QBUS6600 Group Project Report

University of Sydney

QBUS6600: Data Analytics for Business Capstone

Group：113

1. Executive briefing

In order to make Big W more competitive under the harsh environment of the retail industry, this report provides a large number of high-quality practical suggestions based on the modeling and analysis results of six meaningful datasets. The main goal of improving market competitiveness is to increase sales values. This article provides developmental suggestions from four directions based on the results of data analysis. They are geographical, target customer, competitors and promotion and media investment respectively.

The first suggestion direction is proposed based on geographical location:
From a geographical perspective, the current sales level of NSW is relatively high, and remote areas such as Western Australia have lower sales. Which means, Western Australia is a potential market with high purchasing power for customers. Therefore, our suggestion is to open a new store in Western Australia. And opening it in Bunbury city is a good choice. This city is the second largest in Western Australia, with a population up to 90000 which ensures a sufficient number of potential customers. In addition, Mandurah is also a good choice for opening a new store, as its market competitiveness is relatively low with only one large store currently. It is also an important transportation hub which is located between Perth and South Australia. In addition, the variable of "same_area_shopping" is the most important feature, which means that consumers prefer to shop near their place of residence. So, it is recommended to open new store addresses in densely populated city centers in these two cities.

The second suggestion direction is proposed based on target customers:
The data analysis results in the customer group section show that BIG W is more attractive for budget-young families that are pursuing high cost-effectiveness, it is important because it’s aligned with BIG W's product positioning. The variables' lifestage_segment_young families' and 'price_segment_premium' also have high feature importance, so young families prefer to shop at BIG W. As Budget young families are our main customer group, it is recommended to establish a young family product area at the entrance of the store to bring convenience to their shopping. At the same time, regular weekly discounts can be given to the target products of Budget young families to increase customer traffic.

The third suggestion direction is proposed based on competitors:
Data analysis results show that whether opening a store in the same mall as a competitor affects sales levels. Through the data analysis results, we concluded that opening a new store less than 1km away from Kmart and about 3 to 5 kilometers away from Target.

The last suggestion direction is proposed based on promotion and media investment:
The Random Forest model demonstrates the importance of investing in media. It will increase the total sales value of Big W. And the data results show that the ACT state has the highest return on media investment. Therefore, it is recommended to invest more in media in the ACT state, and the advertising time is recommended to be concentrated in May and November.

2. Business context and problem formulation

Big W is a discount department store that is one of the giants in the Australian retail industry. It belongs to the Woolworths Group and was established in New South Wales in 1964. With years of development, it now has nearly 180 stores in Australia. This company not only provides Everyday Rewards (EDR) to maintain customer loyalty, but also spends a huge amount of advertising on promoting brands, covering various channels such as online video, television, search engine marketing, outdoor broadcasting, and social media. However, with the continuous development of the retail industry, market competition has become increasingly fierce (Tracy et al., 2017). Kmart and Target continue to exist as strong competitors for Big W, and e-commerce in Australia has been expanding in recent years. Which leads to some online companies such as eBay and Amazon are also gradually posing a threat to Big W. In such a harsh market environment, Big W needs to conduct more market analysis and continuously provide appropriate management opinions to maintain good competitiveness.

This report uses four datasets provided by Woolworths Group that relate to customer sales value in loyalty programs. These four datasets contain data from 2021 to 2023. The names of them are "sales by customer location train", "sales by customer location test", "sales by store location" and "media Investment respectively". Apart from them, this report also used two more external data sets which were downloaded from the Australian Government Data Bureau (ABS). "They are Total personal income (weekly) by state and territory(a), 2021 Census"' and "Usual resident count by state and territory(a), 2016 and 2021 Census". Exploratory Data Analysis (EDA) and feature engineering is conducted on these datasets to find the hidden attributes, features, patterns, and statistical information in the dataset. Subsequently, based on the results of EDA, this report established many models, they are linear regression, lasso regression, ridge regression, decision tree, bagging, bagging, and XGboosting. By fitting different models to find out which factors have the greatest impact on Big W's sales value, and then based it to provide suggestions for Big W's future development and help it survive better in the harsh competitive market.

3. Data processing, EDA, and feature engineering

3.1 Data Processing

To ensure the data integrity and quality in the following exploratory data analysis (EDA) and feature engineering, a few steps are processed before the EDA. Both the test datasets and train datasets are processed in this part.

Missing Value Processing

The missing values of the store dataset is shown in the table below

Variables	Missing amount	Missing percentage
co_location_flag	11	2.588%
distance_to_kmart	1	0.235%
distance_to_target	1	0.235%

The co_location flag is the indication of whether the BIG W and Woolworths are in the same shopping store. The other two variables represent the distance range to the kmart or target. As the dataset is collected by Woolworths group, which they belong to. The missing of this value is assumed to be the same shopping center they are located in. Since the longitude and latitude are given and the missing data amount is small, the real location is checked through google map and Big W website to verify the assumption and finally filled with False. While the missing od distance to kmart and target may be because that store is a delivery only store, which is similar to an inventory store. After checking the real location, it is filled with >5km. This is the recorded distance of a full function store which provides digital and offline service near this store.

As for the customer location train dataset, the missing values are shown below.

Variables	Missing amount	Missing percentage
price_lifestage_segment	13565	0.742%
Customer_postcode	7	0.0004%

The price_lifestage_segment may miss due to the sensitivity of the information for customers. It is filled by the mode of different customer states as it is assumed that the customer segment's shopping behavior is probably similar in the same area (Appendix A). The customer postcode is missing as the related state is other. This indicates that the customer may be in other territories. Hence, 0 is filled to enable it to be separated from others.
The test set missing values is the same as the train set, therefore, it is filled by the same method. And the media dataset doesn't have missing values.

Data Type Processing

The data type of the date variables week_ending (media investment dataset) and financial week_end_date (customer location train dataset) is set to be object. In order to enhance the time related analyzability of it, it is transformed to the datetime type.

External Dataset Engaging

Two external datasets, 2021 total weekly personal income dataset by state and 2021 state population dataset, are included. These datasets will help to figure out how the income and population will affect the total sale value. And it will also give instructions on state related recommendations based on these basic information.

Other processing

To obtain a better understanding of customer segment patterns, the price_lifestage_segment is split to price_segment and lifestage _segment. To have a more convenient analysis on the total sale and media dataset. The customer train set is merged with the media dataset by grouping by the week_end_date. The media dataset is also combined with the store data based on the state and date.

3.2 EDA

Geographic Analysis

The Geographic can be analyzed through sales channel and state perspective.
Sales Channel (Appendix B): BIG W currently have three channels, which can be divided into two types, store shopping and online shopping. Online shopping includes click and collect and delivery. According to Appendix B, although the store and click and collect service account for the same percentage of the channels, the store makes up the largest sale amount which is 98.47%. And 97.9% of customers choose the offline shopping in BIG W. This indicates that Big W customers still prefer the traditional store shopping. And based on the box plot, people are also probably purchasing more during the store shopping compared with the online shopping, as they don't have clam time and are easier to be attracted by the discount.
State (Appendix C): New South Wales (NSW) has the most stores as well as the largest sale range compared with other stores, while Northern Territory (NT) has the lowest store number and the total sale according to the box plot and bar chart. Noteworthily, although Western Australia (WA) only ranked fifth in the state store number and fourth in the total sale value, WA stores have the highest average sale as well as the average personal sale. This indicates that the potential demand and purchasing power of customers are high in WA.

Customer Segment Analysis

The customer segment is divided by the price stage and life stage. Based on the first bar chart in Appendix D, the main customers of BIG W are from budget-young families, who

have limited budgets and seek high cost-efficiency. While the premium older families, who emphasize more on the life quality, are least likely to purchase in BIG W.
After dividing the whole segment to price and life stage separately to obtain clearer insights, based on the remaining bar chart in Appendix D, with the increased purchasing power of the customer group, the demand for Big W products will decrease. Besides, for the life stage, the young families, older singles/couples, and retries have more willingness to shop in BIG W, as their possible limited budget leads them to more cost-effective products.

Congregation & Competitor Analysis

Congregation: The clustering effect (when BIG W and Woolworths are in the same shopping center) may impact the BIG W total sale. The range of the total sale will increase (Appendix E), as the customer in the BIG W with True co_location_flag is probably just stopping by for a quick stroll instead of shopping specifically in it.
Competitor: The total sale value will be impacted by the distance between BIG W and its competitors. According to the top graph in Appendix F, BIG W suffers an unstable sale when Kmart and BIG W are in the same shopping store as it has the largest range. Either <1 km or > 5km will be a good distance choice for BIG W. However, <1km could be better as its sale is more stable and have higher start and end point.
While the second graph in Appendix F indicates that 3-5km with a target could be a better choice, offering relatively high mean value and narrow range.

Promotion Analysis

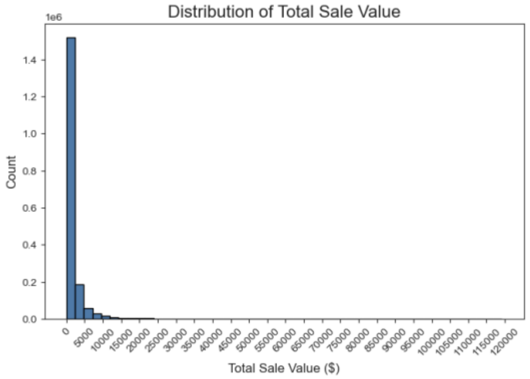
The sale value will increase with the increase of media investment and promotion amount. This indicates the sale can be enhanced by the discount and the special advertising.
As for the promotional value, according to the trend shown in the upper graph in Appendix G. The peak exists in mid-year and end of the year which may be due to the Christmas and midyear sale. The great discounts and the holiday limited edition product drew more people to purchase more cost-effectively especially for BIG W’s target limited budget customer.
While the media peaks appear around May and November. It’s a bit earlier than the sale increase and promotion increase. This is because a buffer time should occur between the media release and actual promotion to enable more people to be informed and increase purchasing. Based on the bar chart, Although NSW spends most on media, the highest return is contributed by ACT, which indicates the sale value in ACT is more sensitive to the media investment.

3.3 Feature Engineering

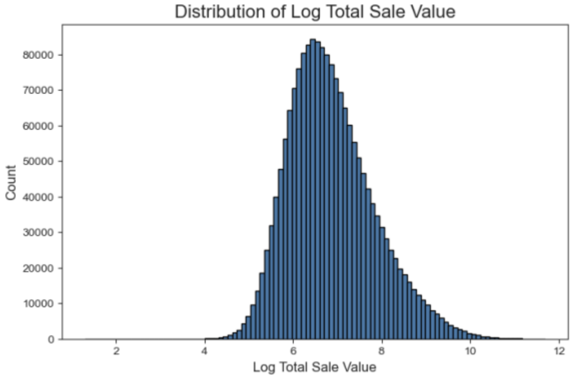
Feature Engineering is done on the merged datasets to select the appropriate features used for the model building. The log transform is used to ensure the normality of the data, and the features are selected based on both the EDA and domain knowledge. And the dummy encoding is used on the categorical valuables at the end.

Data Transformation

After exploring the data, the distribution of target value is found to be right skewed (Figure 2.3.1). To reduce the skewness and build a better model, Log transformation is implicated on the target value to increase its normality. According to Figure 2.3.2, The right skewness of the target value is reduced, it distributes more normalization now.



(Figure 2.3.1)



(Figure 2.3.2)

Feature Selections

To avoid the data leakage, the variables of customer_count, transaction_count, total_sale_value_ex_gst and total promotional sale value is deleted from both tarin dataset and test dataset. These data are hard to explore before the final sale data is received. Therefore, the variable promotional_percentage created based on the promotional sale should also be dropped.

Besides, several other features are also deleted as they have less relevance to explore the total sale value. For the location related data, the store_id is deleted as it is just the identity variable used to combine the datasets. The store_latitude and longitude are also removed since they are duplicated with the store state and postcode and are less useful to predict the target value. While the customer postcode and store postcode are also deleted. This postcode represents the same but more detailed location information compared with the state variables. However, it's hard to explore the specific relation between the detailed area and the target value. The detailed postcode meaning, which is not found in the dataset used, should be included. The high cardinality problem will also occur if the dummies are obtained through the postcode, as although they are numbers, they are still regarded as the categorical value with a large number of categories. In addition, there may also occur a data mismatch between the train data and test data. Besides, as the external income and population dataset engaged is built based on the state instead of the specific postcode area it will be more meaningful to analyze through the state instead of detailed postcode. Therefore, the state variables are chosen to be the feature instead of the postcode. However, to get a sense of whether customers prefer to purchase in the same area shop, the same_area_shopping variable is introduced. It will be True if the customer postcode is the same as the store postcode.

As for the customer segment, since the category amount is large and also to get more detailed customer segment insights from both the price and life stage side, the price_lifestage_segment variable is deleted. The separate price_segment and lifestage_segment are used instead.

As for the time data, it is splitted to year and month to make it more meaningful, which can help to explore the specific relationship among them. While the day is deleted, since it is meaningless without the combination of months. The high sale on a specific day may not repeat every month. Instead, it relates more to the month based on the EDA. For example, the Christmas month always has a higher total salve value than other months.

Finally, based on the EDA and the above analysis, these features in the following table are all have important impact on the final target value total sale value and are selected as the model used features

Store related features	Store_state, co_location_flag, distance_to_kmart, distance_to_target
Customer related features	Customer_state,Price_segemnt,Lifestage_segment,same_area_shopping, Sales_channel
Promotion related features	Media_amount_spend
Other related features	Median income, population, year, month

4. Model Building

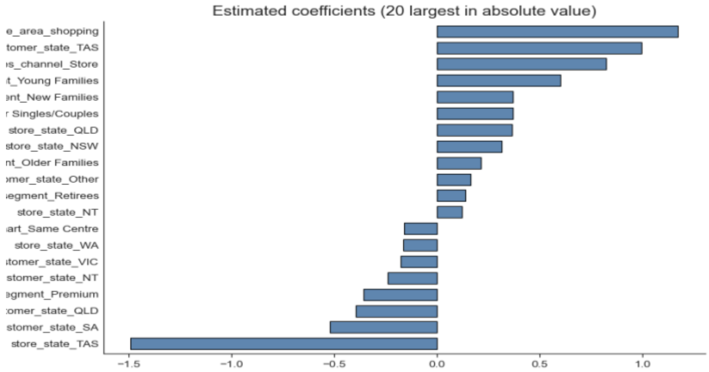
4.1 Model Preparation

To validate the models and select the best model, the original training dataset is divided into a train set and validation set with a train size of 80%. In the following models, root mean square error (RMSE) is used as the evaluation metric. However, since the target value is log transformed, Root mean squared logarithmic error (RMSLE) is also introduced to help evaluate the models.

4.2 Linear Regression

The first model is the linear regression model, it is a simple statistical model that is used to establish and analyze the relationship between two or more variables. Its main goal is to describe the relationship between the independent variables and the dependent variable by fitting a linear model. The coefficients between them represent the relationship between the independent variable and the dependent variable, with a positive value indicating a positive correlation, where the independent variable increases, the dependent variable also increases accordingly. A negative value represents a negative correlation, where the independent variable increases, the dependent variable decreases on the contrary (Heuvelmans et al., 2006).

This study obtained the correlation between various variables and sales values by fitting the Linear Regression of the training set. The following figure shows the estimated coefficients of the 20 variables with the largest correlation. If the negative correlation coefficient is large, it indicates that this variable also has a significant impact on sales values. Therefore, the analysis should consider the absolute value of the estimated coefficients. Through the figure, it can be intuitively observed that ‘same_area_shopping’, 'sales_channel_Store', ‘lifstage_segment’ and 'price_segment_Premium" are variables which have a significant impact on sales value and managers should focus on these variables when making decisions.



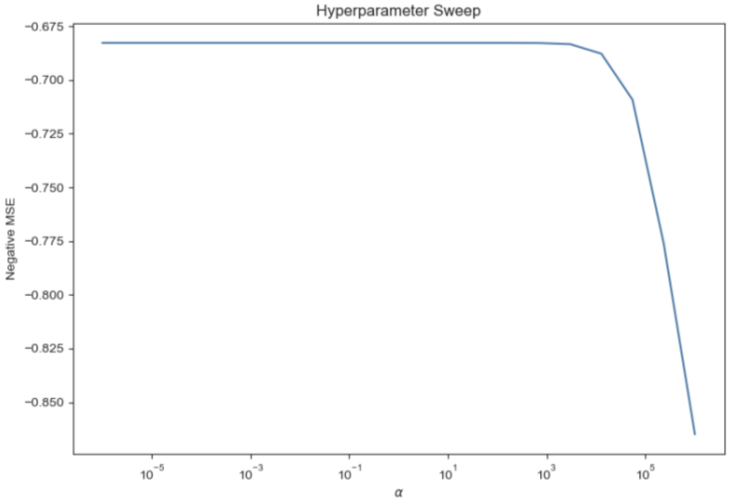
Linear regression has many advantages, such as being a very simple model. It is quite easy to model, the model is very stable and easy to understand. The fitting process is very efficient, especially when the dataset is very large. Through the analysis of coefficients, it can be found that their interpretability is also very strong. However, it still has some limitations, such as the need for a linear relationship between the independent and dependent variables, and the possibility of underfitting, which makes the model unable to capture complex relationships in the data. Besides, the impact of extreme values on the model may be very significant (Jiao et al., 2020).

In the fitting of this study, the RMSE of linear regression is 2503.524 and RMSLE is 0.105 , which is quite large among all models. Which means this is not a suitable model for prediction, but the important variables obtained through the analysis of estimated coefficients still have some reference value.

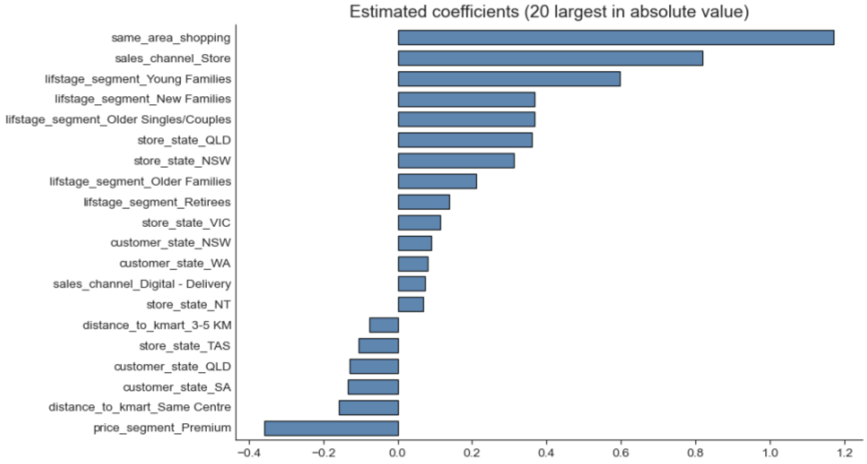
	RMSE	**RMSLE**
Linear Regression	2503.524	0.105
Lasso Regression	2503.519	0.105
Ridge Regression	2503.555	0.105

4.3 Ridge Regression:

The second model is the ridge regression model, it is a variant of linear regression used to solve multicollinearity problems and regularize models (Saleh et al., 2019). It penalizes the coefficients of the model by adding L2 regularization terms to reduce overfitting risks and improve the generalization performance. The specific punishment method is the sum of Ordinarily Least Squares (OLS) and the squared coefficients of the model (Qasim et al. 2021). The sum of the squared coefficients is also multiplied by a regularized strength parameter α. The following figure shows the relationship between α and negative MSE. This study used cross validation (cv=5) to obtain the optimal alpha, because a smaller MSE is better, so for negative MSE a bigger value is preferable. Therefore, the selected best alpha is 8.859.



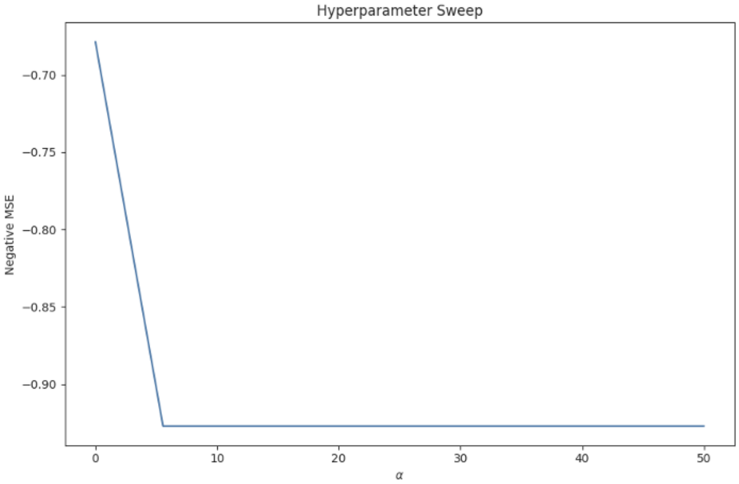
The following figure is the top 20 variables with the largest absolute value of estimated coefficients. It can be seen that the variables that have a significant impact on sales value are the same as simple linear regression. However, the MSE is not the same, it has an RMSE of 2503.555 and RMSLE of 0.105, which is the largest among all models. The largest RMSE of it indicates that this model is not suitable for prediction.



The limitation of ridge regression is that it lacks the ability to perform feature selection and is not suitable for processing a large number of features (Sharma et al., 2023). However, this study has nearly 40 variables, and the modeling method that preserves all features may be the reason for the large RMSE and RMSLE.

4.4 Lasso Regression

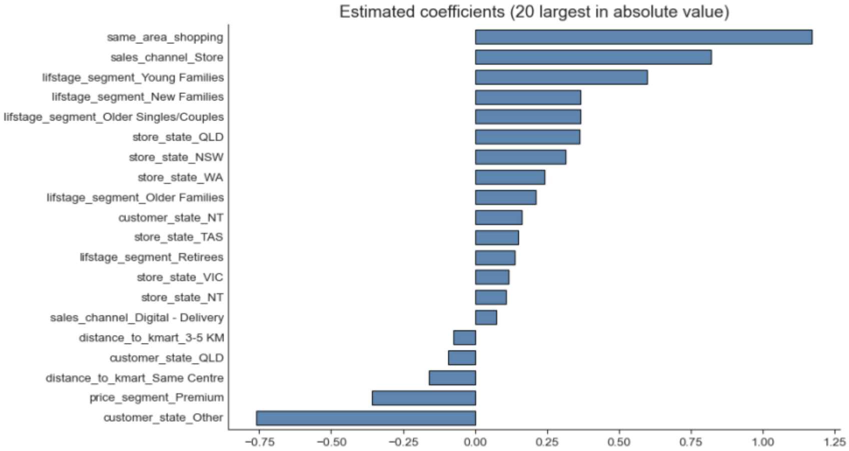
The third model is the lasso regression model, similar to the ridge regression model, Lasso regression punishes the coefficients of the model by adding L1 regularization terms. Specifically, it punishes the sum of OLS and the absolute values of the coefficients (Ranstam & Cook, 2018). Same as lasso, it also used cross validation (cv=5) to obtain the optimal alpha in this study, the following figure shows the relationship of α and negative MSE for lasso regression model and the best alpha obtained is 0.



The RMSE obtained in lasso regression is 2503.519 and RMSLE is 0.105, although it is not a quite small RMSE, it is the smallest one among these three linear regression models.

The goal of L1 and L2 regularization is to minimize the sum of OLS and regularization terms. L2 regularization only reduces their values without reducing the model parameters to zero, which helps to prevent overfitting and improve the generalization performance of the model (Pereira et al., 2016). However, L1 regularization tends to reduce certain model parameters to zero to achieve feature selection, which means it is very effective in handling multicollinearity and feature selection (Laufer et al., 2023). In the lasso regression of this study, the coefficients of variables media amount spent, and median income population was changed to 0. This means that these two features have no help in predicting the sales value. By removing them, the minimum RMSE and RMSLE in linear regression was obtained, which indicates that removing irrelevant variables is crucial for modeling in this study.

Like the previous two models, there is no difference in the important variables obtained through the estimated coefficients figure of the lasso regression.



The limitation of lasso regression is that when there is a high degree of correlation between features, the lasso may select one of the features and reduce the coefficients of other related features to zero. This may result in the model losing some information (Freijeiro-González et al., 2022).

4.5 Decision Tree

The fourth model is the decision tree model. The decision tree model can simulate the human decision-making process. It describes the if-then-else rule sequence of the decision-making process. It is a kind of analysis method by constructing a tree-type decision-making structure based on known various feature values, it is a commonly used supervised algorithm (Coursera, 2023) and one of the most basic tree models.

The advantage of using a decision tree model is that because the structure of the model is based on the variables in the train and test data sets, during the process of building the decision tree, all important variables will be automatically selected from the data set according to the algorithm of the model, and then they will be sorted. In addition, the learning of the decision tree model is fast, and the model highly simulates the human way of thinking for the decision-making process, which leads to the decision-making tree interpretation easily and avoids the black box in the algorithm, so through the decision tree people can quickly understand the importance of variables in the data set to consumers. The limitation of this model is that it uses independent features to try to predict precise probabilistic outcomes, which can cause the model to overfit on the training set, causing the model to fail to accurately predict outcomes on the test set. (EDUCBA, 2023)

Firstly, the model is established. By adjusting the parameters, when assuming that the number is from 1 to 30, select 1 number from every 10 numbers. The model iterates 30 times so that the test results do not have values that are too large or too small. The parameters recommended by the model are that the maximum number of features considered when performing feature segmentation cannot exceed 19, and the minimum number of leaf node samples must exceed 21.

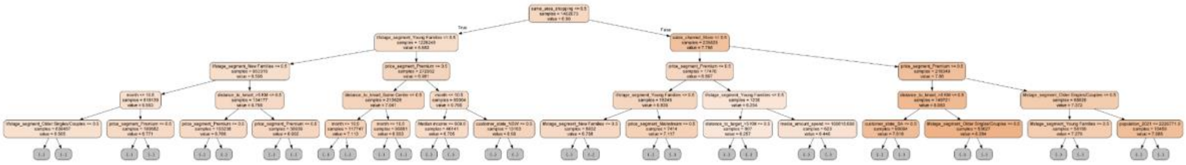
	max_features	min_samples_leaf	random_state
DecisionTreeRegressor	19	21	0

These parameters are appropriate values for the 42 variables in the test data set. By putting the validation set into the model and calculating, the final RMSE value is 2245.305 and 0.098. This value is smaller than the RMSE and RMSLE values of linear models. This is consistent with the logic that the results of the decision tree model should be better than the results of the regression model.

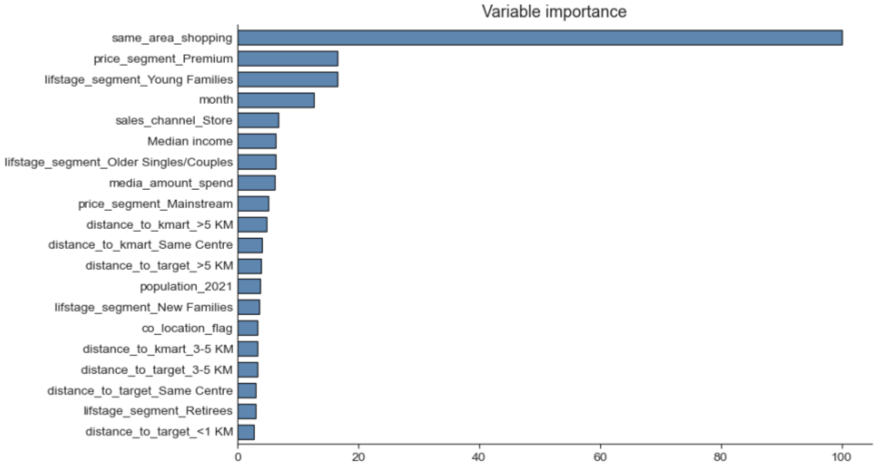
	RMSE	RMSLE
Linear Regression	2503.524	0.105
Lasso Regression	2503.519	0.105
Ridge Regression	2503.555	0.105
Decision Tree	2245.305	0.098

The decision tree model can prioritize the importance of variables. The model believes that the data 'Other' in the variable 'customer_state' has a low correlation with the entire data set and is not important.

This is the decision tree of the test data set. The maximum depth of the decision tree used is 4, and the minimum number of samples of leaf nodes is greater than 21. If the depth is 3, the minimum number of samples for leaf nodes is less than 21. It does not meet the optimal parameters of the model, that is ‘min_samples_leaf=21’.



Through the bar chart of the decision tree model, It shows that the most important data considered by the model are the variable 'same_area_shopping', the 'premium' group in the variable 'price_segment', and the 'young families' group in the variable 'life_stage'.



Therefore, more attention should be paid to this information in sales strategies.

4.6 Bagging Model

The fifth model is the bagging model. The decision tree model trains each tree on the same data set. These trees will be the same because the training trees are deterministic. To increase the predictability of the model, the data of each decision tree needs to be different, so the bagging model will test multiple decision trees and then average the results of these decision trees.

The advantage of the bagging model is that compared to the decision tree model, the bagging model makes the prediction more accurate, it helps reduce the variance of black-box estimators such as decision trees. The limitation of this model is that it is computationally expensive, and will lead to a loss of model interpretability. (CFI, n.d.)

In the process of establishing a bagging model, By adjusting the parameters, when assuming that the number is from 1 to 50, selecting 1 number from every 10 numbers. The model iterates 10 times. This model shows that the maximum depth of each decision tree should be 20. Different data should be used in the bagging model to test 41 decision trees, and then the average of the 41 decision tree results should be taken.

	max_depth	random_state	n_estimators
BaggingRegressor	20	42	41

Bringing the train data set into the bagging model, the RMSE result is 2264.917 and RMSLE is 0.099, which is greater than the RMSE and RMSLE of the decision tree model. The reason for this result might be that the model is overfitted due to parameters.

4.7 Random Forest

The sixth model is the random forest model. In the bagging model, each decision tree uses a different data set, but the decision tree has the characteristic of arranging the importance of variables, so the difference between each decision tree is not big in the bagging model. To reduce the data used by each decision tree with high correlation characteristics, using the random forest model is an option. Random forest is getting many independent trees. This model will randomly select features and samples so that each tree in the forest has both similarities and differences.

The advantage of using the random forest model is that the model is more predictive and suitable for big data analysis. When using the bagging model, overfitting will occur, and the random forest model can resist overfitting. In addition, the random forest model can be processed parallelizable, which can reduce the calculation time of the code. The limitation of this model is that it requires a lot of computing power and time because it builds a large number of decision trees to combine the model. (GreatLearning, 2023)

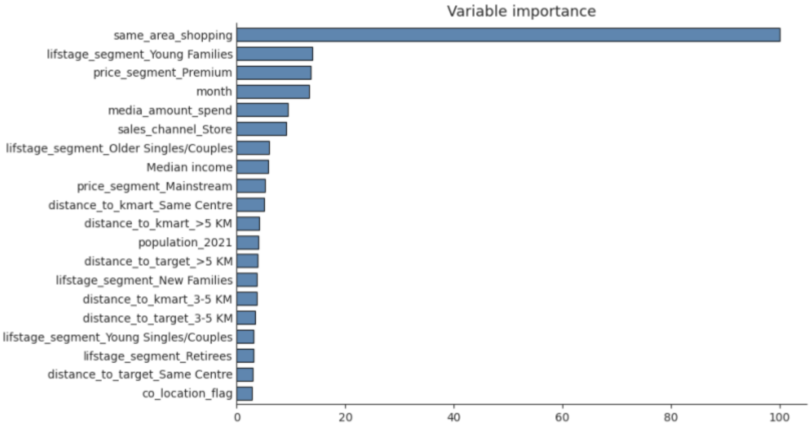
Establishing a random forest model by setting parameters. By adjusting the parameters, when assuming that the model will randomly select one data from every 10 data between 1-50, and the minimum sample number of leaf nodes will be randomly selected between 10-50 so that the result of the decision tree will not be too small or too large. The model iterates 10 times. The parameters recommended by the model are that the maximum number of features considered when performing feature segmentation cannot exceed 20, and the minimum number of leaf node samples must exceed 20, and the number of decision trees used in the model is 31.

	max_features	min_samples_leaf	n_estimators	Random status
Random Forest	20	20	31	42

Fitting the validation set into the model for prediction. The RMSE result is 2232.248 and RMSLE is 0.097, which is smaller than the decision tree model and bagging model. This result is consistent with the logic that the performance of the random forest model is better than the decision tree model and bagging model.

	RMSE	RMSLE
Decision Tree	2245.305	0.098
Bagging	2264.917	0.099
Random Forest	2232.248	0.097

Random forest models can also estimate which variables are important in the classification. Drawing a bar chart through the random forest model. The most important data displayed in the graph is still the variable 'same_area_shopping', the 'young families' group in the variable 'life_stage', and the 'premium' group in the variable 'price_segment', but the difference with the decision tree model are those the positions of the second and third important data are opposite. In addition, the selection of other important data is different from the decision tree model.



These variables will be focused on when providing recommendations.

4.8 XGBoost

XGboost is a model with relatively complex parameters. Similar to the bagging model, the XGboost model calculates multiple decision trees. The difference is that the bagging model will give the same weight to each data in each decision tree, while XGboost will give each data in the decision tree different weights, and the weighted average is based on the importance of the data. In addition, the processing rules of this model are running in sequence, rather than parallelizable processes like the bagging model. Although this operation mode will take more time than the bagging model, it is more consistent with the human decision-making mechanism that gives different weights according to the different importance of data.

The advantage of the XGboost model is that it can prevent overfitting while having a simple training procedure. When setting parameters, it can set the learning rate of the model, which allows the model to obtain optimal parameters when testing the accuracy of the model. The limitation of this model is that this model has many hyperparameters that can be tuned, and finding the optimal parameters can take a lot of time and require expertise. And it takes up a lot of memory on the computer. (Geeksforgeeks, n.d.)

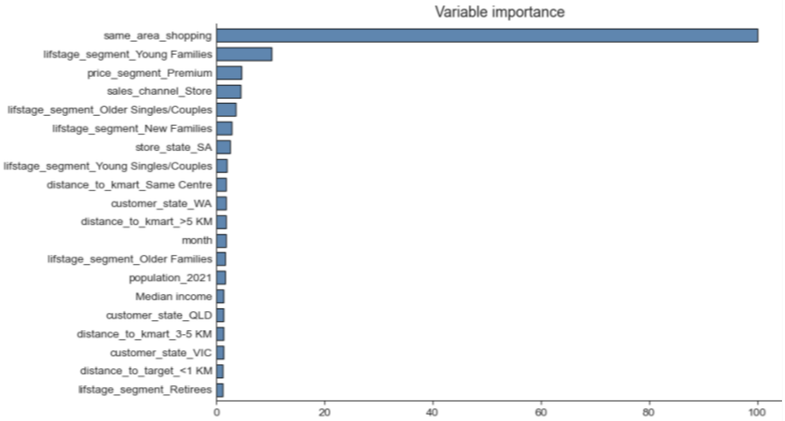
During the process of establishing the model and adjusting parameters, several sets of parameters were tested, and the best-performing parameters and the worst-performing parameters were selected for comparison. Among the parameters of the best-performing model, the learning rate was '0.01 -0.1'. Among the parameters of the worst-performing model, the learning rate is ‘0.1-1’. The final result is that the RMSE of the best-performing model is 2292.9, and the RMSE of the worst-performing model is 2298.454. This shows that the smaller the learning rate, the better the model performs. In the model with the smallest learning rate, the optimal parameters suggested by the model are that the test decision trees should be 50, the maximum depth of the decision tree should be 9, and the learning rate should be 0.1.

	subsample	n_estimators	max_depth	learning_rate
XGBoost	0.5	50	9	0.1

Bringing the validation data set into the model and processing it. The RMSE result is 2292.9 and RMSLE is 0.097, which is bigger than the result of the random forest model. This shows that the performance of XGboost is worse than the random forest model under the existing conditions.

	RMSE	RMSLE
Decision Tree	2245.305	0.098
Bagging	2264.917	0.099
Random Forest	2232.248	0.097
XGboosting	2292.9	0.097

XGboost also has the feature of automatically selecting the importance of variables, so the data that the XGboost model considers important can be displayed through the bar chart.



This figure proves that the variable 'same_area_shopping', the 'premium' group in the variable 'price_segment', and the 'young family' group in the variable 'life_stage' are the most important data, and the results are the same as the random forest model. However, the importance of other variables differed from the results of the random forest model.

4.9 Model Evaluation

Model	RMSE	RMSLE
Linear Regression	2503.524	0.105
Lasso Regression	2503.519	0.105
Ridge Regression	2503.555	0.105
Decision Tree	2245.305	0.098
Bagging	2264.917	0.099
Random Forest	2232.248	0.097
XGBoost	2292.9	0.097

To evaluate the summarized RMSE and RMSLE of all the models. Based on the table. Random Forest model is concluded to be the best model as it has the lowest RMSE and RMSLE, which is 2232.248 and 0.097 respectively. Therefore, the random forest model is chosen to refit with the combined train dataset and then used to predict the target values based on the test dataset. The final RMSE calculated based on the test set is 2266.369 and the RMSLE is 0.097.

5. Conclusion and recommendations

5.1 Conclusion

Based on the above analysis, It is concluded that BIG W should focus on four areas to improve its sales, which are geographical, target customer, competitors and promotion and media investment.

As for the geographical factors, offline shopping is still considered the main source of revenue for Big W which can't be abandoned. And although NSW has higher sales, WA will still be a better location choice as the potential market and customer purchasing power is high.

While the customer group indicates that BIG W's attracts the budget-young families most, as they pursue cost-effective products, which is consistent with Big W's product positioning.

For the Competitors, opening a store in the same shopping mall as WWS will cause unstable sales, but it can attract more customers to shop at Big W. In addition, when choosing the location of offline stores, the distance to Kmart is less than 1 kilometer, and the distance to Target is between 3 and 5 kilometers, which is a reference value.

Regarding the promotion and media investment part, both promotion and media investment will increase the total sales value of Big W. The random forest model proves the importance of media investment. The ACT state has the highest return on media investment.

According to the above key features, linear regression, ridge regression, lasso regression, decision tree, bagging, random forest, and XGBoost are used to build models predicting the total sales value. Based on the RMSE, and RMSLE, the random forest model (RMSE:2232.248, RMSLE:0.097) is regarded as the best model. The final RMSE and RMSLE of the random forest model based on the test set is 2266.369 and 0.097 respectively. Compared with other tree-based models, the random forest model increases the number of simulated decision trees and enhances the randomness of the selected data. .The main advantage of the random forest model is to select random data for each decision tree to differentiate each decision tree. It reduces data bias or contingency caused by data singleness or insufficient simulation times and retains the feature of the model that can arrange data according to the importance of the data. This helps the model perform better and reduces the RMSE value. However, the model can still be improved. Due to computation and time limitations, the hyperparameters setting is small. Besides, since randomized search is used to figure out the best parameters instead of the grid search, time is saved but the opportunity to find the optimal parameter is reduced and probably lead to a higher RMSE. In the future, the model could get better parameters by increasing 'n_estimators', the number of decision trees, and the number of model iterations.

According to the feature importance shown by all the models, the variables 'same_area_shopping', 'lifestage_segment_young families', 'price_segment_premium', 'month', and 'media_amount_spend' are the most important. This reflects that consumers like to shop near where they are.The young families prefer to shop in BIG W. And the customer group of the premium price segment will also have a significant negative effect on the customer willingness to shop in BIG W. Month and Big W’s investment in media will have an important impact on the total sales value. These are all along with the conclusion we obtain through the EDA part.

5.2 Recommendations

After analysis, we have made two suggestions for Bigw. The first suggestion is to open a new store in WA Province. Because WA has a small number of stores and high average sales, it has market potential. The specific suggestion is to open this new store in Bunbury city. Because this city is the second largest city in WA, with a population of approximately 90000, it has a relatively large population size. That means there is a sufficient number of potential customers, and currently there is only one big w store in Mandurah, making the market competitiveness relatively low. Its geographical location is between Perth and South Australia, and it is a transportation hub where many people pass through or travel to the area. This makes Mandurah a strategic location suitable for retailers to open new stores.

Considering the same_ Area_ Shopping is the most important factor in feature importance, indicating that people generally go to shops near their homes to purchase things. Therefore, it is recommended to open new stores in the city center, as there are a large number of students and young families living in the city center. Considering the factors of competitors, the location of the store will be located less than 1 kilometer from Kmart and no more than 3-5 kilometers from the Target. Budget-young families are the main target group. It is recommended to set up a product area for young families at the entrance of the store. When the target group passes by the store, they will first see the products they often purchase, which can stimulate their purchasing desire. Each week, different discounts are introduced for young families' products, which will attract people to consume. Updating products with different discounts on a weekly basis will increase the frequency of customers entering the store to make purchases.

Regarding promotions, the store can set up a monthly promotion theme, such as "Home Decoration Month", "Super Electrical Month", and "Summer Outdoor Activity Month", and provide discounts and promotions for related products. According to EDA analysis, the peak sales amount is reached at the middle and end of each year, and the store needs to provide more targeted holiday promotions. For example, BIG W has launched a holiday gift set, such as a Christmas gift basket or Valentine's Day gift bag. This gift set sales strategy belongs to bundled sales, and some products with poor sales can be added to the set, which can increase the overall sales of the store. Big W can also launch a holiday shopping voucher gift activity, where customers can receive a holiday shopping voucher of 8 Australian dollars when the shopping amount reaches 88 Australian dollars, encouraging them to visit the store again during the holiday period. Our store's promotional products should mainly choose products that are cheap and cost-effective, rather than high-end products, as our main customers come from people with limited budgets and pursuing high cost efficiency.

The second suggestion is to focus on e-commerce operations. In order to achieve higher revenue, Big W should also pay attention to online sales channels, conduct promotional activities online, provide special discounts and discounts for online shopping, improve price attractiveness, and attract online customers. Now Big W provides free delivery services for online shopping with a consumption of over AUD 100. It is recommended to lower the shopping amount standard for free delivery and reduce the shopping amount to over AUD 60 for free delivery. Utilize data analysis and machine learning techniques to provide personalized product recommendations to customers, in order to increase the degree of personalization in the shopping experience. Through price competition and personalized recommendations, the competitiveness of Big W's online channels has been improved, which is conducive to competing with e-commerce platforms such as Amazon in the market. The Big W website allows users to participate in small activities on the webpage, such as daily lucky draws. Users have the opportunity to receive gifts such as shipping vouchers, vouchers, and 9.5 discount coupons, which can increase user engagement and increase user stickiness. Using this method will increase the platform's control over customers. When the platform detects a decrease in customers' recent shopping frequency, in order to prevent customer churn, the platform will give customers a free shipping voucher the next time they click on the webpage, which will be sent in the form of a lottery to better retain users. At the same time, business operations also require investment in the media. Due to our main customer base being budget-young families, our advertising investment will be promoted on YouTube and other social media platforms. The advertising time is mainly concentrated in May and November, and we choose to advertise young families products at affordable prices.

These suggestions may face some risks and limitations. The main risk is that the cost of opening a store in the city center is higher, which may result in a budget shortage. Global and local economic fluctuations may have an impact on the retail industry. Due to the lack of recovery in the market economy, even if promotional activities are launched, the expected sales may not be achieved. The investment in advertising costs is large, and these investments may not effectively translate into customer purchasing power. Online shopping can lead to issues such as lost delivery and delayed refunds, reducing customer satisfaction with the shopping experience. The main limitation is the presence of other large retailers such as ALDI in Mandurah, which means Big W will need to compete, attract local customers, and develop attractive pricing and promotional strategies.

References

Australian Bureau of Statistics. (n.d.). 2021 Mandurah, Census All persons QuickStats | Australian Bureau of Statistics. https://www.abs.gov.au/census/find-census-data/quickstats/2021/LGA55110

CFI. (n.d.). Bagging (Bootstrap Aggregation).https://corporatefinanceinstitute.com/resources/data-science/bagging-bootstrap-aggregation/

Coursera. (2023). Decision Trees in Machine Learning: Two Types (+ Examples).https://www.coursera.org/articles/decision-tree-machine-learning

Delivery offers. (n.d.). BIG W. https://www.bigw.com.au/delivery-pickup/delivery-offers

EDUCBA. (2023). Decision tree limitations. https://www.educba.com/decision-tree-limitations/

Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2022). A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. International Statistical Review, 90(1), 118–145. https://doi.org/10.1111/insr.12469

Geeksforgeeks. (n.d.). XGboost. https://www.geeksforgeeks.org/xgboost/

GreatLearning. (2023). Random forest Algorithm in Machine learning: An Overview. https://www.mygreatlearning.com/blog/random-forest-algorithm/#:~:text=However%2C%20despite%20these%20advantages%2C%20a,trees%20to%20determine%20the%20class.

Heuvelmans, G., Muys, B., & Feyen, J. (2006). Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology (Amsterdam), 319(1), 245–265. https://doi.org/10.1016/j.jhydrol.2005.07.030

Jiao, S., Gao, Y., Feng, J., Lei, T., & Yuan, X. (2020). Does deep learning always outperform simple linear regression in optical imaging? Optics Express, 28(3), 3717–3731.https://doi.org/10.1364/OE.382319

Laufer, B., Docherty, P. D., Murray, R., Krueger-Ziolek, S., Jalal, N. A., Hoeflinger, F., Rupitsch, S. J., Reindl, L., & Moeller, K. (2023). Sensor Selection for Tidal Volume Determination via Linear Regression—Impact of Lasso versus Ridge Regression. Sensors (Basel, Switzerland), 23(17), 7407–. https://doi.org/10.3390/s23177407

Pereira, J. M., Basto, M., & Silva, A. F. da. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. Procedia Economics and Finance, 39, 634–641. https://doi.org/10.1016/S2212-5671(16)30310-0

Qasim, M., Månsson, K., & Golam Kibria, B. M. (2021). On some beta ridge regression estimators: method, simulation and application. Journal of Statistical Computation and Simulation, 91(9), 1699–1712. https://doi.org/10.1080/00949655.2020.1867549

Ranstam, J., & Cook, J. A. (2018). LASSO regression. British Journal of Surgery, 105(10), 1348–1348. https://doi.org/10.1002/bjs.10895

Saleh, A. K. M. E., Kibria, B. M. G., & Arashi, M. (Mohammad). (2019). Theory of ridge regression estimators with applications (1st edition). Wiley.

Sharma, U., Gupta, N., & Verma, M. (2023). Prediction of compressive strength of G GBFS and Flyash-based geopolymer composite by linear regression, lasso regression, and ridge regression. Asian Journal of Civil Engineering, 24(8), 3399–3411. https://doi.org/10.1007/s42107-023-00721-2

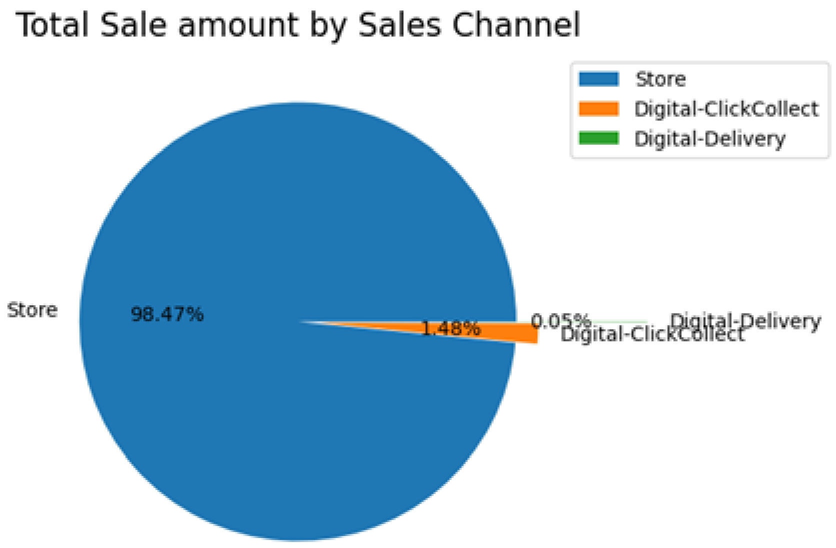
Tracy Grimshaw, Reid Butler, Teresa Rendo, Sophie Elsworth, & David Walker. (2017). A Current Affair.

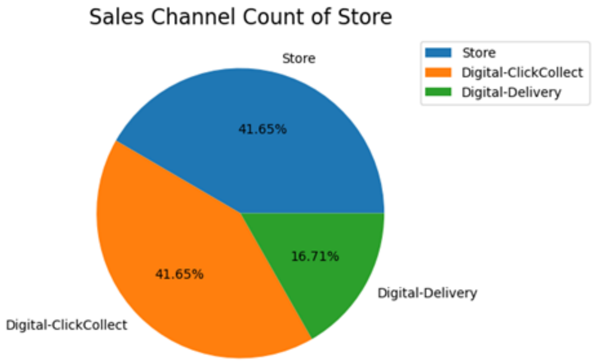
Visit Mandurah. (2022, July 15). 10 Things You Didn't Know About Mandurah https://visitmandurah.com/10-things-you-didnt-know-about-mandurah/

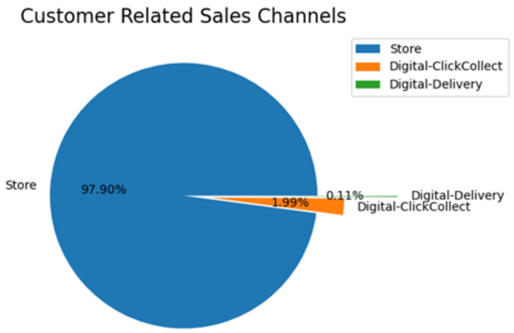
Appendix A

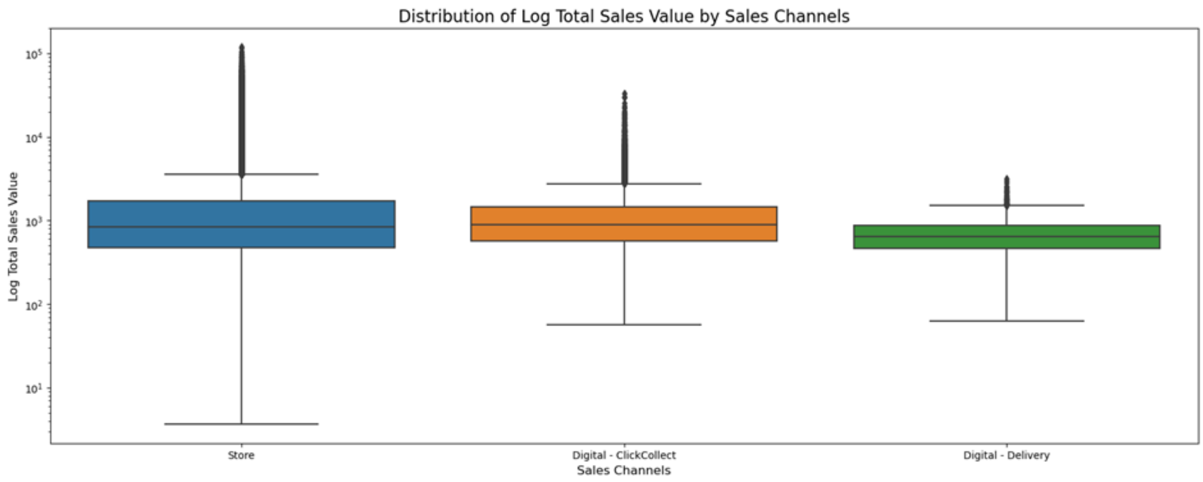
Customer State	Price Life stage Segment
NSW	Budget-Young Families
QLD	Budget-Young Families
VIC	Budget-Young Families
WA	Budget-Young Families
SA	Budget-Young Families
ACT	Mainstream-Young Families
TAS	Budget-Young Families
NT	Mainstream-Older Singles/Couples

Appendix B


Total Sale amount by Sales Channel

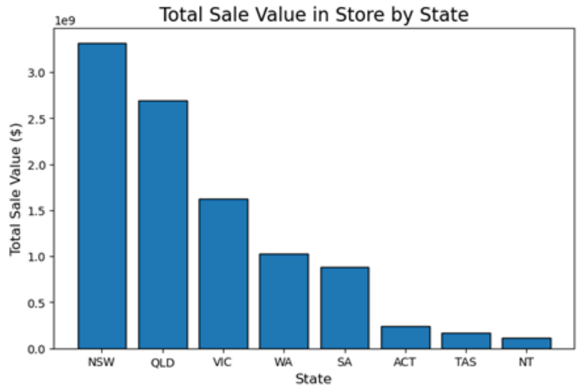

Sales Channel Count of Store


Customer Related Sales Channels

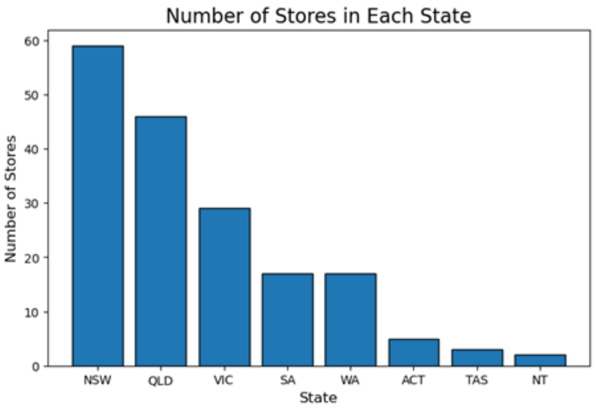


Distribution of Log Total Sales Value by Sales Channels

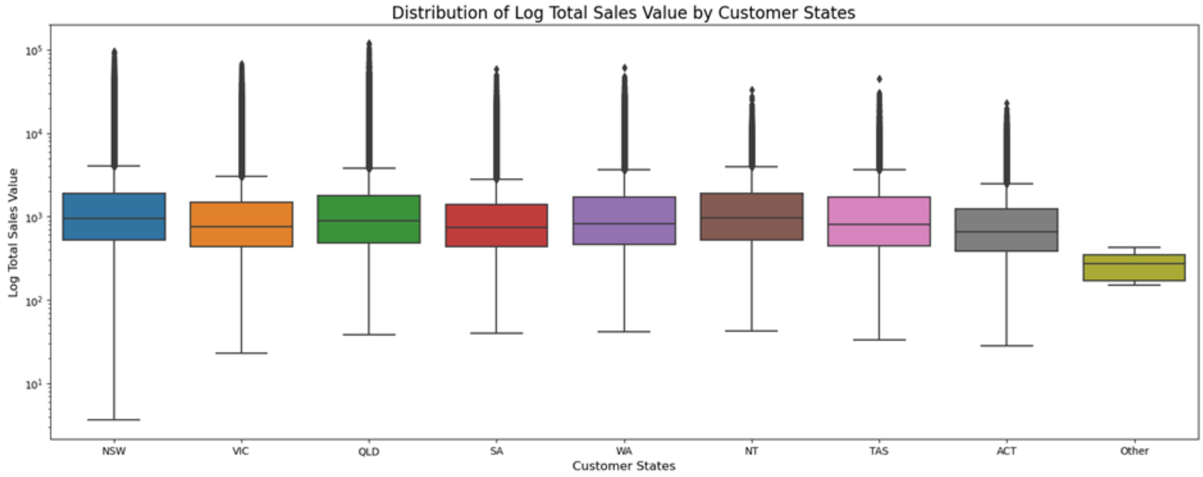
Appendix C



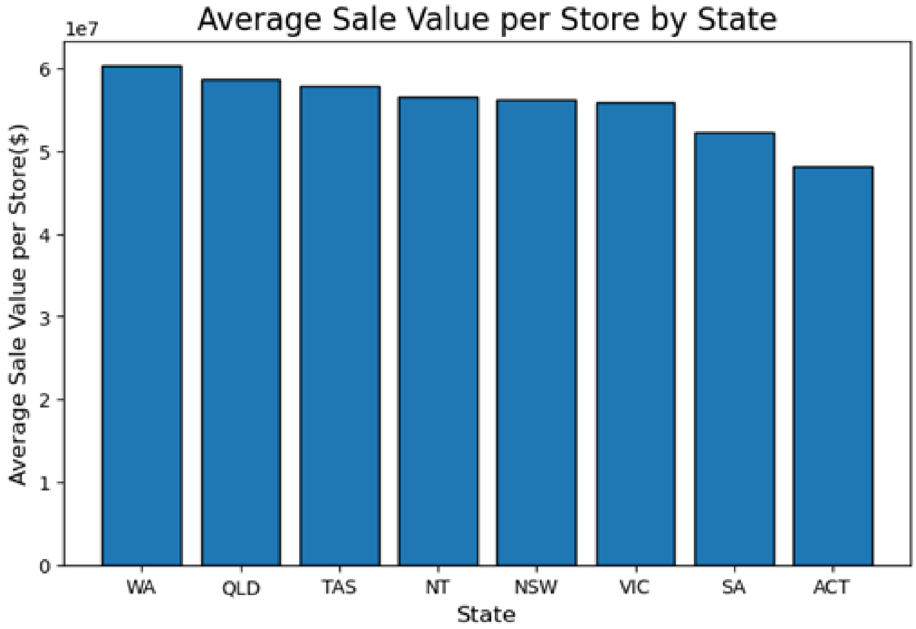
Total Sale Value in Store by State

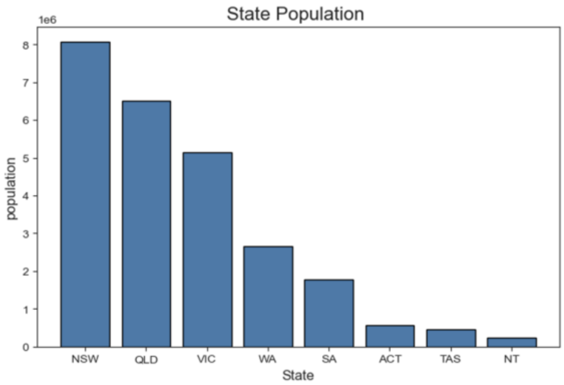


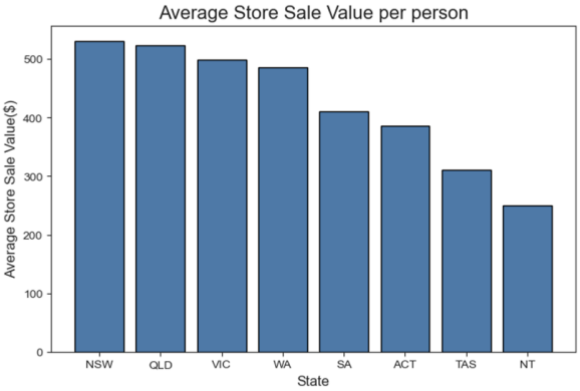
Number of Stores in Each State



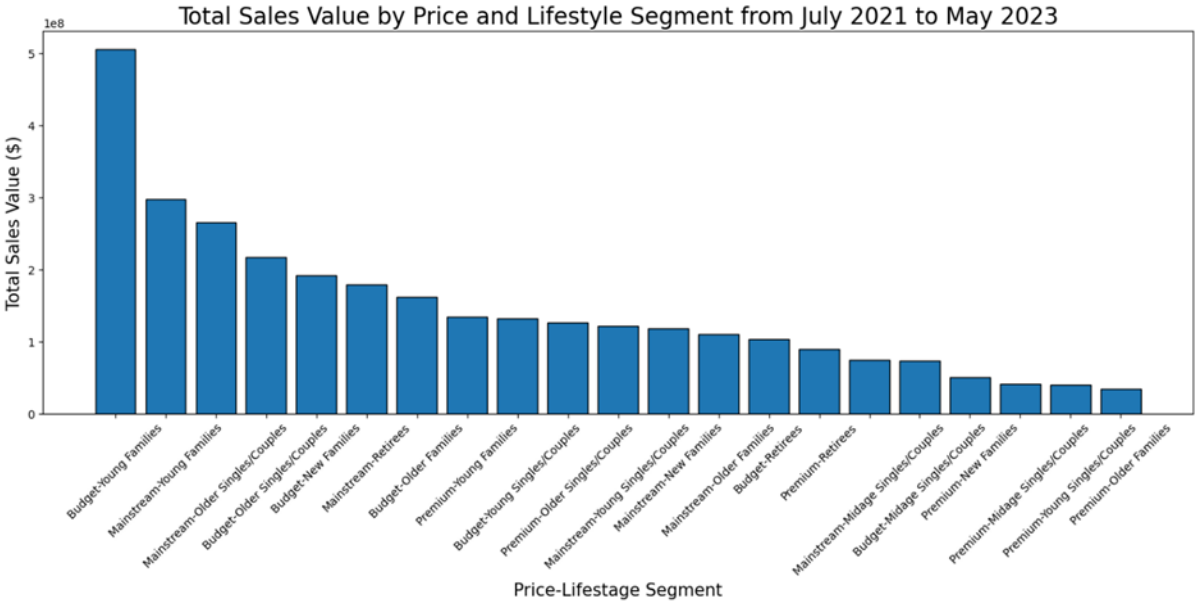
Distribution of Log Total Sales Value by Customer States

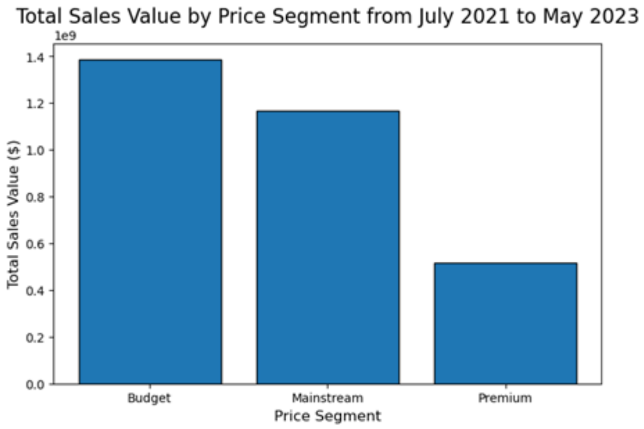

Average Sale Value per Store by State
1e7
Average Sale Value per Store($)
0
1
2
3
4
5
6
WA
QLD
TAS
NT
NSW
VIC
SA
ACT
State

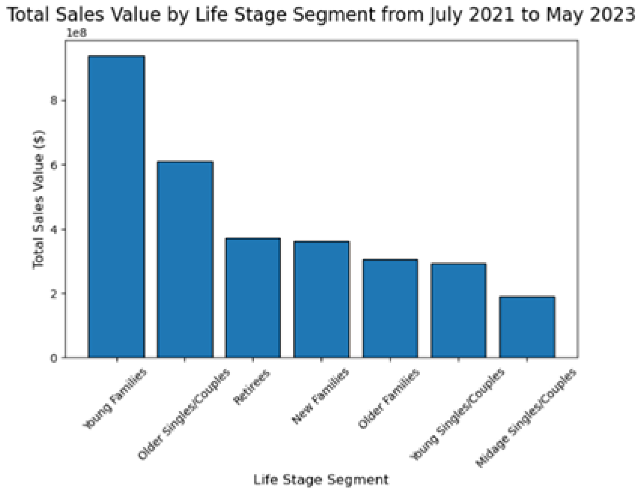

State Population
1e6
population
0
1
2
3
4
5
6
7
8
NSW
QLD
VIC
WA
SA
ACT
TAS
NT
State


Average Store Sale Value per person
Average Store Sale Value($)
0
100
200
300
400
500
NSW
QLD
VIC
WA
SA
ACT
TAS
NT
State

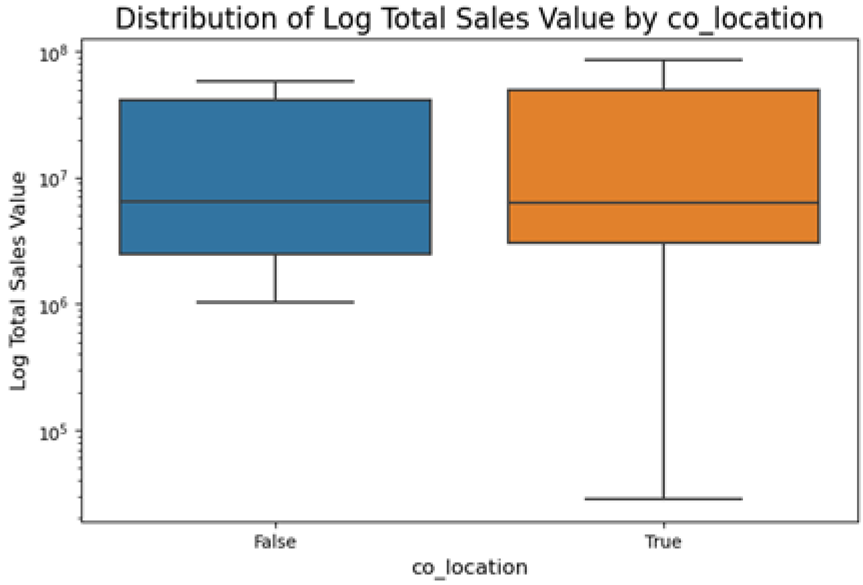
Appendix D


Total Sales Value by Price and Lifestyle Segment from July 2021 to May 2023
1e8
Total Sales Value ($)
0
1
2
3
4
5
Budget-Young Families
Mainstream-Young Families
Mainstream-Older Singles/Couples
Budget-Older Singles/Couples
Budget-New Families
Mainstream-Retirees
Budget-Older Families
Premium-Young Families
Budget-Young Singles/Couples
Premium-Older Singles/Couples
Mainstream-Young Singles/Couples
Mainstream-New Families
Mainstream-Older Families
Budget-Retirees
Premium-Retirees
Mainstream-Midage Singles/Couples
Budget-Midage Singles/Couples
Premium-New Families
Premium-Midage Singles/Couples
Premium-Young Singles/Couples
Premium-Older Families
Price-Lifestage Segment

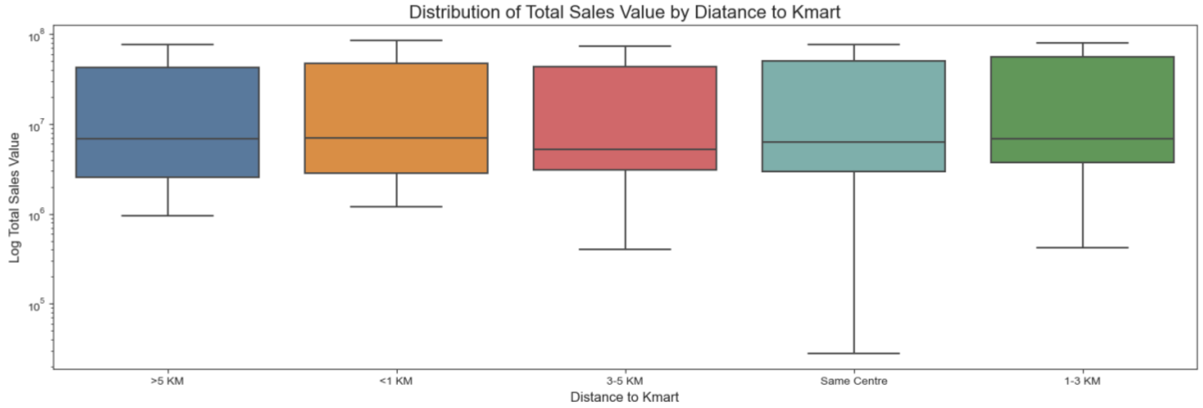

Total Sales Value by Price Segment from July 2021 to May 2023
1e9
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Total Sales Value ($)
Budget
Mainstream
Premium
Price Segment

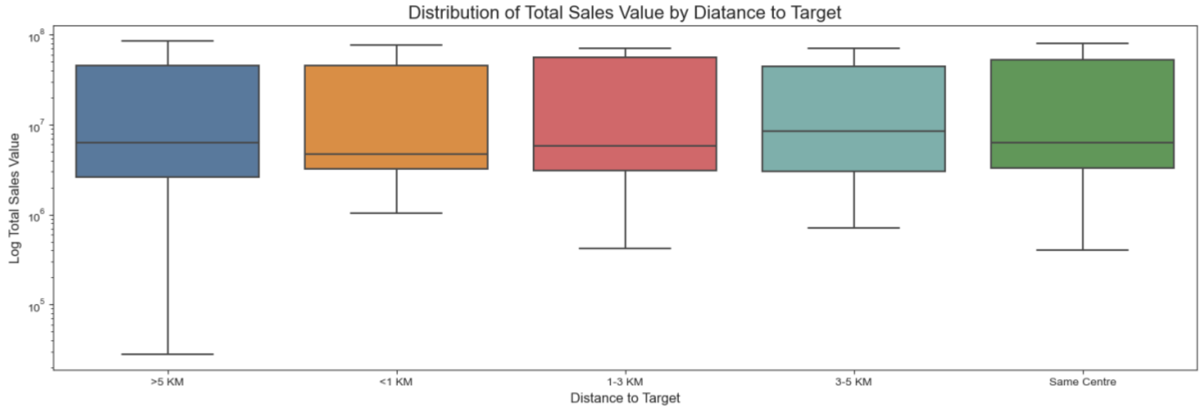

Total Sales Value by Life Stage Segment from July 2021 to May 2023
1e8
8
6
4
2
0
Total Sales Value ($)
Young Families
Older Singles/Couples
Retirees
New Families
Older Families
Young Singles/Couples
Midage Singles/Couples
Life Stage Segment

Appendix E

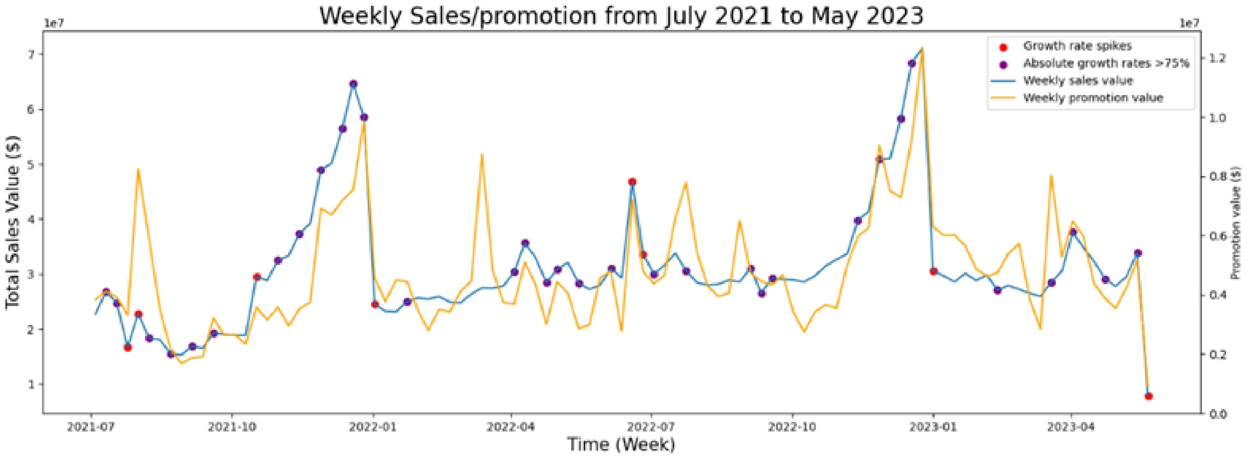

Distribution of Log Total Sales Value by co_location
10^8
10^7
10^6
10^5
Log Total Sales Value
False
True
co_location

Appendix F





Appendix G


Weekly Sales/promotion from July 2021 to May 2023
1e7
Total Sales Value ($)
1e7
Promotion value ($)
Growth rate spikes
Absolute growth rates >75%
Weekly sales value
Weekly promotion value
Time (Week)