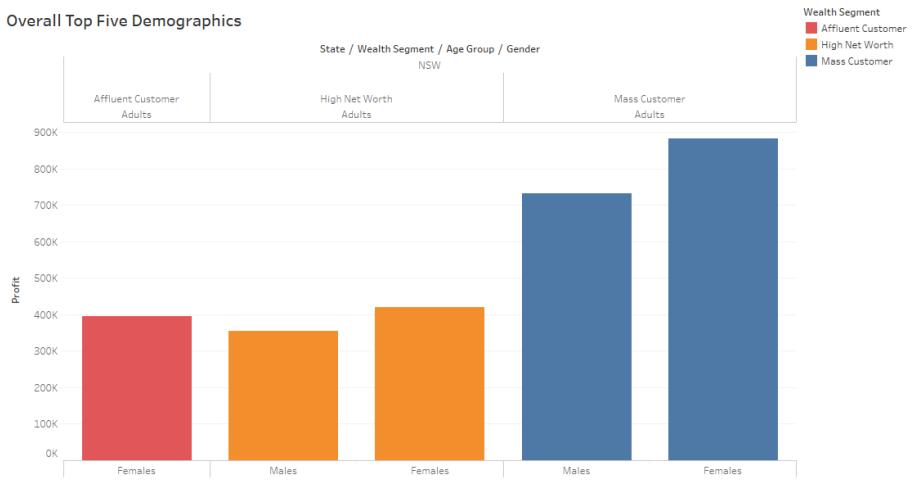
EXECUTIVE SUMMARY

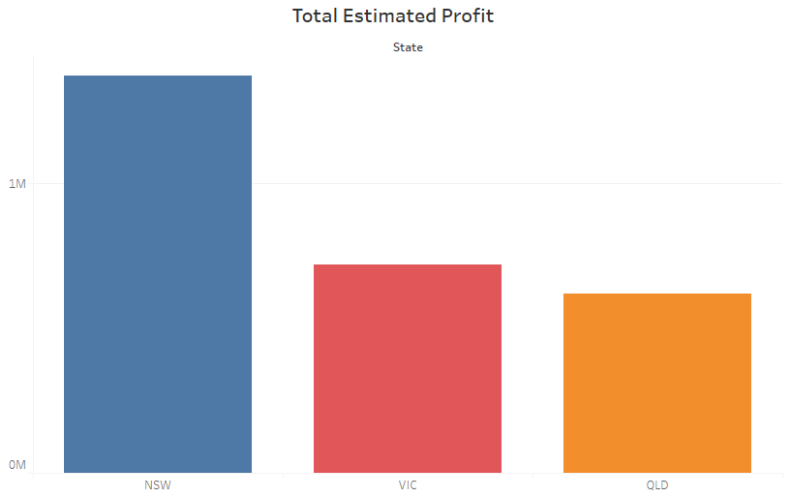
I set out to identify which of the new customers Sprocket Central Pty Ltd just acquired should they target with their marketing campaign based on the historical datasets they have.

I first performed some analysis on the data and came up with three kinds of groupings – 1-factor groups, 3-factor groups and 4-factor groups.

The data shows females, adults, mass customers and New South Wales to be the best set of customers to focus on individually (1-factor grouping). And the overall top 5 customers to focus on to be adults high net worth, mass customers and affluent customers in New South Wales.



Afterwards, I performed predictive modeling on the data to estimate profits from this new customers and this also shows most profit are to be expected from this state – New South Wales.



Writing by Fhuad Balogun

DATA ANALYSIS

Data preparation:

Observation in the raw data.

1. I have the same variables in the old 'CustomerDemographic' and 'CustomerAddress' tables and in the 'NewCustomerList' table. This is good to know, in case I will be building a predictive model, I can easily know variables that won't be available during production.
2. Not all customers that are in the demographics and address tables are in Transaction table.
3. A variable named 'default' in customer demographic table has been deleted due to that it contains information that seem to be corrupt.
4. All gender value stated as 'U' have their DOB missing except one (has year 1843) which is most likely an incorrect input.

For Customer Demographic:

1. Replaced 'F' and 'Femal' to be 'Female' and 'M' to be 'Male'.
2. Created 'age' variable from 'DOB' and subtract the year from 2018 (dividing it by the year 2018 as I assume to be analysing for the year 2018) to derive the age in number of years.
3. Converted all ages older than 75 to NaN to better reflect customers that are young enough to ride a bike.
4. I create age group variable with intervals 16-19 as teen, 20-29 as young adult, 30-50 as adult, 51-64 as elder, 65-80 as senior to better reflect the difference in age group lifestyle.

For Customer Address, I converted all 'New South Wales' to 'NSW', and 'Victoria' to 'VIC' to be consistent with the majority of the data.

Concatenate customer demographics and customer address tables together and dropped missing values (I then have 3908 observations). I don't want to impute missing values for the analysis, I only want to work with complete feature vectors.

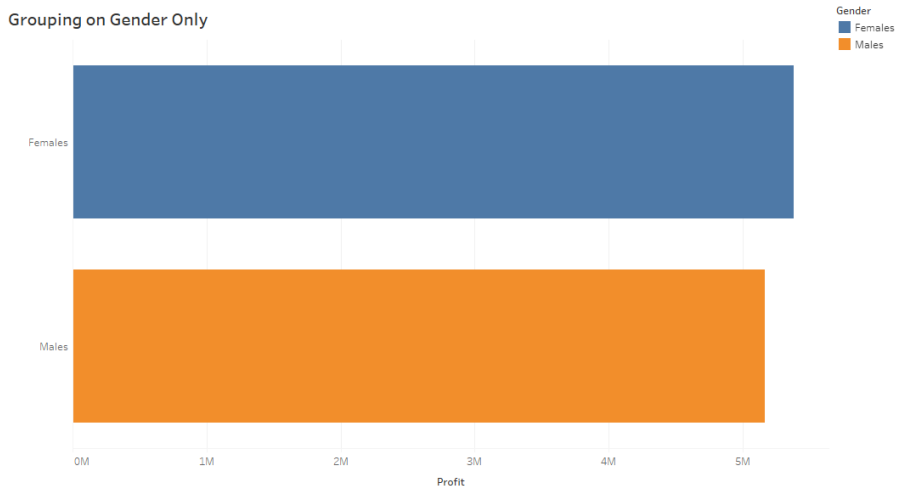
For Transactions, removed cancelled transaction to only have approved transactions.

Merged approved transactions and customers table together to create a table that fills up all transactions with the customer details of the customer.

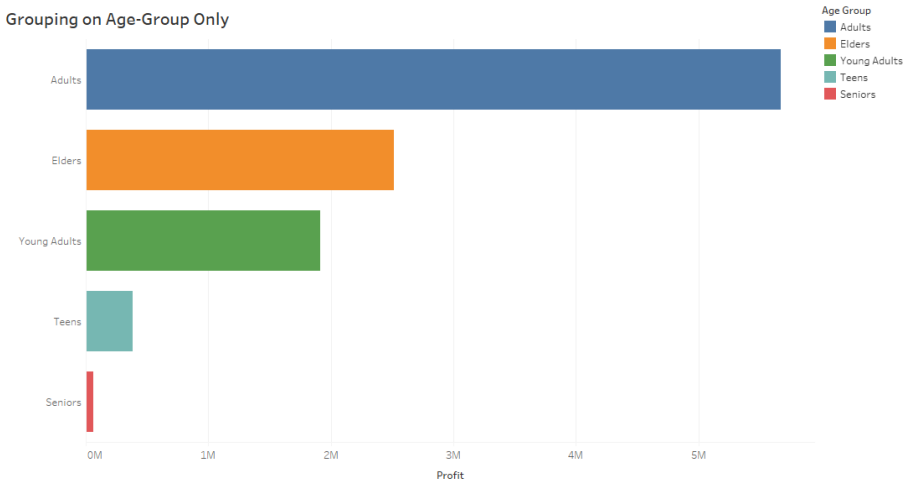
ANALYTICS:

One-factor grouping:

Looking at the gender alone, it seems to not be much difference in the profit made from male and female customers. I could go deeper to ascertain if there is a mean population difference between the two using confidence intervals.

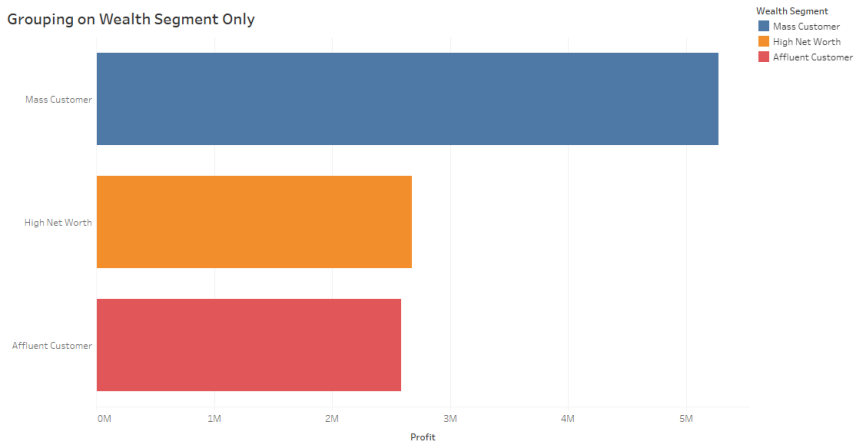


Looking at age groups alone, the top three age groups are 'adults', 'elders' and 'young adults' with 'adults' profit being heavily high. This might be as a result of the 'adults' group having the largest interval, however, the interval illustrates better a stage in the lifestyle of people during which they generally have found their career and life purpose and are likely to get a bike for recreation and exercise only.

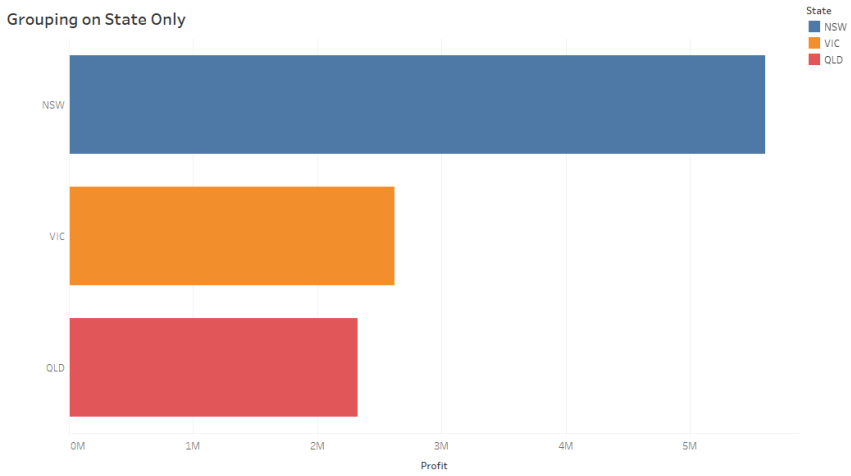


In wealth segment, 'mass customers' has the highest profit above 5 million dollars, followed by 'high net worth' customers and the 'affluent customers' – they are both a little above 2.5 million dollars in profit in 2017. Mass customers have the highest number

of orders placed in 2017 (9600) as opposed to the other two customer groups with just almost 5000 orders each. Although, this might have also influenced the amount of profit made from Mass Customers, it still points to the fact that more of the companies customers are within this group.



Similar situation is obvious when I look at the state segmentation, the highest profit was above 5 million dollars from New South Wales, followed by Victoria State (about 2.6 million) and then Queensland which was less than 2.5 million dollars. In this data, the total number of customers in New South Wales (NSW) is 2,141, in Victoria (VIC) is 1021 and in Queensland (QLD) is 838. There might have been some bias when pulling this data sample from the source where NSW happened to have been pulled out more. Another possibility could be that the organization started business there and so had focused a lot of resources in gaining more customers in the state early on who still happens to still shop with the company. More so, this early marketing campaign in NSW might have caused a ripple effect that still fetch in customers in different ways for the company more than in the other two states.

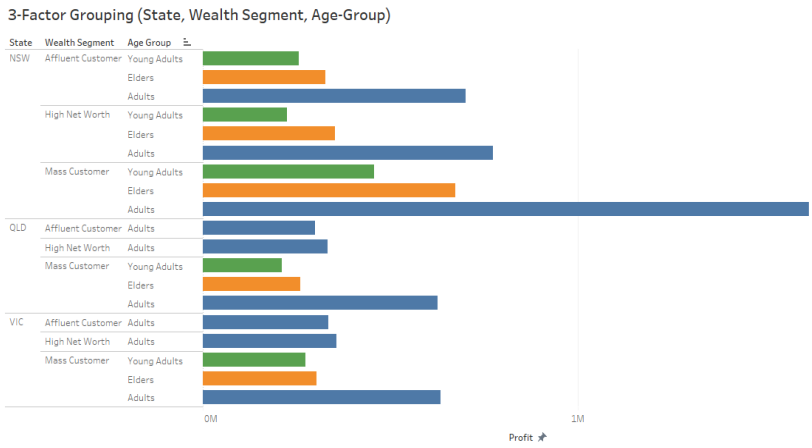


Writing by Fhuad Balogun

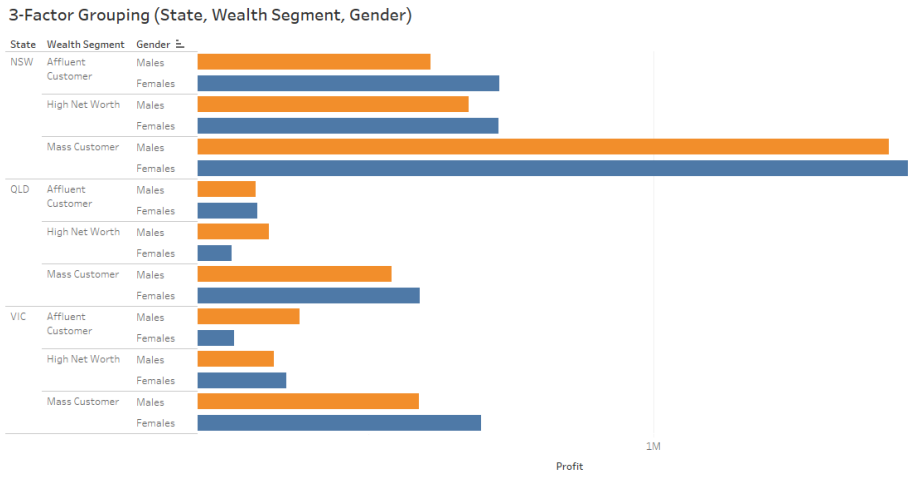
Three-factor grouping:

After creating the visualization, I filtered to only show customer groups with total sum of profit higher than $200,000. This is so that I can easily find information from the topmost customer groups.

First, I grouped the customers by states, wealth segments and age groups only. Generally, the most profited customers were the Adults across all three states. When I looked a little bit deeper, I found that the most profited state still was NSW; and its top three age groups (Adults, Elders and Young Adults) were Mass Customers and each were higher than their counterparts in other states. This was followed by the same age groups in NSW that were High Net Worth customers.



Secondly, I grouped the customers again by states, wealth segment and gender only. Female customers brought more profit to the company than male customers across all wealth segments in the three states except in 'QLD-High Net Worth' customers and 'VIC-Affluent' customers. The best state by far was still NSW and the Mass Customers were the best within the state.



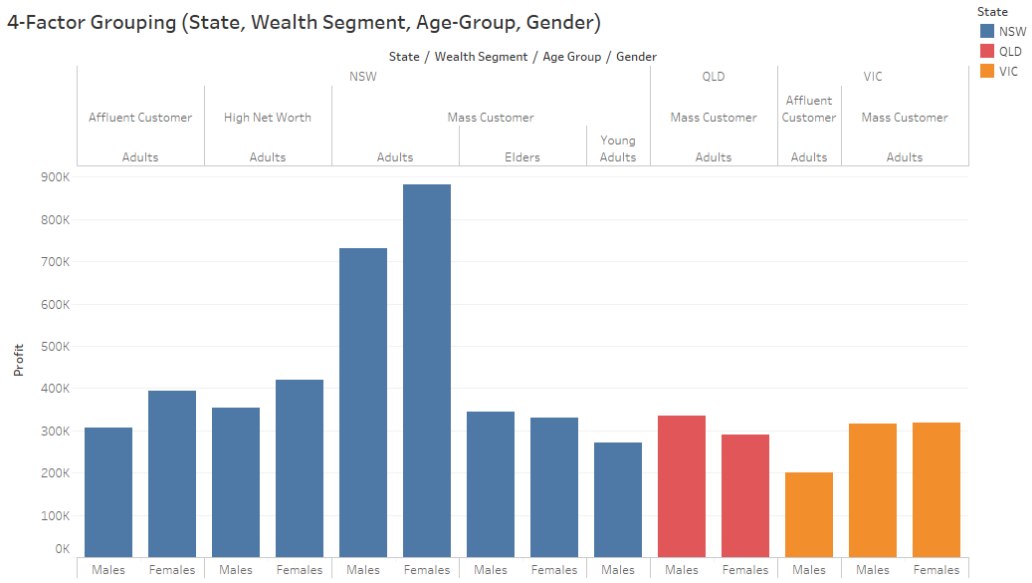
Writing by Fhuad Balogun

Four-factor grouping:

I also filtered to only show customer groups with total sum of profit higher than $200,000.

I grouped the customers by states, wealth segments, age groups and then gender. In NSW, all top customer wealth segments have only adult customers (both male and female) except for mass customers that also included elders (male and female) and young adults (male only).

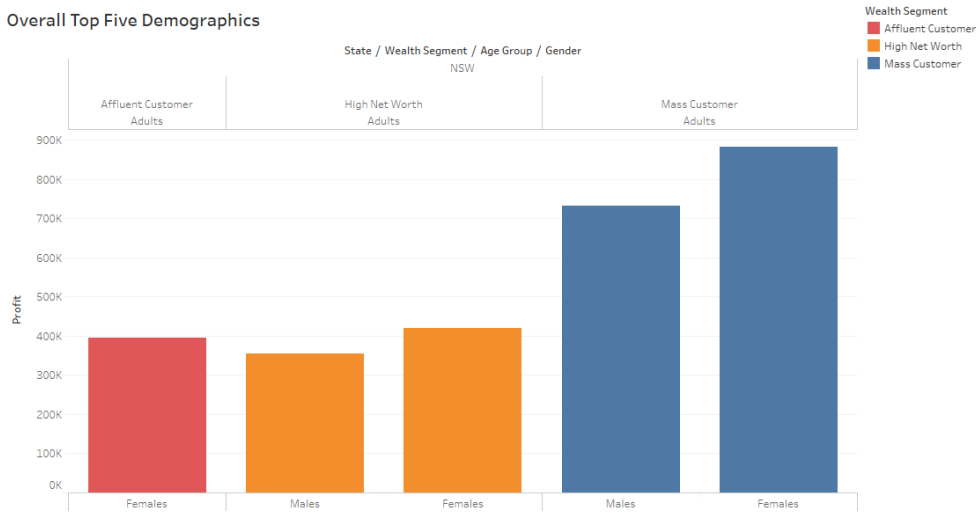
In QLD, only male and female adult mass customers appeared (i.e. have profits higher than $200K). In VIC, male and female adult mass customers also appeared along with only adult male affluent customers. The adult mass customers in NSW had the most profit.



In summary, I would say that in 2017, looking at gender, female customers were more profitable than male customers. In age groups, adult customers were the most profitable. In wealth segment, mass customers were the most profitable. In state, NSW was the most profitable state in 2017. And putting them together in a four-factor grouping, they all manifested in the result to show that the most profitable demographic are the adult female mass customers that live in New South Wales. The overall 5 most profitable customer demographics are:

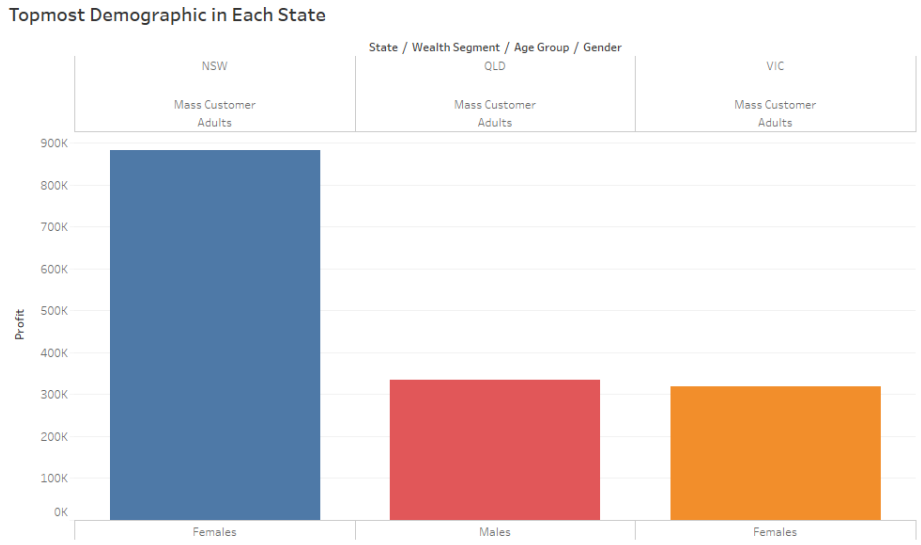
1. Female adult mass customers in New South Wales.
2. Male adult mass customers in New South Wales.
3. Female adult high net worth customers in New South Wales.
4. Female adult affluent customers in New South Wales.
5. Male adult high net worth customers in New South Wales.

Writing by Fhuad Balogun



Most profitable customer demographics one from each state are:

1. Female adult mass customers in New South Wales.
2. Male adult mass customers in Queensland.
3. Female adult mass customers in Victoria.



CONCLUSION

Moving forward, I estimated the population mean difference between the two best 4-factor demographics using confidence intervals at 95% confidence level. This is to see if customers in the topmost demographic in this sample are likely to be entirely higher than the second top demographic in the population on average.

When I estimated the population mean difference of between the two best 4-factor demographics with 95% confidence level, the resulting interval was between -27.7595

Writing by Fhuad Balogun

and 44.1437. There are both positive and negative in the range of reasonable values, which includes zero within this range. This means that the profit from these two groups could be equal, or either one being more than the other if we look at the population.

RECOMMENDATION

I recommend that your company go as deep as this with its customer grouping and focus on these top 5 most profitable demographics with its new customer details acquire. You should create marketing campaigns that cater specifically for each group by researching on what provokes buying within these groups and things that could connect to them emotionally and provoke them to buy or even decides to change the company they buy from to your company.

If you are to look at creating marketing campaigns specific for each states store and customers, then you can also focus on the best demographic of each state as listed above. However, since NSW is likely to be your major region, the campaigns and resources can be focused on the top 5 in that state as also listed above.

PREDICTIVE MODELING

Data Preparation:

I merged customer demographics and customer address tables together to create a new table.

I then created a profit variable by subtracting 'list_price' and 'standard_cost' variables in the transactions table. After this, I added up all profit for each customer ID from the transaction table and appended it to its corresponding customer ID in the newly merged customers table.

I dropped some variables based on the following reasons:

1. Customer ID: Cannot be used for prediction.
2. First name: Cannot be used for prediction.
3. Last name: Cannot be used for prediction.
4. Job Title: Don't want to go too deep in the relationship between job title and profit made.
5. Deceased indicator: Cannot be used for prediction. I have to be sure the customer to be predicted on isn't dead.
6. Address: Don't want to go too deep in the relationship between address and profit made.
7. Postcode: I want to stay afloat within the states only.
8. Country: Zero variance. All customers are in the same country – Austraila.

I replaced all 'F' and 'Femal' with 'Female' and also replaced 'U' and '247' with the most occurring gender in the data – Female.

I created an age variable with information from the DOB variable.

I replaced all NAs in 'job industry category' variable with most occurring category in the variable – Manufacturing.

I replaced all NAs in 'tenure' variable with the median of the variable – 11.

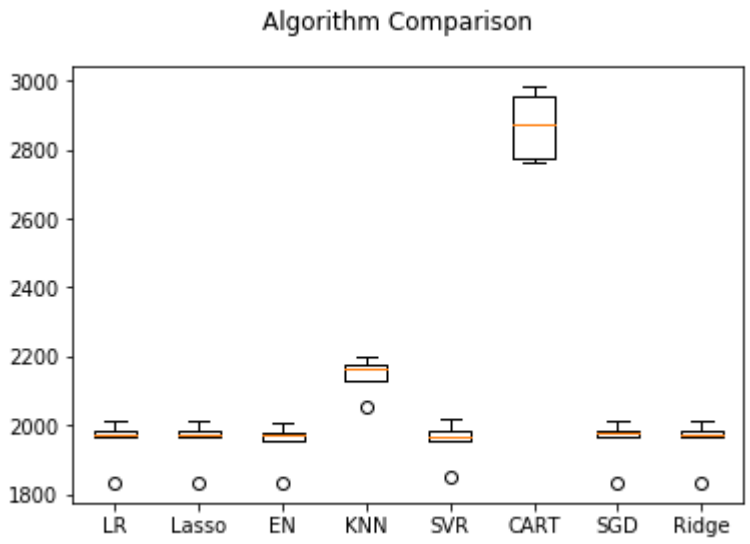
I replaced all NAs in 'ag' variable with the median of the variable – 41.

I then standardised the numerical variables using Scikit-Learn's StandardScaler class. I converted the categorical variables into dummy variables using Scikit-Learn's OneHotEncoder class.

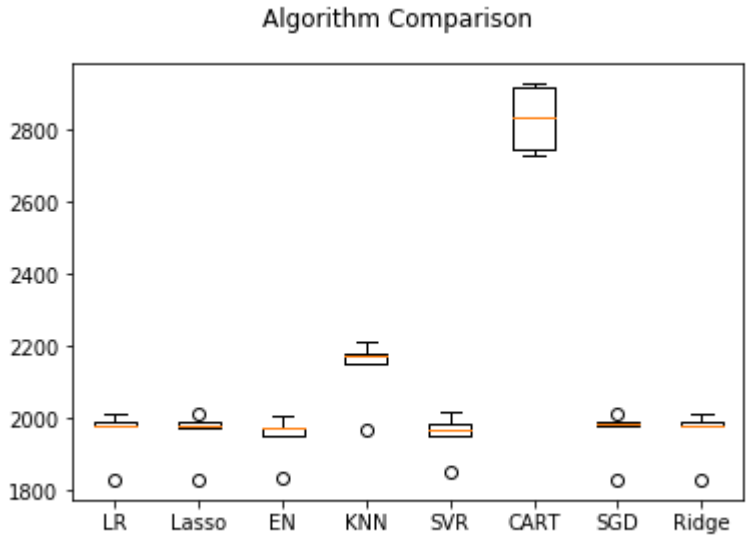
Modeling and Evaluation:

I spot-checked 6 traditional algorithms initially – Linear Regression, Lasso, Elastic Net, K-N earest Neighbours, SVR and Decision Tree Regressor, with 'tenure' and 'property_valuati

on' variables used as numerical variables. I used RMSE as my metric. The best model was the Elastic Net with the lowest RMSE score of 1948.529.



In the second trial, I used 'tenure' and 'property_valuation' as discrete variables and it improved and ElasticNet (1947.655) was still the best algorithm and it also improved its performance than in the first trial.



In both trials:
1. All linear models are around a close range. The problems seems to be one that linear models might be best fit for.
2. ElasticNet has the lowest RMSE score. Shows to be the best model to try and look into.

3. ElasticNet got better result in the second trial. So I leave tenure and property_valuatio n to be in discrete form.

After fine-tuning ElasticNet, the best hyperparameter values were 'alpha': 10.0, 'l1_ratio': 0.7, 'max_iter': 2000. When I retrain the algorithm with these hyperparameter values and test them on the test set, I had an RMSE score of 1931.061.

Deployment:

I used the trained model to estimate profit on 1000 new sets of customers. In this group, t he most customers are from New South Wales (506), followed by Victoria (266) and lastl y, Queensland (228). This is also apparent in the estimated total profit. However, looking at the estimated mean profit, Queensland has the highest mean ($2,707.88) followed b y New South Wales ($2,703.15) and then Victoria ($2,700.14). This, however, doesn't app ear to have much difference.