|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | logo_black.gif | | |
|  | 200925978 |  | 1000 |
|  | GEOG5917M Big Data and Consumer Analytics |  |  |
|  | Assignment 1 |
|  | Prof. Lex Comber |

|  |
| --- |
|  |
| 1. |
| 2. |
| 3. |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

**Q1: What is the household penetration of ‘PRIVATE LABEL THIN SPAGHETTI’? That is, out of all customers purchasing Pasta, what percent purchase this brand? (10 marks)**

After installing and loading the necessary packages and reading in the dunnhumby data, converting it to tibble format in the process, the first step is to find the index of ‘Private Label Thin Spaghetti’ (PLT Spaghetti) and ‘pasta’ within the product lookup data (‘pl’). This is carried out using the ‘which’ function which returns the positions of the requested parameter (see Fig.1).

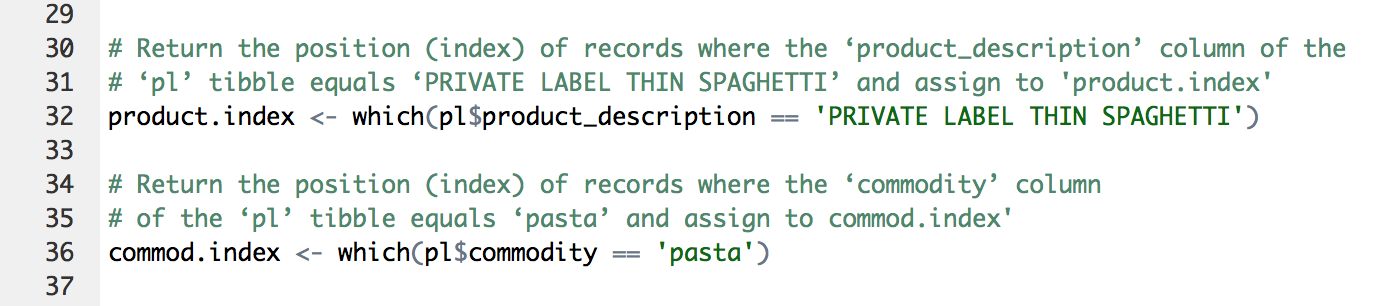


Figure : Product/Commodity Index Lookup

Based on the indexes returned, the Universal Product Code (UPC) for the relevant records can be obtained using the unlist function, which takes the list of codes and simplifies it to a single vector containing the list’s contents (see Fig.2).

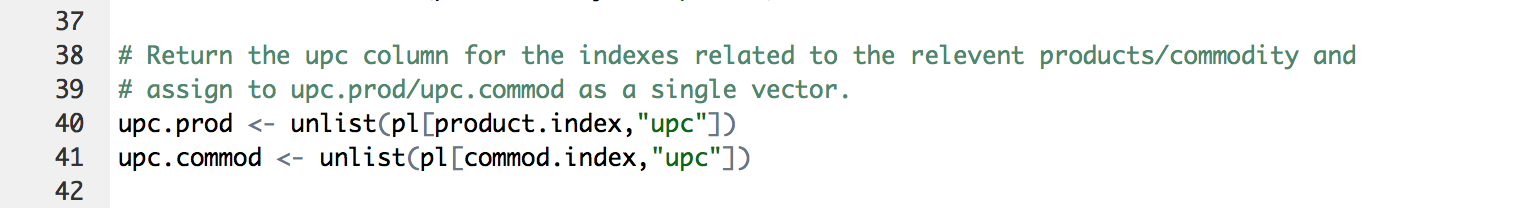


Figure : Product/Commodity UPC Lookup

Using the filter function, which returns records that match specified conditions, purchases of the product/commodity can be identified within the transaction data by searching for records where the UPC matches those identified as being associated with PLT Spaghetti and pasta (see Fig.3). With ‘NA’ records excluded we are left with a dataset of all purchases of PLT Spaghetti (tmp1) and all purchases of pasta (tmp2).

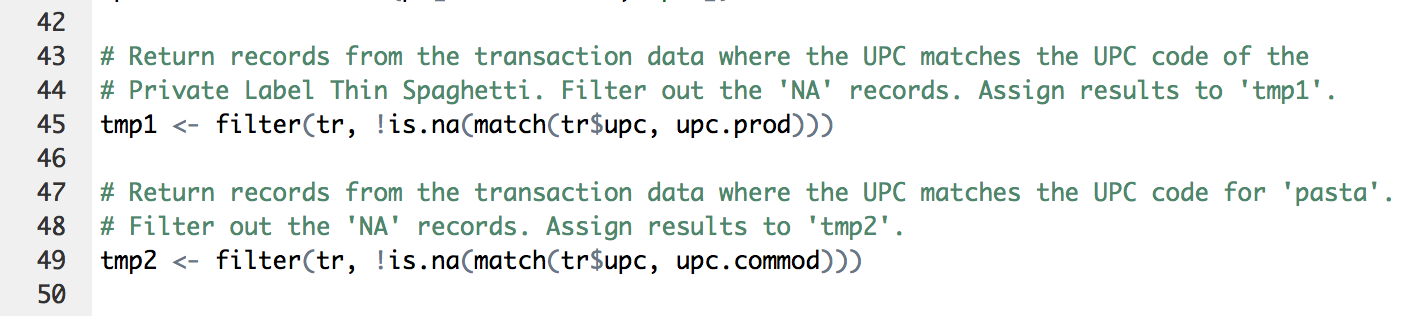
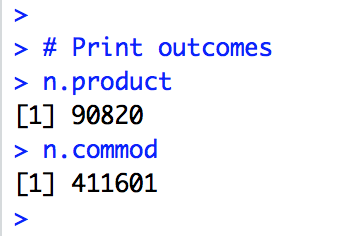
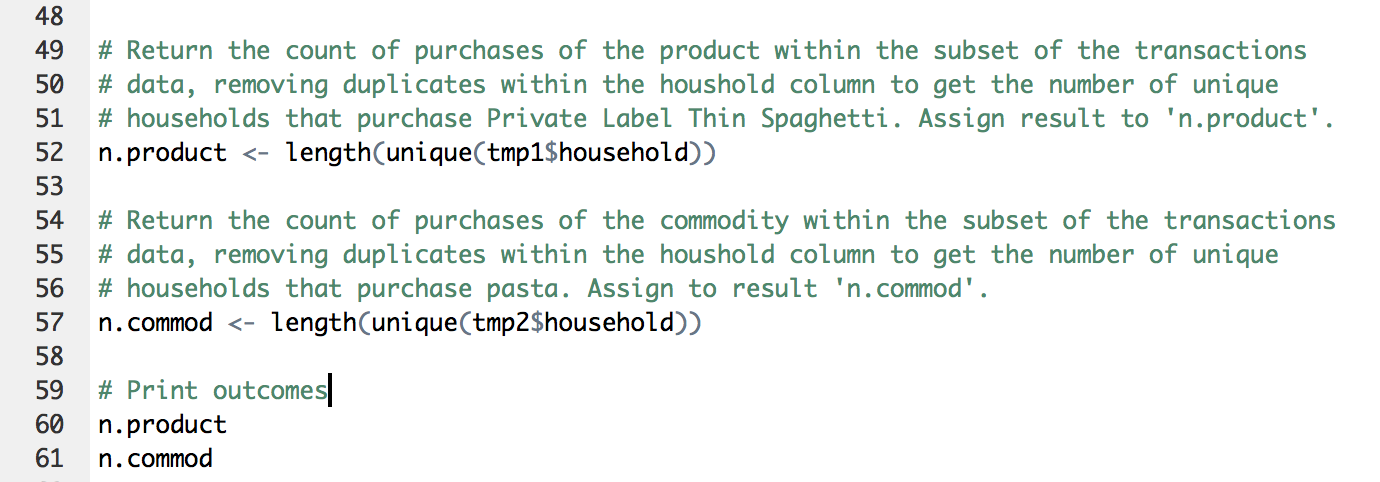


Figure : Product/Commodity Transactions Lookup

**The next step involves both the ‘length’ and the ‘unique’ function (see Fig.4). Length returns the count of purchases of either the product or the commodity whilst ‘unique’ removes duplicates, in this case of the household identifier, meaning that each household that has purchased either the product or commodity is only counted once.

*Figure 4: Households purchasing the Product/Commodity*

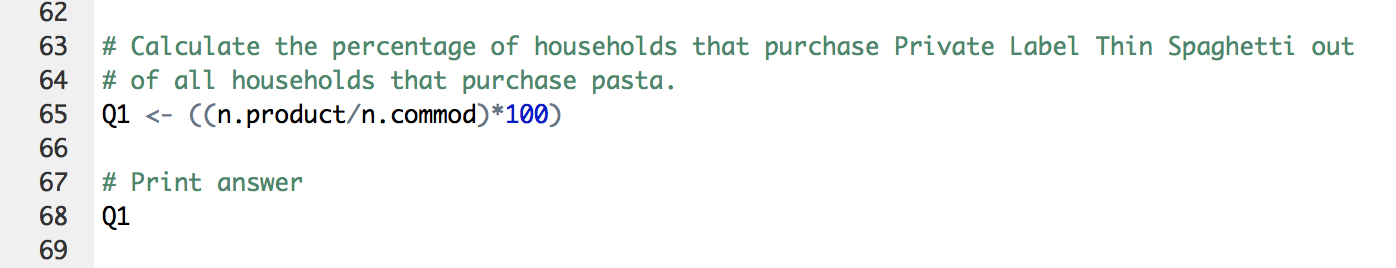
Having done this a simple equation can be carried out to calculate that **22.06%** of households that buy pasta, purchase Private Label Thin Spaghetti:

Figure : Calculating Household Penetration of PLT Spaghetti

**Q2: How does the household penetration of the product PRIVATE LABEL THIN SPAGHETTI vary within the two regions, relative to the sales of Pasta in each region? (20 marks)**

In order to calculate household penetration for each region, the subsets of transaction data that have already been filtered to include only the relevant product, are filtered by geography to create two further subsets of the data, one for each region. A test is in place to ensure that the subsets include all of the data (see Fig.6).

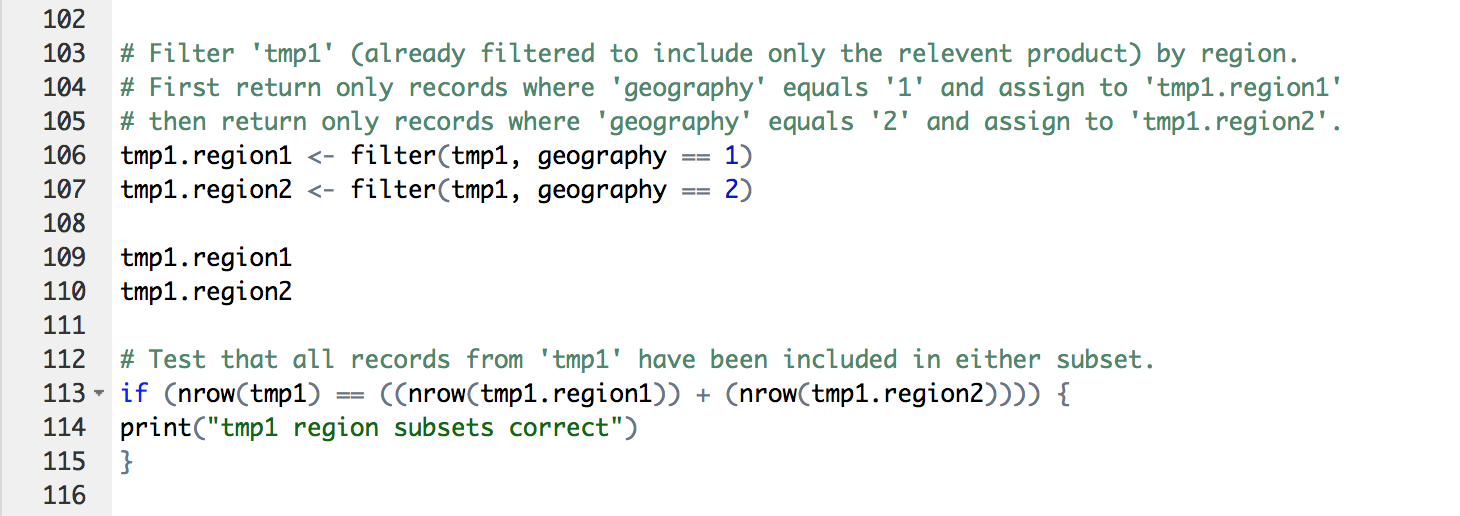


Figure : Filter Transaction Data by Region

As in Q1 we can then use the ‘unique’ function to establish the number of unique households that occur in each subset i.e. how many households in each region purchased Private Label Thin Spaghetti.

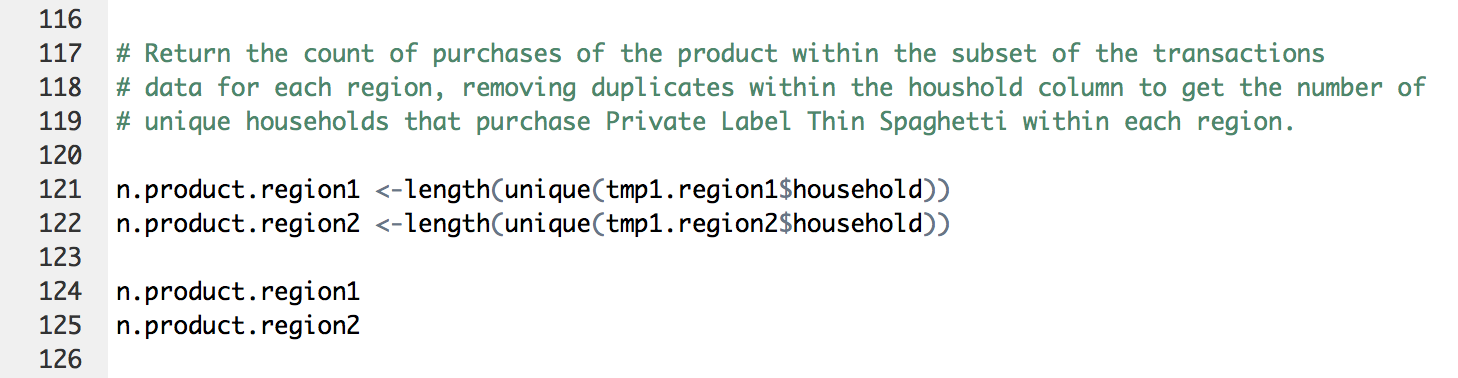


Figure : Households Purchasing PLT Spaghetti by Region

This process can then be repeated (see Fig.8), working this time with the transaction data filtered to include all pasta purchases rather than just the specific product. This gives us the total number of households within each region that purchase any kind of pasta.

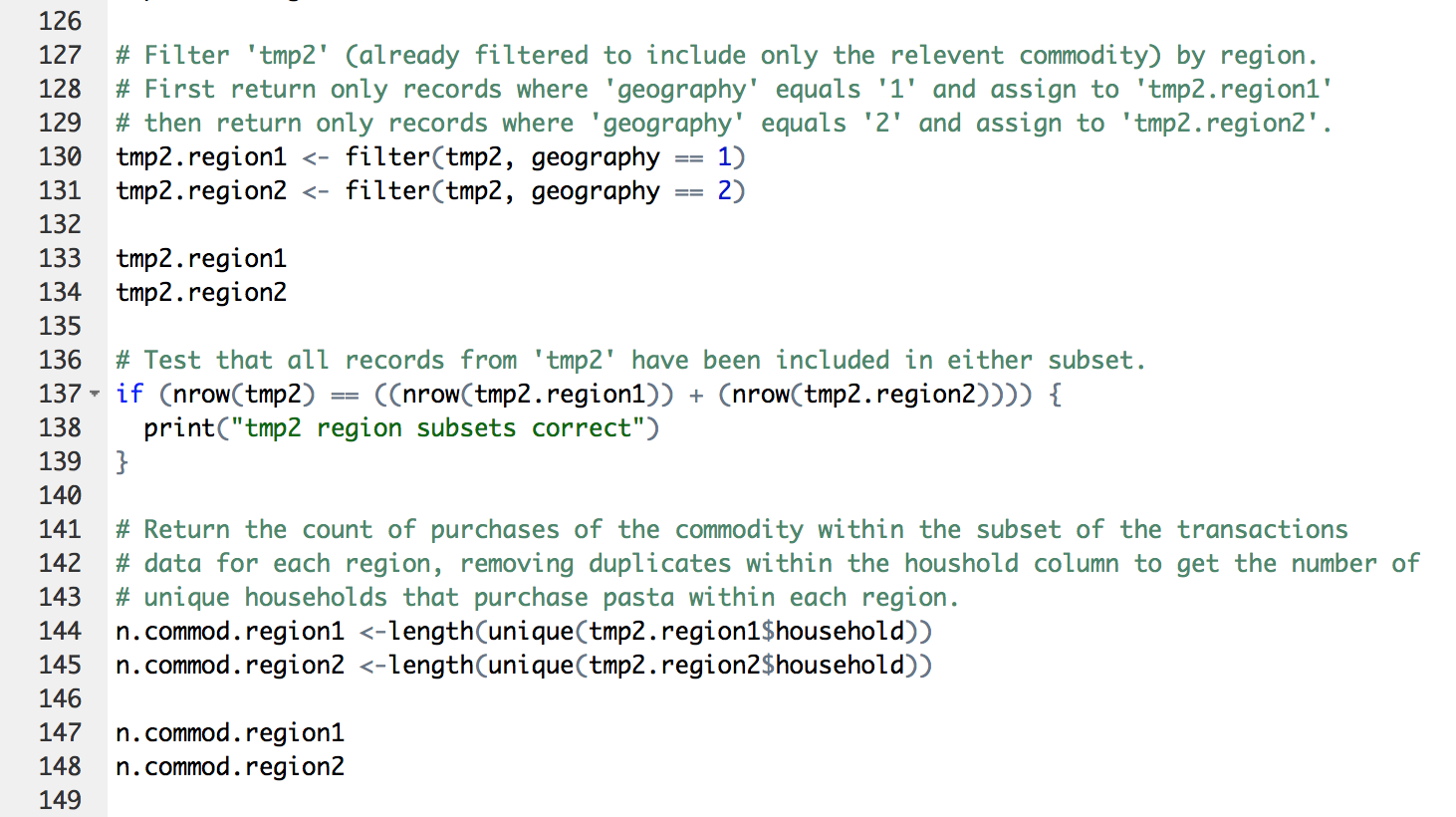


Figure : Households Purchasing Pasta by Region

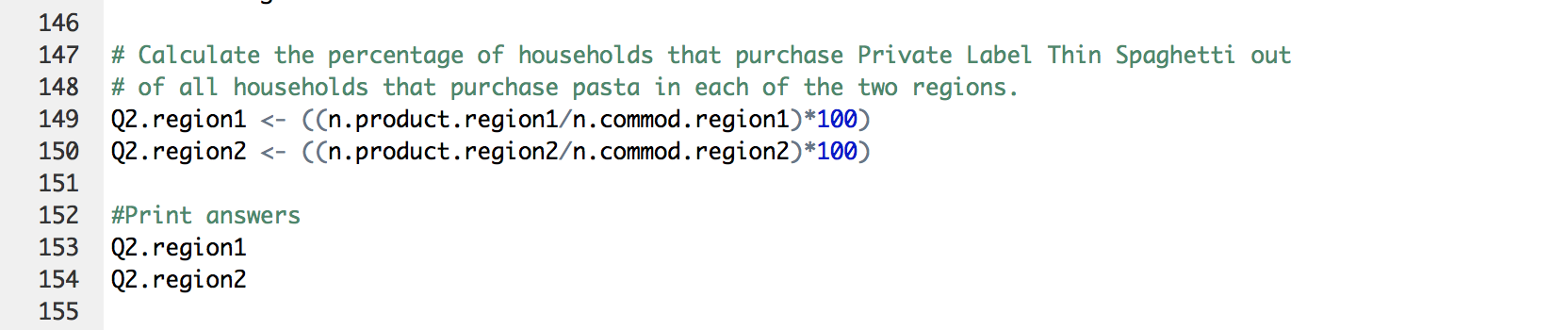
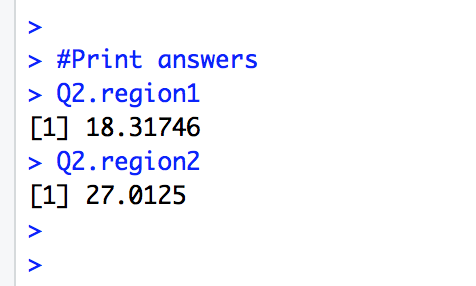
The final step involves a simple calculation (see Fig.9) to work out the percentage of those households that purchase pasta, who purchase Private Label Thin Label Spaghetti in each region.

Figure : Calculating Household Penetration of PLT Spaghetti in each Region

The outcome of this analysis (summarised in Tab.1) shows that the household penetration of PLT Spaghetti is greater in Region 2, with 27% of households who purchase pasta, purchasing PLT Spaghetti compared to 18% in Region 1.

Table : Household Penetration of Private Label Thin Spaghetti by Region

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total number of households who purchase pasta | Number of households who purchase Private Label Thin Label Spaghetti | Household penetration of Private Label Thin Label Spaghetti |
| Region 1 | 238925 | 43765 | **18.32%** |
| Region 2 | 174497 | 47136 | **27.01%** |

In order to visualise this, a function was defined using ggplot2 (see Fig.10), to produce two pie charts (see Fig.11 & 12). The input can be set to the data-frame containing figures for either Region 1 or 2 in order to plot a bar chart, which is then converted to a pie chart with appropriate titles and labels.

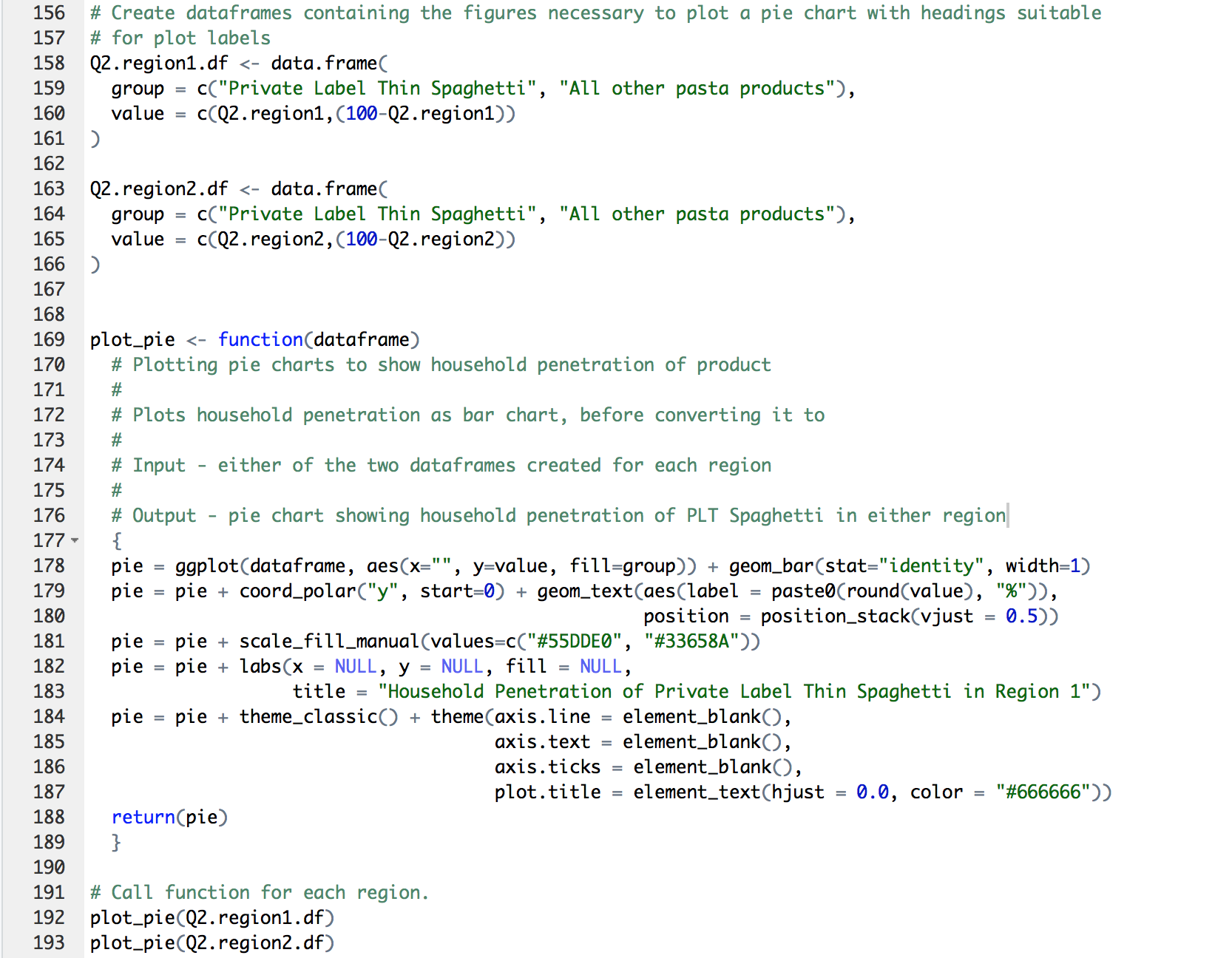


Figure 10: Plotting Pie Charts

****

Figure 11: Household Penetration of PLT Spaghetti in Region 1

****

Figure 12: Household Penetration of PLT Spaghetti in Region 2

**Q3: For which category of products (pasta, pasta sauces, syrups and pancake mixes) does the provision of coupons appear to have a positive impact? For this you should consider which customers first purchased an item in a category using a coupon, and then how many of these customers made additional purchases of the item in the category. (30 marks)**

As the same code will need to be run for each of the 4 commodities being studied, a function has been defined to allow for easy repetition. The input of the ‘Q3.function’ function is set as ‘commod’ so that the function can be called for each commodity (see Fig.13).



Figure : Calling the Function

Given the large size of the datasets being used, a subset of the transaction data, containing the first 10,000 rows, was created (see Fig.14). This allowed ‘mini.tr’ to be used in the space of ‘tr’ within the function in order to save time when running tests and checks as the processing time is reduced.



Figure : Subsetting the Transaction Data

The function first links the product data with the transaction data. This allows us to filter the transaction data by commodity as well as by the presence of a coupon. By extracting the household column of the filtered data, using the ‘unique’ function we are left with a list of households who have purchased the commodity in question using a coupon.

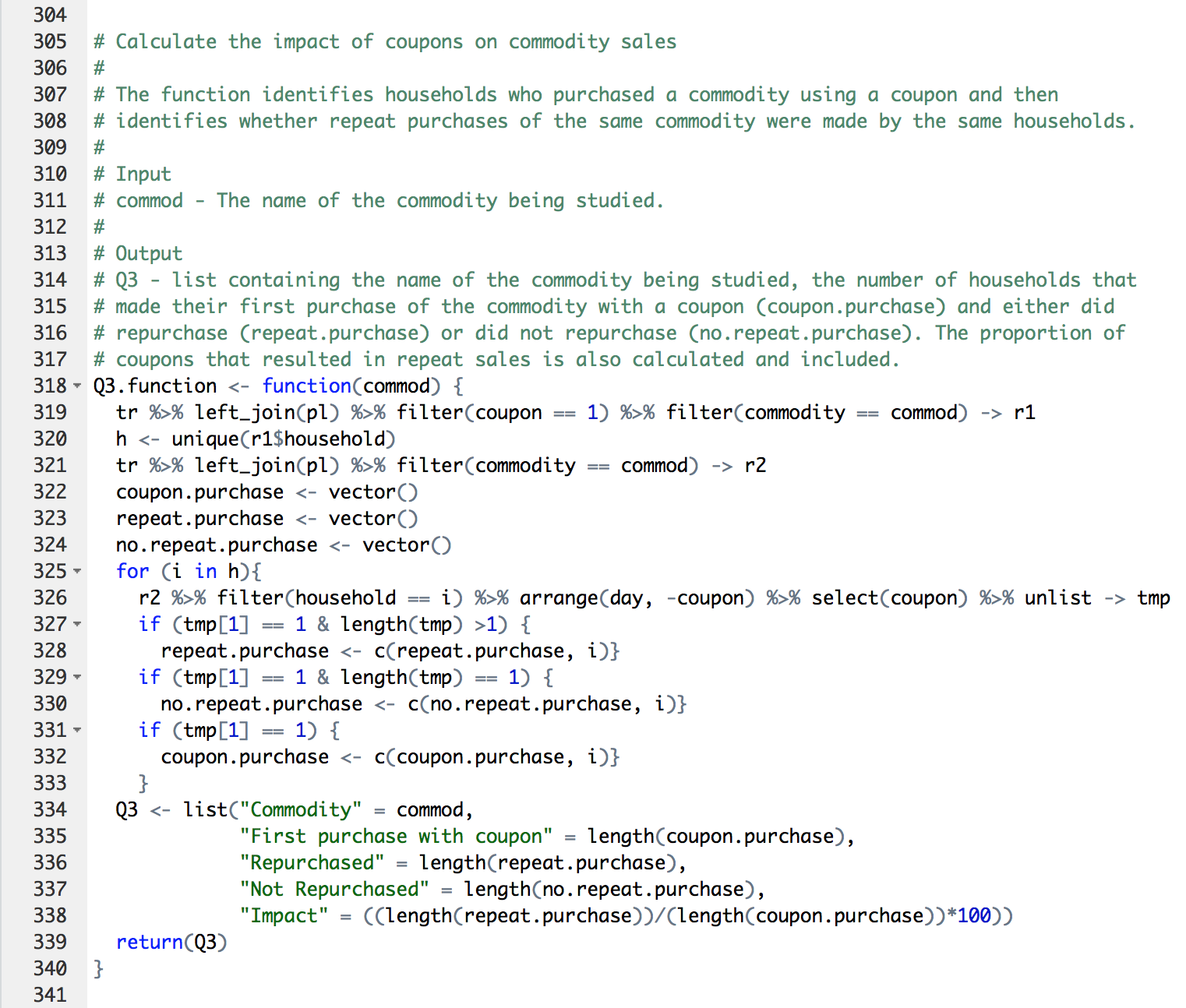


Figure : Households that have Purchased Commodity using Coupon

After linking the transaction and product again, to produce a complete copy of all commodity transactions, we are able to examine each household within the list created above (‘h’) by creating a ‘for’ loop. This iterates through the transaction data, each time filtering it to include the data for only one of the household from ‘h’ before arranging it by day (i.e. first transaction is the first entry) and inversely by coupon (i.e. coupon purchases are first). The coupon column for each household is then assigned to a temporary variable as a single vector.

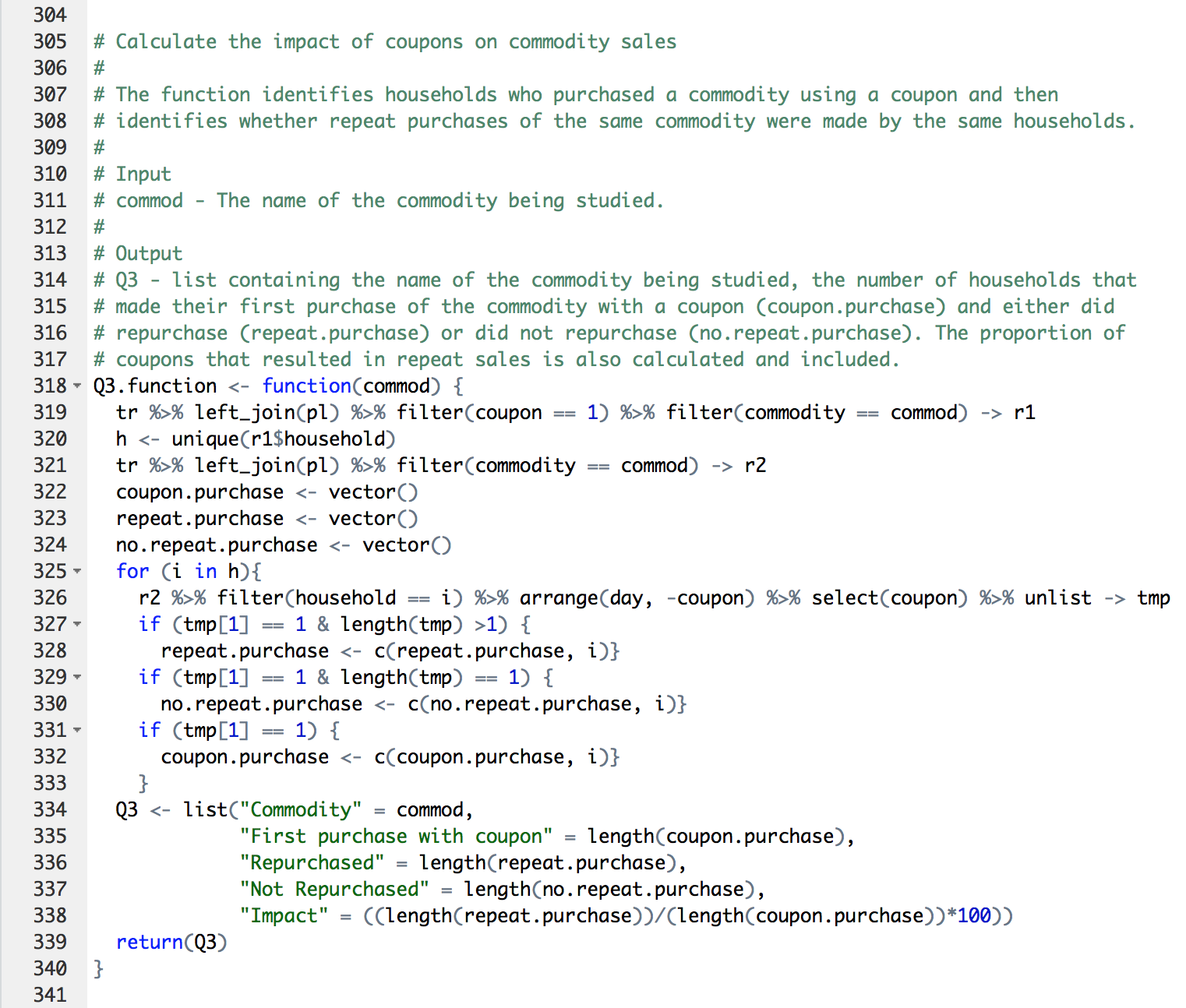


Figure : Iterating through Households

‘If’ statements then allow us to append the household to predefined lists; either ‘repeat.purchase’ if the first entry in the temporary variable equals 1 (i.e. the households first purchase was made using a coupon) and the length of the temporary variable is greater than 1 (i.e. the household made further purchases) or append the household to ‘no.repeat.purchase’ if the first entry equals 1 but the length also equals 1 (i.e. the households first and only purchase was made using a coupon). This process is repeated for all the unique households within ‘h’.

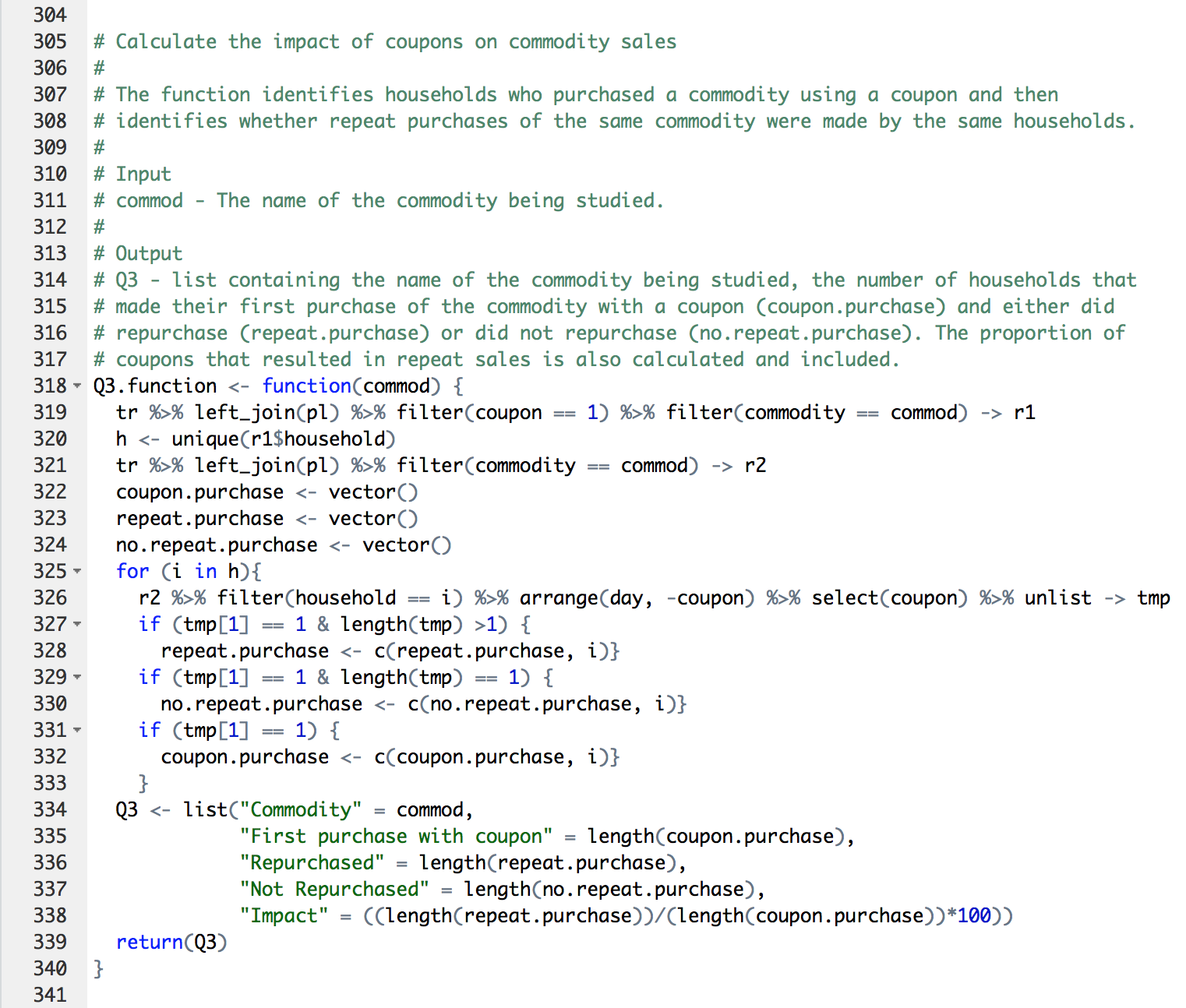


Figure : Appending Households based on IF statements

A third if statement records the total number of households that used a coupon to make their first purchase, regardless of any further purchases, allowing for the proportion of coupon sales that led to further purchases to be calculated.

As functions cannot return multiple elements, all the necessary figures and calculations, plus the commodity name, are appended to the ‘Q3’ list (see Fig.18) which is returned when the complete function (see Fig.19) is called.

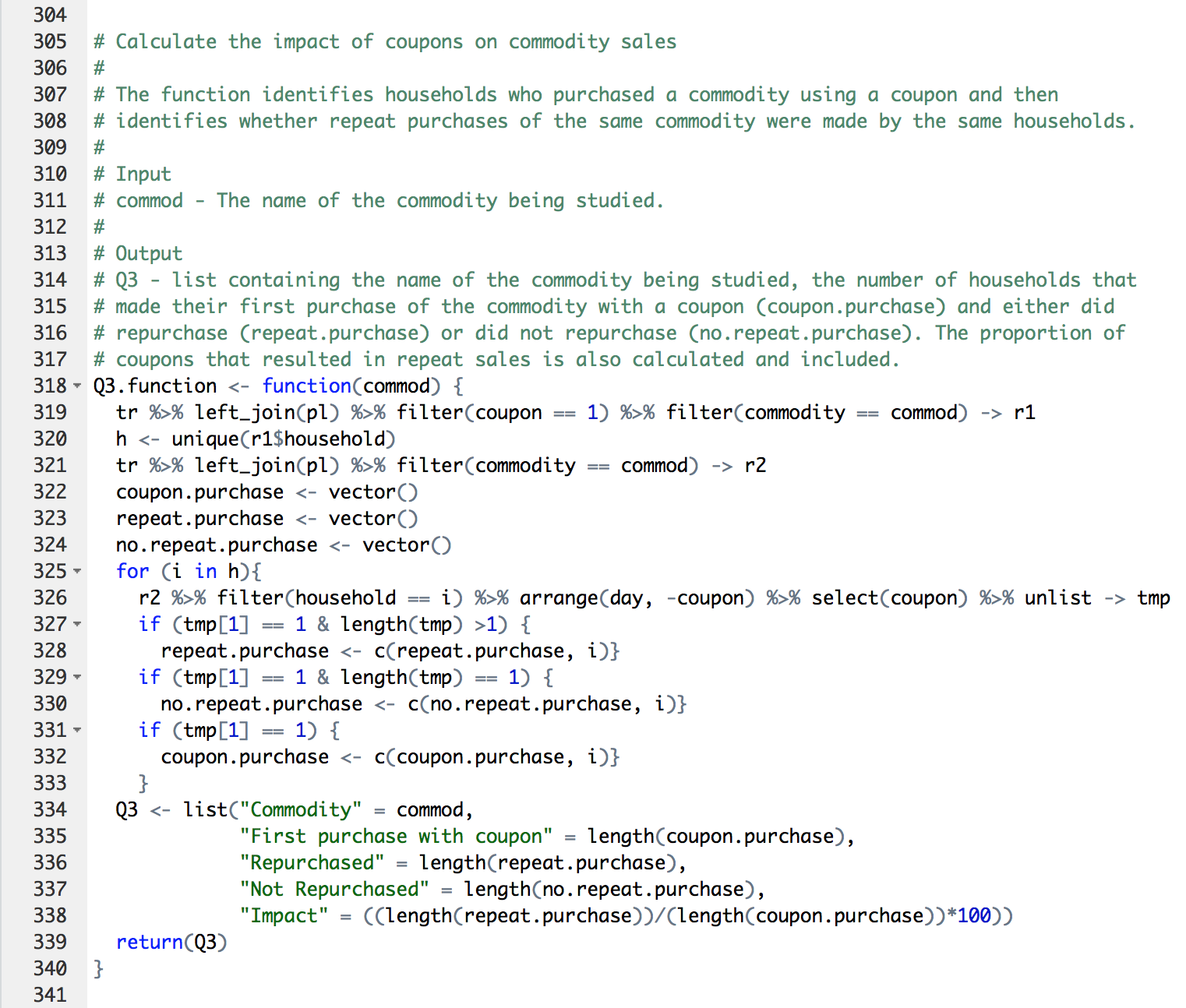


Figure : Defining and Calculating Function Output

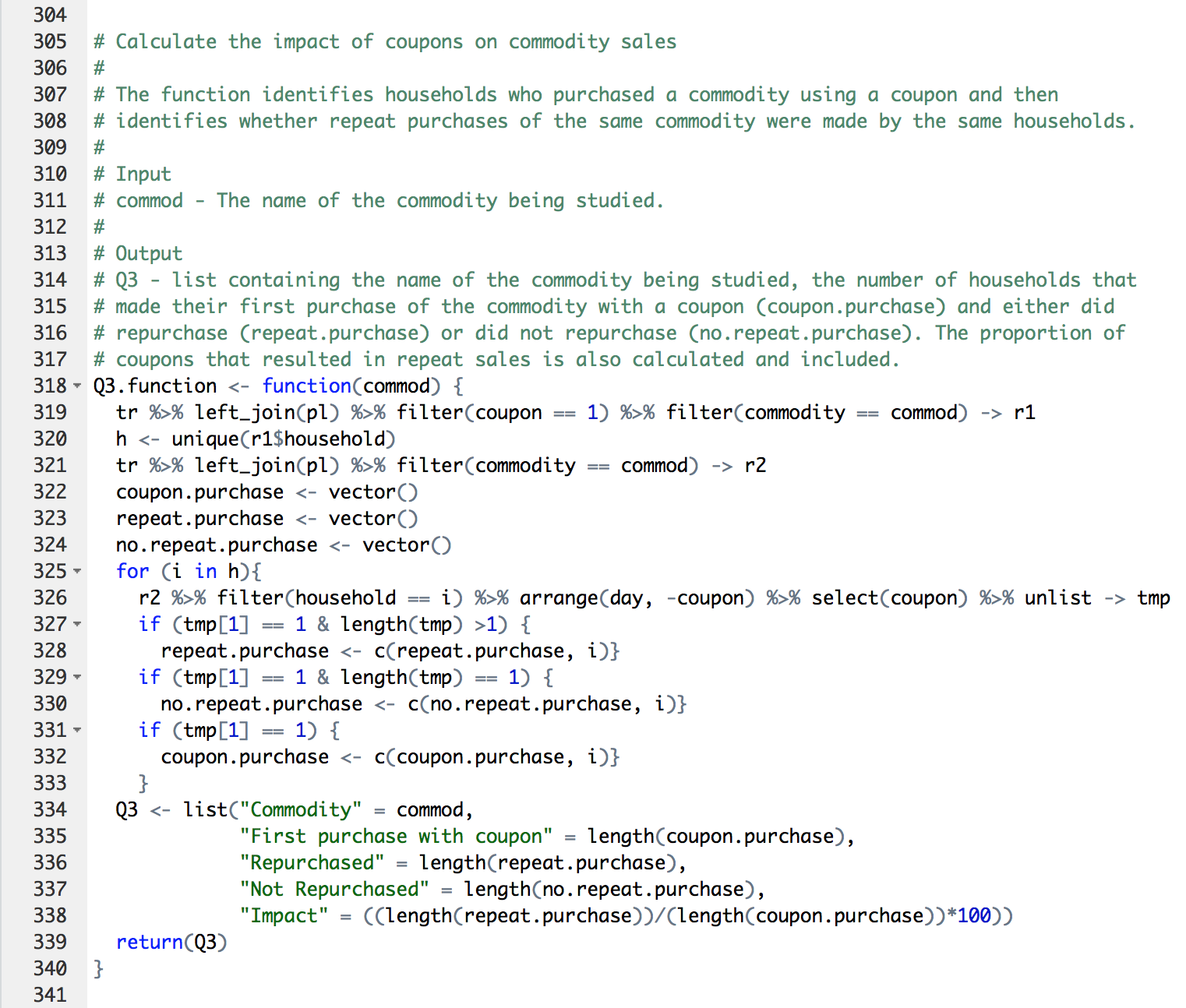


Figure : Complete 'Q3.function' Function

The outcomes of calling the function for the 4 commodities (see Fig.20 & Tab.2) reveal that the provision of coupons has a positive impact on the sale of all commodities with an average repeat purchase rate of 68%. The impact is greatest for ‘pasta’ with 85% of customers who first bought pasta using a coupon repurchasing the commodity, compared to 46% for pancake mixes, where the impact was the least notable.

Figure : Console Outputs on Calling Function for each Commodity

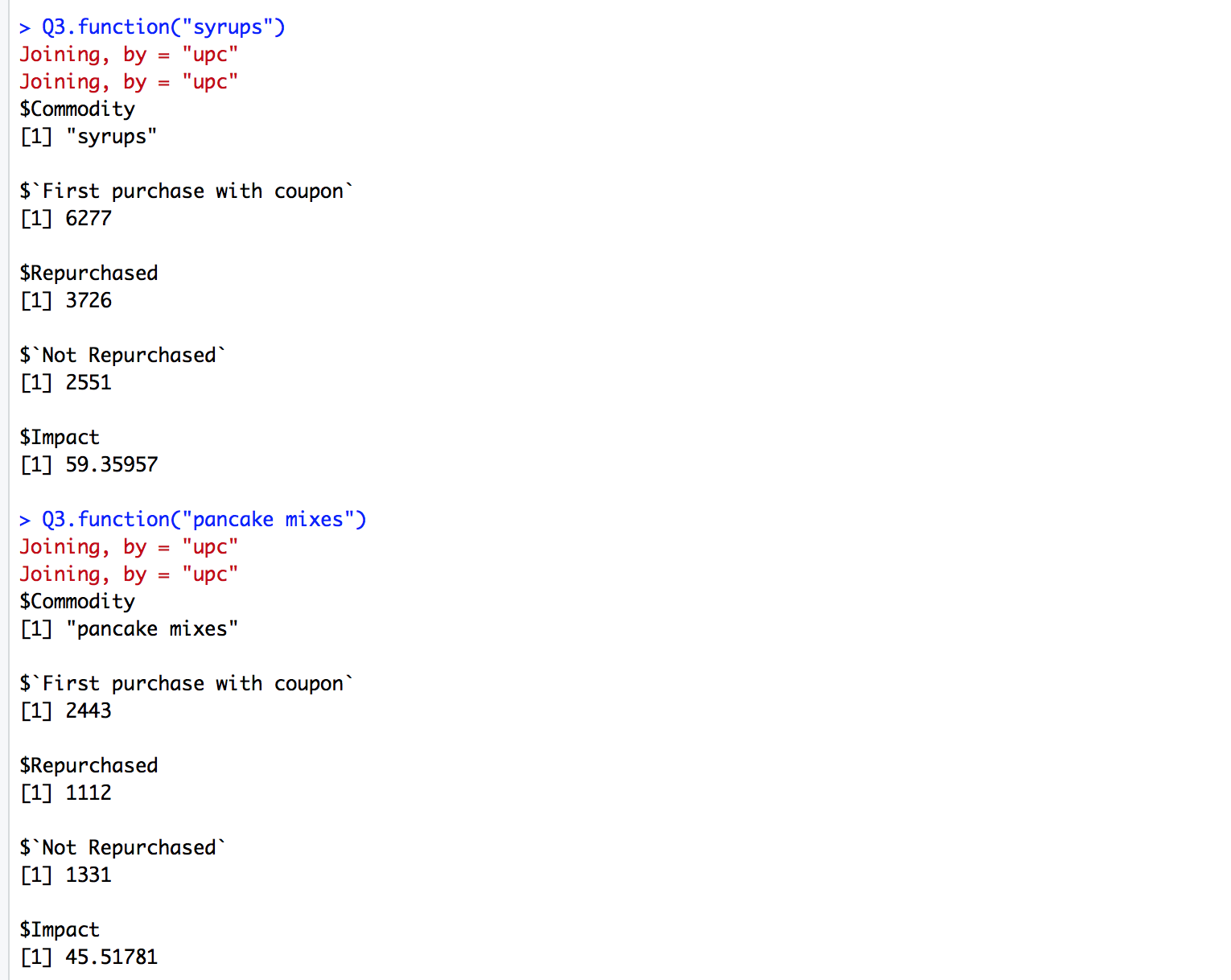
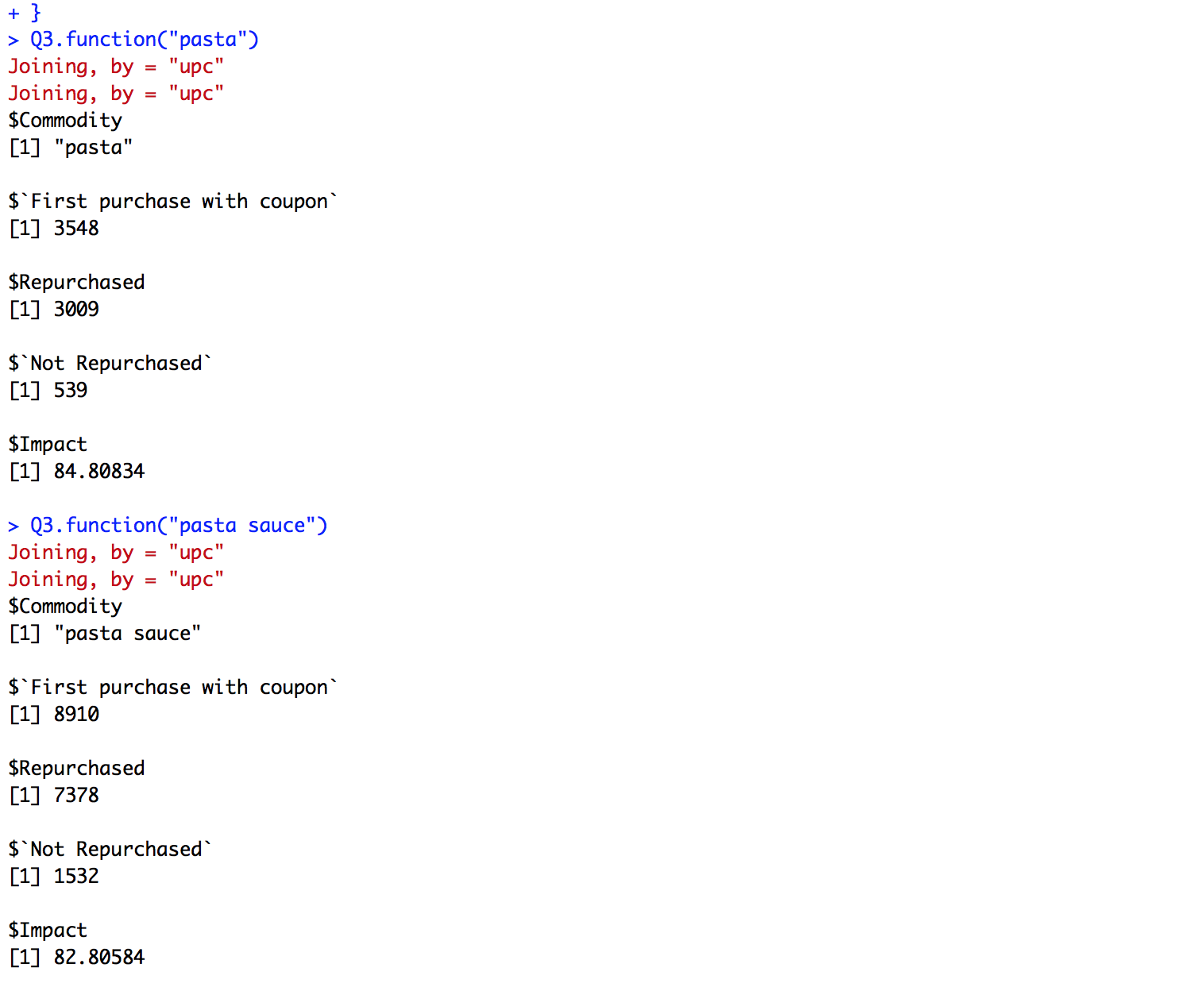


Table : Impact of Coupons on Commodity Sales

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | No. of households that made their first purchase of the commodity with a coupon | No. of households that repurchased after initial coupon purchase | No. of households that did not repurchase after initial coupon purchase | Proportion of households that repurchased after initial coupon purchase |
| Pasta | 3548 | 3009 | 539 | 84.81% |
| Pasta Sauce | 8910 | 7378 | 1532 | 82.81% |
| Syrups | 6277 | 3726 | 2551 | 59.36% |
| Pancake Mixes | 2443 | 1112 | 1331 | 45.52% |
|  |  |  | **Average** | **68.13%** |