Retail Case Study Model Presentation

Virtual Internship

December 30, 2023

Background - Retail Case Study

Problem: *X* is a company that has a beverages business in Australia. They sell their products through various super-markets & also engage in heavy promotions throughout the year. Their demand is influenced by various factors like holidays & seasonality. They need a forecast for each of their products at the item level in weekly buckets.

Business Understanding: The objective is to build a multivariate machine learning model that will be able to forecast sales weekly. The data that is required to build the model may need to be re-coded so that the dates are in weekly buckets.

Data Understanding

1. What type of data have you got for analysis?

The data that I am working with is numeric, with the exception of the features: Product, date, & Price.Discount. The exceptions are strings. Features In.Store.Promo, Catalogue.Promo, Store.End.Promo, Covid_Flag, V_DAY, EASTER, & CHRISTMAS are boolean values: 1 denoting True, 0 denoting False.

2. What are the problems in the data (number of NA values, outliers, skewed, etc)?

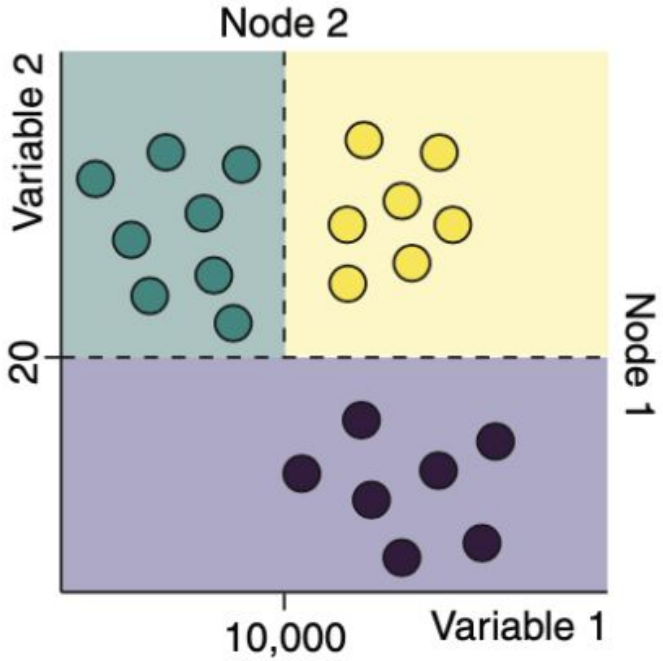
The problems I see in the dataset are the features Price.Discount.... & date. The Price.Discount.... feature needs to be renamed. The date feature needs to be re-coded into weekly buckets.

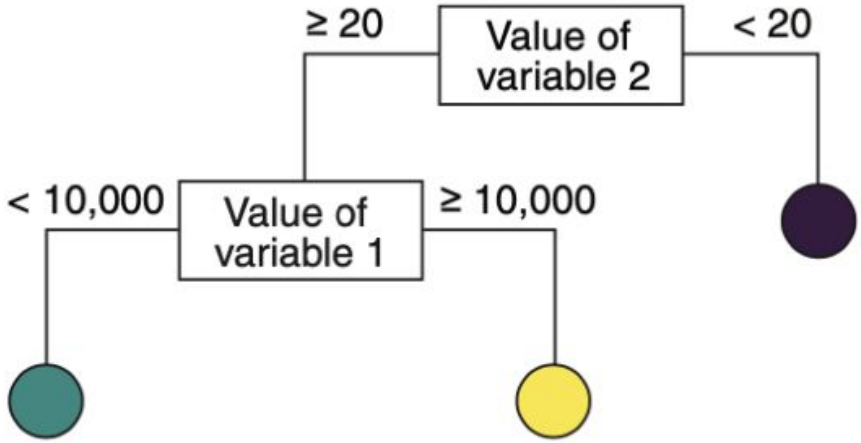
3. What approaches will you apply to your dataset to overcome problems like NA values, outliers, etc. & why?

If there are NA values, I will remove the rows with them. It's better to train a model when we don't have missing data. Although there are many methods to resolve missing data, I feel that doing so will not reflect the true population values of the features. I believe that for this project, the model could be slightly overfit, since only this company will be using this data for its forecasting. Outliers in the sales feature could be the results of promotions, seasonality, holidays, etc. & I want to take that into account when training the model. As such, I will try a model with outliers & without outliers. When building the model without outliers, only extreme outliers will be removed. I believe the systematic way to identify extreme outliers in a statistical research setting is anything over 75th quartile + 3 IQR or anything under 25th quartile - 3 IQR. Anything within will be retained. Extreme outliers will affect the model training process. This model that has removed outliers will be more generalizable. We will determine which is better with the models' predictive accuracy & context.

Chosen Model

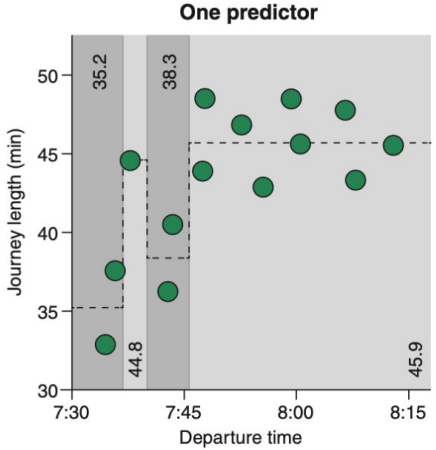
I chose random forests model. The random forests algorithm splits the feature space into separate regions, one split at a time. The feature space refers to the n-dimensions where our variables reside. It tries to learn the binary splits that result in regions that are regions that are as pure as possible (containing mostly items of a singular class).

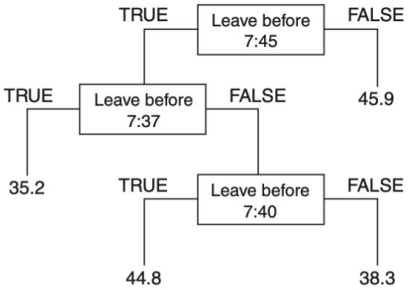


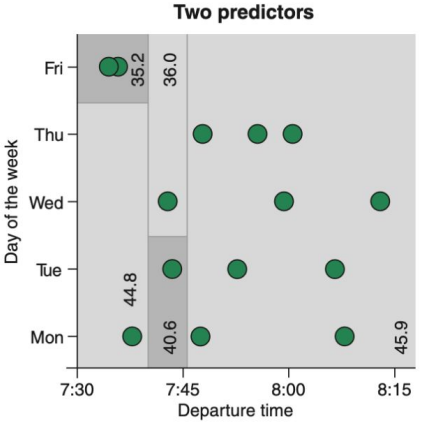


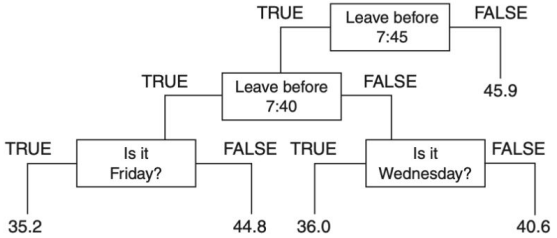
Chosen Model (continued)

The partitioning is much alike herding animals into their respective pens. One pen is for the chickens, the other for the pigs, another for the cows. But how would this work if the outcome variable isn't so clear cut? What if the outcome variable is continuous? Well, in the same way, the only difference is that instead of each region representing a class, it represents a value of the continuous outcome variable.



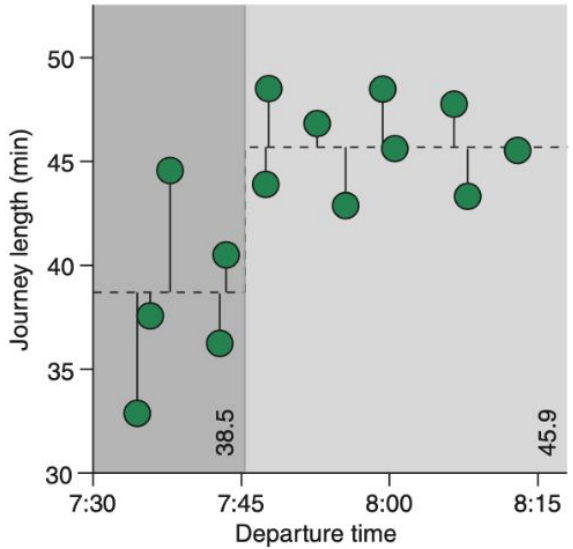


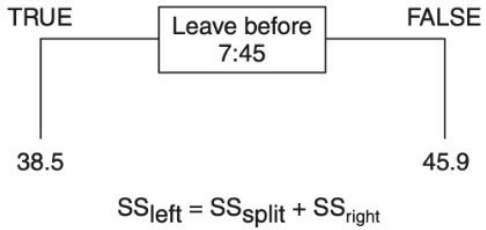




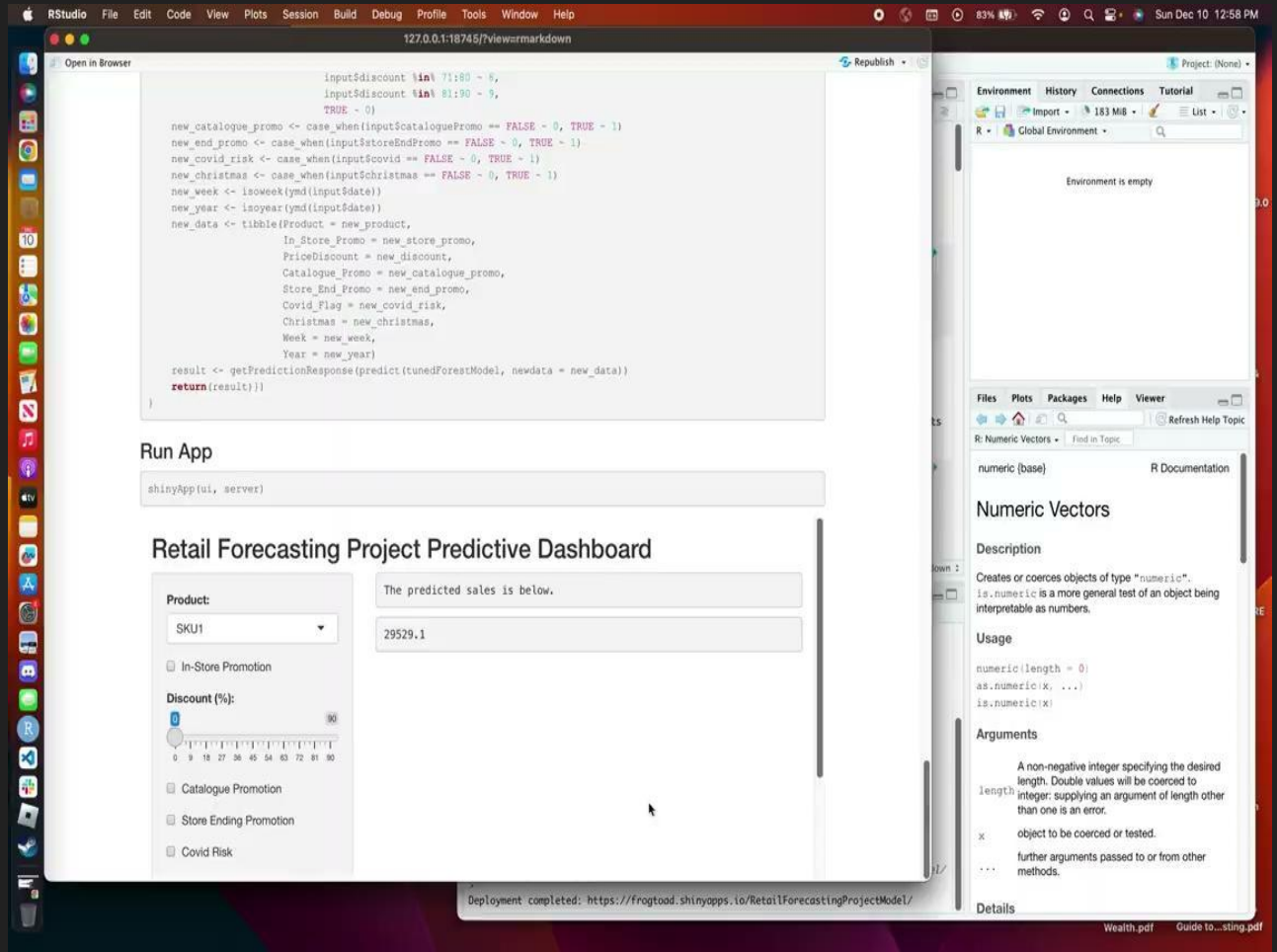
Chosen Model (continued)

The nodes of the regression tree split splits the feature space into distinct regions. Each region represents the mean of the outcome variable of the cases inside it. When making predictions on new data, the model will predict the value of the region of the new data falls into. Although the figures illustrate situations for 1 or 2 predictor variables, it extends to any number of predictors. The way the splits are decided for any variable is by looking for the split with the lowest sum of squares.



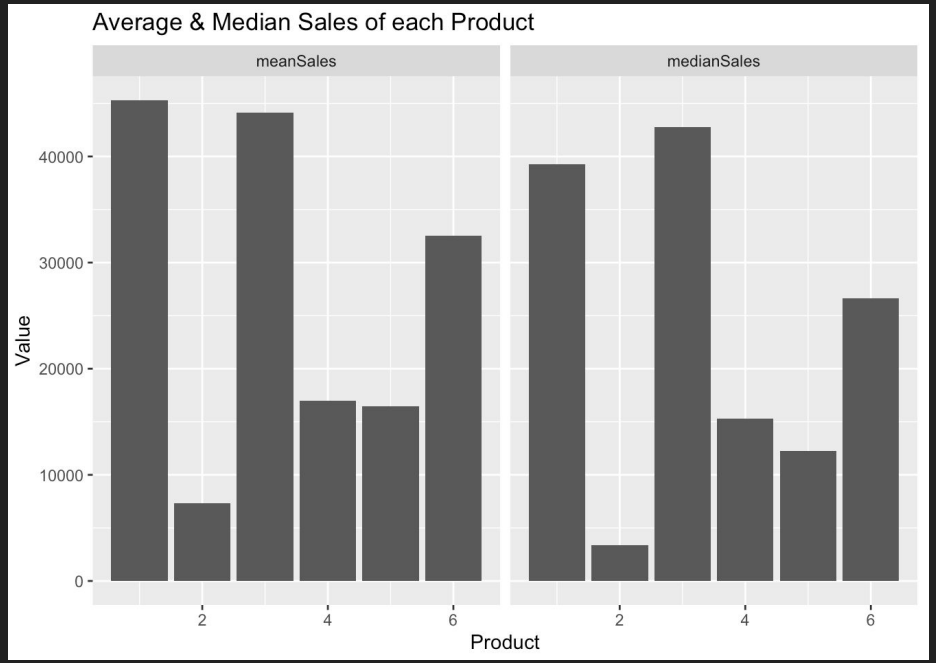


Model Dashboard



Product Sales

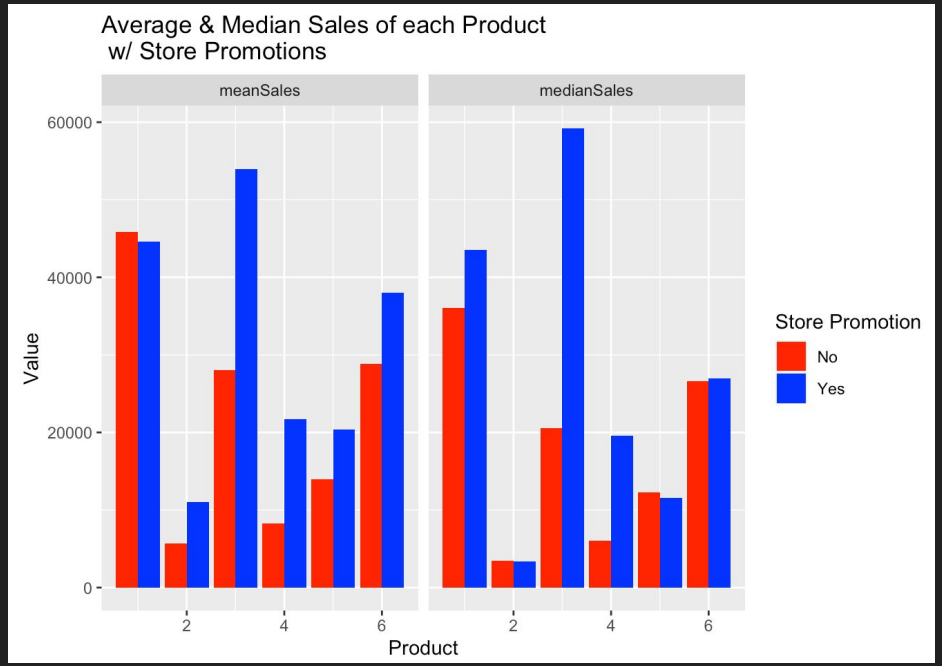
Beverages 1 & 3 have the highest grossing median sales. I provided both median & average sales because of the skewness of the sales distribution. Beverage 6 comes after 1 & 3 in tier, then 4 & 5, & lastly 2. Based off the visualisation, we recommend that company X should abandon beverage 2 & invest in their other beverages that average higher sales.



	Product	Measure	Value
	<dbl>	<chr>	<dbl>
1	1	meanSales	45300.
2	1	medianSales	39290
3	2	meanSales	7305.
4	2	medianSales	3392.
5	3	meanSales	44132.
6	3	medianSales	42768.
7	4	meanSales	16977.
8	4	medianSales	15270
9	5	meanSales	16471.
10	5	medianSales	12249
11	6	meanSales	32528.
12	6	medianSales	26648

Product Sales w/ Store Promotions

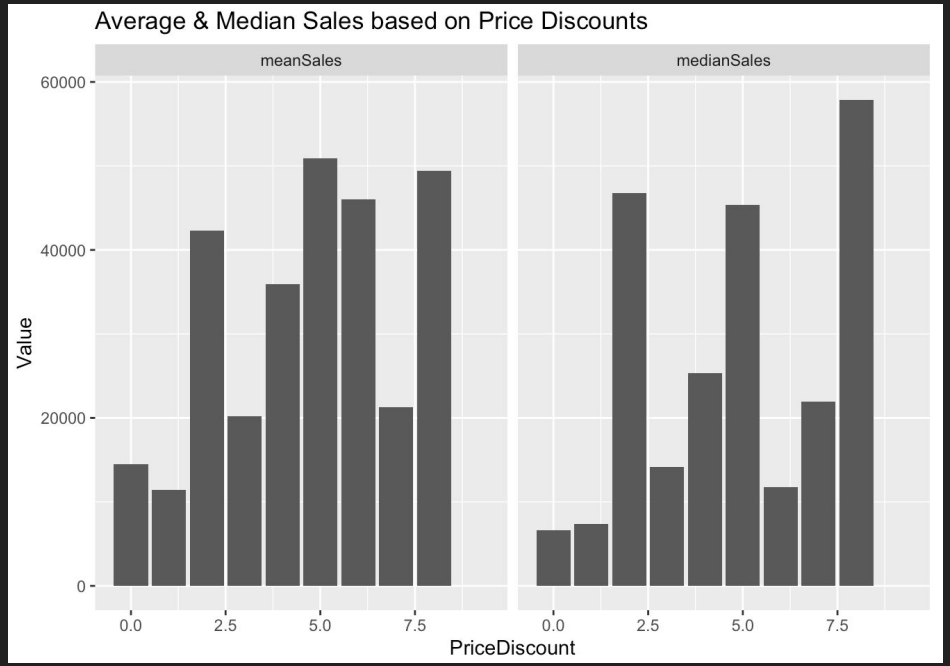
Based on average sales, having store promotions seem to have a significant effect on the sales numbers for beverages 2-6. Sales numbers are not significantly affected for beverage 1. When looking at median sales, beverages 1, 3, & 4 see significant differences in sales where there is a store promotions. Beverages 2, 5-6 see no significant difference.



	Product	In_Store_Promo	Measure	Value
1	1	0	meanSales	45831.171
2	1	0	medianSales	36072.000
3	1	1	meanSales	44560.714
4	1	1	medianSales	43572.000
5	2	0	meanSales	5658.234
6	2	0	medianSales	3416.000
7	2	1	meanSales	10989.540
8	2	1	medianSales	3322.000
9	3	0	meanSales	28025.389
10	3	0	medianSales	20540.500
11	3	1	meanSales	53959.517
12	3	1	medianSales	59189.000
13	4	0	meanSales	8300.236
14	4	0	medianSales	6020.500
15	4	1	meanSales	21709.045
16	4	1	medianSales	19570.000
17	5	0	meanSales	13945.742
18	5	0	medianSales	12307.500
19	5	1	meanSales	20385.213
20	5	1	medianSales	11546.500
21	6	0	meanSales	28847.351
22	6	0	medianSales	26577.000
23	6	1	meanSales	38048.921
24	6	1	medianSales	26968.500

Product Sales w/ Price Discounts

Intuition leads us to think that increase discounts leads to increased sales. However upon visualising, it does not seem to be the case. There are spikes in sales when price discounts are 21-30%, 41-50%, & 71-80%. These spikes in sales could be because of particular promotions that drove sales higher. Company X should definitely revisit those promotions again to see promotion sales numbers across time.

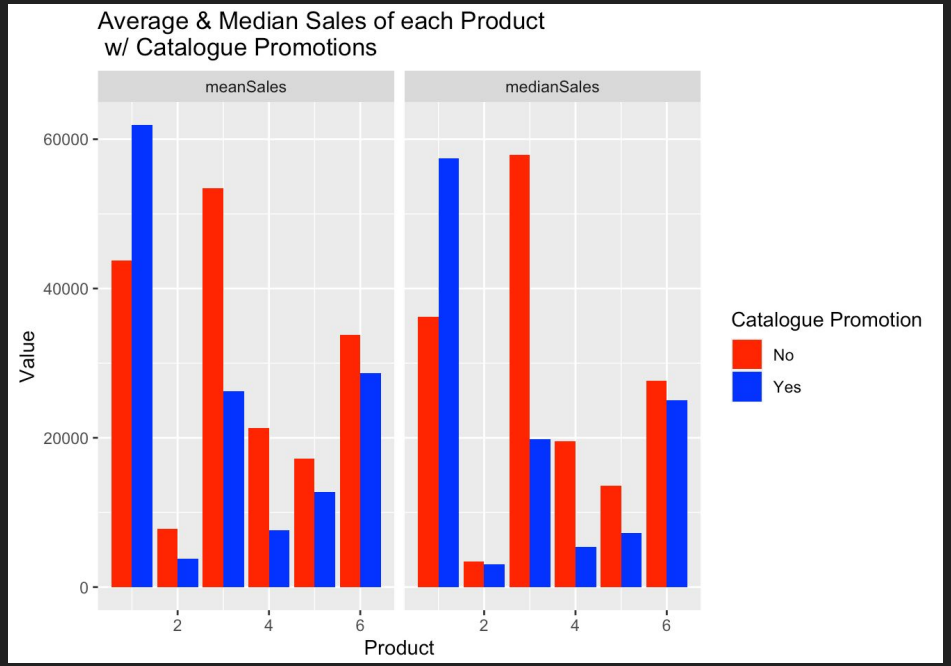



```
## # A tibble: 10 × 3
##    PriceDiscount meanSales medianSales
##            <dbl>     <dbl>       <dbl>
##  1             0    14470.        6622
##  2             1    11388.        7398
##  3             2    42278.       46728.
##  4             3    20193.       14198.
##  5             4    35927.       25314.
##  6             5    50887.       45348
##  7             6    46032.       11742
##  8             7    21291.       21936.
##  9             8    49431.       57867
## 10             9        0            0
```

Product Sales w/ Catalogue Promotions

Catalogue Promotions seem to have significant effect on sales especially for beverage 1. We see a significant increase in sales where there is a catalogue promotion for beverage 1. For all the other beverages, we see higher sales where there is no catalogue promotion – this is surprising. It could be that beverage 1 is the only beverage to ever have a catalogue promotion. Or that when company X did have catalogue promotions for their other beverages, they weren't really appealing.

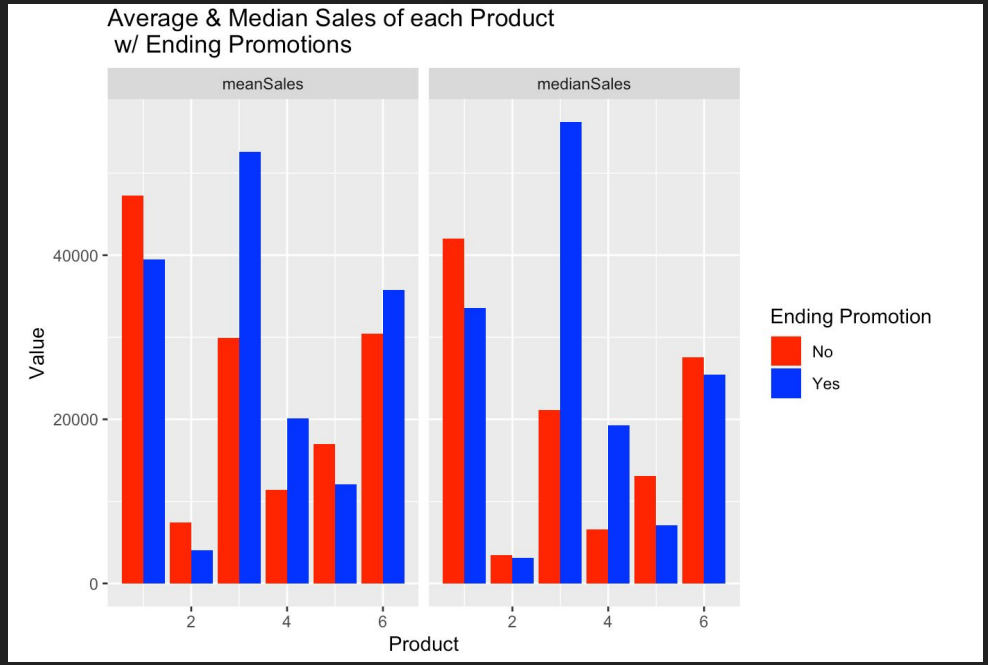
	Product	Catalogue_Promo	Measure	Value
1	1	0	meanSales	43767.125
2	1	0	medianSales	36196.000
3	1	1	meanSales	61893.882
4	1	1	medianSales	57446.000
5	2	0	meanSales	7811.837
6	2	0	medianSales	3444.500
7	2	1	meanSales	3832.500
8	2	1	medianSales	3054.000
9	3	0	meanSales	53431.784
10	3	0	medianSales	57867.000
11	3	1	meanSales	26247.354
12	3	1	medianSales	19818.000
13	4	0	meanSales	21334.259
14	4	0	medianSales	19561.000
15	4	1	meanSales	7657.677
16	4	1	medianSales	5342.000
17	5	0	meanSales	17251.538
18	5	0	medianSales	13577.000
19	5	1	meanSales	12702.257
20	5	1	medianSales	7285.000
21	6	0	meanSales	33832.732
22	6	0	medianSales	27664.000
23	6	1	meanSales	28668.083
24	6	1	medianSales	25017.000



Product Sales w/ Ending Promotions

Ending promotions has a major effect on beverage 3, significantly increasing the number of sales. Same can be said about beverage 4, just to a less degree. Beverage 1 & 5 see lower sales when there is an ending store promotion. Beverage 3 seems to be a popular beverage that is highly affected by ending store promotions. Beverage 1 doesn't seem to be affected by ending promotions to the same degree but still is one of company x's best sellers.

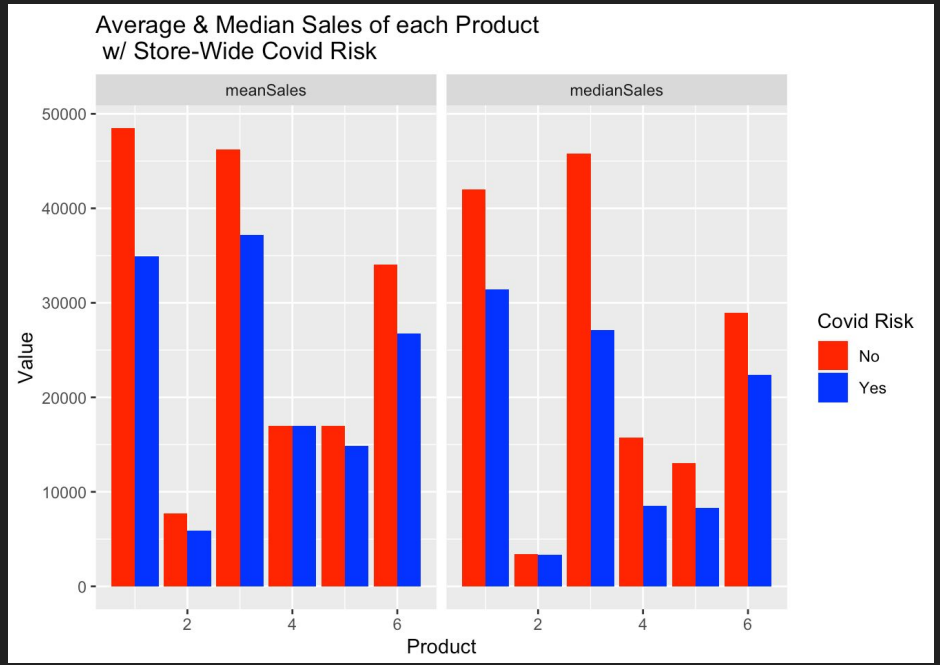
	Product	Store_End_Promo	Measure	Value
1	1	0	meanSales	47282.240
2	1	0	medianSales	42015.500
3	1	1	meanSales	39470.804
4	1	1	medianSales	33604.000
5	2	0	meanSales	7456.272
6	2	0	medianSales	3411.000
7	2	1	meanSales	4019.889
8	2	1	medianSales	3081.000
9	3	0	meanSales	29899.986
10	3	0	medianSales	21108.000
11	3	1	meanSales	52623.126
12	3	1	medianSales	56218.000
13	4	0	meanSales	11408.054
14	4	0	medianSales	6578.500
15	4	1	meanSales	20146.269
16	4	1	medianSales	19286.000
17	5	0	meanSales	17002.044
18	5	0	medianSales	13131.500
19	5	1	meanSales	12078.045
20	5	1	medianSales	7106.000
21	6	0	meanSales	30418.226
22	6	0	medianSales	27569.000
23	6	1	meanSales	35762.933
24	6	1	medianSales	25478.000



Product Sales w/ Store Covid Risk

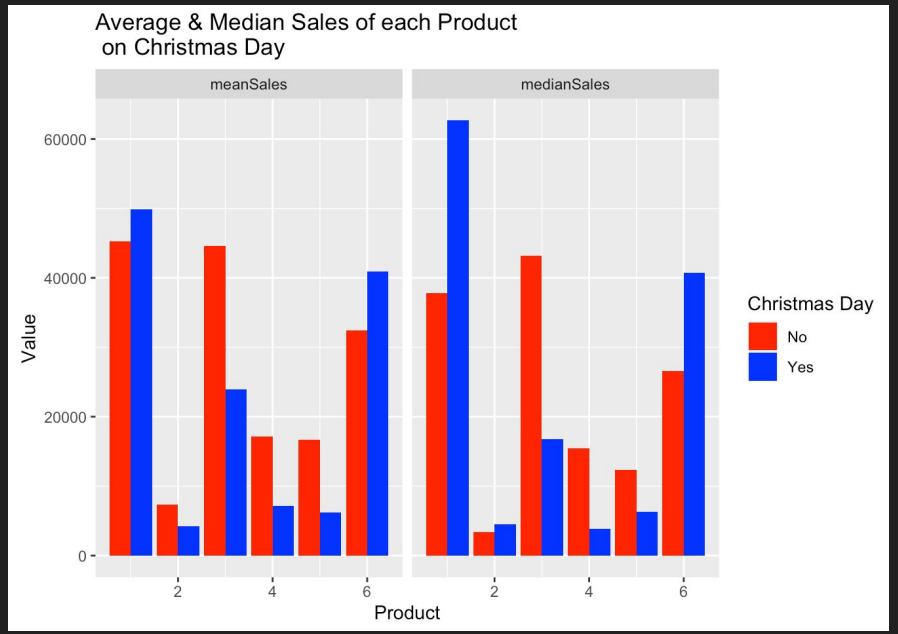
This one makes sense. We see a overall decrease in sales numbers when the store is a covid-risk. I recall when we experienced covid-19, we were told to stay at home, so a reduction in the number at the supermarkets could result in a reduce in sales in beverages for company x. Beverage 2 doesn't see much effect on sales based on covid risk, possible because it wasn't a good seller anyway. Beverage 1 & 3, Company X's best sellers were certainly hit & reduced in the number of sales where the store was a covid risk.

	Product	Covid_Flag	Measure	Value
1	1	0	meanSales	48477.747
2	1	0	medianSales	41970.000
3	1	1	meanSales	34888.809
4	1	1	medianSales	31406.000
5	2	0	meanSales	7724.255
6	2	0	medianSales	3395.000
7	2	1	meanSales	5903.064
8	2	1	medianSales	3339.000
9	3	0	meanSales	46235.671
10	3	0	medianSales	45801.500
11	3	1	meanSales	37150.977
12	3	1	medianSales	27106.000
13	4	0	meanSales	16970.306
14	4	0	medianSales	15705.000
15	4	1	meanSales	16997.298
16	4	1	medianSales	8486.000
17	5	0	meanSales	16952.771
18	5	0	medianSales	13048.000
19	5	1	meanSales	14861.787
20	5	1	medianSales	8282.000
21	6	0	meanSales	34067.980
22	6	0	medianSales	28964.000
23	6	1	meanSales	26752.975
24	6	1	medianSales	22382.000



Product Sales on Christmas Day

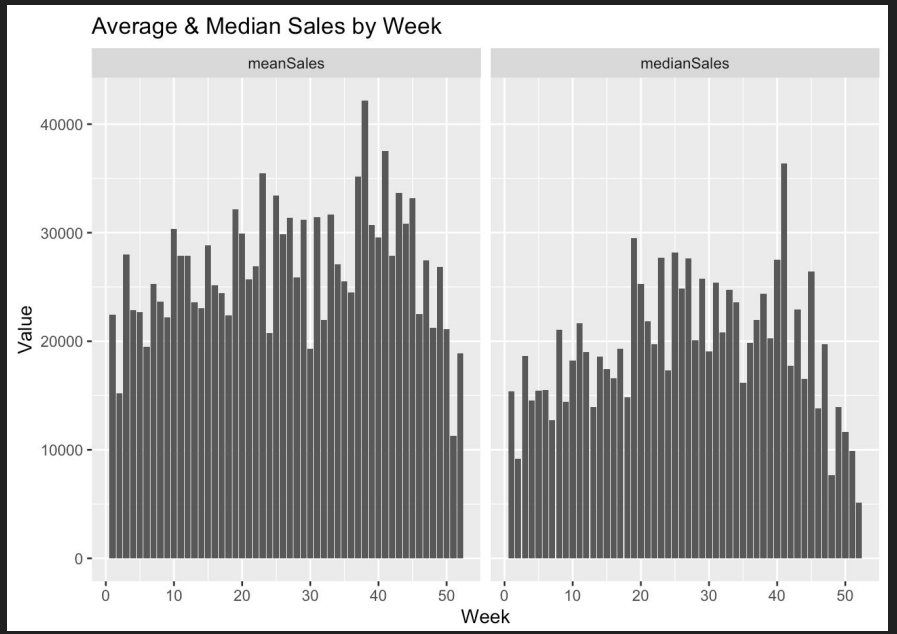
Beverage 1 sees significant increase in median sales on christmas day. Same can be said for beverage 6. Beverage 3 sees significant decrease in median sales on christmas day. Same can be said for beverage 4 & 5. Beverage 1 must be a christmas staple like a holiday champagne or some christmas-themed drink. That makes me think what Beverage 3 might be. It could be more of a summer type of beverage, maybe like a lemonade.

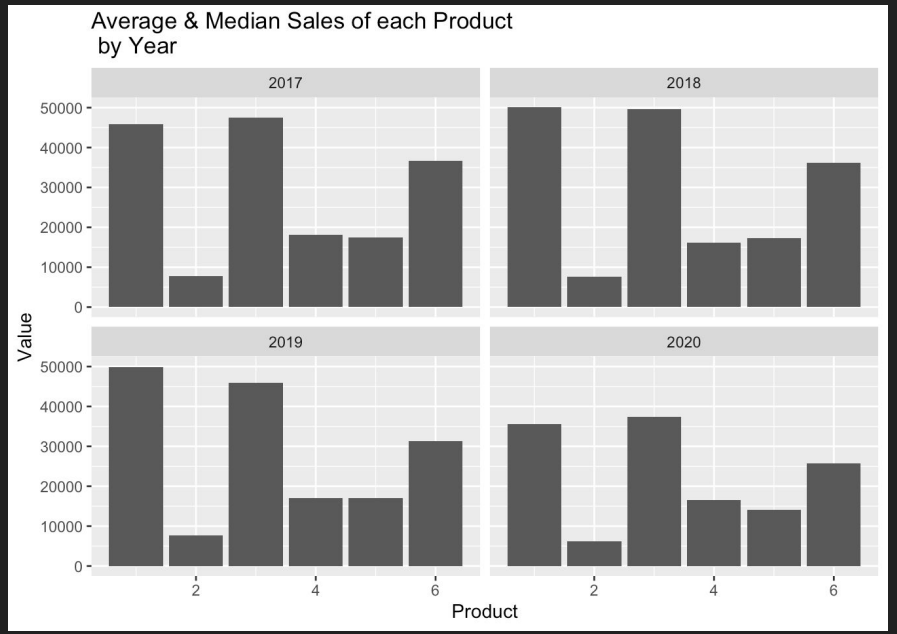


	Product	Christmas	Measure	Value
1	1	0	meanSales	45208.228
2	1	0	medianSales	37845.000
3	1	1	meanSales	49831.500
4	1	1	medianSales	62685.000
5	2	0	meanSales	7367.055
6	2	0	medianSales	3392.500
7	2	1	meanSales	4185.250
8	2	1	medianSales	4525.500
9	3	0	meanSales	44566.333
10	3	0	medianSales	43147.500
11	3	1	meanSales	23928.250
12	3	1	medianSales	16768.500
13	4	0	meanSales	17172.610
14	4	0	medianSales	15488.500
15	4	1	meanSales	7172.250
16	4	1	medianSales	3887.500
17	5	0	meanSales	16675.575
18	5	0	medianSales	12307.500
19	5	1	meanSales	6243.500
20	5	1	medianSales	6305.500
21	6	0	meanSales	32393.005
22	6	0	medianSales	26567.000
23	6	1	meanSales	40941.333
24	6	1	medianSales	40726.000

Product Sales based on Week & Year

Based on the weekly sales, it does not seem that holidays have a major effect on the number of sales. Matter of fact, at the end of the year, when there is the most holidays, it has the lowest sales. This is surprising. It also seems that there yearly sales relatively similar for each beverage. Can't be said about 2020 though, that was when Covid stay-at-home policy was in full effect. This could explain the reduced sales, especially for beverages 1, 3, & 6.





THANK YOU