 1 in 2019-4 compared to 2020-4. The average monthly sales for product 1 has consistently gone down since 2019-4.

## 10. Distributions of Discounts by Product

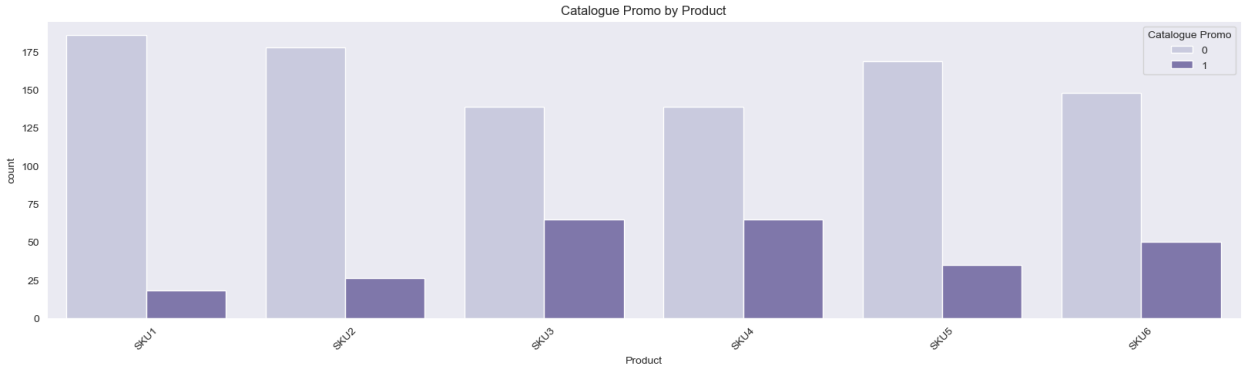

The chart above shows the distributions of discounts for each product. Of the discounts provided from the company, Product 3 and 4 have some of the higher distributions of discounts. Remember that Product 3 had the most amount of total sales over the years.

## 11. In-Store Promo by Product



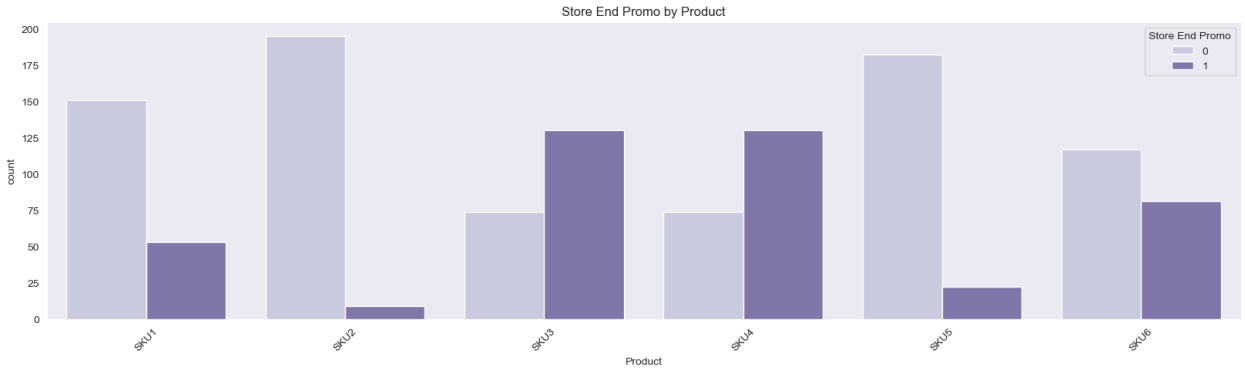

The chart above shows if there was in-store promo or not for each product at the time of sale.Majority of product 3 and 4 sales had in-store promo. It seems customers respond well to in-store promo.

## 12. Catalog Promo by Product



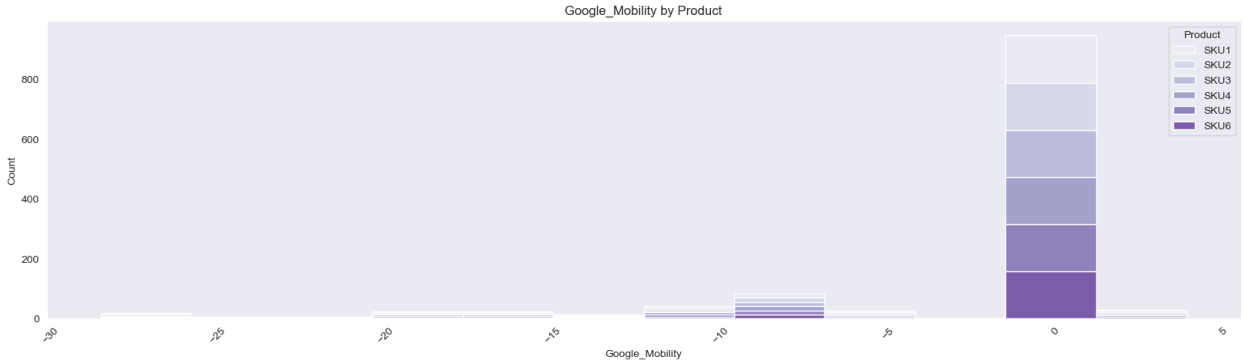

The chart above shows if there was catalog promo or not for each product at the time of sale. Majority of the sales did not have catalog promo. However, of the sales that had catalog promo, product 3 and 4 sales were the highest amount. It seems customers respond pretty well to catalog promo.

## 13. Store End Promo by Product



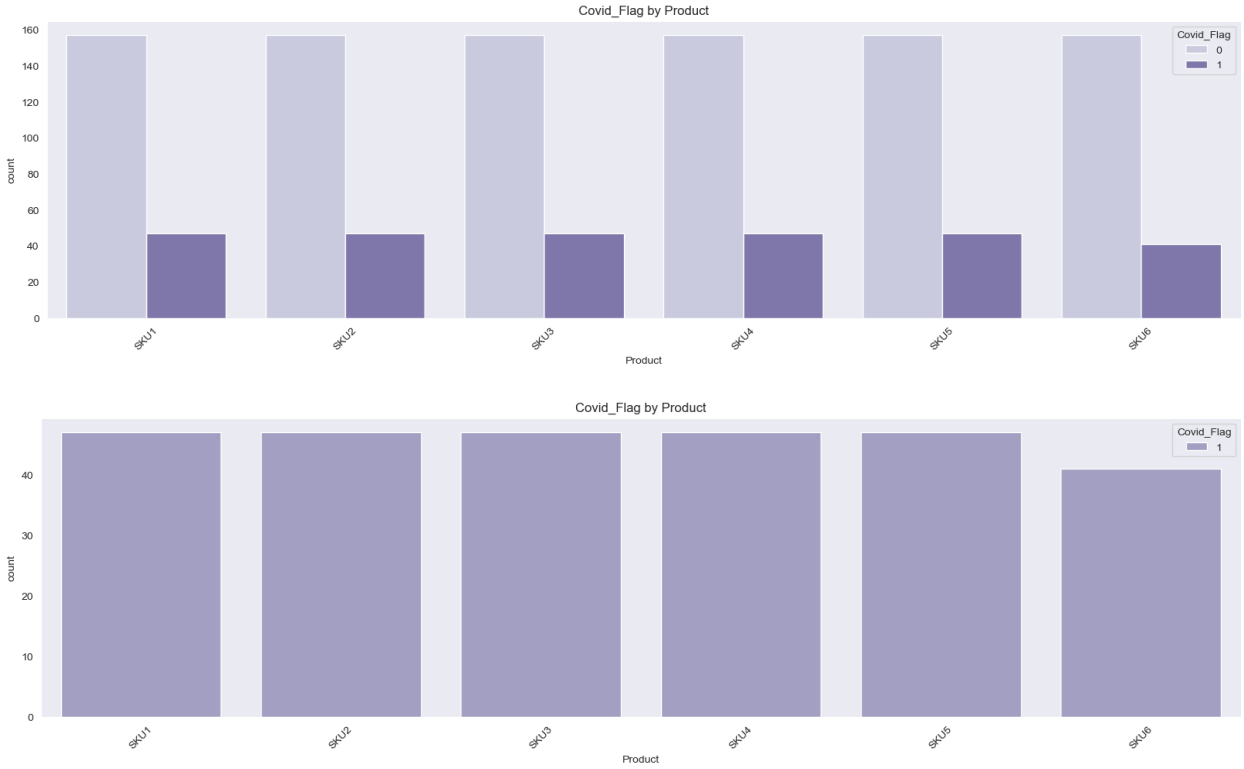

The chart above shows if there was a store end promo or not on the sales date. Majority of the sales did not have store end promo except for product 3 and 4. It seems customers respond pretty well to store end promo. Even with store end promo customers still did not respond to buy product 2.

## 14. Google_Mobility by Product



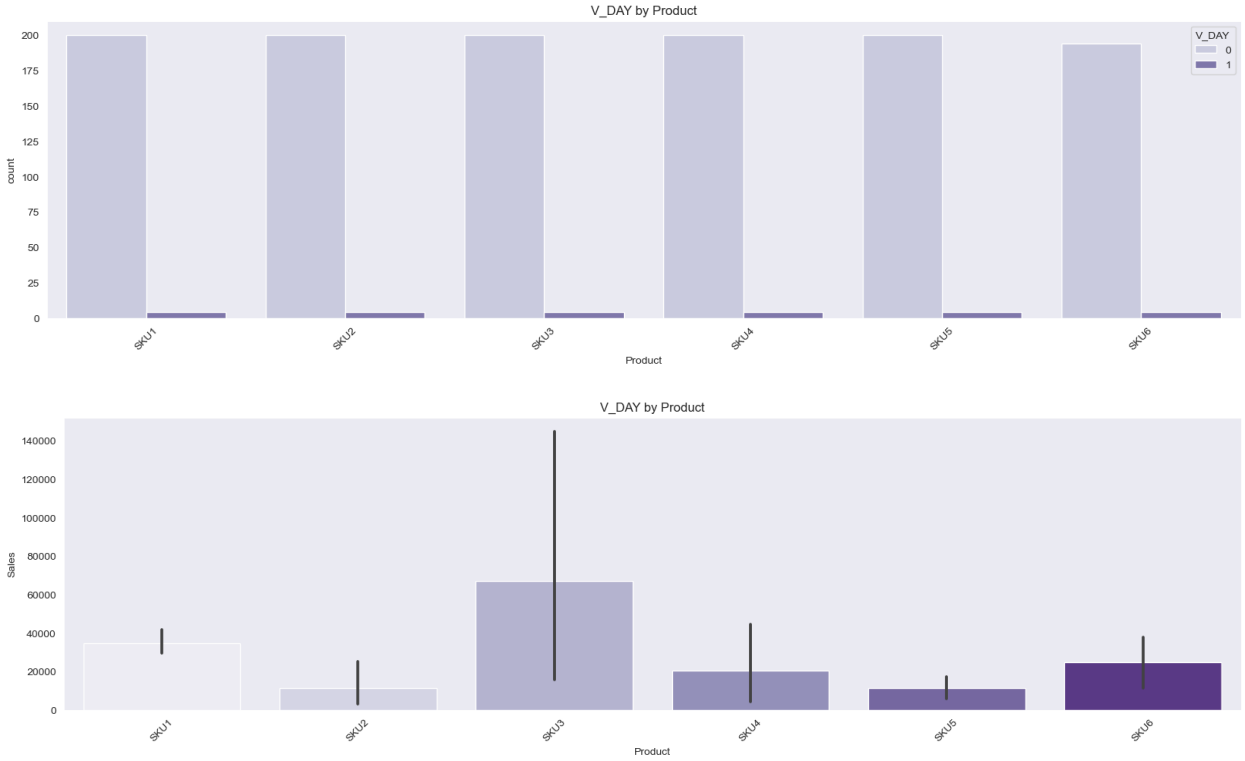

The above chart shows the google mobility of each product. Most of the product mobility is around 0.

## 15. Covid_Flag



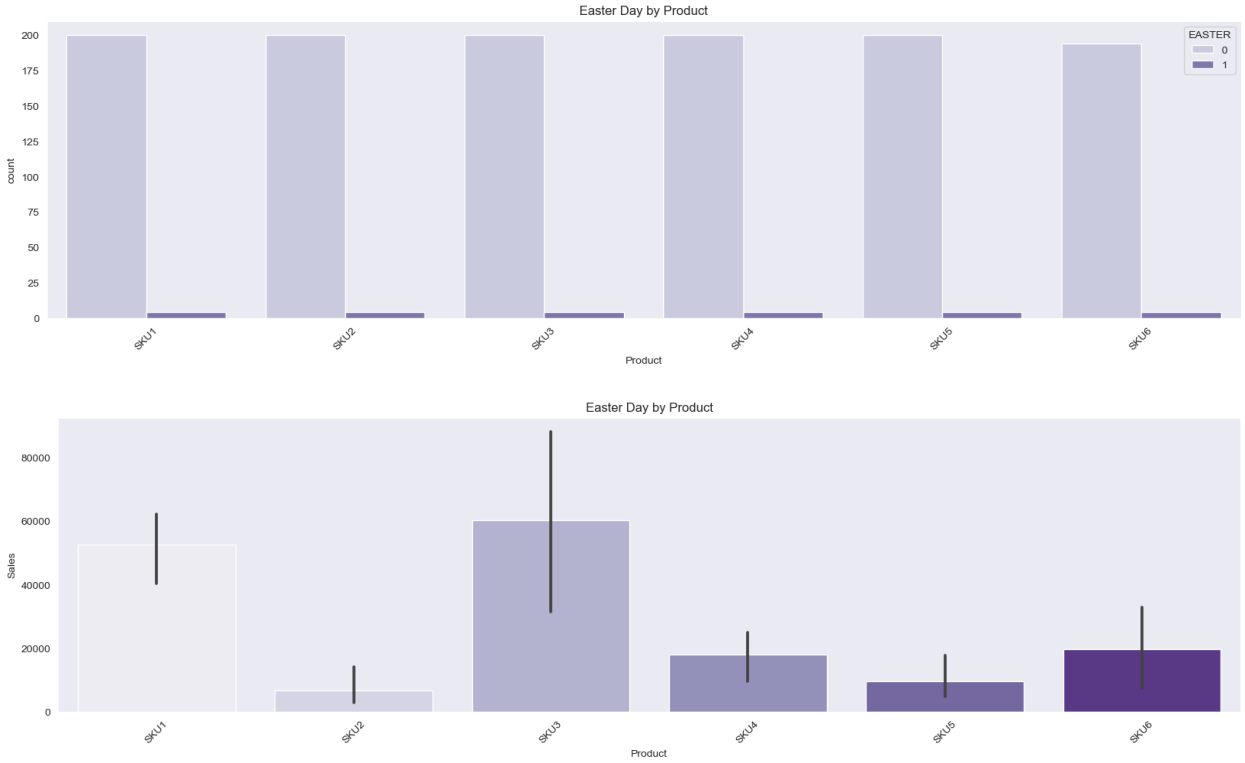

The chart above shows if the sales date had a covid alert. Majority of the sales did not have a covid alert at the time. Product 6 sales had a little bit less covid alerts compared to the others.

## 16. V Day



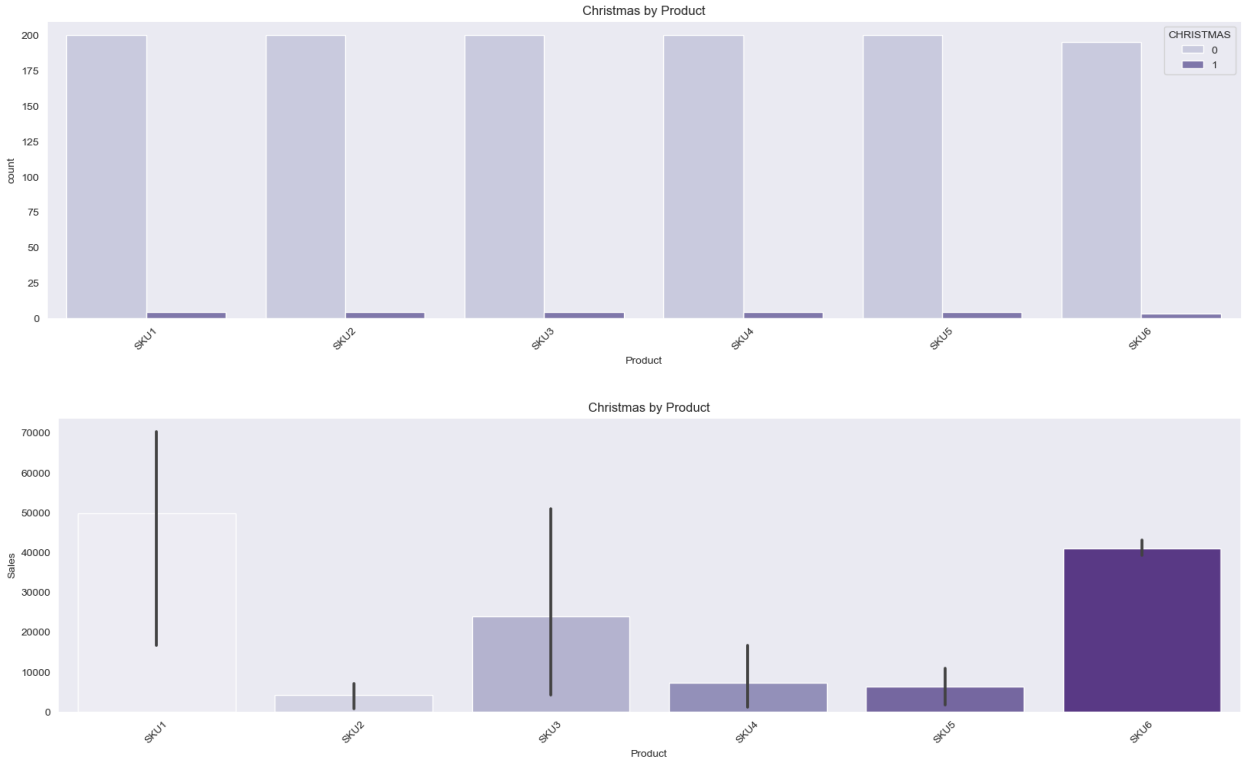

The above graph shows whether or not the sales date was on Valentine’s Day. Majority of the dates weren’t Valentine’s Day. However, when it was Valentine's Day, product 3 had the most sales count.

## 17. Easter




The above graph shows whether or not the sales date was on Easter. Majority of the dates weren't on Easter. However, when it was Easter, product 3 had the most sales count. Product 1 had the second highest amount of purchases.

## 18. Christmas




The above graph shows whether or not the sales date was on Christmas. Majority of the dates weren't on Christmas. However, when it was Christmas, product 1 had the most sales count. Product 6 had the second highest amount of purchases.

## 19. Statsmodel For Each Product

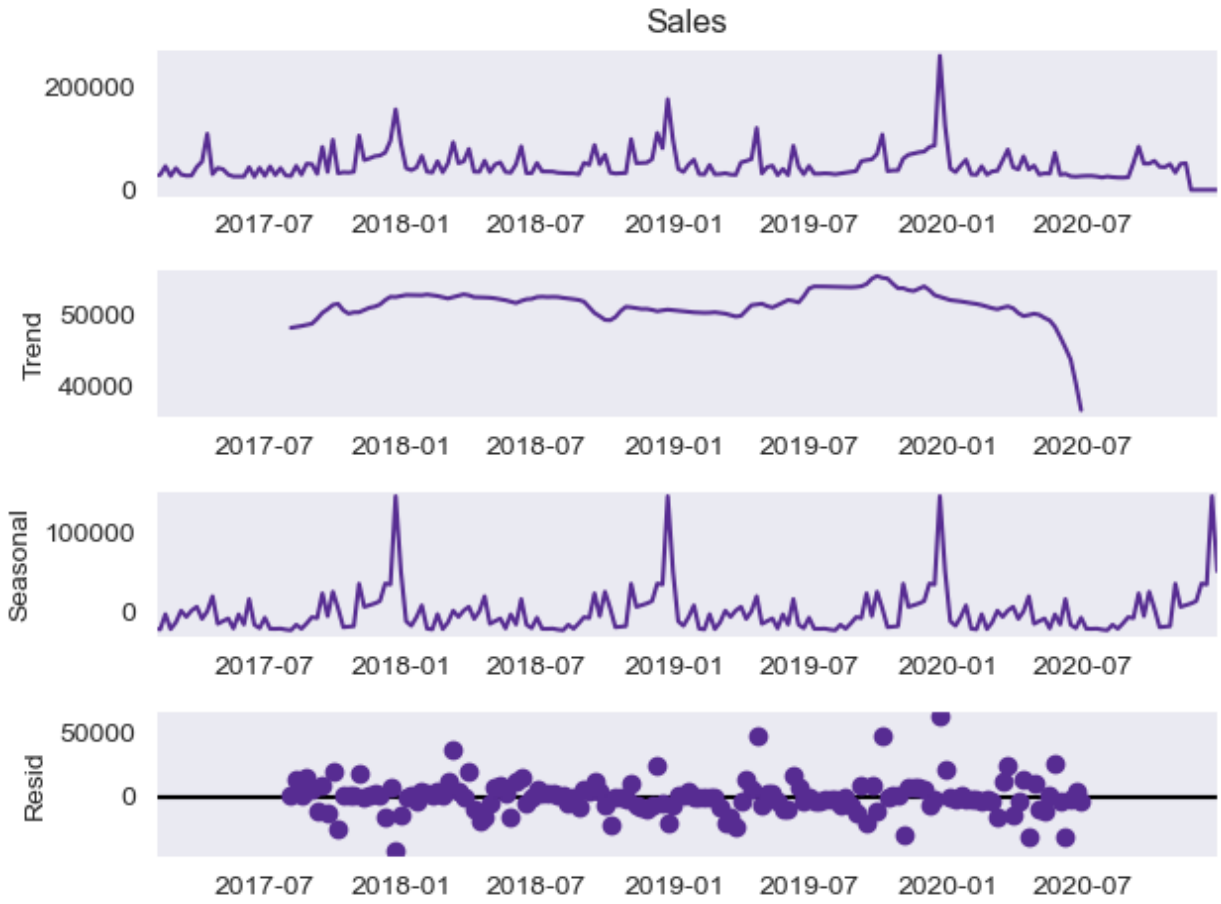
Time series decomposition is a technique that splits a time series into several components, each representing an underlying pattern category, trend, seasonality, and noise.

1. **Level:** describes the average values in time series data.
2. **Trend:** describes the increasing or decreasing trend in data.
3. **Seasonality:** The repeating cycles in data, could be monthly or weekly, etc depending on the granular level of data

4. **Noise:** The random variations in the time series data.

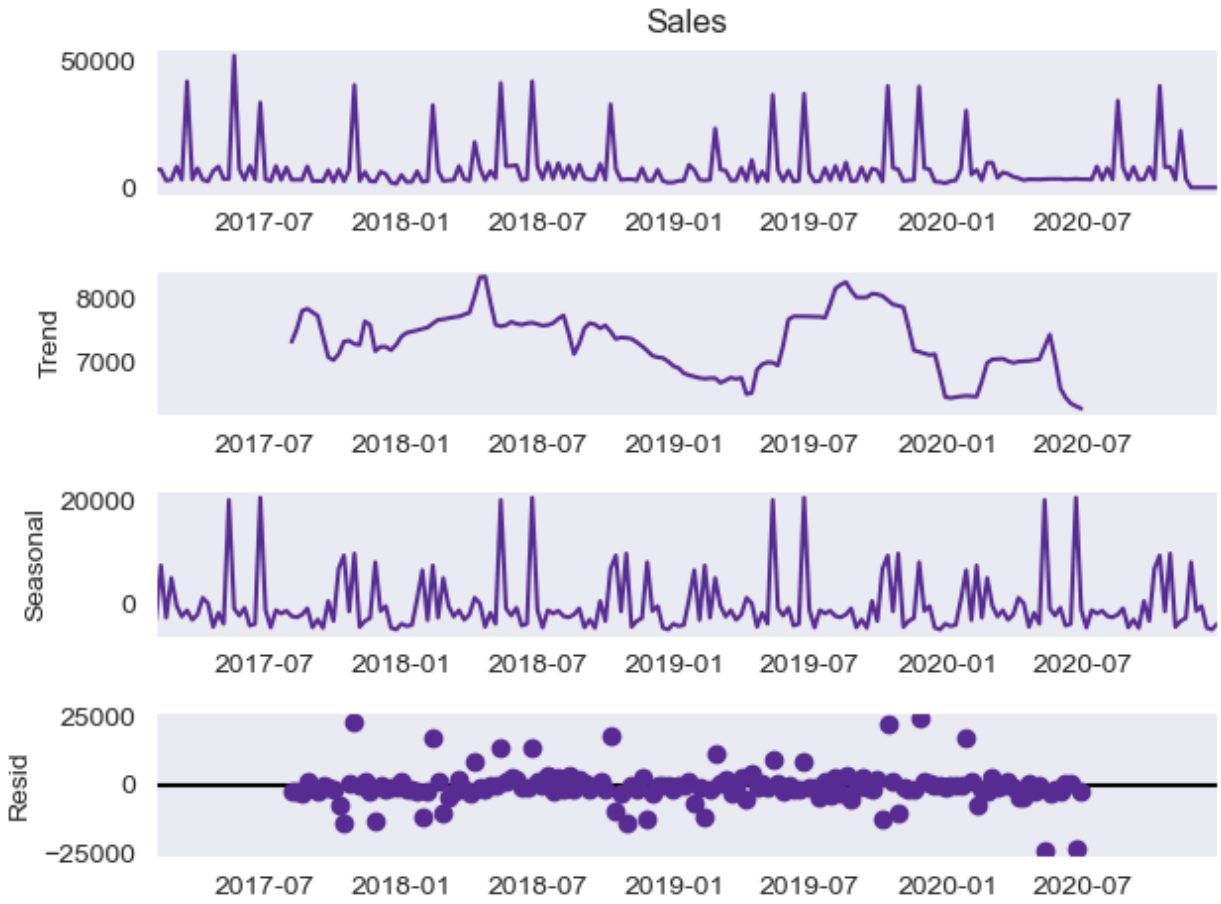
I used an additive model.

## a. Product 1



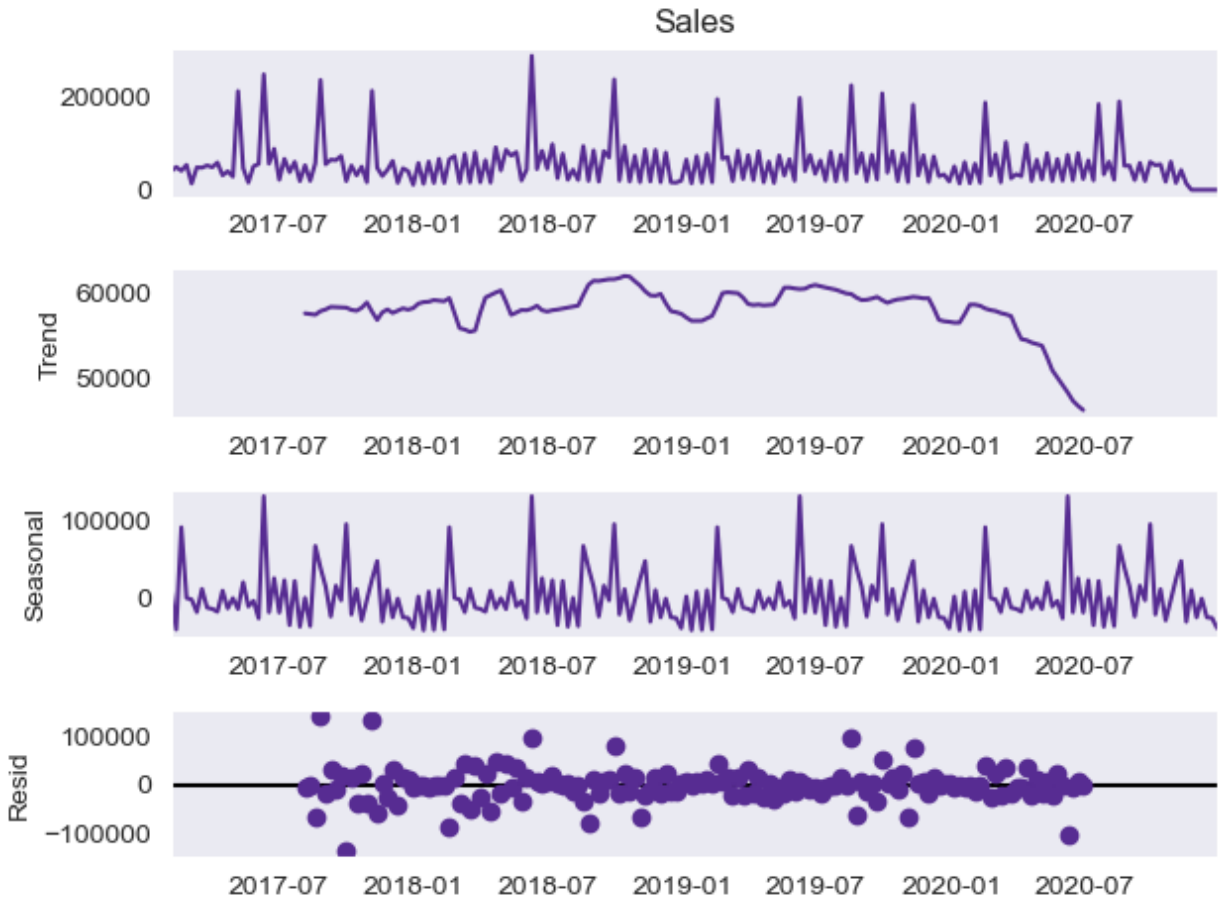

About January 2020 was the highest amount of average sales. I can see a trend near the beginning of the year. After 2020, it went down.

## b. Product 2



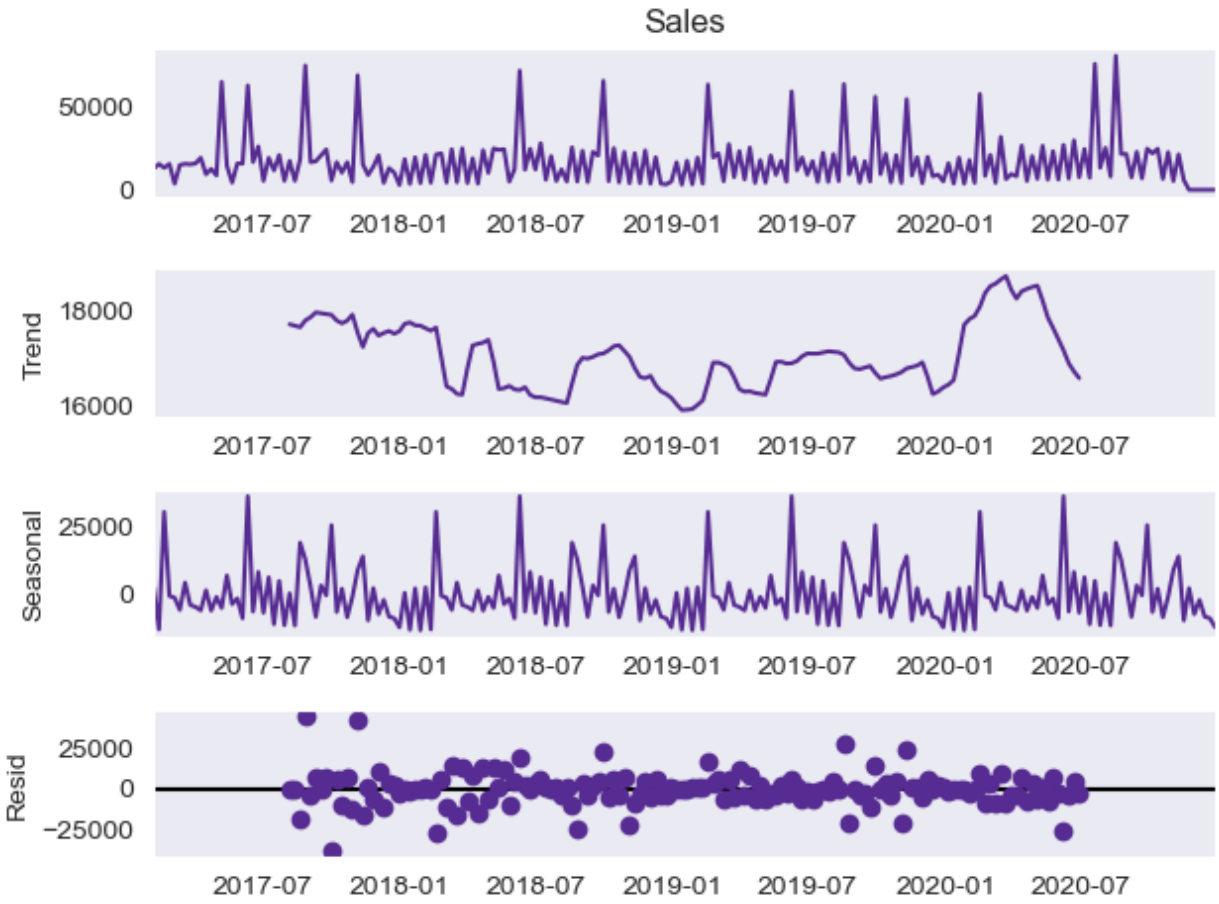

There is a consistent dip between 2018-07 and January 2019, rose again and then fell again around January 2020.

## c. Product 3



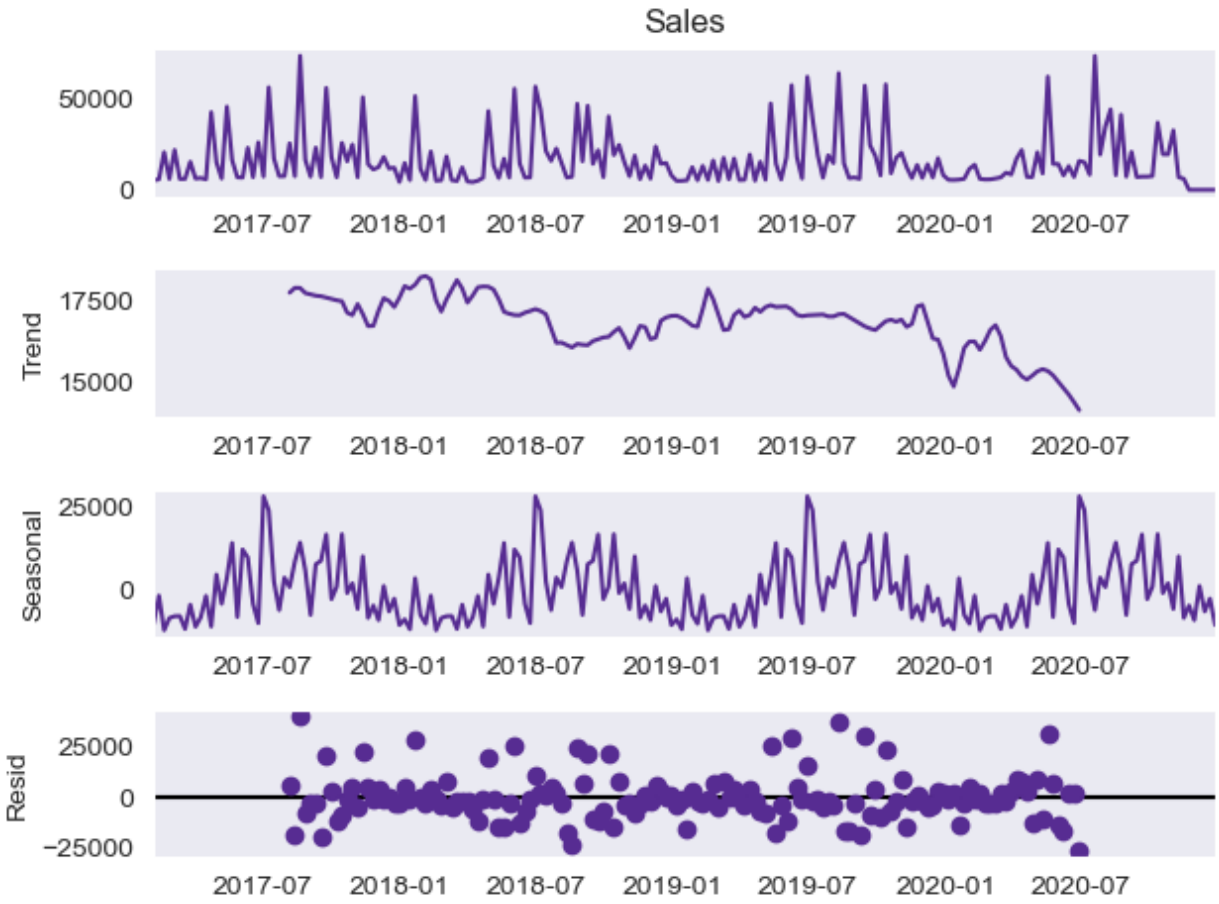

Product 3's trend has started to go downward around January 2020. Before it had minimal dips until around 2020.

## d. Product 4



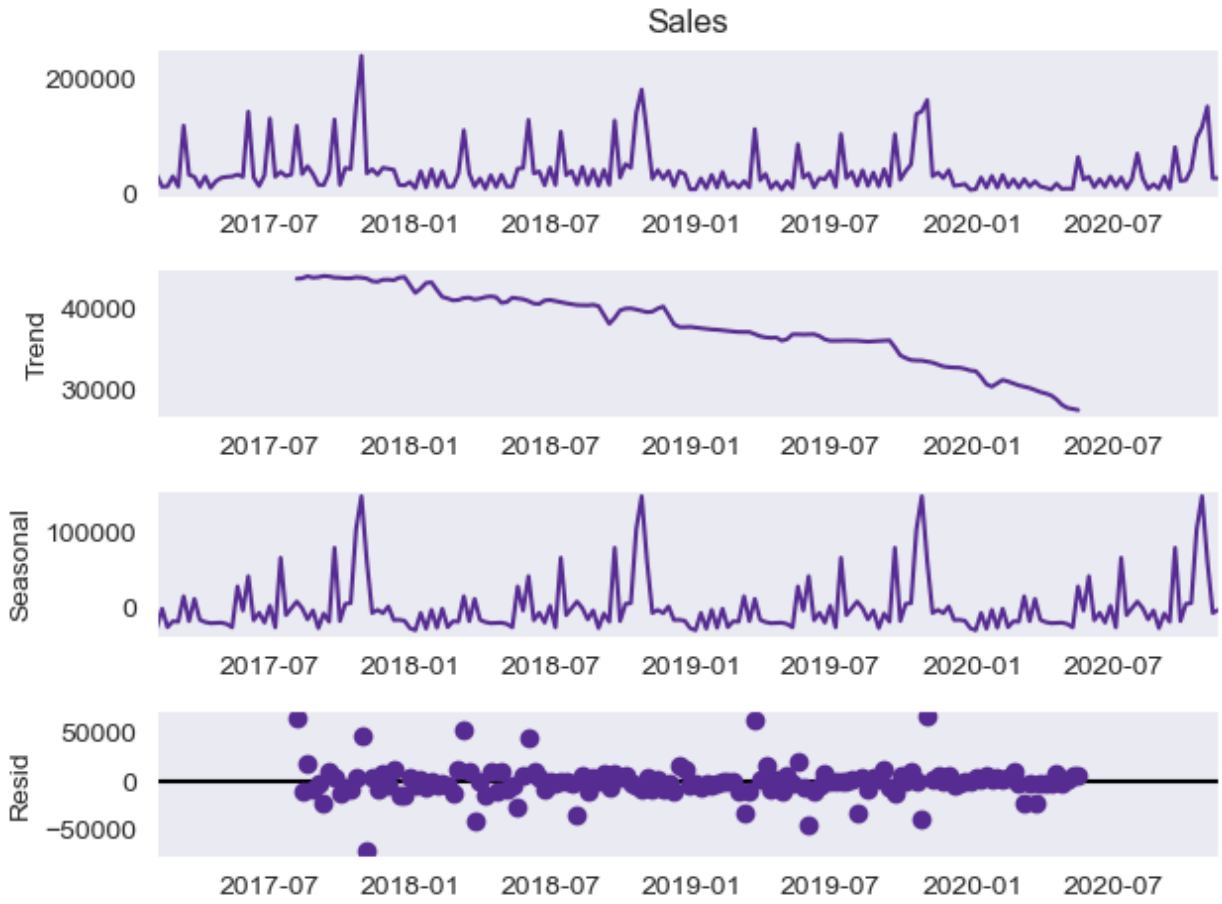

Around January 2018 it started to turn downward and spiked in 2020 and started to take a sharp decrease.

## e. Product 5




Product 5 seems to have the most visible errors. Based on its trend, it has gone down.

## f. Product 6




Based on its trend, it has consistently gone down.

# Modeling

## Overview

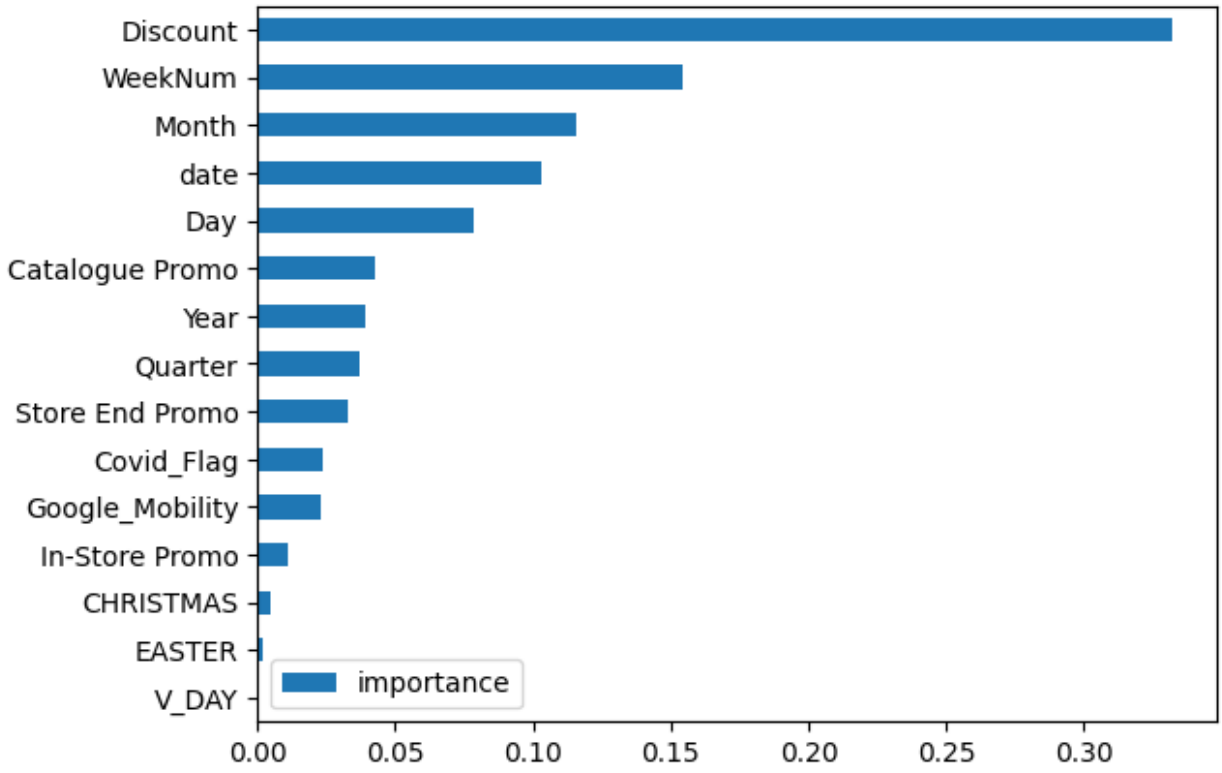
I trained 5 different models to see which one was performing the best based on mean absolute error. I used linear regression as my based model. Then I tested Random Forest, ARIMA, SARIMA, and SimpleRNN. For the model selection I only tested for product 1.

| Model | Mean Absolute Error |
|---|---|
| Linear | $12,602.134 |
| **Random Forest** | **$7,520.47** |
| ARIMA | $26,532.32 |
| SARIMA | $26,299.80 |
| SimpleRNN | $13,856.30 |

Based on the Mean Absolute Error, the Random Forest Model performed the best. Below is the Mean Absolute Error for each product.

| Product | Mean Absolute Error |
|---|---|
| SKU1 | $7,256.62 |
| **SKU2** | **$1,067.46** |
| **SKU3** | **$13,994.45** |
| SKU4 | $4,536.77 |
| SKU5 | $1,588.22 |
| SKU6 | $7,980.13 |

Product 2 performed the best using Random Forest, and Product 3 performed the worst. Below is a chart that shows the feature importance for Product 1. Customers respond well to discounts.




## Key Findings

- Discount and WeekNum are the two most important features.
- The Random Forest model performed the best with a Mean Absolute error of 7,616.85 for product1.
- Product 2 and 5 performed the best with the random forest model. Product 2 had the least amount of total sales, so I wonder if that is a contributing factor.
- Product 3 had the worst performance with random forest. Product 3 also had the highest amount of total sales. I wonder if that is a contributing factor.

## Final Recommendations

The ARIMA model would be great to use to predict upcoming weeks; however, it is not the best performing model.

I would suggest getting more data for each product to help build a better model. Instead of getting sales weekly, maybe collect daily sales in order to get more data.