Rossmann Store Sales Prediction

Springboard Capstone Project 2 by Cuthbert Lo

INTRODUCTION

- To forecast sales of future 6 weeks for 1,115 drug stores of Rossman across Germany.

- With accurate forecast to enable store managers to create effective staff schedule that increase productivity and motivation as planning for stock level as well

Dataset

- The dataset consist of two parts
  a. train.csv - historical data including Sales
  b. store.csv - supplemental information about the stores

In [50]:
```
df_store.info()
```

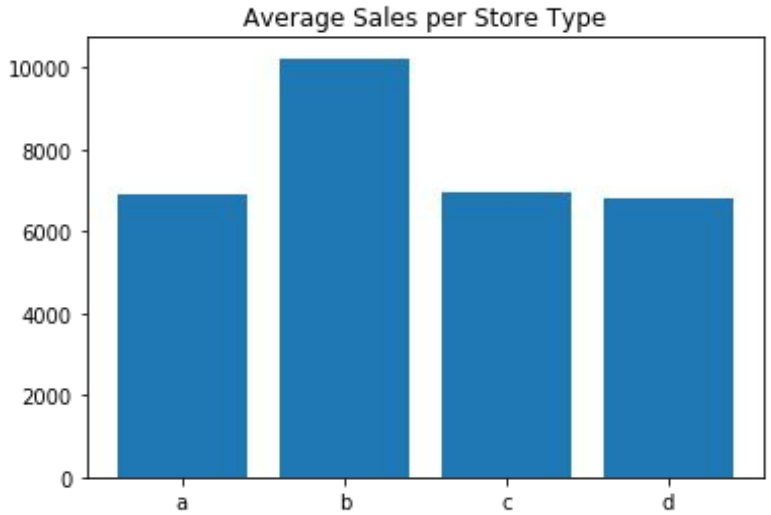
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1115 entries, 0 to 1114
Data columns (total 10 columns):
Store                        1115 non-null int64
StoreType                    1115 non-null object
Assortment                   1115 non-null object
CompetitionDistance          1112 non-null float64
CompetitionOpenSinceMonth    761 non-null float64
CompetitionOpenSinceYear     761 non-null float64
Promo2                       1115 non-null int64
Promo2SinceWeek              571 non-null float64
Promo2SinceYear              571 non-null float64
PromoInterval                571 non-null object
dtypes: float64(5), int64(2), object(3)
memory usage: 87.2+ KB
```

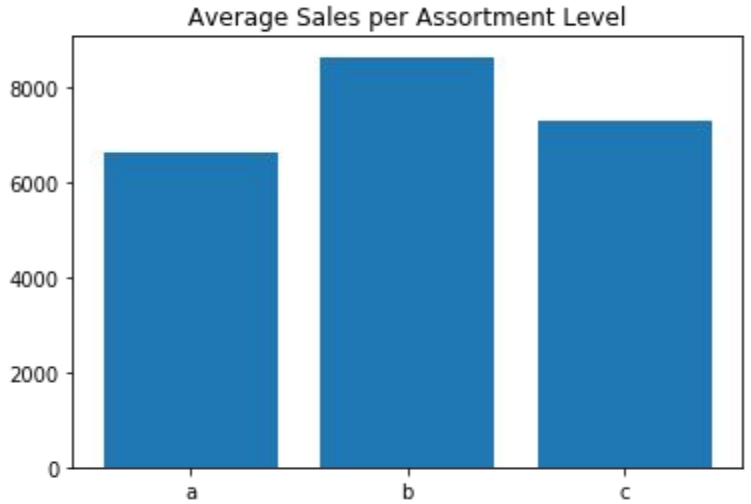
:
```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1017209 entries, 2015-07-31 to 2013-01-01
Data columns (total 8 columns):
Store            1017209 non-null int64
DayOfWeek        1017209 non-null int64
Sales            1017209 non-null int64
Customers        1017209 non-null int64
Open             1017209 non-null int64
Promo            1017209 non-null int64
StateHoliday     1017209 non-null object
SchoolHoliday    1017209 non-null int64
dtypes: int64(7), object(1)
memory usage: 69.8+ MB
```

EXPLORATORY ANALYSIS

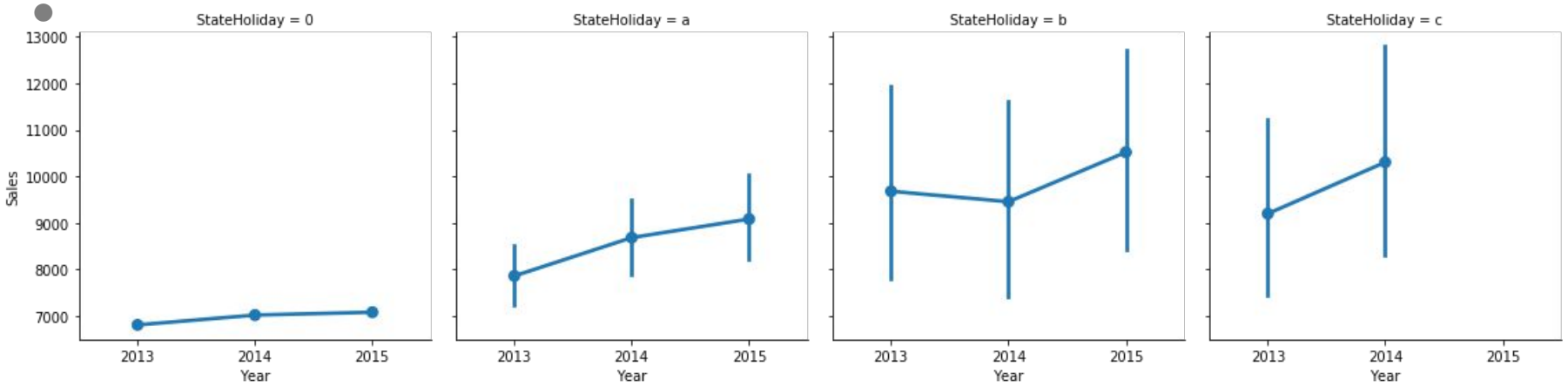
- Store type B and Assortment Level B have the highest average sales





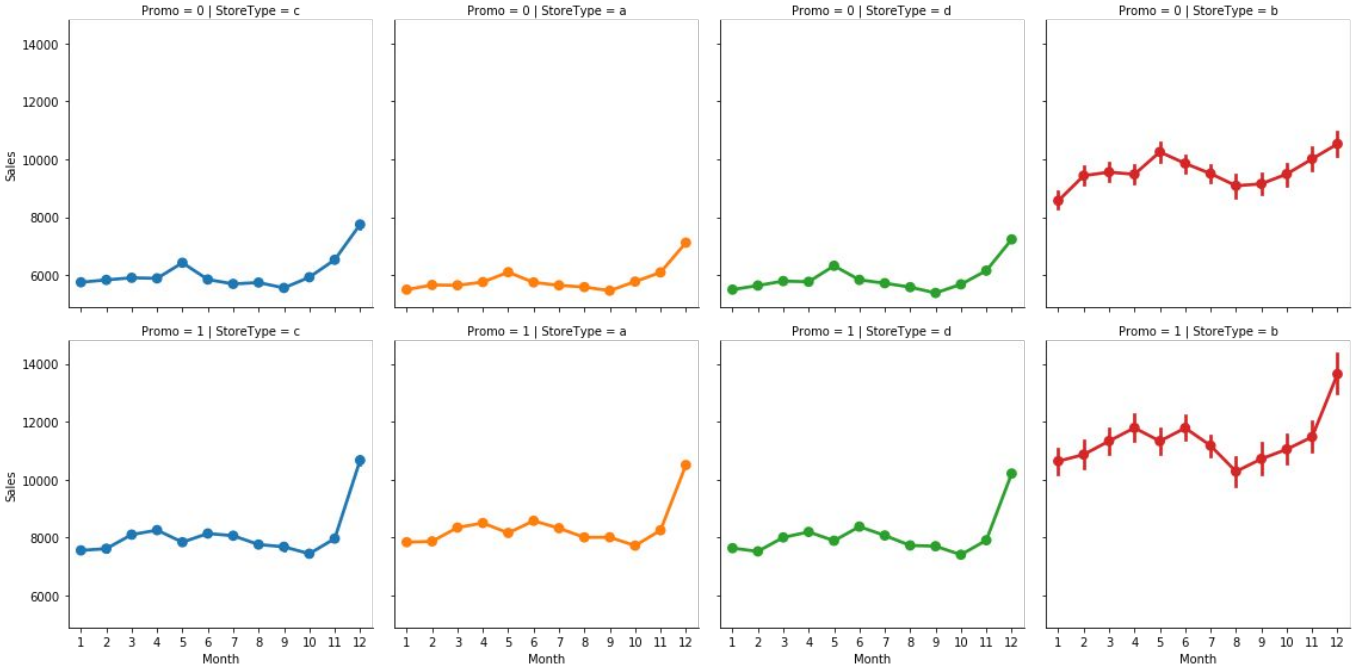
EXPLORATORY ANALYSIS

- The dataset categorized holidays into 4 groups a = public holiday, b = Easter holiday, c = Christmas, 0 = None. From Fig 4 below, state holidays are definitely increase sales, public holiday has a lower sales than Easter holiday and Christmas while Christmas has slightly higher than Easter



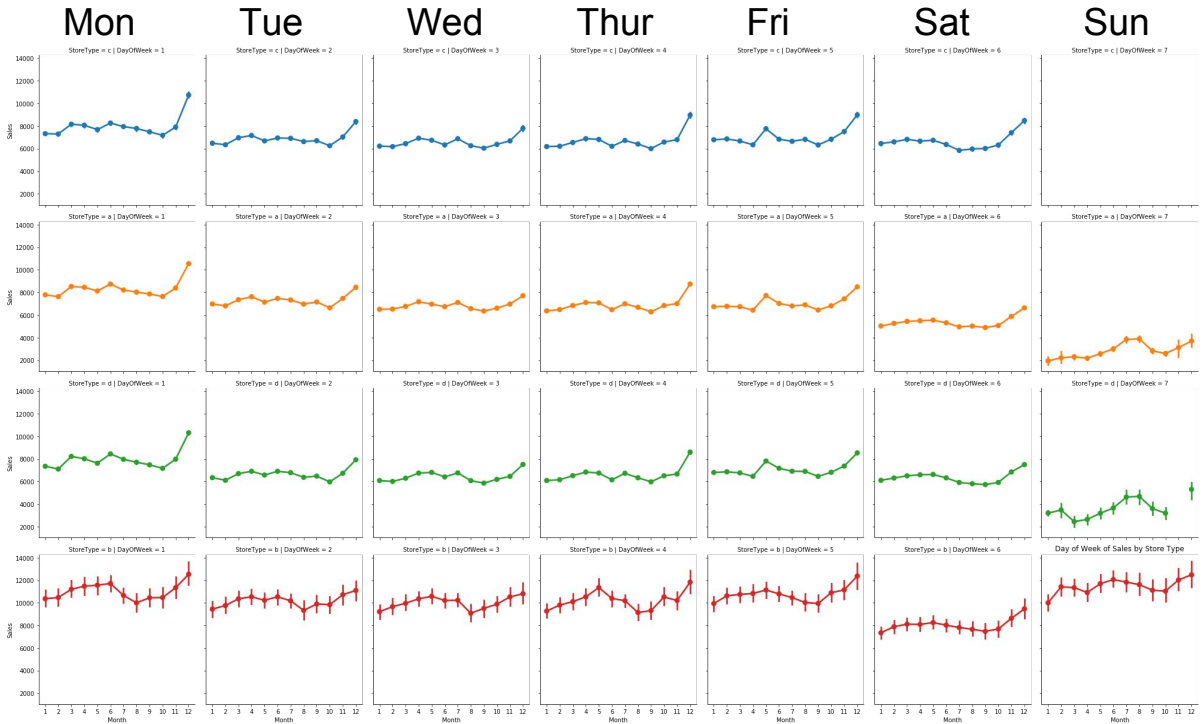
EXPLORATORY ANALYSIS

- Promotion increase sales significantly



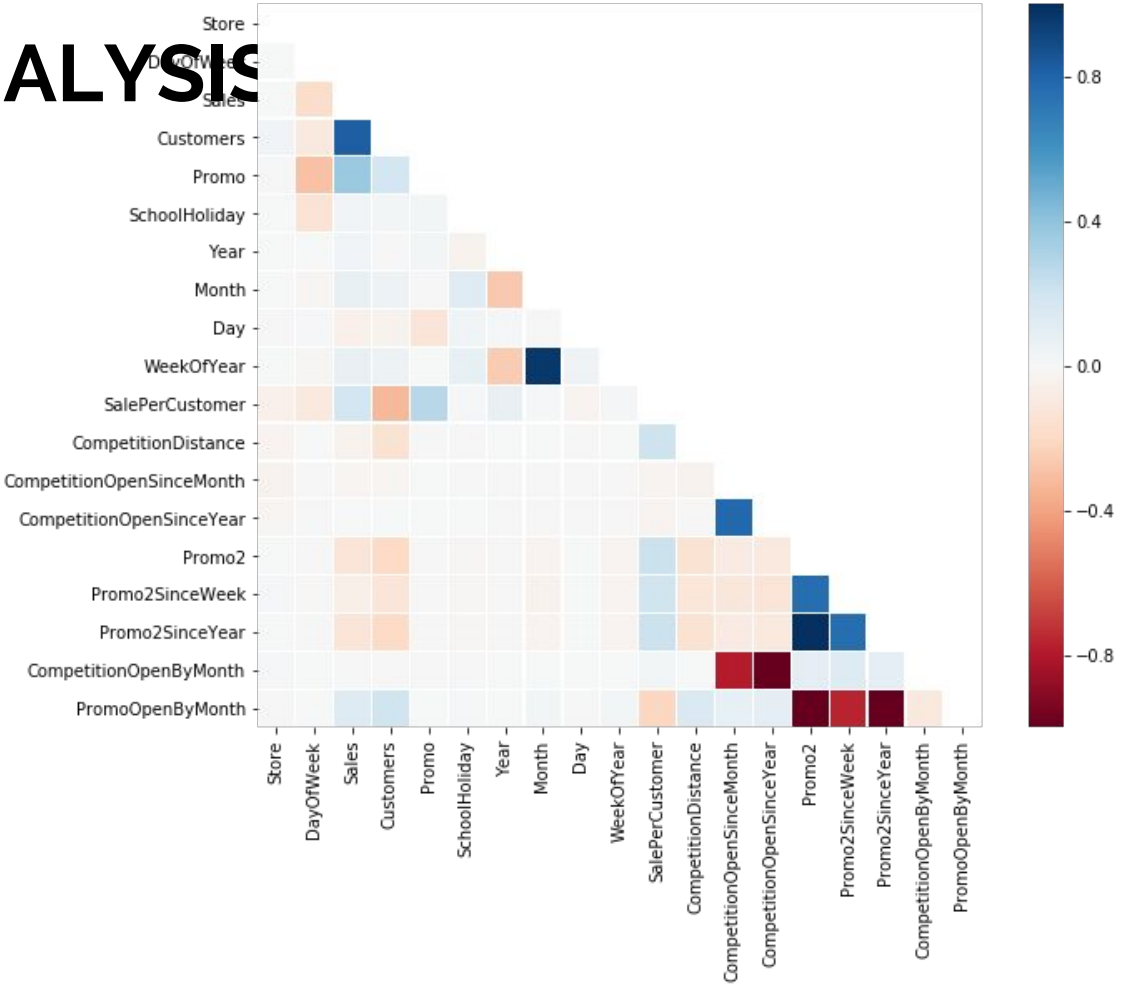
EXPLORATORY ANALYSIS

- Monday tends to has higher sales, while weekend has lower



EXPLORATORY ANALYSIS

- Sales, at column 3 in heatmap, is more correlated (blue) to number of customers and promotion
- week of year and promotion open by month (how long the long term promotion has been running) also have the higher correlation to sales.



FEATURE ENGINEERING

- Breakdown date into Year, Month and Day, add Day of Week
- Convert CompetitionOpenSinceMonth/Year into CompetitionOpenByMonth
- 17 Features for modeling

```
['Store',
 'DayOfWeek',
 'Open',
 'Promo',
 'StateHoliday',
 'SchoolHoliday',
 'Year',
 'Month',
 'Day',
 'WeekOfYear',
 'StoreType',
 'Assortment',
 'CompetitionDistance',
 'Promo2',
 'PromoInterval',
 'CompetitionOpenByMonth',
 'PromoOpenByMonth']
```

DATA PROCESSING

- Label Encoding
  - Convert categorical features into labels
- Data splitting
  - Test set: Last 6 weeks,
  - Train set: reminding
- Hyperparameters Tuning
  - Grid Search for each model

In [10]:

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
tmp = le.fit_transform(df_train_store['PromoInterval'])
df_train_store['PromoInterval'] = tmp

tmp1 = le.fit_transform(df_train_store['StateHoliday'])
df_train_store['StateHoliday'] = tmp1

tmp2 = le.fit_transform(df_train_store['StoreType'])
df_train_store['StoreType'] = tmp2

tmp3 = le.fit_transform(df_train_store['Assortment'])
df_train_store['Assortment'] = tmp3
```

MODELING

- Time series regression problem
- 5 different regressors were used to search for best performing models:
  - Linear Regression, Gradient Boosting, XGBoost, Random Forest and Neural Network
- Prediction target is sales amount
- 29.5 months data as train set, latest 6 weeks of data as train set
- Root Mean Square Percentage Error (RMSPE) is the metric used for performance evaluation

MODELING

- Linear Regression as baseline
- Gradient Boosting performs the best at 17.37% RMSPE

Model	RMSPE
Linear Regression	49.67%
Gradient Boosting	17.37%
XGBoost	17.64%
Random Forest	21.52%
Neural Network	51.92%

CONCLUSION AND IMPROVEMENT

- Add new perspective like weather, seasonal flu epidemic distribution which would help the model to be more predictive in solving the sales prediction problem

- Used Root Mean Square Percentage Error (RMSPE) as performance evaluation metric, we found that the Gradient Boosting regressor gives the the best model performance.

- Achieved the RMSPE to be about 17.34% while the baseline Linear Regression model is 49.67%