NavarreteFinalProject

Rebeca Navarrete

4/26/2022

## Introduction

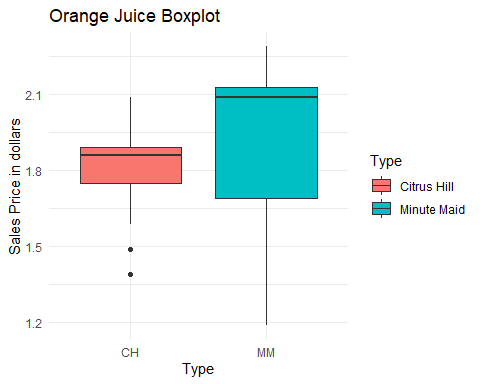
In this report, we analyze the *Orange Juice Data*. This data contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice. This data set contains the following characteristics: Purchase, WeekofPurchase, StoreID, PriceCH, PriceMM, DiscCH, DiscMM, SpecialCH, SpecialMM, LoyalCH, SalePriceMM, SalePriceCH, PriceDiff, Store7, PctDiscMM, PctDiscCH, ListPriceDiff, and STORE. Of all these characteristics, we will only use PriceCH (price charged for CH), PriceMM (price charged for MM), SalesPriceMM (sale price for CH), and SalesPriceCH (sale price for CH) because our objective is to analyze the distribution and skewness of the price of the orange juices, the distribution and skewness of the sales price of the orange juices, and the relationship between the price and sales price of the orange juices.

The abbreviated data set can be found at <https://github.com/rnavarrete2002/OrangeJuice-Project>, and the full data set can be found at <https://vincentarelbundock.github.io/Rdatasets/datasets.html>

To fulfill our objective, we will use a two-sample t-tests and a boxplot to analyze the distribution and skewness of the mentioned variables, and we will create a linear model to analyze the relationship of the same mentioned variables.

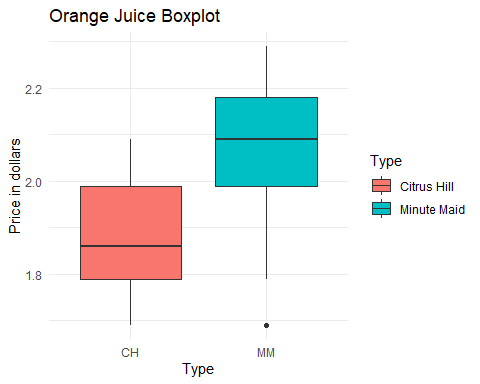
## Sales Price Visualization

The following visualization is a boxplot of the price of Citrus Hill and Minute Maid orange juice. This visualization shows that Citrus Hill price is lower than Minute Maid price. Likewise, it shows there are some outliers among the Minute Maid price, and that the range of the price of Minute Maid is a little bit greater than the range of the prince of Citrus Hill. In this visualization, is important to mention that Citrus Hill sales price distribution is almost symmetric while Minute Maid sales price distribution is left-skewed (negative).



## Price Visualization

The following visualization is a boxplot of the price of Citrus Hill and Minute Maid orange juice. This visualization shows that Citrus Hill price is lower than Minute Maid price. Likewise, it shows there is an outlier among the Minute Maid price, and that the range of the price of Minute Maid is greater than the range of the sales prince of Citrus Hill. In this visualization, is important to mention that Citrus Hill price distribution is right-skewed (positive) while Minute Maid sales price distribution is almost symmetric.



## Sales Price Two-sample t-test

The following two-sample t-test shows that the sales price means are equal. Citrus Hill has a mean sales price of 1.82 dollars while Minute Maid has a mean sales price of 1.96. Likewise, it shows that the p-value is less than 2.2e-16, which means that the results are significant (the p-value results could be due to the size of the sample).

##   
## Welch Two Sample t-test  
##   
## data: CH$SalePrice and MM$SalePrice  
## t = -16.492, df = 1692.7, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.1639071 -0.1290649  
## sample estimates:  
## mean of x mean of y   
## 1.815561 1.962047

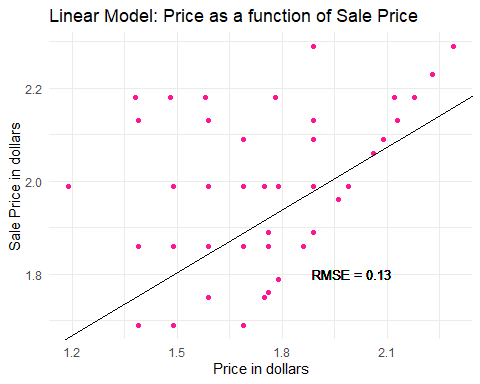
## Price Two-sample t-test

The following two-sample t-test shows that the price means are equal. Citrus Hill has a mean price of 1.87 dollars while Minute Maid has a mean price of 2.09. Likewise, it shows that the p-value is less than 2.2e-16, which means that the results are significant (the p-value results could be due to the size of the sample).

##   
## Welch Two Sample t-test  
##   
## data: CH$Price and MM$Price  
## t = -42.27, df = 1993.5, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.2281045 -0.2078768  
## sample estimates:  
## mean of x mean of y   
## 1.867421 2.085411

## Linear Model: Price as a function of Sale Price

The linear model shows that the model equation is: Price = 0.45SalePrice + 1.13 These coefficients have quite low p-values so we can say they are definitely zero. The RMSE (Root Mean Squared Error) is 0.13 so we can say that the average error of this model when predicting the Price is about 0.13. Which means it is a good model.



## Conclusion

In conclusion, the distribution and skewness of the sales price of Citrus Hill and Minute Maid are different. The first one is almost symmetric while the second one is left-skewed. Besides, Minute Maid’s sales price range is greater than Citrus Hill’s sales price range. Here is important to mention that Minute Maid has an average sales price of 1.81 dollars while Citrus Hill has an average sales price of 1.96 dollars. The distribution and skewness of the price of Citrus Hill and Minute Maid are different. The first one is almost right-skewed while the second one is almost symmetric. Besides, Minute Maid’s price range is greater than Citrus Hill’s price range. Here is important to mention that Minute Maid has an average price of 1.87 dollars while Citrus Hill has an average sales price of 2.10 dollars. Regarding the linear model, the equation found to predict price (Price = 0.45SalePrice + 1.13) had a low root mean squared error; thus, it is a great model to predict the price of Citrus Hill and Minute Maid orange juice. Likewise, it has a positive slope. This means that as the price increases, the sale price increases too.