Rossmann Data Exploration Report

Liqi Gu, Shuai Jiang, Shengtao Zhong, Runying Jiang
lglu20@soton.ac.uk, sj4n20@soton.ac.uk, sz3g20@soton.ac.uk, rj1u20@soton.ac.uk

Abstract: In this paper, we tried to dig deeper into the data exploratory phase of a real-world problem — drug store sales forecasting. **Rossmann Store Sales Prediction** is a problem based on a time series dataset. A time series is a sequence of observations taken sequentially in time. Our findings were documented from three aspects: problem analysis, data exploration and project reflection. Python data wrangling tools (Pandas, Numpy, Seaborn, Matplotlib etc.) were used to analyze and visualize the dataset.

1. Problem Analysis

1.1. Problem Description

Dirk Rossmann GmbH (usual: Rossmann) is one of the largest drug store chains in Europe with around 56,200 employees and more than 4000 stores across Europe. In the Kaggle competition, competitors were challenged to predict daily sales of six weeks for 1,115 stores located across Germany. The model performance was evaluated by Root Mean Square Percentage Error: RMSPE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{y_i}\right)^2}$.
From time to time, the store will hold short-term promotional activities and continuous promotional activities to increase sales. Moreover, store sales are also affected by many factors, including promotions, competition, school and national holidays, seasonality and periodicity.

1.2. Dataset Description

The dataset can be found online. To start off, a simple preview and statistics of our dataset was generated.

Name	Type	Description
Store	Int	a unique Id for each store
Sales	**Int**	**the turnover for any given day (target)**
Customers	Int	the number of customers on a given day
Open	Categorical	whether the store was open
DayOfWeek	Categorical	indicates the day of week
StateHoliday	Categorical	indicates a state holiday
SchoolHoliday	Categorical	indicates if the data was affected by the closure of public schools
Promo	Categorical	indicates whether a store is running a promo on that day

Table I: Description of train dataset

Table.1 shows the description of the train dataset, which contains the historical sales data from 2013-01-01 to 2015-07-31. There are 1017209 rows and has no missing data. Our target attribute is the sales, which is the turnover for any given day. And in test dataset, the same attributes except for Customers and Sales are given for the period from 2015-08-01 to 2015-09-17.

Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
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

Table II: Preview of train dataset

Store	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday
1	4	2015/09/17	1	1	0	0
3	4	2015/09/17	1	1	0	0
7	4	2015/09/17	1	1	0	0
8	4	2015/09/17	1	1	0	0
9	4	2015/09/17	1	1	0	0

Table III: Preview of test dataset

Apart from the train and test dataset, the competition also provides us with the store dataset, which provides any identity store's attribute.

Name	Type	Description
StoreType	Categorical	indicates the store type
Assortment	Categorical	describes an assortment
CompetitionDistance	Int	distance to the nearest competitor store
CompetitionOpenSince	Date	the approximate year and month of the time the nearest competitor was opened
Promo2	Categorical	a continuing promotion for some stores
Promo2Since	Date	describes the year and calendar week when the store started participating in Promo2
PromoInterval	Categorical	describes the intervals Promo2 is started

Table IV: Description of store dataset

Table.2 shows the supplemental information about the stores, which has 1115 rows with some missing data in features related to Competition and Promo.

Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	Promo2
1	c	a	1270.0	9.0	0
2	a	a	570.0	11.0	1
3	a	a	14130.0	12.0	1
4	c	c	620.0	9.0	0
5	a	a	29910.0	4.0	0
6	a	a	310.0	12.0	0

CompetitionOpenSinceYear	Promo2SinceWeek	Promo2SinceYear	PromoInterval
2008.0	NaN	NaN	NaN
2007.0	13.0	2010.0	Jan,Apr,Jul,Oct
2006.0	14.0	2011.0	Jan,Apr,Jul,Oct
2009.0	NaN	NaN	NaN
2015.0	NaN	NaN	NaN
2013.0	NaN	NaN	NaN

Table V: Preview of store dataset

2. Data Exploration

2.1. Time Series Exploration Data Analysis

In this section, we analyse the trend and seasonality of the data. The trend is the linear increasing or decreasing behavior of the series over time while seasonality is the repeating patterns or cycles of behavior over time.

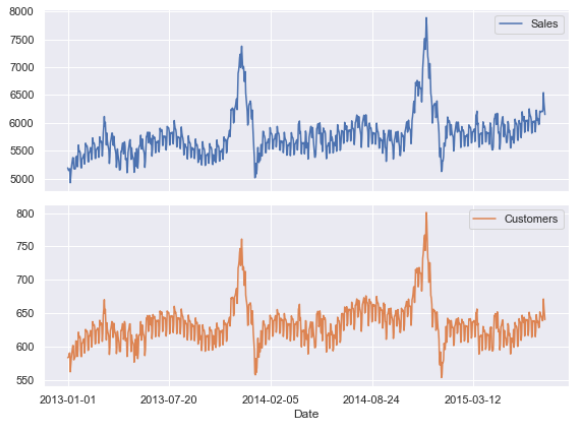

Figure 1: The 30-day Moving Average Sales and Customers

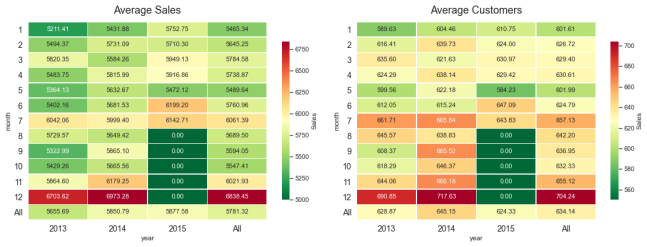

Figure 2: The Heatmap of Monthly Average Sales and Customers

Figure 1 and **Figure 2** show the average sales and customers of each day and month. What can be clearly seen in figures above is the general pattern of monthly average sales and customers that sales and customers usually peak in December and reach a low point in January.

2.2. Categorical Data Analysis

The objective for categorical data analysis is to know the unique values and their corresponding count. As shown in **Figure 3**, most of stores closed on State Holidays and Sundays while School Holidays have some impact on store opening.

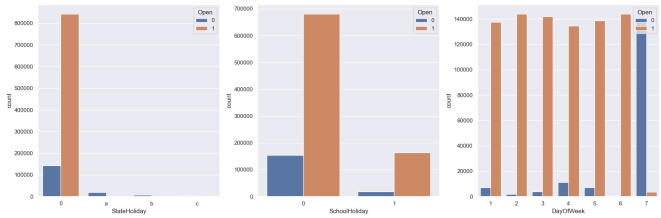

Figure 3: Counts of the openings on holidays and Day of Week: Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None.

As can be seen in **Figure 4**, the market is mainly occupied by Type **a** and Type **d** stores. Moreover, most varieties are basic ones or extended ones. Almost no extra variety.

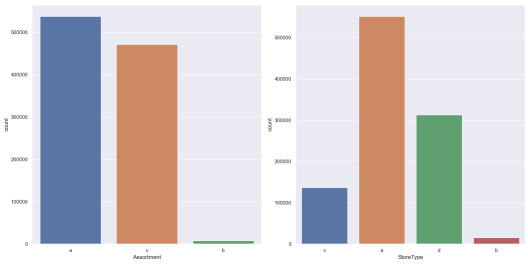

Figure 4: Counts of Assortment and StoreType: Assortment describes an assortment level: a = basic, b = extra, c = extended; StoreType differentiates between 4 different store models: a, b, c, d.

We can learn from **Figure 5** that about half of the stores participate in discount promotions, while the other half do not. In the stores with discounts, the months of their promotions are mainly January, April, July, October.

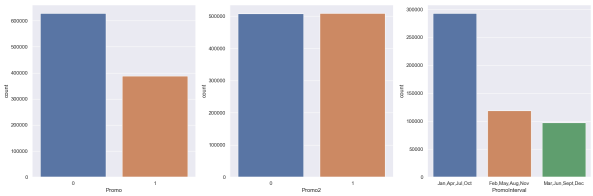

Figure 5: Counts of categories related to Promo : Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating; PromoInterval describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew.

2.3. Numeric Data Analysis

2.3.1. Univariate Distribution of Numeric data

The basic features of numeric data (Sales, Customers, Competition Distance) are provided in the following **Table VI**. To better understand the variates, we go into deeper with the distributions. As shown in **Figure 6**, all the three variates have positively skewed distributions, moreover, they are approximately log-normally distributed.

	count	mean	std	min	25%	50%	75%	max
Sales	83897.0	6071.416999	3834.150546	0.0	4120.0	6048.0	8143.0	41551.0
Customers	83897.0	636.519721	451.684626	0.0	428.0	613.0	824.0	4989.0
CompetitionDistance	83673.0	5404.869193	7659.567708	20.0	720.0	2320.0	6880.0	75860.0

Table VI: Univariate Descriptive statistics of numeric data

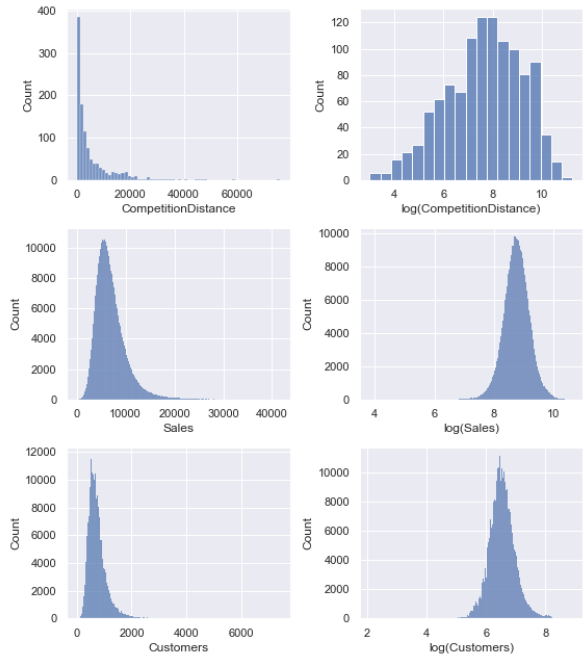

Figure 6: Distributions of CompetitionDistance, Sales and Customers (left side) as well as their log distributions (right side).

2.3.2. Bivariate Distribution of Numeric Data

This section analyzes the distribution of sales across several levels of categorical variables (StoreType, DayofWeek and etc). In the violin plot **Figure 7**, we select Store one as a sample to check the distribution of sales. We can infer that stores closed on state holiday and Sunday, which is the same result we get from the count plots above **Figure 3**. At the same time, it implicates that school holiday has limited influence on sales. The school festival does not increase the number of sales.

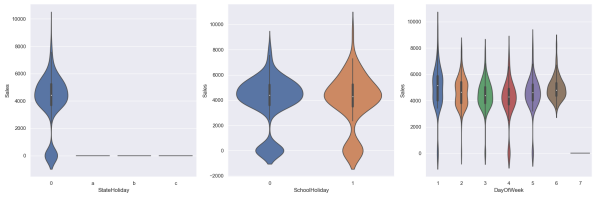

Figure 7: Distribution of Store one sales on holidays and Day of Week

Figure 8 shows that there is almost no difference of sales distribution between Store Type **a** and Store Type **c**, however, stores of Type **b** seems to be open all the year and have better performance in sales. This situation is very similar to extra assortment, although extra assortment has a small proportion.

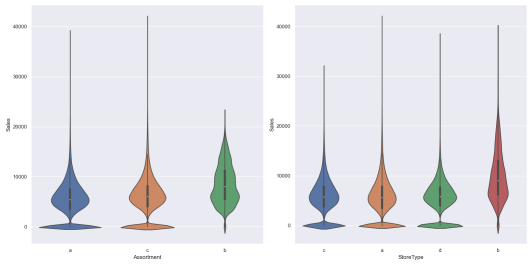

Figure 8: Distribution of sales on Assortment and StoreType

From **Figure 9**, we can find that discounts have a little boost to sales. From the overall situation, whether there are discount activities in shops has almost no impact on sales. Similarly, the discount range has little effect on sales.

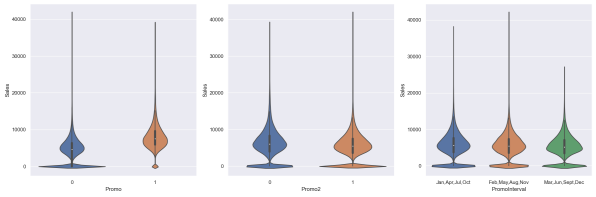

Figure 9: Distribution of sales on Promo, Promo2 and Promo Interval

2.3.3. Pairwise Distribution of Numeric Data

What can be clearly seen in the pairwise joint distribution plot of sales volume and number of customers is that they are positively correlated. There is also a certain degree of correlation between sales volume and competition Distance, Competition Since Year and Promo Since Year.

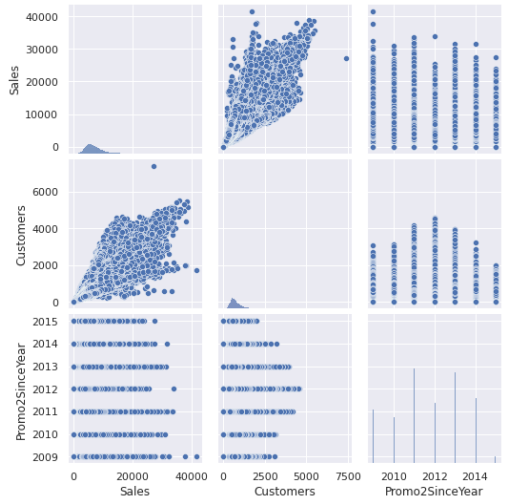

Figure 10: Pairwise Distribution of Sales, Customers, Promo Since Year

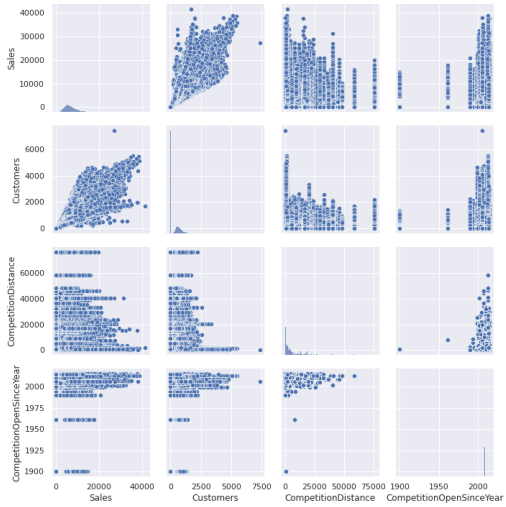

Figure 11: Pairwise Distribution of Sales, Customers, Competition Distance and Competition Since Year

3. Project Reflection

3.1. External Data and Information

Although, in general, the use of external data is prohibited in Kaggle competitions, some public data sets are very reasonable and are indeed helpful for accurate prediction of the model. The followings are the interesting data sets we found to better predict sales:

- **Data on mapping of stores to states**: The data set succeeded in grouping stores in the different state based on the Holidays and Observances in Germany in 2013, since the public holidays in Germany differ state by state. Moreover, this data set can be further used to predict the weather, which is proved to be very effective in predicting sales.
- **Weather data**: Previous research [1] has established that the weather has generally a complex effect on daily sales while the magnitude and the direction of the weather effect depend on the store location and the sales theme. It is shown that weather forecast information improves sales forecast accuracy up to seven days ahead, however, the improvement of the forecast accuracy diminishes with a higher forecast horizon.
- **Google trends data**: Combining standard time series models, the researchers has found evidence that using Google Trends data can enhance the prediction performance of conventional models [2]. The data set can be downloaded as a csv file within a google account.
- **World cup dates**: The dataset of 2014 FIFA world cup dates is the most interesting part. In the final, Germany defeated Argentina 1–0 to win the tournament and secure the country's fourth world title, the first after the German reunification in 1990. Though so far there is evidence to prove that there is a clear relationship between the World Cup and drug store sales, we believe that the influence of significant historical time on drug store sales cannot be ignored.

3.2. Project Outcome

Traditionally, we can use auto-regression to predict time series data. That is only relying on past sales to predict sales for the next six weeks. However, the pattern of sales in the Rossmann data sets is not that clear, and many factors are contributing to sales volume. Therefore, only applying the auto-regression model manually on Rossmann data sets is not wise.

To conclude, the purpose of time series analysis is to understand and observe the random pattern of the sequence, and predict the future value of the sequence based on the observed pattern.

References

[1] F. Badorf and K. Hoberg, "The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores," *Journal of Retailing and Consumer Services*, vol. 52, p. 101921, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969698919303236

[2] B. Fritzsch, K. Wenger, P. Sibbertsen, and G. Ullmann, "Can google trends improve sales forecasts on a product level?" *Applied Economics Letters*, vol. 27, no. 17, pp. 1409–1414, 2020. [Online]. Available: https://doi.org/10.1080/13504851.2019.1686110