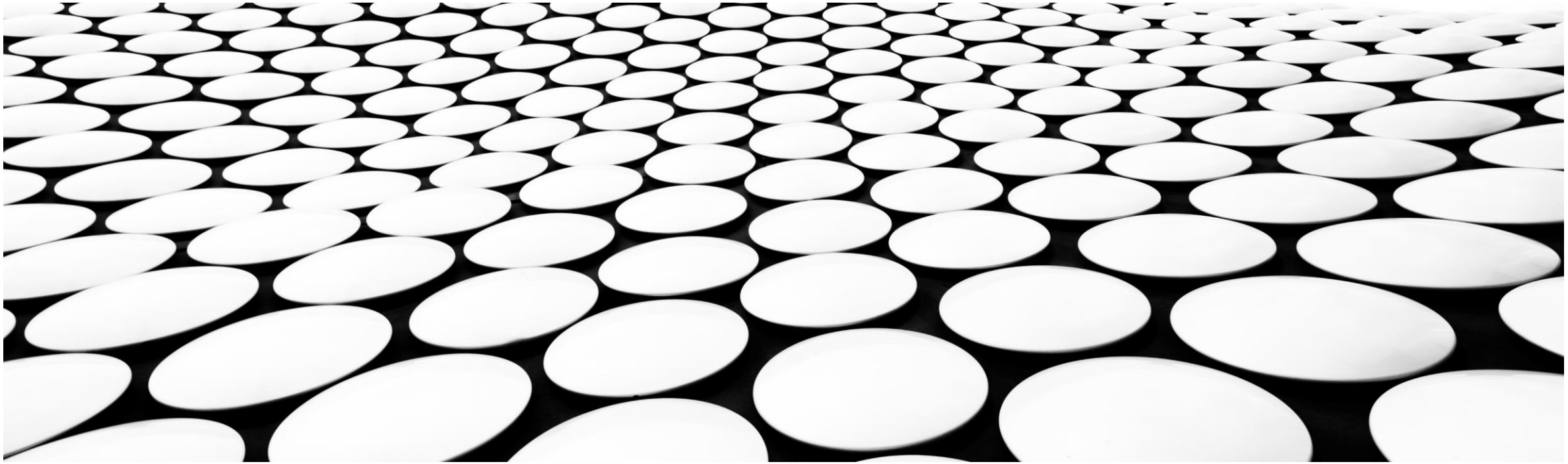
PHARMACEUTICAL SALES PREDICTION ACROSS MULTIPLE STORES

ROSSMAN STORE SALES REPORT



DATA AND FEATURES

- **Id** - an Id that represents a (Store, Date) duple within the test set
- **Store** - a unique Id for each store
- **Sales** - the turnover for any given day (this is what you are predicting)
- **Customers** - the number of customers on a given day
- **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
- **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
- **StoreType** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended. Read more about assortment here
- **CompetitionDistance** - distance in meters to the nearest competitor store
- **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened
- **Promo** - indicates whether a store is running a promo on that day
- **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
- **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

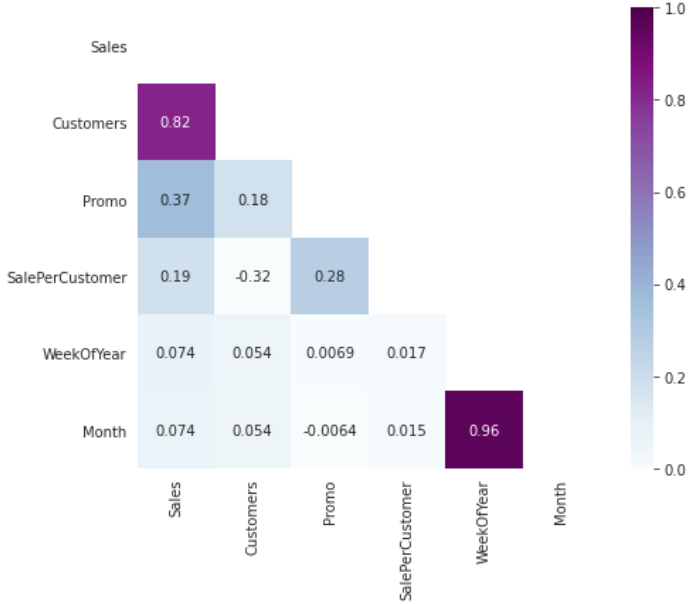
NUMERICAL ANALYSIS

	count	mean	std	min	25%	50%	75%	max
Store	844338.0	558.421374	321.730861	1.000000	280.000000	558.00	837.000000	1115.000000
DayOfWeek	844338.0	3.520350	1.723712	1.000000	2.000000	3.00	5.000000	7.000000
Sales	844338.0	6955.959134	3103.815515	46.000000	4859.000000	6369.00	8360.000000	41551.000000
Customers	844338.0	762.777166	401.194153	8.000000	519.000000	676.00	893.000000	7388.000000
Open	844338.0	1.000000	0.000000	1.000000	1.000000	1.00	1.000000	1.000000
Promo	844338.0	0.446356	0.497114	0.000000	0.000000	0.00	1.000000	1.000000
SchoolHoliday	844338.0	0.193578	0.395102	0.000000	0.000000	0.00	0.000000	1.000000
Year	844338.0	2013.831945	0.777271	2013.000000	2013.000000	2014.00	2014.000000	2015.000000
Month	844338.0	5.845774	3.323959	1.000000	3.000000	6.00	8.000000	12.000000
WeekOfYear	844338.0	23.646946	14.389931	1.000000	11.000000	23.00	35.000000	52.000000
SalePerCustomer	844338.0	9.493641	2.197448	2.749075	7.895571	9.25	10.899729	64.957854
CompetitionDistance	844338.0	5450.044852	7801.082007	20.000000	710.000000	2325.00	6880.000000	75860.000000
CompetitionOpenSinceMonth	844338.0	4.926482	4.283634	0.000000	0.000000	4.00	9.000000	12.000000
CompetitionOpenSinceYear	844338.0	1369.692738	935.556484	0.000000	0.000000	2006.00	2011.000000	2015.000000
Promo2	844338.0	0.498670	0.499999	0.000000	0.000000	0.00	1.000000	1.000000
Promo2SinceWeek	844338.0	11.596159	15.308101	0.000000	0.000000	0.00	22.000000	50.000000
Promo2SinceYear	844338.0	1003.201259	1005.874685	0.000000	0.000000	0.00	2012.000000	2015.000000

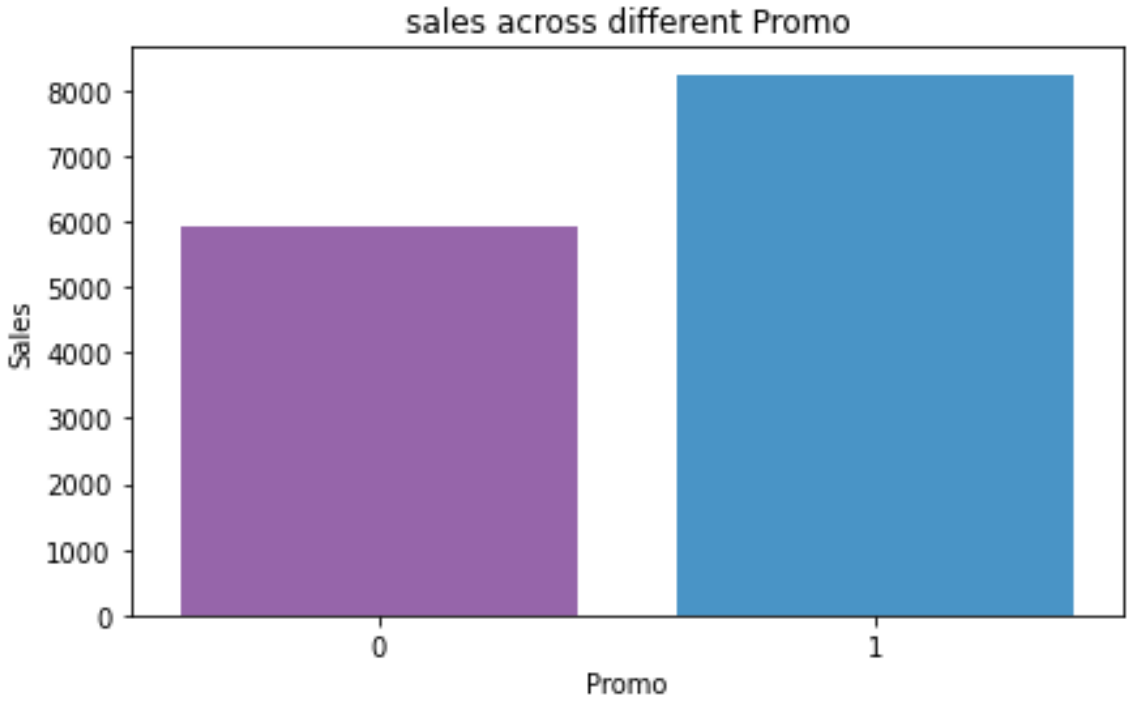
GRAPHICAL ANALYSIS

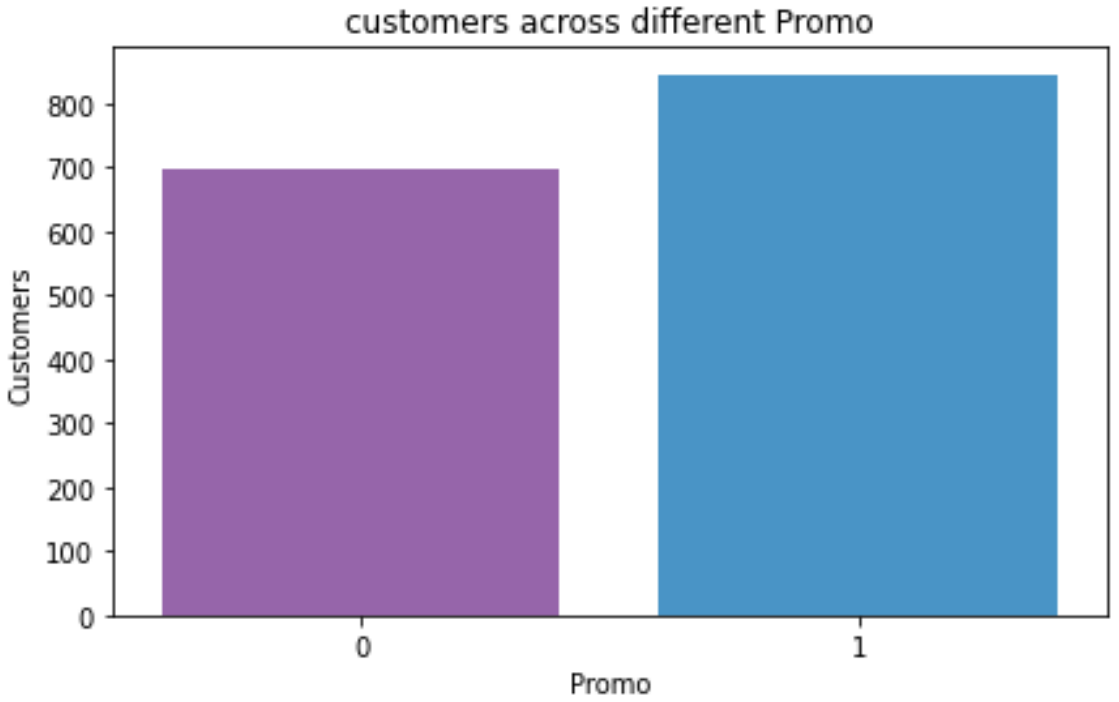
Type c have more sales and customers and as it can be seen more customers gives out more sales

Sales and Customers have a high correlation. In the follow up plots

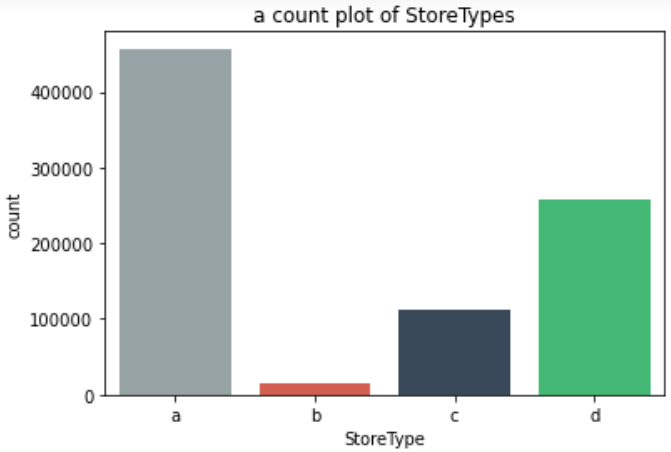


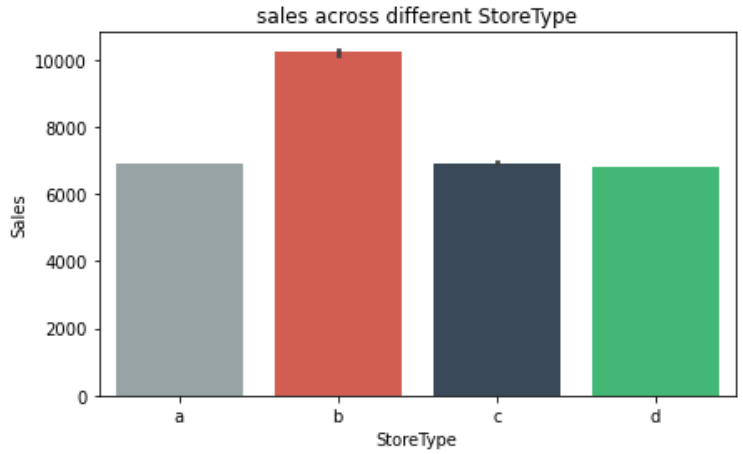
HOW PROMOTIONS AFFECT SALES AND CUSTOMERS

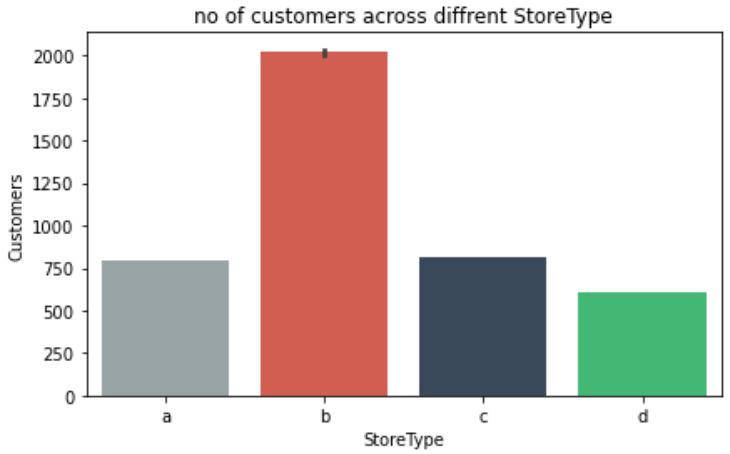




THEY ARE MANY STORES OF TYPE A BUT STORE TYPE B HAVE MUCH CUSTOMERS AND SELLS MORE THAN OTHER STORE

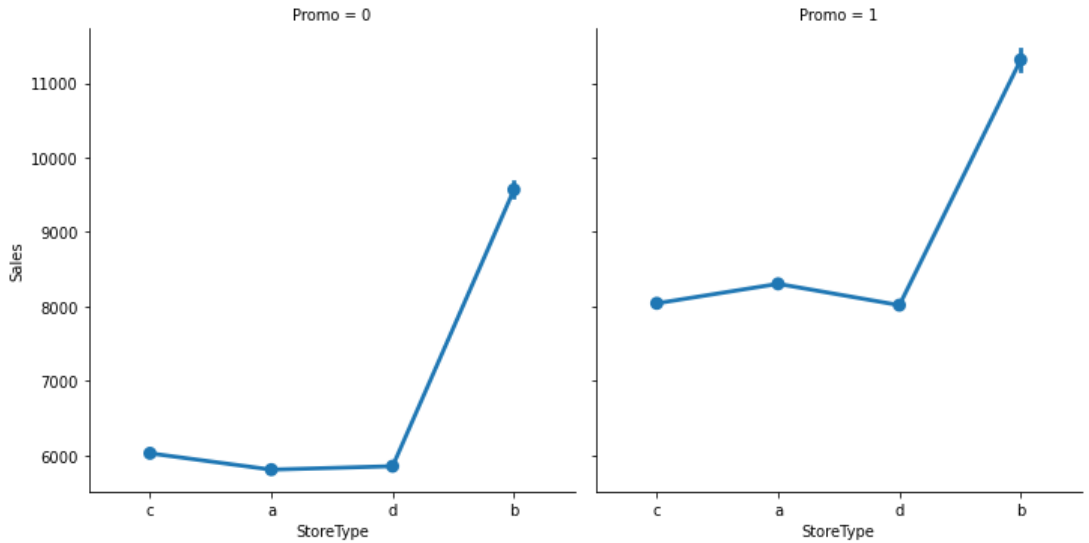




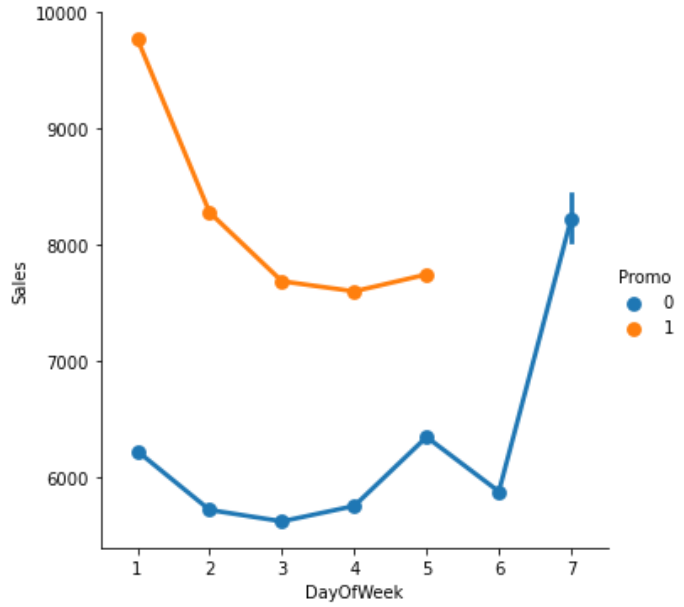


HOW PROMOTIONS AFFECT SALES

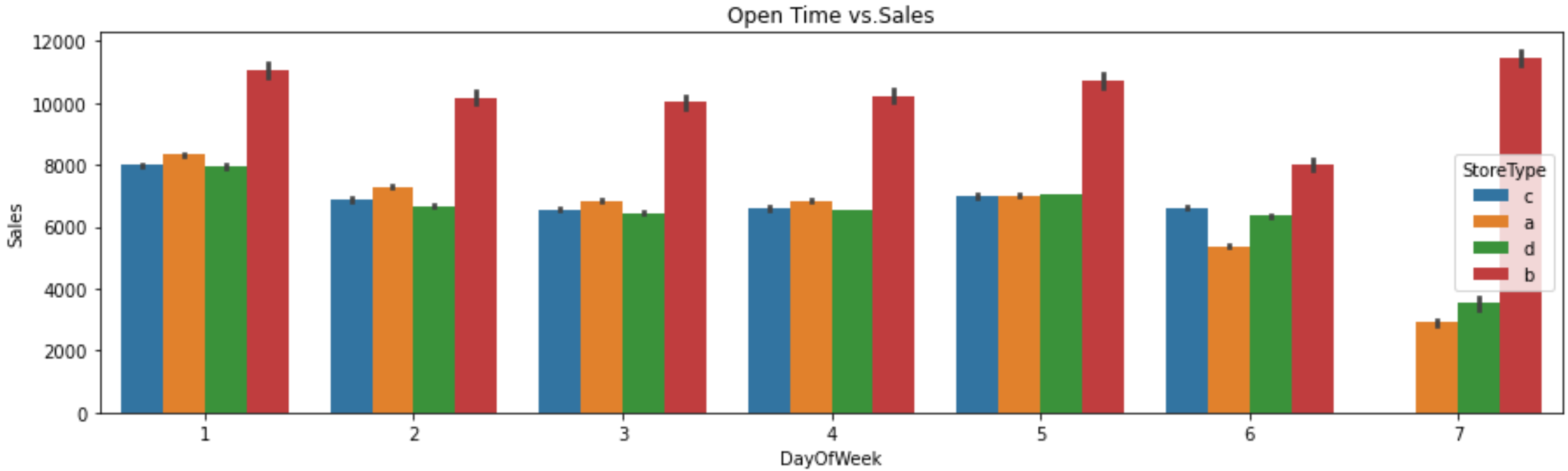
PROMOTIONS INCREASE SALES



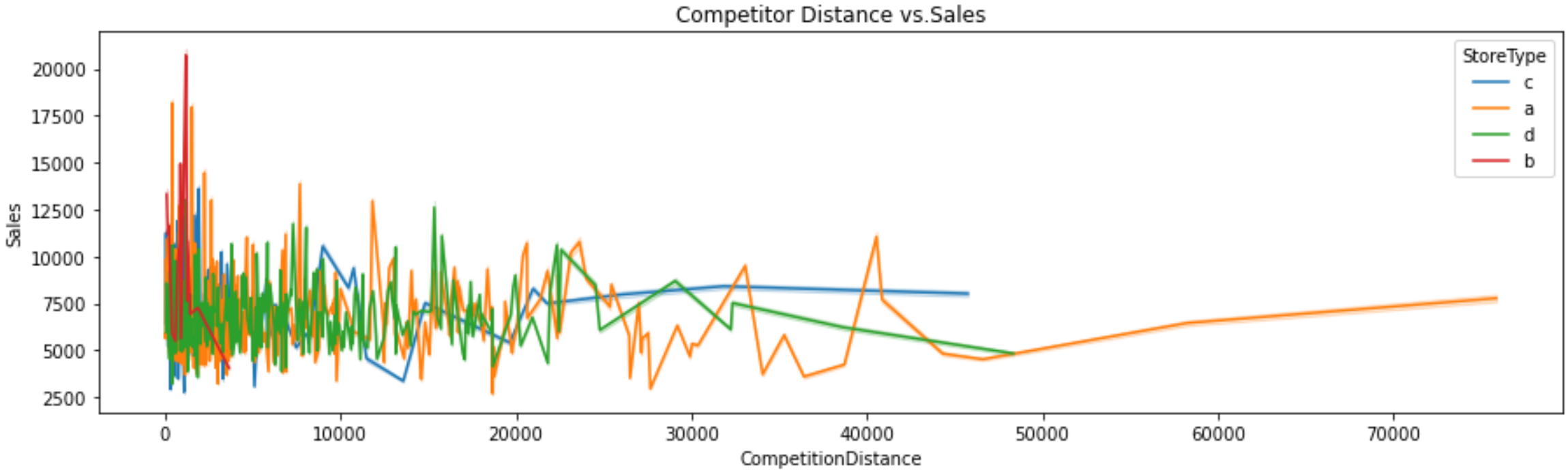
NO PROMOTIONS IN WEEKENDS AND SALES INCRESE ON SUNDAY WHEN NO PROMOTIONS APPLIED



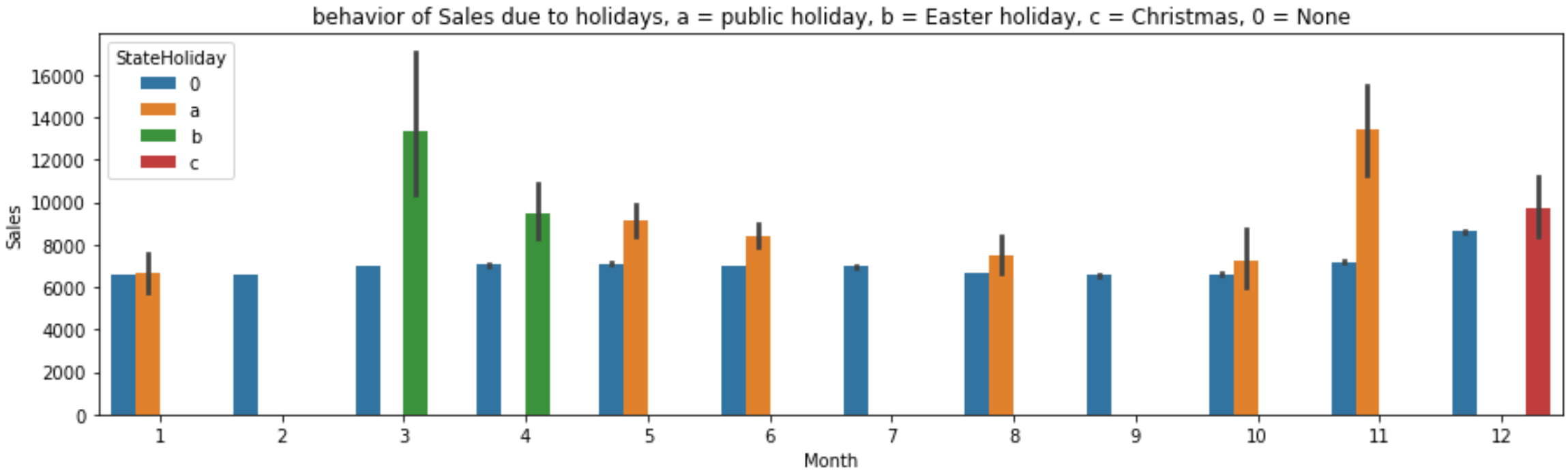
TYPE C STORES DO NOT OPEN ON SUNDAYS AND TYPE B SELLS MORE THAT OTHER STORES



THE SHORTER DISTANCE BETWEEN COMPETITORS THE MORE SALES ARE SOLD AND WHEN STORES A LIKELY TO BE IN THE SAME CITY THE DISTANCE DOESN'T MATTER



THEY ARE MUCH PUBLIC HOLIDAYS IN NOVEMBER, MUCH SALES IN MARCH AND APRIL DUE TO EASTER AND IN DECEMBER THERE IS CHRISTMAS WHICH INCREASES SALES.
MUCH SALES ON PUBLIC HOLIDAYS HAPPEN IN NOVEMBER MAY BE BECAUSE THEY ANTICIPATE CHRISTMAS BEFORE



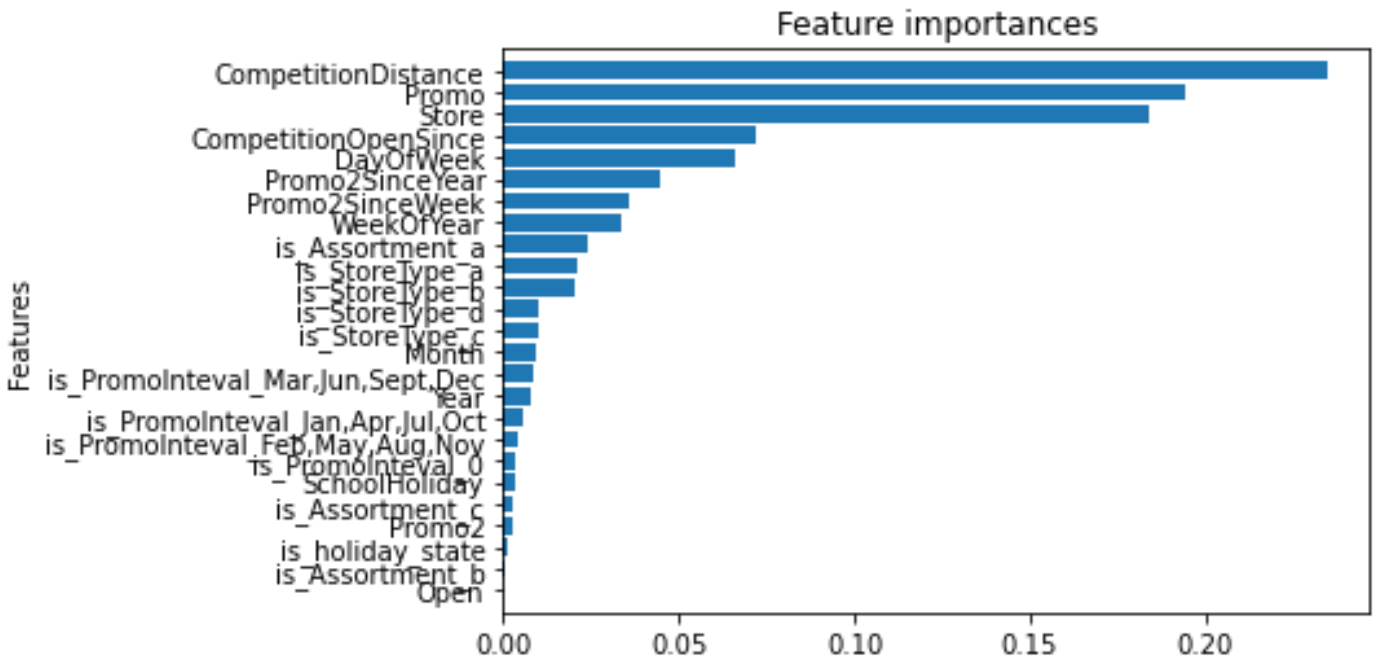
PREDICTIONS FROM RANDOM FOREST MODEL

The most important features are:

Competition Distance : the stores furtherest away from competition will make more sales than stores that are surrounded by competitors

Promo: As we saw earlier stores that had a promo (The first one) had more sales than the ones that didn't have any all

Store: The Store itself represents a unique identificator for the algorithm to recognise which store has what attributes and indeed better accounts for the forecasting of those same stores in a future timeline.



REFERENCES

- https://www.kaggle.com/c/rossmann-store-sales/notebooks
- https://www.kaggle.com/thie1e/exploratory-analysis-rossmann
- https://www.kaggle.com/elenapetrova/time-series-analysis-and-forecasts-with-prophet
- https://www.kaggle.com/shearerp/interactive-sales-visualization
- https://www.kaggle.com/michaelpawlus/obligatory-xgboost-example
- https://www.kaggle.com/stefanozakher94/eda-and-forecasting-with-rfregressor-final-updated
- https://www.kaggle.com/emehdad/time-series-linear-models-tslm
- https://www.kaggle.com/sammyshen/exploratory-and-randomforest