Do more with data.

CREWASIS

New York | San Francisco | Toronto

UNNA BAKERY

UNNA BAKERY

About Unna Bakery

- Bakery product manufacturing company
- Uses natural ingredients
- Founded in 2015, and headquarters in New York
- Goal: Improve their sales

Agenda

- Cleaned Data
- Data visualization
- Hypothesis testing
- Build models
- Conclusions
- Recommendation

Cleaned Data

Combine three dataframes

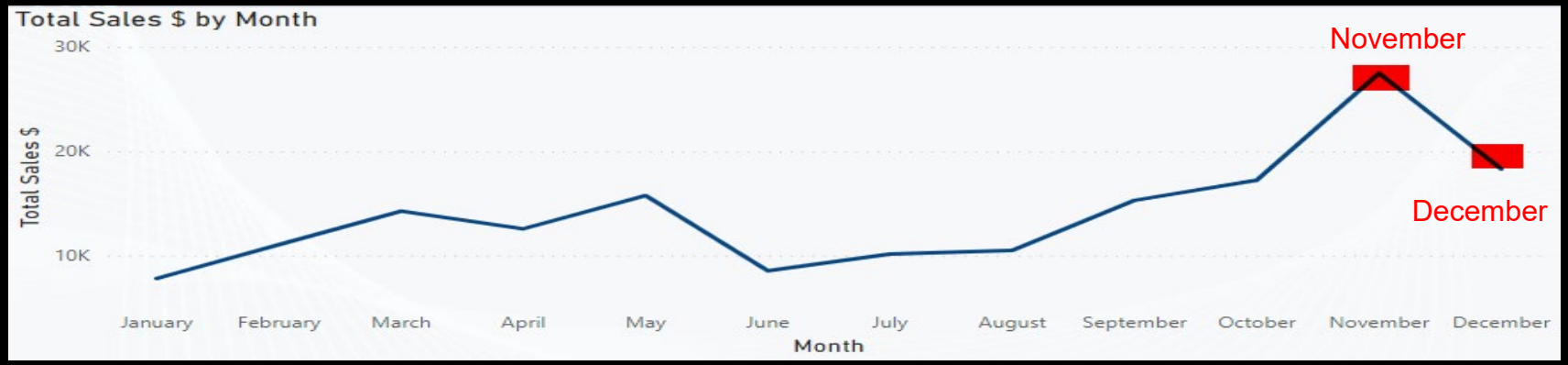
```
Combine_ALL = pd.concat([df_old_1,df_old2_1,df_old3_1])
Combine_ALL
```

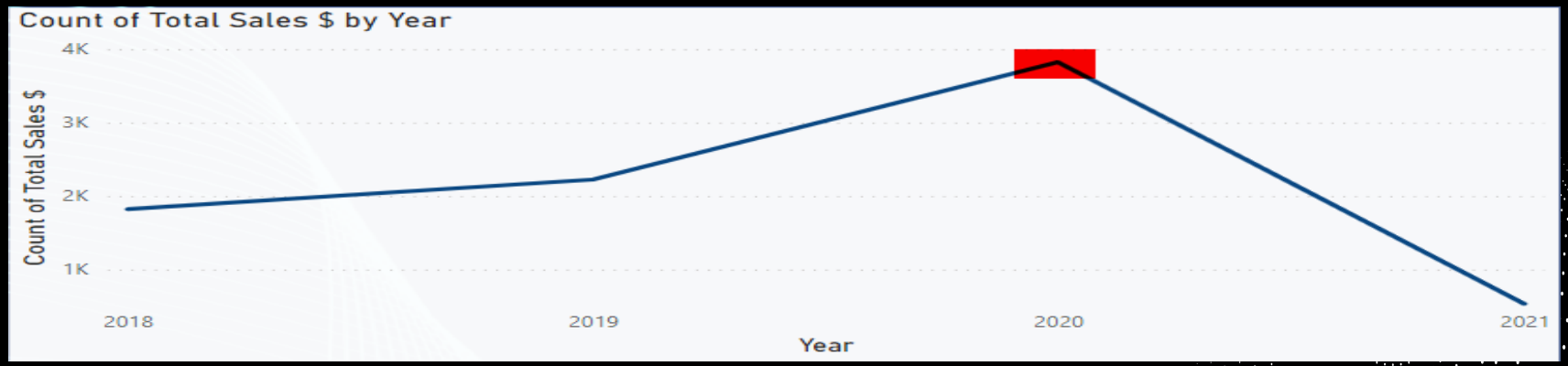
	CUSTOMER/VENDOR CODE	POSTING DATE	CUSTOMER/VENDOR NAME	SHIP-TO STREET	ITEM NO.	ITEM/SERVICE DESCRIPTION	QUANTITY	DOCUMENT NUMBER	SHIP-TO CITY	SHIP-TO STATE	SHIP-TO ZIP CODE	CREDIT HOLD
0	C13311	2018-02-03	Quincy	331 Nostrand Ave Brooklyn NY 11216	188101.0	Unna Bake - Raspberry Cave cookie (6 x 3.4 OZ)	1	NaN	NaN	NaN	NaN	NaN
1	C13311	2018-02-03	Quincy	331 Nostrand Ave Brooklyn NY 11216	188102.0	Unna Bake - Farmer's Cookie (6 x 3.4 OZ)	1	NaN	NaN	NaN	NaN	NaN
2	C13311	2018-02-03	Quincy	331 Nostrand Ave Brooklyn NY 11216	188103.0	Unna Bake - Chocolate Caramel Cookie(6 x 3.4 OZ)	1	NaN	NaN	NaN	NaN	NaN
3	C13311	2018-02-03	Quincy	331 Nostrand Ave Brooklyn NY 11216	188104.0	Unna Bake - Vanilla Dream Cookie (6 x 3.4 OZ)	1	NaN	NaN	NaN	NaN	NaN
4	C13311	2018-02-03	Quincy	331 Nostrand Ave Brooklyn NY 11216	188105.0	Unna Bake - Ginger Snap Cookie (6 x 3.4 OZ)	1	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
3490	C12682	2021-02-27	Fresh Direct - Bronx	2 Alexander Avenue	188503.0	Unna Bakery - Coconut Oat - (6 x 5.5 OZ)	1	3772420.0	Bronx	NY	10454.0	N
3491	C12682	2021-02-27	Fresh Direct - Bronx	2 Alexander Avenue	188501.0	Unna Bakery - Vanilla Cookie - (6 x 5.5 OZ)	2	3772420.0	Bronx	NY	10454.0	N
3492	C12682	2021-02-27	Fresh Direct - Bronx	2 Alexander Avenue	188504.0	Unna Bakery - Brown Butter Cookie - (6 x 5.5 OZ)	2	3772420.0	Bronx	NY	10454.0	N
3493	C15233	2021-02-27	WFM - NE WEE - WEEHAWKEN	1400 Waterfront Terrace	188502.0	Unna Bakery - Lemon Lime - (6 x 5.5 OZ)	1	3772426.0	Weehawken	NJ	7086.0	N
3494	C10993	2021-02-27	WFM - NYC TRB - TRIBECA - 10245	270 Greenwich Street	188501.0	Unna Bakery - Vanilla Cookie - (6 x 5.5 OZ)	1	3773117.0	New York	NY	10007.0	N

8405 rows × 17 columns

Total 8405 rows and 17 columns

Sales by Months and Years

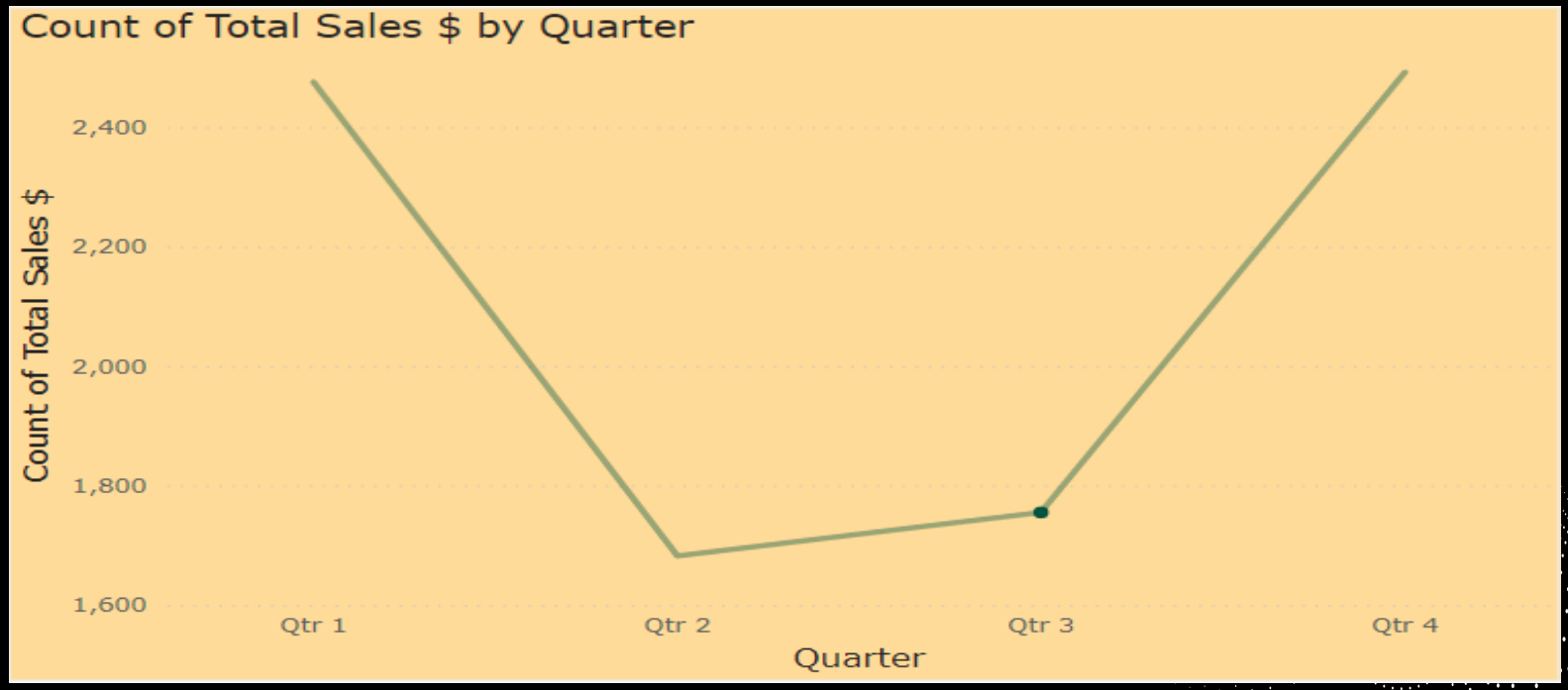




Based on the line chart, we can find the higher sales months are in November and December. The Unna bakery is US company, so their products are mainly sold in North America. However, there are two popular holidays in the United states. One is Thanksgiving, and the other is Christmas. Also, when COVID-19 broke, their sales are impacted.

Sales by Quarter

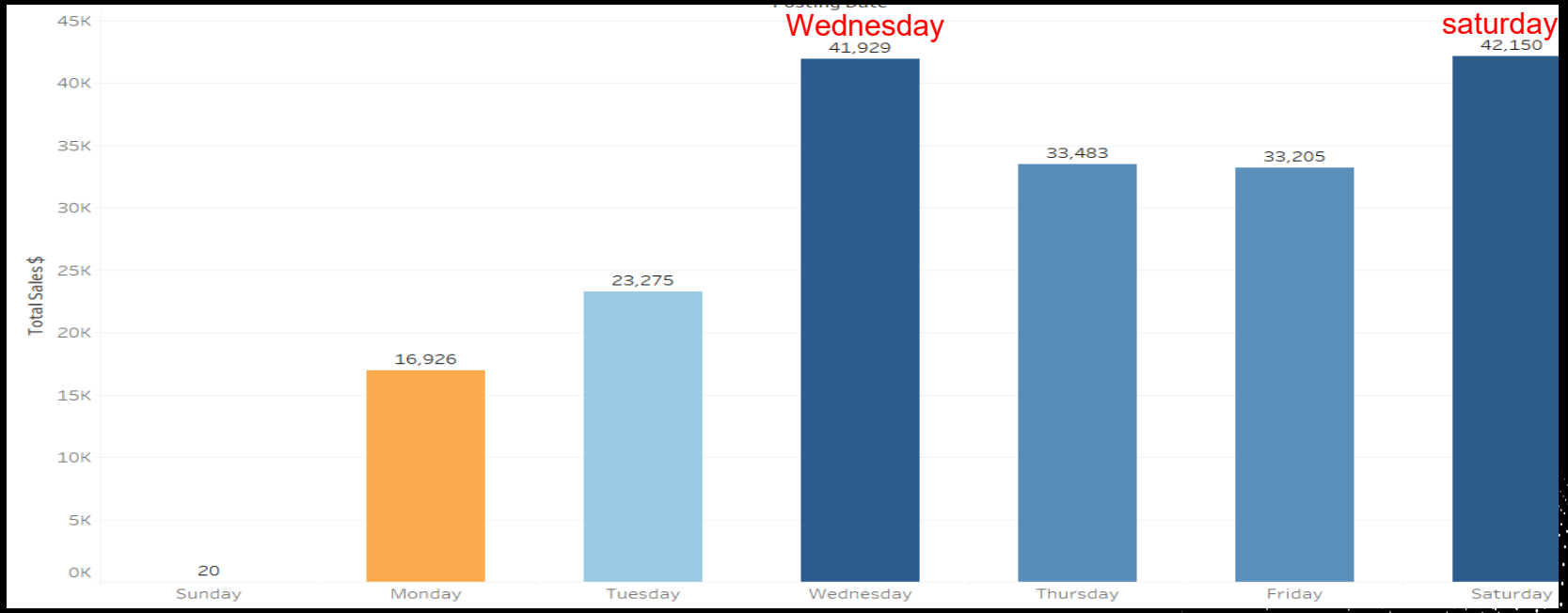
Quarter 2 and Quarter 3 are off-season



- Quarter 2 and Quarter 3 are off-season for this industry. Therefore, the company can combine the different types of products to increase sales.

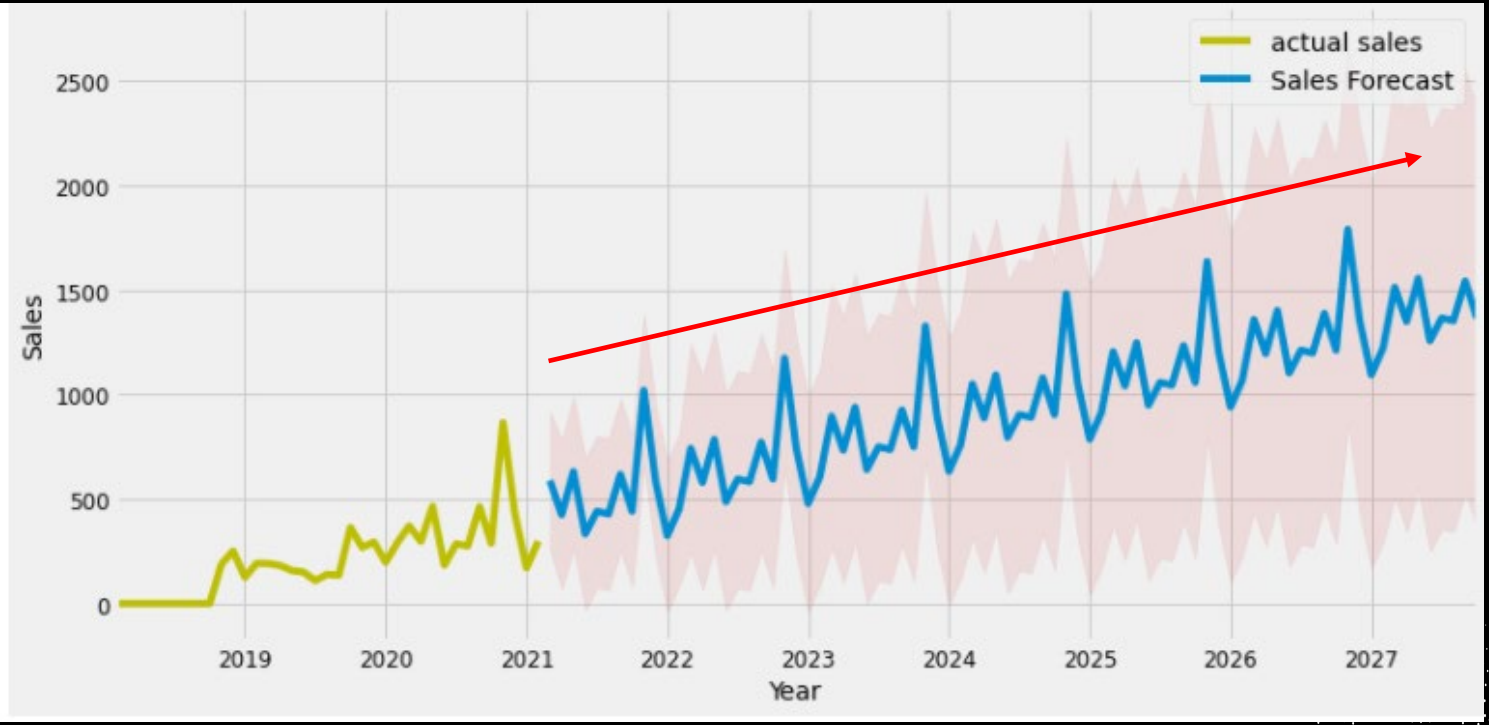
Sales by day

Wed and Sat are the best days to sell



- Wed. and Sat. is a good time to promote their new products to customers.
- Need to adjust inventories on Mon and Tue

Sales Forecast



- The trend of Unna sales grow from 2019 to 2027.
- It means that the sales of company's product strategies are correct. Therefore, the company to grow steadily every year.

Sales Regions

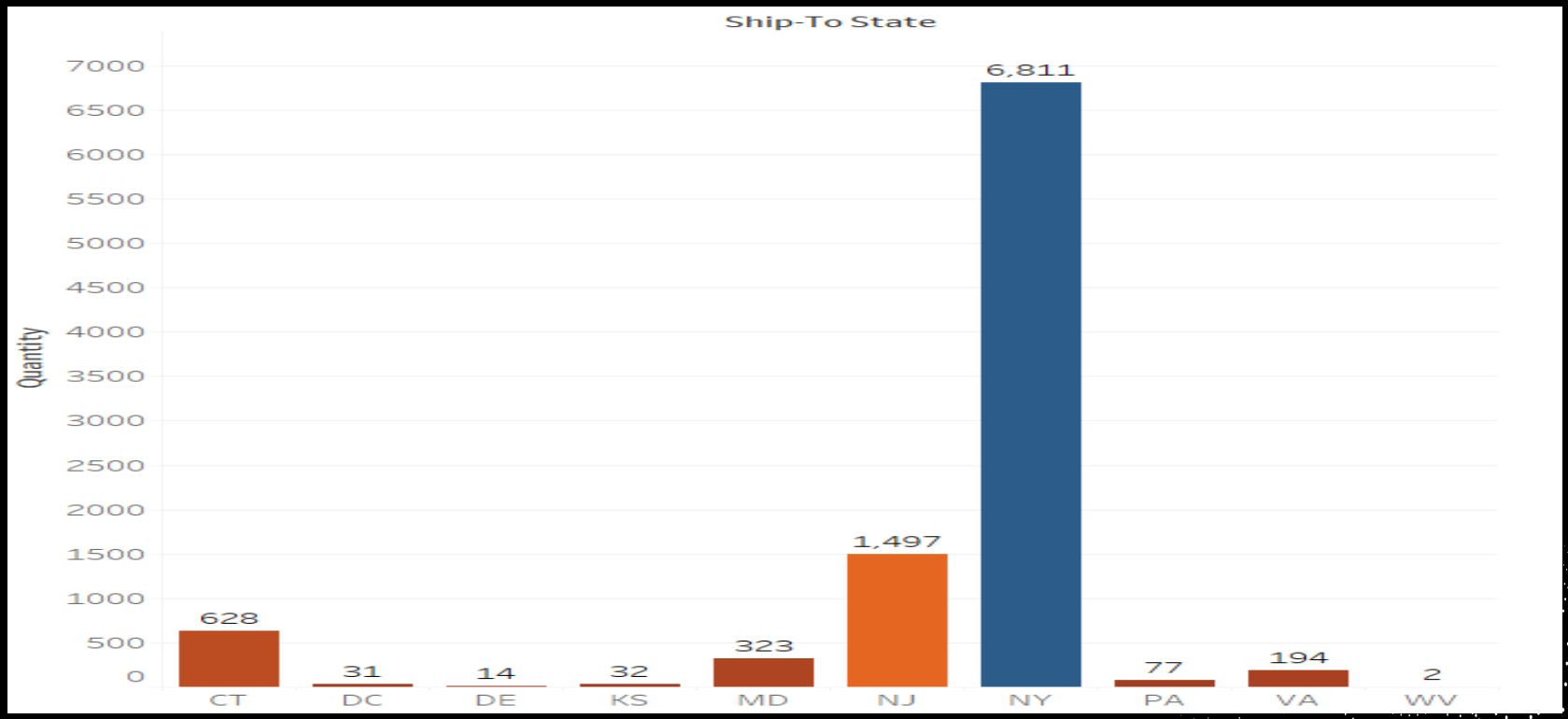
NY area represents 86% of the total income

86% of the total income

State	Total Sales
NY	$132951
NJ	31732
CT	$13894
MD	$ 4945
VA	$ 3406

Map graph can present marketing sales. The dot is bigger. It means that the sale is higher. Based on the Map graph, we can find that the higher sales are closed to NY because the company headquarter is in NY. Also, NY area represents 86% of the total income.

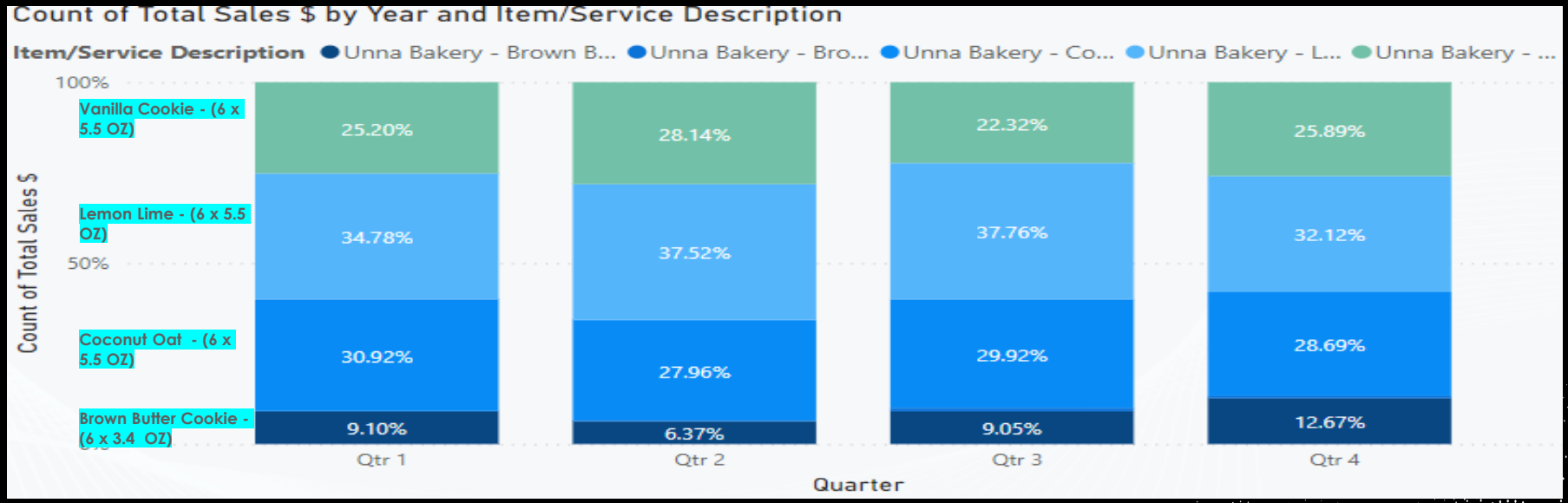
State and Quantity



From bar chat, we can find that the largest export place is New York. The second on is NJ. The third one is CT. This can be used as an indicator that is for the company's future expansion stores.

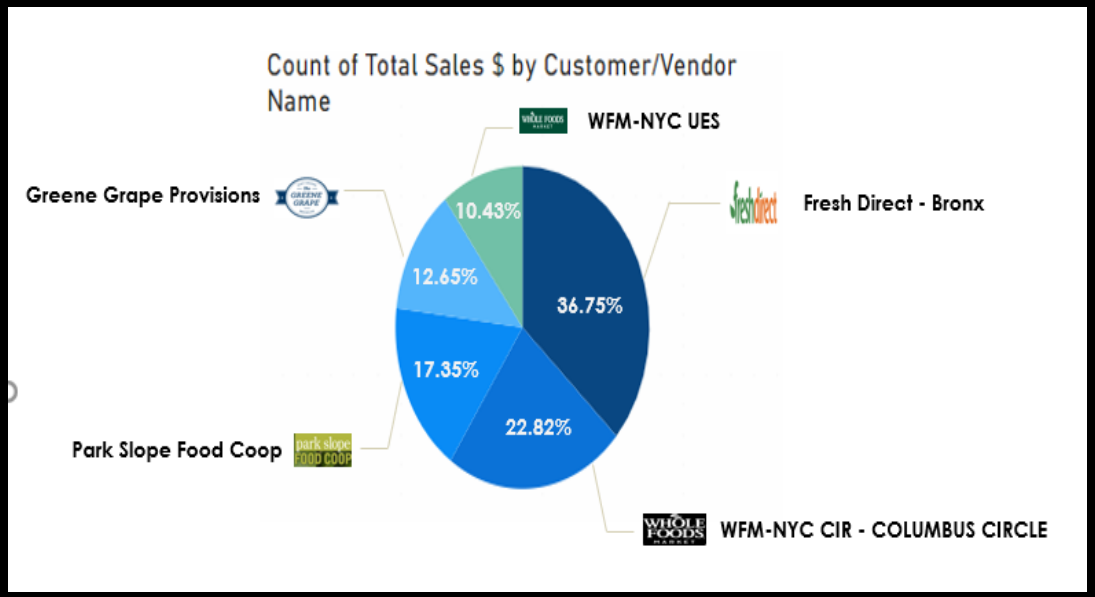
Popular product

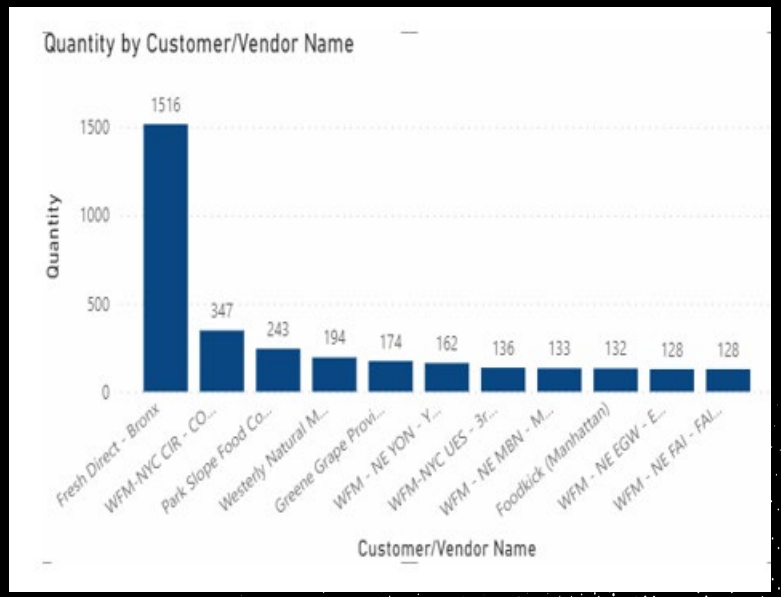
The percentage of products



Based on a stacked column chart, we can find that the best-selling product is Lemon Lime. The second one is coconut oat. The third one is Vanilla Cookie. Therefore, the company can adjust the inventory of their products based on products sales.

Top 5 Selling Stores of Unna bakery

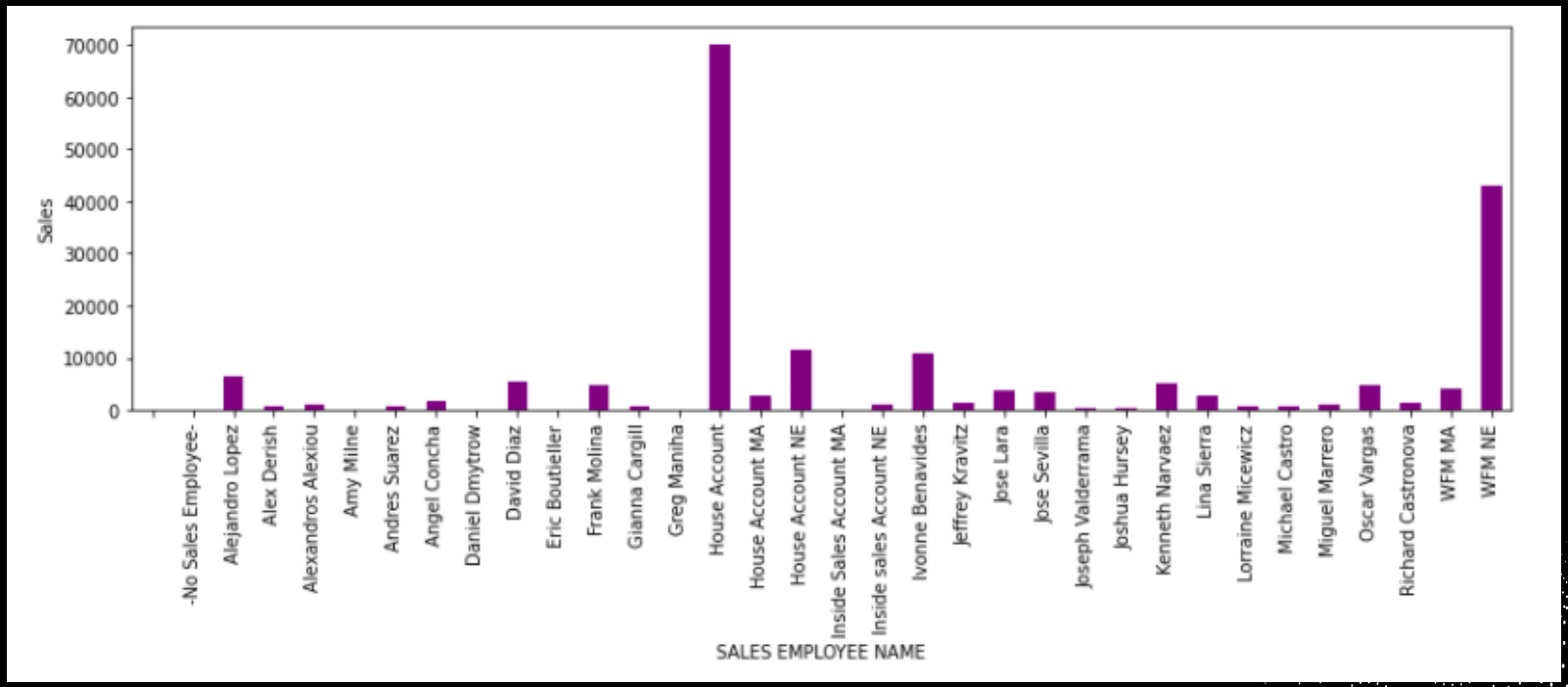




Based on a pie chart and a bar chart, we can find that the top1 selling store is Fresh Direct. The second one is WFM-NYC CIR. The third one is park slope food. Therefore, the Unna bakery should build good relationships with these stores.

Top sales

Salesperson's sales amount



Based on a bar chart, we can find that amount of products are sold by each salesperson. The company can understand who salesperson has better sales.

Hypothesis testing

OLS Regression Results

Dep. Variable:	y	R-squared:	0.873
Model:	OLS	Adj. R-squared:	0.869
Method:	Least Squares	F-statistic:	210.1
Date:	Wed, 09 Feb 2022	Prob (F-statistic):	0.00
Time:	20:18:16	Log-Likelihood:	-23029.
No. Observations:	6051	AIC:	4.644e+04
Df Residuals:	5858	BIC:	4.774e+04
Df Model:	192		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1 Posting date	3.6198	1.143	3.166	0.002	1.378	5.861
x2 Quantity	693.8200	4.832	143.584	0.000	684.347	703.293
x3 City	-15.0266	3.565	-4.215	0.000	-22.015	-8.038
x4	8.4378	7.721	1.093	0.274	-6.698	23.573
x5	0.3440	2.576	0.134	0.894	-4.706	5.394

Based on a hypothesis testing, posting date, quantity, and city are both significant because their P values are less than 0.005. Also, quantity and total sales are the most relevant because it has the highest correlation.

Analyze results of KNN

I split 80% training set, and 20% testing set. After that the training set split 10% data to validation set.

Based on figure 1, we found that the flavor cookies across different time and places and total sales were positively correlated. It means that the total sales gets increased or decreased when customer's flavors for products gets changed.

According to Table 1, the value of MSE was 216.4428, the value of MAE =5.736, and the value of R2 was 0.77. The R2 value indicated 77% of the variation in total sales could be explained by customer's flavors for products .

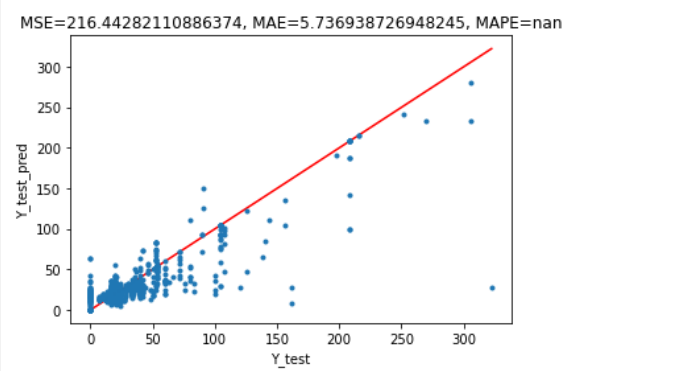


Figure 1 customer's flavors for products VS total sales

Table 1. The performance of KNN algorithm

Algorithms	MSE	MAE	R2
KNN regression	216.4428	5.736	0.77

Analyze results of Random Forest

Based on figure 2, we found that the different flavor cookies across different time and places and total sales were positively correlated. It means that the total sales gets increased or decreased when customer's flavors for products gets changed.

According to Table 2, the value of MSE was 43.52, the value of MAE =1.843, and the value of R2 was 0.956. The R2 value indicated 95.6% of the variation in total sales could be explained by customer's flavors for products .

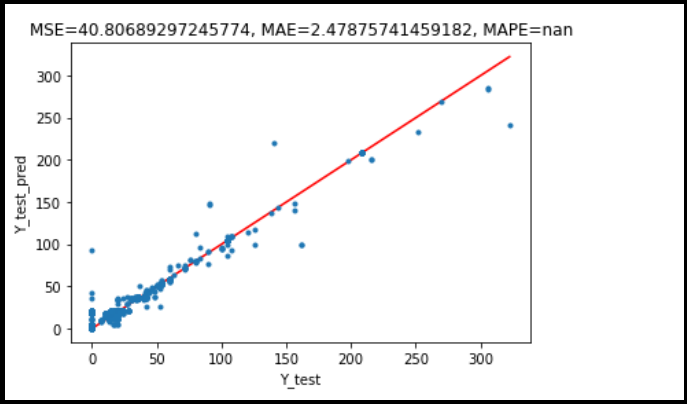


Figure 2 customer's flavors for products VS total sales

Table 2. The performance of Random forest algorithm

Algorithms	MSE	MAE	R2
Random Forest regression	43.52	1.843	0.956

Analyze results of XGBoost

Based on figure 3, we found that the flavor cookies across different time and places and total sales were positively correlated. It means that the total sales gets increased or decreased when customer's flavors for products gets changed.

According to Table 3, the value of MSE was 1.807, the value of MAE =1.8422, and the value of R2 was 0.914. The R2 value indicated 91.4% of the variation in total sales could be explained by customer's flavors for products.

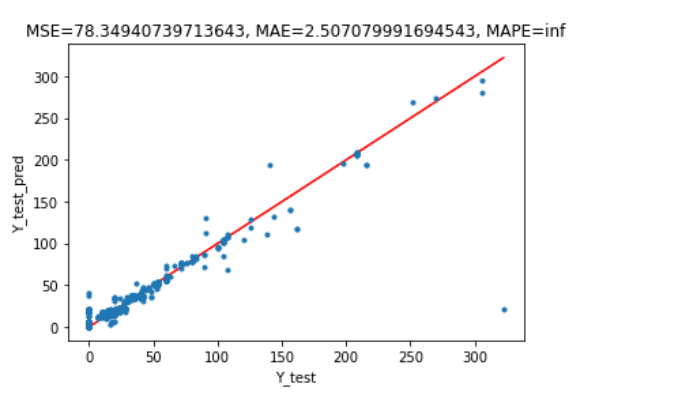


Figure 3 customer's flavors for products VS total sales

Table 3. The performance of XGBoost algorithm

Algorithms	MSE	MAE	R2
XGBoost	1.807	1.8422	0.914

Compare the performance of three algorithms

Based on the MSE and MAE indicators, XGBoost model had the highest accuracy, followed by Random Regression and KNN Regression models.

Random Forest regression had the largest amount of the variability explained by the model, but KNN regression had the smallest amount of the variability explained by the model.

Table 4. The performance of three algorithms

Algorithms	MSE	MAE	R2
KNN regression	216.4428	5.736	0.77
Random Forest regression	43.52	1.843	0.956
XGBoost	1.807	1.8422	0.914

Conclusions

- The three algorithms can precisely predict total sales, which can further help the company to predict total sales for their products.
- According to the results, Random Forest regression has the largest amount of the variability explained by the model and XGBoost has higher accuracy.
- Regression algorithms not only can be used in the field of food industry but also can be used in other fields, such as financial and health care.

Recommendations

- I suggest using a XGBoost algorithm to predict total sales of the products because the highest accuracy.
- We can also use others regression algorithms to analyze the data, such as SVM, and decision tree.
- Most business sales are in New York State. Therefore, they can focus more on their stores on the third state (CT) to increase their sales.
- The best sales are on Wednesday and Saturday. Therefore, it is a good time to promote their new products to customers.

Recommendations

- Quarter 2 and Quarter 3 are off-season for this industry. Therefore, the company can combine the different types of products to increase sales.
- Companies should adjust business strategies with the holidays and COVID-19 reasons.

THANK YOU

Do more with data.

CREWASIS

New York | San Francisco | Toronto