Data Exploration & Machine learning project : Rossman Sales

ROSSMANN

Mein Drogeriemarkt

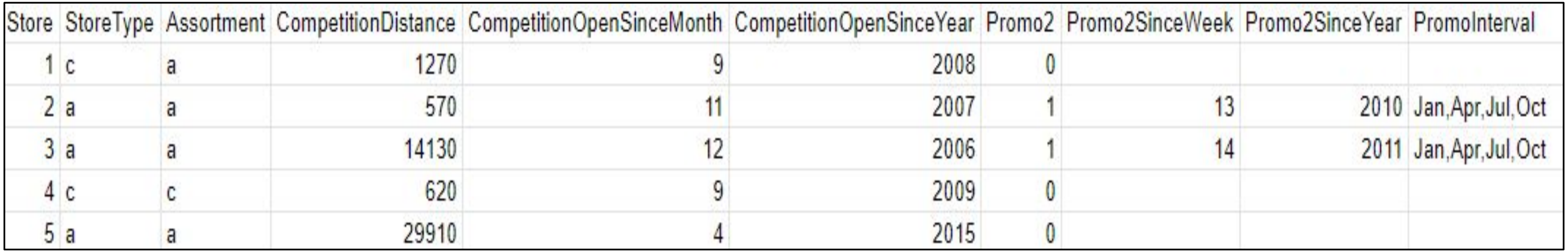
Presentation Guidelines

- Reading through data
- Data Preparation and cleaning
- Exploring and visualising data
- Predictive machine learning model

Reading through data

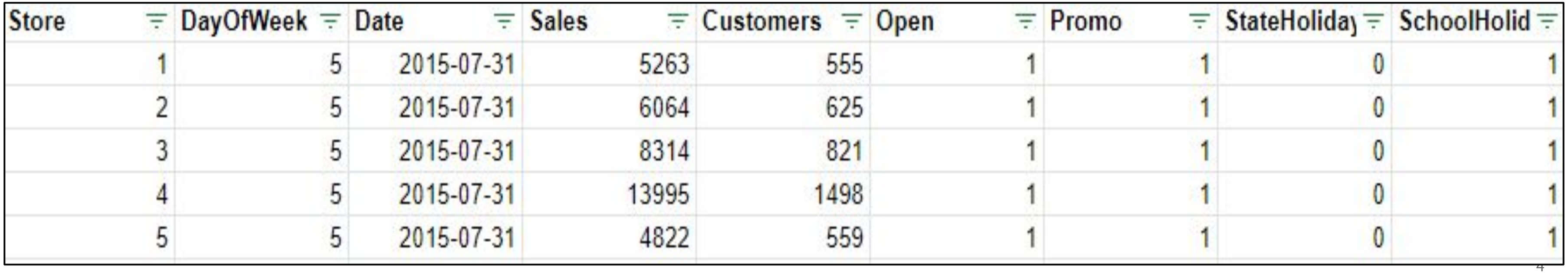
Reading through data

rossman_store



Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
1	c	a	1270	9	2008	0			
2	a	a	570	11	2007	1	13	2010	Jan,Apr,Jul,Oct
3	a	a	14130	12	2006	1	14	2011	Jan,Apr,Jul,Oct
4	c	c	620	9	2009	0			
5	a	a	29910	4	2015	0			

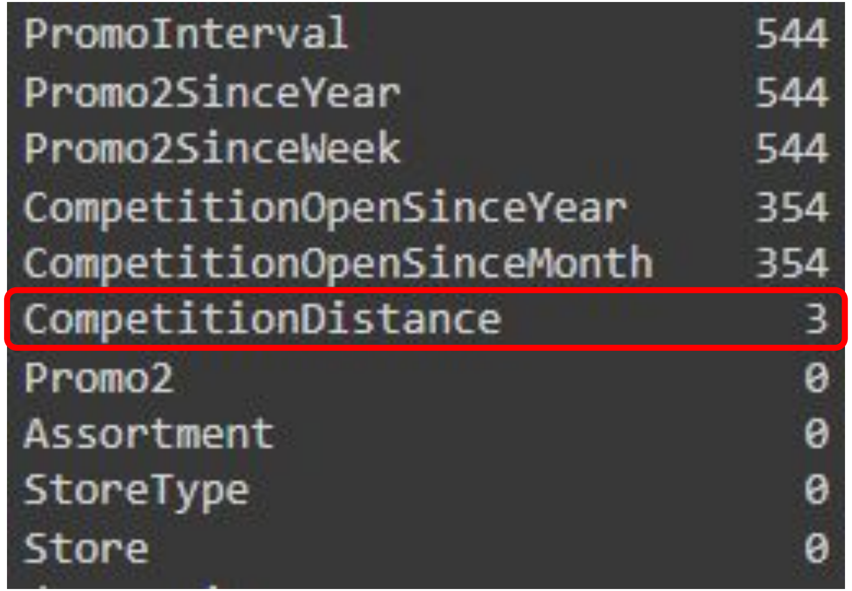
rossman_train



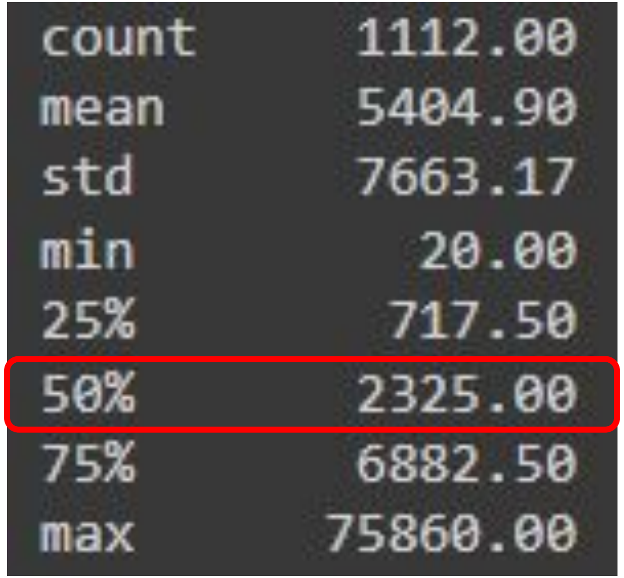
Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHolid
1	5	2015-07-31	5263	555	1	1	0	1
2	5	2015-07-31	6064	625	1	1	0	1
3	5	2015-07-31	8314	821	1	1	0	1
4	5	2015-07-31	13995	1498	1	1	0	1
5	5	2015-07-31	4822	559	1	1	0	1

Data Preparation and Cleaning

Data cleaning

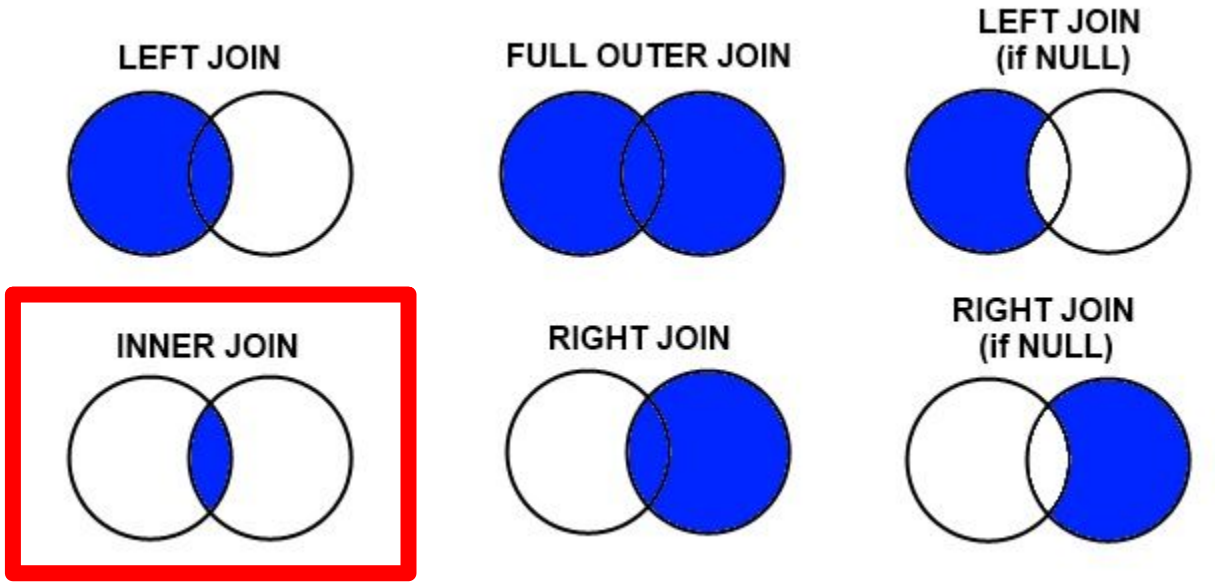


```
PromoInterval                544
Promo2SinceYear              544
Promo2SinceWeek              544
CompetitionOpenSinceYear     354
CompetitionOpenSinceMonth    354
CompetitionDistance            3
Promo2                         0
Assortment                     0
StoreType                      0
Store                          0
```



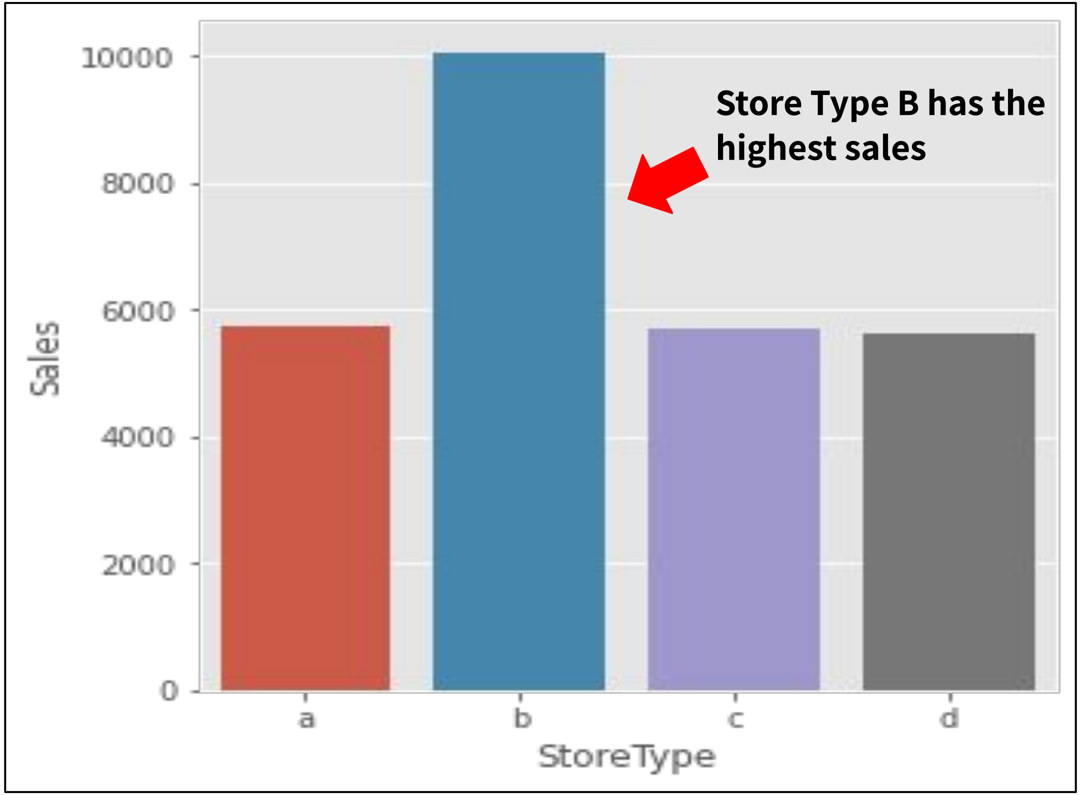
```
count     1112.00
mean      5404.90
std       7663.17
min         20.00
25%        717.50
50%       2325.00
75%       6882.50
max      75860.00
```

Merge data sets



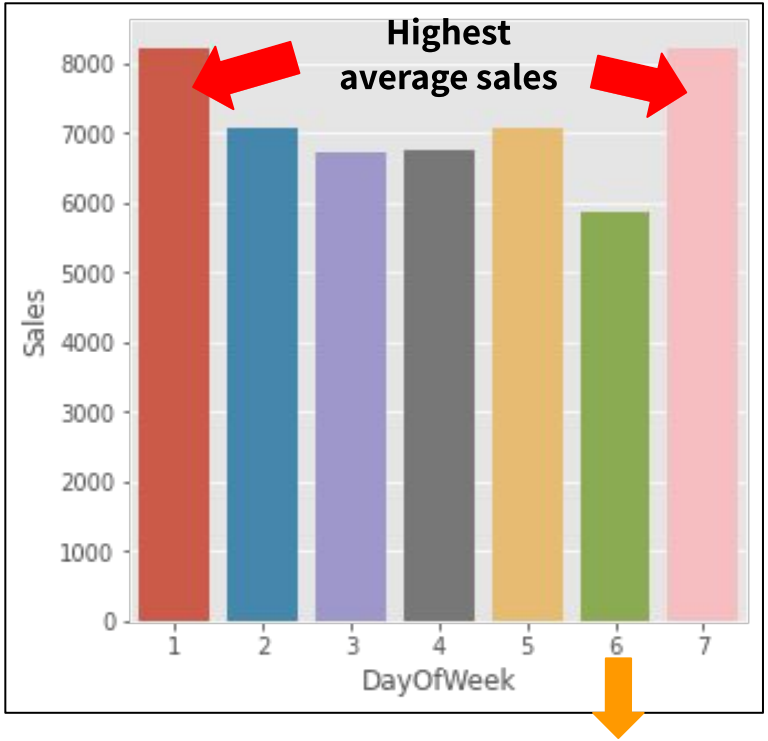
Exploring and visualising data

The average sales in each store type



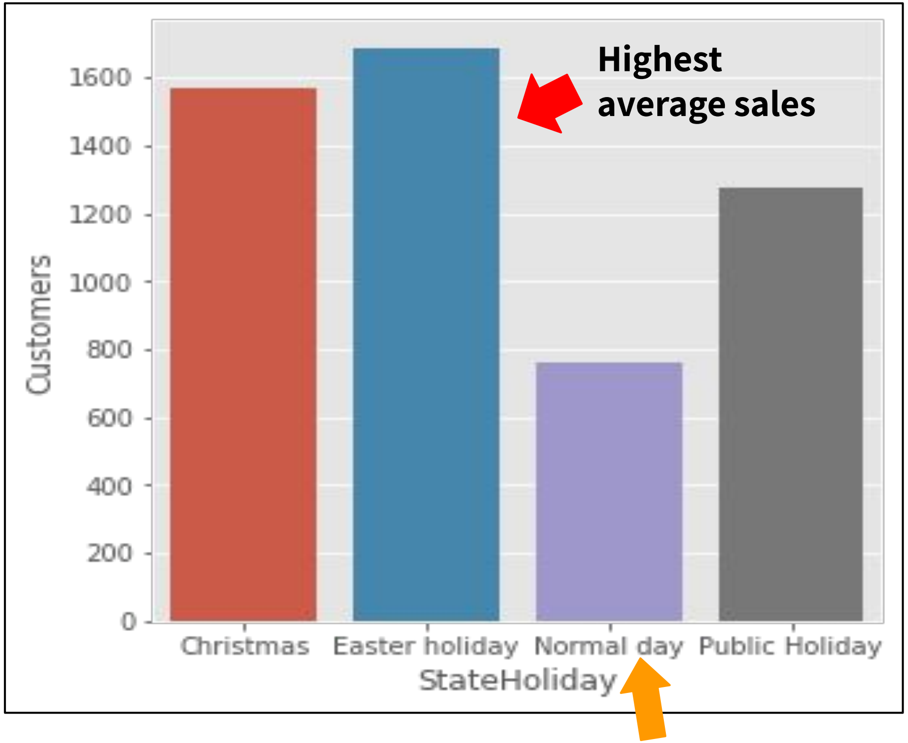
Store Type B has the highest sales

Sales in: Holidays , weekdays and weekend



Highest average sales

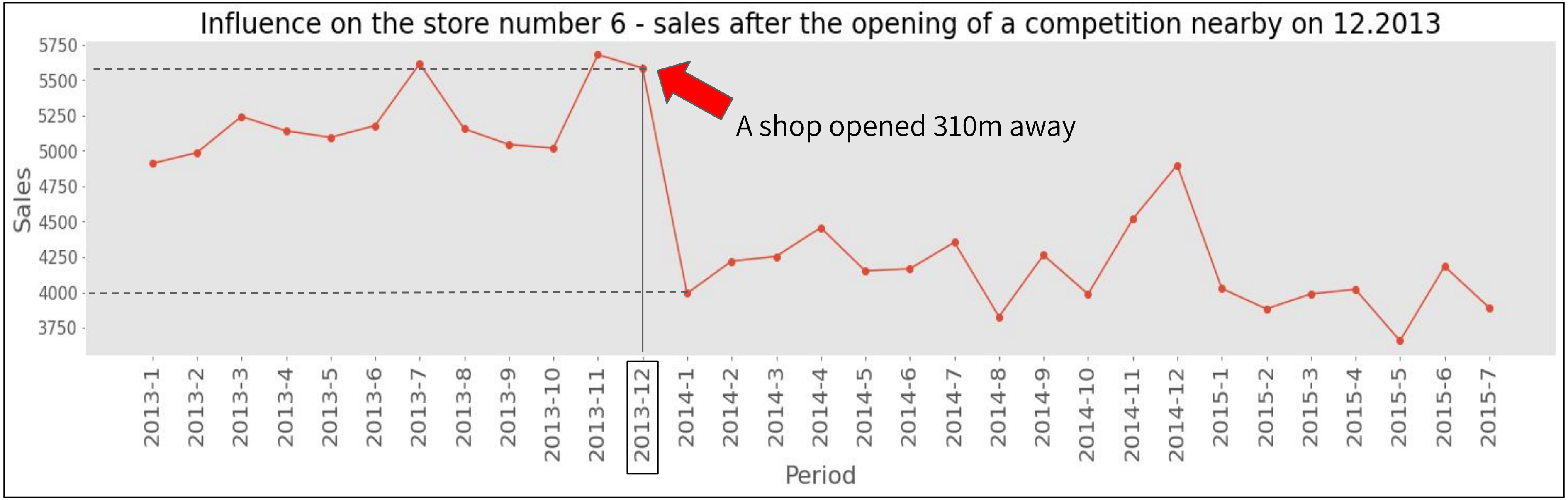
Best day for shopping



Highest average sales

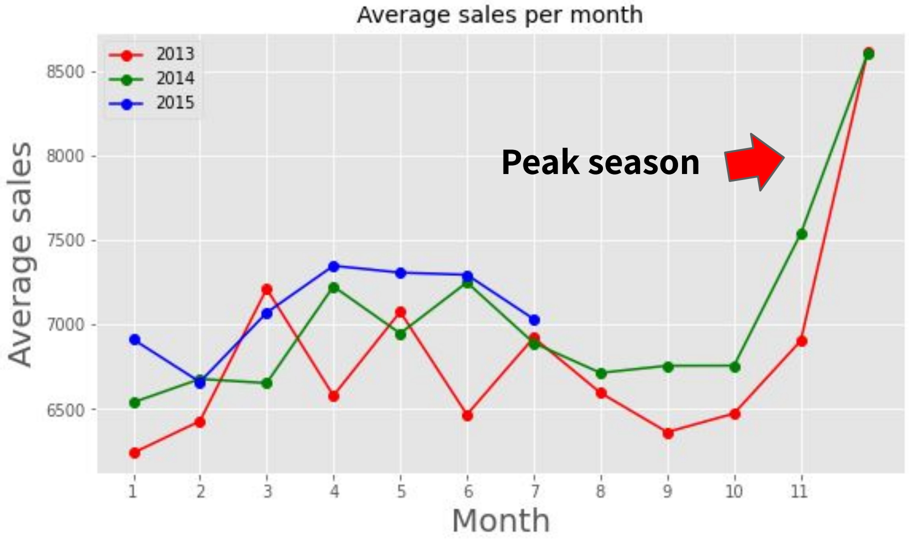
Best time to do shopping

Sales after Competition shop open

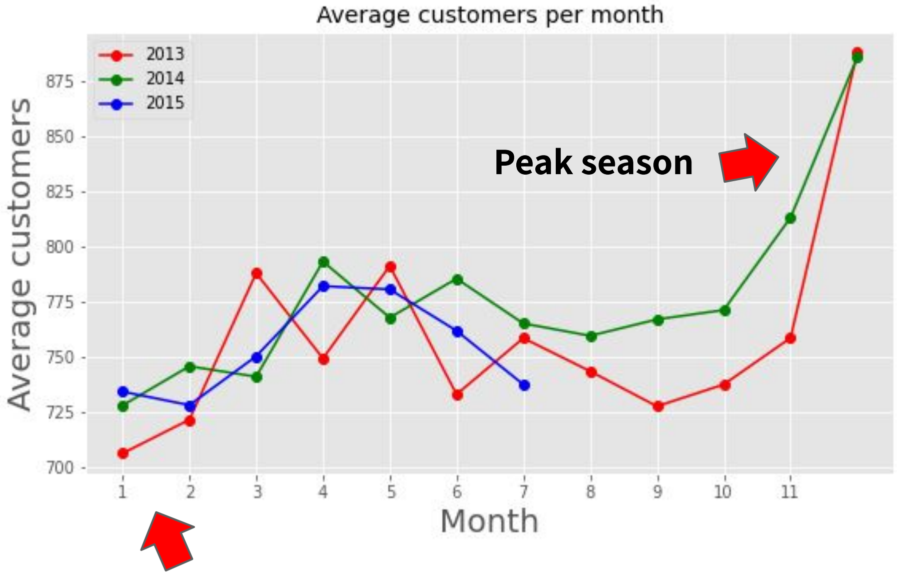


A shop opened 310m away

Average sales & customers per month (2013 - 2015)



Peak season

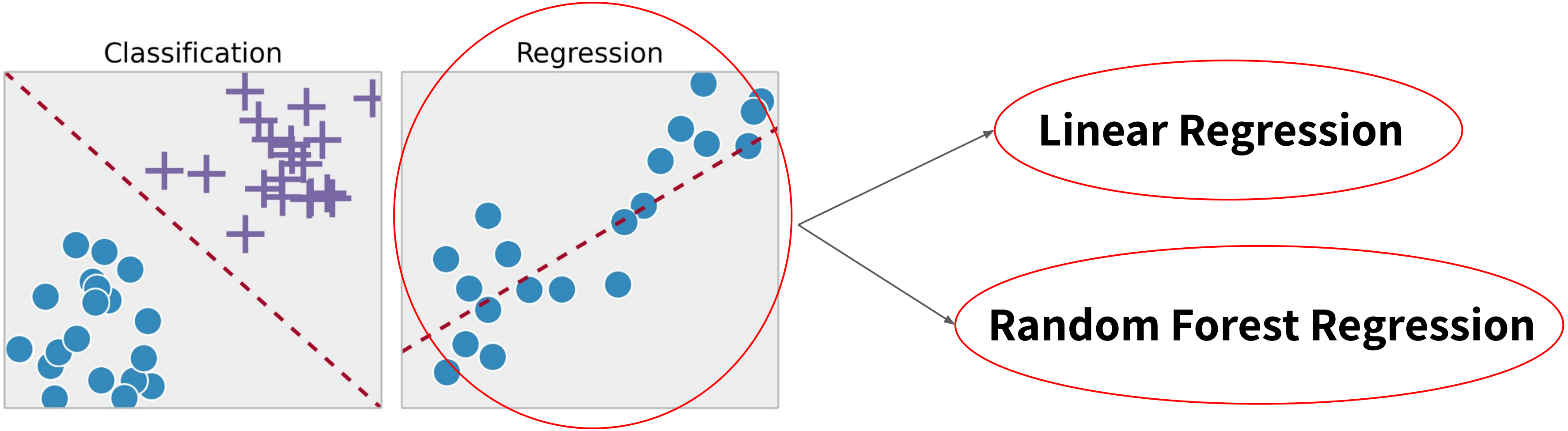


Peak season

Low customers

Predictive machine learning model

Methodology



Classification

Regression

Linear Regression

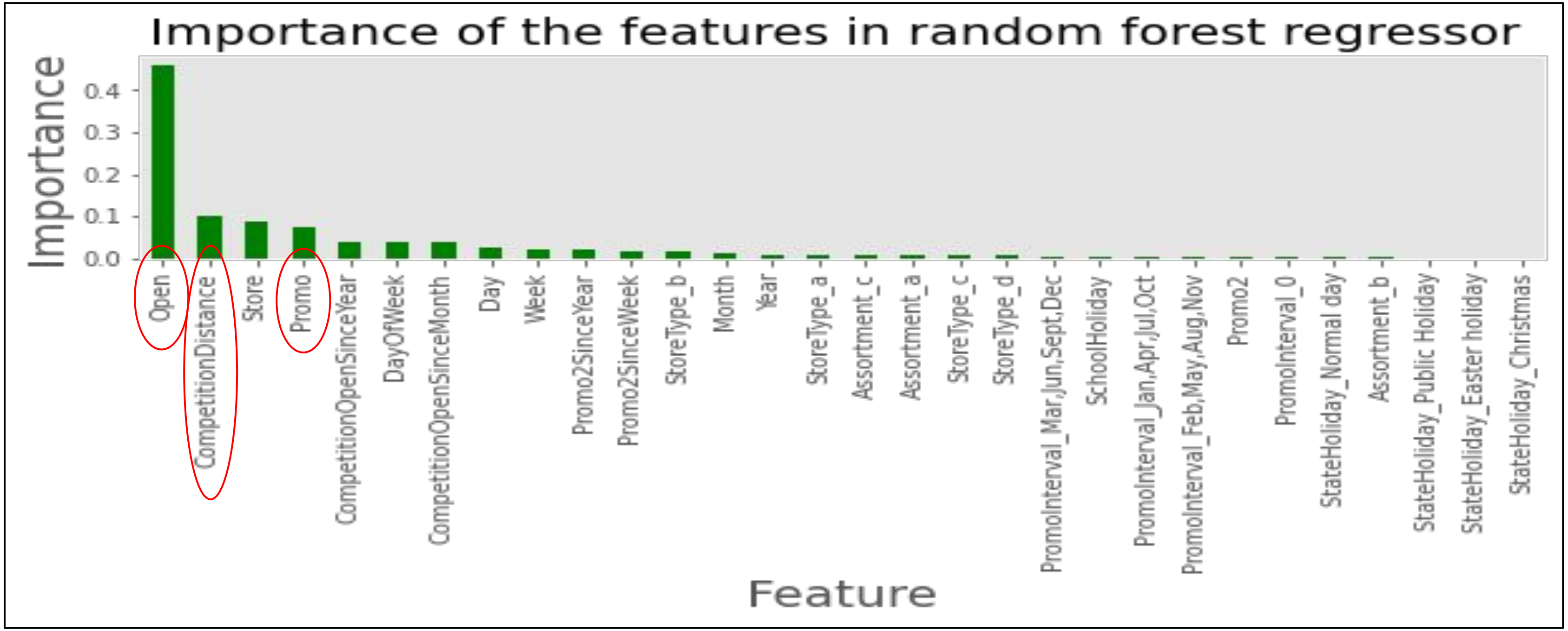
Random Forest Regression

Process of Random Forest Regression Model

- Dependent / Independent Variable(s) Specification
  - Dependent Variable(Y) : Sales
  - Independent Variables(X1, X2, …): (Customers), StoreType_a, …

- Split the data: 80% (model training) & 20% (data testing)
- Create model with random regression trees (100x)
- Fit the output / data
- Get the importance of the features
- Get predictions using our test data
- Calculate the root mean square error

Features Importance - excluding ‘Customers’



Random Forest V.S. Linear Regression

Random Forest Regression result:

- Took 10 min for 100 samples
- RMSE = 828
- More accurate for our case!!

Linear Regression result:

- Runs very fast
- But RMSE = 2508
- Works better when data is linear which not our case!!

Conclusion and Outlook

- Through data analytics:
  - shows a more complete picture
  - use data to build knowledge

- What is next:
  - Continue learning
  - Apply to work

Thank you!!