Tesco

HanChenyue

2024-04-07

Limitations: - The dataset represents purchases by loyalty cardholder only. This may not be representative of the entire population' purchasing patterns.

I need to find 2 insights from the data and present them in a clear and concise manner. - Explore the correlation between demographics and purchasing patterns such as gender, age, and population density. - Examine the relationship between food category purchases and socio-economic indicators such as average age or population density.

EDA

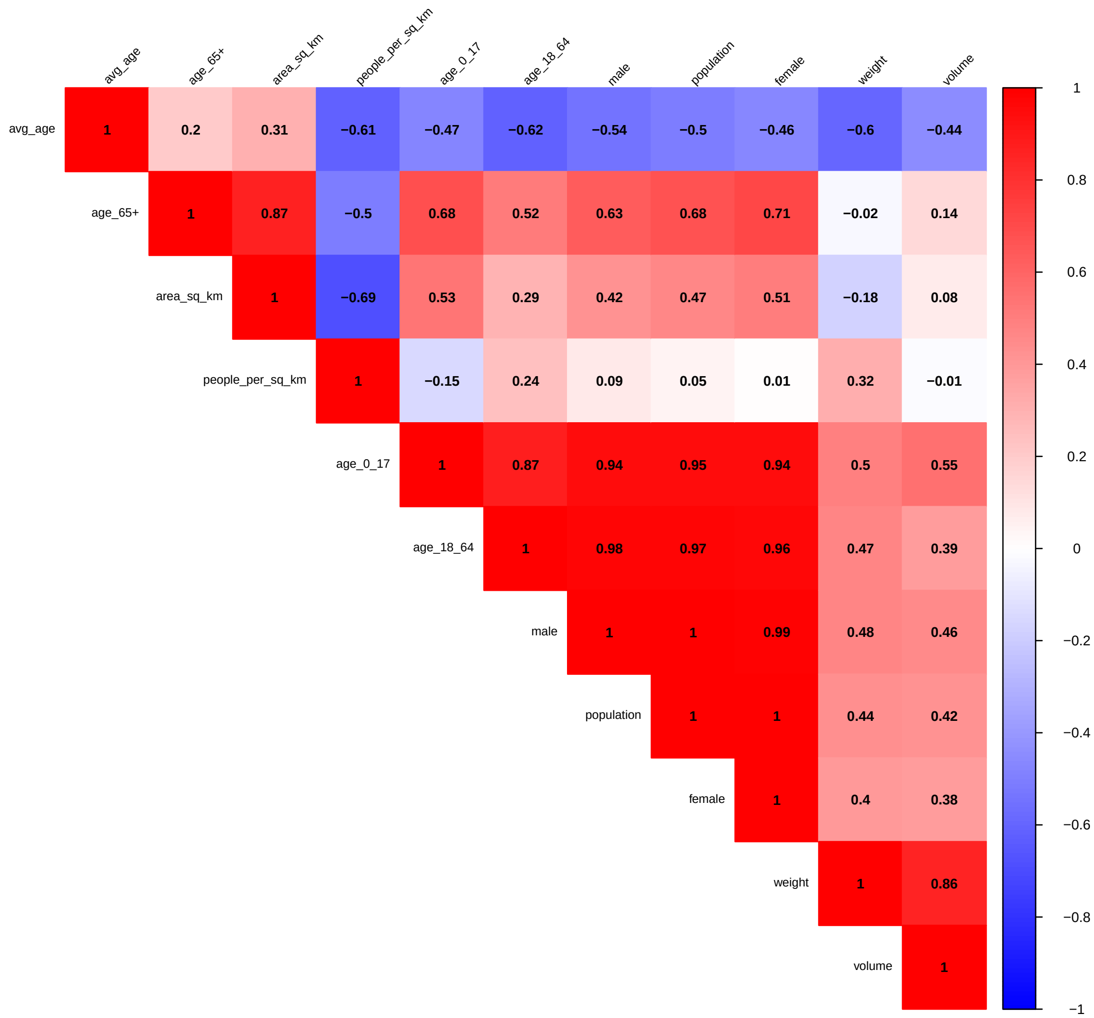
```r
# Check the type of food categories that are available to us
food_categories_unique <- unique(food_categories$category)
kable(food_categories_unique,
      col.names = "Food Categories",
      caption = "Food Categories Available",
      ) %>%
  kable_styling("striped", full_width = F, position = "center")
```

Table 1: Food Categories Available

Food Categories
fruit_veg
grains
sweets
sauces
fats_oils
fish
dairy
readymade
water
eggs
soft_drinks
meat_red
tea_coffee
beer
wine
spirits
poultry

```r
# Check the correlation between demographics and purchasing patterns
# Borough level
borough_year_cor <- borough_year[c('population', 'male', 'female', 'age_0_17', 'age_18_64', 'age_65+',
correlation_matrix <- cor(borough_year_cor, use = "pairwise.complete.obs")
corrplot(correlation_matrix, method = "color",
         type = "upper", # Only upper triangular part of the matrix
         order = "hclust", # Hierarchical clustering order
         tl.col = "black", # Text label color
         tl.srt = 45, # Text label rotation
         tl.cex = 0.6, # Text label size
         addCoef.col = "black", # Add coefficient colour
         title = "Correlation Matrix of Demographic Factors and Purchasing Patterns",
         cl.cex = 0.7, # Color legend text size
         number.cex = 0.7, # Correlation coefficient text size
         number.digits = 2, # Number of digits in correlation coefficient
         mar = c(0, 0, 1, 0), # Margins around the plot
         col = colorRampPalette(c("blue", "white", "red"))(200)) # Change colour scheme
```

Correlation Matrix of Demographic Factors and Purchasing Patterns




```r
kable(correlation_matrix,
      digits = 2,
      caption = "Correlation Matrix of Demographic Factors and Purchasing Patterns") %>%
  kable_styling("striped", full_width = F, position = "center")
```

Table 2: Correlation Matrix of Demographic Factors and Purchasing Patterns

	population	male	female	age_0_17	age_18_64	age_65+	avg_age	area_sq_km	peop
population	1.00	1.00	1.00	0.95	0.97	0.68	-0.50	0.47	
male	1.00	1.00	0.99	0.94	0.98	0.63	-0.54	0.42	
female	1.00	0.99	1.00	0.94	0.96	0.71	-0.46	0.51	

age_0_17	0.95	0.94	0.94	1.00	0.87	0.68	-0.47	0.53
age_18_64	0.97	0.98	0.96	0.87	1.00	0.52	-0.62	0.29
age_65+	0.68	0.63	0.71	0.68	0.52	1.00	0.20	0.87
avg_age	-0.50	-0.54	-0.46	-0.47	-0.62	0.20	1.00	0.31
area_sq_km	0.47	0.42	0.51	0.53	0.29	0.87	0.31	1.00
people_per_sq_km	0.05	0.09	0.01	-0.15	0.24	-0.50	-0.61	-0.69
weight	0.44	0.48	0.40	0.50	0.47	-0.02	-0.60	-0.18
volume	0.42	0.46	0.38	0.55	0.39	0.14	-0.44	0.08

Borough Level Correlation Matrix - Population and Purchasing Volume/Weight: There's a correlation between the population size of a borough and both the weight and volume of purchases. This is expected as larger populations would naturally lead to more purchases. - Age Groups and Purchases: Different age groups (0-17, 18-64, 65+) show varying degrees of correlation with purchasing patterns. This could suggest that the age composition of a borough influences the types and amounts of groceries purchased. - Area Size and Density: The area of the borough (area_sq_km) and the population density (people_per_sq_km) also show interesting correlations with purchasing patterns.

Details Population Size - Weight: Correlation coefficient of 0.44, indicating a moderate positive correlation. As the population size increases, the weight of purchases also increases. - Volume: Correlation coefficient of 0.42, indicating a moderate positive correlation. As the population size increases, the volume of purchases also increases.

Gender Distribution - Male has a slightly higher correlation compare to female in terms of weight and volume of purchases. - This make sense given that usually male tend to consume more food compare to female

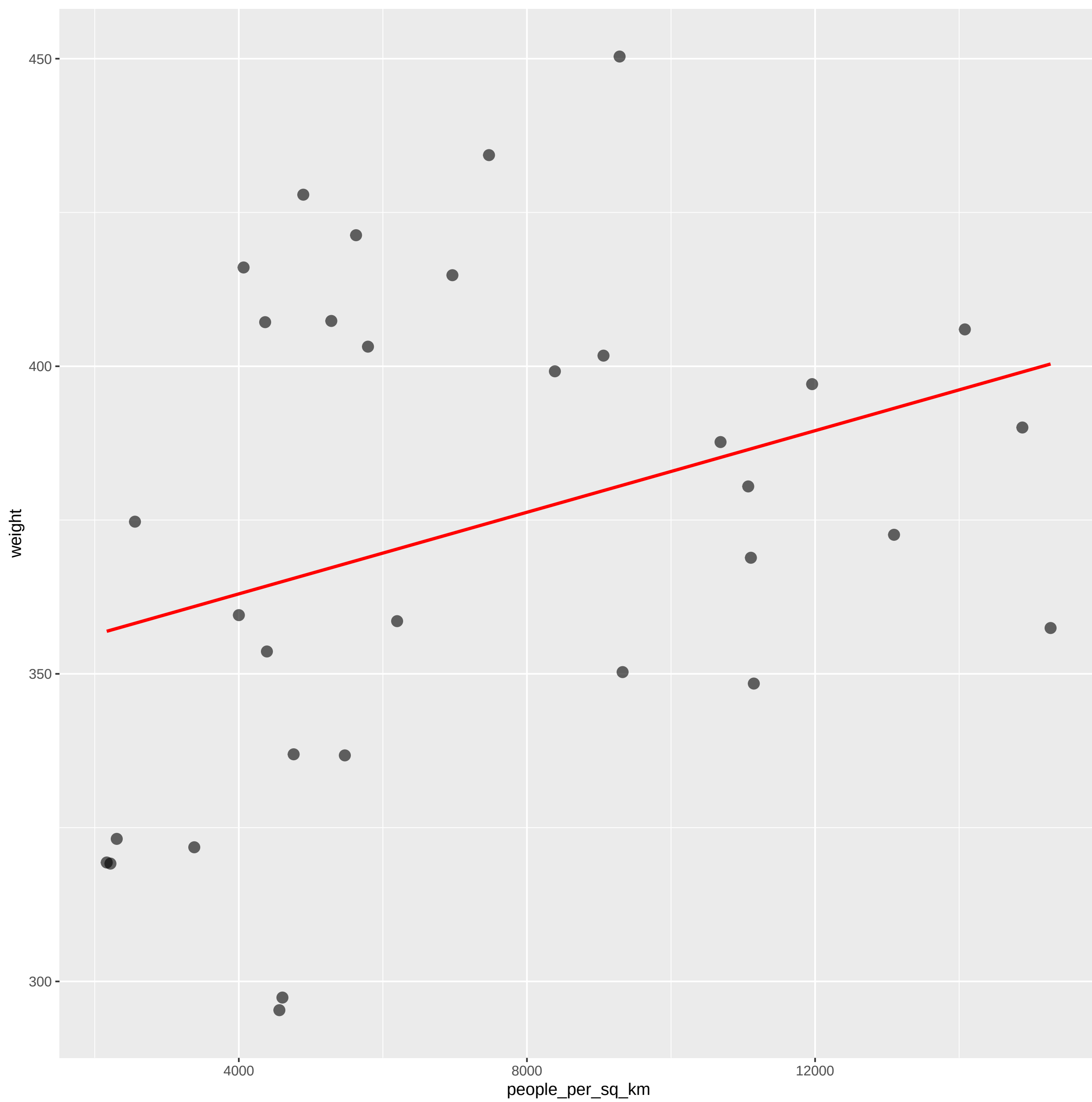
Age Groups - Age 0-17: Strong positive correlation with weight (0.50) and volume (0.55), indicating that areas with a higher proportion of children and teenagers tend to have higher purchase volumes. - Age 18-64: Moderate positive correlation with weight (0.44) and volume (0.42), suggesting that the working-age population contributes significantly to the weight and volume of purchases. - Age 65+: Weak negative correlation with weight (-0.02) and weak positive correlation with volume (0.17), indicating that the elderly population has a smaller impact on purchasing patterns.

Average Age - There is a strong negative correlation between average age and weight (-0.60) and volume (-0.44), suggesting that younger populations tend to purchase more groceries (in terms of weight and volume).

Area Size and Density - Area_sq_km: Weak negative correlation with weight (-0.18) and weak positive correlation with volume (0.08), indicating that the size of the borough has a small impact on purchasing patterns. - People_per_sq_km: Shows a moderate positive correlation with weight (0.32) and negligible correlation with volume (-0.01), suggesting that population density has a moderate impact on the weight of purchases and population density doesn't significantly affect the volume of purchases.

```r
demographics_borough_year <- borough_year[c('avg_age', 'people_per_sq_km')]

# Scatter plot for Weight vs. Population Density
ggplot(borough_year, aes(x = people_per_sq_km, y = weight)) +
  geom_point(alpha = 0.6, size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



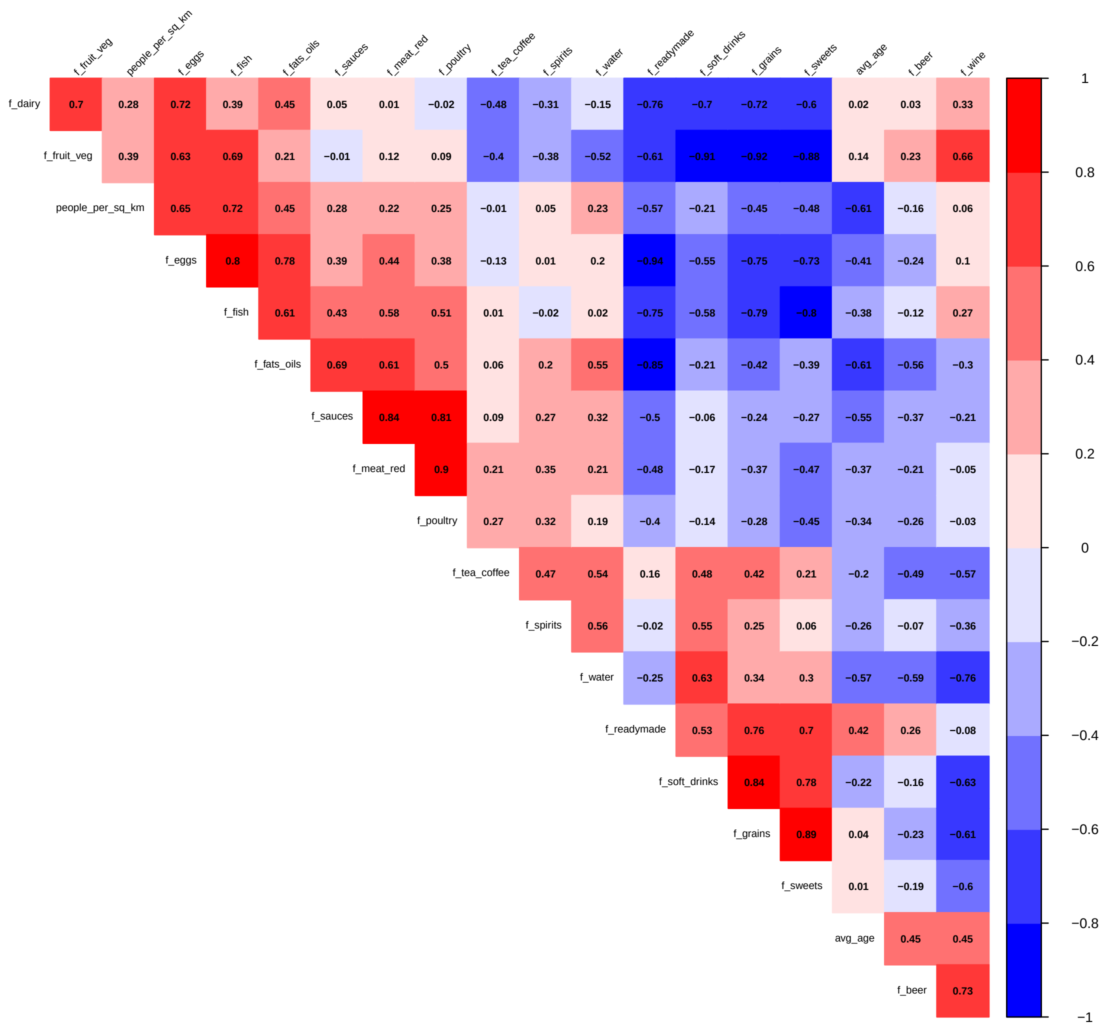

```
#
# # Scatter plot for Weight vs. Average Age
# ggplot(borough_year, aes(x = avg_age, y = weight)) +
#   geom_point(alpha = 0.6, size = 3) +
#   geom_smooth(method = "lm", se = FALSE, color = "red")
#
# # Scatter plot for Volume vs. Population Density
# ggplot(borough_year, aes(x = people_per_sq_km, y = volume)) +
#   geom_point(alpha = 0.6, size = 3) +
#   geom_smooth(method = "lm", se = FALSE, color = "red")
#
# # Scatter plot for Volume vs. Average Age
# ggplot(borough_year, aes(x = avg_age, y = volume)) +
#   geom_point(alpha = 0.6, size = 3) +
```

```
#   geom_smooth(method = "lm", se = FALSE, color = "red")
```

Explore the relationship between food category purchases and socio-economic indicators

```
# Check the correlation between food category purchases and socio-economic indicators
# Borough level
socio_economic_indicators <- borough_year[c('avg_age', 'people_per_sq_km')]
food_categories_borough_year <- borough_year[c('f_beer', 'f_dairy', 'f_eggs', 'f_fats_oils', 'f_fish',

data_for_correlation <- cbind(food_categories_borough_year, socio_economic_indicators)
correlation_matrix_food <- cor(data_for_correlation, use = "pairwise.complete.obs")
corrplot(correlation_matrix_food, method = "color",
         type = "upper", # Only upper triangular part of the matrix
         order = "hclust", # Hierarchical clustering order
         tl.col = "black", # Text label color
         tl.srt = 45, # Text label rotation
         tl.cex = 0.5, # Reduce text label size for space
         addCoef.col = "black", # Add coefficient color
         title = "Correlation Matrix of Food Category Purchases and Socio-Economic Indicators",
         cl.cex = 0.7, # Color legend text size
         number.cex = 0.5, # Reduce correlation coefficient text size
         number.digits = 2, # Number of digits in correlation coefficient
         sig.level = 0.05, # Only show significant correlations
         diag = FALSE, # Do not show diagonal
         mar = c(0, 0, 1, 0), # Margins around the plot
         col = colorRampPalette(c("blue", "white", "red"))(10)) # Simplify color scheme
```

Correlation Matrix of Food Category Purchases and Socio−Economic Indicators



Average Age: - Positive correlations are observed with f_beer (0.45), f_readymade (0.42), and f_wine (0.45), suggesting these items are more popular in areas with an older population. - Strong negative correlations with f_fats_oils (-0.61), f_sauces (-0.55), and f_water (-0.57) indicate these items are less frequently purchased in older populations.

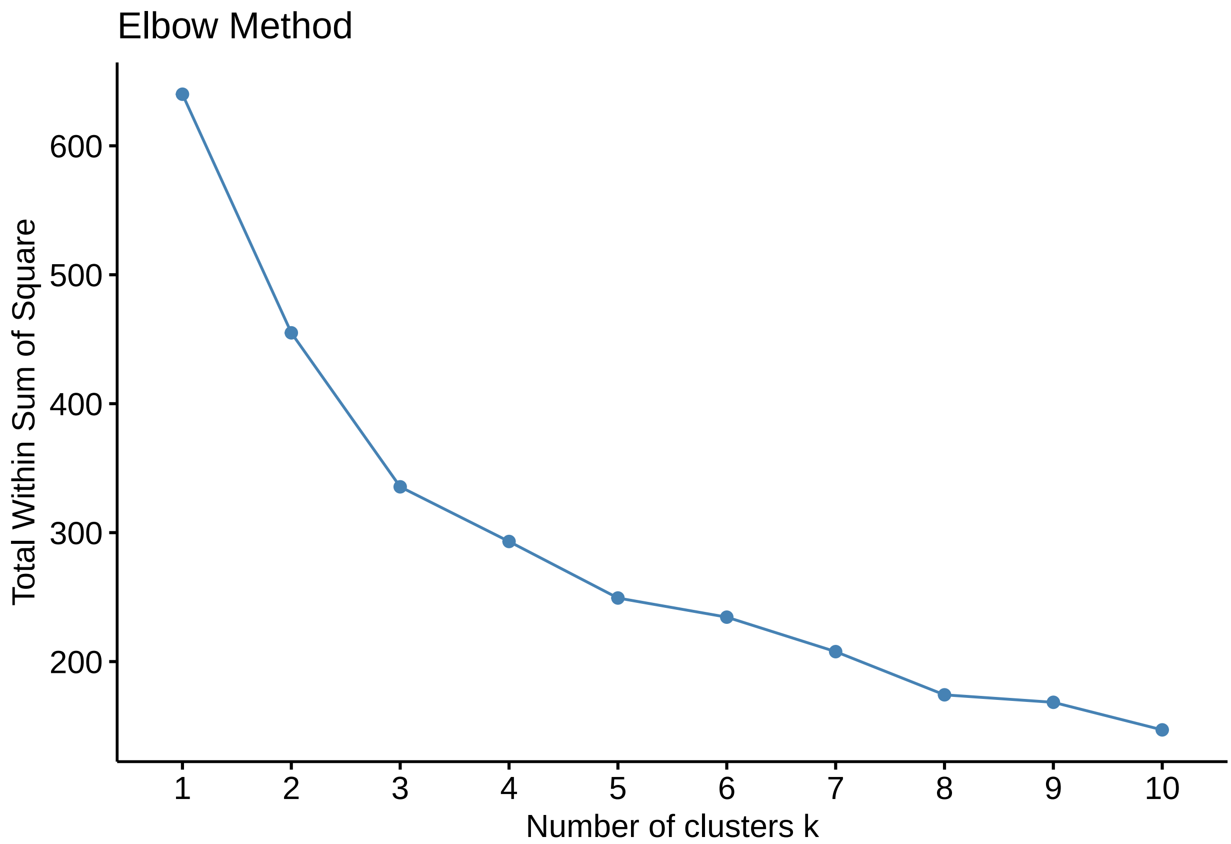
Population Density: - f_fish shows a strong positive correlation (0.72), suggesting higher purchases in densely populated areas. - f_eggs also has a high positive correlation (0.65) with population density. - f_readymade exhibits a strong negative correlation (-0.57), suggesting lower purchases in denser areas. - f_grains and f_sweets also show negative correlations (-0.45 and -0.48 respectively), indicating lower purchases in more densely populated areas.

```
# Use cluster analysis to identify possible purchasing pattern worth exploring
# Elbow method allows us to determine the optimal number of clusters
# If it matches the number of age groups/gender, it could indicate distinct purchasing patterns for eac
# Products categories in csv file
```

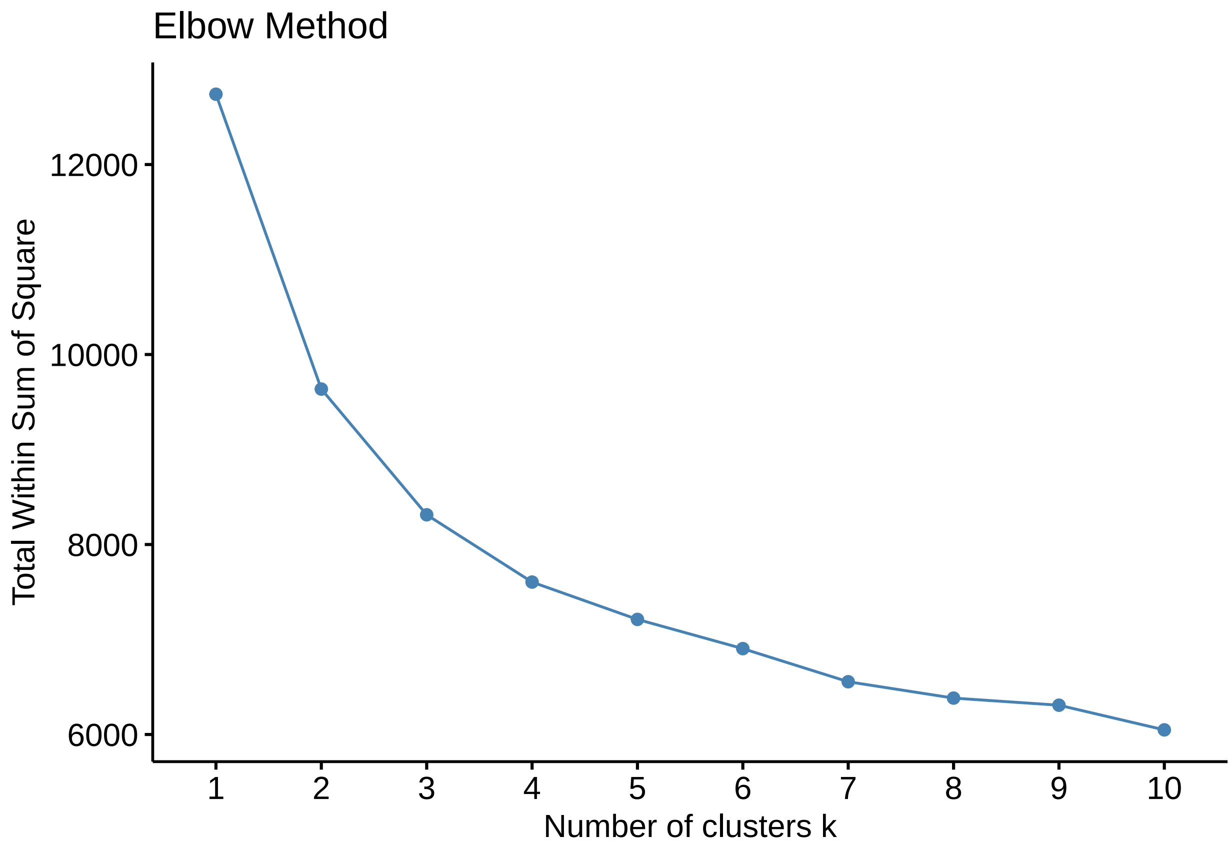
```r
product_categories <- c('f_beer', 'f_dairy', 'f_eggs', 'f_fats_oils', 'f_fish',
                        'f_fruit_veg', 'f_grains','f_meat_red', 'f_poultry',
                        'f_readymade', 'f_sauces', 'f_soft_drinks', 'f_spirits',
                        'f_sweets', 'f_tea_coffee', 'f_water', 'f_wine')

age_columns <- c('age_0_17', 'age_18_64', 'age_65+')
gender_columns <- c('male', 'female')

# K-means clustering to identify patterns in the data
# EDA to check age group vs food categories
# Function to check elbow method
purchasing_patterns_cluster_function <- function(data, hand_picked_features){
  # Selecting the specific age group and purchasing patterns
  features <- data %>% select(hand_picked_features, 'f_beer', 'f_dairy', 'f_eggs', 'f_fats_oils', 'f_fi
                'f_meat_red', 'f_poultry', 'f_readymade', 'f_sauces', 'f_soft_drinks', 'f_spirits',
                'f_sweets', 'f_tea_coffee', 'f_water', 'f_wine')
  # scaled_data <- scale(data[, features])
  scaled_data <- scale(features)

  # Determine the optimal number of clusters using the elbow method
  set.seed(0)  # Ensure reproducibility
  wcss <- map_dbl(1:10, function(k) {
    kmeans(scaled_data, centers = k, iter.max = 50, nstart = 25)$tot.withinss
})

  fviz_nbclust(scaled_data, kmeans, method = "wss") + labs(title = 'Elbow Method')
}
```

```r
purchasing_patterns_cluster_function(borough_year, age_columns)
```

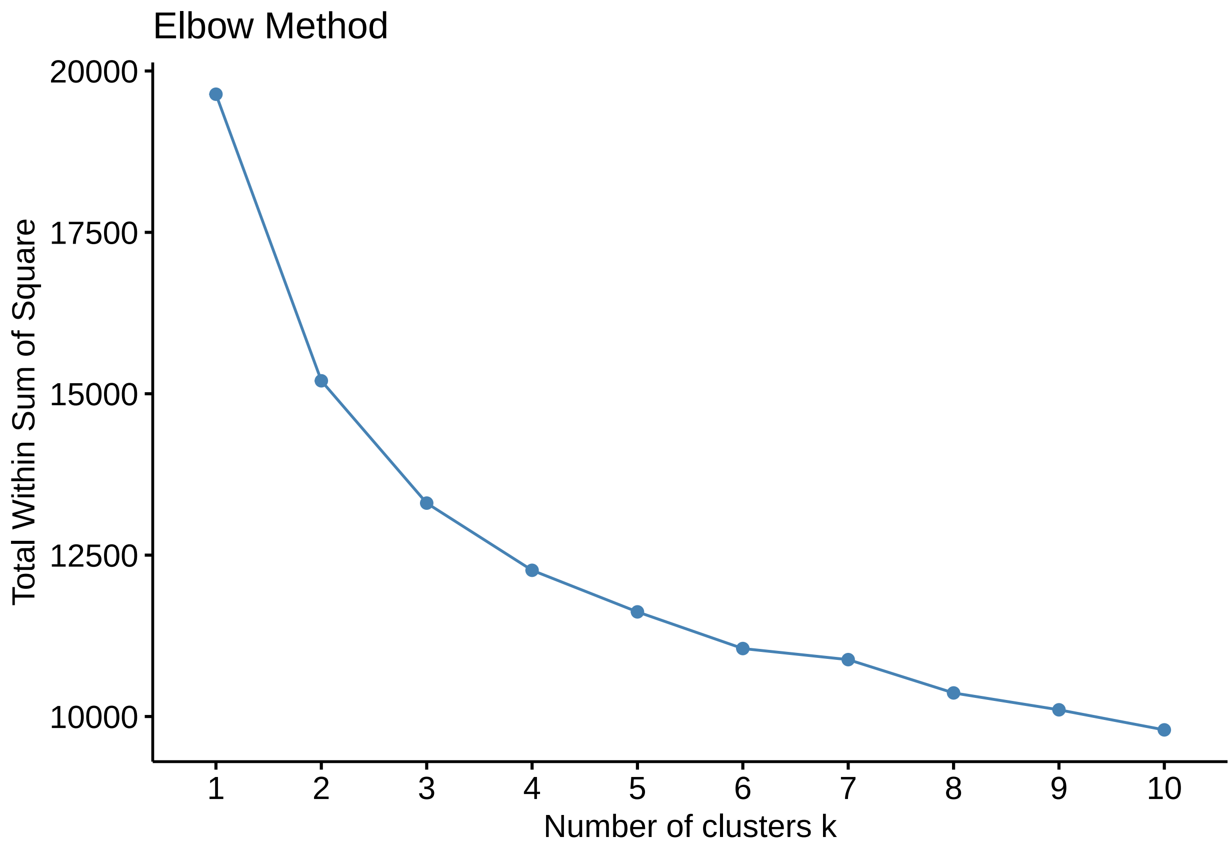
```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(hand_picked_features)
##
##   # Now:
##   data %>% select(all_of(hand_picked_features))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
purchasing_patterns_cluster_function(osward_year, age_columns)
```

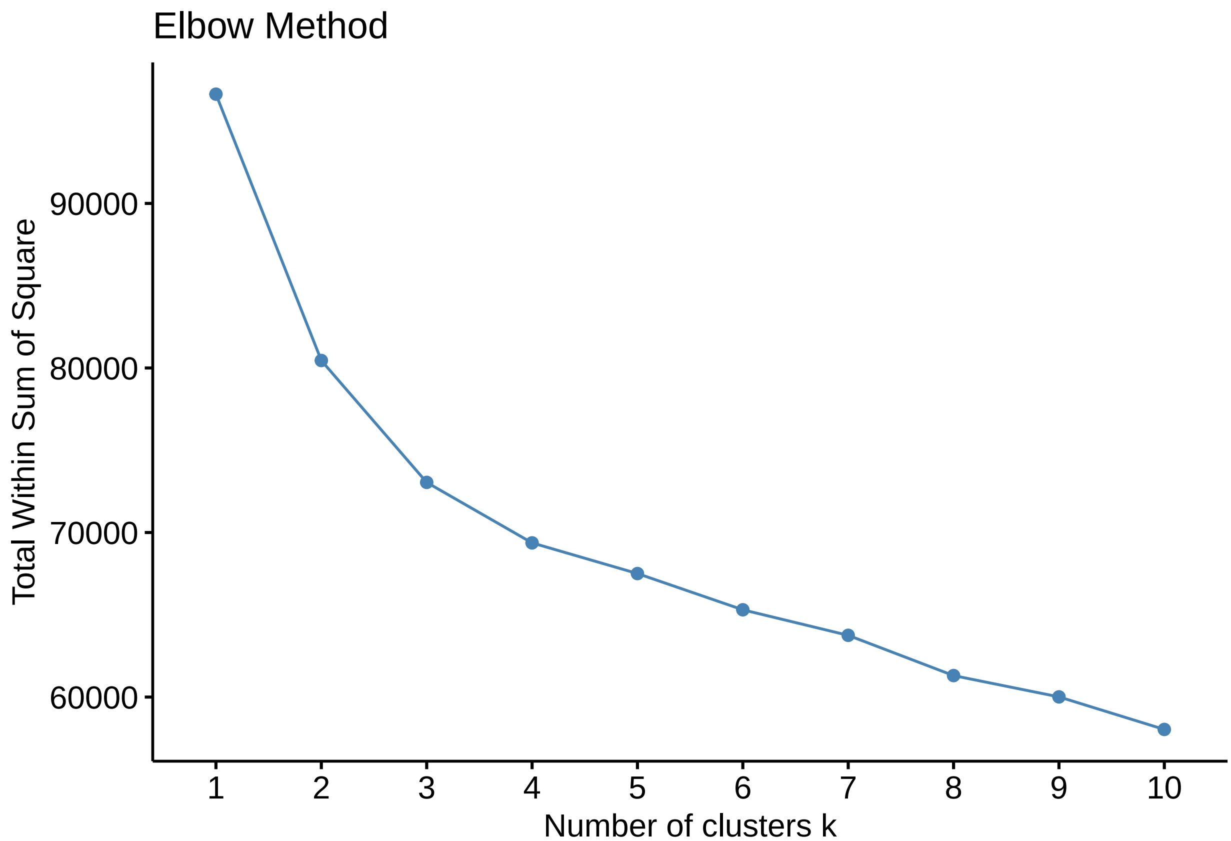



```r
purchasing_patterns_cluster_function(msoa_year, age_columns)
```

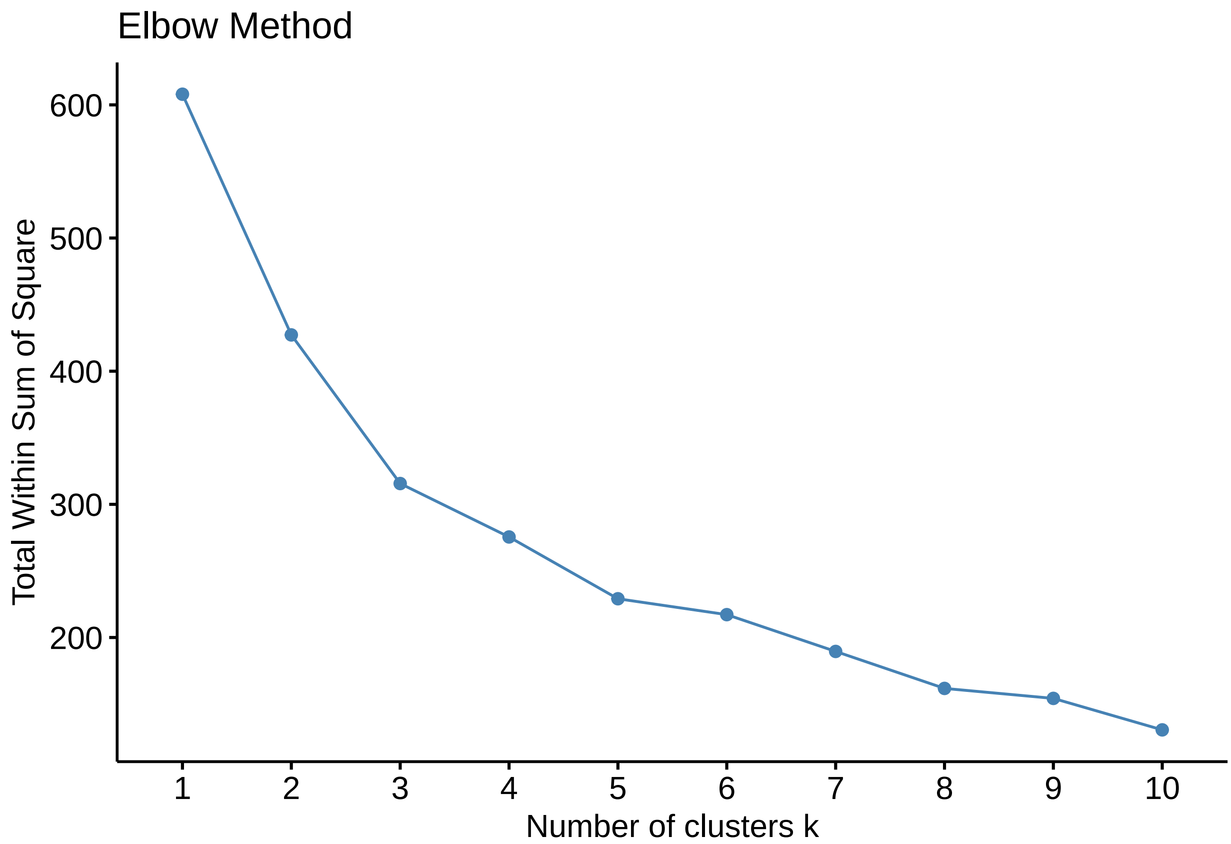



```r
purchasing_patterns_cluster_function(lsoa_year, age_columns)
```

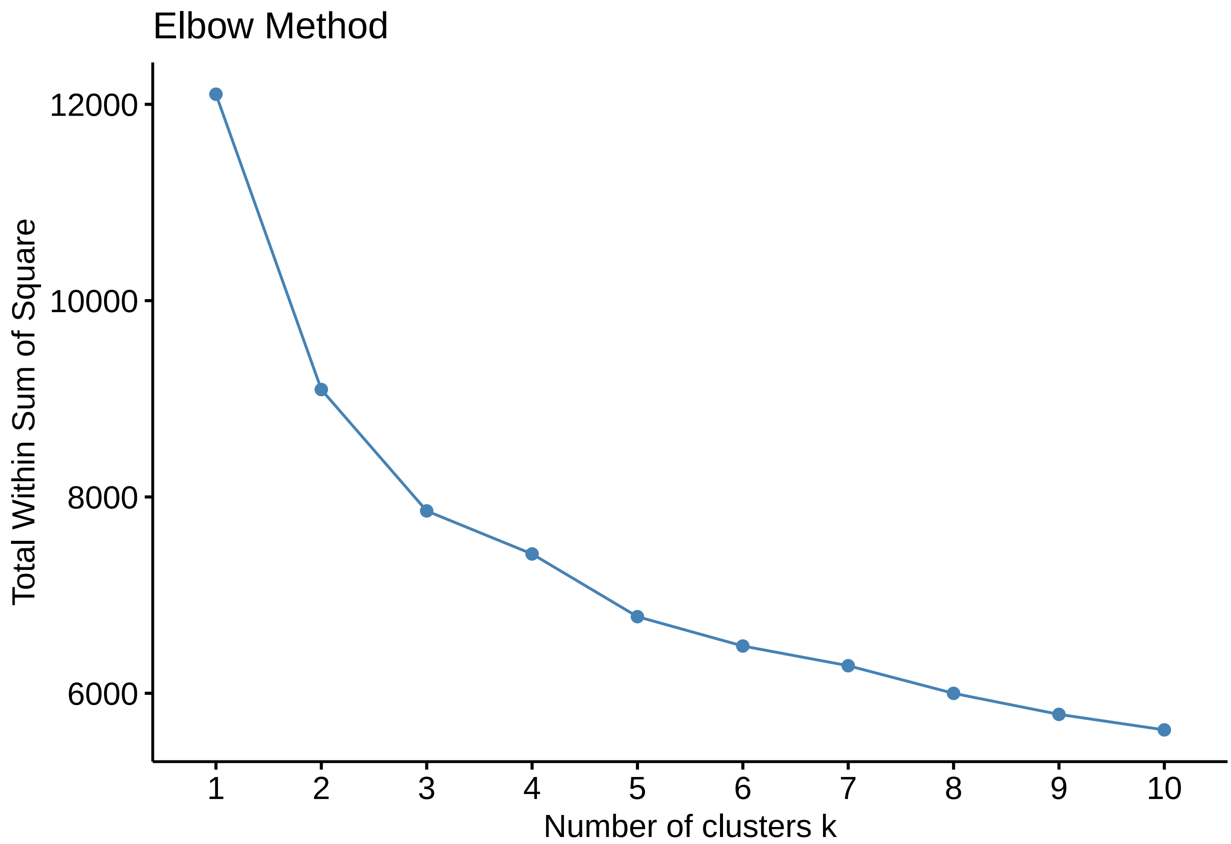
```
## Warning: did not converge in 10 iterations
```



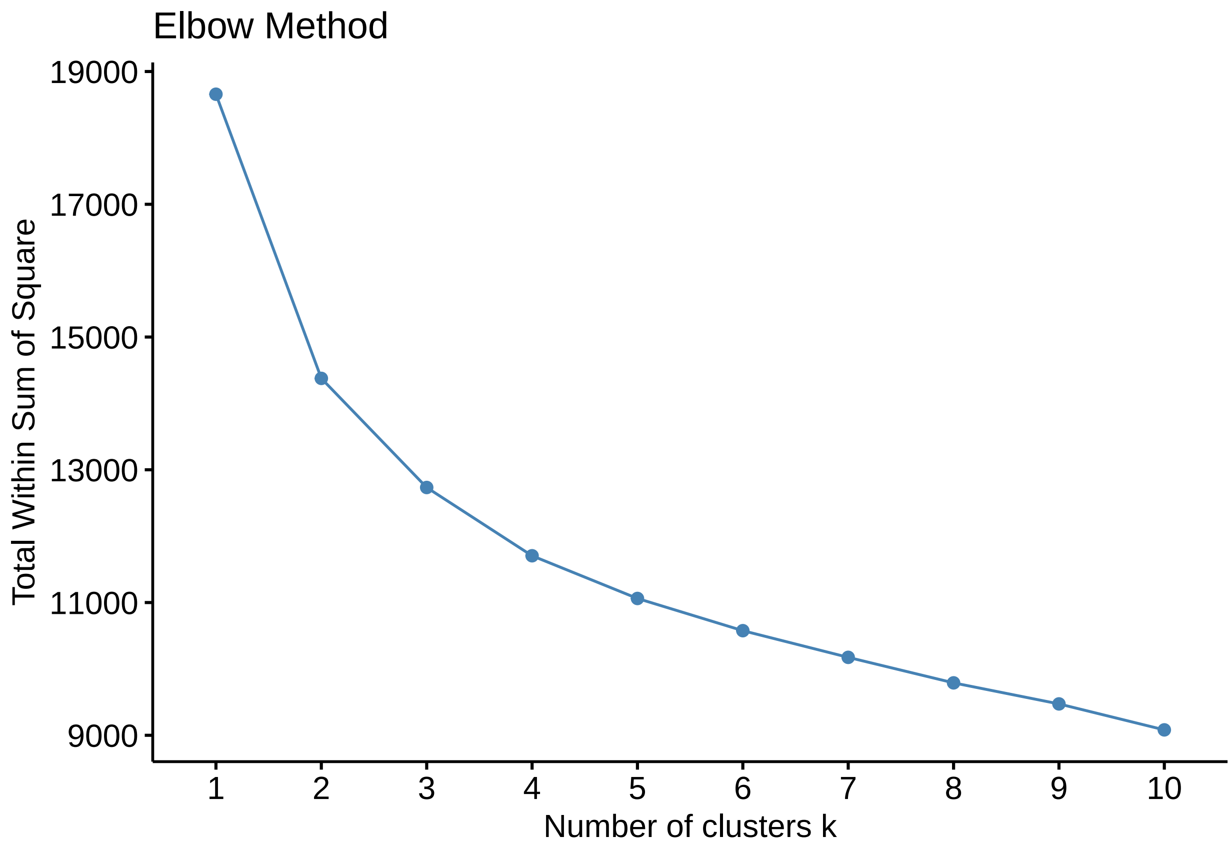

```r
purchasing_patterns_cluster_function(borough_year, gender_columns)
```



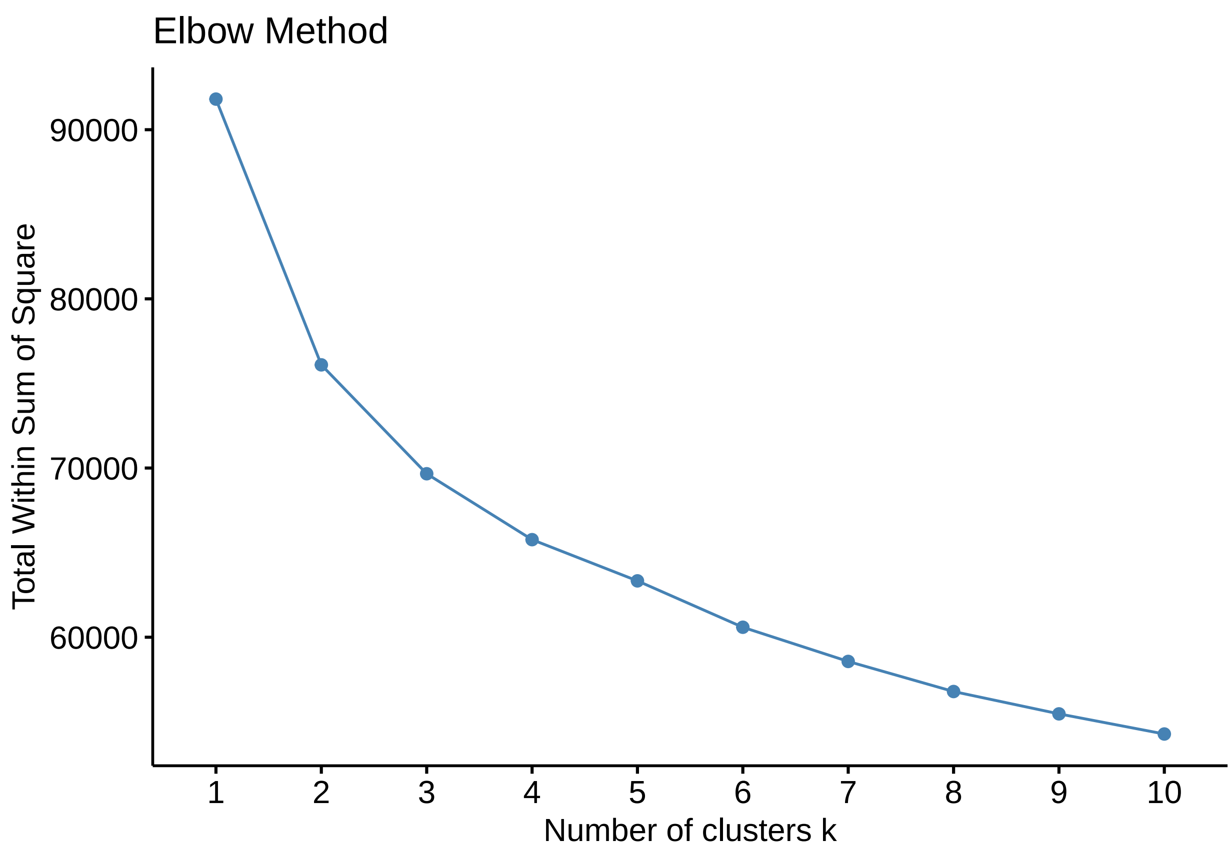

```
purchasing_patterns_cluster_function(osward_year, gender_columns)
```



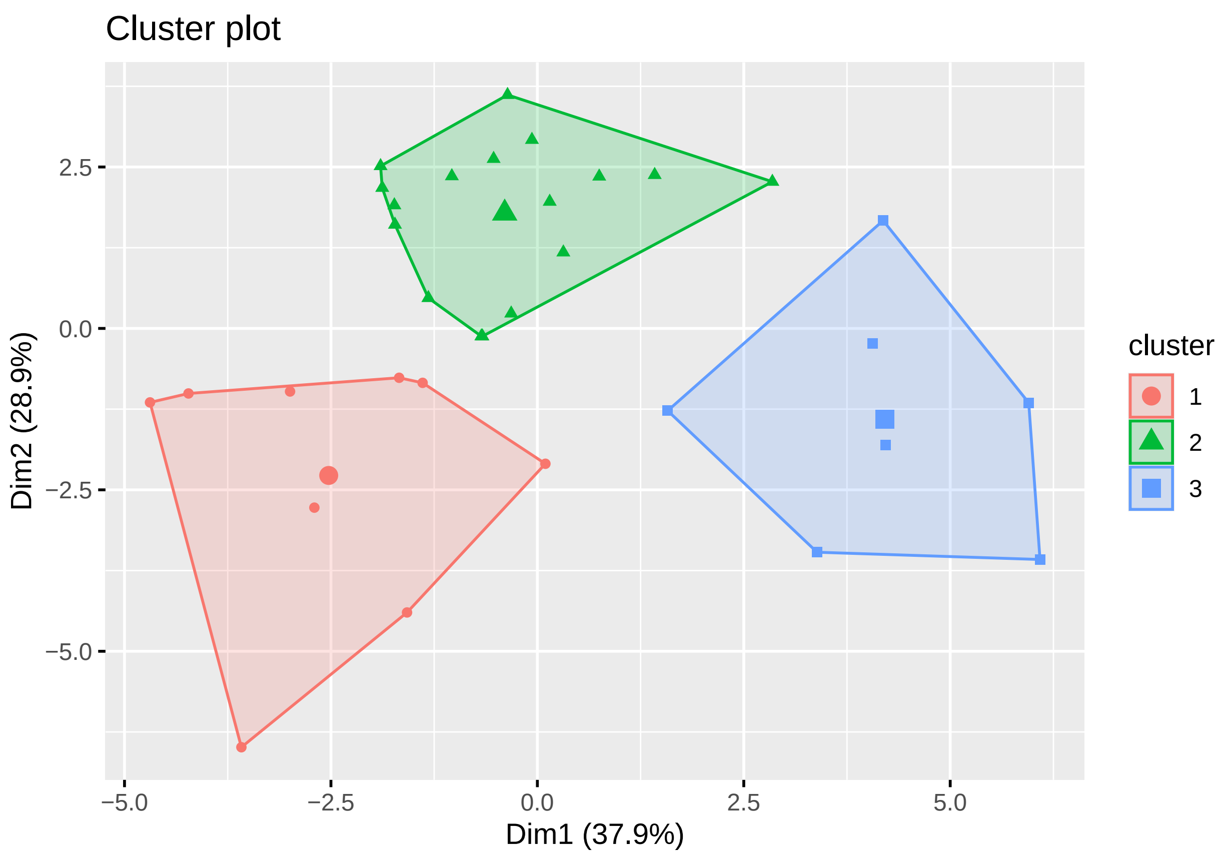

```
purchasing_patterns_cluster_function(msoa_year, gender_columns)
```



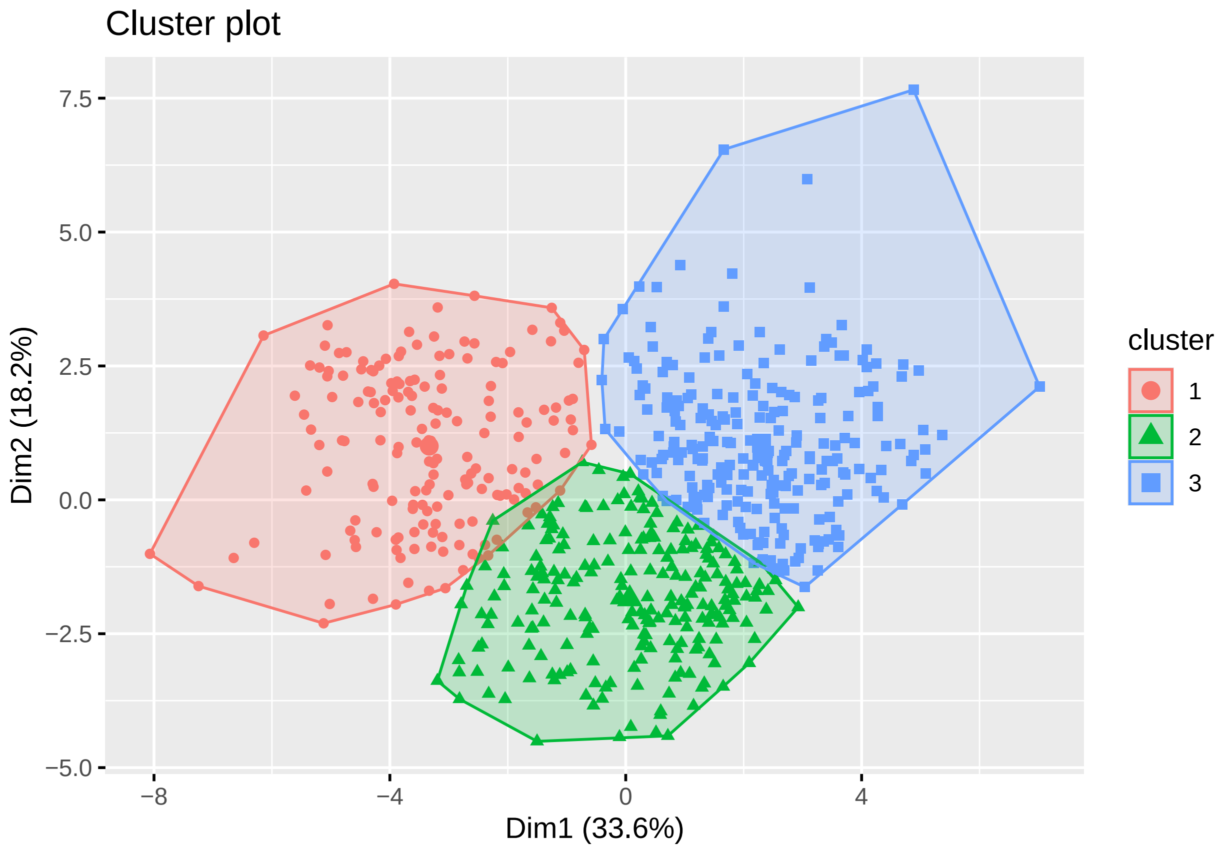

```
purchasing_patterns_cluster_function(lsoa_year, gender_columns)
```



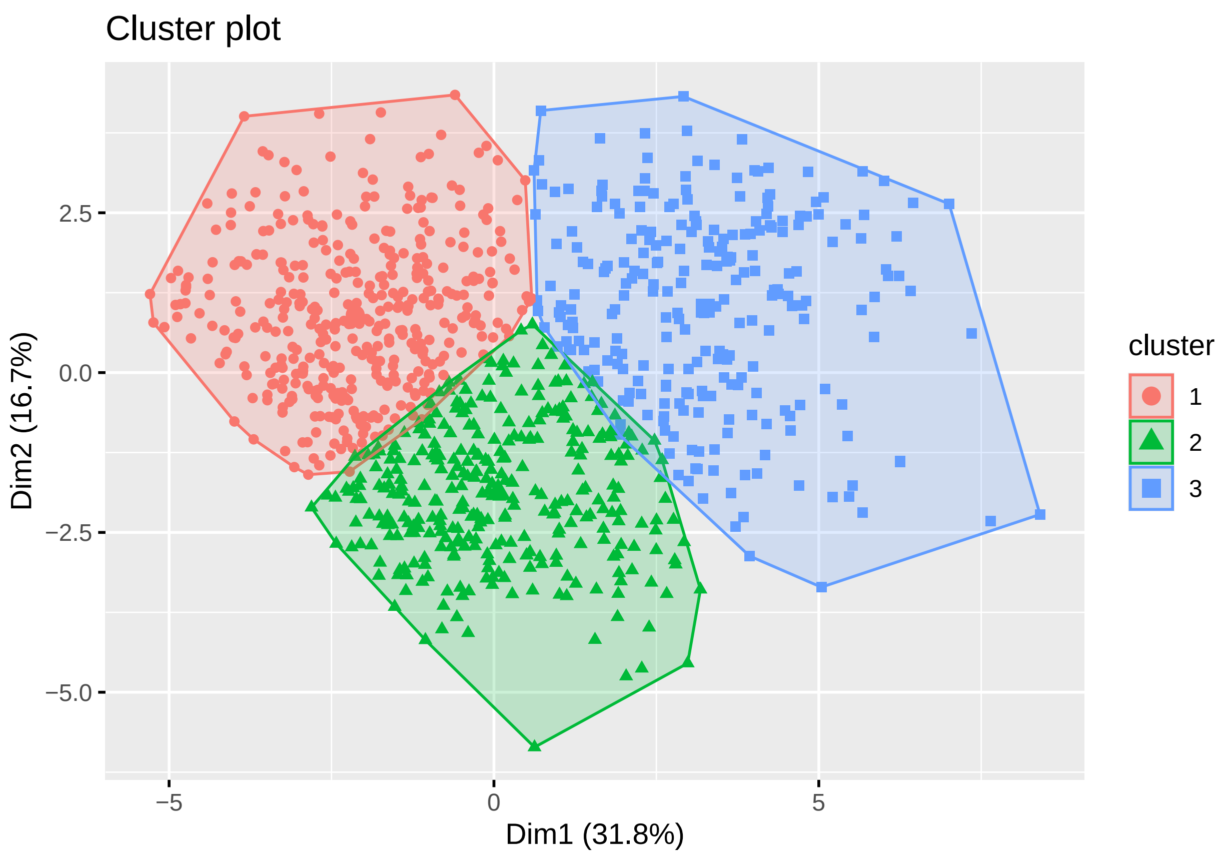

```r
# Function to visualise the clustering results using fviz_cluster
visualise_cluster <- function(data, hand_picked_features, n_clusters) {
  # Selecting the specific age group and purchasing patterns
  features <- data %>% select(hand_picked_features, 'f_beer', 'f_dairy', 'f_eggs', 'f_fats_oils', 'f_fi
                'f_meat_red', 'f_poultry', 'f_readymade', 'f_sauces', 'f_soft_drinks', 'f_spirits',
                'f_sweets', 'f_tea_coffee', 'f_water', 'f_wine')
  # scaled_data <- scale(data[, features])
  scaled_data <- scale(features)

  # Applying k-means clustering
  set.seed(0)
  kmeans_result <- kmeans(scaled_data, centers = n_clusters, iter.max = 50, nstart = 25)

  # Visualizing the clustering results
  fviz_cluster(kmeans_result, data = scaled_data, geom = "point", stand = FALSE, ellipse.type = "convex
}

visualise_cluster(borough_year, age_columns, 3)
```



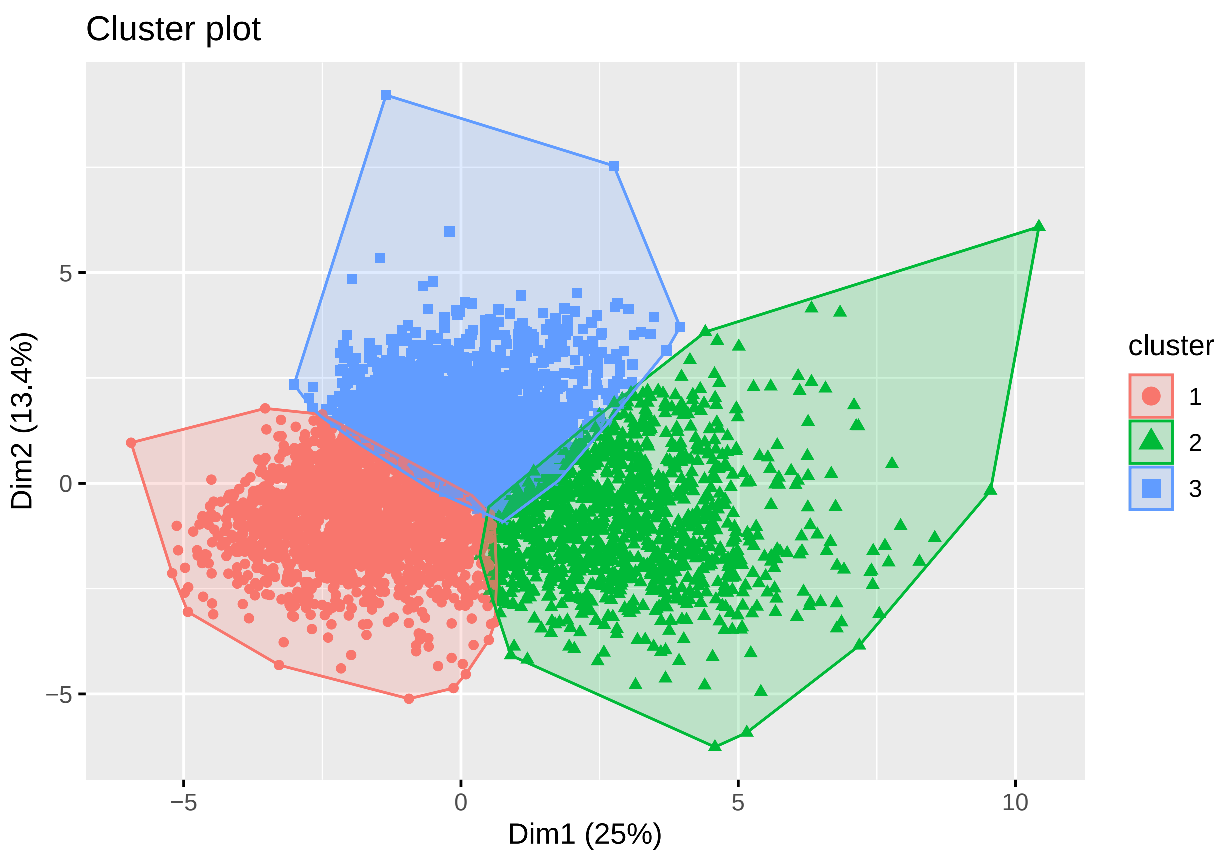

```
visualise_cluster(osward_year, age_columns, 3)
```



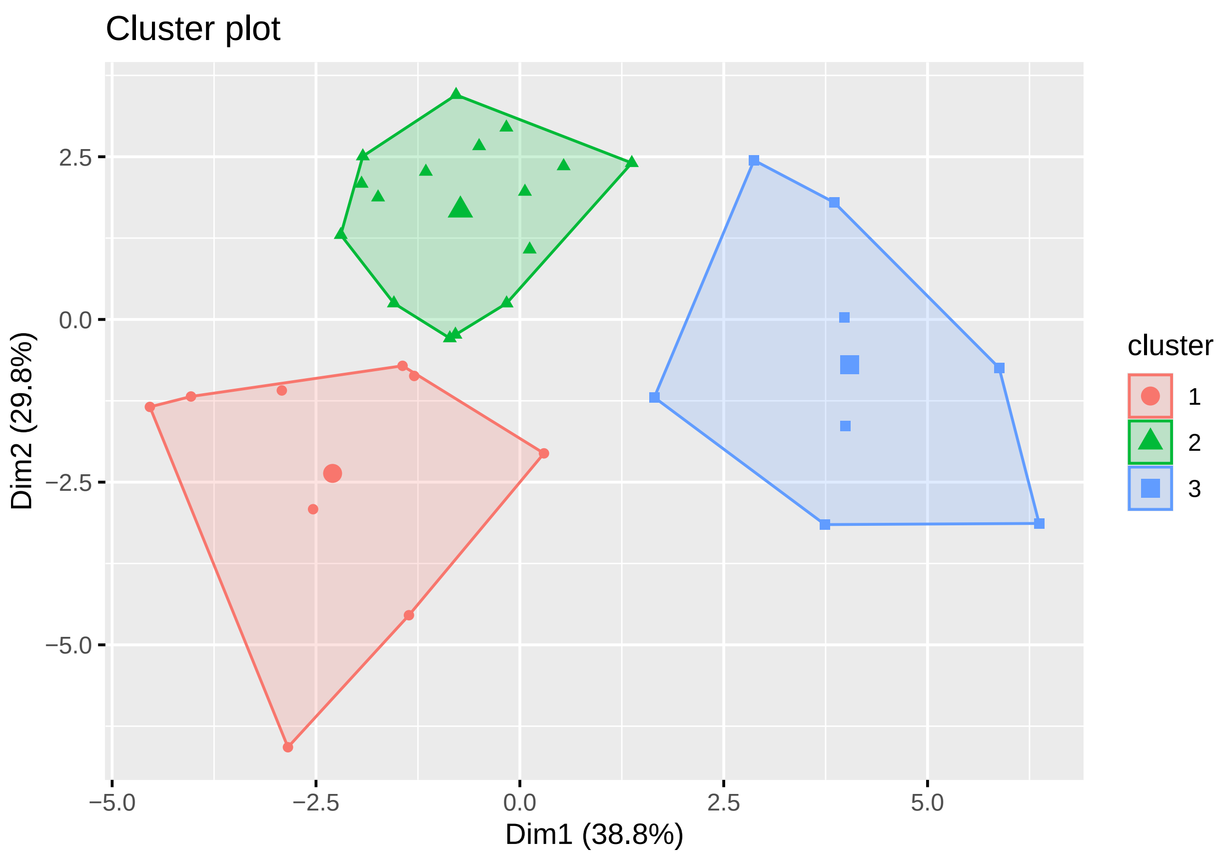

```
visualise_cluster(msoa_year, age_columns, 3)
```



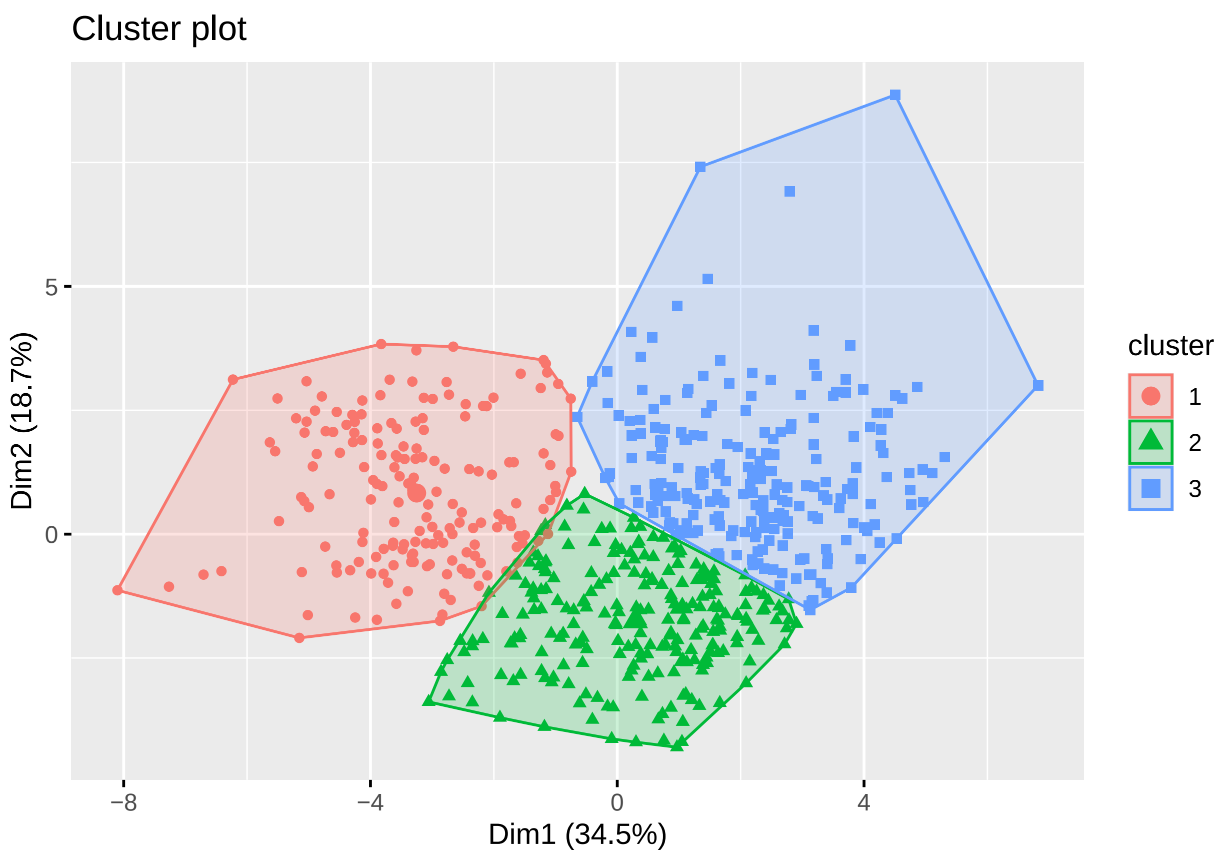

```r
visualise_cluster(lsoa_year, age_columns, 3)
```



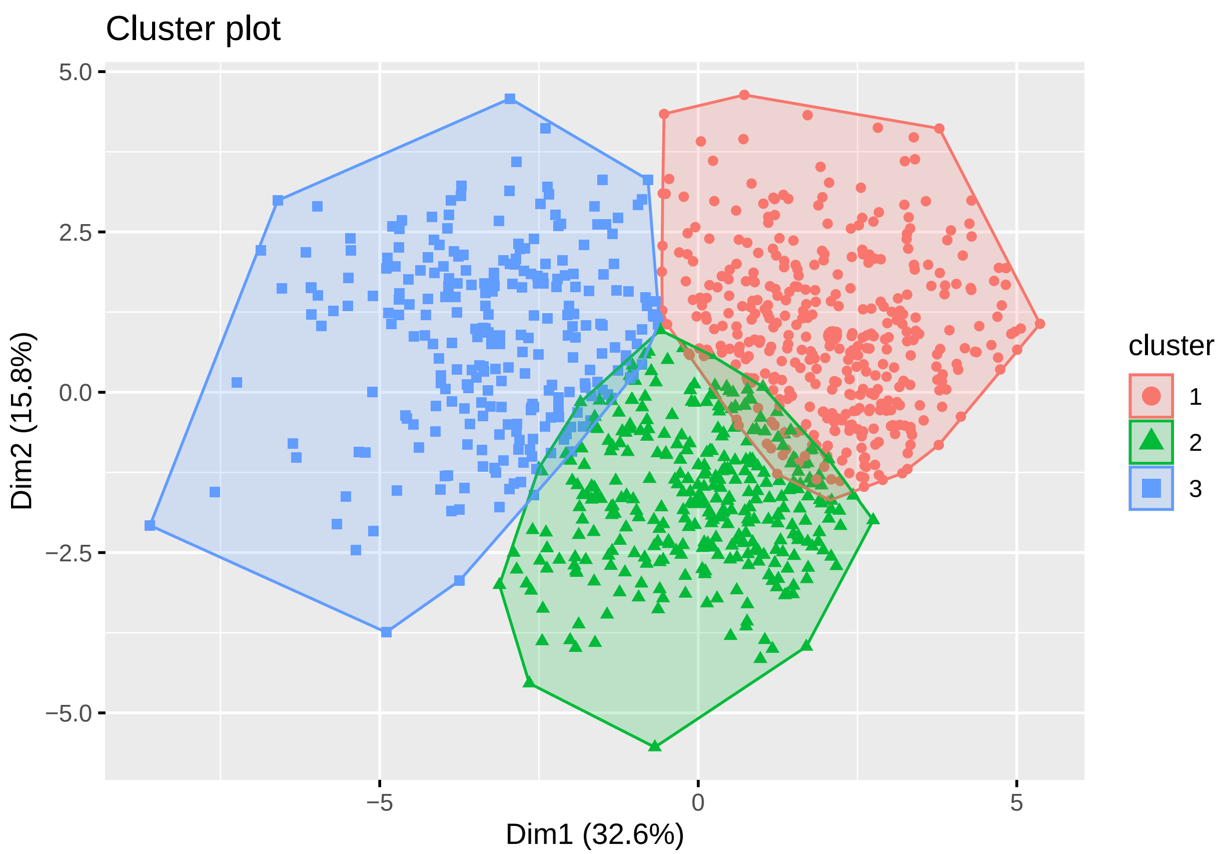

```r
visualise_cluster(borough_year, gender_columns, 3)
```



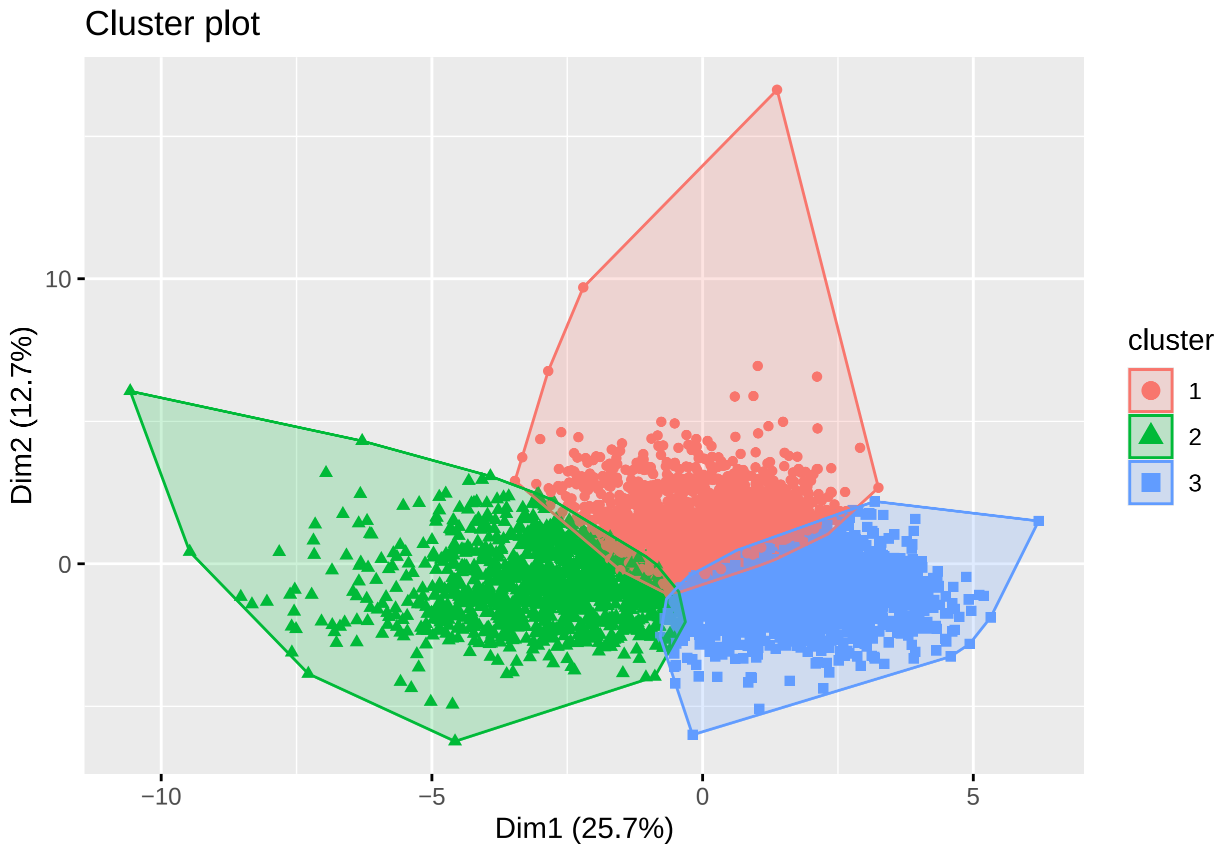

```r
visualise_cluster(osward_year, gender_columns, 3)
```



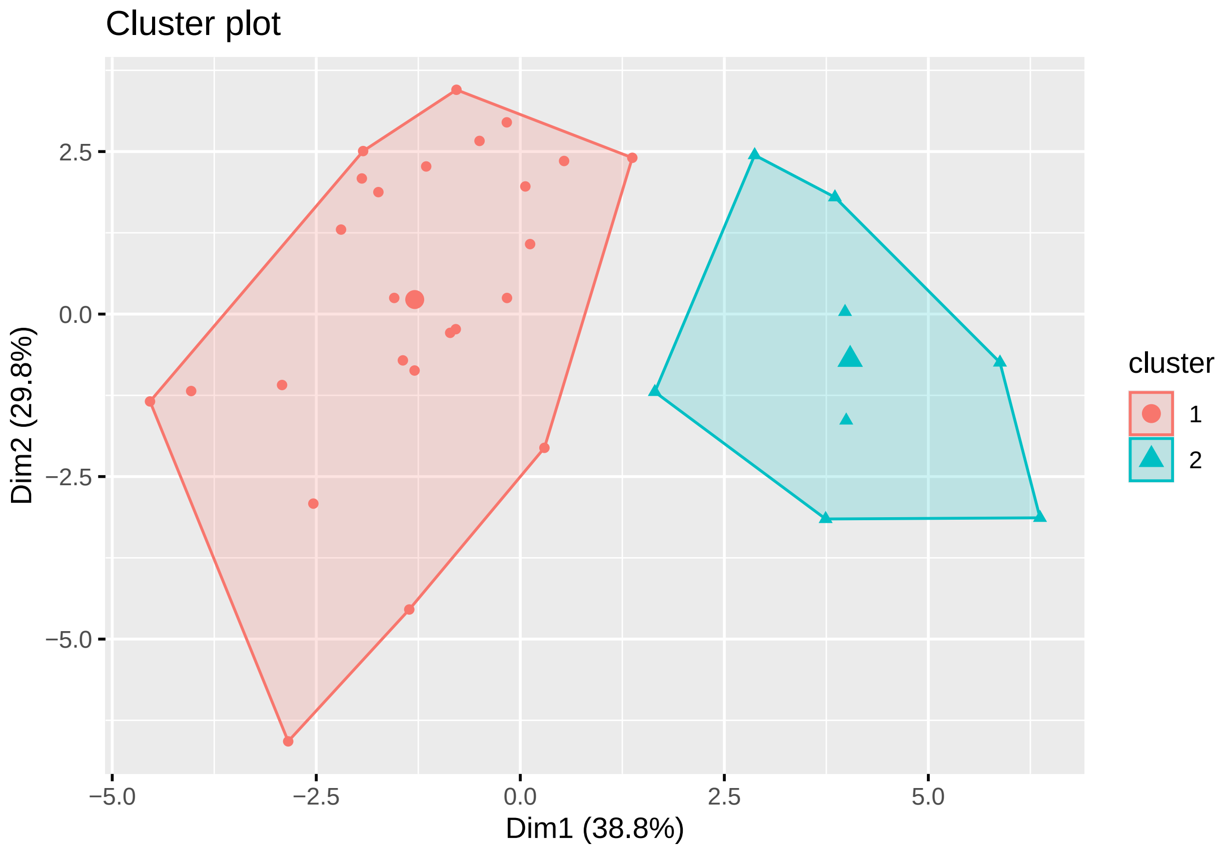

```
visualise_cluster(msoa_year, gender_columns, 3)
```



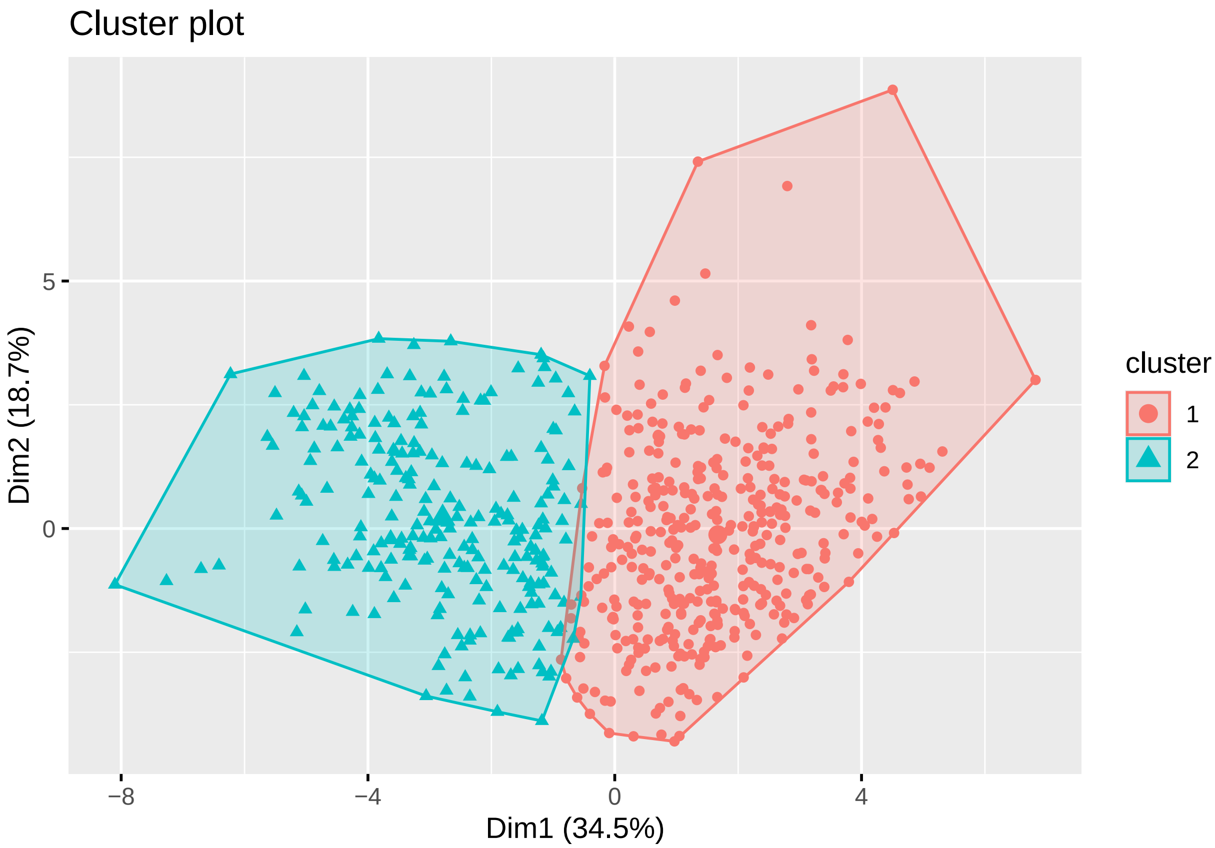

```r
visualise_cluster(lsoa_year, gender_columns, 3)
```




```r
visualise_cluster(borough_year, gender_columns, 2)
```

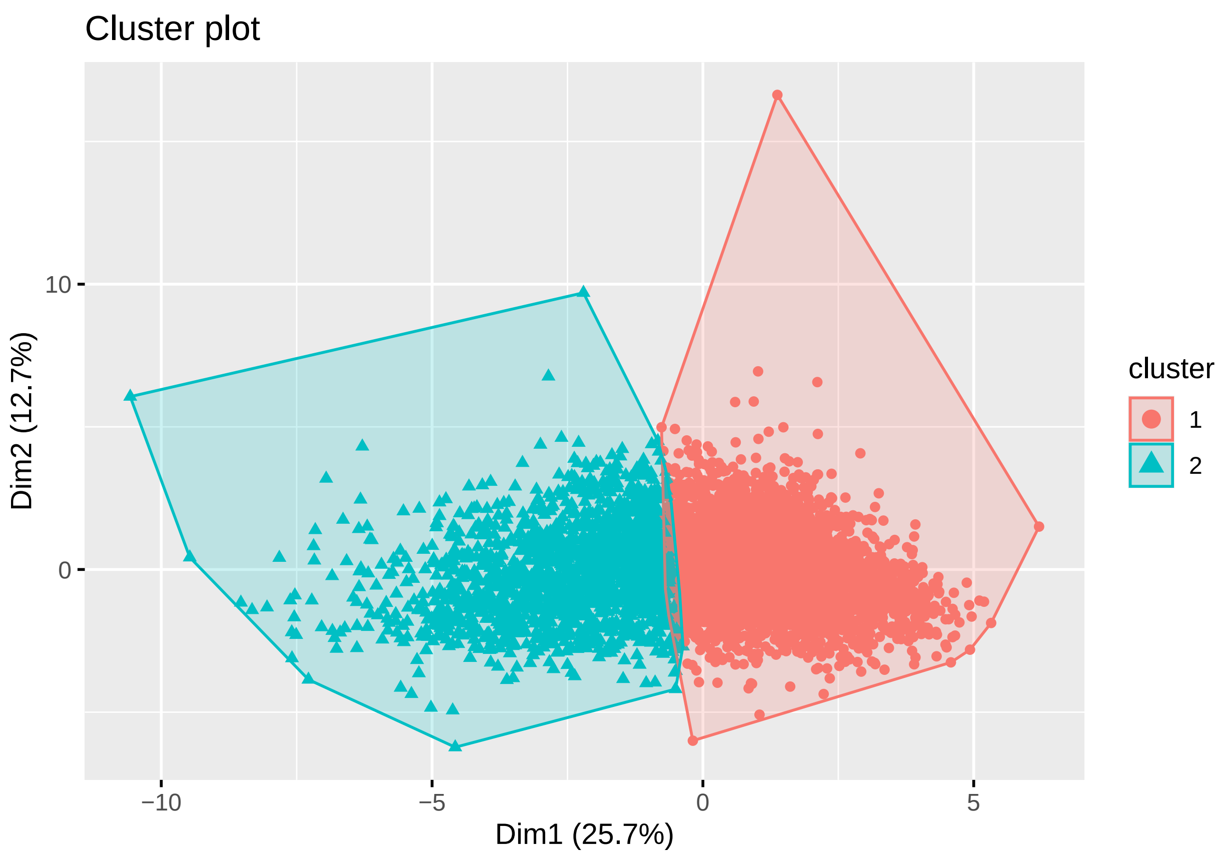



```
visualise_cluster(osward_year, gender_columns, 2)
```




```r
visualise_cluster(msoa_year, gender_columns, 2)
```


```
visualise_cluster(lsoa_year, gender_columns, 2)
```

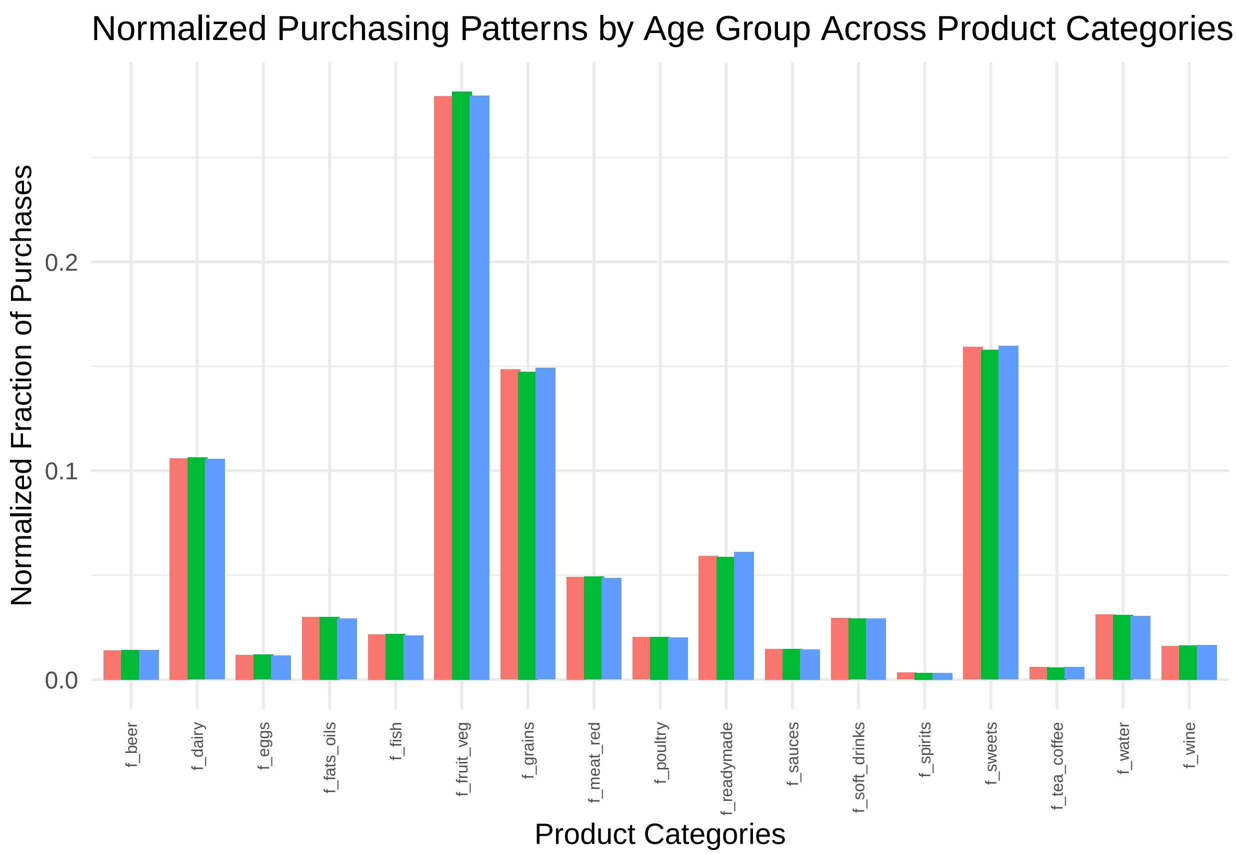


Analysis: - Cluster seems to shows it is worth investigating as there are distinct purchasing patterns shown for each age groups as well as gender. Elbow analysis shows that 3 clusters might be optimal for age groups, while 2 or 3 clusters might be optimal for gender (we will choose 2 in this analysis). - We will then investigate the purchasing patterns using normalisation to understand the preferences of each group.

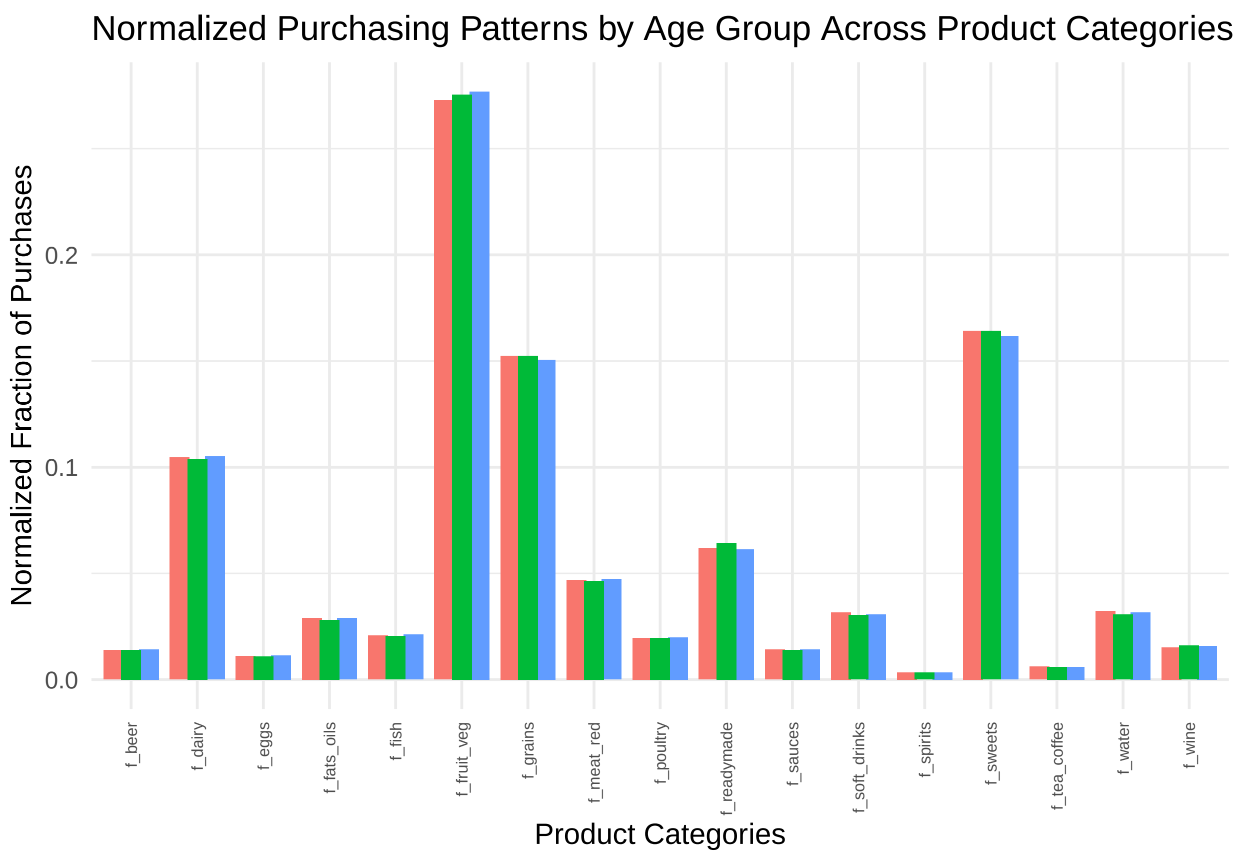
```
# Purchase preference for each age group
# Function to calculate normalized purchase sums by age group and plot the data
calculate_and_plot_purchases <- function(data, product_categories, age_columns) {
  # Calculating the sum of purchases for each product category by age group
  purchase_sums_by_age <- lapply(age_columns, function(age) {
    colSums(data[product_categories] * data[[age]], na.rm = TRUE)
  })
  names(purchase_sums_by_age) <- age_columns

  # Normalizing these sums by the total count for each age group
  normalized_purchases <- lapply(names(purchase_sums_by_age), function(age) {
    purchase_sums_by_age[[age]] / sum(data[[age]], na.rm = TRUE)
  })

  # Transforming the data for visualization
  normalized_purchases_df <- as.data.frame(normalized_purchases)
  rownames(normalized_purchases_df) <- product_categories
  normalized_purchases_df <- normalized_purchases_df %>%
    tibble::rownames_to_column(var = "Product")

  melted_data_age <- normalized_purchases_df %>%
```

```
    pivot_longer(cols = -Product, names_to = "Age_Group", values_to = "Fraction")

  # Plotting the data
  ggplot(melted_data_age, aes(x = Product, y = Fraction, fill = Age_Group)) +
    geom_bar(stat = "identity", position = position_dodge(width = 0.8)) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 6)) +
    labs(x = "Product Categories", y = "Normalized Fraction of Purchases", fill = "Age Group") +
    ggtitle("Normalized Purchasing Patterns by Age Group Across Product Categories") +
    guides(fill = FALSE)
}

# Function to calculate and plot normalized purchase sums by gender
calculate_and_plot_purchases_gender <- function(data, product_categories) {
  # Calculating the sum of purchases for each product category by gender
  purchase_sums_by_gender <- list(
    male = colSums(data[product_categories] * data[['male']], na.rm = TRUE),
    female = colSums(data[product_categories] * data[['female']], na.rm = TRUE)
  )

  # Normalizing these sums by the total count for each gender
  normalized_purchases_gender <- list(
    male = purchase_sums_by_gender$male / sum(data$male, na.rm = TRUE),
    female = purchase_sums_by_gender$female / sum(data$female, na.rm = TRUE)
  )

  # Transforming the data for visualization
  normalized_purchases_gender_df <- as.data.frame(normalized_purchases_gender)
  rownames(normalized_purchases_gender_df) <- product_categories
  normalized_purchases_gender_df <- normalized_purchases_gender_df %>%
    tibble::rownames_to_column(var = "Product")

  melted_data_gender <- normalized_purchases_gender_df %>%
    pivot_longer(cols = -Product, names_to = "Gender", values_to = "Fraction")

  # Plotting the data
  ggplot(melted_data_gender, aes(x = Product, y = Fraction, fill = Gender)) +
    geom_bar(stat = "identity", position = position_dodge()) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    labs(title = "Normalized Purchasing Patterns by Gender Across Product Categories",
         x = "Product Categories", y = "Normalized Fraction of Purchases",
         fill = "Gender") +
    guides(fill = FALSE)
}

calculate_and_plot_purchases(borough_year, product_categories, age_columns)
```

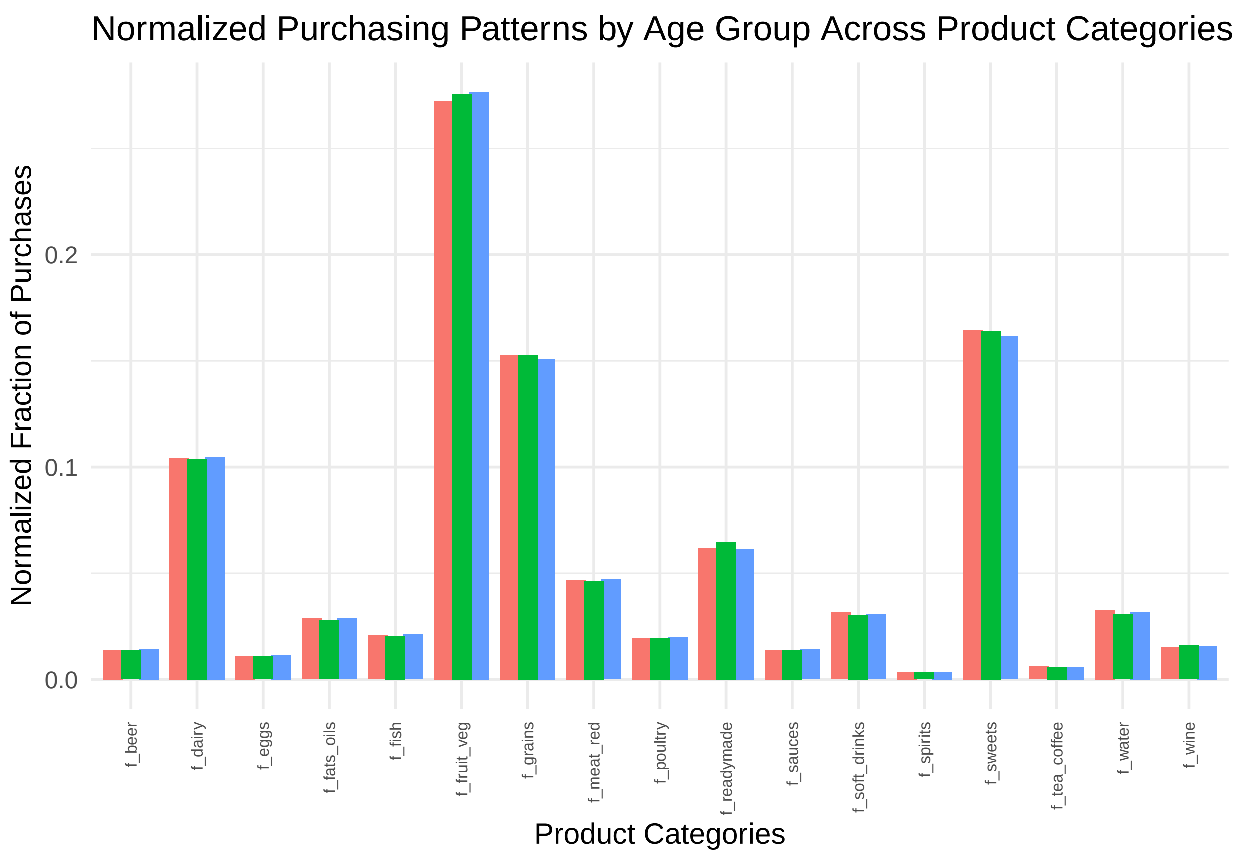
```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



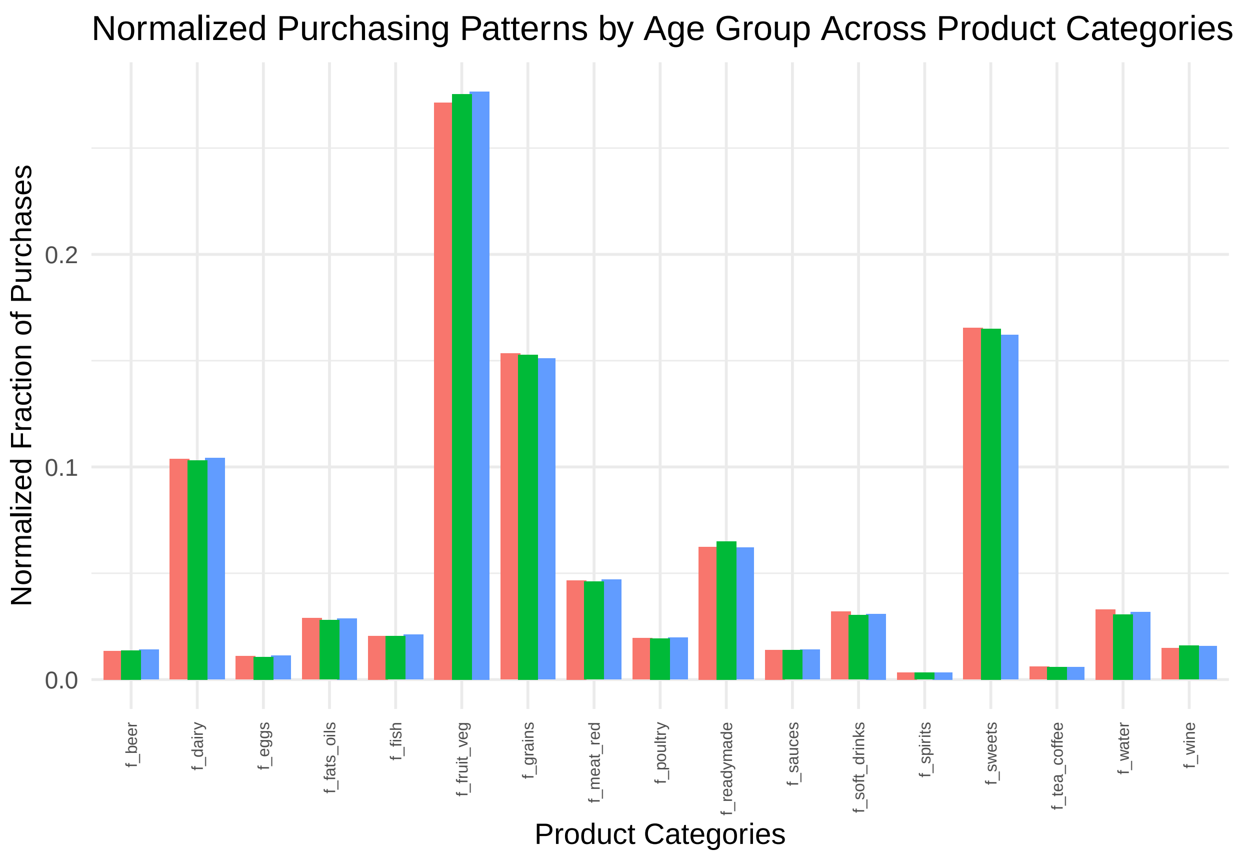

```
calculate_and_plot_purchases(osward_year, product_categories, age_columns)
```



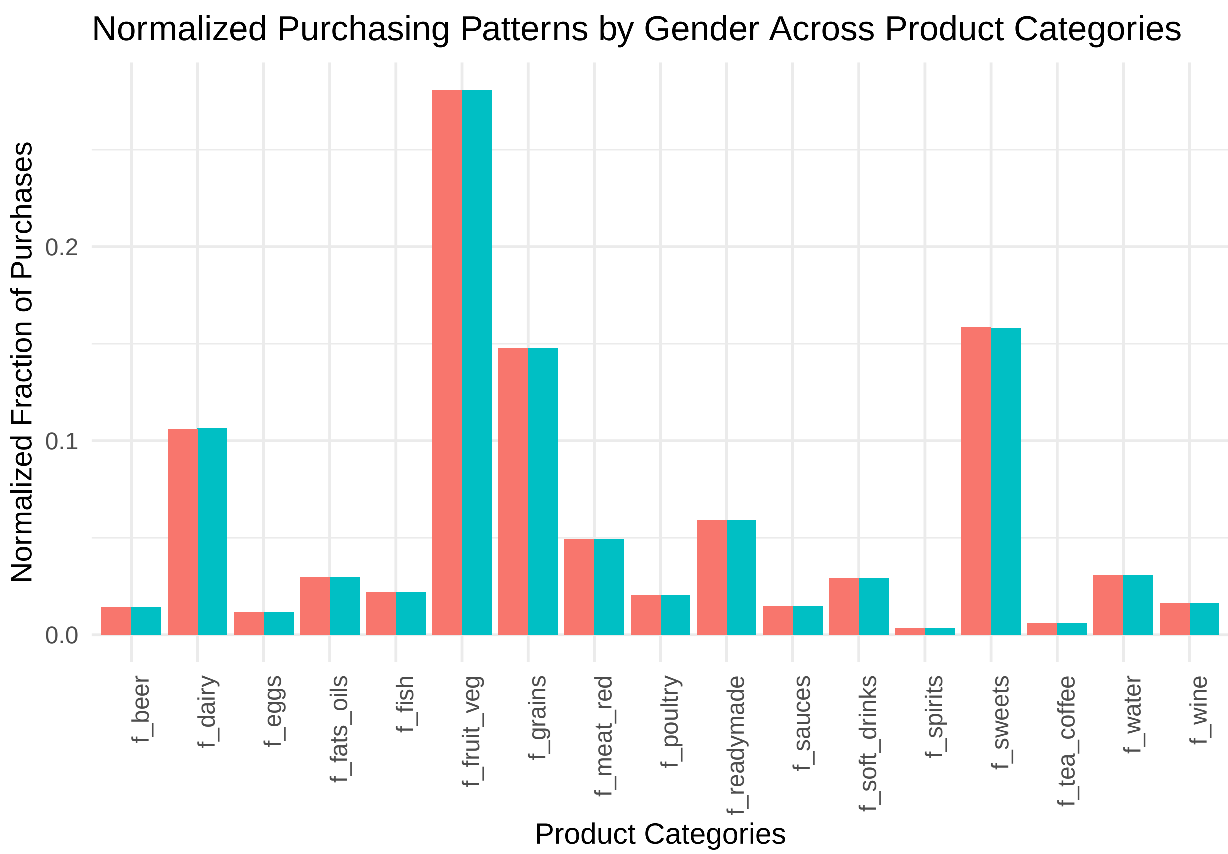

```
calculate_and_plot_purchases(msoa_year, product_categories, age_columns)
```



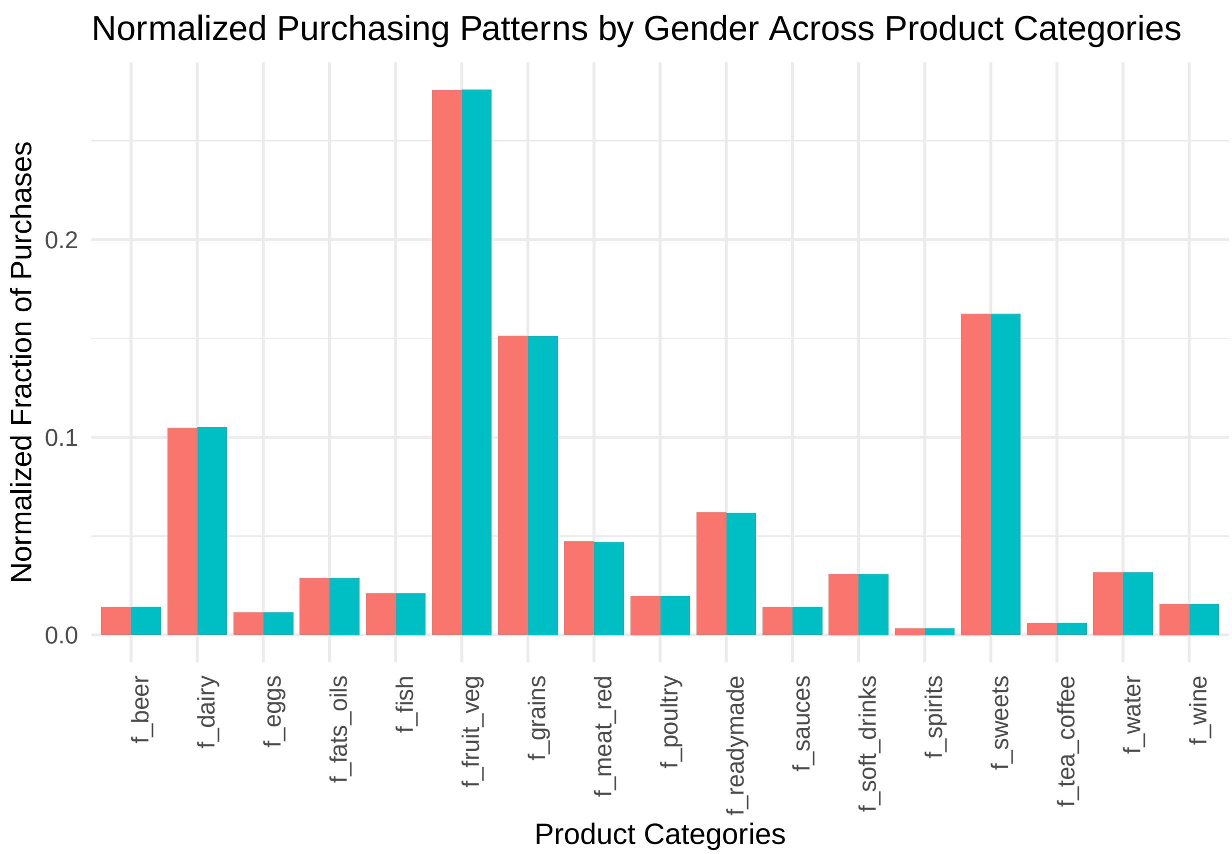

```
calculate_and_plot_purchases(lsoa_year, product_categories, age_columns)
```



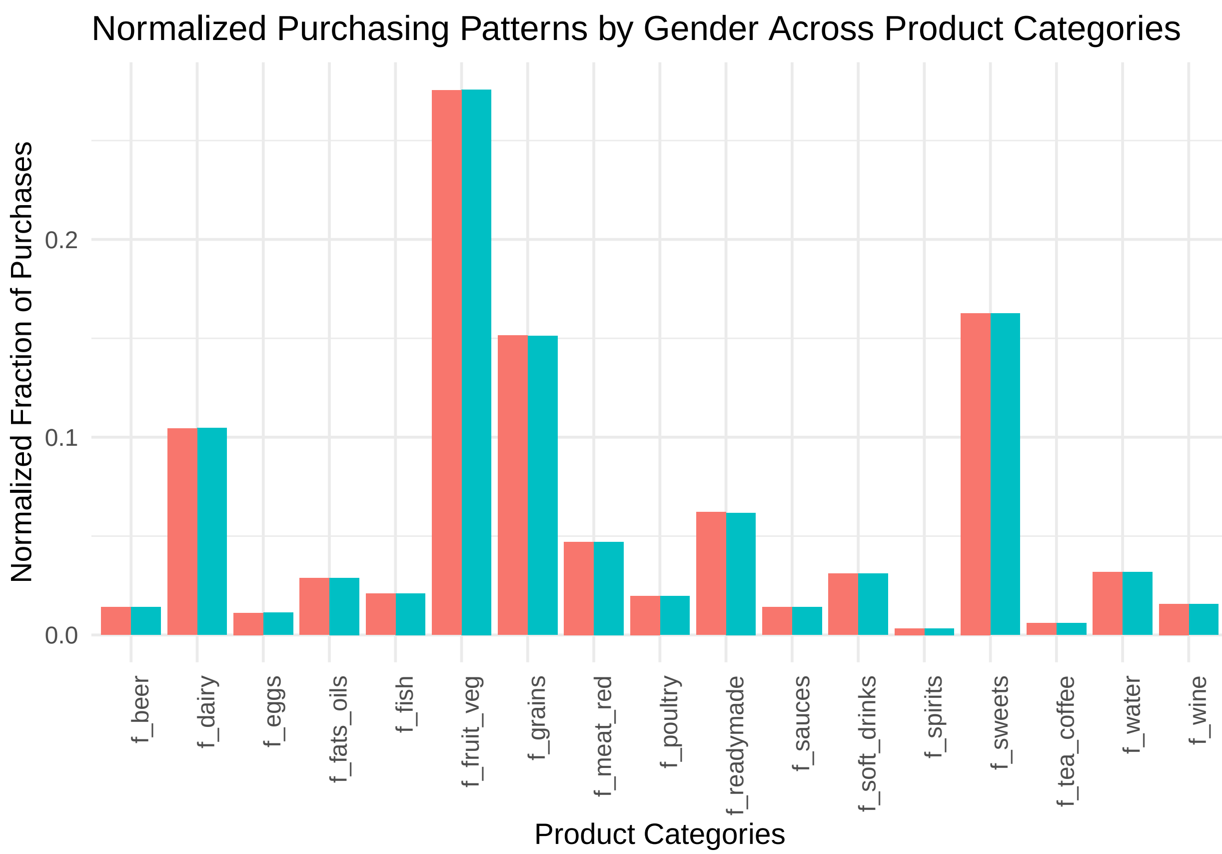

```
calculate_and_plot_purchases_gender(borough_year, product_categories)
```



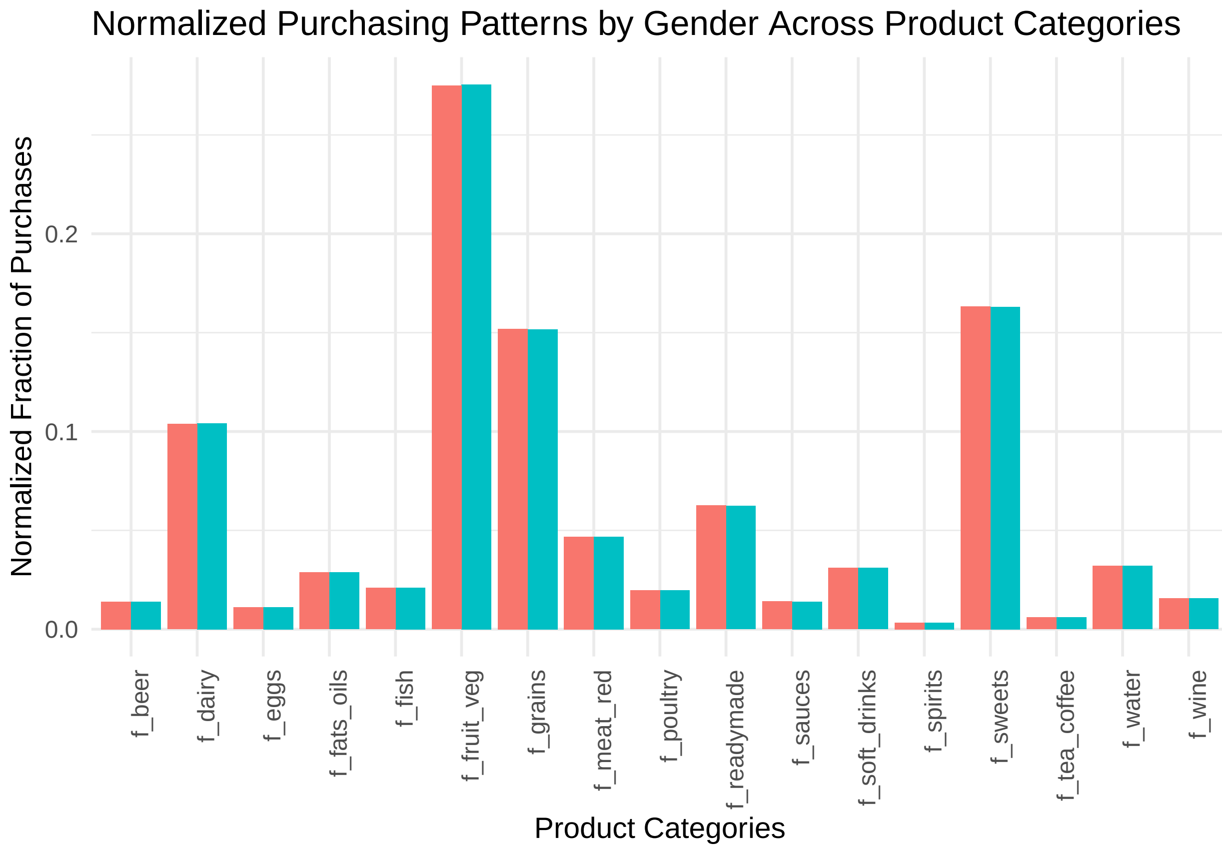

```
calculate_and_plot_purchases_gender(osward_year, product_categories)
```




```
calculate_and_plot_purchases_gender(msoa_year, product_categories)
```




```
calculate_and_plot_purchases_gender(lsoa_year, product_categories)
```



Key Observations: Age Group Differences: - Younger Age Group (0-17 years): This group shows relatively lower purchasing fractions across most categories, likely reflecting their lesser purchasing power or dependence on adults for buying decisions. Noticeable interests might be in categories like f__sweets and f__soft__drinks. - Middle Age Group (18-64 years): Dominates most categories, reflecting their broader economic activity and varied preferences. This group shows higher fractions in categories like f__beer, f__wine, and f__spirits, which are adult-oriented products. - Older Age Group (65+ years): Shows interest in categories that might be considered necessities or health-oriented, such as f__fruit__veg and f__dairy. There's also a noticeable fraction in f__tea__coffee.

Key Observations: Gender Differences: - Certain categories like beer, spirits, and wine show a higher purchasing fraction among male customers compared to female customers. - Female customers tend to have a higher fraction of purchases in categories like f__fruit__veg, f__dairy, and f__sweets, indicating possible preferences for these items. - Shared Interests: Some categories such as f__soft__drinks and f__tea__coffee appear to have relatively balanced fractions between genders, suggesting these items are universally popular.