Discriminant Analysis

Felipe Montealegre

2022-10-23

Overview of the strategy:

Suppose our client is a Wholesale Distributor that has information on current and past clients expenditure on different products categories. We are required to design a tool to discriminate between smaller clients (i.e., the "Horeca" channel) and bigger clients (i.e., the "Retail" channel) depending on the expenditure per category of products. First, we present a general overview of the theoretical aspects of Discriminant Analysis techniques. Then we present a general overview of the structure of the data. We show the frequency of observations across channels and the distributions of the variables of interest. We study the assumptions of the model (namely normality and homoskedasticity between classes), comment on them, and transform the data accordingly for the analysis. After that, we perform discriminant analysis using Linear Discriminant Analysis, Quadratic Discriminant Analysis and Bayes Discriminant Analysis. We compare different measures of accuracy across models and recommend one of them based on the results.

Descriptive statistics

The table below contains the descriptive statistics of the dataset. [*** general comment on the min max means variances].

```
##          class region        fresh         milk      grocery       frozen
## nbr.val     NA     NA 4.400000e+02 4.400000e+02 4.400000e+02 4.400000e+02
## nbr.null    NA     NA 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## nbr.na      NA     NA 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## min         NA     NA 3.000000e+00 5.500000e+01 3.000000e+00 2.500000e+01
## max         NA     NA 1.121510e+05 7.349800e+04 9.278000e+04 6.086900e+04
## range       NA     NA 1.121480e+05 7.344300e+04 9.277700e+04 6.084400e+04
## sum         NA     NA 5.280131e+06 2.550357e+06 3.498562e+06 1.351650e+06
## median      NA     NA 8.504000e+03 3.627000e+03 4.755500e+03 1.526000e+03
## mean        NA     NA 1.200030e+04 5.796266e+03 7.951277e+03 3.071932e+03
## SE.mean     NA     NA 6.029377e+02 3.518457e+02 4.530455e+02 2.314375e+02
## CI.mean     NA     NA 1.185003e+03 6.915113e+02 8.904077e+02 4.548631e+02
## var         NA     NA 1.599549e+08 5.446997e+07 9.031010e+07 2.356785e+07
## std.dev     NA     NA 1.264733e+04 7.380377e+03 9.503163e+03 4.854673e+03
## coef.var    NA     NA 1.053918e+00 1.273299e+00 1.195174e+00 1.580332e+00
##          detergents_paper delicatessen
## nbr.val      4.400000e+02 4.400000e+02
## nbr.null     0.000000e+00 0.000000e+00
## nbr.na       0.000000e+00 0.000000e+00
## min          3.000000e+00 3.000000e+00
## max          4.082700e+04 4.794300e+04
## range        4.082400e+04 4.794000e+04
## sum          1.267857e+06 6.709430e+05
## median       8.165000e+02 9.655000e+02
```

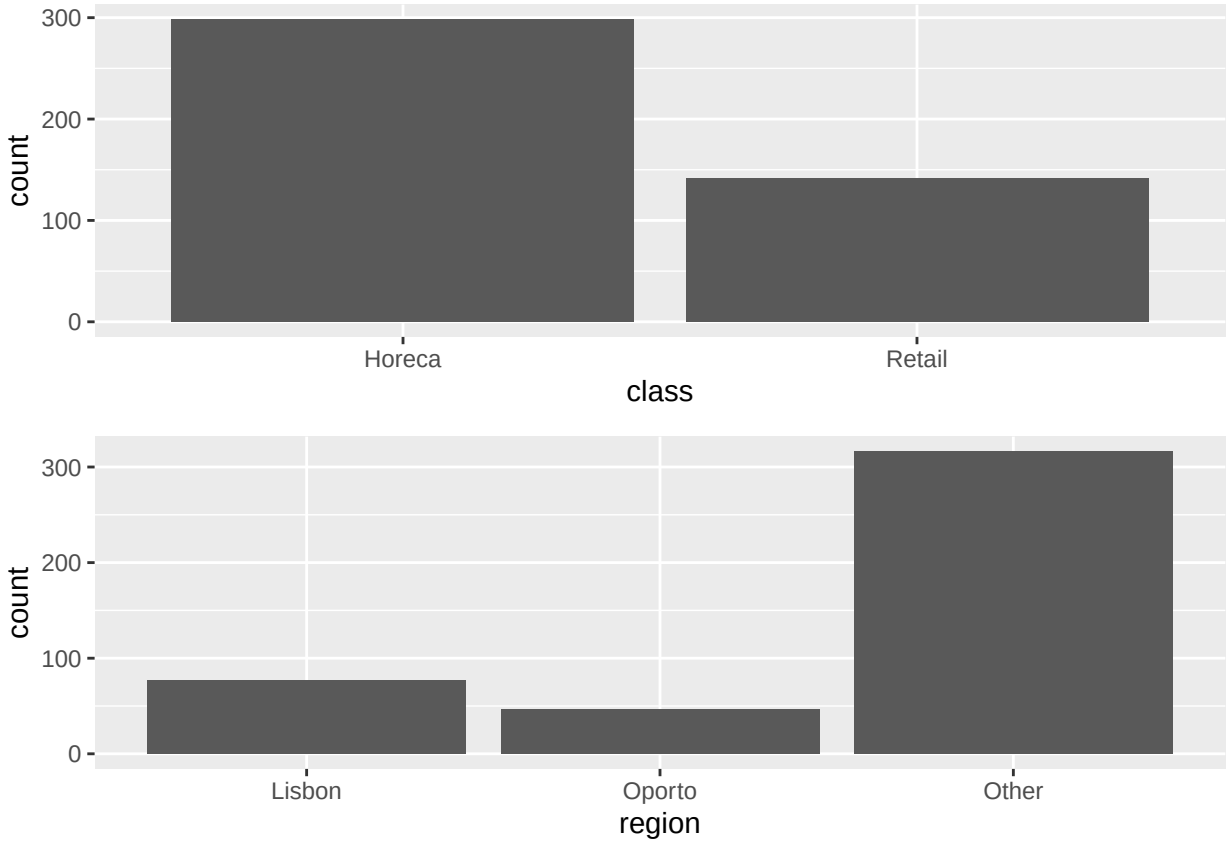
```
## mean         2.881493e+03 1.524870e+03
## SE.mean      2.272985e+02 1.344433e+02
## CI.mean      4.467286e+02 2.642325e+02
## var          2.273244e+07 7.952997e+06
## std.dev      4.767854e+03 2.820106e+03
## coef.var     1.654647e+00 1.849407e+00
```

The categorical variables are class and region, both of them denoting the type of client (i.e., if the client is a hotel, restaurant, or café ("horeca")) and the region where the client is based (i.e., Lisbon, Oporto, or other region).67% of the clients belong to the "Horeca" category, and 32% belong to the retail channel. 17% of the clients are located in Lisbon, 11 percent of them are located in Oporto, and the remaining 72% are located in a different location.

```
##
## Horeca Retail
##    298    142
```

```
## data$class :
##         Frequency Percent Cum. percent
## Horeca        298    67.7         67.7
## Retail        142    32.3        100.0
##   Total       440   100.0        100.0
```

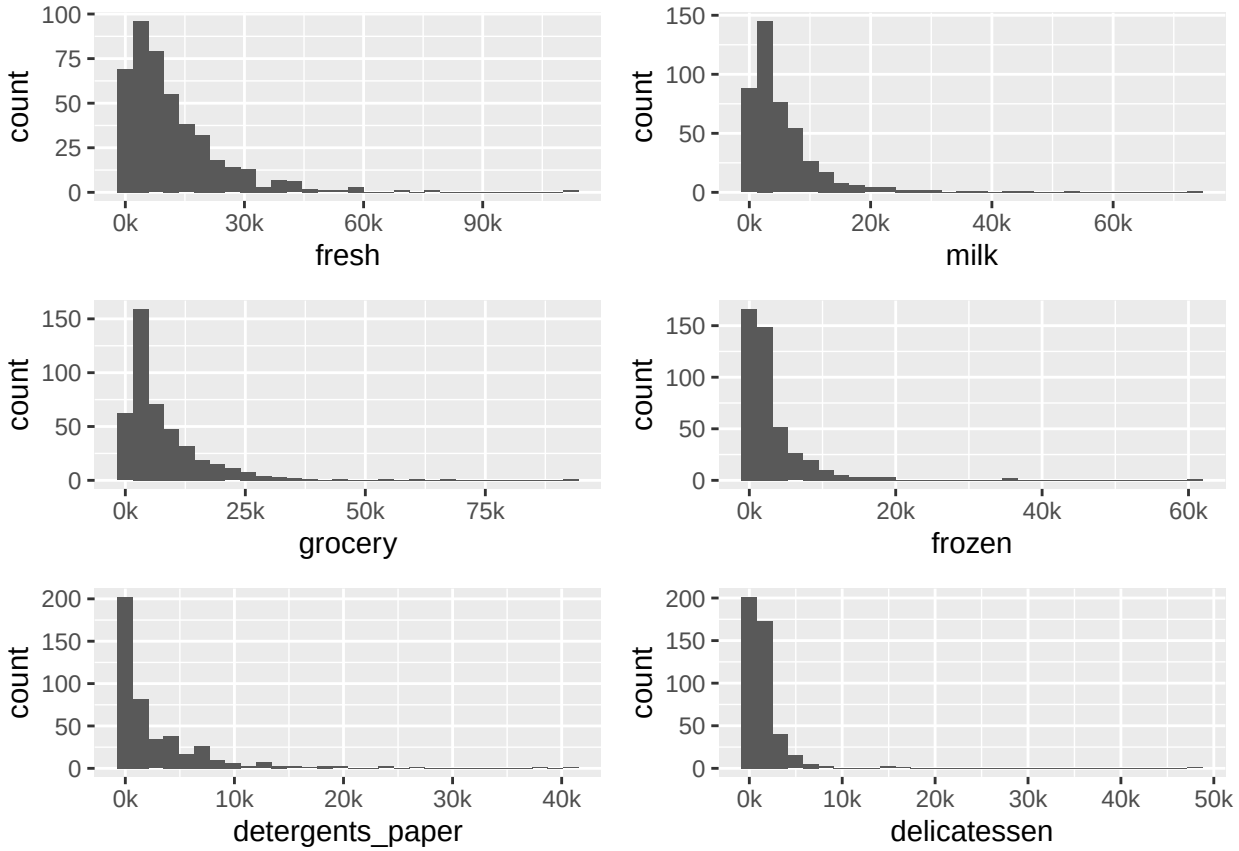
```
## data$region :
##         Frequency Percent Cum. percent
## Other         316    71.8         71.8
## Lisbon         77    17.5         89.3
## Oporto         47    10.7        100.0
##   Total       440   100.0        100.0
```



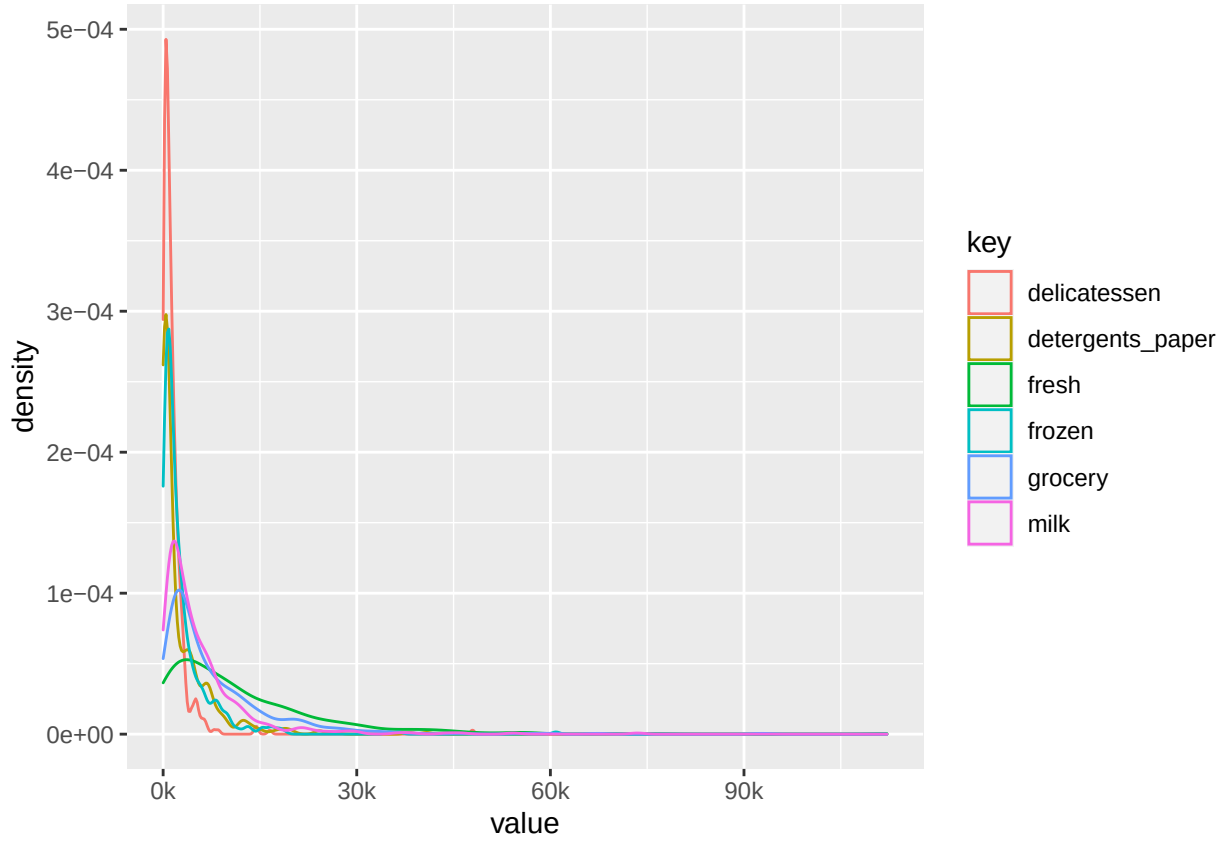
The covariates contain information about expenditure in monetary units across 6 different categories:

1. Fresh: annual spending on fresh products.
2. Milk: annual spending on milk products.
3. Grocery: annual spending on grocery products.
4. Frozen: annual spending on frozen products.
5. Detergents_paper: annual spending on detergents and paper products.
6. Delicatessen: annual spending (m.u.) on and delicatessen products.

The figure below displays the distributions of each of the covariates. In general, most clients spend between 0k and 30k monetary units a year on all categories. Some clients spend much more than the average, with annual expenditures ranging between 30k and 100k sin some categories. Expenditure on detergents and/or paper, frozen products, and delicatessen products ranges from 0k to 10k approximately. This is expected as they are either luxury goods or goods which frequency is not as high as basic needs food such as fresh products, milk and grocery products in general.



The figure below condenses the distributional information across covariates for ease of interpretation.



Testing the assumptions of the model on the data

LDA assumes that the observations come from a normally distributed DGP with constant variance between classes. This generates decision rules that are linear in the covariates (i.e., it is possible to discriminate observations by a line or a plane, depending on the number of classes). If the homoskedasticity assumption is violated, then it is better to apply QDA. QDA generates decision rules by weighting . ***Bayes***

We first explore variance equality between individual covariates using Bartlett's test for equal variance across samples for individual variables. Then we perform Cai TT, Ma Z (2013) two sample test of equality of covariance matrices. Bartlett's test strongly rejects the null hypothesis of equality of variance between groups for each of the covariates. This gives us a hit that for each of the covariates, the distribution of the data is quite different across classes. For the Cai and Ma test we strongly reject the null hypothesis that the covariance matrices between classes is equal.

Bartlett's test:

```
## # A tibble: 6 x 2
##   bart_names       bart_pvalues
##   <chr>                   <dbl>
## 1 fresh                2.15e-  8
## 2 milk                 7.48e- 31
## 3 grocery              1.33e- 69
## 4 frozen               5.50e- 39
## 5 detergents_paper     1.72e-124
## 6 delicatessen         1.72e-124
```

Cai TT, Ma Z test:

```
## $statistic
## [1] 12.10429
##
## $threshold
## [1] 9.300059
##
## $reject
## [1] TRUE
```

Next, we explore the normality assumptions. First, we compute the Shapiro-Wilk test of normality for each of the covariates individually. For all the individual covariates we strongly reject the null hypothesis of normality. We then perform a Generalized Shapiro-Wilk test for Multivariate Normality. We again strongly reject the null hypothesis of joint normality in the sample.

Shaphiro wilk test:

```
## # A tibble: 6 x 2
##   shap_names       shap_pvalues
##   <chr>                   <dbl>
## 1 fresh                7.92e-24
## 2 milk                 9.76e-30
## 3 grocery              3.91e-28
## 4 frozen               1.29e-32
## 5 detergents_paper     1.91e-30
## 6 delicatessen         1.91e-30
```

Generalized Shapiro-Wilk test

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  .
## MVW = 0.6652, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.36823, p-value < 2.2e-16
```

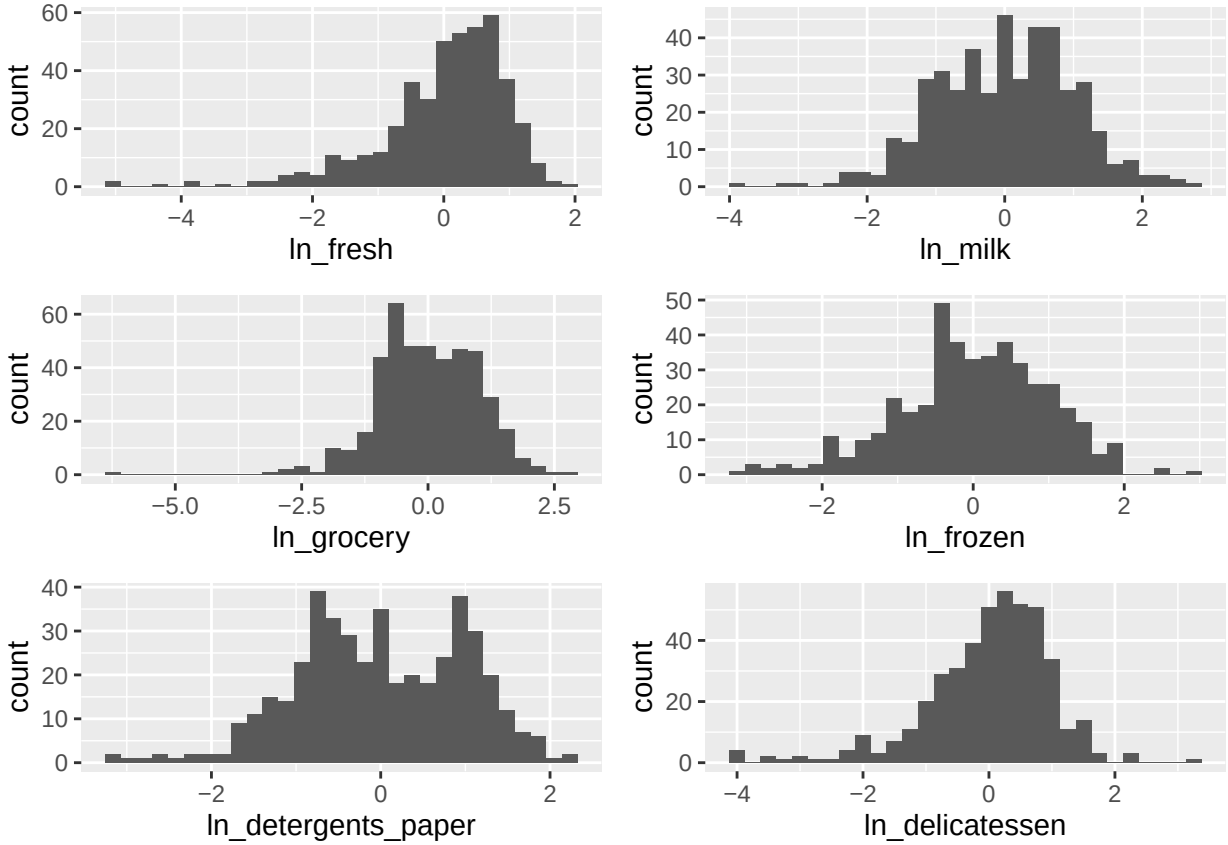
```
##          Beta-hat      kappa p-val
## Skewness 230.1227 16875.6651     0
## Kurtosis 416.4123   394.3618     0
```

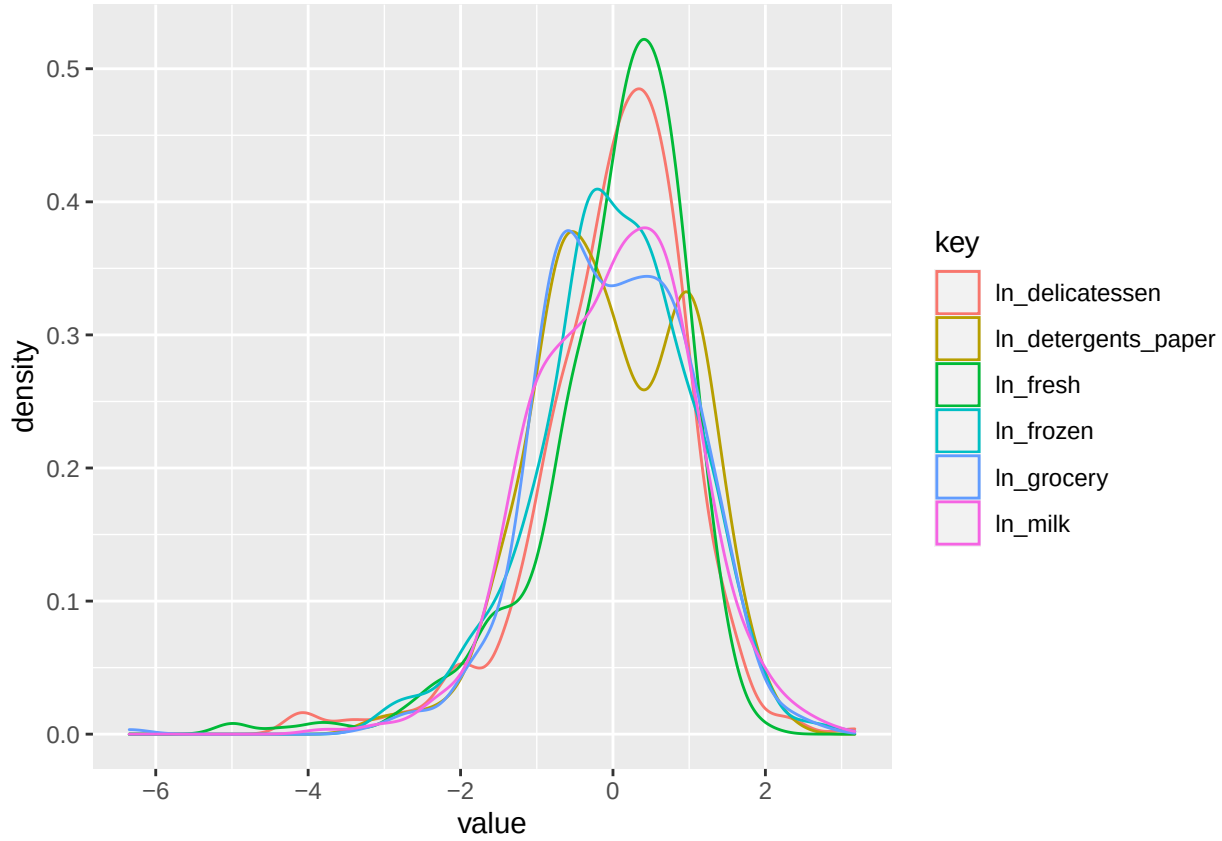
We decide to apply a log transformation and standardize the data for ease of manipulation and to comply with the model assumptions. It is worth emphasizing that QDA does not require the assumption of equality of covariances between classes, and that Bayes Discriminant rule does not require the assumption of normality in the data. As expected, once transformed and standardized, the mean and variance of the transformed covariates are equal to 0 and 1 respectively.

```
##                  ln_fresh       ln_milk    ln_grocery     ln_frozen
## nbr.val       4.400000e+02  4.400000e+02  4.400000e+02  4.400000e+02
## nbr.null      0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
## nbr.na        0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
## min          -4.995530e+00 -3.790609e+00 -6.347964e+00 -3.155526e+00
## max           1.968421e+00  2.853334e+00  2.695213e+00  2.896796e+00
## range         6.963952e+00  6.643943e+00  9.043176e+00  6.052322e+00
```

```
## sum           1.487092e-14 -3.172480e-13  2.301878e-13  1.294988e-13
## median        2.145970e-01  6.923666e-02  2.254774e-02  2.177675e-02
## mean          3.383931e-17 -7.206887e-16  5.227853e-16  2.942505e-16
## SE.mean       4.767313e-02  4.767313e-02  4.767313e-02  4.767313e-02
## CI.mean.0.95  9.369593e-02  9.369593e-02  9.369593e-02  9.369593e-02
## var           1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
## std.dev       1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
## coef.var      2.955143e+16 -1.387562e+15  1.912831e+15  3.398465e+15
##               ln_detergents_paper ln_delicatessen
## nbr.val              4.400000e+02    4.400000e+02
## nbr.null             0.000000e+00    0.000000e+00
## nbr.na               0.000000e+00    0.000000e+00
## min                 -3.161992e+00   -4.084207e+00
## max                  2.237671e+00    3.173741e+00
## range                5.399663e+00    7.257948e+00
## sum                 -4.451474e-14   -1.494378e-14
## median              -5.003716e-02    1.565625e-01
## mean                -1.015064e-16   -3.386653e-17
## SE.mean              4.767313e-02    4.767313e-02
## CI.mean.0.95         9.369593e-02    9.369593e-02
## var                  1.000000e+00    1.000000e+00
## std.dev              1.000000e+00    1.000000e+00
## coef.var            -9.851596e+15   -2.952768e+16
```

The distribution of transformed variables now resembles a normal distribution:





Nonetheless, the same normality tests as before are performed on the transformed data. The proportion of observations in each class should be considered when creating the priors. If not and the probabilities of classification do not match the class proportions, the classification procedure will be erroneous. Even though for two out of the six variables the hypothesis of homoskedasticity is still dejected, the generalized two-sample test of equality of covariances strongly rejects the null hypothesis of homoskedasticity.

Bartlett's test:

```
## # A tibble: 6 x 2
##   bart_names          bart_pvalues
##   <chr>                      <dbl>
## 1 ln_fresh                   0.211
## 2 ln_milk                    0
## 3 ln_grocery                 0
## 4 ln_frozen                  0.428
## 5 ln_detergents_paper        0
## 6 ln_delicatessen            0
```

Cai TT, Ma Z test:

```
## $statistic
## [1] 29.43113
##
## $threshold
## [1] 9.300059
##
## $reject
## [1] TRUE
```

A similar story is told for the normality tests. The hypothesis of normality, both at the individual variable level and as a joint distribution of the covariates is strongly rejected. In a nutshell, the data is neither normally distributed nor homoscedastic across classes.

Shaphiro wilk test:

```
## # A tibble: 6 x 2
##   shap_names          shap_pvalues
##   <chr>                      <dbl>
## 1 ln_fresh                7.10e-17
## 2 ln_milk                 9.04e- 2
## 3 ln_grocery              3.54e- 8
## 4 ln_frozen               5.46e- 3
## 5 ln_detergents_paper     3.62e- 5
## 6 ln_delicatessen         3.62e- 5
```

Generalized Shaphiro wilk test:

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  .
## MVW = 0.94112, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.9031, p-value = 4.065e-16
```

```
##          Beta-hat     kappa p-val
## Skewness 14.02010 1028.1410     0
## Kurtosis 87.89748   42.7077     0
```

Discriminant Analysis

The total number of observations are split into two groups: the "training" group and the "testing" group using a split ratio of 0.65. The training subsample is used to create the discriminant rules while the testing subset is used to determine the different accuracy measures. After splitting the data, it is verified that the proportions of observations in the two classes are maintained both in the training set and in the testing set. The decomposition proportions are presented below.

```
## # A tibble: 3 x 6
##   prop_names       n_values_1 prop_values_1 n_values_2 prop_values_2 total
##   <chr>                 <int>         <dbl>      <int>         <dbl> <int>
## 1 Complete dataset        298         0.677        142         0.323   440
## 2 Training dataset        194         0.678         92         0.322   286
## 3 Testing dataset         104         0.675         50         0.325   154
```

The assumptions analysis in the previous code suggested that Bayes DA was more suited for analyzing the data and that QDA better than LDA. The table below presents accuracy measures for each of the three methods for discriminating between classes. Note that these accuracy measures use the training subset. The discriminating method with the highest accuracy rate is LDA (95%) while QDA and BAYES have accuracy rates of 93.7% and 94.4% respectively. The false positive rate is quite low for all discriminating methods, all range between 3% and 5%. Note, however, that because the accuracy is based on the same observations that were used to compute the discriminant rules, this method of evaluation is naively optimistic.

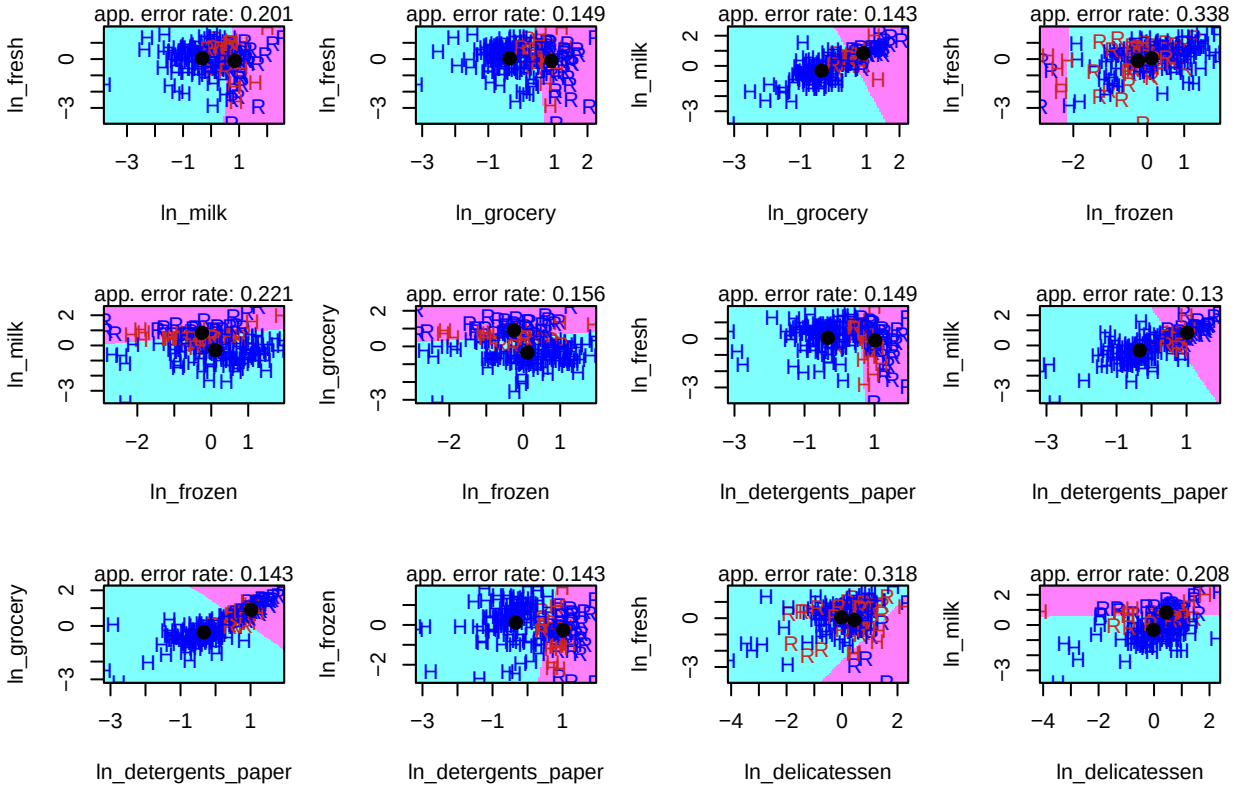
```
## # A tibble: 7 x 4
##   fit_names                   fit_scores_LDA_~ fit_scores_QDA_~ fit_scores_BAYE~
##   <chr>                                  <dbl>            <dbl>            <dbl>
## 1 Accuracy: (TP+TN)/total                0.951            0.937            0.944
## 2 Misclassification Rate: (F~            0.0490           0.0629           0.0559
## 3 Sensitivity, True Positive~            0.935            0.913            0.896
## 4 False Positive Rate: FP/(T~            0.0412           0.0515           0.0316
## 5 Specificity, True Negative~            0.959            0.948            0.968
## 6 Precision: TP/(predicted y~            0.915            0.894            0.935
## 7 Prevalence: (True yes)/tot~            0.322            0.322            0.336
```

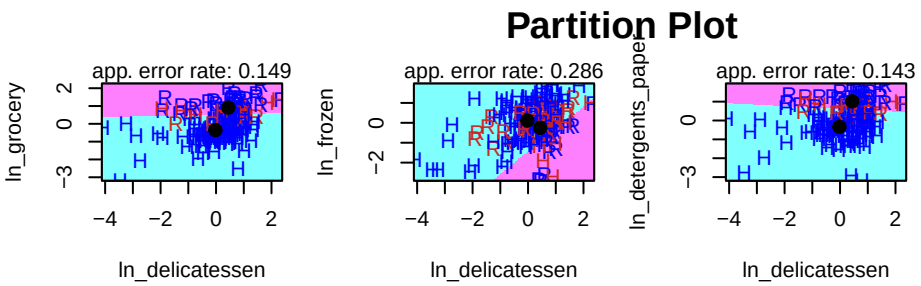
The cross-validation criterion uses the training subset we had previously defined. The table below presents the accuracy measures under the testing subset. The accuracy rate of LDA is around 86% while QDA and BAYES are the same at 87%. The misclassification rates range around 14% to 16% being the highest for the LDA. In general, the measures of accuracy are very close across classification methods which suggest that means between classes are quite different while variance within classes is low. Note also that the similarity of accuracy measures between QDA and BAYES could Indicate that the estimates of the variance/covariance matrices are accurate enough.

```
## # A tibble: 7 x 4
##   fit_names                   fit_scores_LDA_~ fit_scores_QDA_~ fit_scores_BAYE~
##   <chr>                                  <dbl>            <dbl>            <dbl>
## 1 Accuracy: (TP+TN)/total                0.857            0.877            0.877
## 2 Misclassification Rate: (F~            0.143            0.123            0.123
## 3 Sensitivity, True Positive~            0.84             0.88             0.754
## 4 False Positive Rate: FP/(T~            0.135            0.125            0.0430
## 5 Specificity, True Negative~            0.865            0.875            0.957
## 6 Precision: TP/(predicted y~            0.75             0.772            0.92
## 7 Prevalence: (True yes)/tot~            0.325            0.325            0.396
```

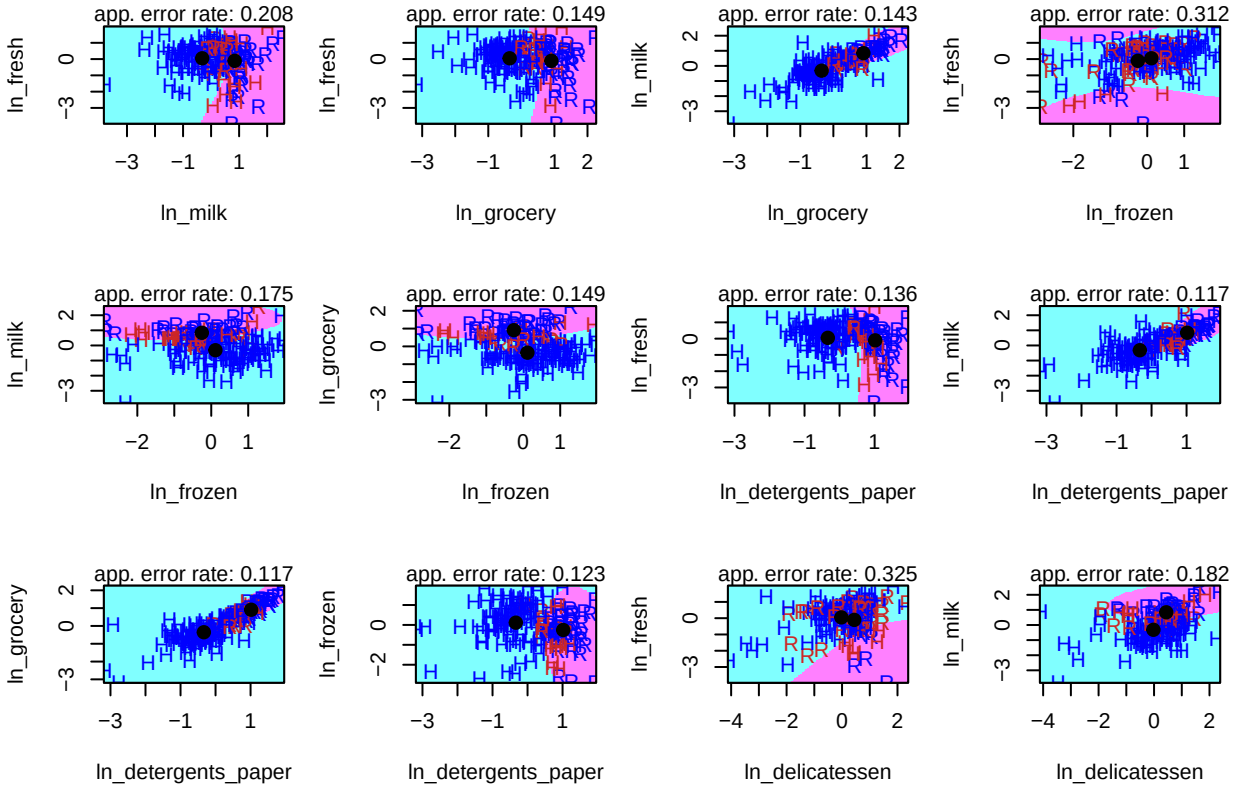
The partitioning rules for each of the partitioning methods are presented below. The boundary between the pink and light blue areas is the partioning rule. The observations marked in Blue are correctly classfied. The observations in red are missclassified. "H" observations are the ones that belong to the "Horeca" class, while "R" observations belong to the "Retail" class.

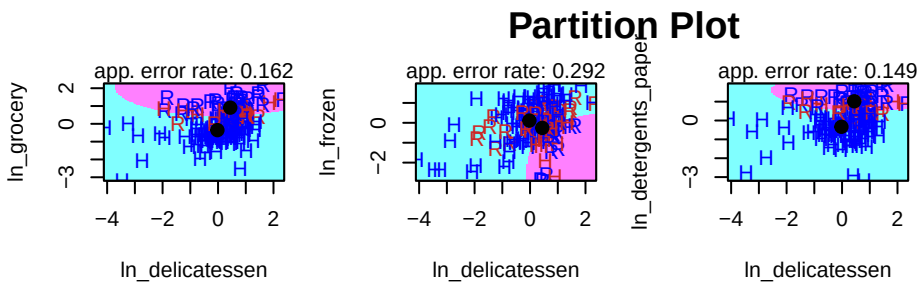
LDA:


app. error rate: 0.201
ln_fresh
ln_milk
app. error rate: 0.149
ln_fresh
ln_grocery
app. error rate: 0.143
ln_milk
ln_grocery
app. error rate: 0.338
ln_fresh
ln_frozen
app. error rate: 0.221
ln_milk
ln_frozen
app. error rate: 0.156
ln_grocery
ln_frozen
app. error rate: 0.149
ln_fresh
ln_detergents_paper
app. error rate: 0.13
ln_milk
ln_detergents_paper
app. error rate: 0.143
ln_grocery
ln_detergents_paper
app. error rate: 0.143
ln_frozen
ln_detergents_paper
app. error rate: 0.318
ln_fresh
ln_delicatessen
app. error rate: 0.208
ln_milk
ln_delicatessen

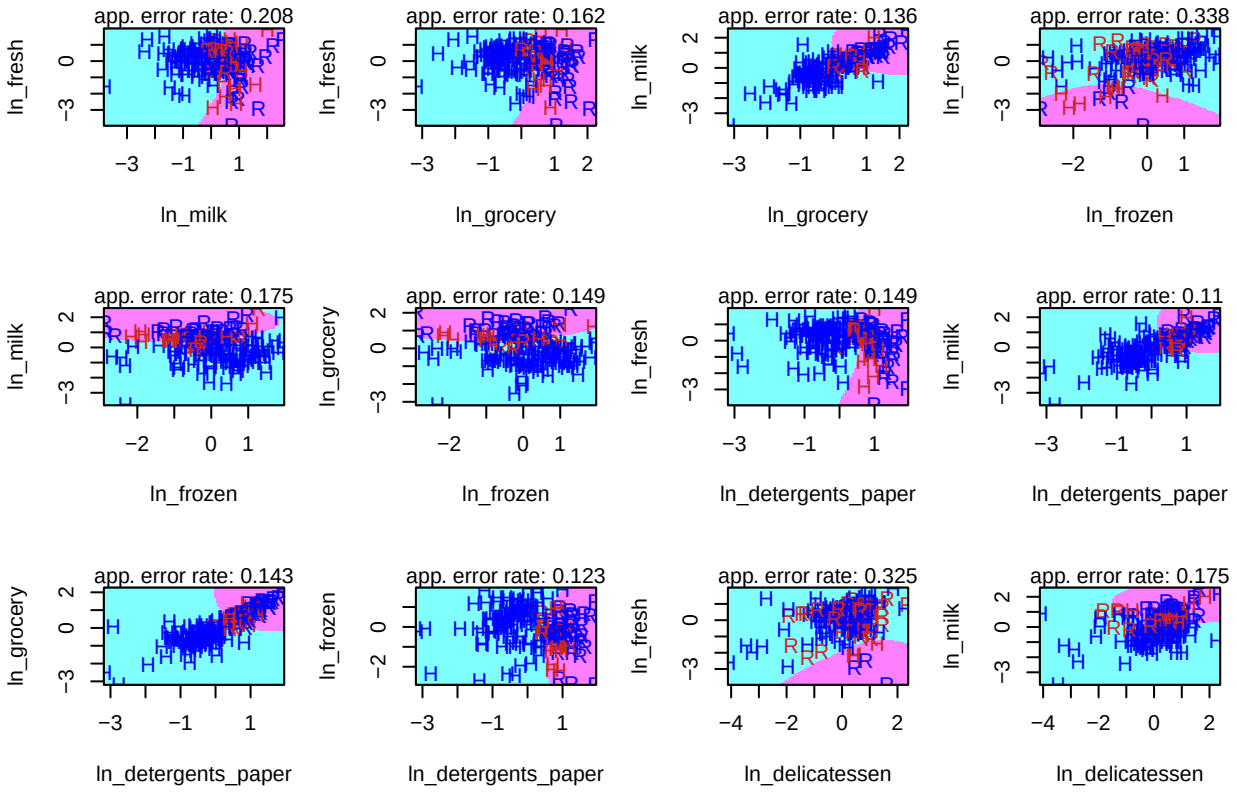

Partition Plot
app. error rate: 0.149
app. error rate: 0.286
app. error rate: 0.143
ln_grocery
ln_frozen
ln_detergents_paper
ln_delicatessen
ln_delicatessen
ln_delicatessen

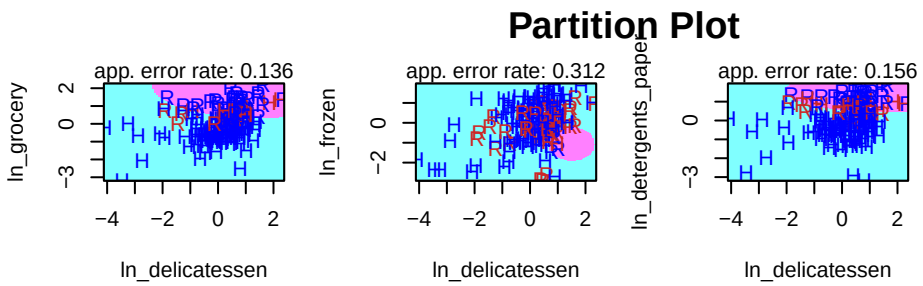
QDA:


app. error rate: 0.208
ln_fresh
ln_milk
app. error rate: 0.149
ln_fresh
ln_grocery
app. error rate: 0.143
ln_milk
ln_grocery
app. error rate: 0.312
ln_fresh
ln_frozen
app. error rate: 0.175
ln_milk
ln_frozen
app. error rate: 0.149
ln_grocery
ln_frozen
app. error rate: 0.136
ln_fresh
ln_detergents_paper
app. error rate: 0.117
ln_milk
ln_detergents_paper
app. error rate: 0.117
ln_grocery
ln_detergents_paper
app. error rate: 0.123
ln_frozen
ln_detergents_paper
app. error rate: 0.325
ln_fresh
ln_delicatessen
app. error rate: 0.182
ln_milk
ln_delicatessen


Partition Plot
app. error rate: 0.162
app. error rate: 0.292
app. error rate: 0.149
ln_grocery
ln_frozen
ln_detergents_paper
ln_delicatessen
ln_delicatessen
ln_delicatessen

Bayes:


app. error rate: 0.208
ln_fresh
ln_milk
app. error rate: 0.162
ln_fresh
ln_grocery
app. error rate: 0.136
ln_milk
ln_grocery
app. error rate: 0.338
ln_fresh
ln_frozen
app. error rate: 0.175
ln_milk
ln_frozen
app. error rate: 0.149
ln_grocery
ln_frozen
app. error rate: 0.149
ln_fresh
ln_detergents_paper
app. error rate: 0.11
ln_milk
ln_detergents_paper
app. error rate: 0.143
ln_grocery
ln_detergents_paper
app. error rate: 0.123
ln_frozen
ln_detergents_paper
app. error rate: 0.325
ln_fresh
ln_delicatessen
app. error rate: 0.175
ln_milk
ln_delicatessen



Additional Analisys

We applied the three methods for discriminating observations to the raw data. This is an additional measure of the efficiency of the models and gives us a hint wheter transforming and standardizing the data affects the results in any way. The table below shows the accuracy rates for the raw data.

```
## # A tibble: 7 x 4
##   fit_names                   fit_scores_LDA_~ fit_scores_QDA_~ fit_scores_BAYE~
##   <chr>                                  <dbl>            <dbl>            <dbl>
## 1 Accuracy: (TP+TN)/total                0.838            0.857            0.864
## 2 Misclassification Rate: (F~            0.162            0.143            0.136
## 3 Sensitivity, True Positive~            0.56             0.72             0.854
## 4 False Positive Rate: FP/(T~            0.0288           0.0769           0.133
## 5 Specificity, True Negative~            0.971            0.923            0.867
## 6 Precision: TP/(predicted y~            0.903            0.818            0.7
## 7 Prevalence: (True yes)/tot~            0.325            0.325            0.266
```

THE END