# OVERVIEW

01 **Business Problems**

02 **Data Cleaning**

03 **EDA**

04 **Modeling**

05 **Conclusion**

06 **Recommendations**

# Business Problems

- Help PUMA predict future sales
- How to do inventory management in advance to avoid both inventory storage and overstocking

# Data Cleaning

- Merged the two datasets: “Activity” and “Product”
- Removed some useless columns, including some descriptive columns
- Converted the time variables (Year and Week) into the standard datetime variable
- Converted the missing values to 0 in all variables involving order quantity and order value

# Data Cleaning

**The Cleaned dataset has 2,064,371 rows and 10 effective variables**

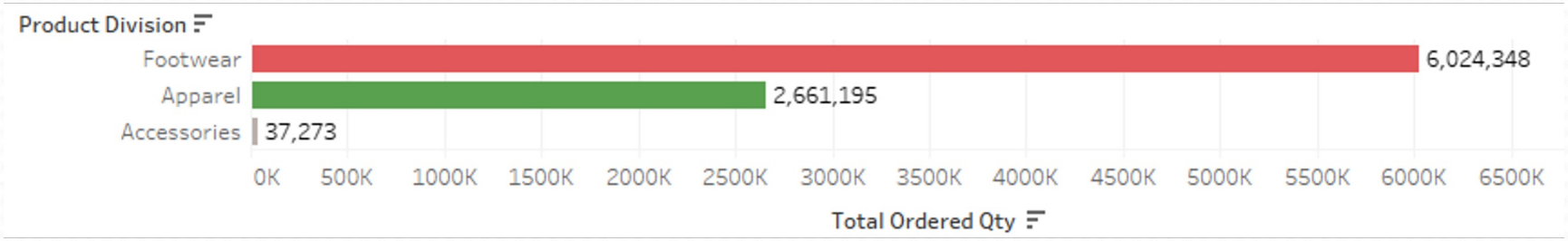
**Target variable:**

Total Ordered Qty = On hand Qty + Canceled Qty + Shipped Qty

**Effective variables:**

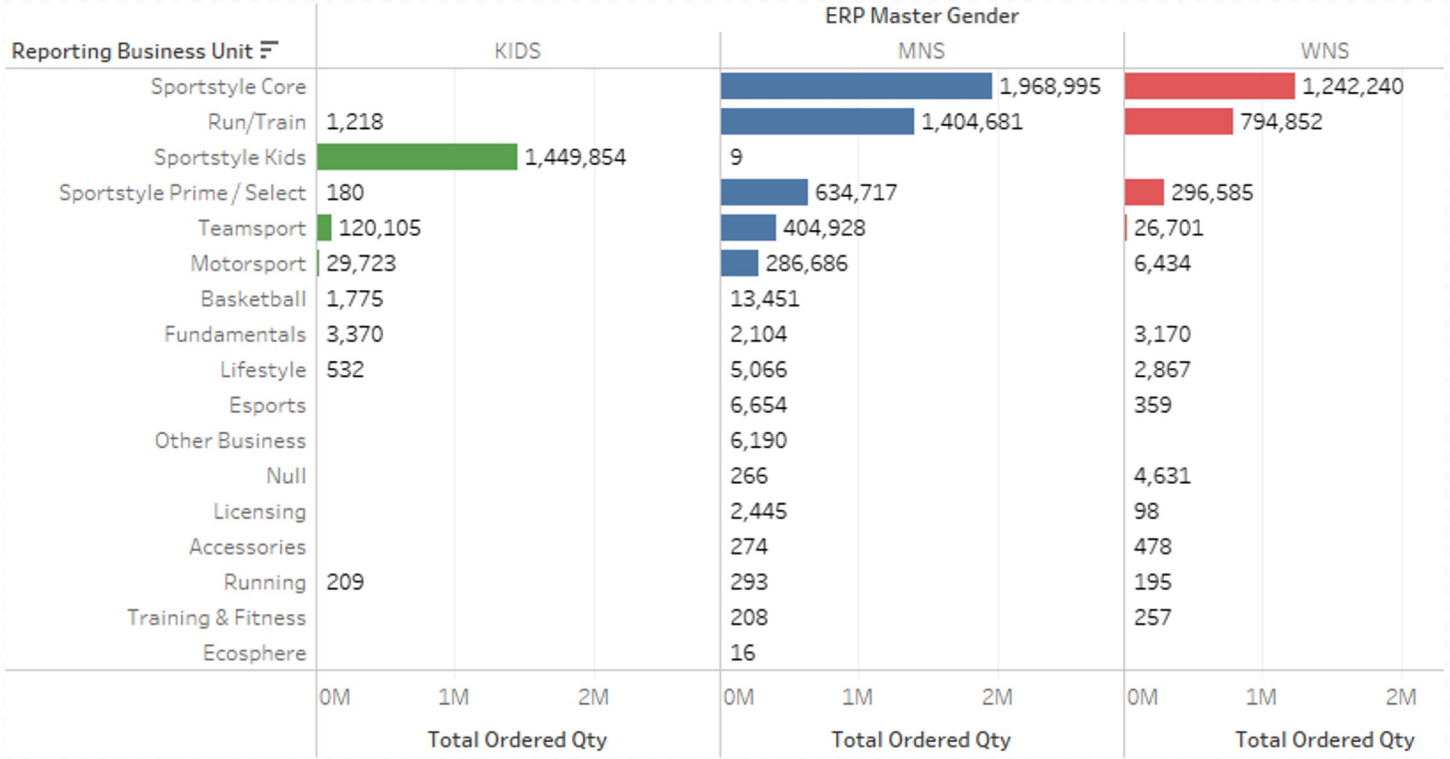
Date (converted from Year and Week)
Product Division
Master Gender
Reporting Business Unit
Main Color Group
US Size
Style Family
Season Opened
Style Num (6 identifer for the product)

# EDA (Product Division)

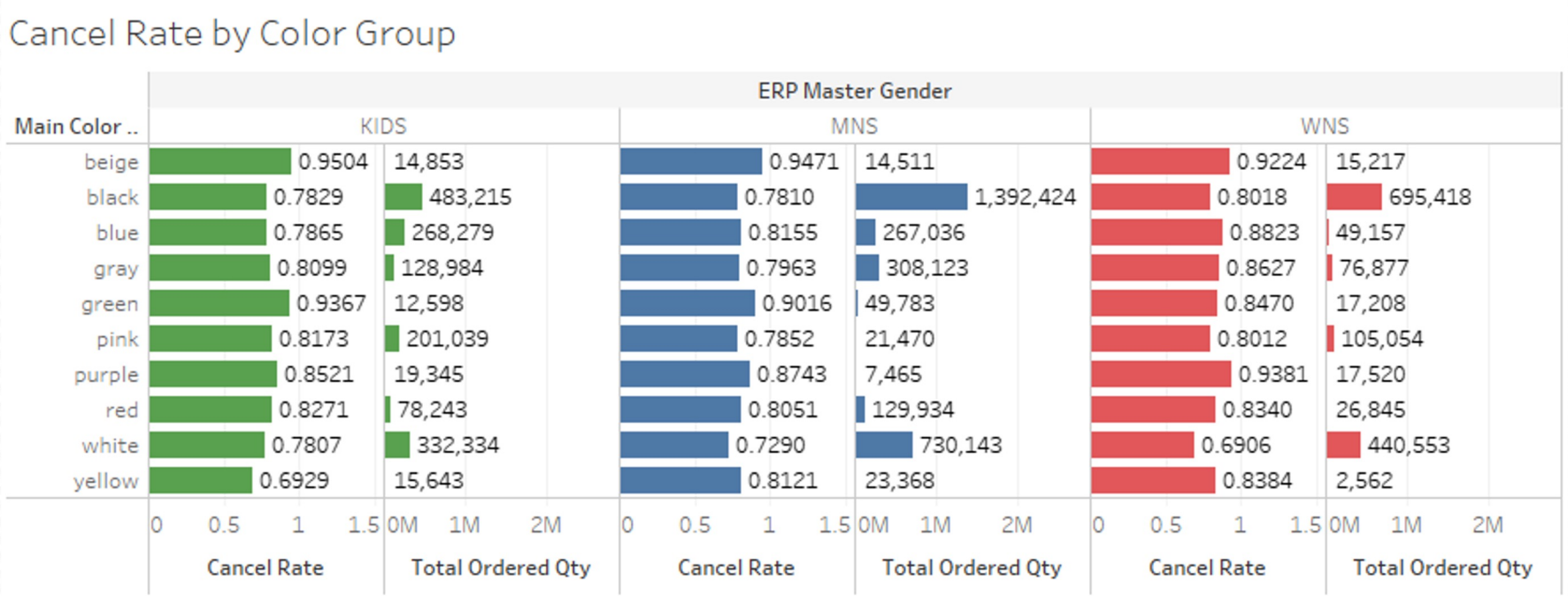



# EDA (Reporting Business Unit for each Gender Group)

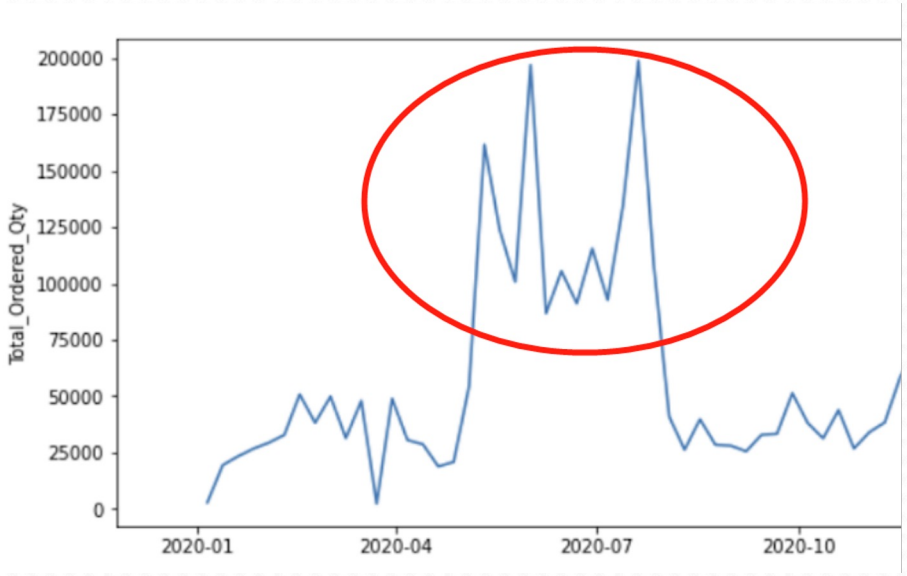
- There is a clear imbalance in distribution (provided for specific groups of people).
- In some specific areas such as Motorsport, PUMA also maintains a high degree of competitiveness.



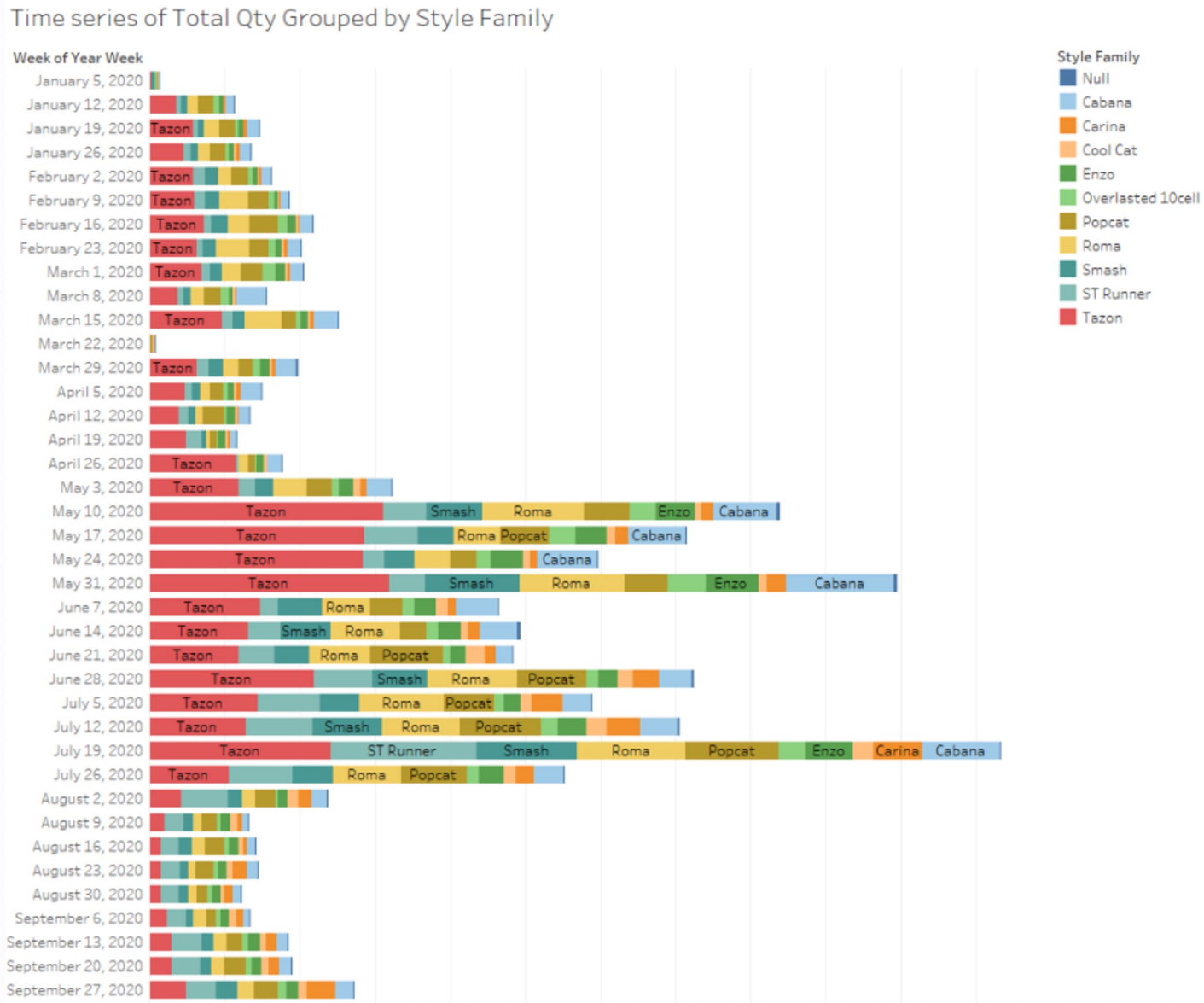
# EDA (Cancel Rate)



# EDA (Top 10 Style Families in Time Series of Total Qty)

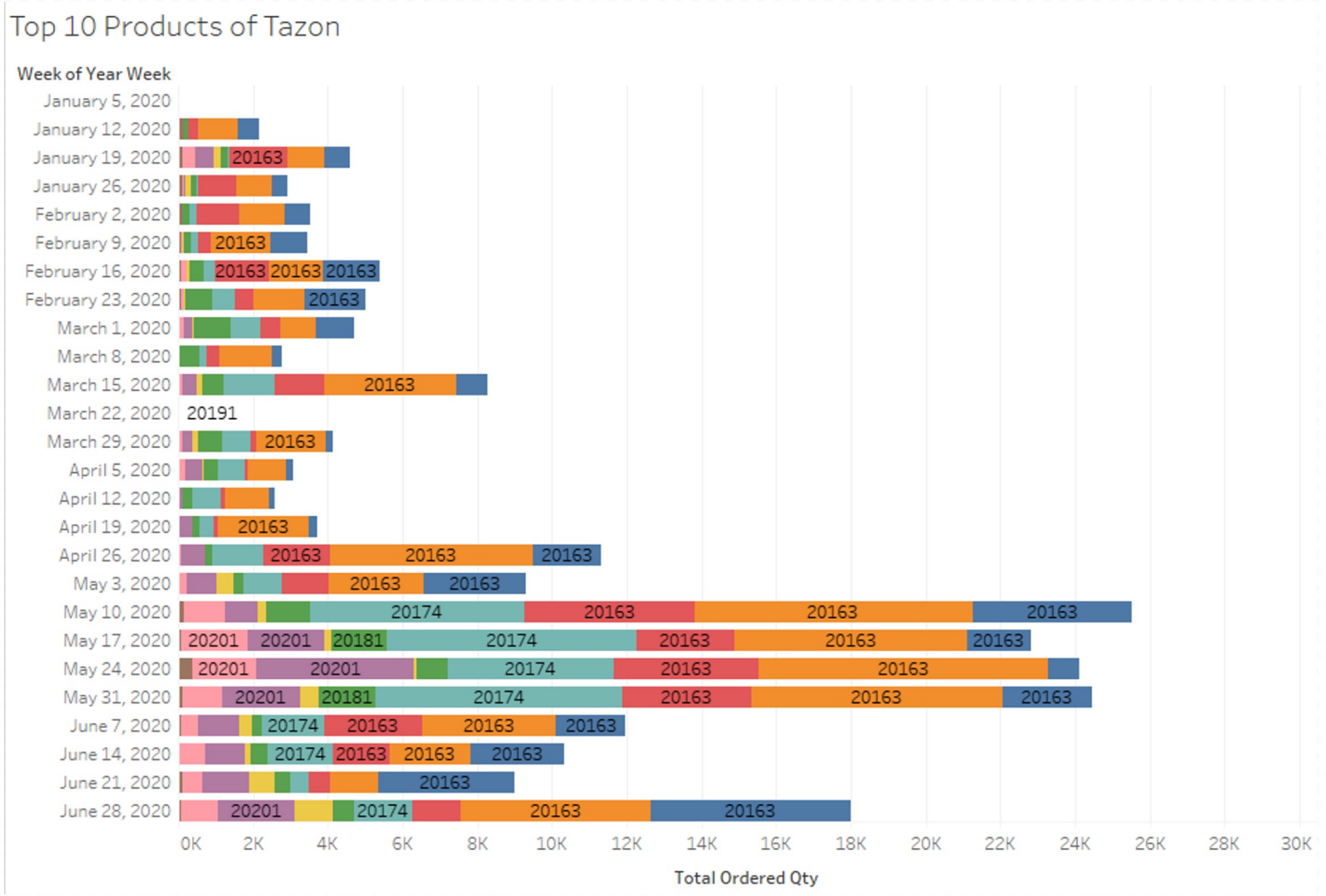



- We want to know if it was because of a few popular products that drove sales.

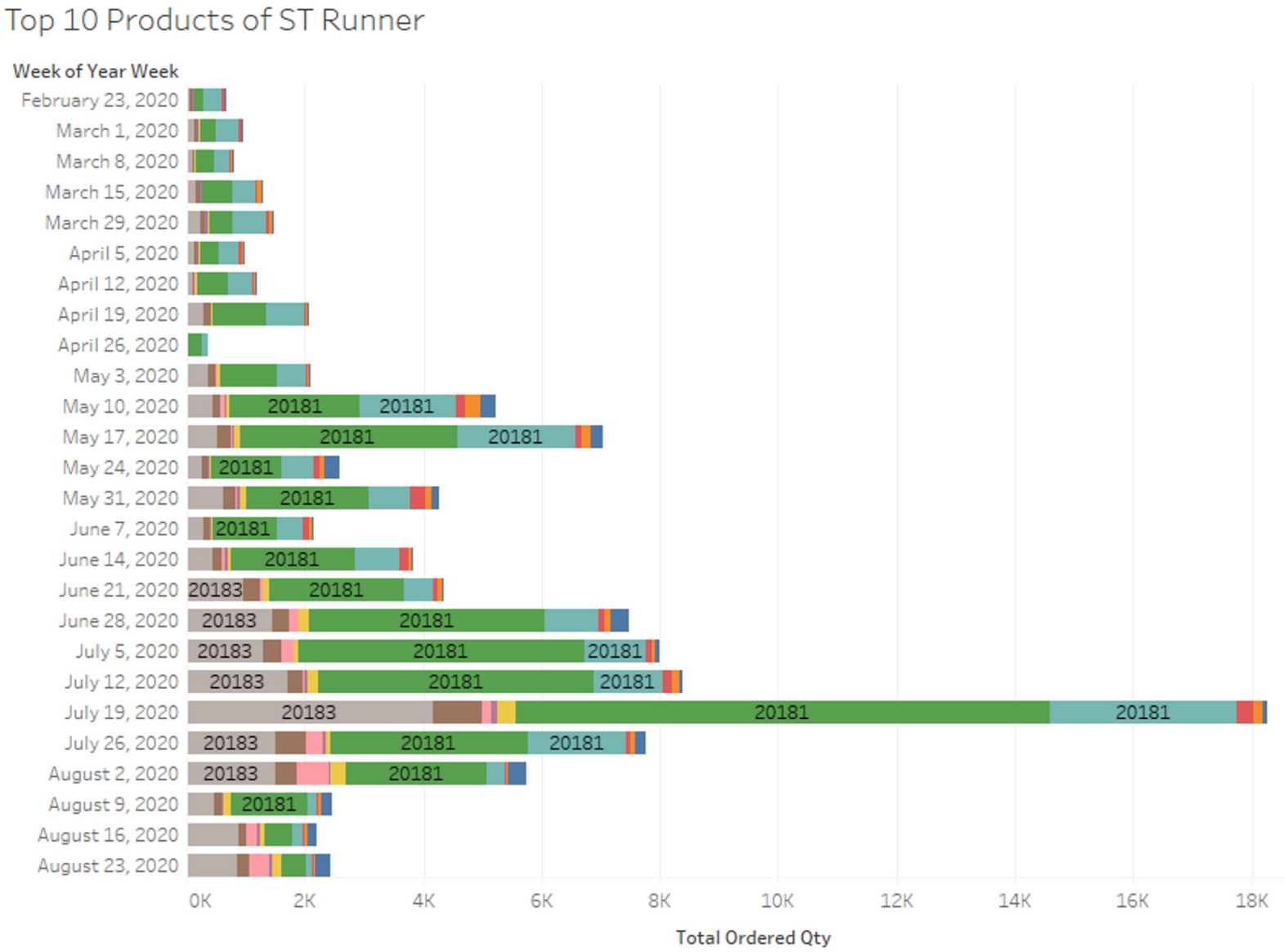


# EDA (Top 10 Products of Tazon)

- Sales rise or fall together, and there is no evidence that sales were driven by the popularity of certain products.



# EDA (Top 10 Products of ST Runner)



# Modeling (ARIMA)

- 80% of the data is used as the training set and the remaining 20% as the test set.


- Do the stationarity of the training set time series test.

  The p-value (0.002373) < 0.05, so the data is stationary. The order of differencing (d) = 0.

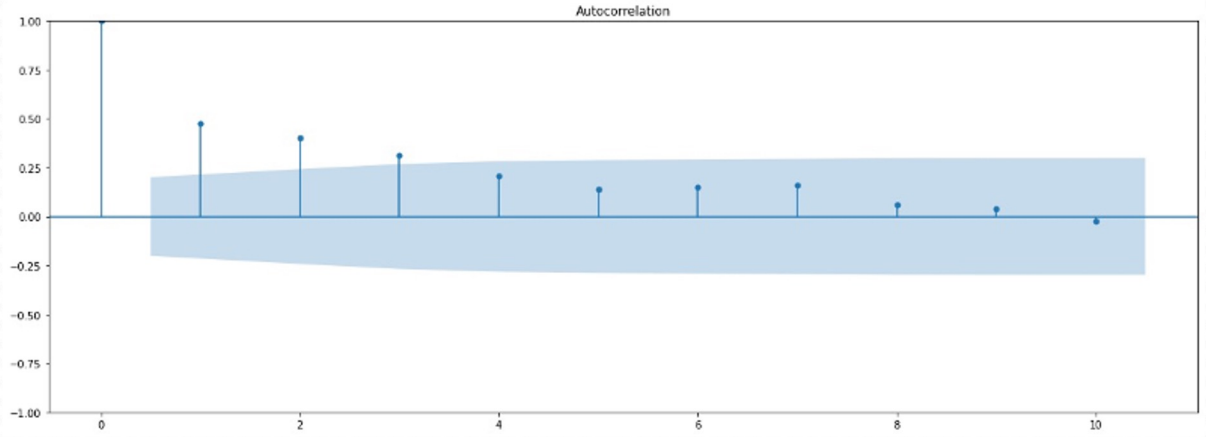
```
# Stationarity Test
result = adfuller(train)
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])

# p-value is less than 0.05, the time series in training set is stationary.
```

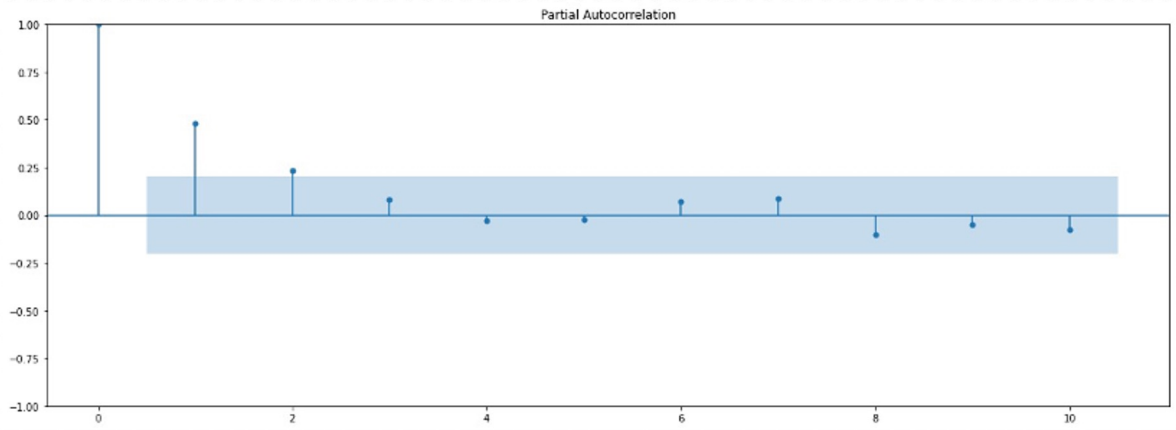
```
ADF Statistic: -3.857206
p-value: 0.002373
```

# Modeling (ARIMA)

- Determine the parameters q and p of ARIMA through autocorrelation (ACF) and partial autocorrelation (PACF)
- Parameter q should be within 3 and parameter p should not exceed 2



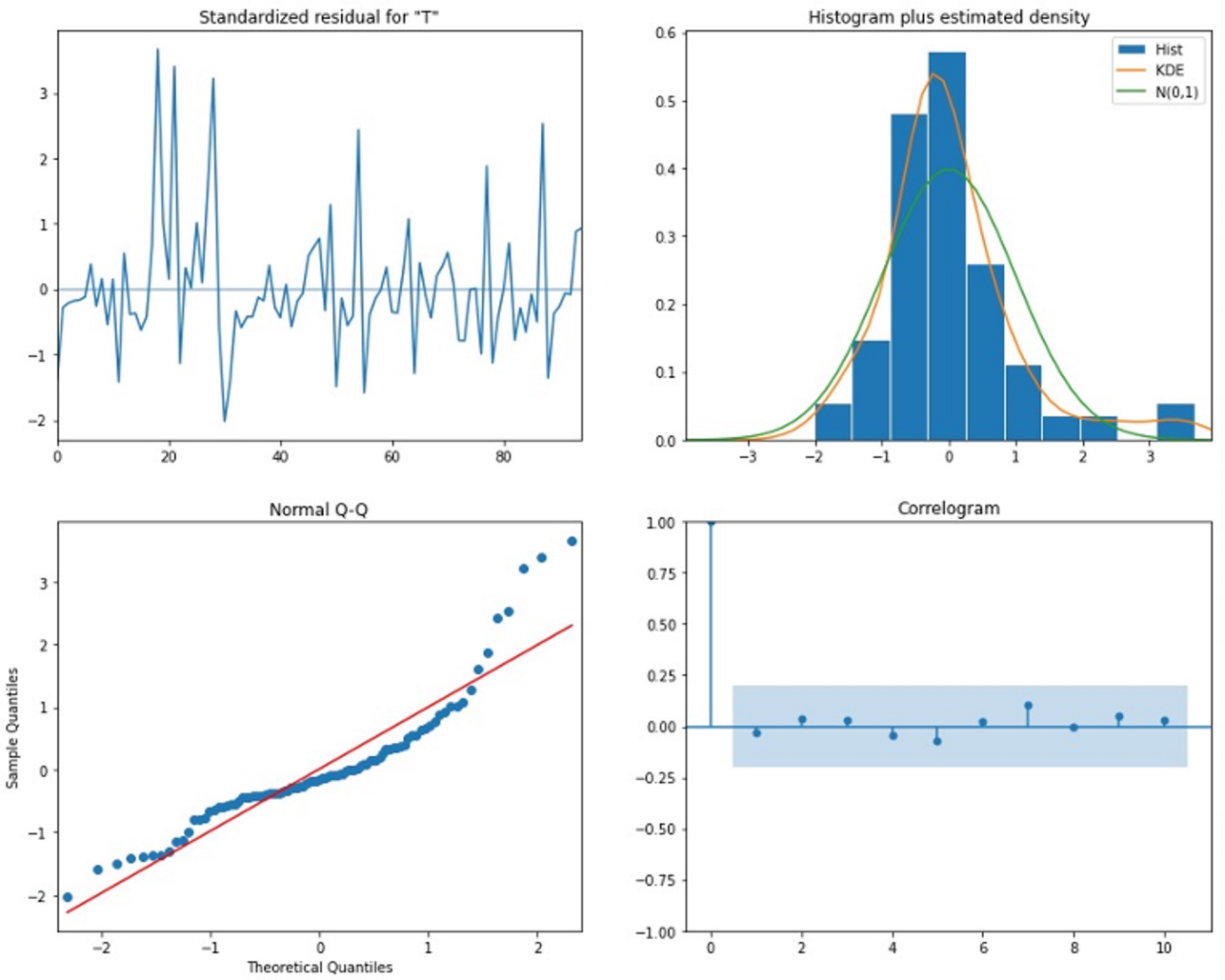

Autocorrelogram




Partial Autocorrelogram

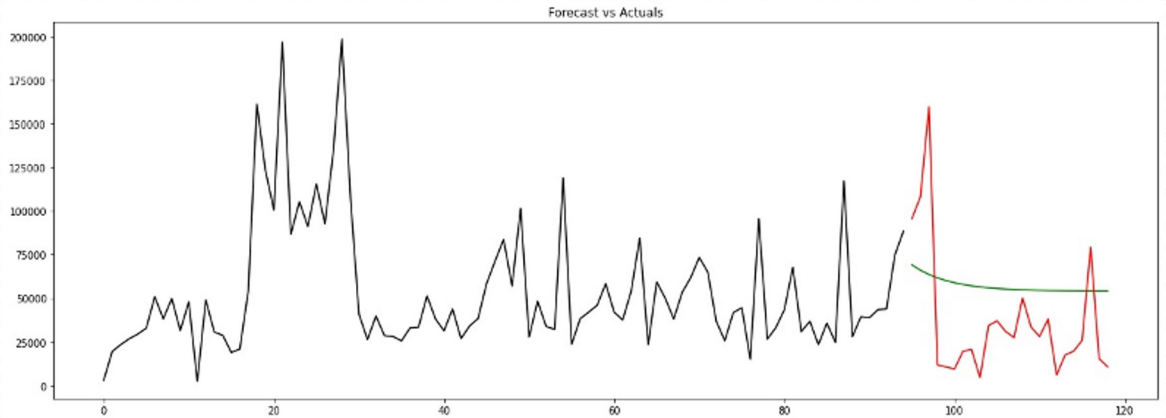
# Modeling (ARIMA)

- Using grid search to find the combination of parameters that results in the lowest AIC. The optimal combination of parameters (p, d, q) is (1, 0, 1).
- The residual term was diagnosed and found to be close to normal distribution, so it was white noise.

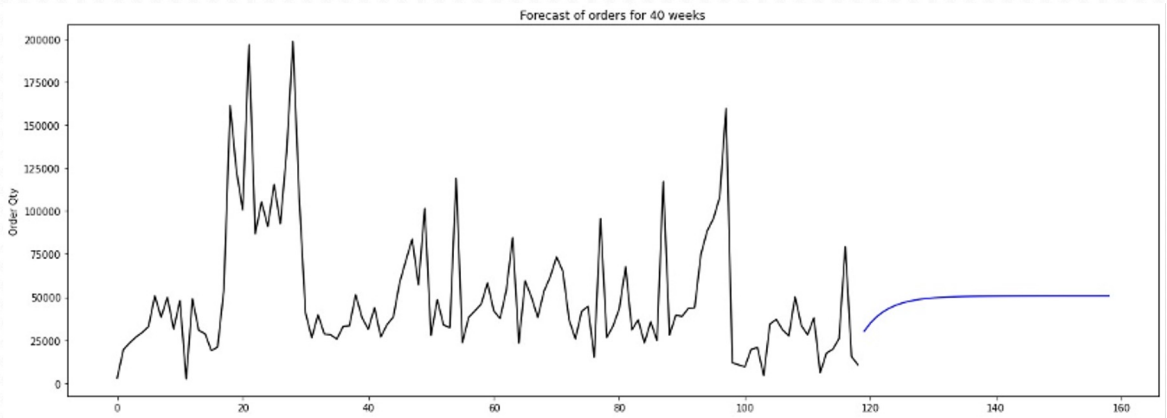


# Modeling (ARIMA)

- The mean square error (MSE) of the predicted values for test set is 2,183,179,120.3 and the root mean square error (RMSE) is 46,724.5.
- The predicted values have a large difference with the actual values of the test set, so the ARIMA model does not perform well.
- Predicted the total sales for the next 40 weeks, and the model still performed poorly.




Predictions on the test set

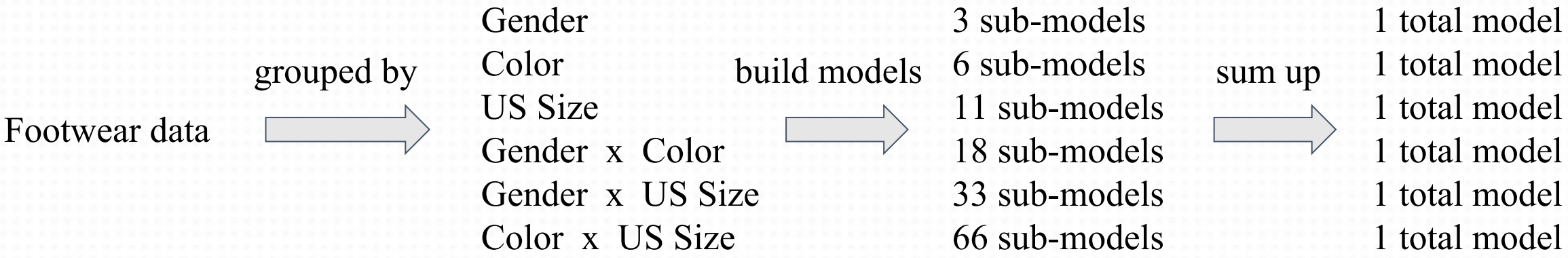



Sales forecast for the next 40 weeks

# Modeling (ARIMA)

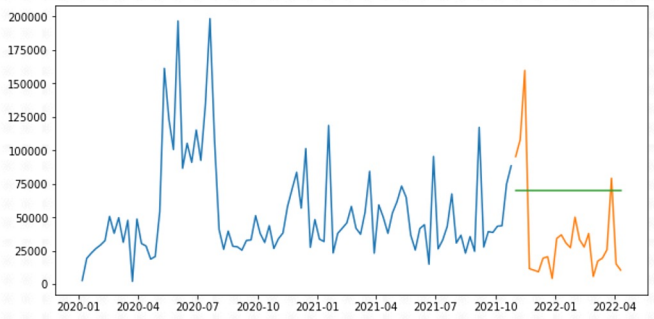
Split the data into subsets by Gender, Color, and US Size, predict separately, and sum the results to get the final prediction.

- Gender (MNS, WNS, KIDS)
- Color (black, white, blue, gray, pink, red)
- US Size (7, 7.5, 8, 8.5, 9, 9.5, 10, 10.5, 11, 11.5, 12)

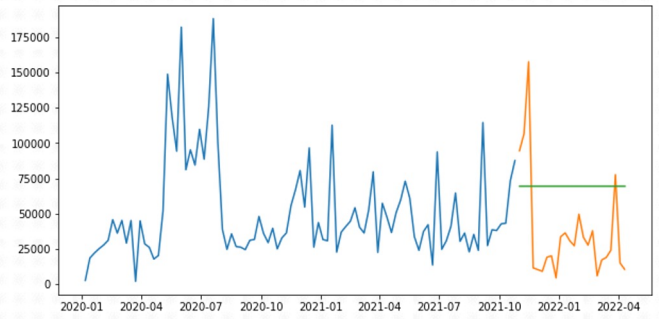


# Modeling (ARIMA)

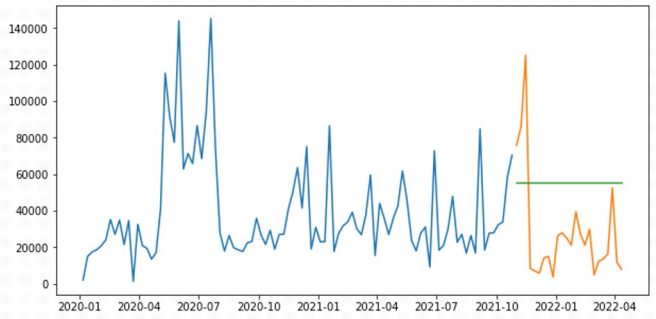
Sales forecast results are still close to a straight line, and models are underperforming.



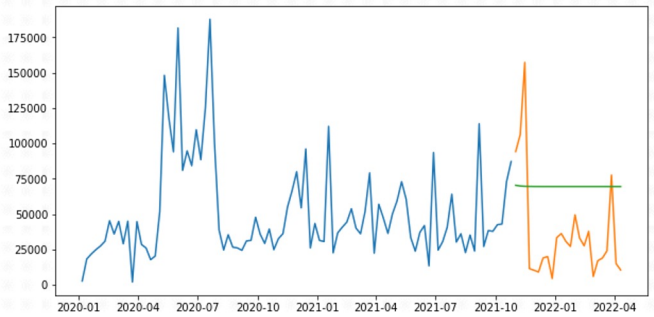

Grouped by Gender



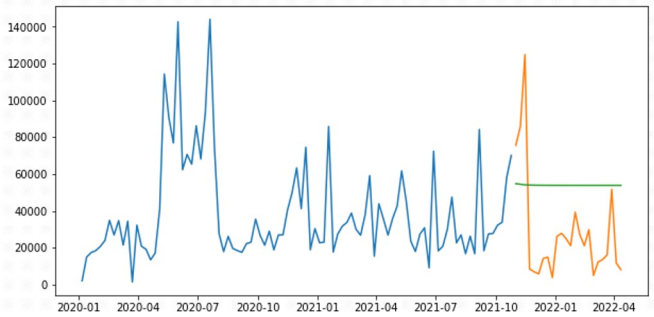

Grouped by Color



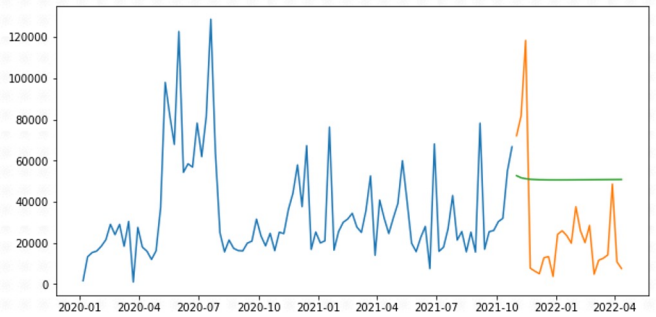

Grouped by US Size




Gender x Color



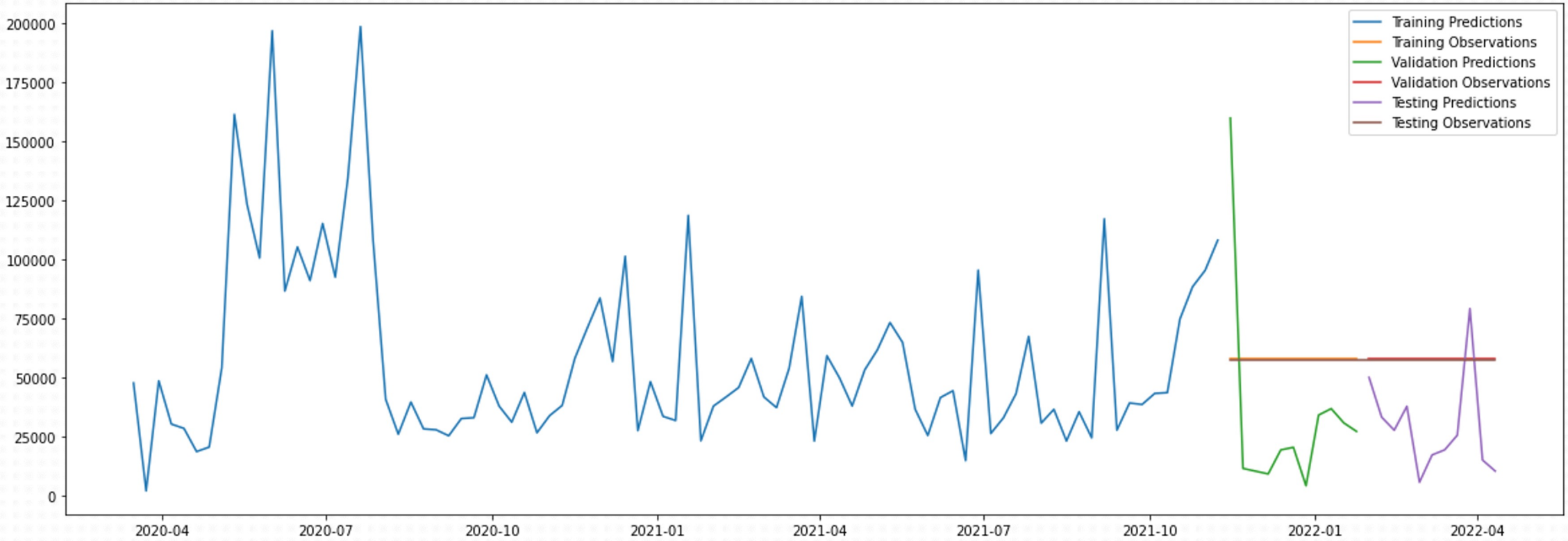

Gender x US Size




Color x US Size

# Modeling (RNN and LSTM)

The performance of the prediction results is still not good.



# Conclusion

- PUMA's products are very competitive in the field of footwear.
- The whole cancellation rate is in a high position.
- There is no evidence that sales were driven by a few popular products or new products.
- There is no seasonality in this time series data, and ARIMA model may not be suitable for predicting future sales.
- The prediction performance of RNN and LSTM models is also not good.
- There may be many other factors affecting sales.

# Recommendations

- Do some research on what causes different peaks, there may be the following reasons: co-branding, some large promotions and social media influencer
- Introduce more variables that may affect sales or create variables through feature engineering
- Try to predict from subdivided dimensions with other categorical variables, such as predicting the sales of a reporting business unit or a style family

# References


*Aditya Shastri, Elaborated Marketing Strategy of Puma + SWOT Analysis,*

*Retrieved from https://iide.co/case-studies/marketing-strategy-of-puma/*

*Autoregressive Integrated Moving Average (ARIMA)*. (2021, October 12). Investopedia. *https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp*

*Theodore Levitt, Exploit the Product Life Cycle,*

*Retrieved from https://hbr.org/1965/11/exploit-the-product-life-cycle*

*Tighe, D. (2022, March 1). Puma: Net sales share, by segment worldwide 2021. Statista. Retrieved April 18, 2022, from https://www.statista.com/statistics/254019/share-of-pumas-consolidated-sales-worldwide-by-segment/*

# Thank you!