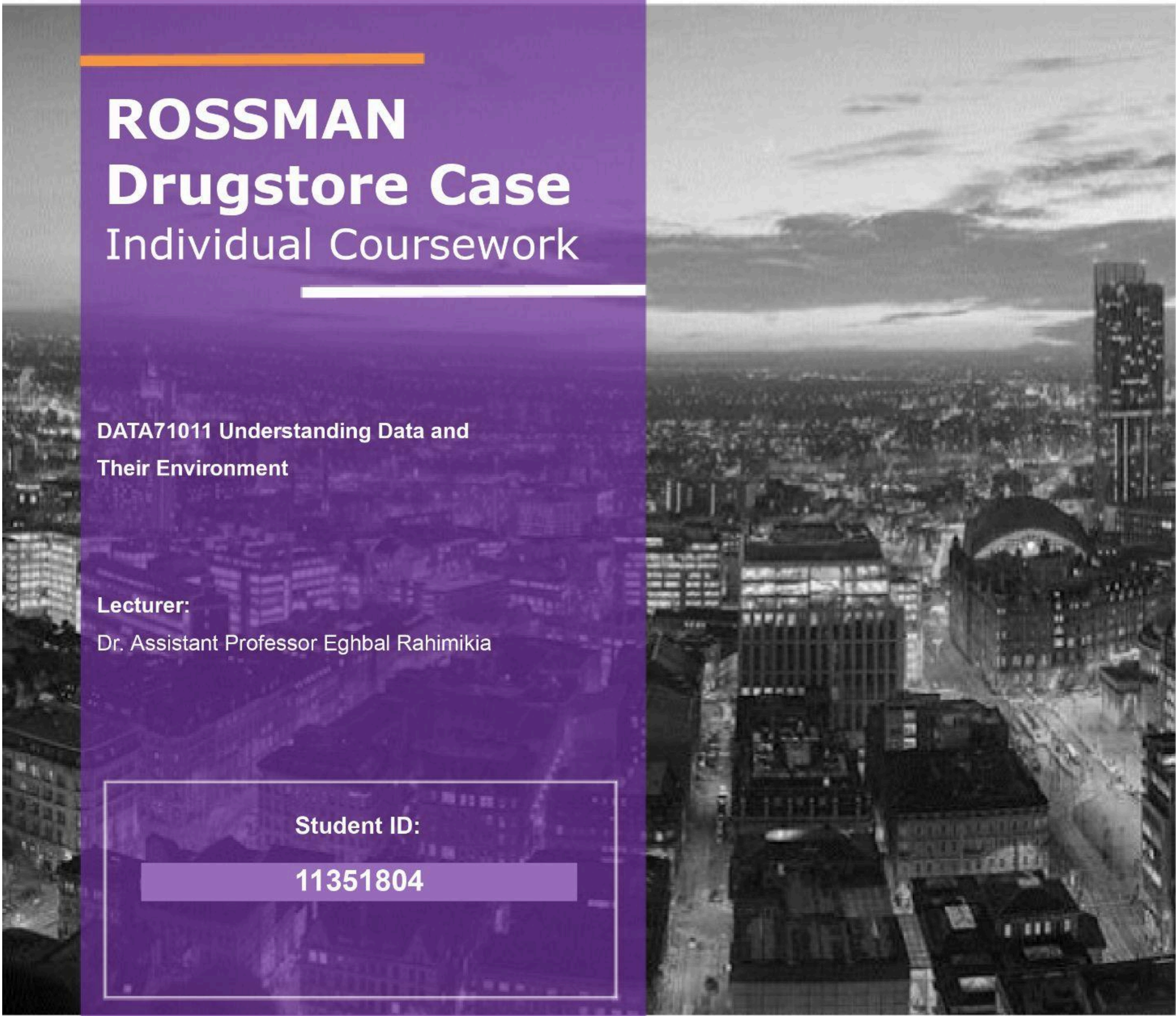
MANCHESTER
1824
The University of Manchester



ROSSMAN Drugstore Case

Individual Coursework

DATA71011 Understanding Data and Their Environment

Lecturer:

Dr. Assistant Professor Eghbal Rahimikia

Student ID:

11351804

2024

Chapter 1: Introduction

1.1 Background & Problem Statement

ROSSMANN, a German drug store chain with 1,115 stores nationwide, faces the challenge of accurately forecasting future sales. Reliable sales predictions are crucial for enhancing store managers' ability to boost overall productivity and profitability, elevate customer satisfaction, and gain a competitive edge in the retail market.

1.2 Scope & Limitation

For this research, several assumptions and limitations are considered:

- External factors, such as force majeure will not be considered.
- All stores have the same characteristics (e.g. geographical, store age) and the same operations strategy.

1.3 Objective

Forecast sales for the period of August 1st to September 17th, 2015, using historical data from January 1st, 2013, to July 31st, 2015 by employing the most appropriate EDA technique, data preprocessing, suitable ML model, and evaluate performance metrics for accurate sales predictions.

Chapter 2: Methodology

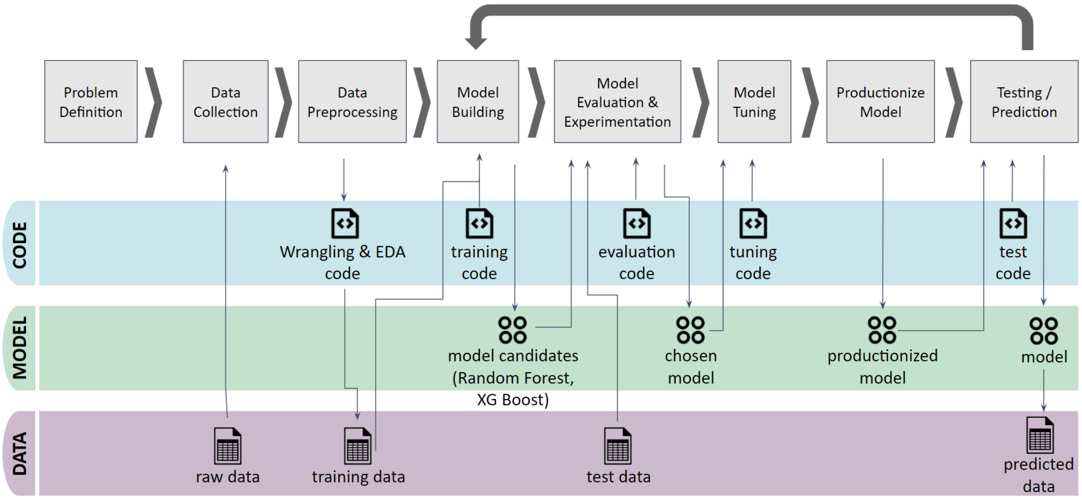


Figure 1 Research Pipeline of ROSSMANN Drugstore

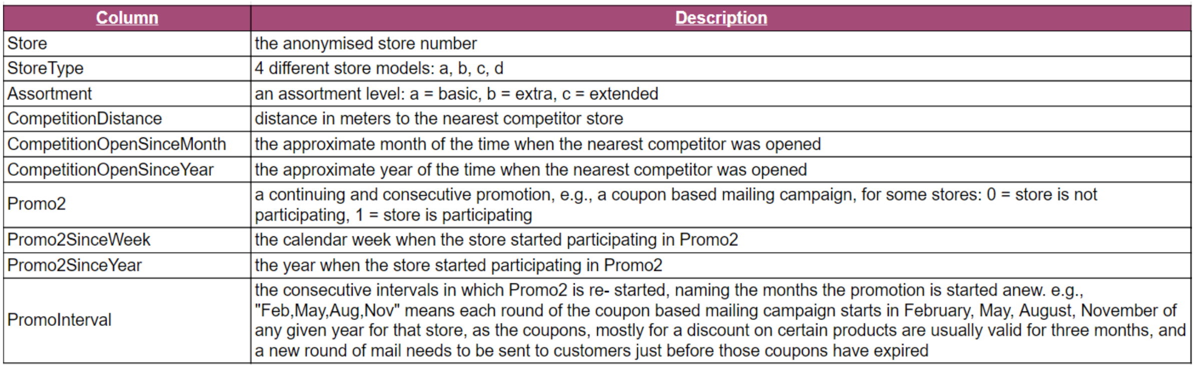
The ROSMANN problem is defined, and data collection involves pre-processing the gathered data for model training. The model-building phase includes training Random Forest Regressor and XGBoost models. Performance assessment, cross-validation & hyperparameter tuning, and testing are conducted systematically, ensuring a complete and efficient ML research pipeline.

Chapter 3: Dataset Review

Three raw datasets were given for this analysis;

a. Store.csv (store_df) : Contains the supplementary information for the 1,115 Rosmann drugstores as displayed on Table 1.

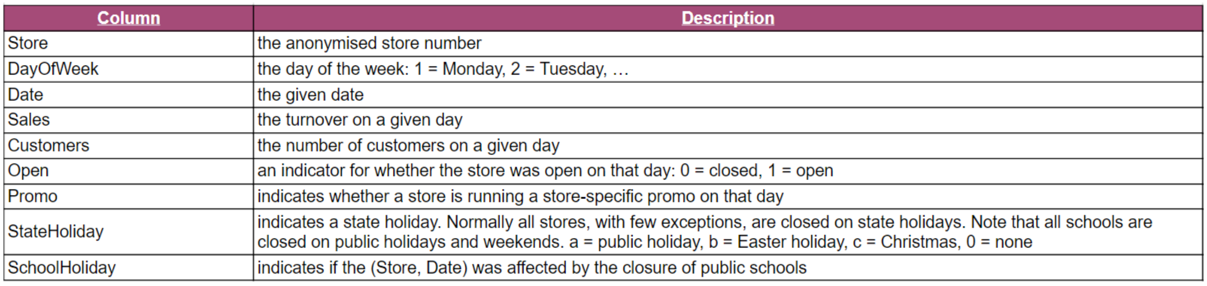
Table 1 Store.csv content and description



Column	Description
Store	the anonymised store number
StoreType	4 different store models: a, b, c, d
Assortment	an assortment level: a = basic, b = extra, c = extended
CompetitionDistance	distance in meters to the nearest competitor store
CompetitionOpenSinceMonth	the approximate month of the time when the nearest competitor was opened
CompetitionOpenSinceYear	the approximate year of the time when the nearest competitor was opened
Promo2	a continuing and consecutive promotion, e.g., a coupon based mailing campaign, for some stores: 0 = store is not participating, 1 = store is participating
Promo2SinceWeek	the calendar week when the store started participating in Promo2
Promo2SinceYear	the year when the store started participating in Promo2
PromoInterval	the consecutive intervals in which Promo2 is re- started, naming the months the promotion is started anew. e.g., "Feb,May,Aug,Nov" means each round of the coupon based mailing campaign starts in February, May, August, November of any given year for that store, as the coupons, mostly for a discount on certain products are usually valid for three months, and a new round of mail needs to be sent to customers just before those coupons have expired

b. Train.csv (train_df) : Contains the historical sales data, which covers sales from 1st January 2013 until 31th July 2015, as displayed on Table 2.

Table 2 Train.csv content and description



Column	Description
Store	the anonymised store number
DayOfWeek	the day of the week: 1 = Monday, 2 = Tuesday, …
Date	the given date
Sales	the turnover on a given day
Customers	the number of customers on a given day
Open	an indicator for whether the store was open on that day: 0 = closed, 1 = open
Promo	indicates whether a store is running a store-specific promo on that day
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

c. Test.csv (test_df) : Identical with Train.csv, except the Sales and Customers columns are unknown for the period 1st August - 17th September 2015.

Please refer to Appendix A for a complete dataset overview.

Chapter 4: Exploratory Data Analysis & Pre-processing

4.1 Data Cleaning & Preparation

4.1.1 Impute missing value on Store_df

Missing values in ‘*Promo2SinceWeek*’, ‘*Promo2SinceYear*’, and ‘*PromoInterval*’ require imputation, but without additional context, accurate imputation is challenging, therefore opting to fill them with 0.

For ‘*CompetitionDistance*’, a different approach was adopted. Due to the skewed distribution as seen on Appendix B, median imputation is preferred over mean. ‘*CompetitionOpenSinceMonth*’ and *'CompetitionOpenSinceYear'* are imputed with the most frequent month and year, respectively, when competitors opened their stores.

4.1.2 Linking Store_df and Train_df

The 'Store' variable is common to both Store_df and Train_df, with 1115 unique values. Hence, linking these datasets using the 'Store' variable is appropriate.

4.2 Exploratory Data Analysis (EDA)

4.2.1 Analysis on Total Sales

In Figure 2, the monthly average of sales across all stores consistently rises each year, reaching the highest peak in December (Christmas), then typically declining in August and rebounding in early September.

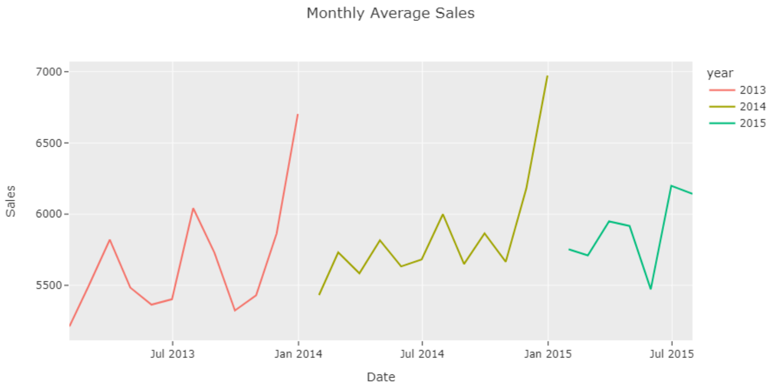


Figure 2 - Monthly Average Sales of Rosmann drugstore

Examining average sales on a weekly basis, the peak sales are always on Monday and lowest on Sunday. Interestingly, if we’re subsetting DayOfWeek based on Promo and consecutive Promo (‘*Promo2*’), Sunday recorded the highest sales in a week.

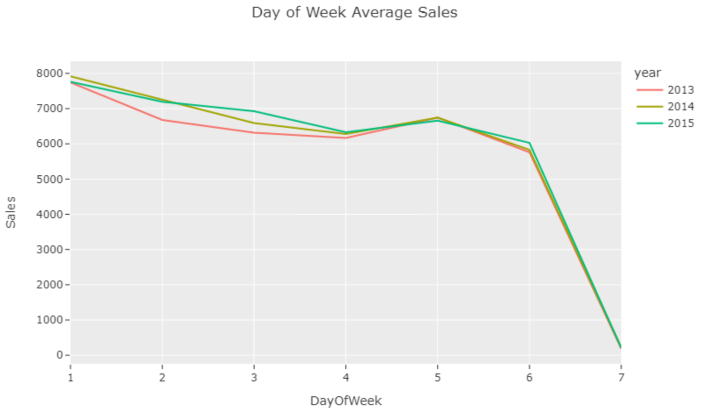


Figure 3 - Day of Week Average Sales of Rosmann drugstore

4.2.2 Sales based on Store Type

Sales based on store type showed an increased trend and hit the peak at Christmas, except Store Type-C who were relatively plummeting as shown on Figure 4. Looking more detailed at daily StoreType sales, there were patterns where Type-A had sales peak every 12-days and Type-B every 7-days.

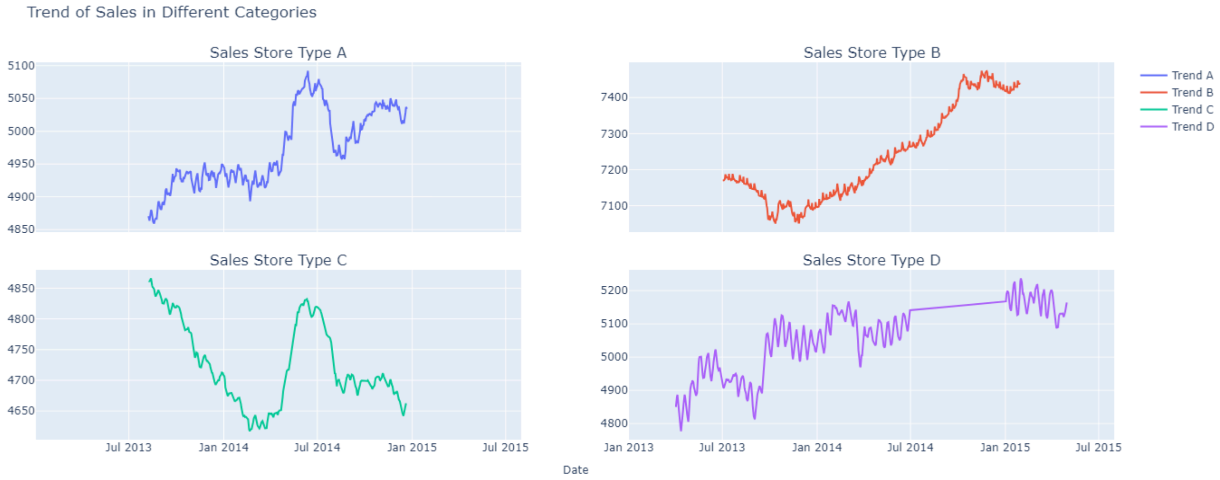


Figure 4 - Sales based on StoreType

4.2.3 Customer Analysis

Customers exhibit a straightforward positive correlation, as seen in Figure 5. However, the causal nature of this relationship will be further validated at the correlation phase.

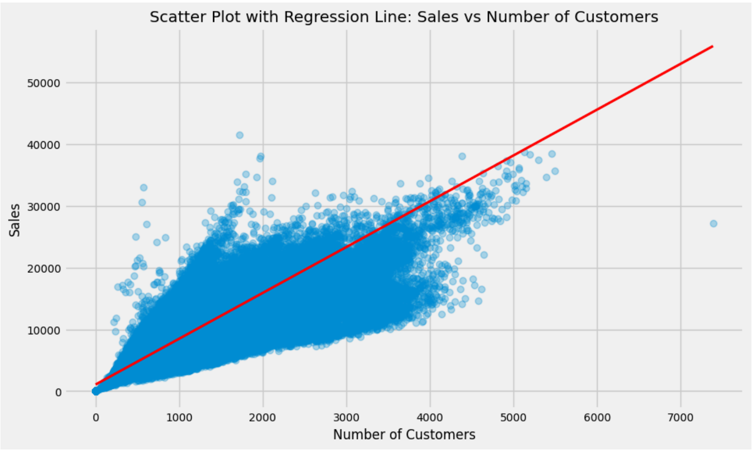


Figure 5 - ‘Customer’ vs ‘Sales’ variables scatter plot with regression line

4.2.4 Promo Analysis

The Promo variable is relatively having a moderate-high impact on the sales, looking at with highest Promo based on the Store Type, it can be seen the most impact to the Sales was during December as shown on Figure 6.

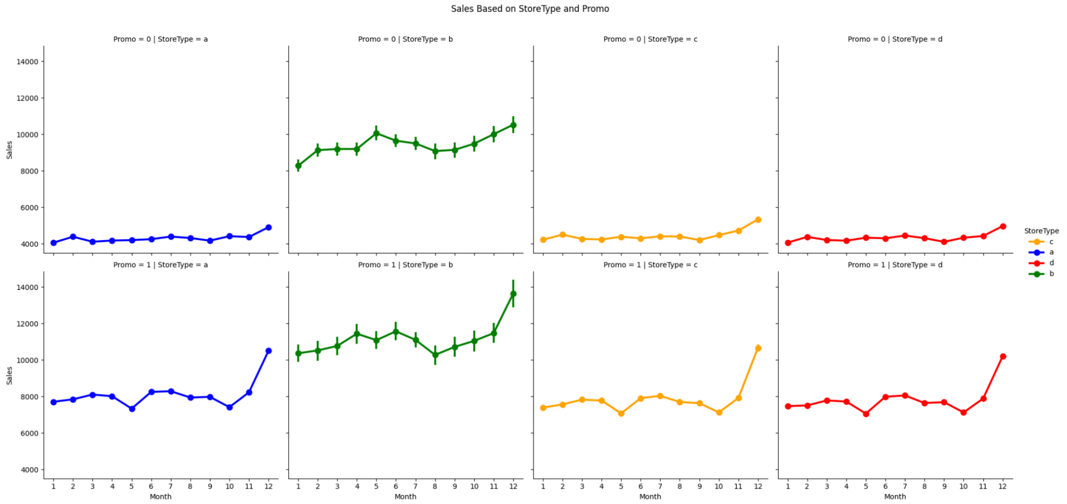


Figure 6 - Sales based on Promo and StoreType (store ‘a’ to ‘d’ from left to right) (top; without promo, bottom; with promo)

4.2.5 Sales Opening Days

Total sales across all stores consistently continue throughout the year, indicated by a constant line at value 1 (Please refer to Appendix C). This suggests that, despite SchoolHoliday/StateHoliday variables, ROSSMAN never stops generating sales.

4.2.6 Store with zero Sales

However, there were stores that opened but no sales on working days. Averagely, stores had 62-days per-year with zero sales, with no consistent pattern (see Appendix-D).

4.2.7 ‘CompetitionDistance’ variable analysis

As seen in Figure 7, CompetitionDistance exhibits a higher sales that often occurs when there is a nearby competitor, indicating a negative correlation with Sales.

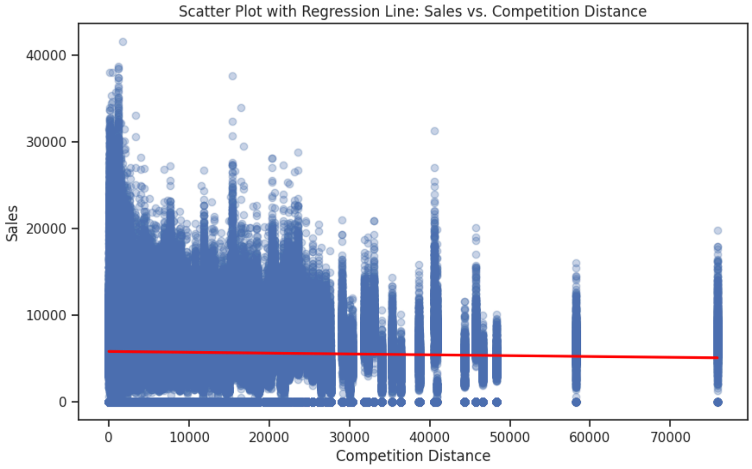


Figure 7 - Scatter Plot with Regression Line: Sales vs CompetitionDistance

4.2.9 Correlation Analysis

As seen in Figure 8, the Sales are having high correlation with Customers, Promo, and Weekly Sales/DayOfWeek, proving more the analysis before. However, the Promo seems not correlated with Promo2 (a consecutive promotion), looks like the promo-continuation is counterproductive compared to seasonal sales strategy.

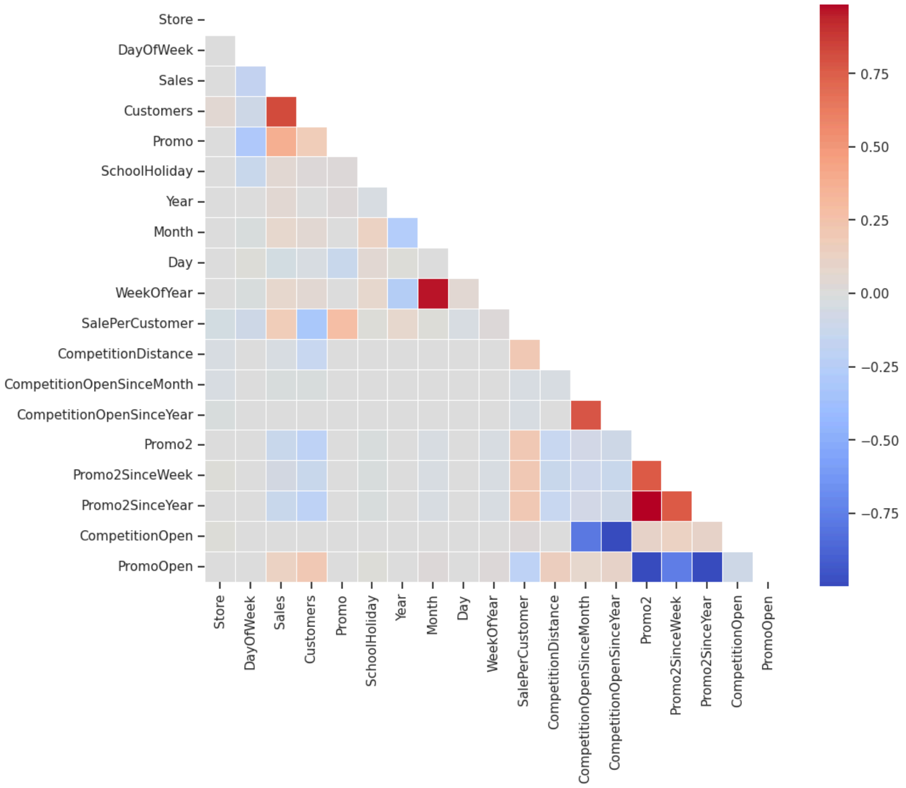


Figure 8 - Correlation Analysis

4.2.10 Conclusion from EDA

ROSSMAN displays unique annual and weekly/day-of-week sales patterns, with consistent daily activity including holidays. Sales can be distinctly differentiated based on its StoreType, there is a negative correlation between Sales with competition distance and stagnan Sales if promo continuation was applied.

4.3 Data Pre-processing

4.3.1 Delete Unnecessary Records

EDA identified variables with unnecessary records that should be deleted, hence we remove closed stores and those with no sales.

4.3.2 Feature Engineering - Sales lag

Effective time-series sales forecasting requires sales lag features (sales from previous time units). As a different sales pattern is distinctly shown by StoreType, the sales_lag is created based on them, Type-A with lags every 12 days, Type-B every 7 days, while Type-B and C are combinations from every 7 and 12 days.

4.3.3 Feature Engineering on the forecast dataset (Test_df)

Test_df underwent similar treatment as Train_df, with the creation of lags variables as predictors in the modelling phase.

4.3.4 Principal Component Analysis

Apply PCA for dimensionality reduction by capturing the maximum variance in the variables.

4.3.5 Encode

Encoding was applied to 'Assortment', 'StoreType', and 'PromoInterval' variables using scikit-learn's label encoder to convert categorical variables into integers, ensuring compatibility with ML algorithms.

4.3.6 Split the Train_df into Train & Validation sets

The Train_df is divided into training (80%) and validation sets (20%) to facilitate robust model evaluation and prevent data leaking.

Chapter 5: Modelling & Prediction

5.1 Modelling and Feature Assessment & Selection

Random Forest Regressor (RFR) and XGBoost (XGB) are selected for time-series experimentation. The initial model ran with default hyperparameters, and evaluated as the results are shown in Table 3.

Table 3 Performance Metrics Evaluation Comparison RFR and XGB

Metrics Evaluated		**RFR**	**XGB**
RMSPE	Train set	9.59%	25.66%
	Validation set	22.42%	23.07%
MAE	Train set	331.63	808.90
	Validation set	923.95	943.70
R2	Train set	0.97	0.85
	Validation set	0.81	0.80

The RFR outperforms XGBoost, RMSPE and MAE showing better performance on the training both sets indicating a lower errors, then also a higher R2, indicating better accuracy and capturing the underlying patterns in both training and validation datasets, hence will be used on the prediction phase.

Feature importance analysis in Table 4 reveals that StoreType and Assortment are less significant in predicting sales. Consequently, these features were removed from the model to prevent overfitting.

Table 4 - RFR Feature Importance sorted from the most to least (before hyperparameter tuning)

Importance level	**Feature (RFR)**
1	sales_lag12
2	sales_lag24
3	sales_lag14
4	sales_lag7
5	Promo
6	DayOfWeek
7	*encoded_StoreType*
8	*encoded_Assortment*

Following that, hyperparameter tuning using time series cross-validation (TSCV) was performed on the RFR model using combinations as shown on Table 5.

Table 5 - RFR Hyperparameter value tested during tuning phase

No	RFR Hyperparameter	Tuning value
1	N_estimators	[50,100]
2	Max_depth	[10,15]
3	Max_features	[0.1, 0.5]

5.2 Prediction and Evaluation

5.2.1 Prediction on Training Set

Using RFR, the model predicts sales figures on the training dataset. Figure 9 illustrates the accurate performance of prediction sales compared to actual sales. Figure 10 provides a detailed view of the model's predictions at the Store level, Store-1 as example.

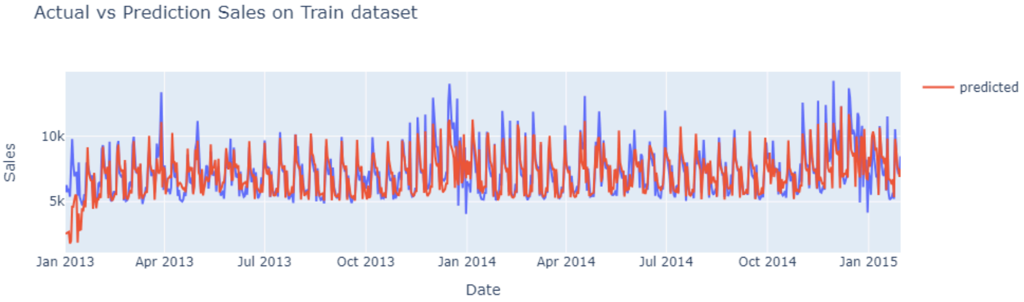


Figure 9 - Actual vs Prediction sales graph (Training set)

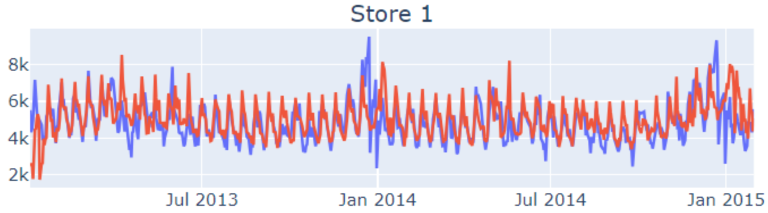


Figure 10 - Actual vs Prediction sales graph at Store-1 level (Training set)

5.2.2 Prediction on Validation set

To assess the model's generalisation ability, sales are predicted on the Validation set (unseen dataset), as depicted in Figure 11 (overall sales) and Figure 12 (store-level sales) that perform a good generalisation, refer to Appendix E for the complete result.

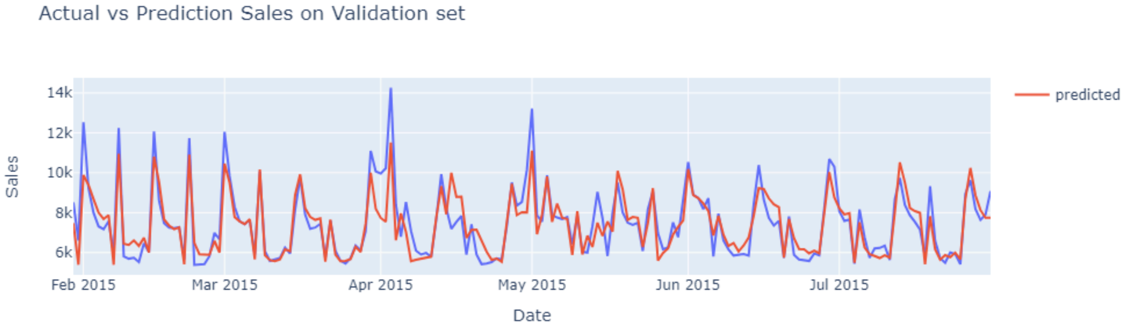


Figure 11 - Actual vs Prediction sales (Validation set)

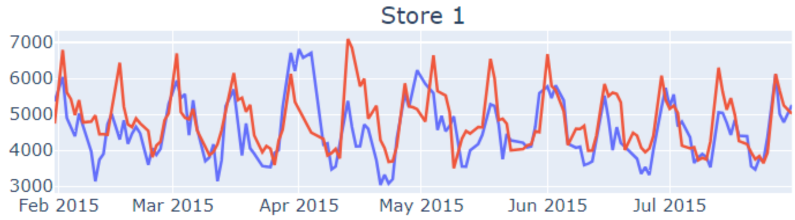


Figure 12 - Actual vs Prediction sales graph at Store-1 (Validation set)

5.2.3 Prediction on the Test set

Run the model on the Test_df that becomes the main objective of this study that will forecast the Sales from 1 August - 17 September 2015 as shown on the Figure 13 and Figure 14.

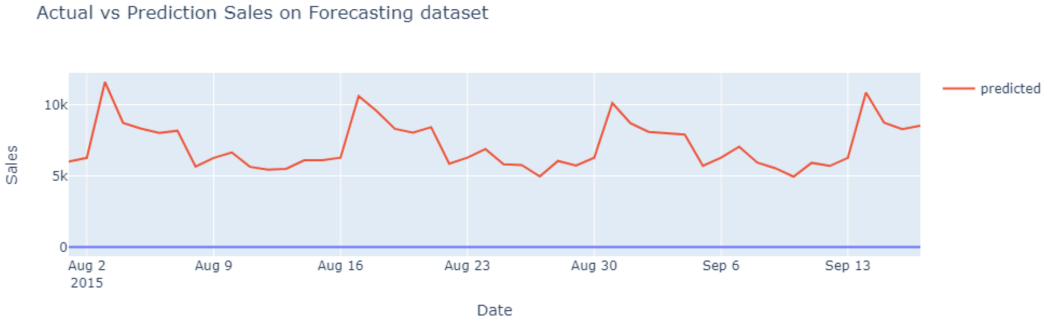


Figure 13 - Actual vs Prediction sales (Test set)

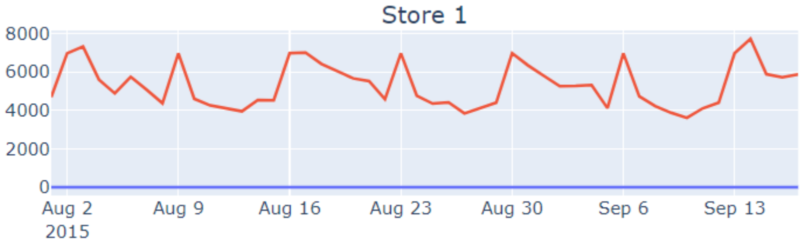


Figure 14 - Actual vs Prediction sales graph at Store-1 level (Test set)

Chapter 6: Conclusion, Implication, and Recommendation

6.1 Conclusion

This research successfully forecasted Rosmann's sales for August 1st to September 17th, 2015, leveraging comprehensive EDA and data preprocessing (Imputation, PCA, Feature Engineering, Encode). Employing RFR and XGBoost models, it achieved accurate predictions, validated through robust metrics like RMSPE, MAE, and R2 using Time Series Cross-Validation. These insights not only provide a reliable sales forecast but also lay the foundation for informed decision-making and strategic planning in the specified period.

6.2 Business Implication

Implementing time-series forecasting holds significant promise for Rosmann, optimising operations, boosting sales, retaining customers, and enhancing overall business performance.

6.3 Recommendations

1. Store-specific modelling
   Developing tailored models for each store (e.g *location, %conversion, and store age*) can effectively optimise the predictions.

2. Optimising time lagging
   A dedicated study is needed to determine the optimal time lagging for accurate forecasts.

3. Assortment and Store Type segmentation
   Specific assortment and store types may exhibit distinct trends due to factors like seasonal variations and could lead to more precise forecasts.

References

Allwright, S. (2022). What is a Good MAE Score? (Simply Explained). Available at: https://stephenallwright.com/good-mae-score/(Accessed: 29 January 2024)

D. Sun, J. Xu, H. Wen, D. Wang. Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and randomforest. Eng Geol, 281 (2021), Article 105972

Gonçalves, João N.C., Carvalho, M.S., Cortez, P. (2020). Operations research models and methods for safety stock determination: A review, Operations Research Perspectives, 7: 100164.

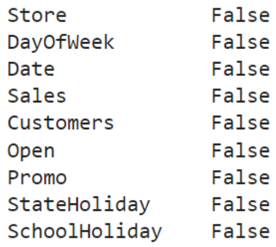
M. Belgiu, L. Drăgu. Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogramm Remote Sens, 114 (2016), pp. 24-31

W. Zhang, D. Lee, J. Lee, C. Lee. Residual strength of concrete subjected to fatigue based on machine learning technique. Struct Concr, 23 (2022), pp. 2274-2287

Appendix

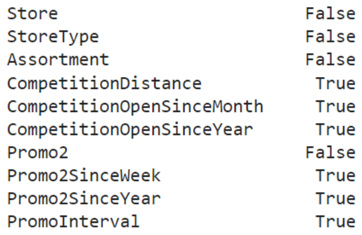
Appendix A

- Missing values at Train_df column:



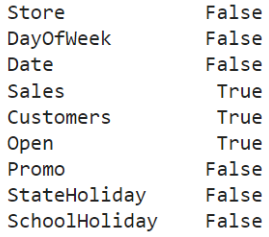
```
Store            False
DayOfWeek        False
Date             False
Sales            False
Customers        False
Open             False
Promo            False
StateHoliday     False
SchoolHoliday    False
```

- Missing values at Store_df column:



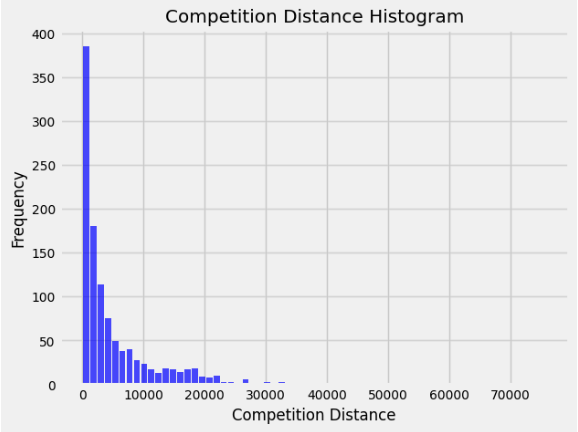
```
Store                        False
StoreType                    False
Assortment                   False
CompetitionDistance           True
CompetitionOpenSinceMonth     True
CompetitionOpenSinceYear      True
Promo2                       False
Promo2SinceWeek               True
Promo2SinceYear               True
PromoInterval                 True
```

- Missing values at Test_df column:



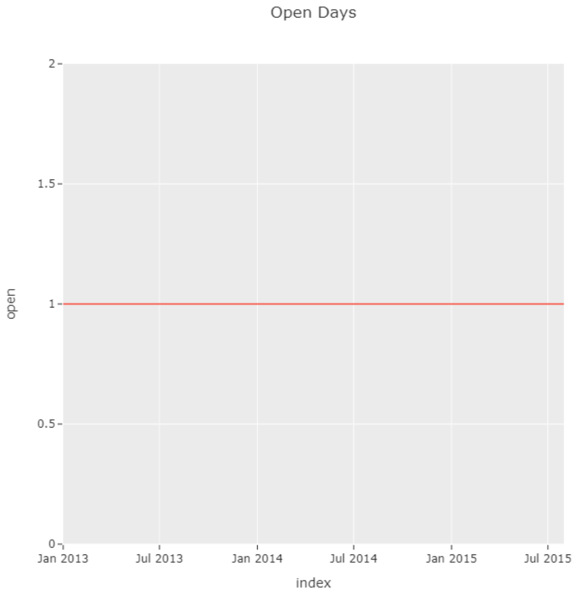
```
Store            False
DayOfWeek        False
Date             False
Sales             True
Customers         True
Open              True
Promo            False
StateHoliday     False
SchoolHoliday    False
```

Appendix B



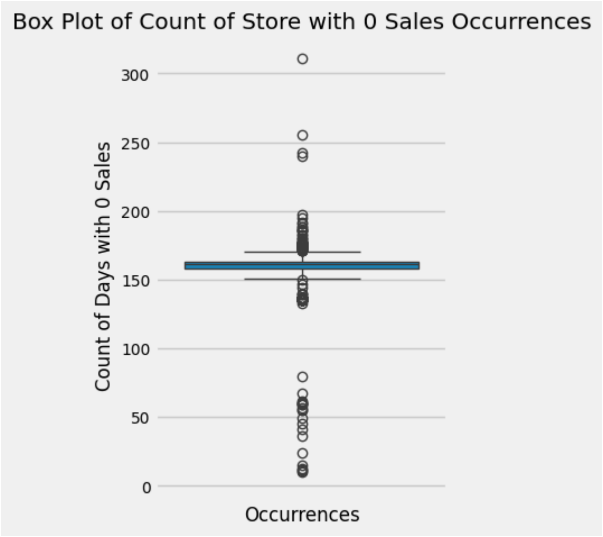
Appendix B - Skewed Histogram of ‘CompetitionDistance’ variable from Store_df

Appendix C



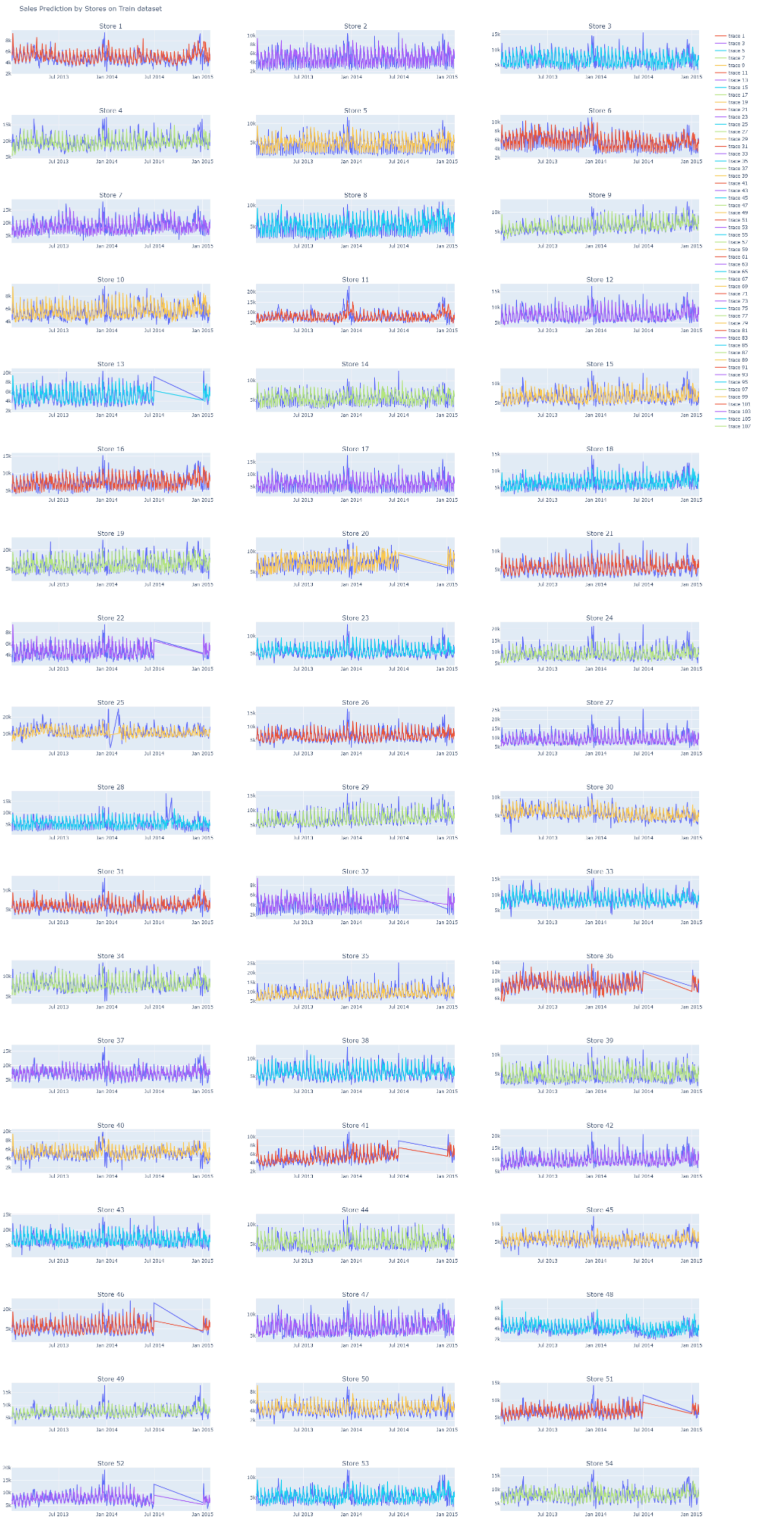
Appendix C - Rosmann Store's Opening Days

Appendix D



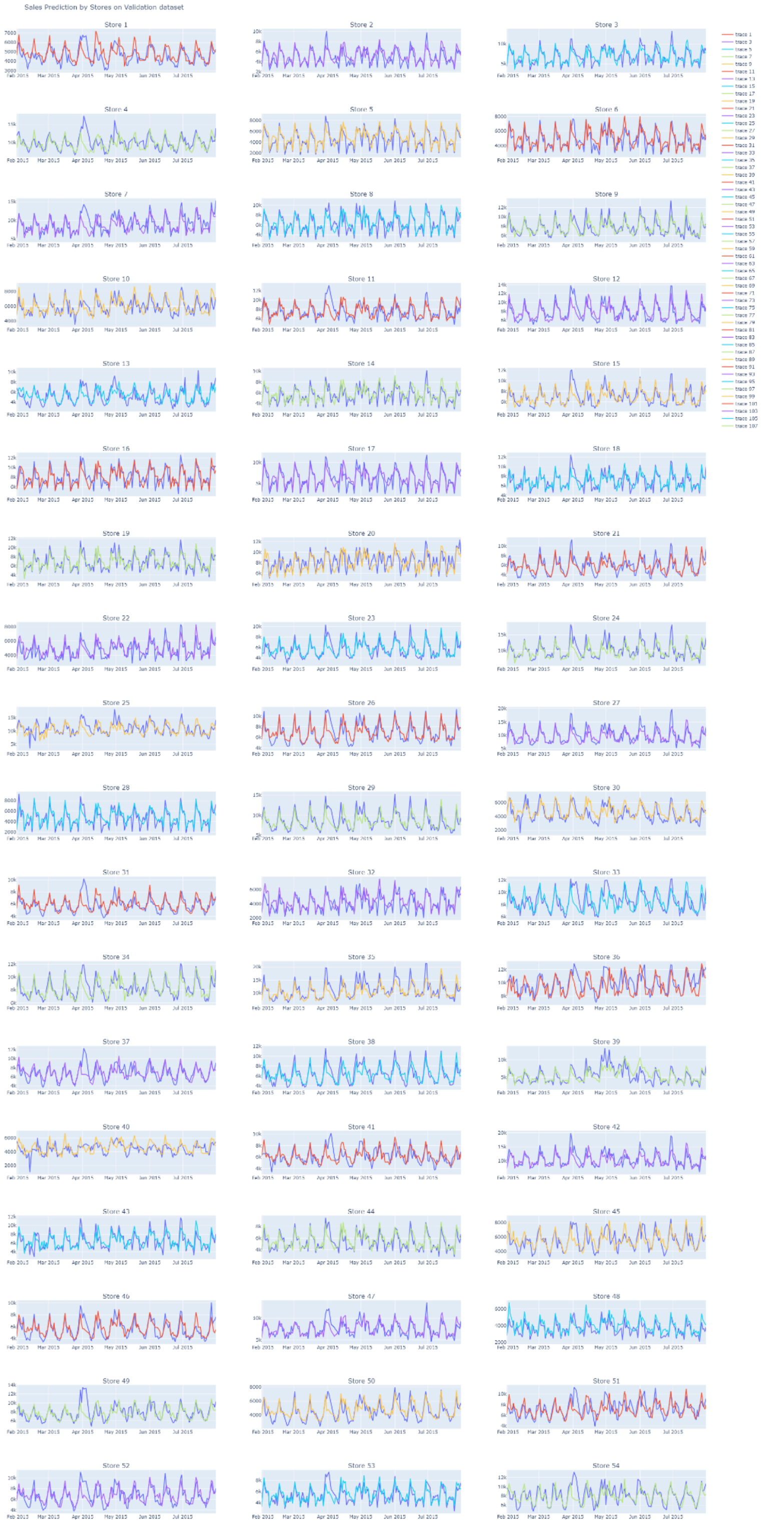
Appendix D - Box Plot of Store with 0 Sales Occurrences

Appendix E



Appendix E - Actual vs Prediction Sales of Training Set

Appendix F



Appendix F - Actual vs Prediction Sales of Validation Set