FEBRUARY 9, 2024

DATA71011 - UNDERSTANDING DATA AND THEIR ENVIRONMENT

COURSEWORK REPORT

11356880

1. Introduction

This report analyses the Rossmann Sales Dataset, focusing on the critical phases of data pre-processing, namely missing values, cleaning, transformation, integration, and feature engineering to enhance the accuracy of sales forecasting in the retail domain. This report addresses missing values, applying targeted imputation techniques to preserve the integrity of the dataset for subsequent analysis. Data cleaning efforts remove extraneous information, and transformations adapt the dataset for analytical use.

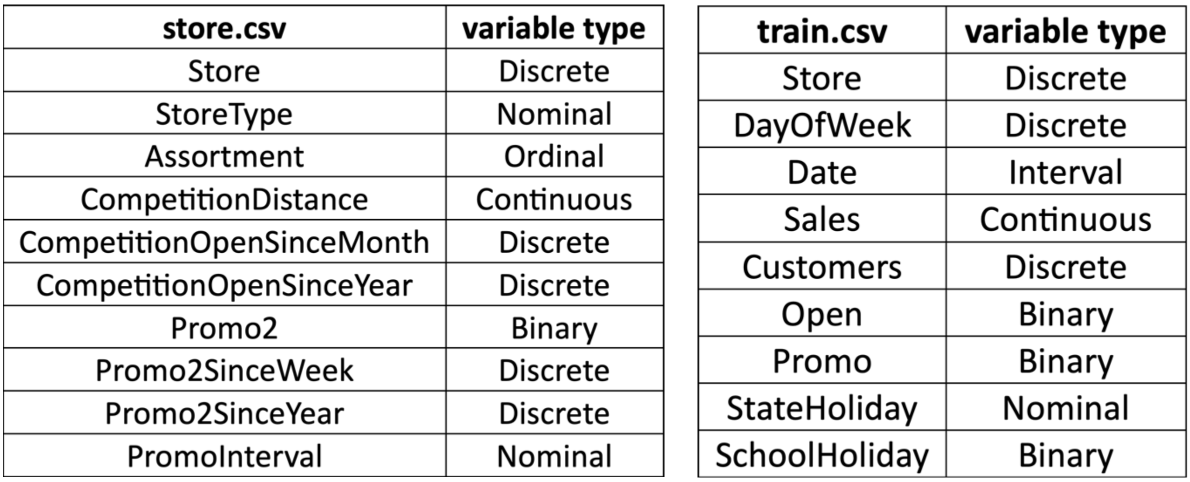
Through feature engineering, the analysis identifies and leverages inherent seasonal and trend patterns within the sales data, essential for developing an effective forecasting model. The selection of the Extreme Gradient Boosting model, aka, XGBRegressor model reflects a strategic decision, motivated by its robustness to multicollinearity, efficiency with large datasets, and built-in regularization to prevent overfitting, demonstrating a comprehensive and methodical approach to overcoming the dataset's analytical challenges.

This attempt not only showcases predictive analytics techniques but also aims to contribute to the strategic decision-making and operational efficiency in the retail sector by providing a reliable sales forecasting model.

2. Data Pre-Processing

2.1 Data Description

Three datasets were provided which contained information of historical sales and stores from 01/01/2013 to 31/07/2015 in *train.csv* and 01/08/2015 to 17/09/2015 in *test.csv*. The *store.csv* details 1,115 stores, including their assortment types, competitor proximity, and promotional activities. The *train.csv* contains daily sales and customer numbers, alongside information on store operations and holiday effects.

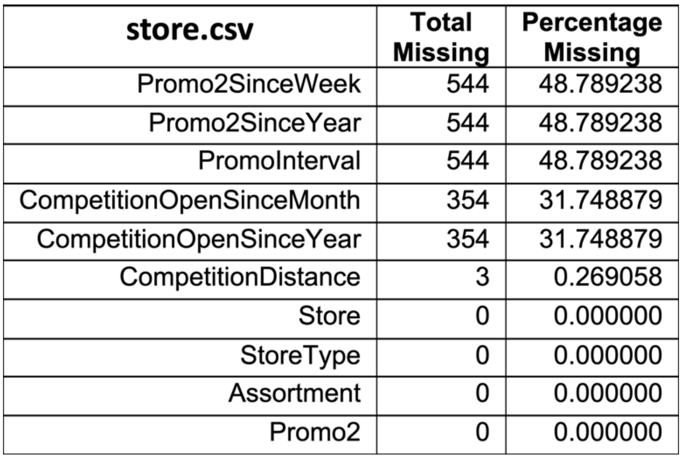


store.csv	variable type
Store	Discrete
StoreType	Nominal
Assortment	Ordinal
CompetitionDistance	Continuous
CompetitionOpenSinceMonth	Discrete
CompetitionOpenSinceYear	Discrete
Promo2	Binary
Promo2SinceWeek	Discrete
Promo2SinceYear	Discrete
PromoInterval	Nominal

train.csv	variable type
Store	Discrete
DayOfWeek	Discrete
Date	Interval
Sales	Continuous
Customers	Discrete
Open	Binary
Promo	Binary
StateHoliday	Nominal
SchoolHoliday	Binary

Table 1 – Given *store.csv* and *train.csv* datasets

2.2 Handling Missing Values



store.csv	Total Missing	Percentage Missing
Promo2SinceWeek	544	48.789238
Promo2SinceYear	544	48.789238
PromoInterval	544	48.789238
CompetitionOpenSinceMonth	354	31.748879
CompetitionOpenSinceYear	354	31.748879
CompetitionDistance	3	0.269058
Store	0	0.000000
StoreType	0	0.000000
Assortment	0	0.000000
Promo2	0	0.000000

Table 2 – store.csv Missing Values Information

In this data pre-processing stage, specifically handling missing values within the *store.csv* dataset, *Promo2SinceWeek*, *Promo2SinceYear*, and *PromoInterval* variables had significant missing data. In business context, these variables are critical for understanding promotional strategies to boost sales. It was observed that whenever Promo2 = 0, these values were missing, categorising them as Not Missing At Random (NMAR). I chose to impute 0 for the missing values to maintain the dataset's integrity and usefulness.

For *CompetitionOpenSinceMonth* and *CompetitionOpenSinceYear*, approximately 32% of the values were missing. These values were imputed using MICE (Multiple Imputation by Chained Equations) technique, enabling a nuanced fill-in based on the dataset's broader patterns. These values can be categorised as Missing At Random (MAR) working on the assumption that stores may not know when their competition opened.

The handling of *CompetitionDistance* missing values, where only 3 values are unreported, I work on the assumption that the absence indicates a significant distance from competitors. To reflect this in the dataset without distorting the overall data distribution, missing values were imputed with a value higher than the highest recorded *CompetitionDistance*.

train.csv	Total Missing	Percentage Missing
Store	0	0.0
DayOfWeek	0	0.0
Date	0	0.0
Sales	0	0.0
Customers	0	0.0
Open	0	0.0
Promo	0	0.0
StateHoliday	0	0.0
SchoolHoliday	0	0.0

Table 3 – train.csv Missing Values Information

Following with evaluating *train.csv* dataset, which did not exhibit missing values, simplifying the pre-processing steps for this part of the data. This detailed focus on handling missing values lays a solid foundation for the subsequent data analysis and model training phases, potentially increasing the forecasting model's accuracy and reliability.

2.3 Data Cleaning

In the Data Cleaning stage, I refine the *train.csv* dataset by removing data entries for days when stores were closed, as these instances contribute no value to sales forecasting due to zero sales on such days. By focusing only on days with stores operating and sales above zero, it enhances the dataset's relevance for the analysis. This step includes the removal of the *open* column post-filtering, streamlining the dataset for more efficient processing.

2.4 Data Transformation

In the Data Transformation stage, I've adjusted key variables in *store.csv* for analytical precision. *Promo2SinceWeek*, *Promo2SinceYear*, *CompetitionOpenSinceMonth*, and *CompetitionOpenSinceYear* were converted to integers to accurately reflect time components. The *date* column in *train.csv* was transformed into a datetime format, facilitating time-based feature extraction in later stages. To enhance model interpretability and performance, *StateHoliday, StoreType* and *PromoInterval* underwent One-Hot encoding, providing a clear, binary distinction for the model. Additionally, *Assortment* was treated with Ordinal Encoding, recognising the inherent hierarchy within assortment levels to aid the model's understanding.

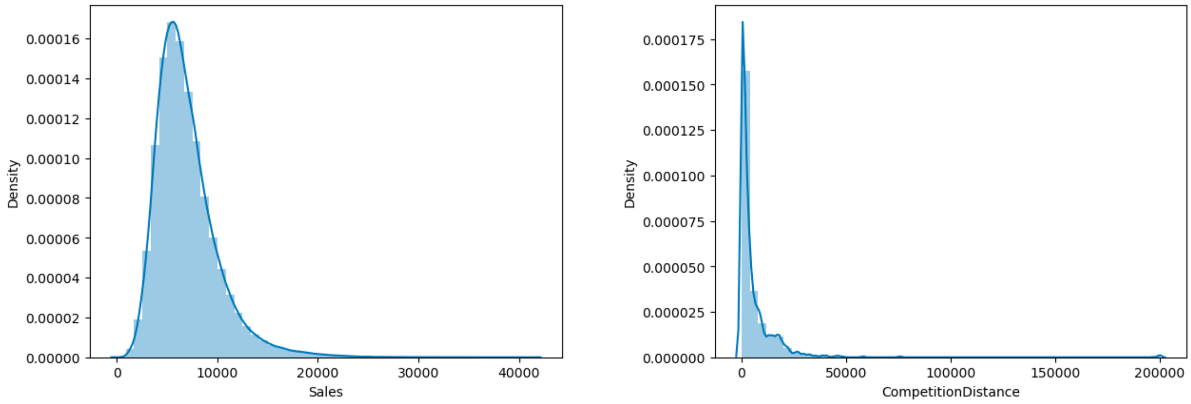


Figure 1 – KDE and Histogram Plot for *Sales* and *CompetitionDistance*

Lastly, a Log Transformation and Box-Cox Transformation was applied to *Sales* and *CompetitionDistance* respectively to reduce the influence of outliers, reduce skewness as seen by the Figure-1 above, and optimizing model's performance, ensuring the forecasting approach is both refined and robust.

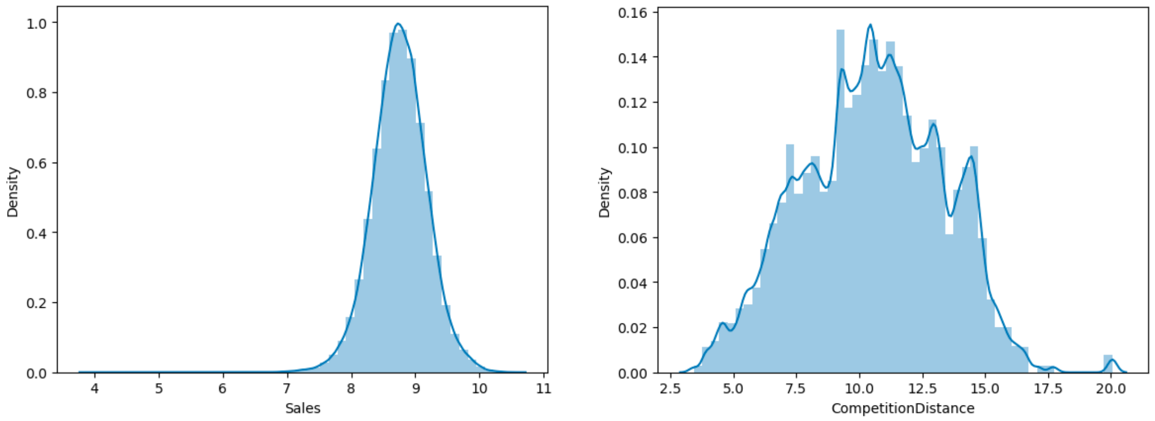


Figure 2 – *Sales* and *CompetitionDistance* after Transformation

2.5 Data Integration

Following the Data Transformation, I merge the *train.csv* and *store.csv* datasets. This integration leverages the *store* column, a shared identifier across both datasets, to merge them accurately. By setting the date as the index post-merger, I transform the combined dataset into a time series format, optimally positioning it for the forthcoming model training stage.

3 Feature Engineering

3.1 Feature Extraction

Leveraging the *date* index in *train.csv*, I've extracted critical time components—month, day of the month, weak of year, and year—to enhance the model's ability to predict sales and by observing repeating trends.

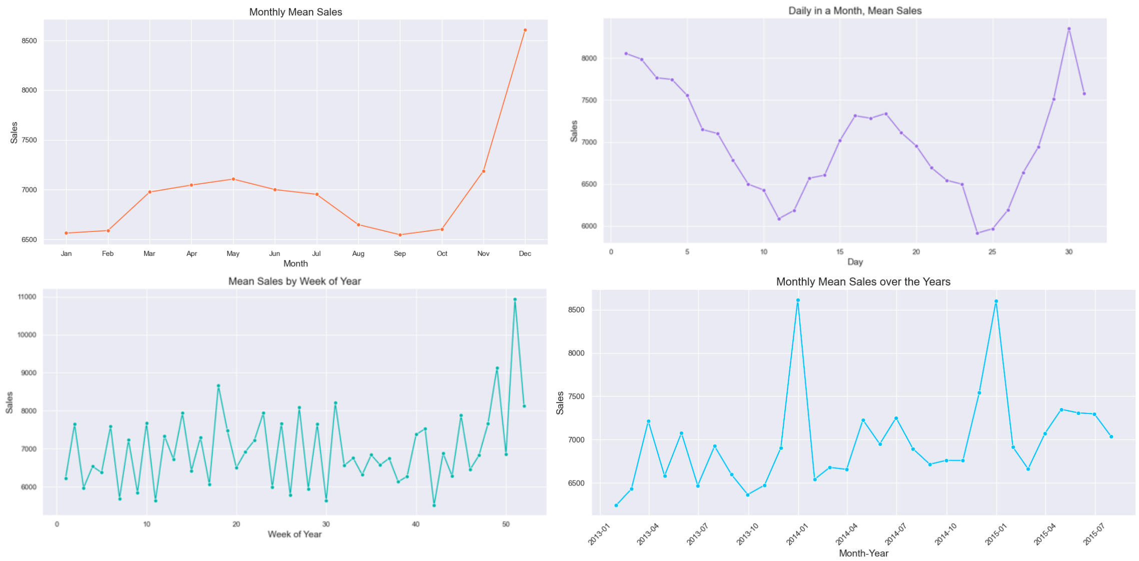


Figure 3 – Graphs Showing Trends/Seasonality in derived time components

This extraction, supported by the graphs above allowed me to observe distinct patterns and trends through time series analysis, where mean sales visibly fluctuate with these time variables. By incorporating *month*, *day*, *week_of_year* and *year* as features, the model adeptly captures the inherent seasonality and trend in the sales data, potentially improving its forecasting accuracy.

3.3 Feature Selection

In the feature selection process, I omit the *customers* variable from training due to its absence in the *test.csv*, acknowledging the impracticality of predicting store visits in advance. I then applied Pearson Correlation Coefficient to evaluate the correlation between independent and dependent variables.

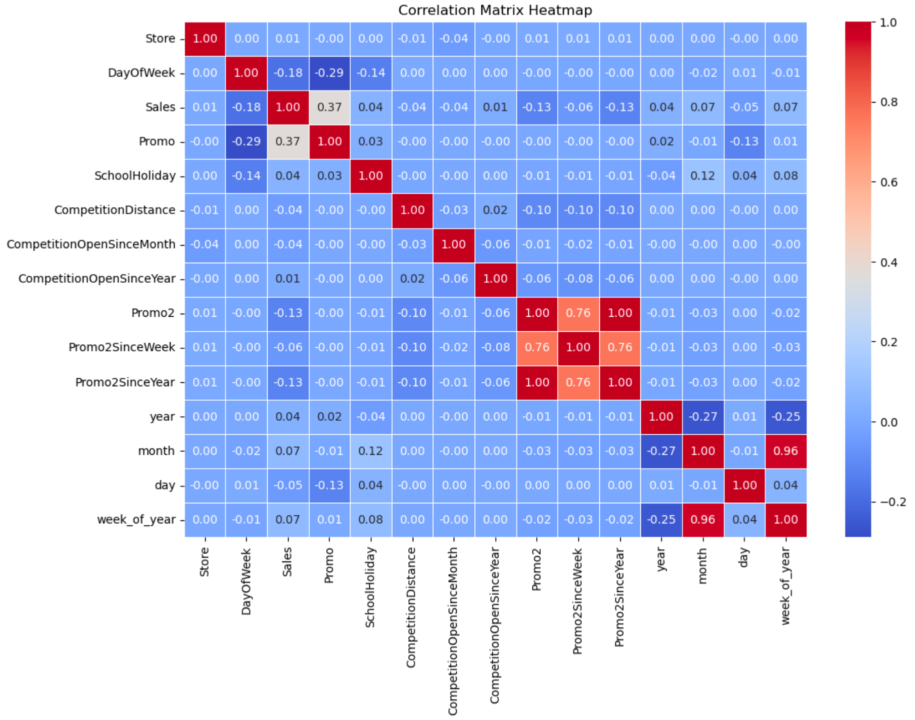


Figure 4 – Correlation Heatmap

Visualized through a correlation heatmap, the figure above aids in identifying impactful features. The *Promo2SinceWeek* and *Promo2SinceYear* having strong multicollinearity with *Promo2*, were removed. *week_of_year* was also removed as it has strong multicollinearity with *month*. Furthermore, the introduction of encoded dummy variables for categorical data raises concerns of multicollinearity, a challenge I plan to address in the modelling section with an appropriate choice of Model selected for forecasting.

4 Forecasting Sales

4.1 Model Selection

For forecasting the sales, I chose XGBRegressor model due to its multiple strengths suitable for this dataset's challenges. Its robustness to multicollinearity makes it ideal for handling encoded dummy variables, crucial given the pre-processing steps. The efficiency and effectiveness of XGBRegressor, alongside its quick training times, facilitate handling the dataset's volume without sacrificing performance. Its capability to manage various data types aligns great with the dataset's condition. The inclusion of regularization within the model parameters aids in avoiding overfitting, enhancing the model's reliability. Analysing feature importance is streamlined through the model's built-in functions, offering insightful analytics on driving factors of sales. Validation was conducted through hold-out method, using a portion of the training data as a validation set. TimeSeriesSplit enabled cross-validation was used that respects the chronological order of data, preventing the model from using future sales to predict past sales. Hyperparameter tuning through RandomisedSearchCV further refined the model, ensuring optimized performance. This comprehensive approach supports the choice of XGBRegressor as a fitting solution for sales forecasting challenge, balancing technical capability with practical application needs.

4.3 Model Evaluation

The performance of the sales forecasting model was evaluated on 3 training folds and the validation set using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). MAE and RMSE, being in the same units as the target variable (*sales*), allow for direct assessment of prediction deviations, while MAPE offers insights into the relative accuracy of the predictions, making it easier to understand the model's performance in terms of percentage errors.

Training Folds	RMSE	MAE	MAPE
Fold 1	783.8451722399269	522.5773278805857	0.0790311517139918
Fold 2	737.8145726073258	495.67395258460317	0.07440243991900326
Fold 3	757.3544605178033	533.6064981955474	0.07721373517547749

Table 4 – Model's performance on Training Set

	RMSE	MAE	MAPE
Validation Set	1010.8943306973766	533.6064981955474	0.07721373517547749

Table 5 – Model's performance on Validation Set

The increase in RMSE is expected as models tend to perform slightly worse on unseen data. However, MAE and MAPE being exactly same in Validation and 3rd fold suggested that the model has a stable average error and average percentage error. This indicates, on average, the model shows relatively stable performance and good generalisability. The built-in feature importance function was used to extract the same, and it shows almost all the features being used and having importance in the model's performance in the figure below.

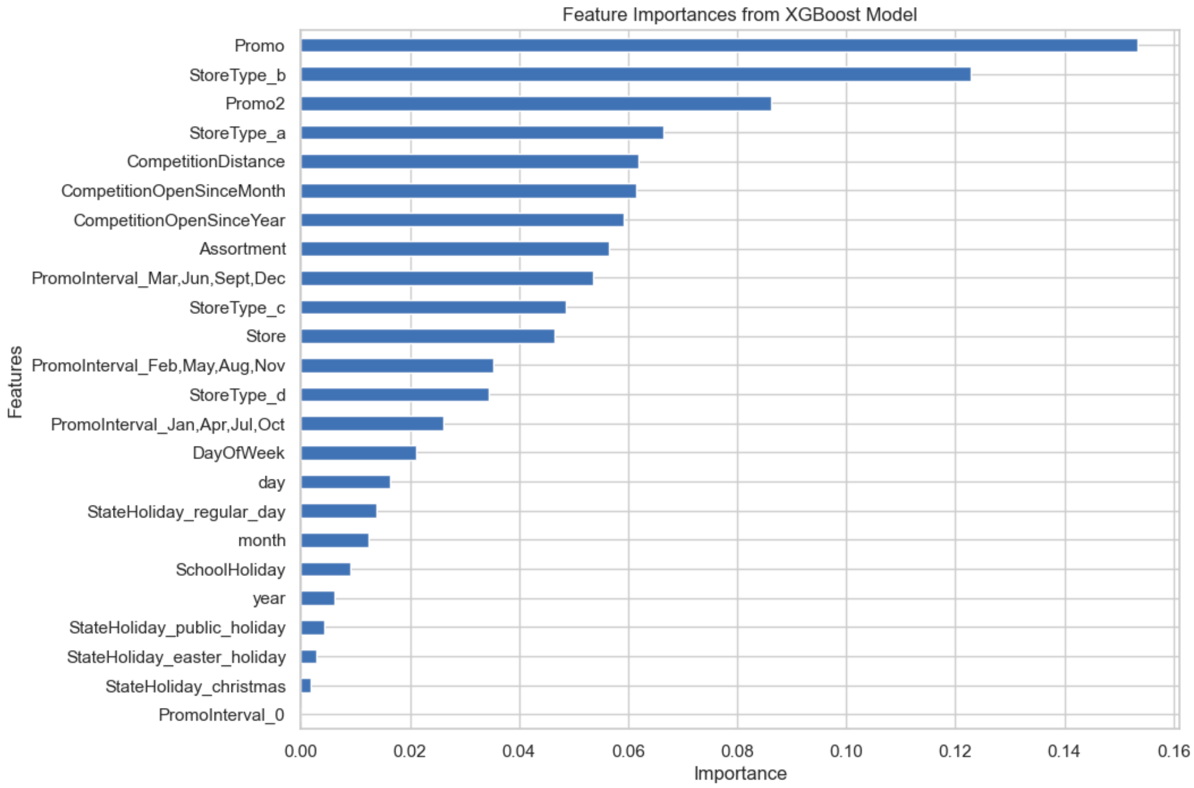


Figure 5 – Feature Importance Graph of used Features

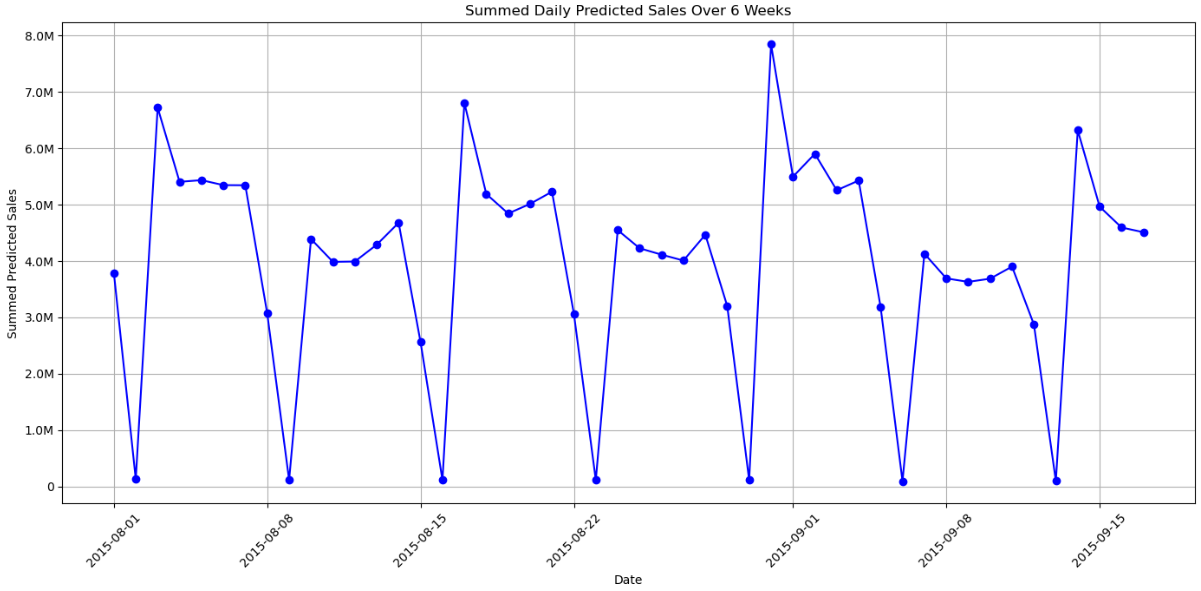


Figure 6 – Graph of Predicted Sales for *test.csv*

Figure-6 shows the predicted sales summed up for every day on *test.csv*. This shows a repeating pattern on a weekly basis where Sunday is generating the least revenue because of having very less stores open and Monday being the best sales day. This follows with the weekly trend that I observed with the time components for training dataset, hence giving an indication that predictions follow the general trend, and they are relatively stable.

5. Conclusion and Recommendations

The analysis presented in the report highlights the significance of in-store promotion as key driver of sales with the Rossmann drugstore chain. It is recommended that Rossmann capitalise on this insight by customising and enhancing store-specific promotions to boost sales effectively. Furthermore, the data suggests an opportunity to increase sales by extending working hours on Sunday due to its current underperformance.

Seasonal trends also emerge as crucial sales determinants, with end of the year and month showing a marked increase in sales volumes. This seasonal insight should prompt Rossmann to adjust stock levels accordingly to meet the heightened demand during these periods.

However, the model's predictive scope is inherently limited by the data quality, information provided and computational resources at-hand. For more nuanced forecasting, future exploration could benefit from incorporating additional data such as customer feedback, social media sentiment analysis to gauge consumer attitudes and preferences. Additionally, considering the impact of local occurrences of events and weather data on consumer behaviour could offer valuable insights into sales fluctuations.

By addressing these recommendations and potential areas for improvement, Rossmann can enhance its strategic planning and operational efficiency, leveraging data-driven insights for optimized sales performance.