Rossman Stores Forecasting Sales Project

ADS1002 Semester 2 2023

Final Project Report

	Name, Surname	Contribution (description)
1.	Shekinah Robins	Date pre-processing, data exploration, presentation, report introduction, report pre-processing, report exploration analysis
2.	Batheendra Nanayakkara	Data exploration, presentation, report introduction, report pre-processing, report exploration analysis.
3.	Jaehong Kang	Data exploration, presentation, report exploration analysis.
4.	Salihah Nisha	Data exploration, presentation, report, data modelling, report exploration analysis.
5.	Sanugi Fernando	Data exploration, presentation, report, data modelling, report conclusion, report modelling, report exploration analysis.
6.	Hieu Nguyen	Data exploration, presentation, report exploration analysis.

Part 1: Description of the project

Rossmann is a well-known pharmacy and retail chain that operate over 3000 drug stores in 7 European countries. They provide a wide range of products including cosmetics, personal care, health and wellness items, and home goods. In this project, the Rossmann Store dataset was used from a Kaggle competition, consisting of three files: the training dataset, the testing dataset, and the store dataset. However, for the purpose of our investigation we only explored the training and store dataset. These datasets contain information and statistics for 1115 stores across Germany. For the 1115 stores, a variety of both numerical and categorical data were provided, which gave information on competition, promotions, sales, customers, store types, holidays, and store assortments.

The main objective of our project is to be able to forecast sales using different models and identify the factors that impact the profitability and sales of Rossmann stores. There are many features that can impact sales, including different holidays and events, seasonal variations, store types and product assortment, competition stores and promotions, all of which are vital to understanding the fluctuations in sales and profitability.

To conduct our exploratory analysis, we investigated the key patterns, trends, and cycles in terms of sales, the factors effecting the volume of customers, the effect competitors have on profitability, the store types that are most successful, and the effect of promotions, assortment types and holidays on sales. We then utilised this knowledge to aid in our data modelling and forecasting.

Part 2: Details of pre-processing and manipulation of data in Python:

The first part of the investigation involved the cleaning and pre-processing of the store and training datasets, that involved transforming the data into a more accurate and usable format that can be utilised for a comprehensive exploratory analysis and modelling.

In the state holiday variable, two zeros were found, one represented as a string and the other as an integer. To ensure consistent data types for all unique values, the first step taken was to address these duplicates by converting the integer zero to a string type.

Next, it was important to deal with any missing values in the datasets. It was found that the Rossman store dataset had missing values in six different variables regarding competition distance, competition opening month, competition opening year, promotion two since year, promotion two since week and promotion two interval. By filtering the missing values in promotion two since year, promotion two since week and promotion two interval to only consider instances where the promotion two variable was equal to zero, it was found that the missing values in these columns were present due to there being no promotion two occurring in these stores. As such, the missing values in these columns were replaced with zero. To deal with the missing values in the columns regarding competition, the missing values were inputted with the mode, as we found that the median and mean is not suitable for imputation considering two of the three competition variables have month or year values which are discrete categories.

Following this, the rows for closed stores and rows with zero sales were removed from the training data set. This was done so that the data would only include relevant data points that would be used for analysis,

as including data from when the store is closed or not making any sales can introduce noise and make it more difficult to extract meaningful insights.

Subsequently, the date column in the training data set was converted to a datetime object and set as the index of the data frame. This was done so that the data can easily be manipulated and filtered for specific time periods. Following this, the year, day of the week name, and month were extracted from the datetime index and added to new variables for easier filtering, segmentation, and visualisation of the dataset.

Furthermore, we created a python function which maps months to seasons, and then added this information to a new variable season which allows us to analyse the data based on seasons. Another new variable called average purchase was also implemented into the data frame, which was calculated by dividing sales by customers and provides insights into the purchasing behaviour of customers.

Finally, the training dataset was filtered to contain only the top twenty percent of stores in terms of sales by grouping the sum of sales by stores and finding the top twenty stores with the highest sales. The training dataset was then merged with the store dataset, using the store variable as the key for the merge operation.

Part 3: Summary of exploratory data analysis and any significant conclusions

By analysing the key trends and cycles in terms of sales, several key patterns were revealed. It was evident that spring had the highest median sales followed by winter. These patterns can be linked to several events, such as the spring allergy season and the winter holiday season, which can influence consumer purchasing. Additionally, another significant pattern that was observed occurred in the difference between the median sales and median customers on different days of the week. As seen in figure 1, Mondays had the highest median sales and Saturdays had the lowest median sales followed by Sundays. However, as illustrated in Figure 2, Sundays had the largest median number of customers. This suggests that Sundays are often a busier day for business, yet customers tend to spend less money. Furthermore, the median monthly sales trend, as seen in Figure 3, shows a consistent downward trend every September, which corresponded with the start of the German school year, and a consistent upward trend every December, during the Christmas break, before declining once more in January. Finally, upon analysing the semi-annual median sales trend, as illustrated in Figure 4, a long-term growth trend is prominent from 2013 to 2015, suggesting the possibility of further company development.

From figure 5, it is evident that customers and sales are highly correlated, thus, more investigation into factors driving customers was performed.

Surprisingly, the side-by-side boxplots from figure 6 suggest that customers are not significantly influenced using promotions. This highlights the idea that the Rossman store's appeal appears to transcend the immediate need for medication, as customers frequent the store even when not personally afflicted.

For effect of competitors, the negative skewness from figure 7 and 8 implies that stores situated closer together tend to attract a higher volume of customers. This correlation indicates that the stores in closer proximity are typically located in densely populated areas. Thus, even with competition nearby, these stores maintain a steady stream of sales due to the high local demand. This contradicts the initial hypothesis that reduced competition would correspond to increased store traffic. It suggests that other variables play a crucial role.

Considering the impact of competitors on profitability, the histograms from figure 7 and 8, plotting customer footfall against sales, accentuates the significance of this relationship. Further analysis, as seen in

figures 9 and 10 reveals that stores facing new competitors between 2012 and 2014 recorded the highest sales figures.

To delve deeper, the time series graph from figure 11 specifically focusses on stores with competitors that emerged in 2014. The comparative analysis of sales before and after 2014 highlights a considerable decline post-2014, particularly noticeable during December, a trend likely influenced by the winter season.

The density of sales data further confirms the influence of competitors, revealing a noticeable decrease in the frequency of high sales figures post-competition entry. This suggests a notable impact on the overall profitability of the stores in question.

Our analysis of the impact of state holidays and school holidays on sales revealed several key findings. In As shown in figure 15, during state holidays, sales predominantly fell within the range of 5000 to 15000, with public holidays consistently generating the highest sales, followed by Easter and Christmas holidays. However, no clear trend was observed from 2013 to 2015, as sales for each holiday type fluctuated (Figure 16). Examining store types and correlated variables, we noted that "basic assortment" stores consistently reported the highest sales, "promo" had a more pronounced effect than "promo 2," and store type "B" consistently achieved the highest sales across all state holidays (Figure 17). Moreover, school holidays appeared to have no significant impact on sales throughout the period from 2013 to 2015.

The variable assortment type was also investigated for the explanatory analysis. Every store in the data set has been labelled with an “assortment type”, “a” for “basic”, “b” for “extra” and “c” for “extended”. No information was given on what the actual details of the assortment types are, but it was assumed that they represent various levels of products and merchandise. Firstly, the average sales for each assortment type were calculated. Figure 12 shows that a store labelled “b”, would on average have the highest number of sales, below that would be a store labelled “c” and lastly assortment type “a”. Next, the average purchase for each assortment type was calculated and the information was presented in a boxplot. From figure 13, stores labelled “c” have the highest average purchase, below that are stores labelled “a” and the stores with assortment type “b” have the lowest average purchase per customer. From figures 12 and 13 stores labelled “b”, have the highest average sales, but have the lowest average purchase per customer. This can be explained by the average number of customers for each assortment type. From figure 14 which presents the average number of customers for each type of store, we can conclude that assortment type “b” stores have many customers coming in that do not spend a lot of money, but due to the high number of customers, it results in high sales numbers.

Rossman dataset also consisted of promo column which had 0s and 1s as input. 0’s representing no promotion and 1’s representing stores undertaking promotions. Due to it containing categorical variable, the count function was used to plot the count of 0’s and 1’s against the average sales This showed that when promotion was going on the average sales were higher and when there was no promotion, it declined the average sales as shown in figure 21.

To further analyse the promo analysis, promo2 column was used that had data to represent continuity of the promotion. This column also had 0s and 1s as its entry, therefore, another bar graph was plotted to see if continuous periods of promotion resulted in better average sales, however this was not the case, as seen in figure 22, the longer the promotion continued the less sales it conquered.

To further investigate the correlation between promotions and sales, dummy variables were used to investigate the correlation of promotion with sales was created which gave a high correlation coefficient of 0.43, depicted by figure 23, second higher correlation after customers. Other promotion related variables were also a part of the heat map with dummy variables for other categorical variables, however, none of the other variables depicted a great correlation with sales. Since Rossman is a pharmaceutical company, it can be concluded that promo may have an impact on the sales for certain products like cosmetics, but it

doesn't necessarily affect the other important product sales like medications as it is a necessity and consumers buy it as needed not as it is put to sale.

Finally, the variables store type was investigated in relation to sales and customers. Rossmann separates all the stores into types "a", "b", "c" and "d". Our purpose in analysing this variable is to acknowledge customers' behaviours and store performance of the different store types. Firstly, the total sales and customers of each store type are examined. Figure 18 shows that store type A tends to have the highest sales and customers, followed by store type D, store type C, and store type B. However, the average sales and customers of store type B are the highest, followed by store type A, store type C, and store type D. The reason for the best store performance of store type B is with only 12 stores (figure 20), but the total sales is nearly two hundred thousand.

Part 4: Summary of any undertaken modelling and any significant conclusions

The main objective of our analysis was to gain insights into the key factors affecting sales in Rossman stores, as well as to come up with a model that can forecasts the sales for the next 4 weeks. From the 5 research questions in our exploratory analysis, we were able to come up with many valuable insights that assisted in the creation of our model.

In our data analysis and modelling for Rossman stores, our primary objective was to predict sales for the next four weeks. To achieve this goal, we conducted an in-depth exploratory data analysis, which allowed us to identify and select the most relevant variables for our modelling process. The variables we deemed significant for affecting sales were Competition Distance, Promo, and Customers.

We began our modelling journey with multivariate linear regression, which is a fundamental approach for predicting numerical values. However, our initial attempts with linear regression yielded an R-squared (R2) score of 0.66, indicating that our model explained only a moderate portion of the variance in sales. Moreover, we explored linear regression using other variables, which resulted in even lower R2 scores. We also experimented with decision trees, but these models proved to be less effective in capturing the complex relationships within the data. After multiple testing and experimentation, we arrived at the decision to employ the RandomForestRegressor algorithm with the features Competition Distance, Promo, and Customers. This choice led to our best-performing model, boasting an impressive R2 score of 0.90. This high R2 score is a strong indication that our model successfully explains a substantial proportion of the variance in sales.

During the modelling, we executed a train-test split of our data, partitioning it into two subsets. The first, X1, contained the predictor variables, while the second, y1, encompassed our target variable, 'Sales.' This division enabled supervised learning, allowing us to use the predictor variables to forecast sales. Our next step involved selecting the RandomForestRegressor algorithm, known for its proficiency in handling regression tasks. We instantiated the model with 100 ensemble members, providing a robust foundation for capturing intricate data relationships. Thereafter, we applied the trained model to make predictions on the testing data, and these forecasts were stored in the 'y1_pred' variable. To evaluate our model's performance, we calculated the R-squared (R2) score, in which we received a score of 0.900 (Figure 21).

During our forecasting process for the upcoming four weeks, we encountered an intriguing and insightful observation regarding sales for the third week. In our predictions, we held the Competition Distance constant, introduced a promotional offer only in the third week, and anticipated a decrease in the number of customers. Surprisingly, our model predicted the lowest sales value for the third week, even in the presence of a promotion. We attribute this phenomenon to historical data, suggesting that such

promotions may coincide with special occasions, during which our competitors might be offering more compelling deals than Rossman stores.

In conclusion, our modelling efforts, guided by the high R2 score of 0.90, have equipped us with a robust tool for forecasting sales over the next four weeks. This model, built upon key variables, empowers us to make informed business decisions and strategic planning. The insight gained from the third-week prediction underscores the importance of considering external factors and market dynamics when interpreting our forecasts and optimizing promotional strategies.

Part 5: Conclusions in the relation to the original problem

Our primary aim in conducting this analysis was to uncover the fundamental factors that significantly impact sales within Rossman stores and, consequently, to develop an effective forecasting model predicting sales for the next four weeks. Throughout our exploratory analysis, we addressed five distinct research questions, which, in turn, provided us with a wealth of valuable insights that have substantially contributed to the creation of our predictive model.

One of the fundamental insights we gained was the effect of seasonal variations in sales. Our data revealed a distinct pattern in sales performance across different seasons, with spring and winter exhibiting the highest median sales. This observation was instrumental in understanding the influence of external factors, such as the spring allergy season and the winter holiday season, on consumer purchasing behaviour.

Furthermore, we analysed the dynamics between sales and customer behaviour on various days of the week. This revealed a noteworthy contrast – Mondays consistently recorded the highest median sales, while Saturdays marked the lowest, even though Sundays attracted the most significant median number of customers. This intriguing observation in customer behaviour on weekends highlighted the need for effective strategies to maximize sales on these days.

In our journey to develop an effective forecasting model, we examined the impact of promotions and the presence of competitors on sales. The revelation that promotions had limited influence on customer visits suggested a broader appeal of Rossman stores beyond mere immediate need, reflecting positively on the brand's reputation and customer loyalty. Equally fascinating was the influence of competitors, which contradicted initial assumptions. Our analysis demonstrated that stores situated in close proximity to competitors tended to attract more customers, challenging the traditional notion that reduced competition led to higher store traffic. This relationship between competition and customer volume emphasizes the critical role of additional variables in shaping customer behaviour.

To further enhance our forecasting model, we delved into assortment types and their impact on sales and customer spending. Stores labelled "b" exhibited the highest average sales, while stores designated as "c" recorded the highest average purchase per customer. Notably, "b" stores drew a significant number of customers who, despite not spending as much individually, collectively drove high sales figures.

During our forecasting modelling, we uncovered a valuable insight: despite introducing a promotion in the third week, our model predicted the lowest sales, likely due to historical data indicating competitors offering better deals during such promotions. In conclusion, our model, guided by the high R2 score, equips us for informed decision-making.

In conclusion, our exploratory analysis illuminated a multitude of key insights that not only enhanced our understanding of the critical factors influencing sales in Rossman stores but also guided the development of a beneficial forecasting model. These insights encompass seasonal patterns, customer behaviour on different days, the influence of promotions and competitors, and the role of assortment types. Ultimately, our findings led us to the conclusion that the most influential factors on sales were customers, promotions, and competition distance. These were the primary variables incorporated into our modelling approach. By

leveraging this knowledge, we are better equipped to make data-informed decisions and develop strategies that align with the dynamics of the retail landscape.

For further data science approaches, we could consider using individual stores as the focus of our data modelling. This method would improve the accuracy of our sales estimates by customising predictive models to the unique attributes and data of each store.

Figures

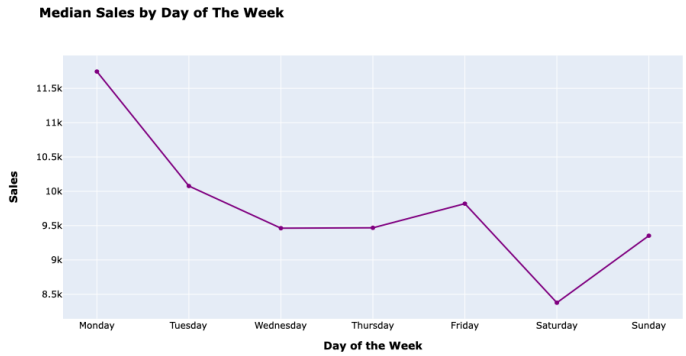


Figure 1: Displays a line plot of median sales by day of the week.

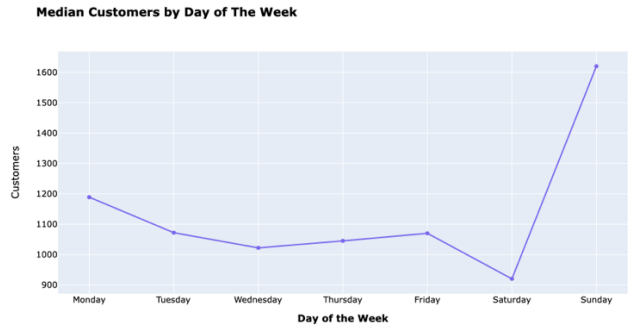


Figure 2: Displays a line plot of median customers by day of the week.

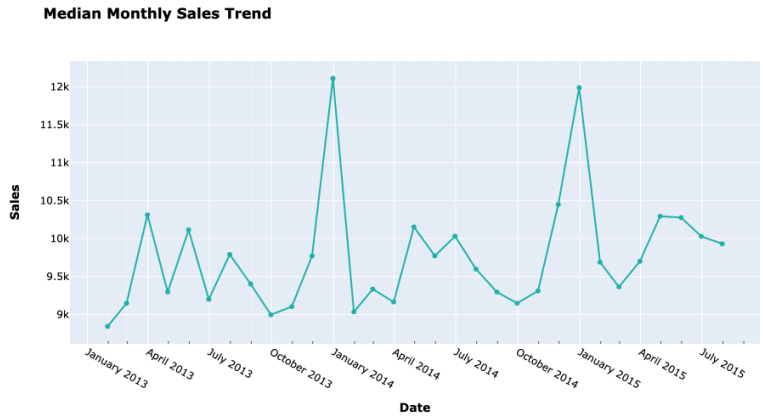


Figure 3: Displays a line plot of the median monthly sales trend.

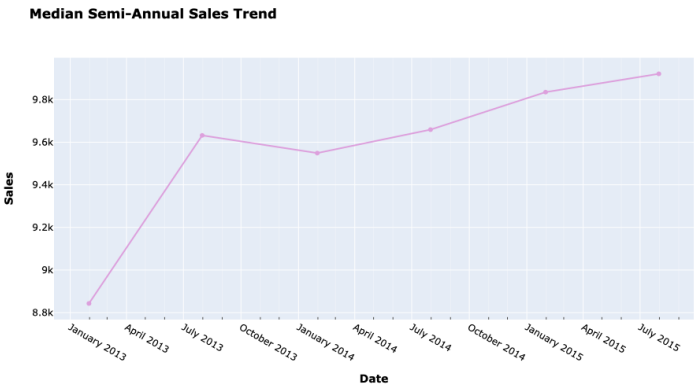


Figure 4: Displays a line plot of the median semi-annual sales trend.

Figure 5: Scatterplot of sales and customers

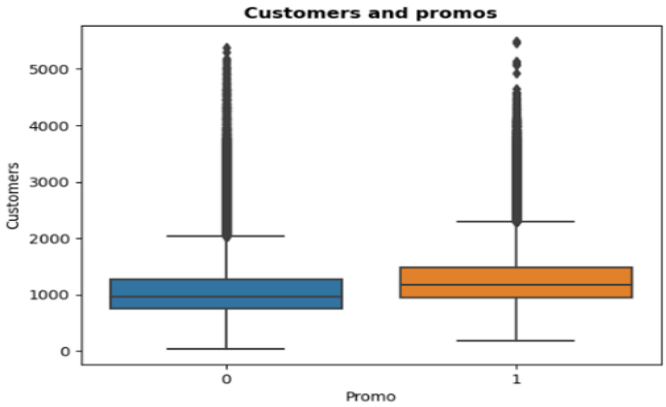


Figure 6: Boxplot of sales figures of stores that participated in promotions and those that didn't.

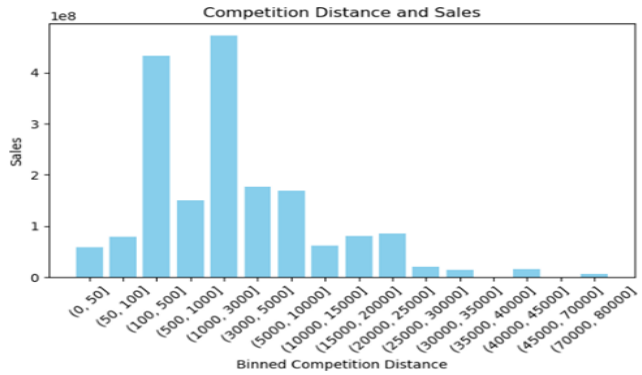


Figure 7: Binned competition proximity and sales

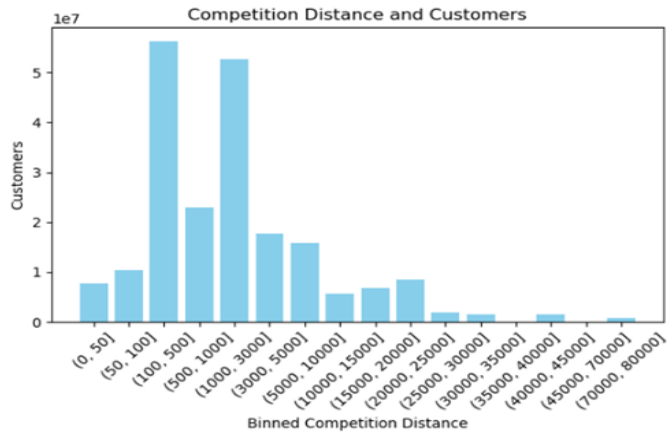


Figure 8: Binned competition proximity and customers

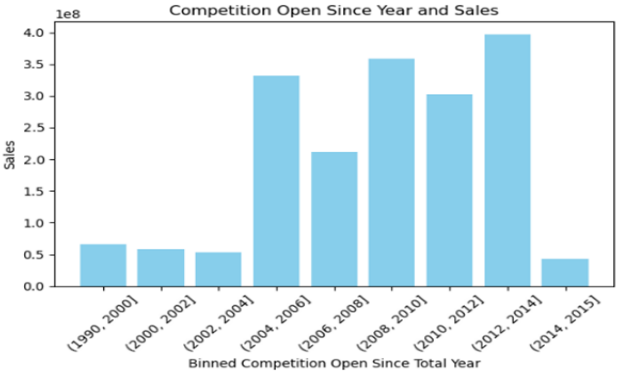


Figure 9: Binned competition opening year and sales.

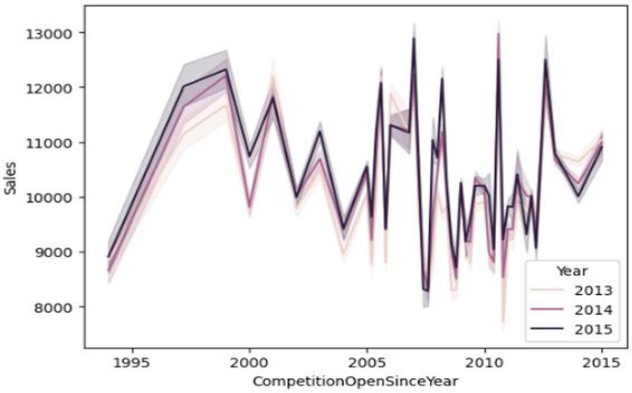


Figure 10: Line plot of when competition opened to their sales figures.

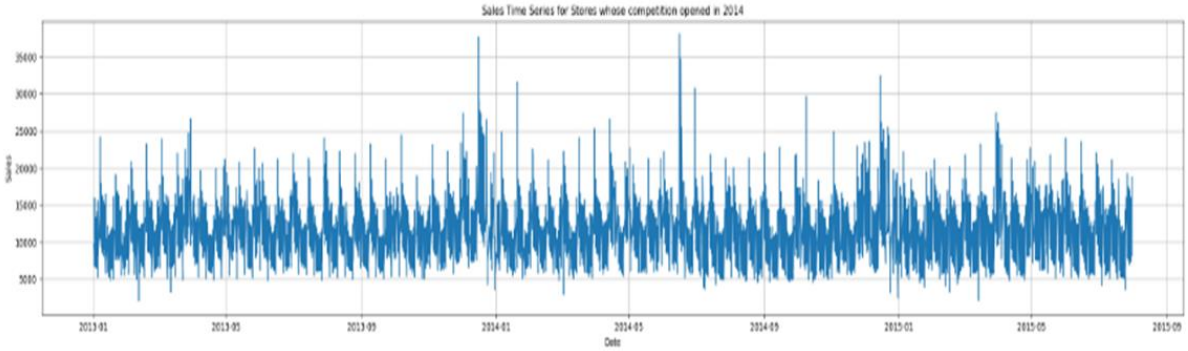


Figure 11: Displays a line plot of sales for stores who had a competition open in 2014.

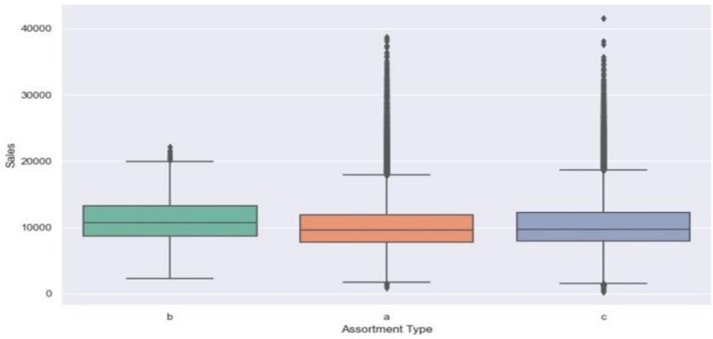


Figure 12: Boxplot of the average sales for each assortment type.

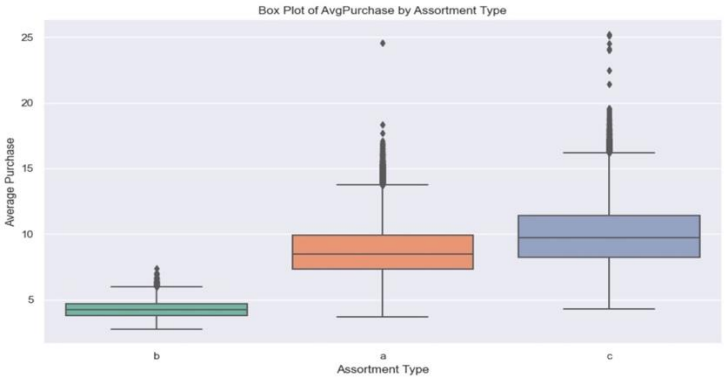


Figure 13: Boxplot of the average purchase for each assortment type.

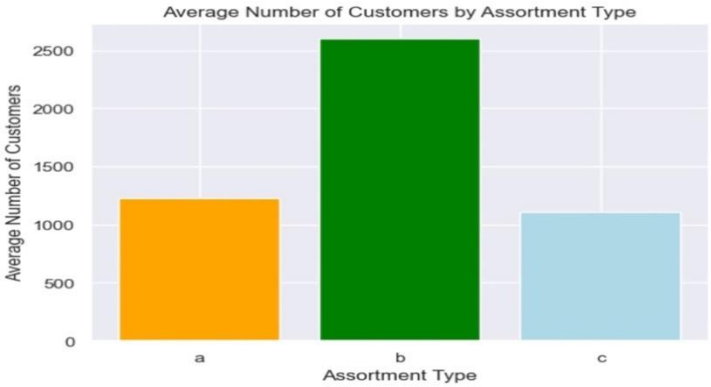


Figure 14: Bar chart of the average number of customers for each assortment type.

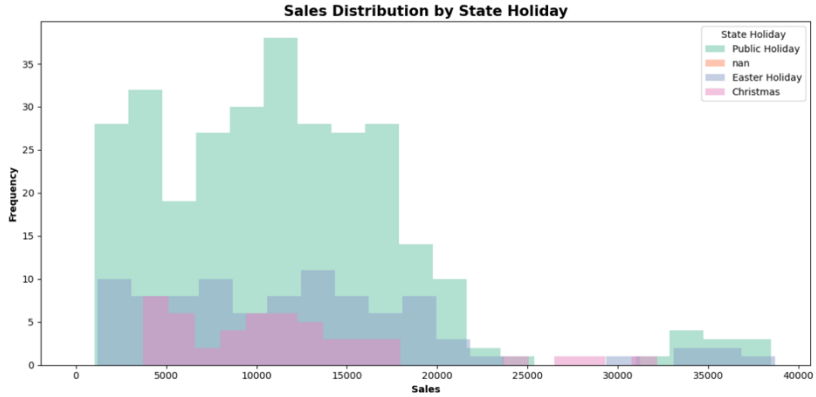


Figure 15: Histogram of Sales distribution by State Holiday

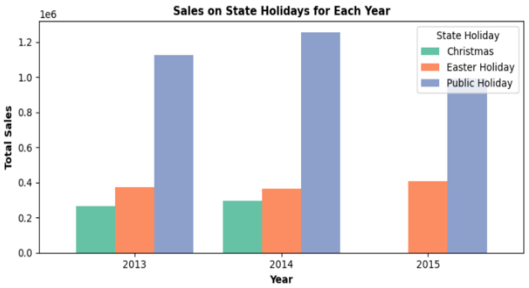


Figure 16: Bar Chart of Total Sales for each year by State Holiday

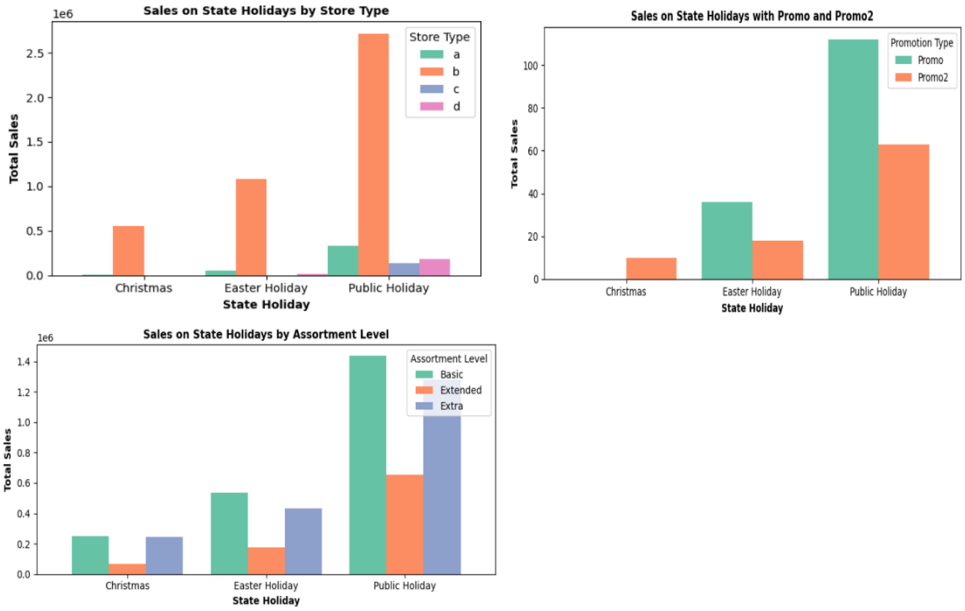


Figure 17: Bar Charts showcasing the effect of State holidays on Sales based on various other variables

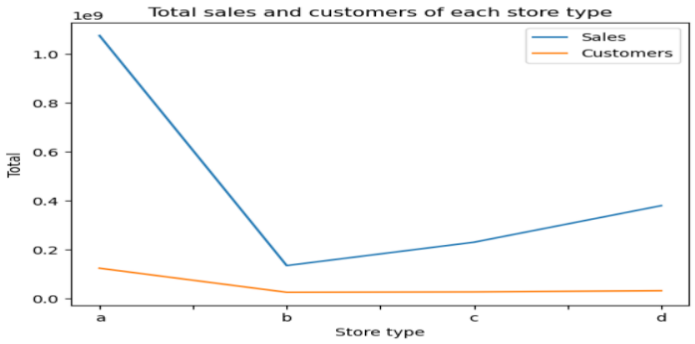


Figure 18: Line chart of total sales and customers for each store type

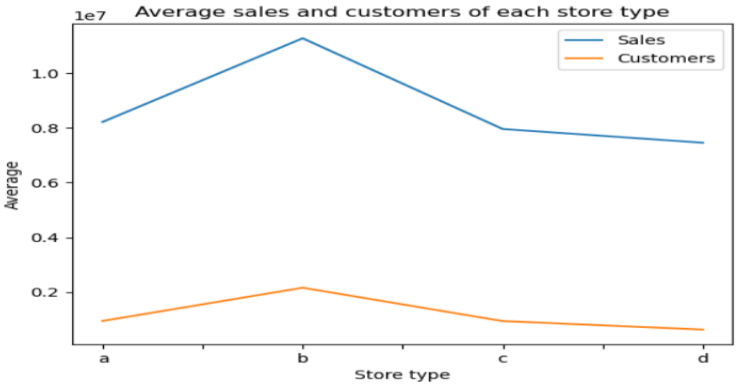
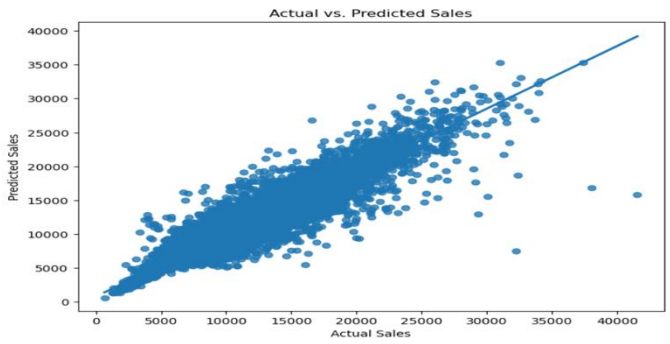


Figure 19: Line chart of average sales and customers for each store type

	Store
StoreType	
a	131
b	12
c	29
d	51

Figure 20: Total number of stores for each store type



	Date	PredictedSales
0	2015-08-01	4716.676667
1	2015-08-08	4461.610000
2	2015-08-15	3368.360000
3	2015-08-22	4425.910000

R-squared (R2) Score: 0.900

Figure 24: Shows a scatterplot of predicted sales against actual sales, a table of the predicted sales for 4 weeks, and an R-squared score of 0.900.

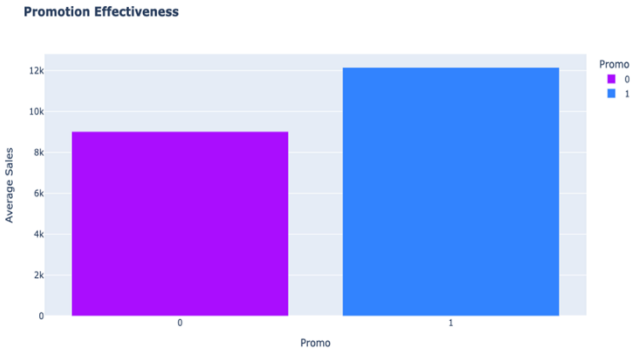


Figure 21: Bar plot for promotion against average sales

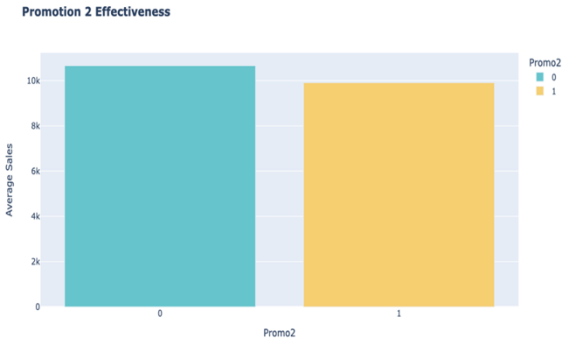


Figure 22: Bar plot for continuity of promotion against sales

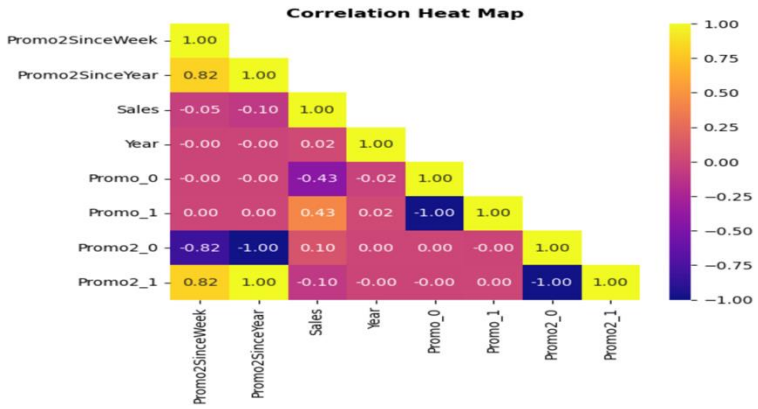


Figure 23: Correlation Heat Map to show correlation between sales and promotion