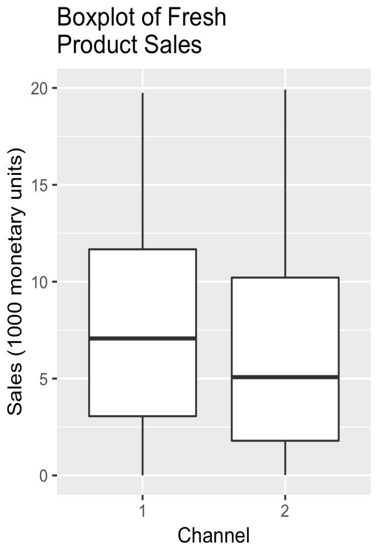
## Introduction:

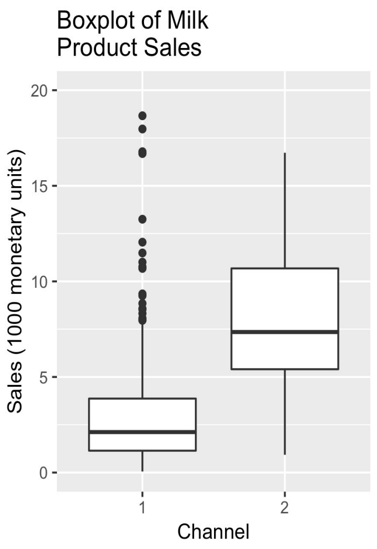
This project will be looking at the wholesale spending data in Portugal taken from the UCI Machine Learning repository. The data comes with 8 features, 6 of which pertain to annual monetary sales of different types of products and 2 categories. One category is called "Region" and denotes which city the wholesale stores customers are located in. The last feature is called Channels, Hotel/Restaurant/Cafe vs Retail, which refers to which type of establishment each observation was. This project will attempt to see if the true average spending of the two channels are different and if these channels can be clustered using linear principal component analysis.

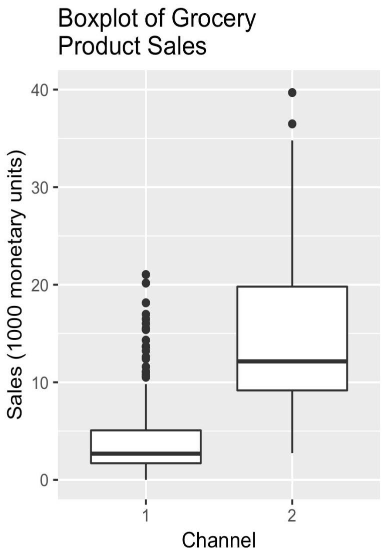
For the following analysis we will assume that our data is multivariate normal and that both channels have the same population covariance. For this project the "Region" category will also be ignored so that the focus can be on the numeric features.

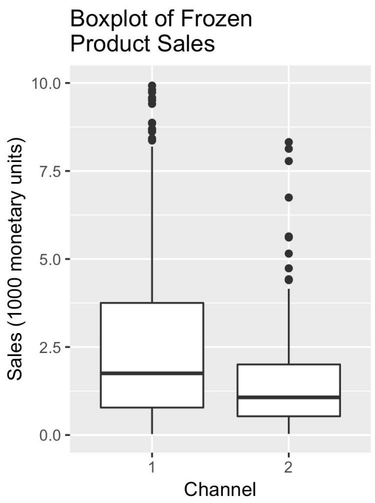
## Summary:

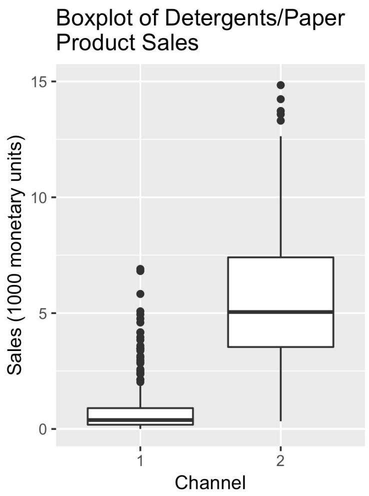
Initial box plots were made to see the spread and comparison of the different features.

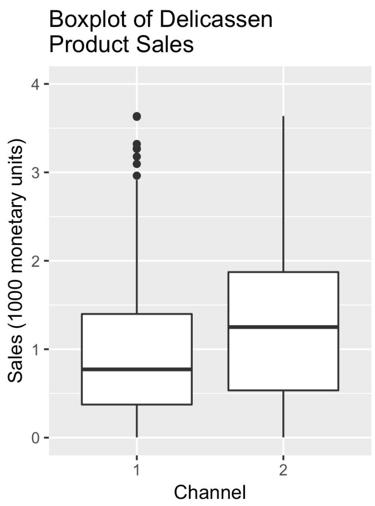

Boxplot of Fresh Product Sales


Boxplot of Milk Product Sales


Boxplot of Grocery Product Sales







Note that these boxplots had their y limits set so that we could better see the spread of the data. We can see that 4 out of 6 of the boxplots are skewed upward. These would be the Milk, Grocery, Frozen, and Detergents/Paper sales. In addition, 3 of these 4 categories have channel one being shown as having a smaller median than channel 2. This could be an indicator that their population means are not equal. However, because of the skew in the data this may not be certain.

## Analysis:

The data was split into two samples, channel 1 (Ho/Re/Ca) and channel 2 (Retail) and the following summary statistics were found:

$n_1 = 298$, $n_2 = 142$

$$\vec{\bar{x}}_1 = \begin{bmatrix} 13.48 \\ 3.45 \\ 3.96 \\ 4.75 \\ 0.79 \\ 1.42 \end{bmatrix} \quad S_1 = \begin{bmatrix} 191.32 & 14.84 & 11.56 & 26.21 & -0.11 & 11.08 \\ 14.84 & 18.94 & 9.30 & 10.04 & 1.20 & 8.62 \\ 11.56 & 9.30 & 2.57 & 5.18 & 2.15 & 4.99 \\ 26.21 & 10.04 & 5.18 & 31.85 & -0.19 & 7.61 \\ -0.11 & 1.20 & 2.15 & -0.19 & 1.22 & 0.27 \\ 11.08 & 8.62 & 4.99 & 7.61 & 0.27 & 9.91 \end{bmatrix}$$

$$\vec{\bar{x}}_2 = \begin{bmatrix} 8.90 \\ 10.72 \\ 16.32 \\ 1.65 \\ 7.27 \\ 1.75 \end{bmatrix} \quad S_2 = \begin{bmatrix} 80.78 & 20.60 & 9.76 & 4.37 & 1.30 & 4.88 \\ 20.60 & 93.70 & 78.21 & 3.06 & 37.87 & 6.49 \\ 20.60 & 93.70 & 78.21 & 3.06 & 37.87 & 6.49 \\ 4.37 & 3.06 & 0.99 & 3.29 & 0.19 & 1.12 \\ 4.37 & 3.06 & 0.99 & 3.29 & 0.19 & 1.12 \\ 4.88 & 6.49 & 3.80 & 1.12 & 0.84 & 3.82 \end{bmatrix}$$

$$S_{pooled} = \begin{bmatrix} 155.73 & 16.69 & 10.98 & 19.18 & 0.34 & 9.09 \\ 155.73 & 16.69 & 10.98 & 19.18 & 0.34 & 9.09 \\ 10.98 & 31.48 & 56.97 & 3.83 & 24.4 & 4.60 \\ 19.18 & 7.80 & 3.83 & 22.66 & -0.07 & 5.52 \\ 0.34 & 13.01 & 24.41 & -0.07 & 13.57 & 0.45 \\ 9.09 & 7.94 & 4.60 & 5.52 & 0.45 & 7.95 \end{bmatrix}$$

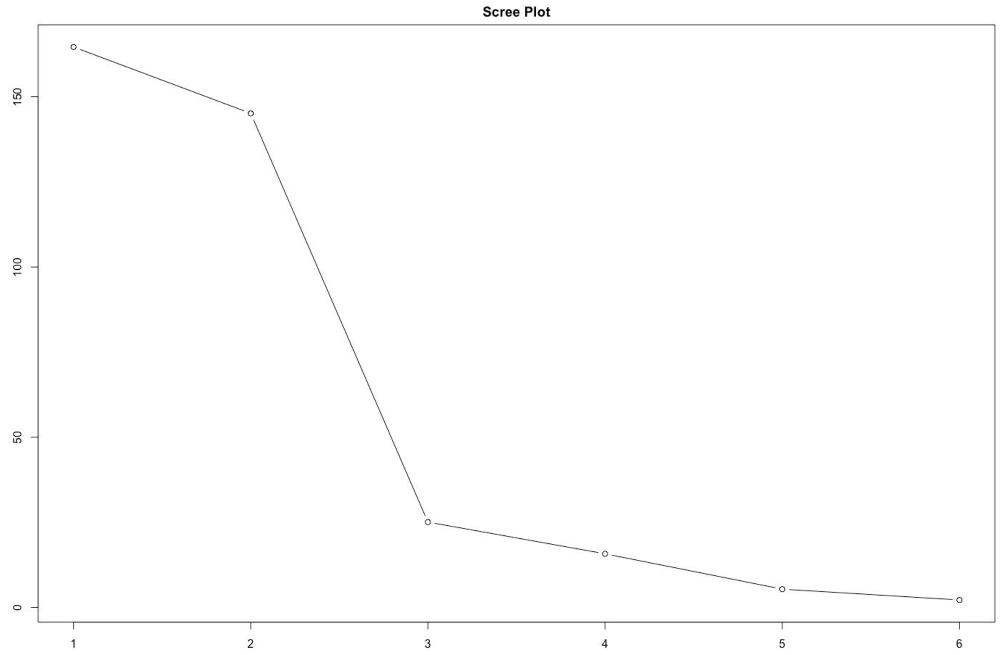
| | Hotelling’s T^2 | | Bonferroni Correction | |
|---|---|---|---|---|
| | Lower Bound | Upper Bound | Lower Bound | Upper Bound |
| Fresh | 0.01 | 9.14 | 1.20 | 7.94 |
| Milk | -9.66 | -4.87 | -9.04 | -5.49 |
| Grocery | -15.12 | -9.60 | -14.40 | -10.32 |
| Frozen | 0.35 | 3.84 | 0.81 | 3.38 |
| Detergents/Paper | -7.83 | -5.13 | -7.47 | -5.48 |
| Delicassen | -1.37 | 0.69 | -1.10 | 0.42 |

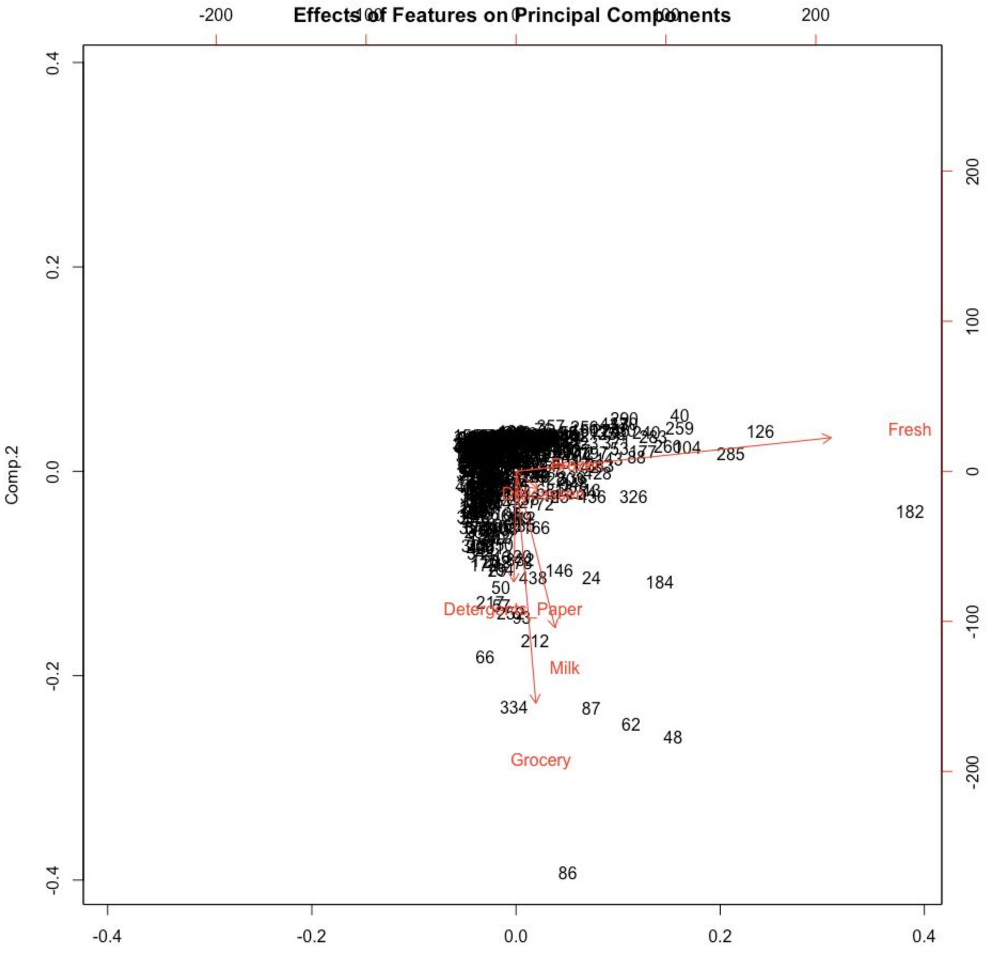
$H_o : \vec{\mu_1} = \vec{\mu_2}$ T^2 = 341.891, $\frac{(438)\,(6)}{435} F_{6,\,436}(.05) = 12.86$, p_val < 0.001

| **Importance of Components** | | | | | | |
|---|---|---|---|---|---|---|
| | Comp 1 | Comp 2 | Comp 3 | Comp 4 | Comp 5 | Comp 6 |
| Standard deviation | 12.83 | 12.05 | 5.01 | 3.97 | 2.32 | 1.48 |
| Proportion of Variance | 0.46 | 0.41 | 0.078 | 0.044 | 0.02 | 0.01 |

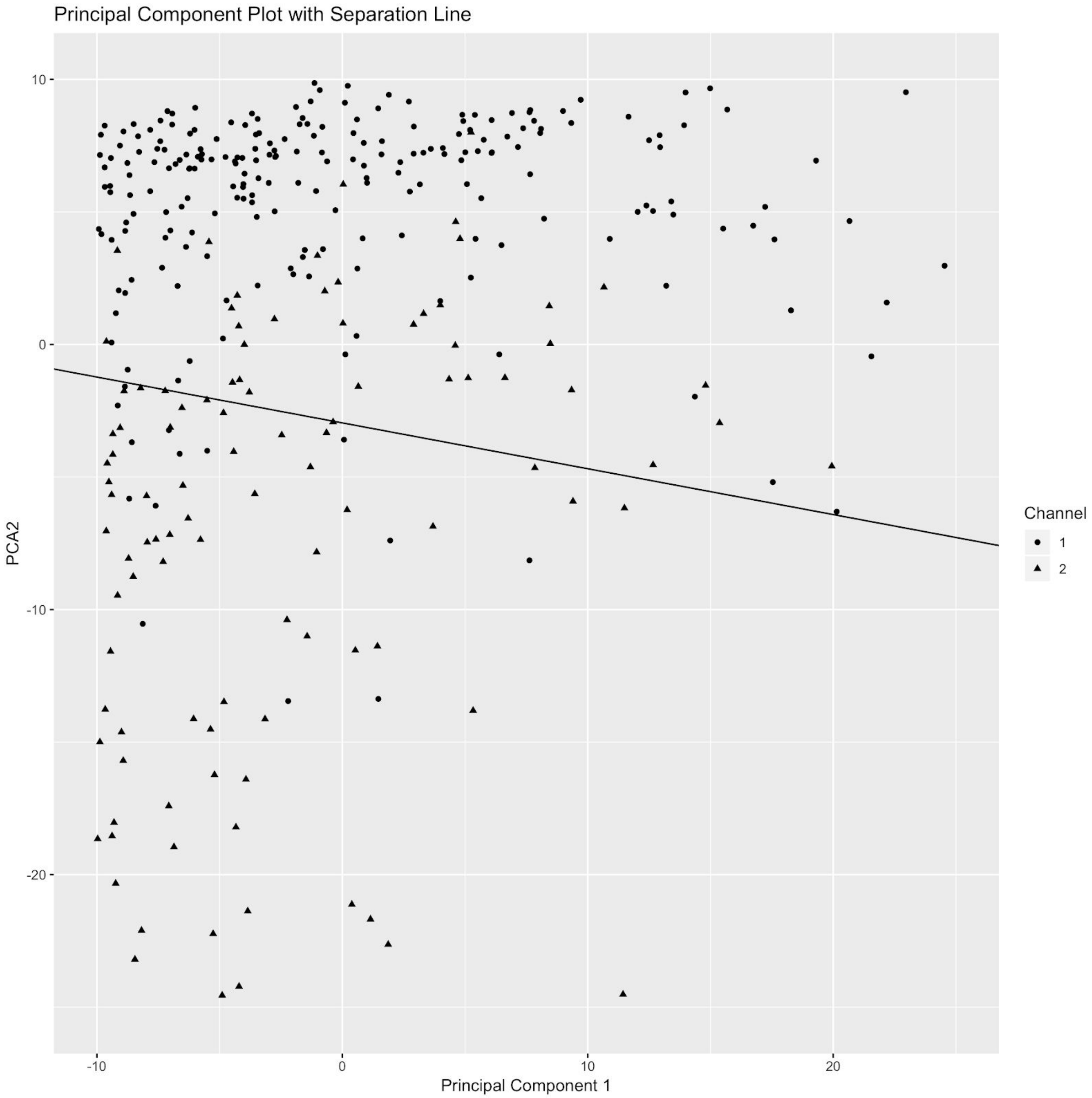
Loadings:

| | Comp 1 | Comp 2 | Comp 3 | Comp 4 | Comp 5 | Comp 6 |
|---|---|---|---|---|---|---|
| Fresh | 0.977 | 0.111 | 0.179 | 0 | 0 | 0 |
| Milk | 0.121 | -0.516 | -0.510 | 0.646 | 0.203 | 0 |
| Grocery | 0 | 0.765 | 0.276 | -0.375 | -0.160 | -0.411 |
| Frozen | 0.152 | 0 | -0.714 | -0.646 | 0.220 | 0 |
| Detergents/ Paper | 0 | -0.365 | 0.204 | -0.149 | 0.208 | 0.871 |
| Delicassen | 0 | 0 | -0.283 | 0 | -0.917 | 0.265 |





|  | Confusion Table |  |  |
|---|---|---|---|
|  |  | Predicted Values |  |
|  |  | Ho/Re/Ca | Retail |
| Actual Values | Ho/Re/Ca | 281 | 17 |
|  | Retail | 38 | 104 |



## Interpretation:

After finding the summary statistics of both samples, simultaneous confidence intervals based on Hotelling's T^2 and Bonferroni corrections were made. For both types of simultaneous confidence intervals, every feature except for "Delicassen" were found to be significantly different at the 95% corrected significance level. Going further, a two sample difference in means test was done using Hotelling’s T^2. Based on both the simultaneous confidence intervals and Hotelling’s test, we found with 95% confidence that the true average annual wholesale spending is significantly different between Hotel/Restaurants/Cafes and Retails, and if in reality they were the same we would observe this data or more extreme with probability less than .001.

We later moved onto the principal component analysis. We found that the first two principal components explained almost 90% of the variance in the data. Moving on to the loadings, fresh, milk, and frozen product spending were the only features that contributed to the first principal component. Based on the loadings, the fresh feature contributes the most to this principal component since it has the highest loading. The fresh, milk, grocery, and detergents/paper product spending contribute to the second component. Grocery product spending contributes the most to this component while the milk and detergents/paper product spending take away from it. So all the features, except Delicassen, are able to explain a majority of the variation in the data since all the features, besides Delicassen, were present in the first two principal components. A scree plot was also done to see how many components should be used to visualize this data and we found that the first two would be enough because of the drop off in the graph after the second principal component. A biplot was also done and labeled “Effects of Features on Principal Components”. From this plot we can once again see that detergents/paper, milk, and grocery products are highly associated with the second component because of their large magnitude and their direction is almost parallel with the component 2 axis. We can see a similar relationship between component 1 with fresh and frozen products.

A linear discriminant analysis was also done after to see how well we could separate these two categories. We used the principal components found earlier in order to perform linear discriminant analysis. A confusion matrix was made and we found that a majority of the Hotels/Restaurants/Cafes were correctly classified with accuracy 94%. However, Retails were not as easily classified with only around 75% of the data being correctly classified resulting in an overall accuracy of 87.5%. The separation line was also graphed on to the principal component graph. In the top portion of the graph , Hotels/Restaurants/Cafes were clustered relatively well, however Retails were somewhat sparsely scattered across the graph creating a not-so-good cluster. The separation line did a good job of separating channel 1 but not a good job of separating channel 2. This may be because of the higher number of channel 1 observations compared to channel 2 observations biasing the algorithm towards channel 1.

## Conclusion:

In conclusion, we found that the Hotels/Restaurants/Cafes have significantly different average annual wholesale spending compared to Retails. We were also able to find a cluster structure in our data and classified the two types of customers based on their principal components using linear discriminant analysis with 87.5% accuracy.

However, this analysis did have some issues. All of the testing and confidence intervals relied on the fact that our data was multivariate normal. Multivariate normality was assumed for each population. In reality it is unknown if the population actually follows this distribution. Next, Hotels, Cafes, and Restaurants were lumped together in one category and compared to the Retail category during testing. It could have been that one of the lumped types of establishments may have been similar to Retails.