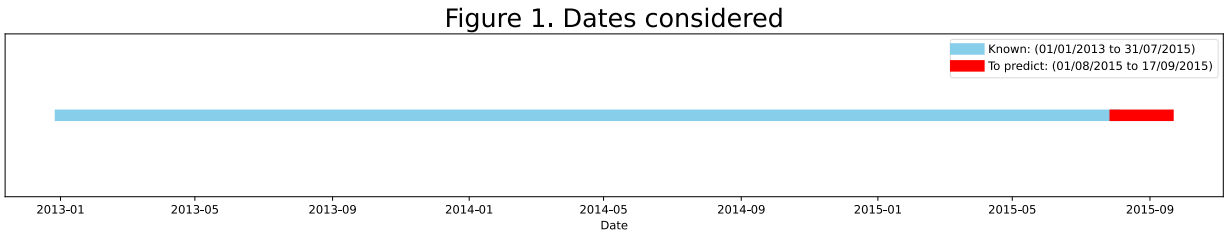


Time Series Forecasting for ROSSMANN stores in Germany (2013-2015).

Sergio Chavez Lazo, MSc

The goal of this project was to predict the sales of ROSSMANN stores in Germany for a then unknown future. ROSSMANN is one of Europe's largest pharmacy chains with 60,500 employees in Europe and 4,514 branches. At the time the data for this project was collected (September 2015), ROSSMANN had more than 1100 stores in Germany. The forecasting process used data related to the specifics of each shop and daily information on its sales performance and other eventualities (e.g., holidays). Figure 1 shows that the daily sales record of 2 years and 7 months were available as potential inputs to predict 48 daily sales performance.



After cleaning, processing, and transforming the data, a successful sales forecasting model with a percentage error less than 2% was developed.

I. Validation

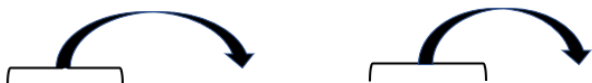
The first step was to validate that the measurement format of the variables was adequate. Specifically, the recording of dates was standardized (YYYY-MM-DD) to facilitate comparisons, future processing, and modeling. Also, the variables related to the type of the store and level of assortment were transformed. StoreType column was converted to category since there is not hierarchy established between types of stores that could allow any kind of logic order. Assortment was converted to integer considering that there is a logic sequence behind the types.

Another aspect of validation was the construction of new date variables. The store database disaggregated in date columns the variables related to the start of the implementation of Promo 2 and the opening of a rival store around it. These columns were processed and joined in such a way as to have an exact reference date with which to do further operations (e.g. determine whether by a specific date, store 'X' already had competition or Promo 2 available). In that sense, two assumptions had to be made:

- Given that the available variables did not allow to identify the exact opening day of the rival store (columns referred to month and year), it was assumed that these events happened on the first day of the month.
- Since the variables did not allow the exact day of adoption of Promo 2 to be identified (columns referred to the week of the year and the year), it was assumed that the event was the first day (Monday) of the specified week.

Finally, the months in which email campaigns related to Promo 2 of each shop were re-released were separated into different columns and converted to date format.

Figure 2. Date transformation (CompetitionSince & Promo2Since)

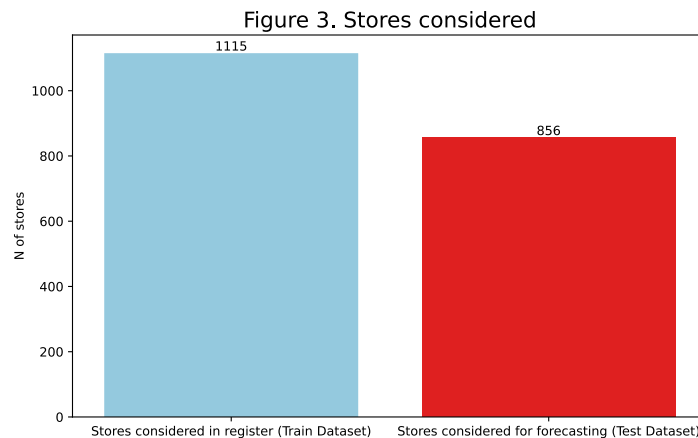


Store	CompetitionOpen SinceMonth	CompetitionOpen SinceYear	Competition Since	Promo2Since Week	Promo2Since Year	Promo2Since	PI-Month1	PI-Month2	PI-Month3	PI-Month4
1	12	2013	1/12/2013	13	2012	19/03/2012	January	April	July	October
2	3	2014	1/3/2014	12	2014	24/3/2014	March	June	September	December
3	4	2010	1/4/2010	3	2014	16/01/2014	January	April	July	October
4	5	2015	1/5/2015	2	2011	10/1/2014	February	May	Agosto	November
5	6	2008	1/6/2015	21	2010	24/05/2010	February	May	Agosto	November

II. Basic exploration and filtering

The first question to explore was: for which stores were we predicting? ROSSMANN is an international chain with more than 4,000 stores today. Each store has several special characteristics and circumstances relevant to his sales performance. In that sense, it would not be rigorous to try to predict the sales of one store based on the characteristics and performance of others. Therefore, the first step was to explore which stores were being considered for the forecasting.

Figure 3 revealed that the historical record considered a larger number of stores than those for which forecasting was desired. Therefore, the first major decision was to filter out from the training base those stores that were not included in the test database. Again, the motivation was to maintain rigor: avoid forecasting performance of stores based on the historical performance of different ones.



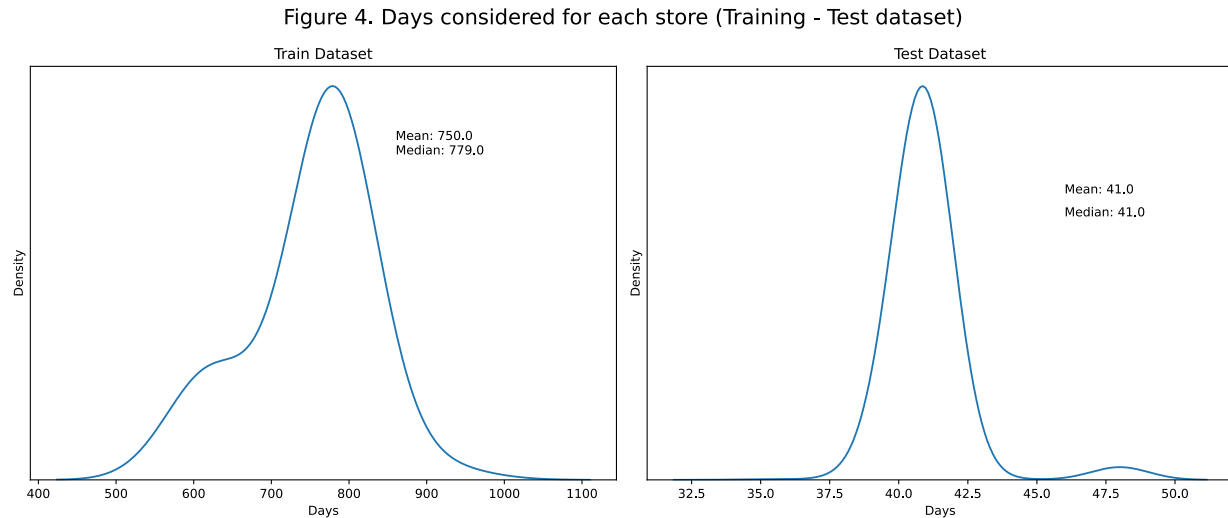
The next step was to evaluate the relevance of considering all rows/dates in the train dataset. It was assumed that a row/date was an interesting input for the project as long as it records sales information. In this sense, we identified which rows did not record sales and proceeded to explore the possible explanations. What was found was:

- While there were no missing records (NaN) for sales, there were a total of 131289 dates where stores recorded 0 sales.
- Of the 131289 dates with a record of 0 sales, 131246 (99.9%) were also dates on which the stores were closed.

Although they were not formally missing values, the relationship between the two variables suggested that the absence of sales was due to a case of Missing at Random (MAR). That is, the absence depends almost exclusively

on another variable. This relationship was logically understandable: if a store does not open, it cannot sell. Therefore, it was decided to filter out from the training and test database all dates on which stores did not open¹

After the filtration was carried out, the number of dates available for each store was analyzed. Figure 4 shows the number of dates/rows available for each store in the train and test datasets.



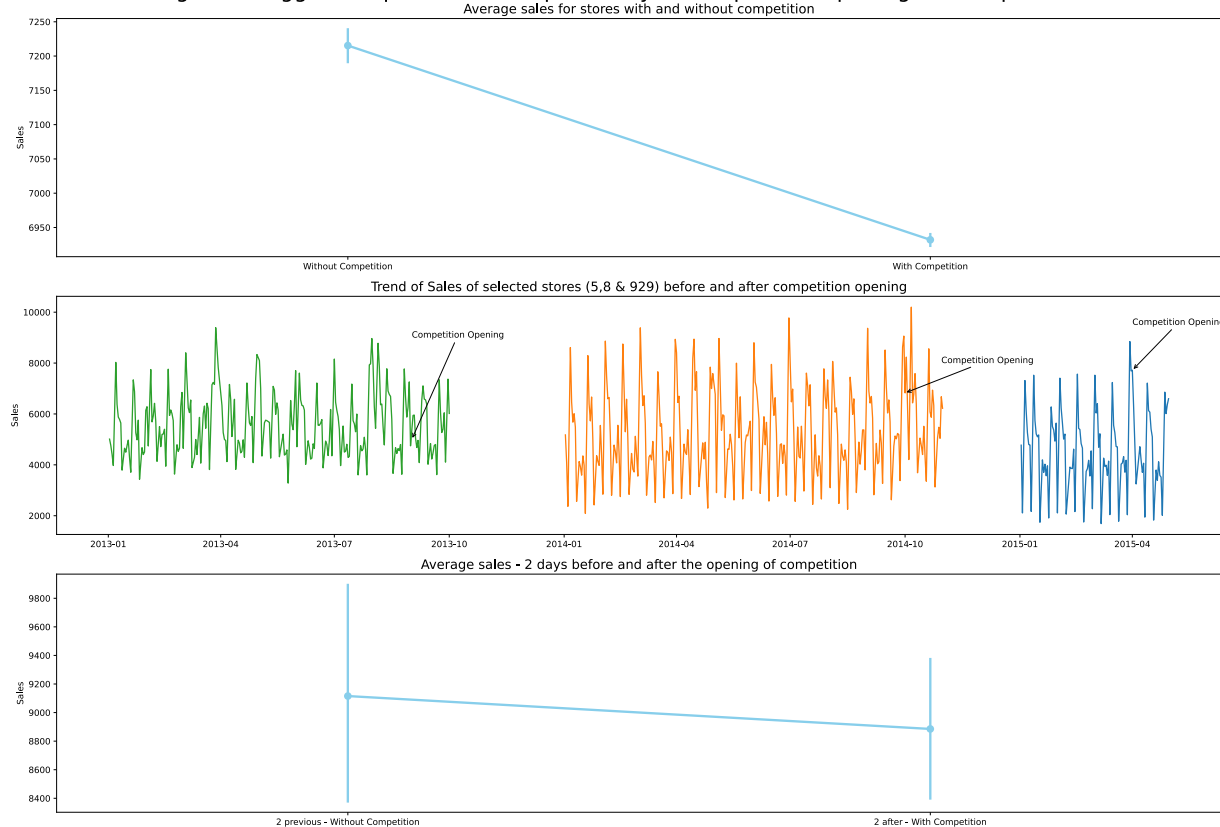
III. Missing values and imputations.

The highlight of this part was to identify that there were many stores (315) that recorded distance from their competitors, but no values for the opening date of their competitors. This relationship was the only one found among NAN's that was inconsistent since it was identified that the missing values for Promo2 related variables were exactly those stores where Promo2 was not available in the first place. On the other hand, it was assumed that those stores that did not record the distance from the competition simply did not have competition identified around them.

The imputation of opening dates for competitors required two approaches. Given that this was a variable of date and stores in different locations, the classic imputations of mean and median were not recommended as they were meaningless (the opening date of a nearby rival store does not suggest when and where another will be opened). For this reason, the initial hypothesis was to establish that a good imputation would be the date when sales began to drop. To confirm this proposal, the sales trends were evaluated for the stores where there was information about the opening date of competition. Figure 4 suggests that this was inadequate. Although subfigure 4.1 shows that the average sales in the post-opening stage is lower, subfigure 4.2 indicates that the opening dates did not coincide with the date in which some stores registered their least number of sales. Moreover, it can be seen in 4.3 that the sales trend does not change in the short term -2 days before and after- for all stores. Therefore, we rejected this form of imputation.

¹ On the test dataset there are few rows with formal missing values for the column 'Open'. Further exploration shows that those rows belong all to the store number 622 and are from the dates between 5-17 of September 2015. After checking that the median (11) of number of open days for those stores in the same condition (same period of days for the same month and year considering also the same condition of non-state nor school holiday) it was decided that the missing values should be imputed as Open days (1).

Figure 5. Biggest drop in sales as a possibility of competition opening date imputation



More practical approach was chosen then. Using the 761 shops that did have a date for the attribute, it was found that more than 75% of the time the rival shop had opened before 01/01/2013. Considering that the historical database provided only has information from that date and that the general sense of imputing data in this case was to observe whether for a particular date a shop had a nearby rival or not, it was considered feasible to impute 01/01/2013 for missing values. It is recognized that it is highly likely that the actual date is not exactly that, but the trend indicates that it is even more unlikely that the date is later. Therefore, in view of the practical objectives of the use of the variable, we proceeded in that way.

When it was decided not to include dates on which stores were closed, a couple of days were identified on which, despite being operational, some shops recorded no sales. It is assumed that this is perfectly possible ('bad business days'). However, it was noted that on 2 of these days, customers did register. This was problematic if a customer is assumed to be a buyer. Therefore, we calculated the average sales per customer for the last 15 days for the shops with this abnormality (stores 948 and 1000) and imputed on that basis.

IV. Exploratory Data Analysis

First, we analyzed the target variables and their fluctuation over time. Figure 6 reveals what was the distribution of average sales and customers. Indeed, in both cases, a distribution like a normal distribution with a skew to the right was observed. As expected, this indicates that there were dates where sales and customer volumes were much higher than expected.

To optimize the forecasting model, sales was included in their logarithmic version². However, in the subsequent descriptive analyses of sales, the original format was maintained to facilitate interpretation.

Figure 6. Distribution and Relationship of Daily sales and customers

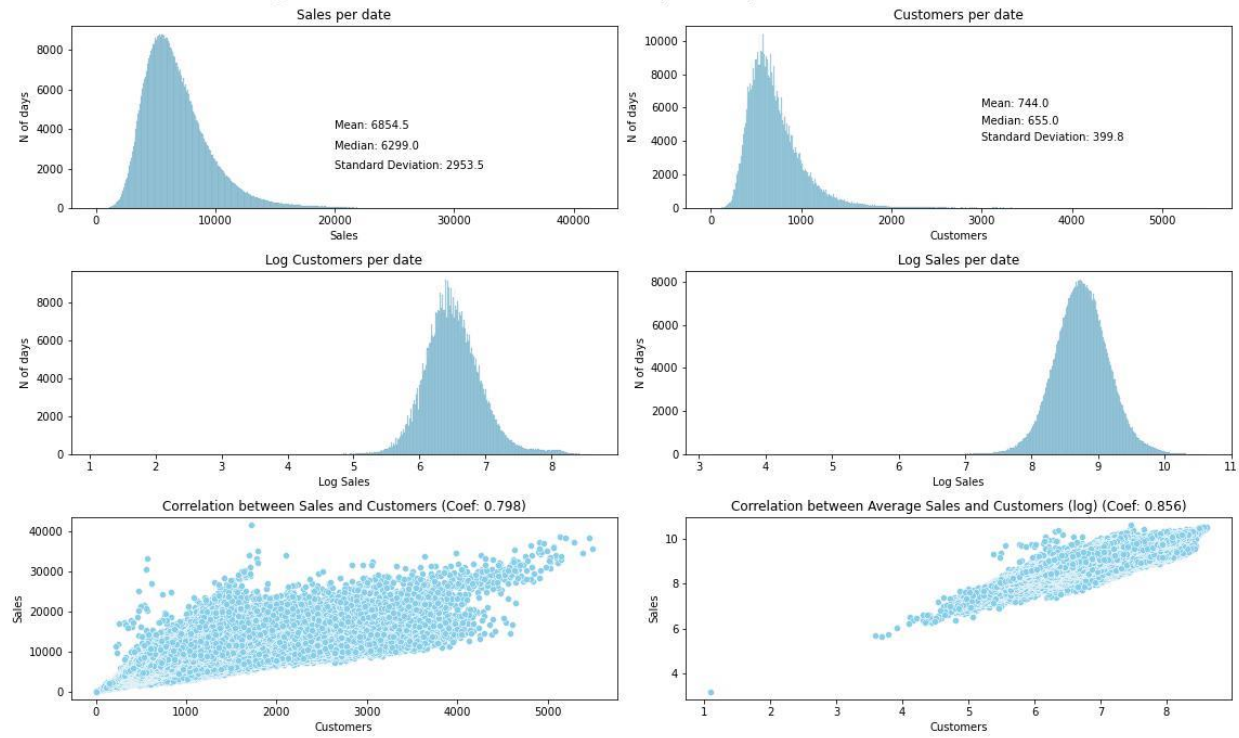
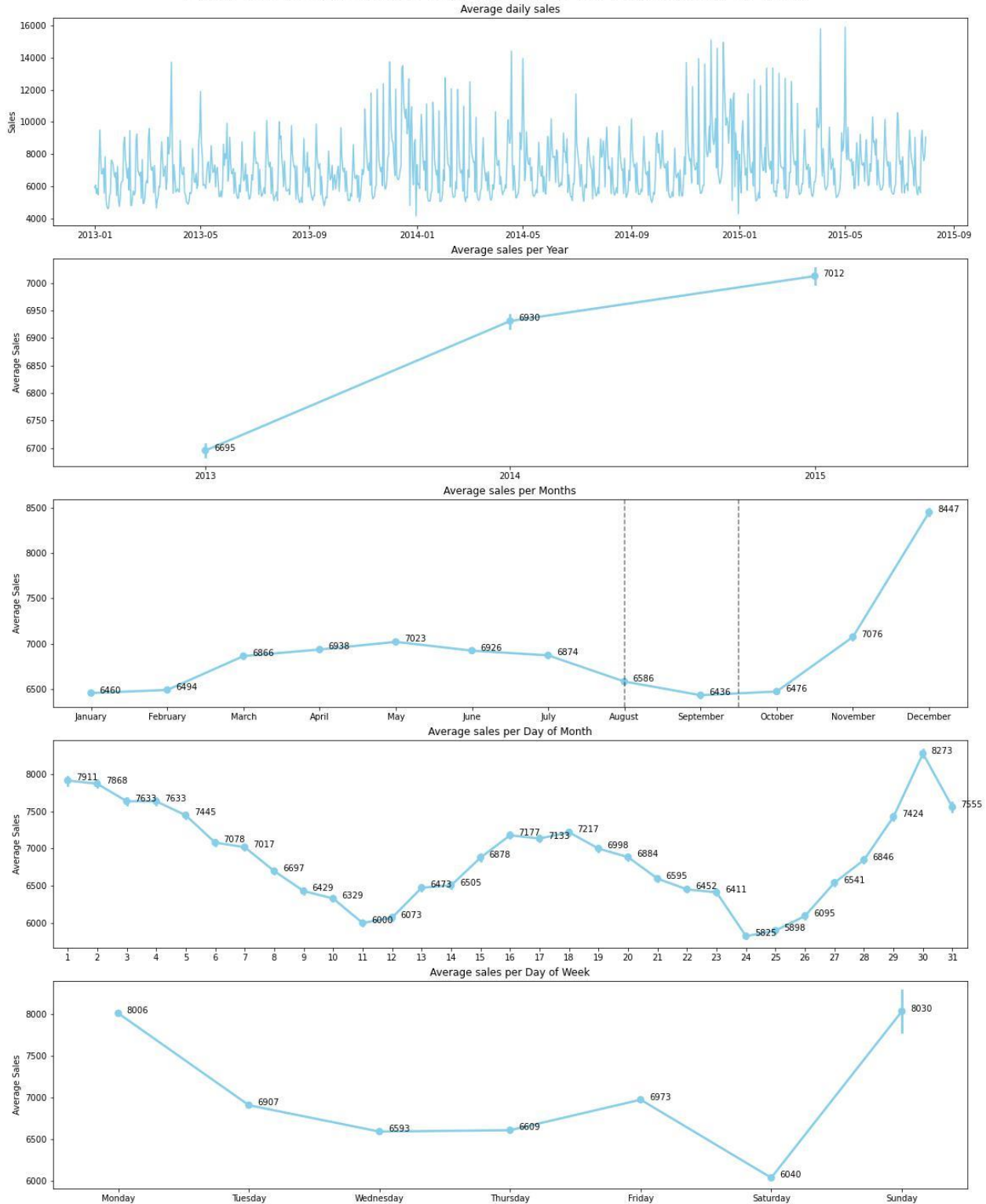


Figure 7 shows the distribution of average sales according to different time periods:

- At the daily level, no clear trend could be observed. However, it did suggest an increased volatility since the beginning of 2014 (very high peaks and lows close together).
- At the annual level, the average level of sales had improved significantly.
- At the monthly level, there was a fluctuation favouring the second quarter months (April, May and June) and, above all, December. The period to be predicted (August and mid-September) had an intermediate-low average performance.
- The fluctuation within each month suggested 2 clear peaks of sales: at the beginning and at the end. The reason for this is likely to be since salary payments are usually made at the close of each month.
- The weekly fluctuation indicated that the best days for sales were at the end and beginning. After that, the average remained similar with a slight drop on Saturdays.

² This ensures a more normal behavior and reduces the bias to the right.

Figure 7. Sales according to Year, Month, Day of Month and Day of Week



The distribution of sales was evaluated in relation to all other available variables. The most important findings are highlighted below.

a. Store Type.

Only 0.01% of stores belong to category B. This was important because these would appear to be better positioned stores given that they have a much higher average than the rest. The difference between the other types of stores is much smaller (Figure A1)

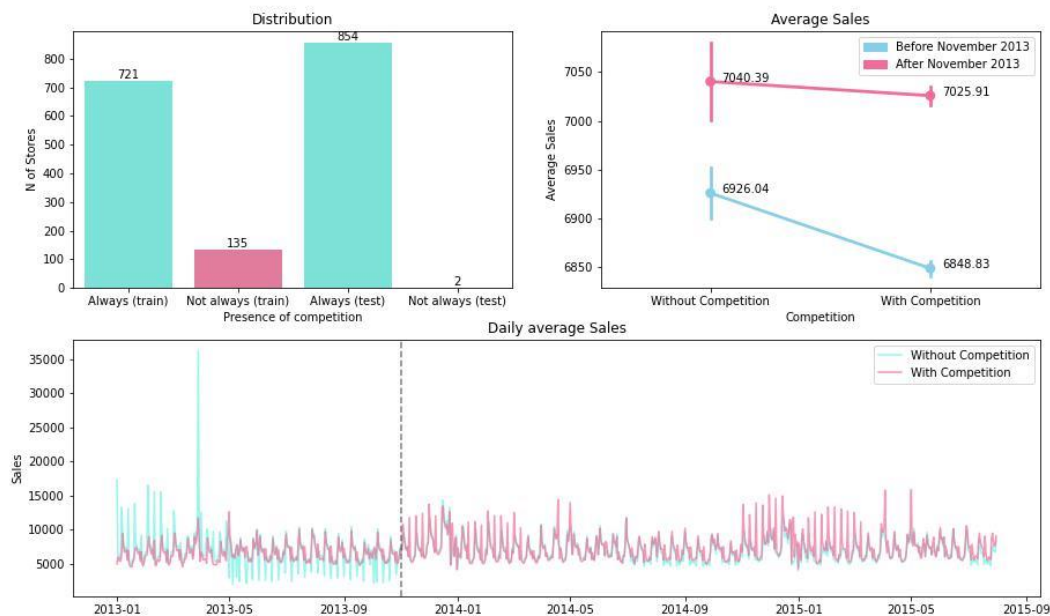
b. Competition

It was observed that more than 99% of stores already had competition for the period that we wanted to forecast. This is a very important aspect because it reveals that the variable will practically be applied as a constant in the forecasting process.

Secondly, we noticed that there was a significant difference between the sales of stores with and without competition in the training database. This aspect complies with the basic intuition that a store without competition will have more sales. However, when analyzed the pattern of sales over time we noticed that the difference in sales just existed until the end of 2013. After that period, the lines of average sales are juxtaposed showing that there were no significant differences between the two groups.

Due to the low variability of the variable and its little relevance to sales closer to the dates we were interested in forecasting, it was decided not to consider this feature in the final model.

Figure 8. Sales according to Presence of Competition



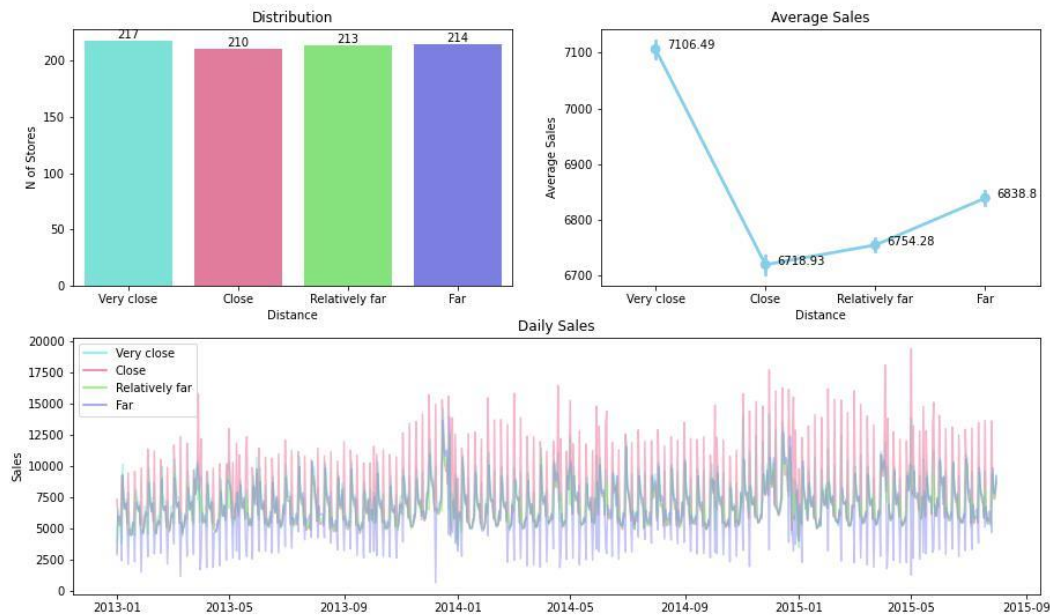
c. Competition Distance.

To optimize the EDA it was decided to categorize the variable 'Distance' according to the quartiles of distance of competition: very close (< 750 meters), close (< 2350 meters), relatively far (< 6880 metres) and far (more than 6880 metres). The first thing worth noting in Figure 9 was that the stores with very close competition were those with the best average sales. Although this may seem counterintuitive, the truth is that it is probably a spurious relationship: a relationship that would be explained by an unconsidered third variable. The hypothesis was that the stores that have very close competition were those that were also better

located (e.g., in a shopping center). Thus, it was the good location that would cause the store to have more competition and sales without meaning that the presence of competition directly generates more sales.

Similarly, it can be understood why stores with close and relatively distant competition lower average sales: they were in less 'strategic' places. However, when it comes to very distant competition, sales pick up again, suggesting that the location-sales relationship is not linear, but rather convex.

Figure 9. Sales according to Competition distance



Special Dates

The period we wanted to forecast included just public holidays. However, we found a significant difference in the average sales in those holidays compared to non-holidays. On the other hand, we noticed that all stores had been affected by school closures in both datasets and that the difference in the average for these days was significant. Therefore, we kept both variables for the model (Figure A3)

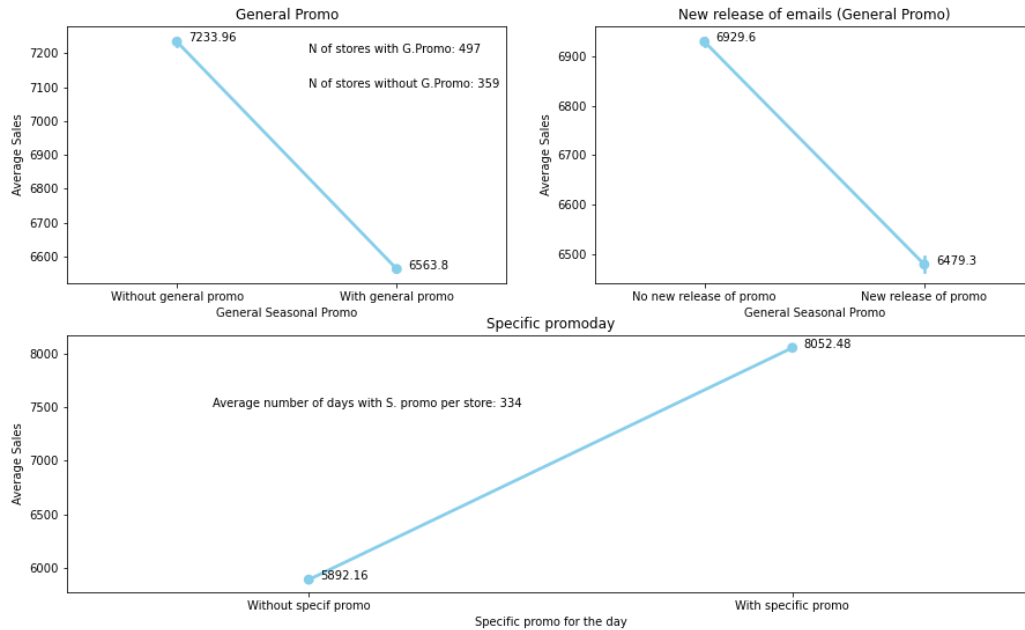
Promotions

Figure 10 shows the relationship between the two types of promotions with the average sales. Those stores that had a general promotion throughout the year had lower average sales. It was observed that in the launch month of those promotions, sales were also significantly lower. All this would suggest that this strategy was not only ineffective, but in some way counterproductive to the objectives of ROSSMANN.

On the other hand, the trend was contrary with the promotions some stores adopt for the day. The last graph shows that sales on those days were significantly higher. Thus, in this case the sales strategy of including a daily promotion would be effective.

Since both variables showed significant differences in sales, we kept them in the predictive model.

Figure 10. Sales according to Promos



V. Feature selection and transformation

Before starting the sales training and forecasting process, the format of the variables to be used was adapted to the algorithm and reviewed. Specifically:

- The ordinal variable Distance was recategorized so that it was numerical (0: Near, 3: Far away).
- One Hot Encoder was used to generate dummy variables for categorical characteristics (Type of holiday and Type of shop).
- Considering that all variables had numerical format (dichotomous/scalar), a correlation matrix was performed to assess and control multicollinearity.
- Considering the numerical format of the features, all of them were standardized using Z scores to regularize the variance and optimized the forecasting process.

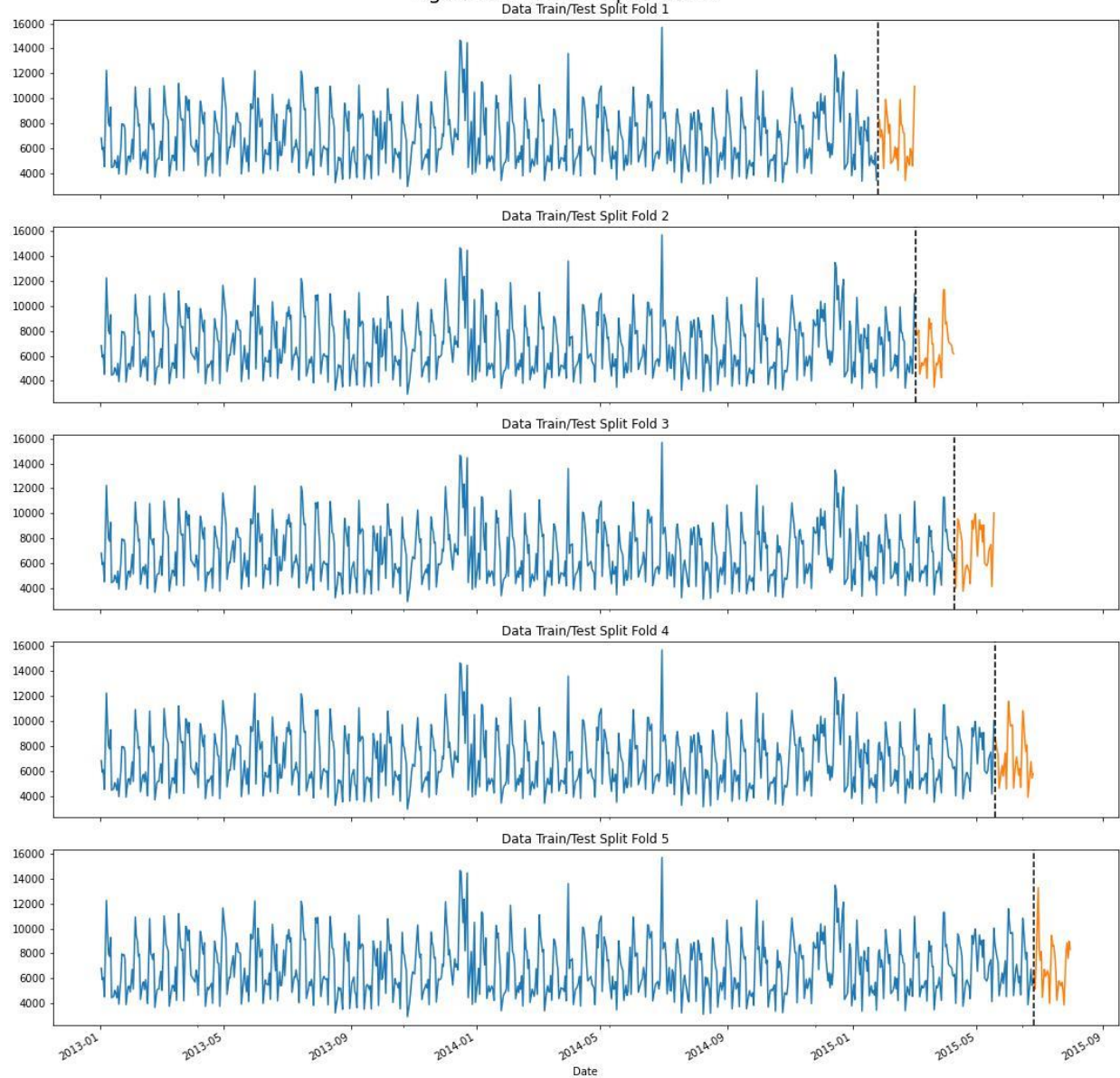
Finally, the training and test databases were filtered to contain only the selected/transformed features and time variables (day of week, day of month, month and year as separated columns).

VI. Model training, optimization, and forecasting

The statistical model to forecast sales was XGBoostRegressor. XGBoost starts from a general sales prediction for all records and progressively optimizes it based on decision trees. In this sense, it is based on a series of iterations that continue if they get closer to the real value of sales (for this reason, it does not require that the variable to be predicted follow such a fixed pattern as a linear regression). The amount of data available facilitated iteration and constant improvement of the algorithm.

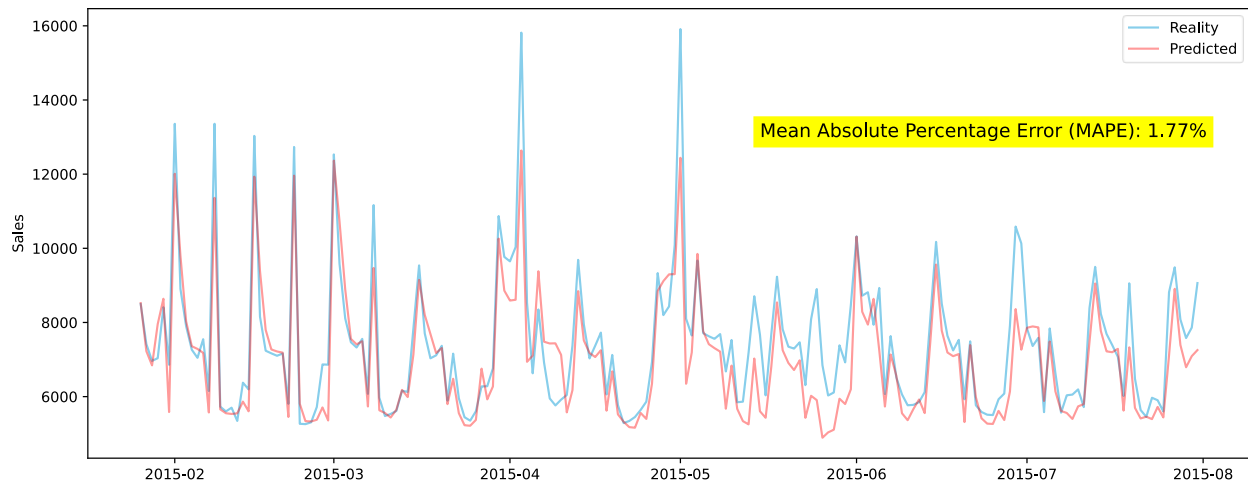
Being Time Series, the segmentation of the data for Cross Validation could not occur randomly. The distribution was to be progressive, always taking data from the past as training to predict the future. Therefore, the separation of the data followed the pattern shown in Figure 11.

Figure 11. Pattern split for CV



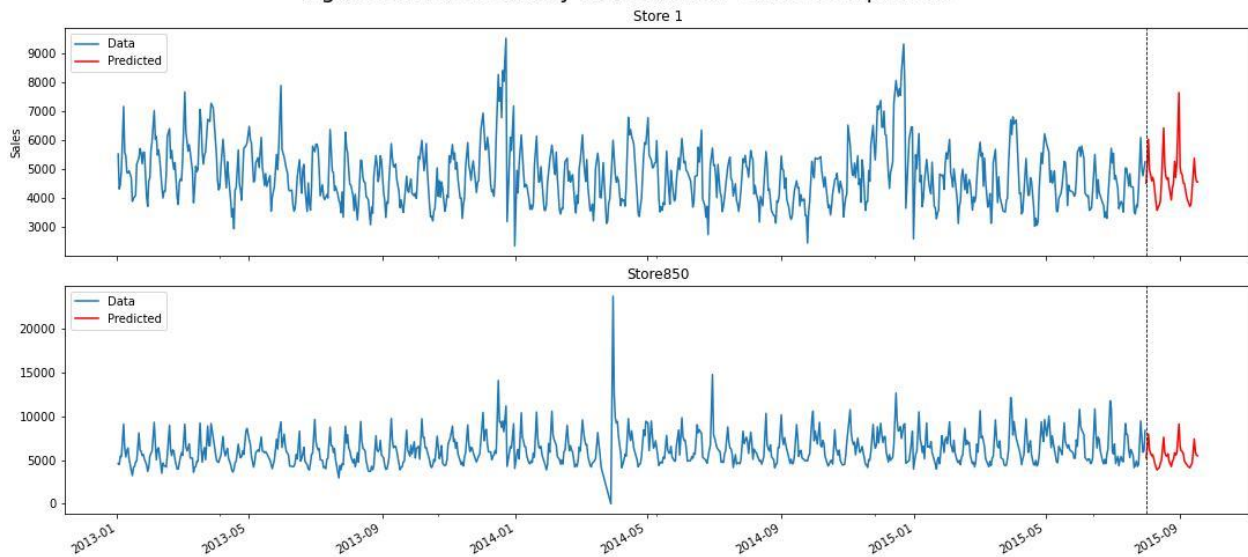
After training and optimizing the hyperparameters of the model, a general average error of 1.77% was obtained. The error metric used was Mean Absolute Percentage Error (MAPE) and it tells us how far our predictions are from the actual sales values in average percentage. Considering the volatility and nature of sales, the model is considered accurate and useful for the goal of the project.

Figure 12. Average Sales per Date - Reality vs Predicted



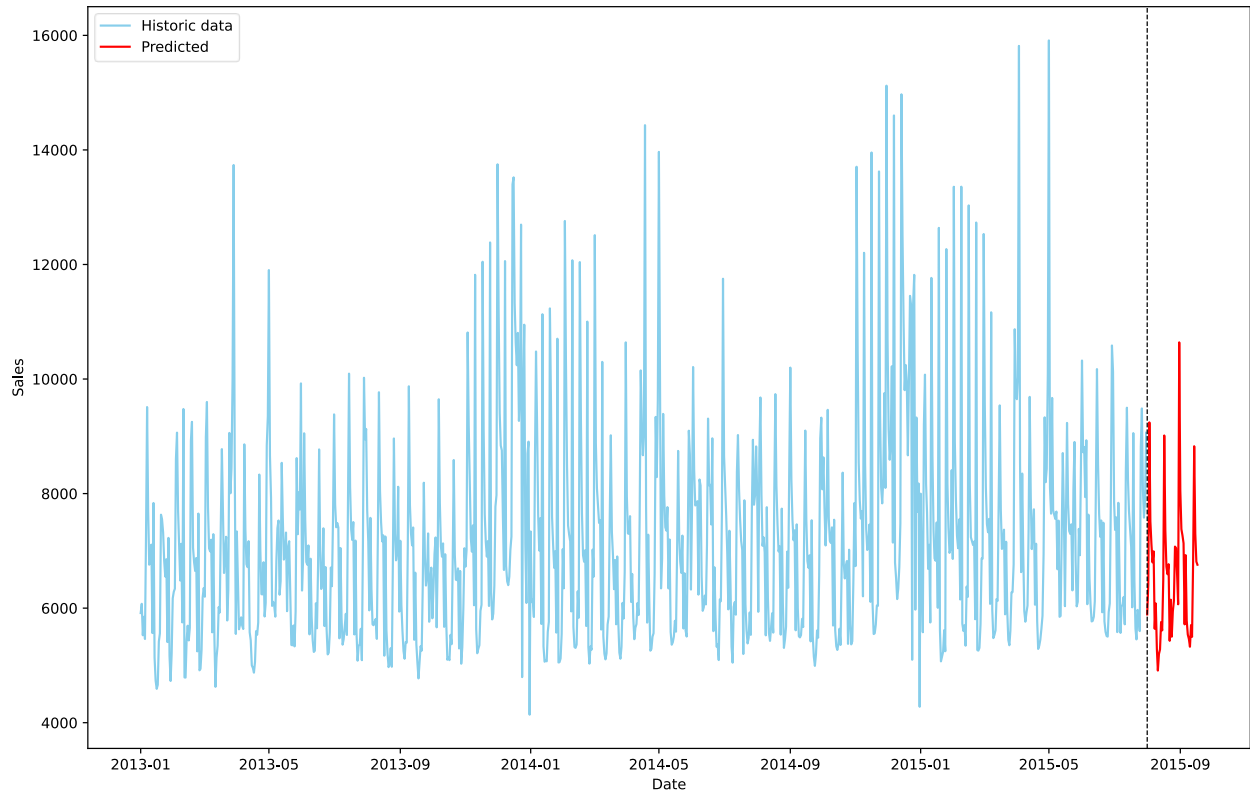
Using the model, the sales performance of all stores was forecast for the desired period. In addition, functions were created to allow comparisons between the predictions of different stores. For example, Figure 13 shows the predictions for Store 1 and 850.

Figure 13. Sales Reality vs Prediction- Store Comparison



Finally, Figure 14 shows the average sales forecast for all the stores considered in this project.

Figure 14. Average Sales per Date - ROSSMANN Stores



Appendix

Figure A1. Sales according to StoreType

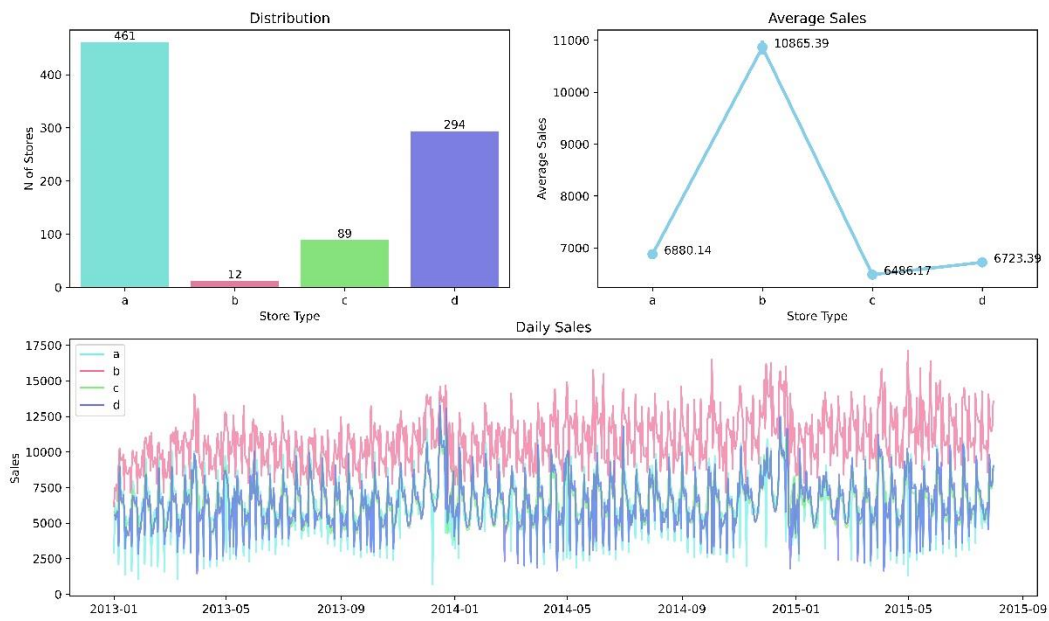


Figure A2. Sales according to Assortment

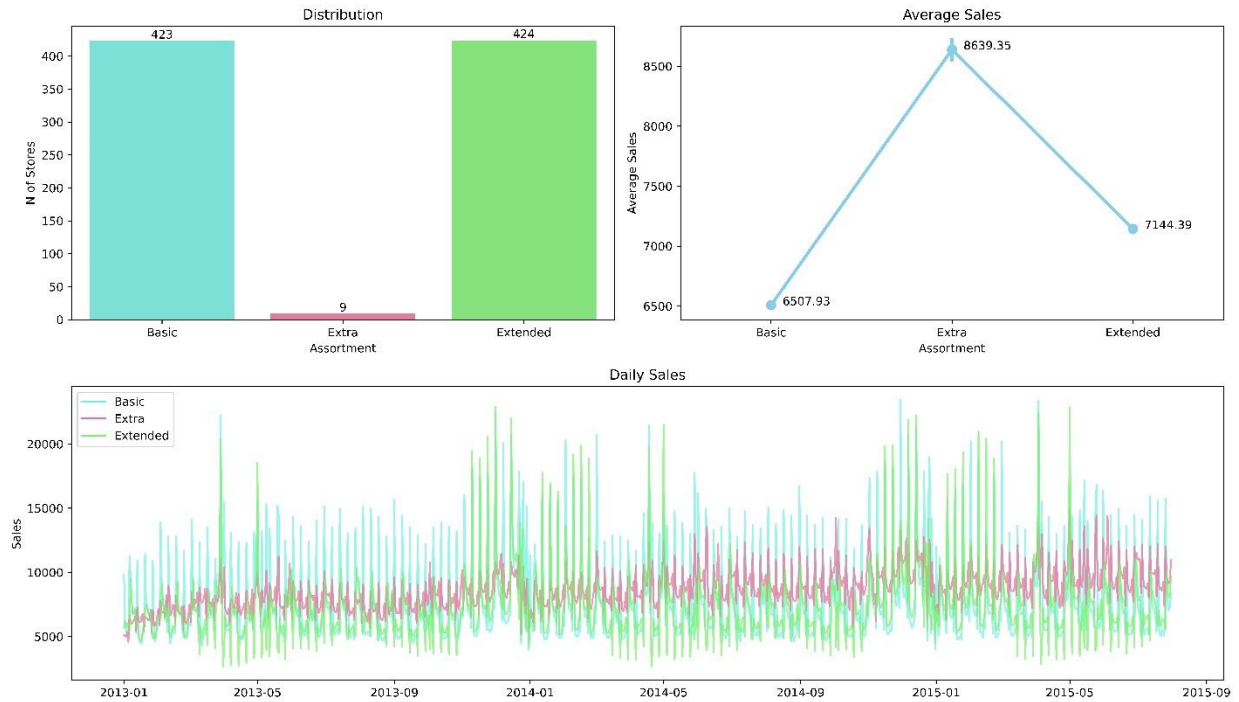


Figure A3. Sales according to Special Dates

