

# Project Report

## Dominicks Orange Juice Data Analysis

SCM 651 Business Analytics by Professor Basu



**Written by:** Liya Wang  
Sascha Hagedorn  
Ruiyang Chen  
Maximilian Ott

**Date:** 12/12/2017

**Institution:** Syracuse University, Whitman School of Management

## Table of Contents

<b>Executive Summary .....</b>	<b>3</b>
<b>Data Preparation .....</b>	<b>3</b>
<b>Demand - Price Analysis .....</b>	<b>4</b>
<b>Demand - Sale Analysis .....</b>	<b>5</b>
<b>Analysis of Demographics .....</b>	<b>6</b>
<b>Price - Brand Analysis.....</b>	<b>7</b>
Price of FL .....	8
Price of HH .....	8
Price of TROP .....	8
<b>Brand-Location Analysis.....</b>	<b>9</b>
FL on-sale locations and frequency .....	9
HH on-sale Frequency.....	10
TROP on-sale frequency .....	10
<b>Linear Regression (Predict Demand) .....</b>	<b>11</b>
SVM Model .....	11
KSVM Model (default).....	12
KSVM Model (laplacedot kernel).....	12
<b>Sales Prediction with Logistic Regression.....</b>	<b>12</b>
Full GLM Model.....	13
Test for WORKWOM .....	13
Test for RETIRED .....	14
Test for UNEMP .....	14
<b>Cross-Brand Price Analysis.....</b>	<b>14</b>
Model for FL .....	15
Model for HH .....	15
Model for TROP .....	15
Cross Price Elasticity Grid.....	15
<b>Pricing Strategy.....</b>	<b>16</b>
<b>Conclusion .....</b>	<b>17</b>

## Executive Summary

In this project, we are given a dataset about orange juice sale acquired from Dominicks database to analyze. Given more than a dozen of different brands, we are to choose only two high-priced brands and one lower-priced one and be more specific about our results. We also have developed different research questions to analyze the relationship between price, brands, locations, and many more. Based on our analysis, we generated models, validate our models, and made predictions.

This report encompasses detailed analysis of our models generated and intuitive visualizations to help our targeted audience to better understand the information we try to convey.

## Data Preparation

Before actually analyzing the data, a preparation is necessary. The information is divided into three different spreadsheets containing information about many brands. We join the three spreadsheets containing information about the product, the weekly movement and the demographics of the store location. Here, the feature “storeweek” acts as primary key. Then, we decide for two high-priced brands and one low-priced brand to analyze. By calculating the average price of each brand, we figured which brands can be considered as high-priced and which cannot. The brands “Tropicana Pure Premium”, “HH” and “Flat Nat Homesq” are selected and a subset containing information about only these three is created.

We also delete the observations containing missing values to make sure, that we conduct our analysis on clean data containing all the necessary information. After that, we also change the type of the features “Storeweek”, “orange\_high\_movement\_upc” and “upcrfj\_upc” to be character and the type of the features “logPRICE”, “PRICLOW”, “PRICMED” and “PRICHIGH” to factor.

Furthermore, we split our data randomly in two parts called training and test data. The ratio is 7:3. The intention for doing that is to exclude data from the data used to fit our regression model. Thus, we have so far unseen data to validate or test our model in the end.

For the Cross-Brand Analysis, we also needed to create a new table. We needed to filter and create three different table, so that each brand has its own table. Afterwards, we joined two brands onto a chosen brand with the STOREWEEK (inner to make sure that we only look at

observations that have all three brands in a given STOREWEEK) as key. Therefore, we had all the individual logMOVEs and logPRICES for each brand in a given STOREWEEK.

## Demand - Price Analysis

After preparing the data, we trained a linear regression model with the following structure:

```
lm(formula = logMOVE ~ orange_high_movement_BRAND + logPRICE +
  orange_high_movement_BRAND * logPRICE + Feat * orange_high_movement_BRAND
  + Feat + QTY + AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME +
  HHSINGLE + HHLARGE + WORKWOM + HVAL150 + SINGLE + RETIRED +
  UNEMP + NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL +
  CPDIST5 + CPWVOL5, data = dfmas)
```

Looking at the coefficients of the intersection between BRAND and logPrice on gets the following results. This answers the question how the price influences the demand for each brand. The coefficients for each brand can be seen in the following table.

logPrice (FL)	= -3.0941		
HH	= -3.0941	+ 0.0582	= -3.0359
TROP	= -3.0941	+ 0.6844	= -2.4097

A first insight which can be derived from the coefficients above is that the price elasticity varies for each brand and that price affects demand. This means that when the price changes one percent, the demand for the brand FL decreases more drastically than for the other brands (-3.0941 %) - the price elasticity for FL is high. Opposed to FL, TROP has a relatively low price elasticity.

We did the Linear Hypothesis test, which identify the price elasticity of demand of a brand is different at a 99% confidence level, because 2.2e-16 is smaller than 0.001.

Linear hypothesis test

Hypothesis:

orange\_high\_movement\_BRANDHH:logPRICE = 0

orange\_high\_movement\_BRANDTROPICANA PURE PREM:logPRICE = 0

Model 1: restricted model

Model 2: logMOVE ~ orange\_high\_movement\_BRAND + orange\_high\_movement\_BRAND \*  
logPRICE + QTY + logPRICE + Feat + Feat \* orange\_high\_movement\_BRAND +  
AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +  
HHLARGE + WORKWOM + HVAL150 + SINGLE + RETIRED + UNEMP +  
NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 +  
CPWVOL5

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	90517	57388				
2	90515	57148	2	240.16	190.19	< 2.2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Demand - Sale Analysis

The coefficients' values for the Demand - Sale Analysis are coming from the same regression model as the values for the Demand - Price Analysis. Here, we just look at other coefficients. The intersection of the features "Feat" and brand show the coefficients found in the table below.

Feat1 (FL)	= 0.2044		
HH	= 0.2044	- 0.0325	= 0.1719
TROP	= 0.2044	+ 0.1654	= 0.3698

A first result is that if a product is on sale or not has an effect on demand for every brand. One can see, that the effect of Sale on demand varies for each brand. If you do the math, you can derive that HH orange juice on sale increases the (log)demand by 0.17 (or move by  $e^{0.2044} = 1.19$ ). The same is applicable for the other two brands: 0.20 for FL (1.23) and 0.36 (1.45) for TROP.

We also did the Linear Hypothesis test, which can identify that the impact of sale for demand is different among three brands.

### Linear hypothesis test

#### Hypothesis:

orange\_high\_movement\_BRANDHH:Feat1 = 0

orange\_high\_movement\_BRANDTROPICANA PURE PREM:Feat1 = 0

Model 1: restricted model

Model 2: logMOVE ~ orange\_high\_movement\_BRAND + orange\_high\_movement\_BRAND \*  
logPRICE + QTY + logPRICE + Feat + Feat \* orange\_high\_movement\_BRAND +  
AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +  
HHLARGE + WORKWOM + HVAL150 + SINGLE + RETIRED + UNEMP +  
NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 +  
CPWVOL5

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	90517	57275				
2	90515	57148	2	126.92	100.52	< 2.2e-16 ***


---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

>

## Analysis of Demographics

This paragraph sheds some light on the question how the demographics of a location of a store influence the demand. The picture below shows the different demographics including their importance. On the left, one can see the demographics which have positive impact on demand and on the right one can see the demographics who influence the demand in a negative way.

1	<u>AGE9</u>	<u>10.137974796</u>		17	SSTRDIST	-0.009504824	
11	<u>SINGLE</u>	<u>5.783274583</u>		15	POVERTY	-0.060422009	
2	<u>AGE60</u>	<u>1.960041770</u>		18	SSTRVOL	-0.120572864	
5	NOCAR	1.040758179		20	CPWVOL5	-0.259330067	
4	EDUC	0.743820515		3	ETHNIC	-0.444413536	
10	HVAL150	0.439346243		12	RETIRED	-1.889511559	
14	NWHITE	0.343886992		7	HHSINGLE	-3.293164277	
16	DRTIME5	0.058527692		9	<u>WORKWOM</u>	<u>-5.586195791</u>	
19	CPDIST5	0.050660852		8	<u>HHLARGE</u>	<u>-9.359367511</u>	
6	INCOME	0.042622294		13	<u>UNEMP</u>	<u>-9.470700760</u>	

AGE9, SINGLE and AGE 60 have the highest positive impact on demand as opposed to UNEMP, HHLARGE and WORKWOM have the highest negative impact on demand. With

AGE9 being the highest impact for orange juice sales, we can interpret that neighborhoods with a higher number of kids usually buy more orange juice. Single households usually might be younger people, such as students, and are also interested in orange juice, because of their childhood or other reasons. The biggest negative factors might indicate that people that are unemployed usually might rather buy soda or other cheap drinks and not our high priced orange juice (2/3 are high priced in the analysis). Having this information, the responsible business unit can decide where to plan new stores and where a closing of stores should be considered. Obviously information about different areas where the store is located is needed to draw the relation. Here the business unit can team up with municipalities or other public organizations to retrieve this information.

The research also found that POVERTY and INCOME have no significant impact. After removing these two variables from the full model, and comparing, we found that the P value is 0.5906, which cannot reject our Null Hypothesis (Poverty and Income have no impact) at a reasonable confidence level.

#### Analysis of Variance Table

```
Model 1: logMOVE ~ orange_high_movement_BRAND + orange_high_movement_BRAND *
logPRICE + QTY + logPRICE + Feat + Feat * orange_high_movement_BRAND +
AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +
HHLARGE + WORKWOM + HVAL150 + SINGLE + RETIRED + UNEMP +
NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 +
CPWVOL5
Model 2: logMOVE ~ orange_high_movement_BRAND + orange_high_movement_BRAND *
logPRICE + QTY + logPRICE + Feat + Feat * orange_high_movement_BRAND +
AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + HHSINGLE + HHLARGE +
WORKWOM + HVAL150 + SINGLE + RETIRED + UNEMP + NWHITE + DRTIME5 +
SSTRDIST + SSTRVOL + CPDIST5 + CPWVOL5
Res.Df  RSS Df Sum of Sq    F Pr(>F)
1  90515 57148
2  90517 57149  -2   -0.66498 0.5266 0.5906
```

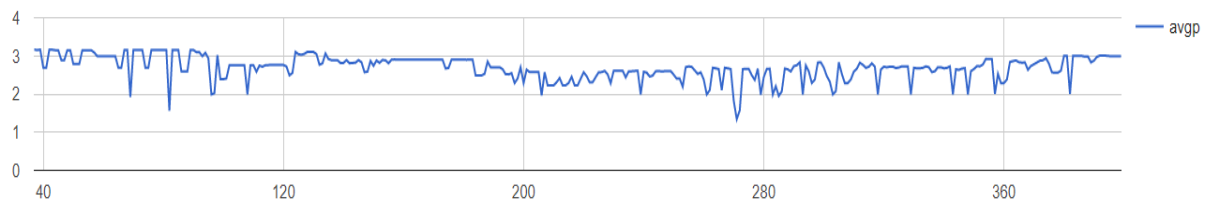
## Price - Brand Analysis

Calculating the average price for each brand shows how the price is related to the brand. FL and Tropicana are our high-priced brands and HH is the low-priced brand. There is a huge difference between these two groups as we see in the picture below.

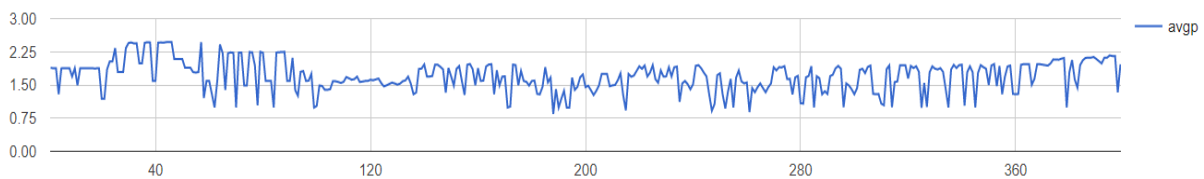
orange_high_movement_BRAND	avglogprice
FL NAT HOMESQ	0.979143691360744
TROPICANA PURE PREM	0.975724224472475
HH	0.4985746999580827

Furthermore, the price for each brand over time is analyzed and visualized in the next three graphics.

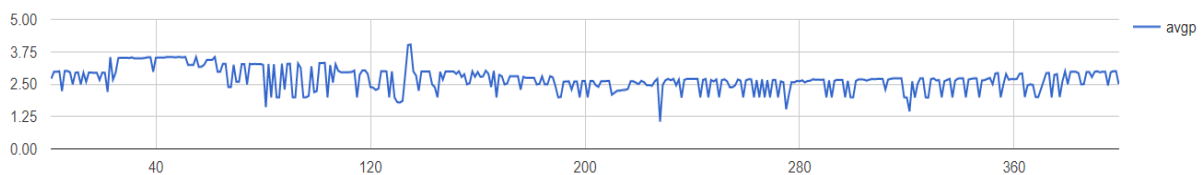
### Price of FL



### Price of HH



### Price of TROP



Having a close look on all three visualizations it seems like the brands started with a rather high price. Over time the price decreased until reaching a point where it started increasing again. At the end of capturing the data it was again higher than in the middle of the captured time span.

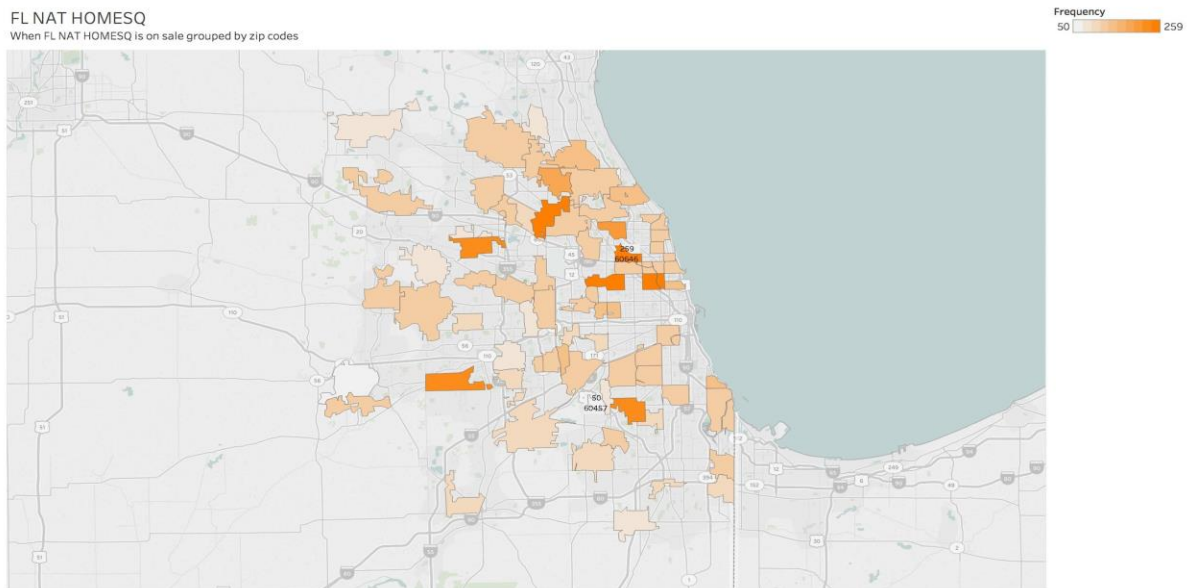


One can also see that the price of each brand has several intense outliers. Here responsible business units can analyze what the reason for these either extremely high or low prices is. If this has positive impact on selling orange juice, the brand can continue doing these price changes. If this has negative impact, actions to avoid these outliers can be made.

## Brand-Location Analysis

We continue our analysis with brand focusing on the frequency of how many times each brand is on sale, grouping by zip codes. We use Tableau to create these maps defined by the following dimensions and measures. Simply stated, these maps consist of location coordinates from our original dataset as latitude and longitude, and each map is colored by the how many times each brand is on sale. Each pair of location coordinates is corresponded to a zip code, and the map is sized by all zip codes in our dataset. Maximum and minimum frequencies are labeled for each brand.

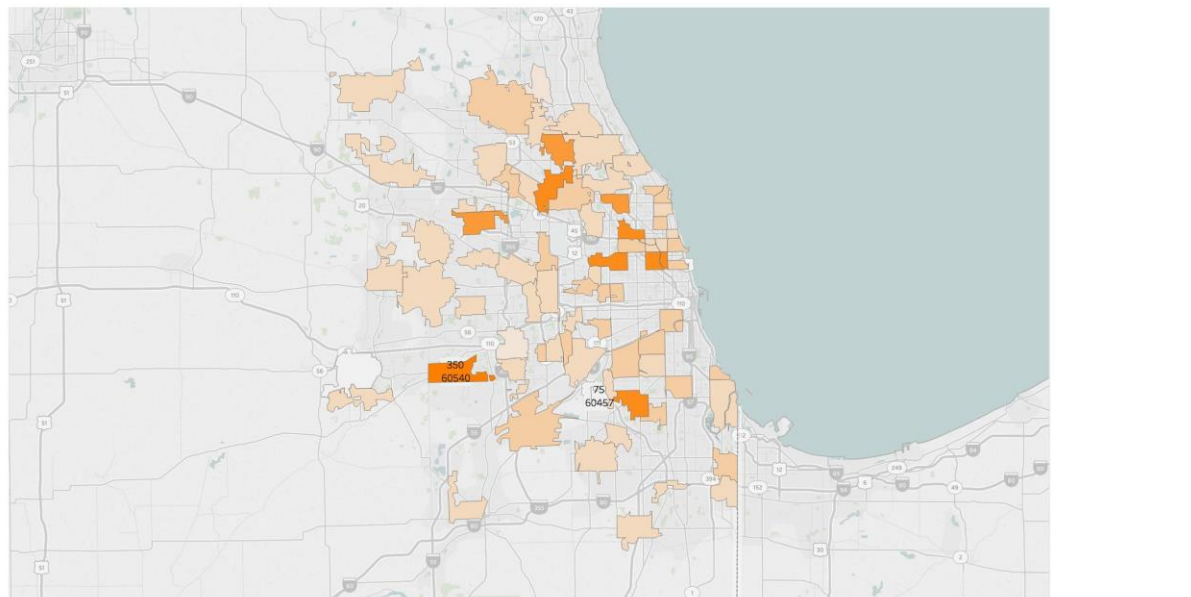
### FL on-sale locations and frequency



## HH on-sale Frequency

### HH ORANGE JUICE

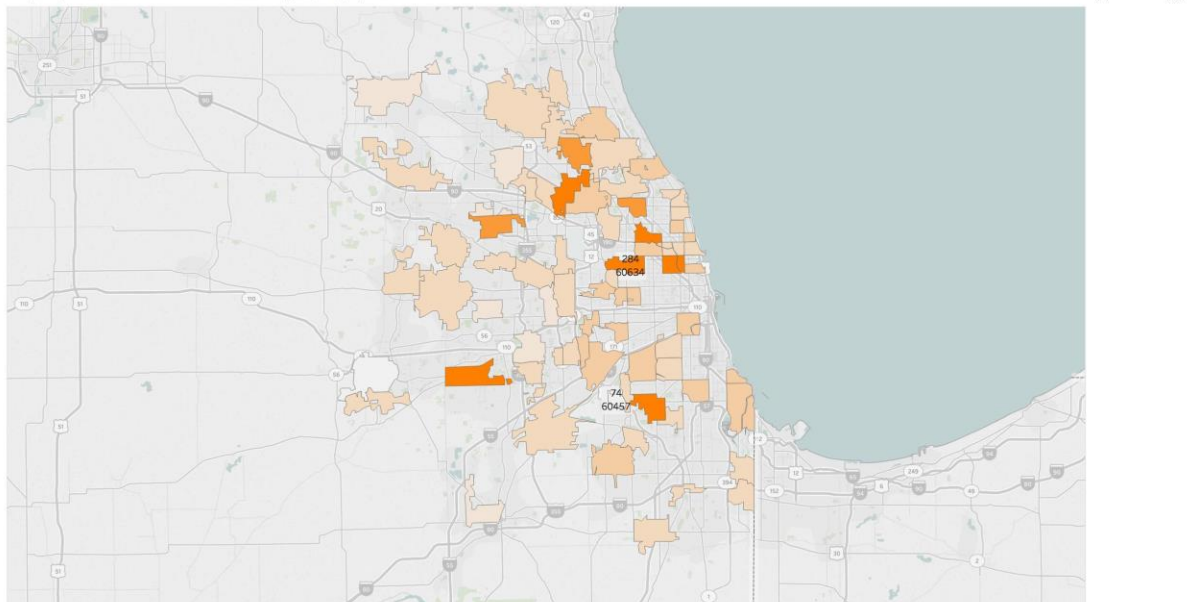
When HH ORANGE JUICE is on sale grouped by zip codes



## TROP on-sale frequency

### TROP

Frequency when TROPICANA PURE PREM is on sale grouped by zip codes



Healthy Humans orange juice is more likely to be on sale in Chicago area, and this pattern aligns with our previous price-brand analysis that HH undergoes more fluctuated prices. Areas close to the north Chicago are more likely to have on-sale orange juice of these three brands, probably because they are where students of Northwestern University live.

## Linear Regression (Predict Demand)

Here the linear regression model is used to actually predict an exact number of demand based on the independent variables. As mentioned before the model is fit with the training data. Estimating a performance on the same data used for training the model does not make sense, because all bias and errors could just be learned by the model. Learning all the noise of data is called overfitting. This leads to a very good prediction of the training data, but the model performs very badly when applying to future data. This is the reason why we split our data into parts called training and validation data. As the names of the two sets suggest, we use the first set to train our model and the second set to validate the model. The benefit is that using the validation data for evaluating, the model predicts demands for observations which were never seen while training the model.

To estimate our performance, we use the following two different indicators. The table below shows the evaluation of both indicators.

Mean Square Error	Min Max Accuracy
0.63008	0.8628
<b>Root Mean Squared Error: 0.7937757</b>	

(Both show that our model performs well and can be used for future predictions. )

Besides the linear regression model, we also tried other models, which are SVM model, KSVM (default value) model, and KSVM model (kernel=laplacdot). We can see the KSVM model (kernel=laplacdot) has the minimum value of Root Mean Squared Error, which is 0.7079102.

## SVM Model

```
> model.svm.train <- svm(logMOVE ~ orange_high_movement_BRAND+ orange_high_movement_BRAND * logPRICE
+
+QTY+logPRICE+Feat+Feat*orange_high_movement_BRAND+AGE9+AGE60+ETHNIC+EDUC+NOCAR
+
+HHSINGLE+HHLARGE+WORKWOM+HVAL150+SINGLE+RETIRED+UNEMP+NWHITE+DRTIME5+SSTRDIST+SSTRVOL+CPDIST5
+
+CPWVOL5,data = oj.train)
> model.svm.predict <- predict_model(model.svm.train, oj.test)
> rmse(oj.test$logMOVE,model.svm.predict$model.predict)
[1] 0.717542
```

## KSVM Model (default)

```
> model.ksvm.train <- ksvm(logMOVE ~ orange_high_movement_BRAND+ orange_high_movement_BRAND * logPRICE
+
+QTY+logPRICE+Feat+Feat*orange_high_movement_BRAND+AGE9+AGE60+ETHNIC+EDUC+NOCAR
+
+HHSINGLE+HHLARGE+WORKWOM+HVAL150+SINGLE+RETIRED+UNEMP+NWHITE+DRTIME5+SSTRDIST+SSTRVOL+CPDIST5
+
+CPWVOL5,data = oj.train)
> model.ksvm.predict <- predict_model(model.ksvm.train, oj.test)
> rmse(oj.test$logMOVE,model.ksvm.predict$model.predict)
[1] 0.7185299
```

## KSVM Model (laplacedot kernel)

```
> model.ksvm.laplacedot.train <- ksvm(logMOVE ~ orange_high_movement_BRAND+ orange_high_movement_BRAND * logPRICE
+
+QTY+logPRICE+Feat+Feat*orange_high_movement_BRAND+AGE9+AGE60+ETHNIC+EDUC+NOCAR
+
+HHSINGLE+HHLARGE+WORKWOM+HVAL150+SINGLE+RETIRED+UNEMP+NWHITE+DRTIME5+SSTRDIST+SSTRVOL+CPDIST5
+
+CPWVOL5,data = oj.train, kernel = "laplacedot")
> model.ksvm.laplacedot.predict <- predict_model(model.ksvm.laplacedot.train, oj.test)
> rmse(oj.test$logMOVE,model.ksvm.laplacedot.predict$model.predict)
[1] 0.7079102
```

## Sales Prediction with Logistic Regression

We also tried to predict sales with Logistic Regression. First, we tried to find an appropriate model to do so. According to our results, the 'WORKWOM', 'RETIRED' and 'UNEMP' are not significant factors to predict sales.

## Full GLM Model

```
Call:
glm(formula = Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE +
    QTY + AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +
    HHLARGE + WORKWOM + HVAL150 + SINGLE + RETIRED + UNEMP +
    NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 +
    CPWVOLS, family = binomial(logit), data = oj.3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5961 -0.7917 -0.5619  0.9386  2.5535

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.799954   1.992791   2.910 0.003609 **
logMOVE         0.367778   0.009975  36.869 < 2e-16 ***
orange_high_movement_BRANDHH
orange_high_movement_BRANDTROPICANA PURE PREM -2.898660   0.032646 -88.791 < 2e-16 ***
logPRICE       -0.643673   0.024765 -25.991 < 2e-16 ***
QTY            -5.095527   0.058889 -86.528 < 2e-16 ***
AGE9           3.842827   0.206664  18.595 < 2e-16 ***
AGE60          -5.207448   1.712469  -3.041 0.002359 **
ETHNIC         -2.360926   1.097100  -2.152 0.031400 *
EDUC            1.627320   0.304009   5.353 8.66e-08 ***
NOCAR          -1.389745   0.234500  -5.926 3.10e-09 ***
INCOME         -1.004565   0.341938  -2.938 0.003305 **
HHSINGLE        -0.567684   0.144127  -3.939 8.19e-05 ***
HHLARGE        2.321528   0.546597   4.247 2.16e-05 ***
WORKWOM        5.591412   1.025450   5.453 4.96e-08 ***
HVAL150       -0.792819   1.153944  -0.687 0.492050
SINGLE          1.110574   0.106854   10.393 < 2e-16 ***
RETIRED        -1.939504   0.692799  -2.800 0.005118 **
UNEMP          1.135197   1.375227   0.825 0.409109
NWHITE         -2.272084   2.123764  -1.070 0.284692
POVERTY        -0.966034   0.294072  -3.285 0.001020 **
DRTIME5       -3.981056   1.282414  -3.104 0.001907 **
SSTRDIST       0.072475   0.011123   6.516 7.22e-11 ***
SSTRVOL       0.052194   0.003456  15.101 < 2e-16 ***
CPDIST5       0.081916   0.022968   3.567 0.000362 ***
CPWVOLS       -0.131756   0.015081  -8.736 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

We defined the Null Hypothesis: 'WORKWOM', 'RETIRED', and 'UNEMP' are not significant. Removing the possibly non-significant factors, and comparing the two models, we found that the P values are at least larger than 0.2, so we cannot reject the Null Hypothesis, so 'WORKWOM', 'RETIRED', and 'UNEMP' are not significant at 99% confidence level.

## Test for WORKWOM

```
> GLM1.1 <- glm(Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY + AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME +
+ HHSINGLE + HHLARGE + HVAL150 + SINGLE + RETIRED +
+ UNEMP + NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL +
+ CPDIST5 + CPWVOLS, family=binomial(logit), data=oj.3)
> anova(GLM1, GLM1.1, test="Chisq") #workwom is not significant
Analysis of Deviance Table

Model 1: Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY +
  AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +
  HHLARGE + WORKWOM + HVAL150 + SINGLE + RETIRED + UNEMP +
  NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 +
  CPWVOLS
Model 2: Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY +
  AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +
  HHLARGE + HVAL150 + SINGLE + RETIRED + UNEMP + NWHITE + POVERTY +
  DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 + CPWVOLS
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      90519      96670
2      90520      96670 -1 -0.47212    0.492
```

## Test for RETIRED

```
> GLM1.2 <- glm(Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY + AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME +
+             HHSINGLE + HHLARGE + HVAL150 + SINGLE + WORKWOM +
+             UNEMP + NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL +
+             CPDIST5 + CPWVOL5, family=binomial(logit), data=oj.3)
> anova(GLM1, GLM1.2, test="Chisq") #retired is not significant
Analysis of Deviance Table

Model 1: Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY +
  AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +
  HHLARGE + WORKWOM + HVAL150 + SINGLE + RETIRED + UNEMP +
  NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 +
  CPWVOL5
Model 2: Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY +
  AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +
  HHLARGE + HVAL150 + SINGLE + WORKWOM + UNEMP + NWHITE + POVERTY +
  DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 + CPWVOL5
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      90519      96670
2      90520      96671 -1 -0.68142  0.4091
```

## Test for UNEMP

```
> GLM1.3 <- glm(Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY + AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME +
+             HHSINGLE + HHLARGE + HVAL150 + SINGLE + WORKWOM + RETIRED + NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL +
+             CPDIST5 + CPWVOL5, family=binomial(logit), data=oj.3)
> anova(GLM1, GLM1.3, test="Chisq") #unemp is not significant
Analysis of Deviance Table

Model 1: Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY +
  AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +
  HHLARGE + WORKWOM + HVAL150 + SINGLE + RETIRED + UNEMP +
  NWHITE + POVERTY + DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 +
  CPWVOL5
Model 2: Feat ~ logMOVE + orange_high_movement_BRAND + logPRICE + QTY +
  AGE9 + AGE60 + ETHNIC + EDUC + NOCAR + INCOME + HHSINGLE +
  HHLARGE + HVAL150 + SINGLE + WORKWOM + RETIRED + NWHITE +
  POVERTY + DRTIME5 + SSTRDIST + SSTRVOL + CPDIST5 + CPWVOL5
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      90519      96670
2      90520      96671 -1 -1.1447  0.2847
```

Afterwards we used the restricted model to predict sales. AGE9 has the highest negative influence on sales, while HHLARGE has the highest positive influence for sales. AGE9 might indicate that, when the neighborhoods have a high number of children, then it is not necessary to have the orange juice on sale, because the sales will be high enough anyways. Large households do not seem to be attracted to orange juice, meaning stores might need to increase the profit with sales, but some more analyses are needed to get to the root of this influence.

Further, our model predicted with an accuracy of 75.8%, which is decent.

## Cross-Brand Price Analysis

We used three different Linear Regression Models to find out the cross price elasticities of our three different brands.



## Model for FL

```
lm(formula = logMOVE_fl ~ logPRICE_fl + logPRICE_hh + logPRICE_trop,
   data = dfcrossbrand)
```

## Model for HH

```
lm(formula = logMOVE_hh ~ logPRICE_hh + logPRICE_fl + logPRICE_trop,
   data = dfcrossbrand)
```

## Model for TROP

```
lm(formula = logMOVE_trop ~ logPRICE_trop + logPRICE_hh + logPRICE_fl,
   data = dfcrossbrand)
```

## Cross Price Elasticity Grid

<i>Demand</i>	AVG PRICE	FL	HH	TROP
FL	<b>2.69</b>	<b>-3.54397</b>	-0.00168	1.21141
HH	1.69	0.47743	<b>-3.2787</b>	0.49664
TROP	<b>2.70</b>	1.33074	0.47601	<b>-3.02489</b>

The price elasticities of each brand are negative, which is normal, so if the price goes up the corresponding demand decreases. All cross price elasticities, except FL-HH, are positive, which indicate an increase in demand if the prices of the other brands increase.

We can also see that the cross price elasticities of FL and TROP for HH are approximately 0.5, thus, rather small. This might indicate that HH and FL/TROP have two different customer groups. There are not that many customers that switch from HH to FL or TROP, when the price of FL or TROP increases. The results of the cross price elasticities for FL and TROP support that argument, because the elasticities of TROP-FL and FL-TROP are the largest in our analysis. This might indicate that FL respectively TROP are a substitute for each other in case of the prices increase. Further, most of the customers of FL and TROP do not care about an increase of the price of HH.

The only exception is FL-HH - if the price of HH increases, then the demand decreases a small amount. A reason for this might be that maybe some customers like to buy some amount of FL, when they come for HH. So, if HH's price increases they might go to another store and this affects the sales of FL. But this would be just a small number of customers.

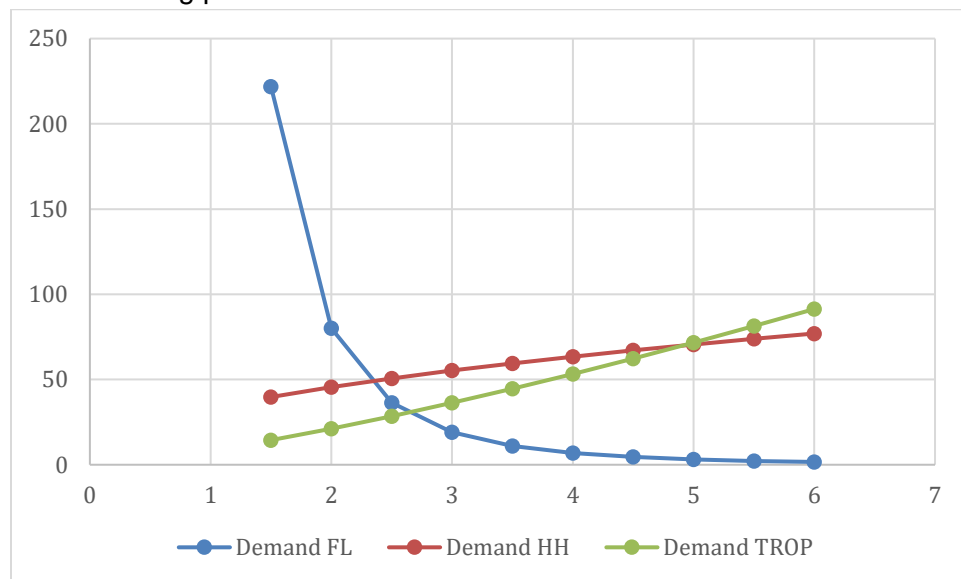
# Pricing Strategy

Our strategy is to increase the price of FL to push our customer base for FL to our brand TROP. There are multiple reasons for this:

1. The price elasticity of TROP is lower than the one of FL, which means we can push customers more easily to our brand TROP
2. Another reason is the cross elasticity of TROP-FL vs. FL-TROP, which explains a bigger amount of customers that substitute TROP for FL (FL-TROP: 1.33074 vs TROP-FL: 1.21141)
3. The average profit margin of TROP is better than the one of FL (12.17% vs. 10.85%)
4. The price increase of FL effects the demand of HH slightly less than an increase of the price of TROP (0.47743 vs 0.49664)

While keeping the price of HH and TROP constant and increasing the price of FL, we can see that the demand increases steadily for HH and TROP. See Figure below.

Demand while increasing price of FL



The prices for our strategy are:

- ➔ FL 4
- ➔ HH 2.32
- ➔ TROP 3.93

The total profit per week (on average across stores) would be 150.81 and therefore the annual profit would be 7842. In comparison to actual profit, which we calculated based on the average absolute profit margin of each brand, would be 7078.52. This means that we would



improve the profit by approximately by 763. It would be possible to increase even further if we keep on increasing the price of FL. For example, another reasonable price of FL would be 4.2, which would not change the price of HH and TROP that much, but increase the annual profit by approximately 1093. This also might mean that we could suggest to remove the brand FL completely and to focus entirely in HH and TROP.

## Conclusion

We did a variety of analysis throughout the project, such as demand-price analysis, demand-sale analysis, analysis of demographics, description of price for each brand, location of sales analysis, prediction of demand, prediction of sales, cross brand analysis and defining a price strategy.

At the end of our analysis, we conclude that:

- Each brand has a different price elasticity while Flat Nat Homesq has the highest price elasticity.
- The demand of Tropicana Pure Premium is affected by sales the most. Variable AGE9 has the biggest positive impact on demand and unemployment rates have the most negative impact on demand. However, poverty and income have no statistical significance and do not affect our predicted demand.
- All three brands of orange juice are on sale more frequently in north Chicago.
- KSVM with Laplace dot kernel has the best prediction accuracy for demand
- Working women, retired people and unemployed people have no statistical significance for the prediction of sales.
- There is high cross price elasticity between the two high price brands FL and TROP. These two brands do not affect the low price brand that much.

Thus, our strategy is to focus on HH & Tropicana Pure Premium and to increase the price of Flat Nat Homesq to push the customers from Flat Nat Homesq to Tropicana Pure Premium.