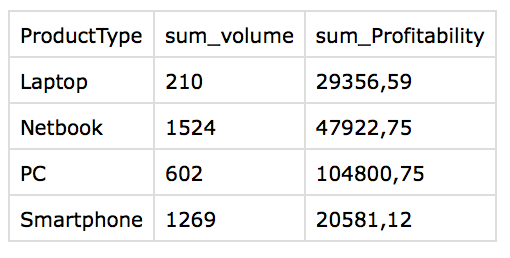
# Chapter 2 Task 3: Multiple Regression in R

**Background:**

Danielle Sherman has been consulted by the sales team regarding product sales in one of Blackwell’s stores. Our team has been asked to analyse historical sales data and then make sales volume predictions for a list of new product types, i.e. PC, Laptops, Netbooks and Smartphones; and try to understand how different product types may or may not affect sales volume predictions.

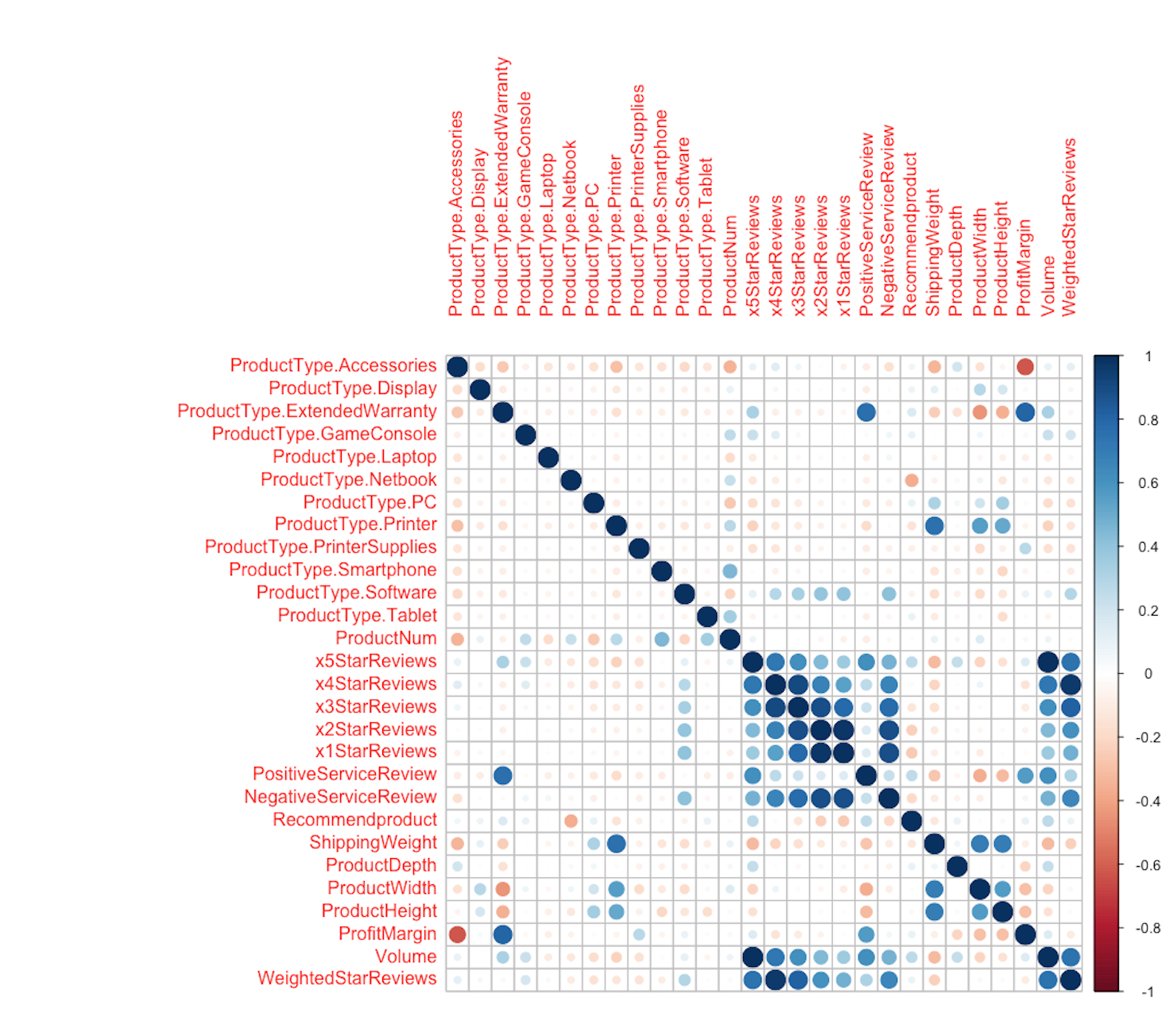
**Executive Summary:**

* Using correlation matrix analysis, we have found out that product types don’t have any effect on predicting sales volume, hence, using product types as attributes does not have any effect on final volume/sales predictions.
* The attributes that have the most predictive power have been identified, i.e. star reviews and positive service reviews. Star reviews have been transformed into weighted star reviews, with weights determined by coefficients of a linear model. Positive service reviews have been kept unchanged.
* Using Random Forest model, that had the best performance metrics and hence the highest predictive confidence, we have predicted sales volume for four product types outlined in the background section.



**Analysis:**

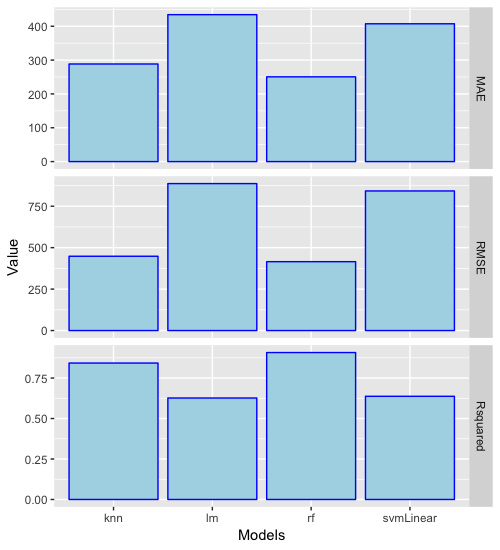
In order to understand which attributes have the most predictive power, we have constructed a correlation matrix and represented it using a heat map.



* As we can see, 1/2/3/4/5 Star reviews and Positive service reviews have the most effect on predicting volume.
* 5 Star reviews have a 1-1 correlation, which is unrealistic and hence introduces unwanted noise into our data. As a conclusion, 5 star reviews have been removed.
* In order to avoid multi-collinearity, we have decided to transform separate star reviews into one single attribute called “weighted star reviews”.

**Method:**

To make our predictions, we have tried a variety of different predictive models, namely SVM, k-NN, Random Forest and Linear model. By analysing respective performance metrics, we have come to a conclusion that Random Forest model fits our data best.



We select Random Forest because it has the highest R-squared value, while having the smallest RMSE and MAE.   
By first training and testing our model on the existing product attributes dataset to understand its performance, we have then used it to make sales volume predictions on the new product attributes dataset. The total number of sales and total profit per product type have been outlined in the executive summary section.

**Conclusions:**

* Our analysis has concluded that weighted star reviews are a much better representation of customer’s opinions on products than separate star reviews.
* Product types do not have any effect on predicting sales volume, but certain product types are more likely to bring more profit to Blackwell.

