Predicting supermarket sales

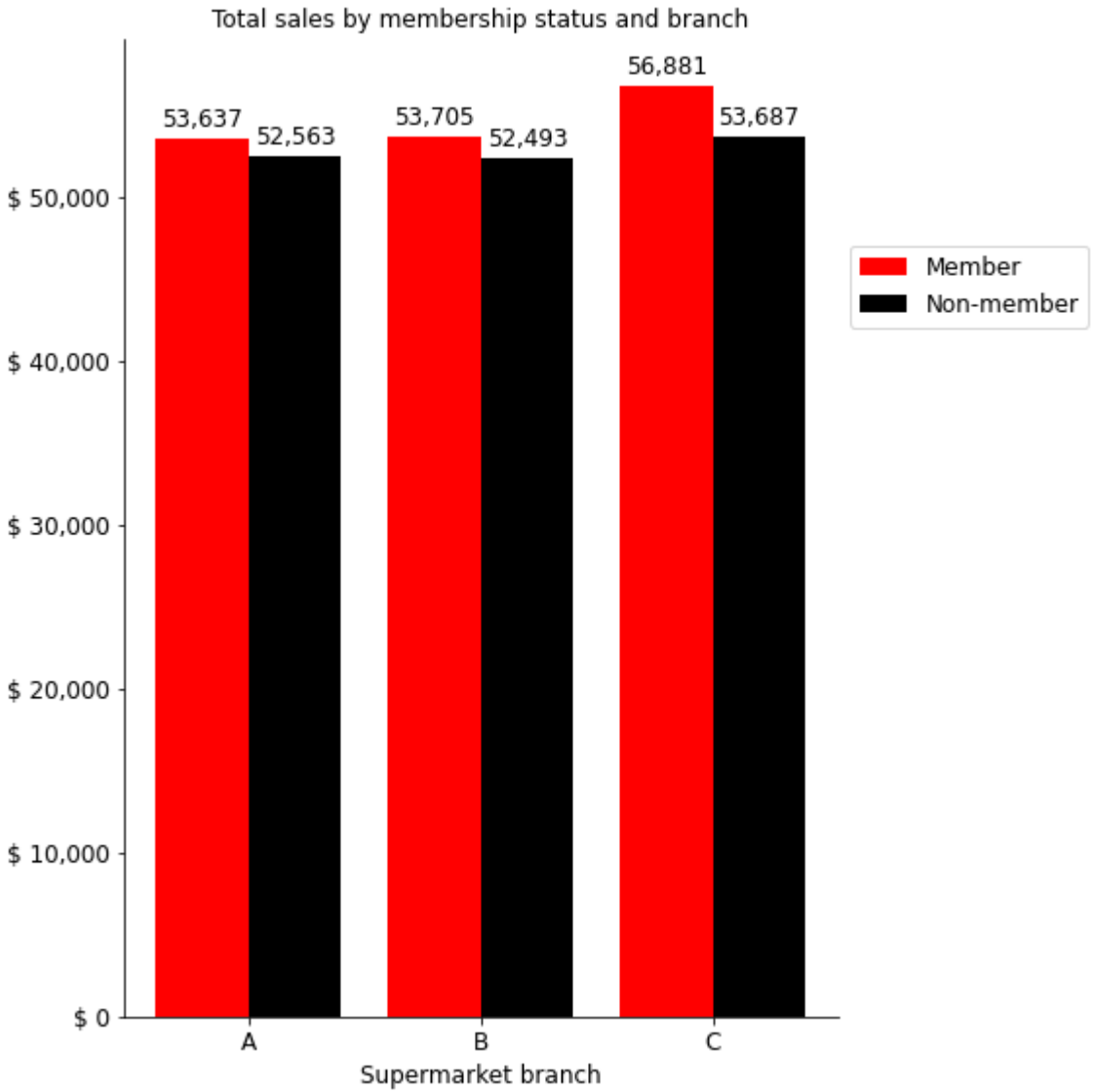
Data analysis, interpretation and prediction

by Amy Birdee

Introduction

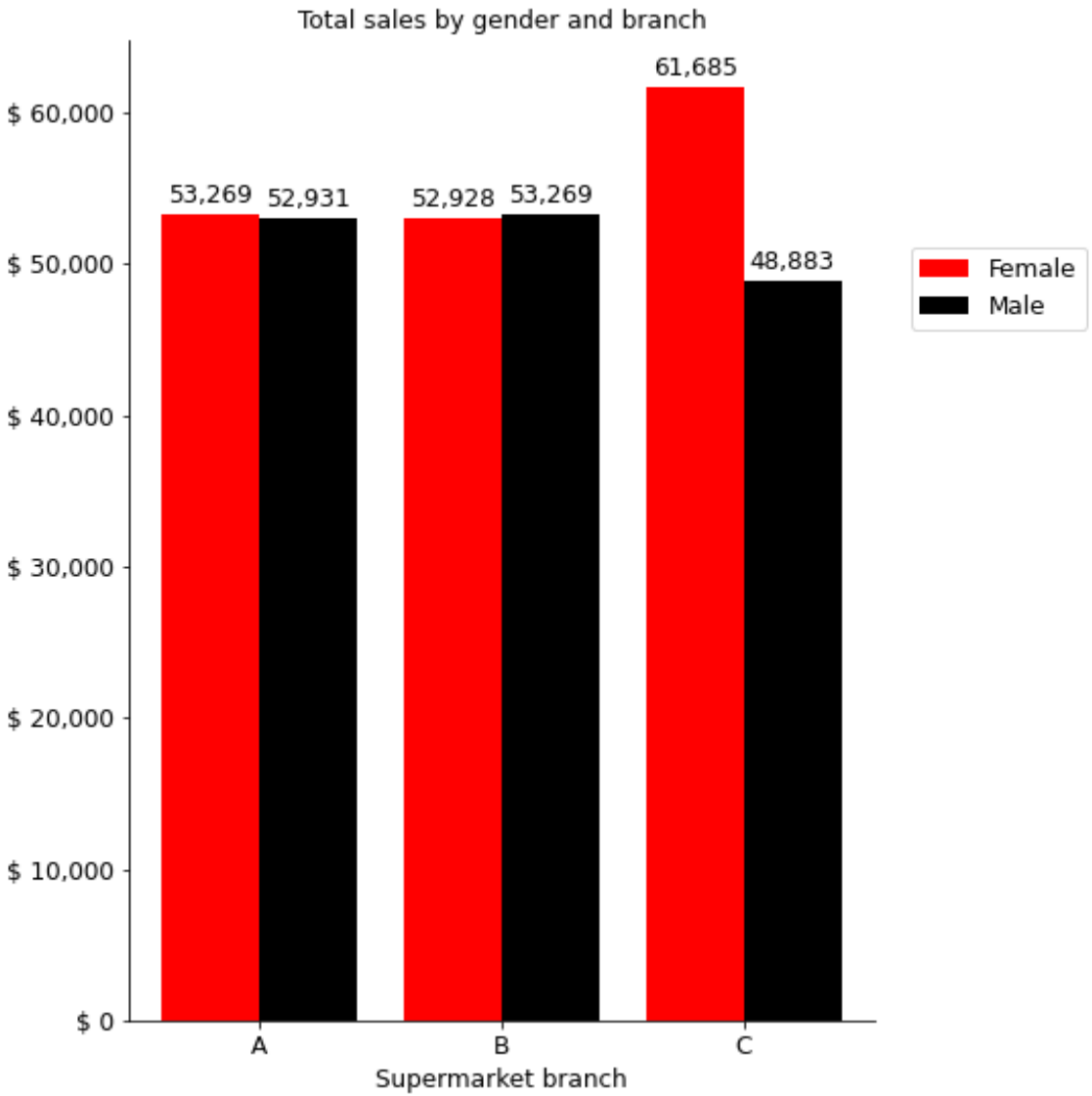
- The data consist of details of sales volumes at three supermarkets based in the USA
- Included in the data are variables such as the customer’s gender, whether they were a member of the supermarket branch, what product line they bought from and how they paid for their purchase
- This project aims to segment the customer data and build a regression model which will predict supermarket sales data
- The main findings in the data have been presented in graphical format and the data analysis has been carried out in Python
- The models used in the analysis are a multiple linear regression and polynomial regression

How do sales differ by members and non-members?



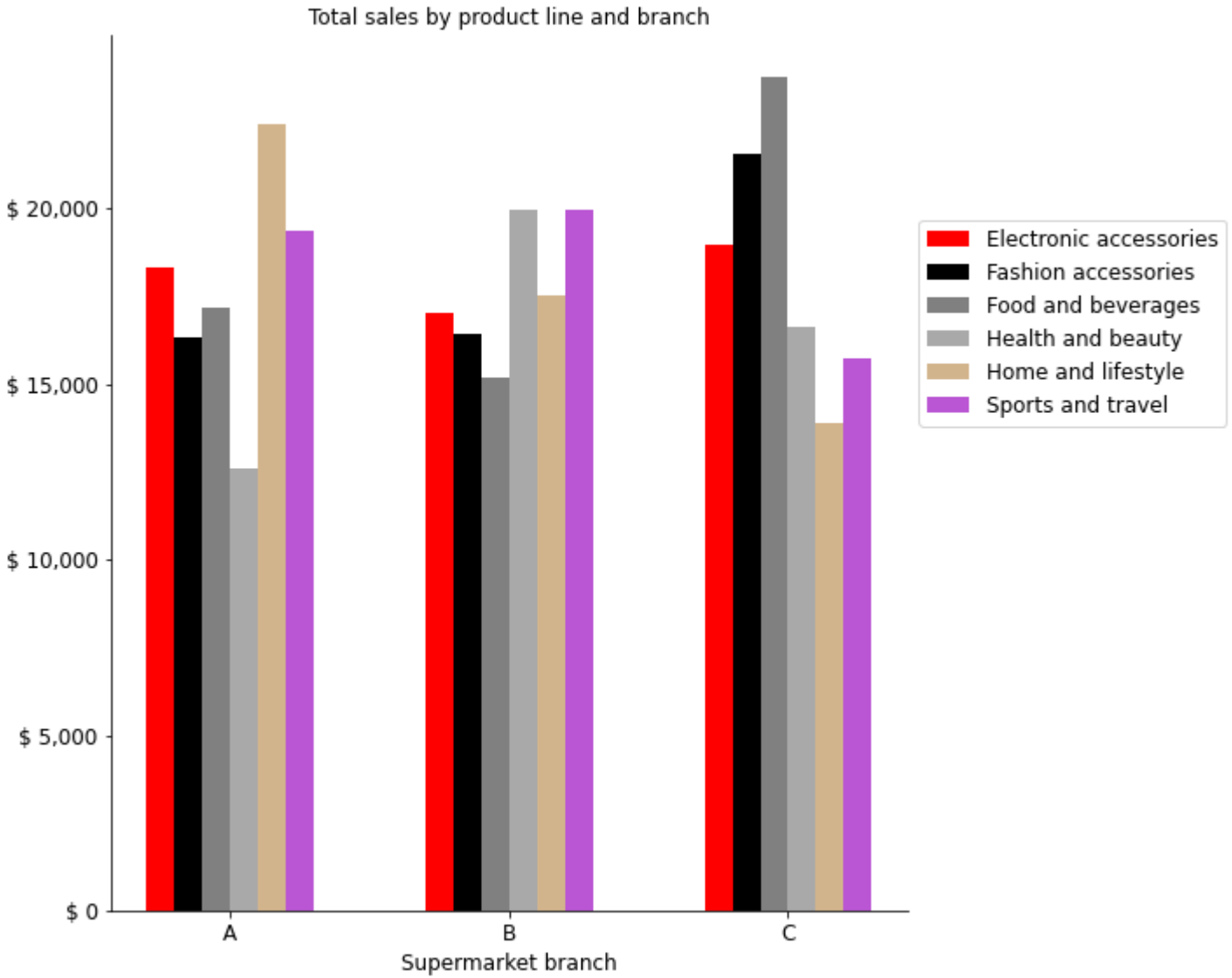
- Members spend more in each store than non-members although numbers are fairly similar for branches A and B
- Branch C has experienced the highest sales volume over the period and also has the largest difference in total spend by members and non-members
- Members spent **6%** more at branch C compared to non-members
- Given that members spend more in all branches, it would be worth trying to recruit more members by undertaking **advertising campaigns**

How does sales differ by gender?



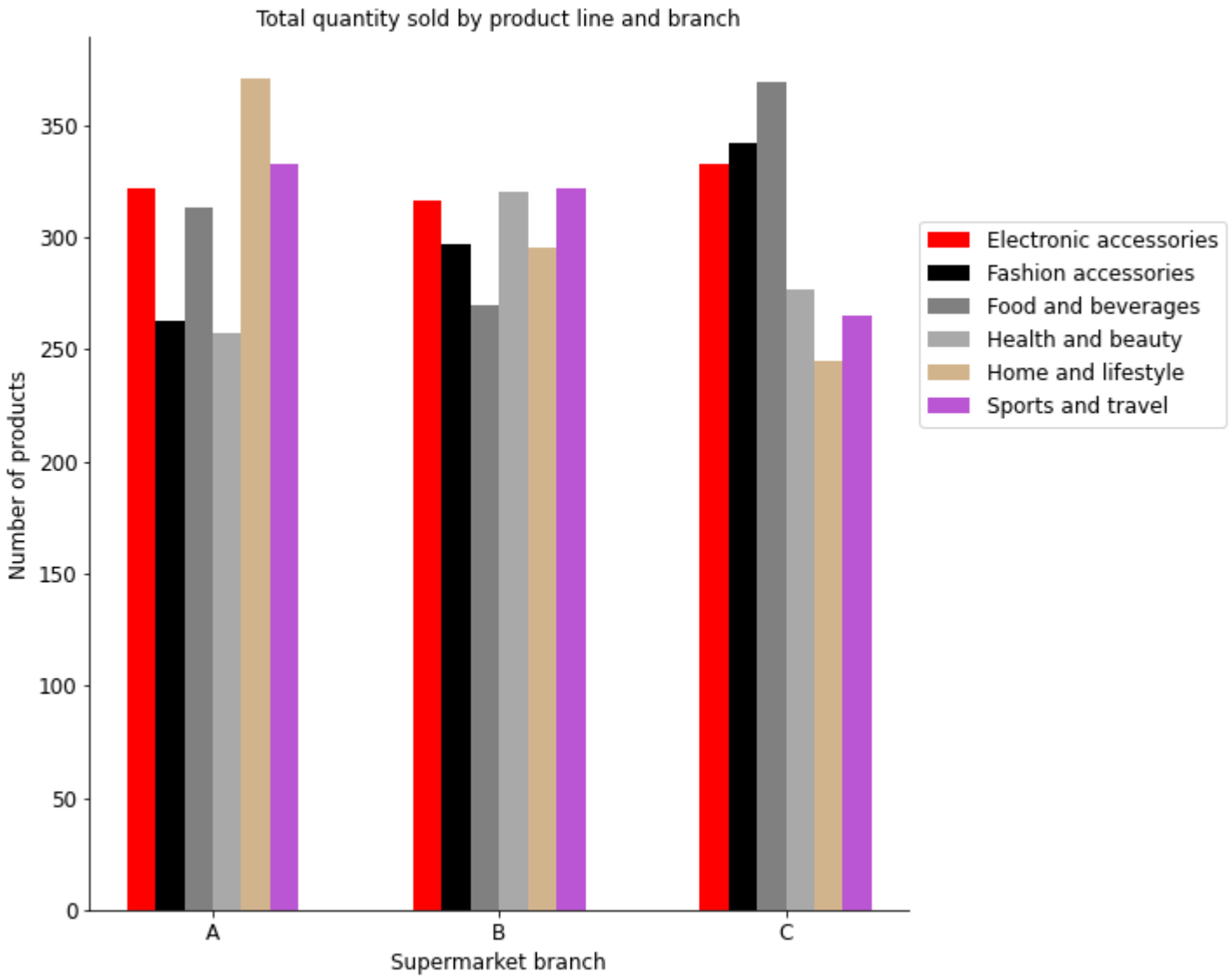
- Male and female shopping spend is roughly equal in branches A and B
- However females spend far more at branch C
- Of the total sales at branch C, **56%** are generated by females
- It is difficult to determine why this might be without knowing the population demographics of the area – if the female population is larger in the branch C location, then these figures could make sense

How do sales differ by product line?



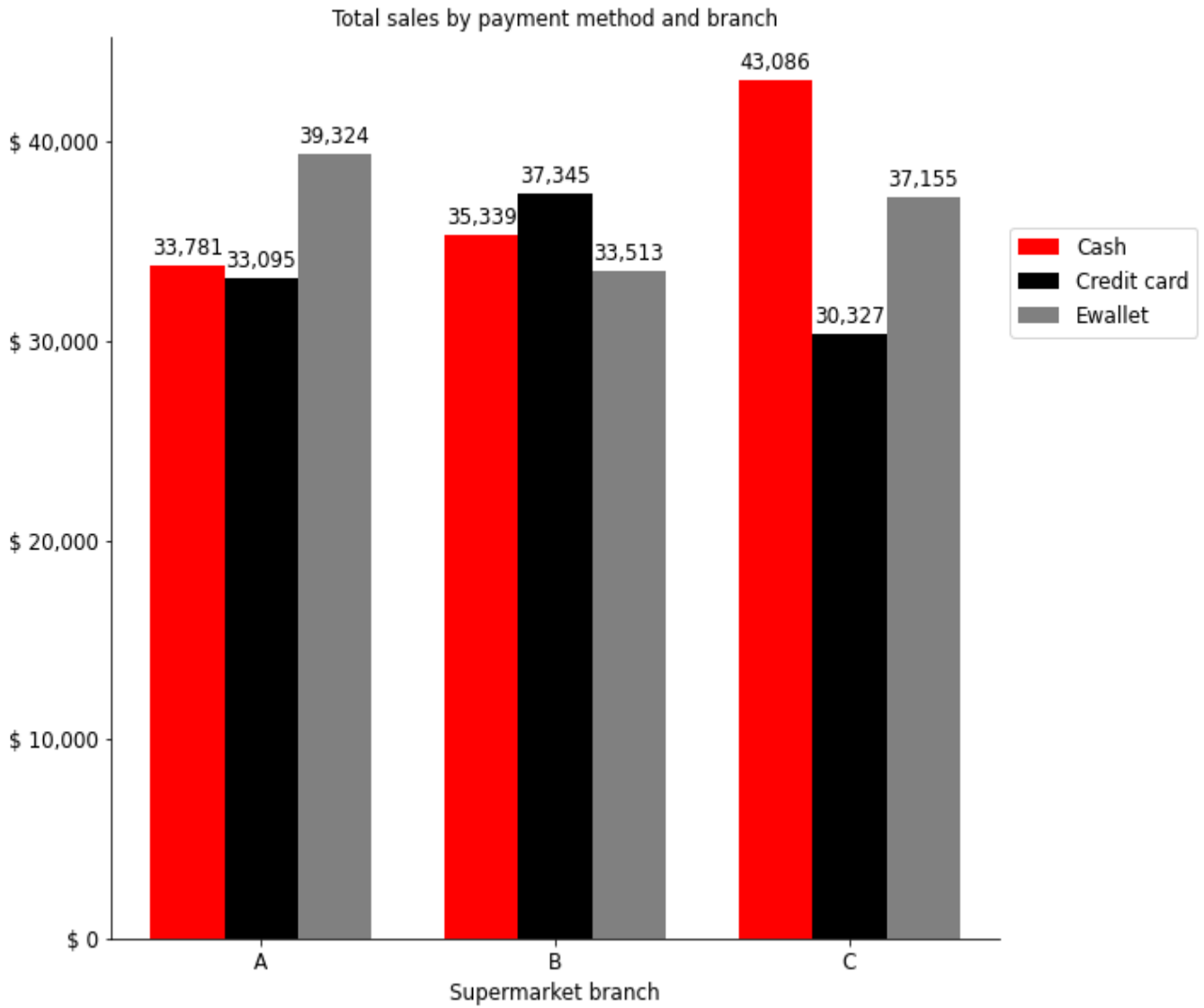
- Home and lifestyle goods generate the most sales at branch A, followed by sports and travel goods
- At branch B, health and beauty and sports and travel related goods generate the most sales
- Food and beverages generate the most sales at branch C which is what one might expect at a supermarket. Fashion accessories follow closely behind
- Sales volumes for food and beverages at branch C are **38%** higher than at branch A and **56%** higher than at branch B

Do the sales quantity data match the sales volume data?



- The sales pattern does appear to be similar – the home and lifestyle department sells the highest quantity of products at branch A
- At branch B, there's a slight difference – electronic accessories sell in similar quantities to health and beauty and sports and travel but generate less revenue, hence indicating these products must be cheaper
- The food and beverage department still dominates at branch C, followed closely by fashion accessories

How do customers prefer to pay?

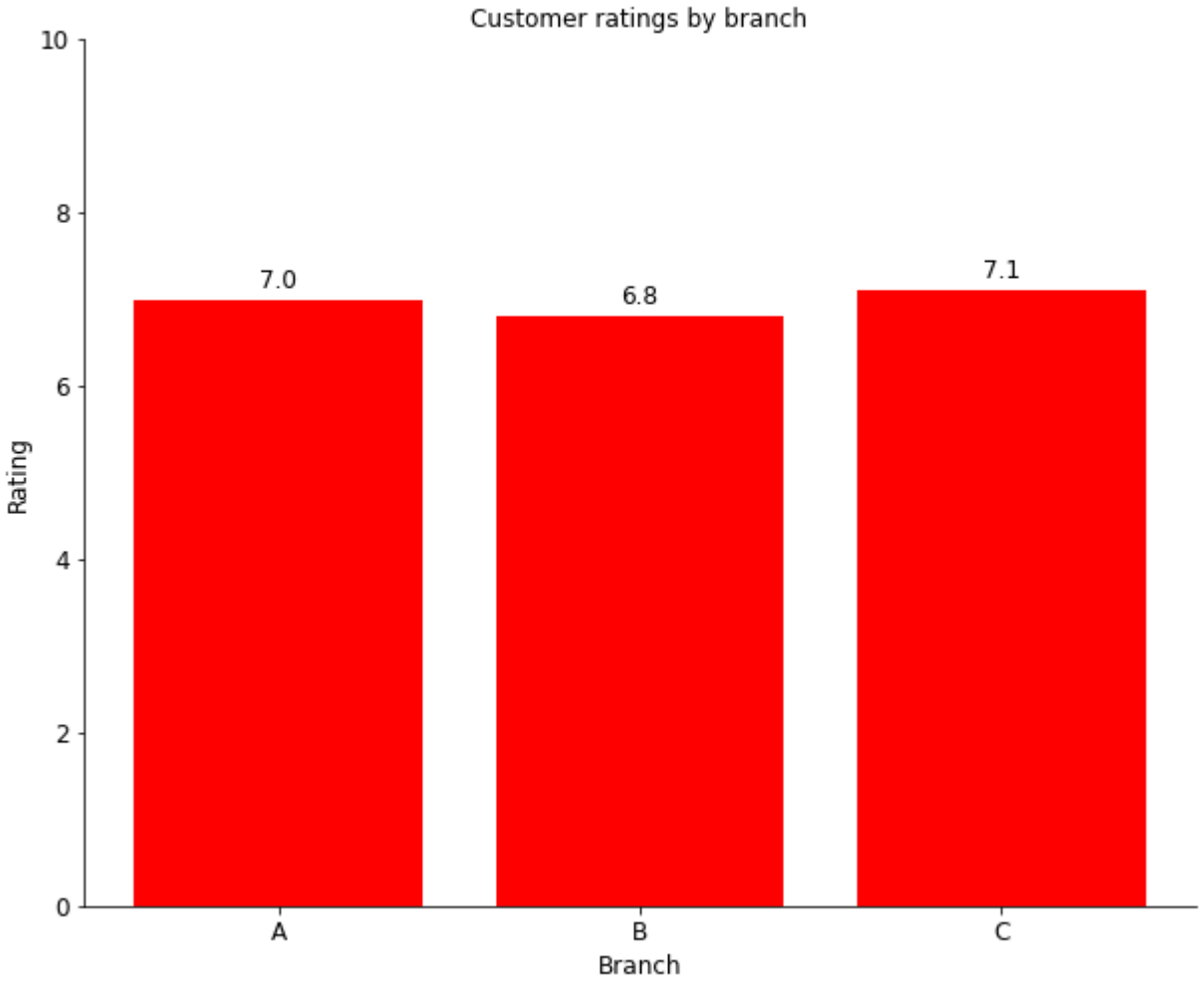


- The majority of sales are generated by customers paying by Ewallet at branch A – this payment method generates **16%** more sales compared to cash and **19%** more than credit card for branch A

- Sales volumes are similar across all payment methods at branch B although credit card has the edge

- Cash is by far the most popular payment method at branch C – **39%** of sales volumes are generated this way – perhaps customers at this branch are of an older generation who tend to rely on cash more heavily

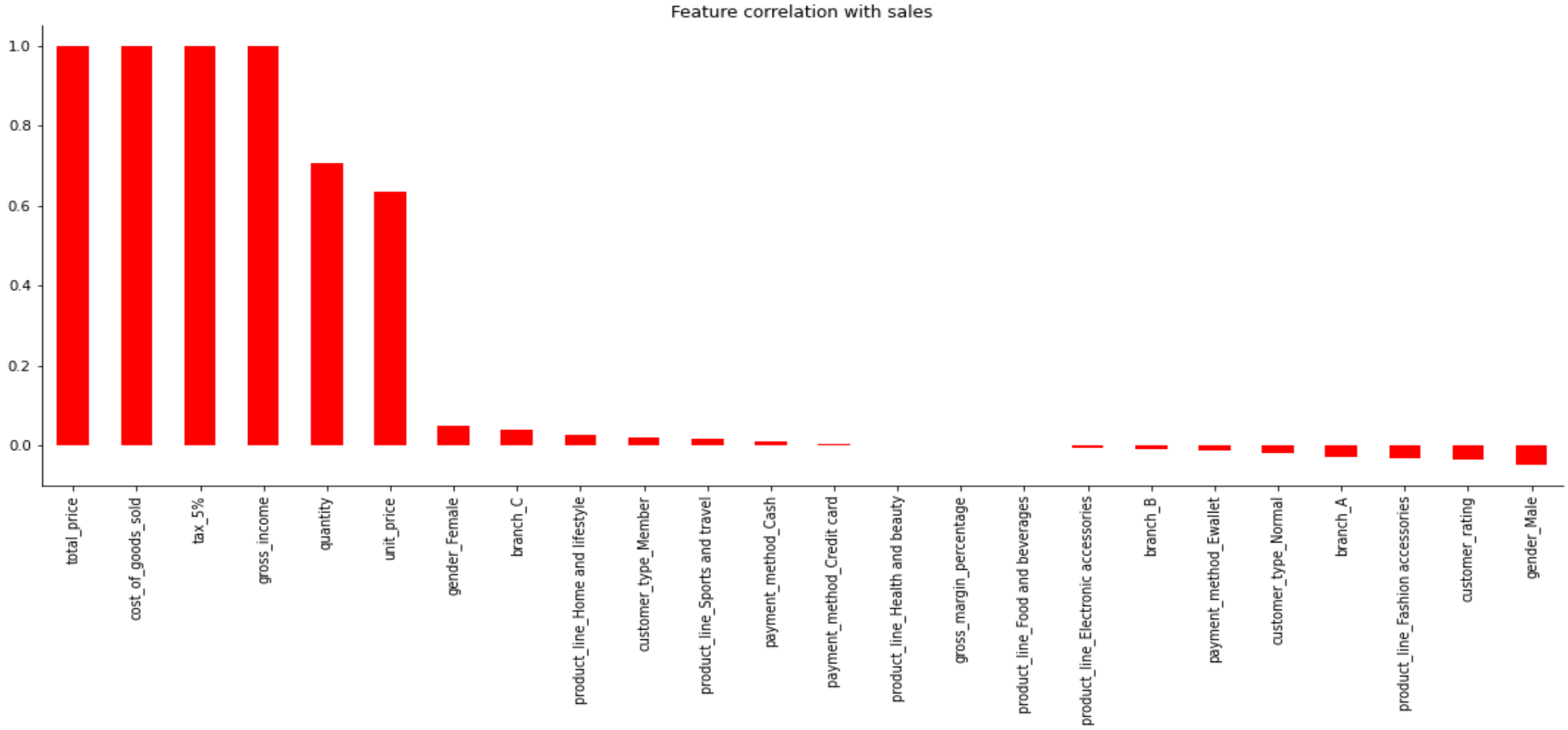
How do customers rate each supermarket?



- Customers are able to rate each supermarket between 1 and 10 in a customer satisfaction survey
- The average ratings are pretty similar across all three branches although slightly higher for branch C
- One might expect the rating for branch C to be a lot higher given that it generates such high sales volumes compared to the other two – **this might be something that management at branch C should work on**

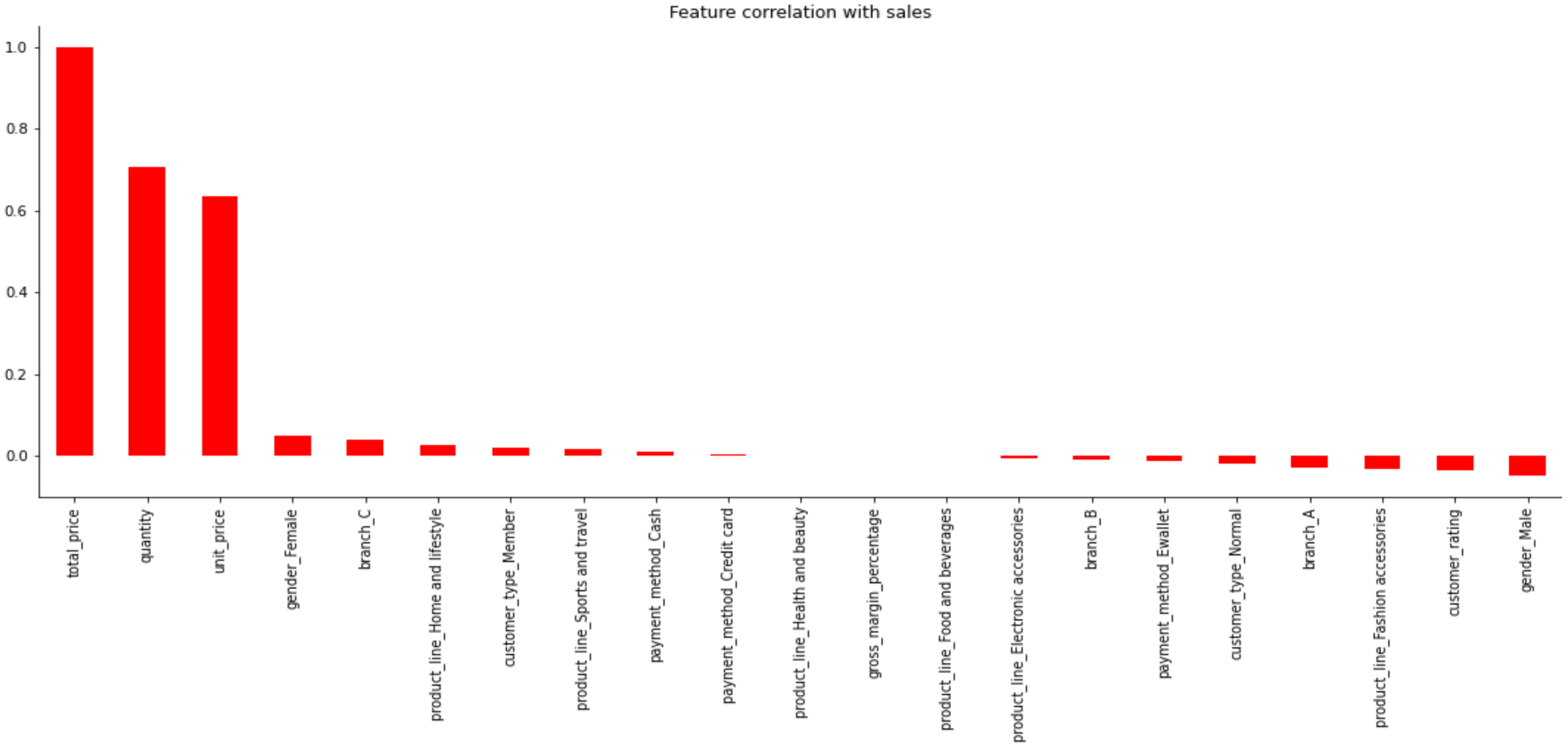
How do all features of the dataset correlate with sales?

- Three features are showing multicollinearity with sales (total price), i.e. they have a perfect positive correlation with sales and can therefore predict sales. These three features will be removed in the next iteration of the correlation analysis so they don't interfere with the model



How do all features of the dataset correlate with sales (2)?

- The quantity bought and unit price are most positively correlated with sales whereas male gender and customer rating are least correlated



Fitting and evaluating the models

Multiple linear regression:

R^2: 0.902
Root mean squared error: 0.077
Mean absolute error: 0.057

Model evaluation:

R^2 is relatively high at 0.902 (circa 90% of the variability in the data is explained by the model). Error vales are also very low – the model is a good fit

Polynomial regression:

R^2: 1.000
Root mean squared error: 0.000
Mean absolute error: 0.000

Model evaluation:

R^2 is 1.0 and there are zero errors - this is either a perfect model or the model has over-fit the data. Before concluding this is the best model, more work should be done on model tuning and perhaps also changing the train and test split (an 80:20 split was used here)

Conclusion

- Branch C generates the most sales and members are bigger spenders than non-members. Advertising campaigns can help to generate more members
- Females are by far the biggest spenders in branch C although the gender ratio is fairly even at branches A and B. The larger proportion of female shoppers at branch C could be related to the population demographics of the area
- There is a difference between product lines at the different branches with no two branches generating their highest sales volume from the same product line
- Preferred payment methods differ across the three branches with cash being most popular at branch C – one reason for this could be because older people shop at branch C and they tend to be more reliant on cash
- Customer ratings were similar across all branches, even branch C where most sales are generated. Management should try and increase the satisfaction ratings at this store in particular given that it generates so much revenue
- The multiple linear regression and polynomial regression models both had high R^2 scores and a low root mean squared error and mean absolute error. However, given that the polynomial regression generated a perfect score, it may have over-fit the data
- The analysis could be improved by further model tuning, adjusting the training and testing splits or perhaps trying an alternative model, e.g. a different type of regression (Ridge, Lasso) or Random Forest

Thank you

Contact details:

Amy Birdee

amybirdee@gmail.com