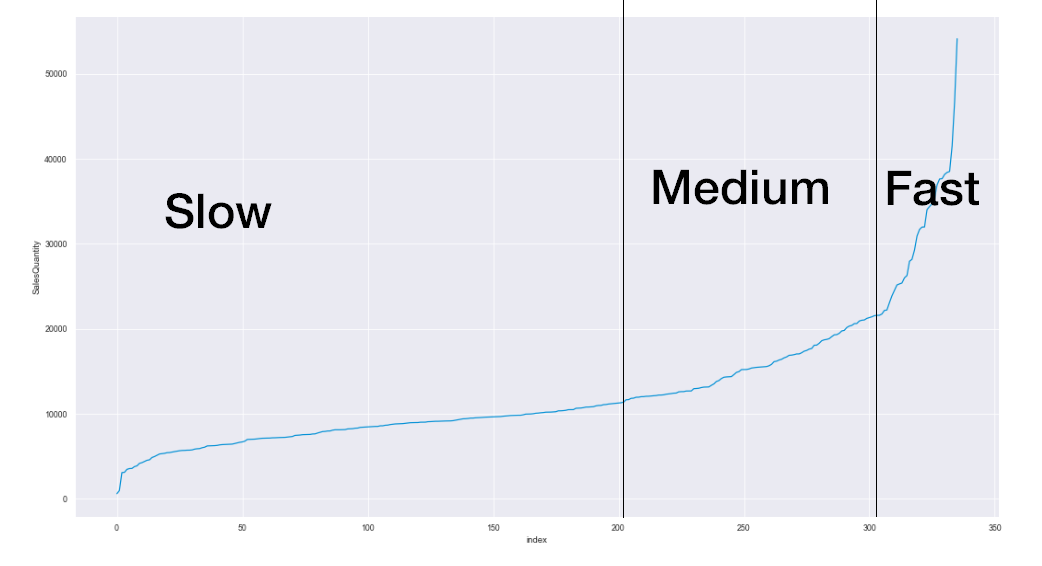
**Promotion Bump Assignment**

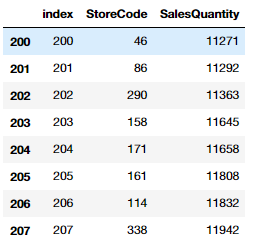
My overall approach, in this case, was as follows:

* Analyze the data using Python and Excel
* Add more variables so that the models could explain more of the unexplained variance.
* Product groups are added in the same manner.
* Item types, store types, day, month, and weekday are added in order for models to explain the seasonality in the data.
* Promotion Dates are added as a binary variable so that models can explain the bump in the sales quantity during promotions. (Also, the promotion dates were wrongly arranged, so I fixed them.)
* In the end, a gradient boosted decision tree, Linear Regression, and Stochastic Gradient Descent regression are tried.
* Because the Linear Regression and SGD didn’t accept the categorical variables as an argument unlike the Decision Tree algorithm, I had to convert to categorical variables to binary variables using OneHotEncoder.

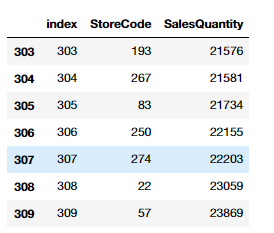
**What are your criteria for separating Fast, Medium and Slow items? Why?**



My criteria to separate items was sales quantity. First I took a look at the chart. It can be seen that the difference between the sales quantity of items can be observed by looking at the derivative line of the point. So, I decided on a decision boundary, as seen above. Then to justify my claim I look at the raw data. The quantity increased immensely between the boundaries, as seen in the below frame.



As seen, before passing the 203rd index sales quantity difference increases from 20~100 to 300~. Thus we call items less than 203rd index “Slow” items.

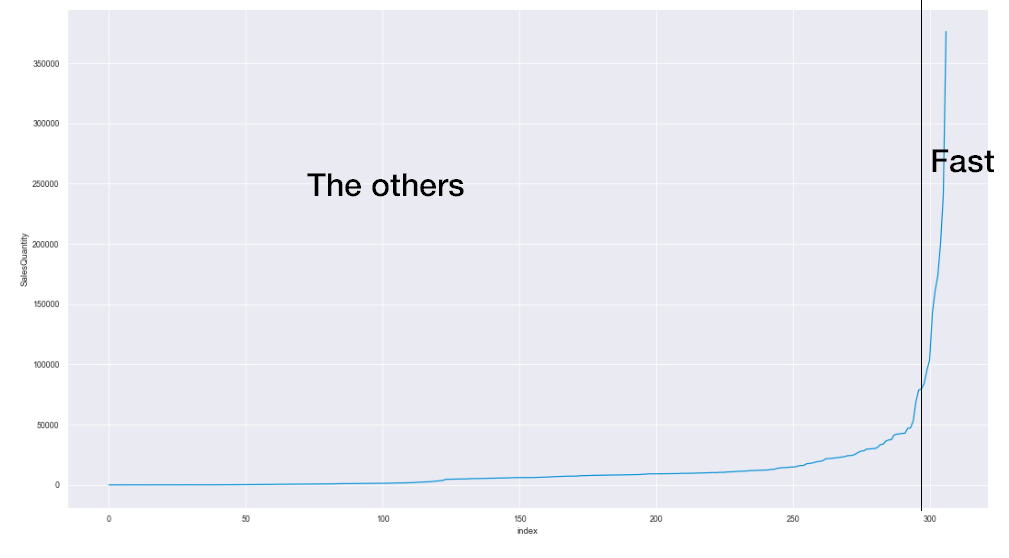


We can use the same claim for separating the fast and medium items. Thus, from the 203rd index to the 307th, we call the items “Medium” items.

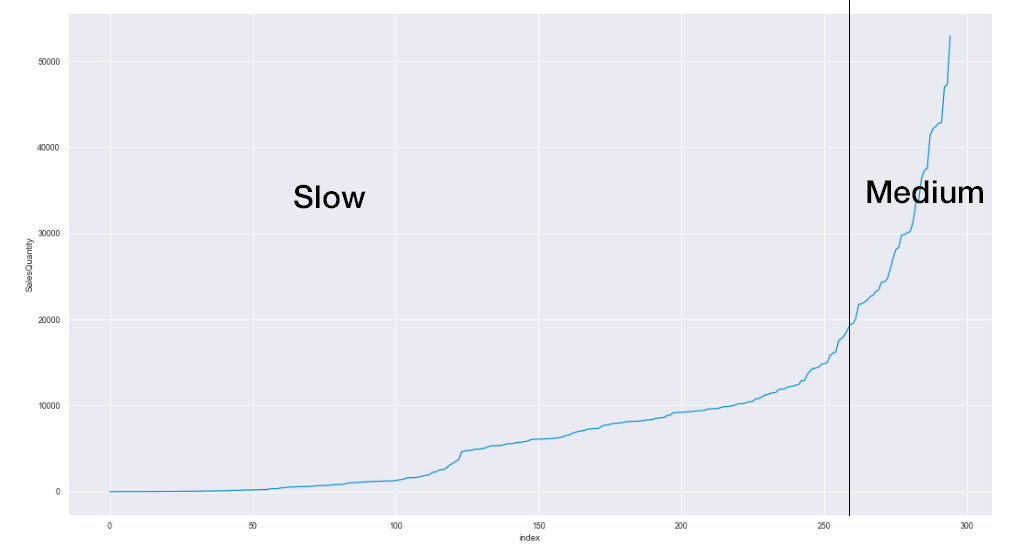
In the data, the time interval is the same so by comparing the sales quantity. The reason why I used such an approach is that the item's type depends on how fast it is sold. We can find this term by using the approach above.

**What are your criteria for separating Fast, Medium and Slow Stores? Why?**

The same approach, as above stays the same for the stores.

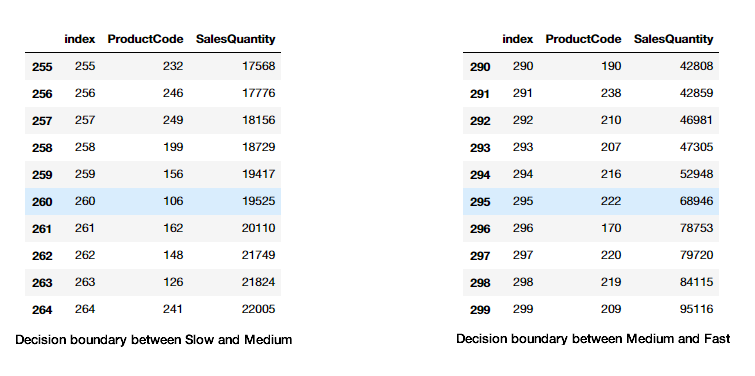


As seen, above the 295th index, the stores have an immense number of sales. This caused the chart not to show the other type of stores, so I pruned the chart a bit:



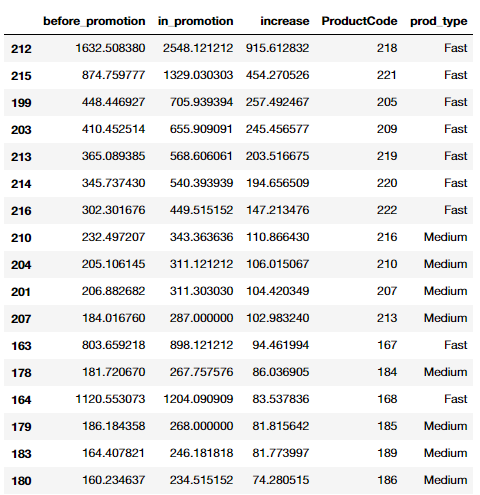
As seen, the difference between the 260th and the 261st store is more compared to the stores that are in the lower index than the 260th.

Again to prove my point, here is the boundaries’ place from the raw data:



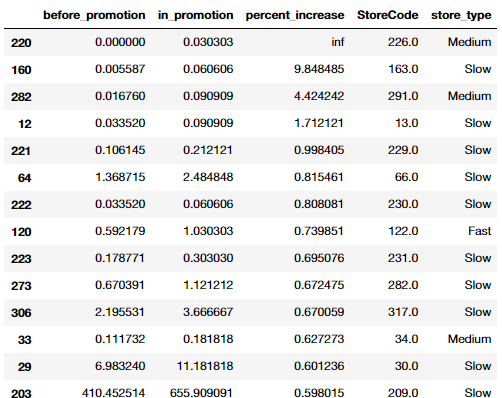
**Which items experienced the biggest sale increase during promotions?**

As expected, fast items experienced the biggest sale increase.

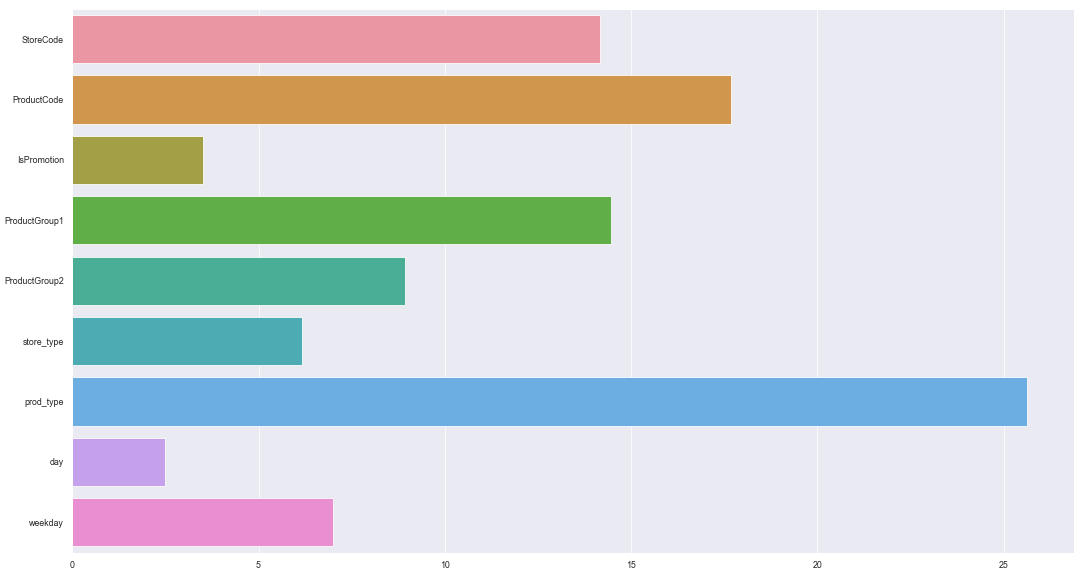


**Are there stores that have higher promotion reaction?**

Yes, as expected, mostly slow stores have an immense percent increase among other types. Actually, here we should emphasize on the last entry of the data below because the others has low sales quantity during non-promotion days.



**What is the biggest effect explaining sales change during promotions?**



Depending on the best-performed model which is a gradient boosted decision tree, the product type affects the sales the most.



**Is there any significant difference between promotion impacts of the Fast versus Slow items?**

When we take a look at the above analysis I made on excel, we see that Medium type items tend to have more sales quantity change during promotions.



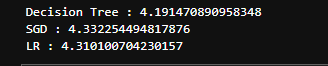
**Is there any significant difference between promotion impacts of the Fast versus Slow stores?**

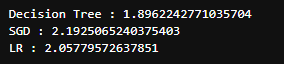
Here, we see that the overall change during promotions for stores is the same for slow and medium, but little lesser for fast stores.

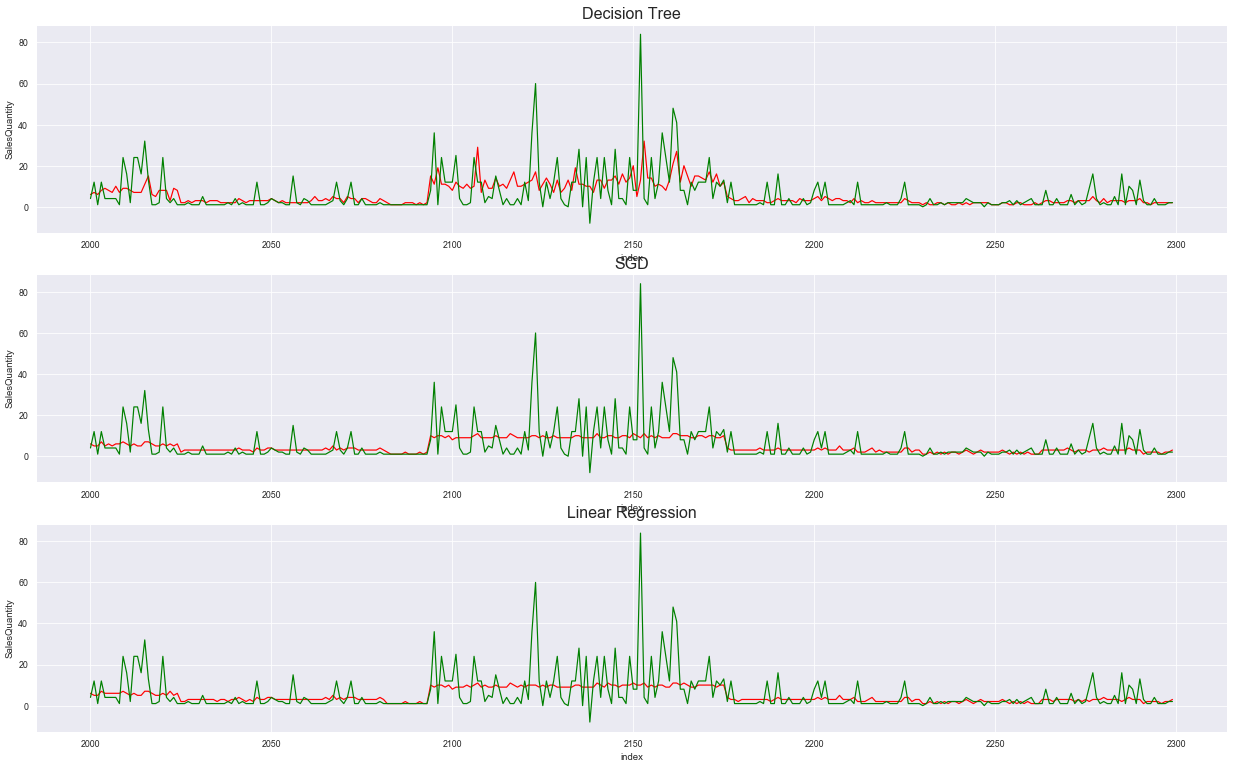
**You are asked to measure how well your model has worked on this new data. Based on the model developed in part A forecast what would the effect of promotion 5 will be on sample store – item pairs. Compare the results of your forecast for promotion5 with the real observed sales during that period.**

As seen, a more complex model proved to be better within the randomly selected frame of time. It gives more bias in the event of promotion thus causing the model to behave better.

The scores are RMSE, and MAE in order:







Furthermore, if we choose only promotion 5, the MAE becomes:



**What measure would you use for goodness of fit?**

The measures I used for this task are RMSE and MAE. However, using MAE is more beneficial because the RMSE error penalizes the closeness more than the MAE error. Since in our case, it is not that important to predict the exact quantity, using MAE is beneficial.

**How good is your model developed in step 1?**

The scores are given above.

**What are the main problem points causing bad fits?**

The main problem is the uncertainty of the quantities during promotions. As you can see, our overall is 1.8, however, during promotion 5 the MAE increases to 2.4. We can comment that the reason for the uncertainty of the predictions is caused by promotions.

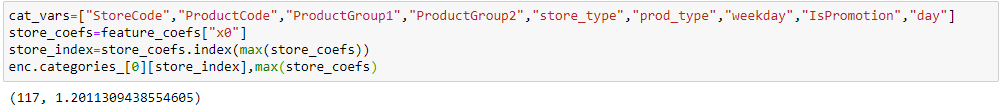
**What would you change in step 1?**

My aim was to use Multivariate time series forecasting (VAR model), by using that I could have used the seasonality effect and using the promotion variable, I believe I could generate better results.

**Conclusion and Report**

Because it is easy to explain the variables using regression, I will emphasize on the Linear Regression when making assumptions.

When we take a look at which stores has the greater coefficient, we see that:



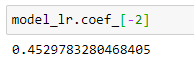
which means that the store with code 117 has the greatest coefficient in order words, has the most expectation in terms of sale quantity.



As seen here, if the product has code 137 the sales quantity is increased by 10~. Actually, we could determine whether the product or store is whether Fast, Medium or Slow by observing the expectations of the variables.



Here we see that the index of 5, which equals Saturday, has the most effect on the response variable. So, when the day is Saturday the sales quantity increases by 0.72, which makes sense because Saturday is a holiday.



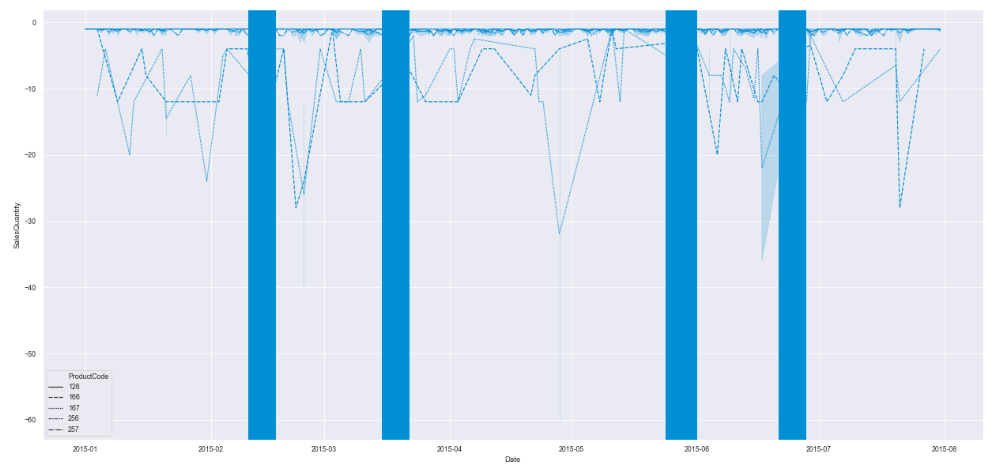
Furthermore, the coefficient of Promotion’s binary variable (Whether the transaction happens on a promotion day or not) is 0.45. This means when there is promotion, a product has an expectation of being sold 0.45 more in terms of quantity. However, this is not biased when it should be, because this number contains all the items, Fast, Slow or Medium, and when there is promotion the Slow and Medium items tend to sell more. Also, this concept applies to stores, as well.

Depending on this issue, I conducted an analysis on Excel (because it is interactive and gives you more flexibility in this case):



As seen, the change percentage changes with the type of stores and products.

Also, we can see that there is an increase in returns after the promotions.



**Future Works and Improvements**

* Creating different models for different types of products or stores will increase predictability during promotion days.
* More sophisticated models, like CatBoost I used, may be used in these kinds of cases but will require more computing power.
* The other sheet shows a different analysis which explains the percent increase including the product groups, store types. We could use these 48 instances and create models.
* Going further: Creating different models for different products or stores. (Also, causes more computing power but beneficial)
* We could prepare a feasibility analysis including the above statement and use the one giving us the optimal output.
* We could add national holidays to the data. So, that we could explain the bumps or diminishes along with the data.
* Also, more information about the stores and items will help.
* Stock data of the products and the availability in the stores of the products would help the model explain the sales quantity.