

How to deal with COVID 19 data?

Prepared as part of the

Industrial Engineering Capstone Experience
University of Arkansas

Submitted on April 7, 2022, by

Carolina Virreira

Supported by Nestlé USA

Edosa Aibangbee, Ecommerce Manager for Frozen Meals
James Gairhan, Category Sales

Technical Report

My team has the pleasure to be currently working with Nestle as our industry partner for 21-22 IE capstone experience. The project we are working on is based on a forecasting tool created by last year's capstone team. We got the tool "broken" at the beginning of the project; however, the tool was not broken. The only issue with the tool was that some lines of code needed to be modified for the user to be able to run a time series forecast. The overall goal this year is to improve the user usage of the tool as well as improving the accuracy of the forecast. Nestlé believes there is still room for improvement in this tool, especially with the user friendliness side of it. The tool was created using R studio and R shinny and has over 600 lines of very complex coding. This tool reads large csv files of data which are hard to manipulate because of the size they have.

An issue that was faced on the project was how to give nestle an accurate forecast taking 2020 into consideration. COVID 19 affected 2020 sales dramatically, leading to a forecast that does not follow the historical trend. We were provided with historical data since 2017 up to 2021 on a csv file called brandlevel.csv. This data is divided into Walmart weeks, store number, category, brand, POS quantity and POS sales. Nestle works with Walmart weeks which are 52 weeks each year. The brand level file contains over 12,000,000 cells, which is not compatible with excel. R studio reads files as big as our Brandlevel.csv file, however I believe that manipulating data is easier to do in excel. So, I decided to find a way in which I could manipulate this data in excel. I did some research, and an interesting article came up. The article was about large sets of data splitted into multiple smaller workbooks. I downloaded a csv splittler called MacUncle csv splitter.

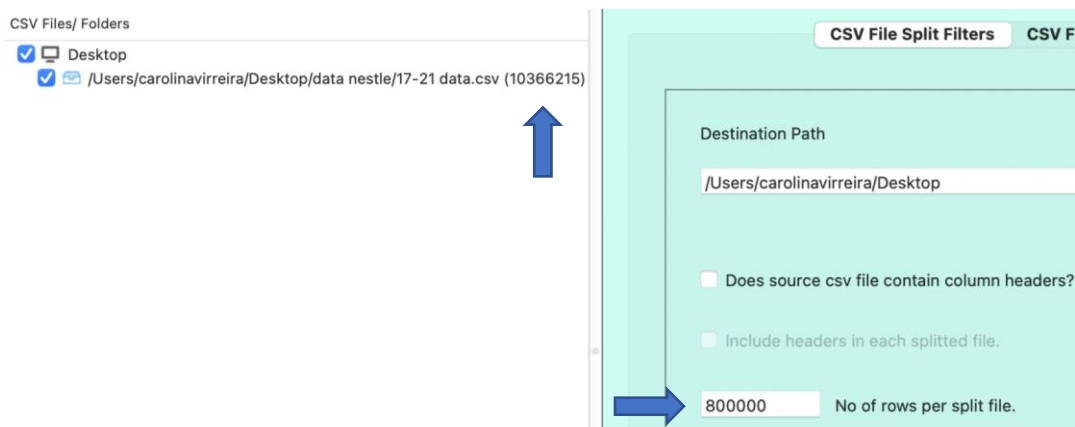



Figure 1. This is an example of the brandlevel.csv file being splitted. The file contains 10,366,215 rows in total, which were splitted into 13 .csv files of 800,000 rows.

A disadvantage of this csv splitter is that each excel workbook has the data unorganized. Every Walmart week contains around 32,280 rows and to analyze the data correctly each workbook needed to end with the last day of any Walmart week. Due to this issue, I had to find manually when a week ended and copy and paste the rest of the next week to the beginning of the next workbook. After doing this, I had a completely new data set in excel that was ready to be manipulated and analyzed.

Subsequently, I worked on which approach I should take to analyze COVID-19 year. I came up with two solutions. The first solution was to compress each year of data. Instead of having 800,000 rows per file, which means around 2,400,00 rows for each year, I ended up with 52 rows for each year we were given. On the new dataset, each row now represents each Walmart week, this means that the data was compressed into a dataset divided by week by category.



	A	B	C	D	E	F
1	WM_Week	StorNbr	NWT_Category	ConsBrand	POSQty	POSSales
2	201701	1	BAKING	NESTLE	384	753.48
3	201701	1	CHILLED CRE	COFFEEMAT	468	1506.34
4	201701	1	CHILLED CRE	NESQUIK	80	102.4
5	201701	1	FROZEN ENT	HOT POCKET	303	1323.96
6	201701	1	FROZEN ENT	LEAN POCKE	57	122.4
7	201701	1	FROZEN ME/	LEAN CUISIN	490	1183
8	201701	1	FROZEN ME/	STOUFFERS	547	2663.45
9	201701	2	BAKING	NESTLE	297	719.36
10	201701	2	CHILLED CRE	COFFEEMAT	626	2128.03
11	201701	2	CHILLED CRE	NESQUIK	46	58.88
12	201701	2	FROZEN ENT	HOT POCKET	376	1775.22
13	201701	2	FROZEN ENT	LEAN POCKE	18	60.78
14	201701	2	FROZEN ME/	LEAN CUISIN	567	1357.5
15	201701	2	FROZEN ME/	STOUFFERS	677	3118.02
16	201701	3	BAKING	NESTLE	180	355.8
17	201701	3	CHILLED CRE	COFFEEMAT	358	1237.94
18	201701	3	CHILLED CRE	NESQUIK	72	84.96
19	201701	3	FROZEN ENT	HOT POCKET	268	1363.99

	A	B	C	D	E	F
1	category					
2	WM week	BAKING	CHILLED CRE	FROZEN ENT	FROZEN ME	Grand Total
3	2017-01	958863	2118655	1273701	3985509	8336728
4	2017-02	1045437	2091583	1413601	3988156	8538777
5	2017-03	1105991	2063963	1347205	3968325	8485484
6	2017-04	891344	2092065	1303472	3880538	8167419
7	2017-05	940986	2145795	1345927	4048668	8481376
8	2017-06	1016994	2176767	1452412	4178667	8824840
9	2017-07	1082010	2132308	1397675	3974916	8586909
10	2017-08	1019367	2049674	1222110	3872979	8164130
11	2017-09	1014619	2002204	1121574	3769308	7907705
12	2017-10	1126172	2104061	1347124	4088066	8665423
13	2017-11	1265680	2110860	1305828	3866977	8549345
14	2017-12	918206	1922517	1127275	3500543	7468541
15	2017-13	748918	1965148	1092433	3720024	7526523
16	2017-14	826680	2055404	1224552	3897145	8003781
17	2017-15	856484	1999808	1248404	3809198	7913894
18	2017-16	816931	1907201	1132403	3566222	7422757
19	2017-17	828784	1957415	1091126	3567807	7445132

Figure 2. Previous dataset compared to the new data set with weighted values.

To simplify each Walmart week, I made sure to move a complete year into one sheet, select the data for a complete year, and by using excel pivot chart feature I took an average of each week. This led me to a very summarized excel workbook that contains 2017-2021 data weighted on separate sheets for each year, making the analysis easier and with very few places on which errors could be made. With this I was able to start the first approach, which was to multiply each category by a percent change for 2019-2021. I made the percent change for each year in a new sheet. I decided to make a percent change of 2019-2021 because this will show the impact covid had in 2021 and how 2020 is different from

2019 which is a regular year. To make this percent change I used a basic percent change formula.

$$\% \text{ change} = \left| \left(\frac{\text{final} - \text{initial}}{\text{initial}} \right) * 100 \right|$$

I believe this is a reasonable approach because the data will be multiplied by a percentage that was gathered from historical data with normal trends as well with a year that had irregular sales because of an unforeseen event. Since we want to take events such as a pandemic into account, COVID-19 gave a different but real trend to the forecast. The results of this approach were positive, a more normal trend came up in the forecast. However, this data needs further division to be able to forecast it by week, by category, by brand.

As a second solution and as part of checking that my assumptions about the data were correct, I decided to create a hypothetical 2020 year. I based this new year on averaging all of the historical data except 2020 and 2021. I did this very simple by using an average formula locked and just drag and drop it into a new excel sheet. Then with those averages per year, I made one last average to make sure I had a new year based on real sales of normal trends. I finished up this approach by graphing the old 2020 data and the new one.

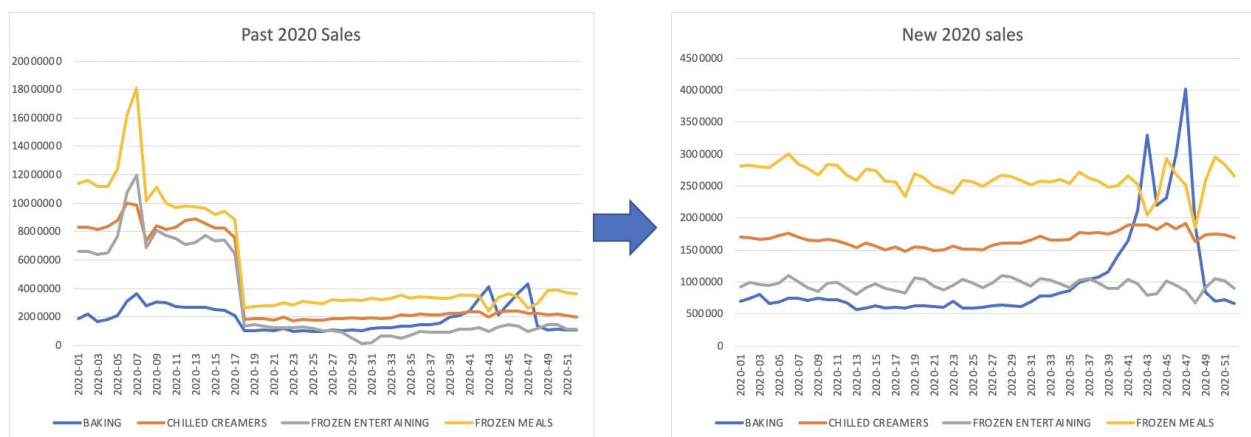


Figure 3. Comparison of old and new 2020 data.

With the visuals, I can conclude that the assumptions about making a hypothetical year with the historical data we were given by Nestle was correct. The new 2020 sales show a normal trend, with noise at the end of the year which indicates high sells around thanksgiving and Christmas. This analysis is still in progress, since creating a new year from historical data is a correct approach but is not the most accurate. As a group we

decided to continue working with my first approach and save the second approach in case the first one shows problems.