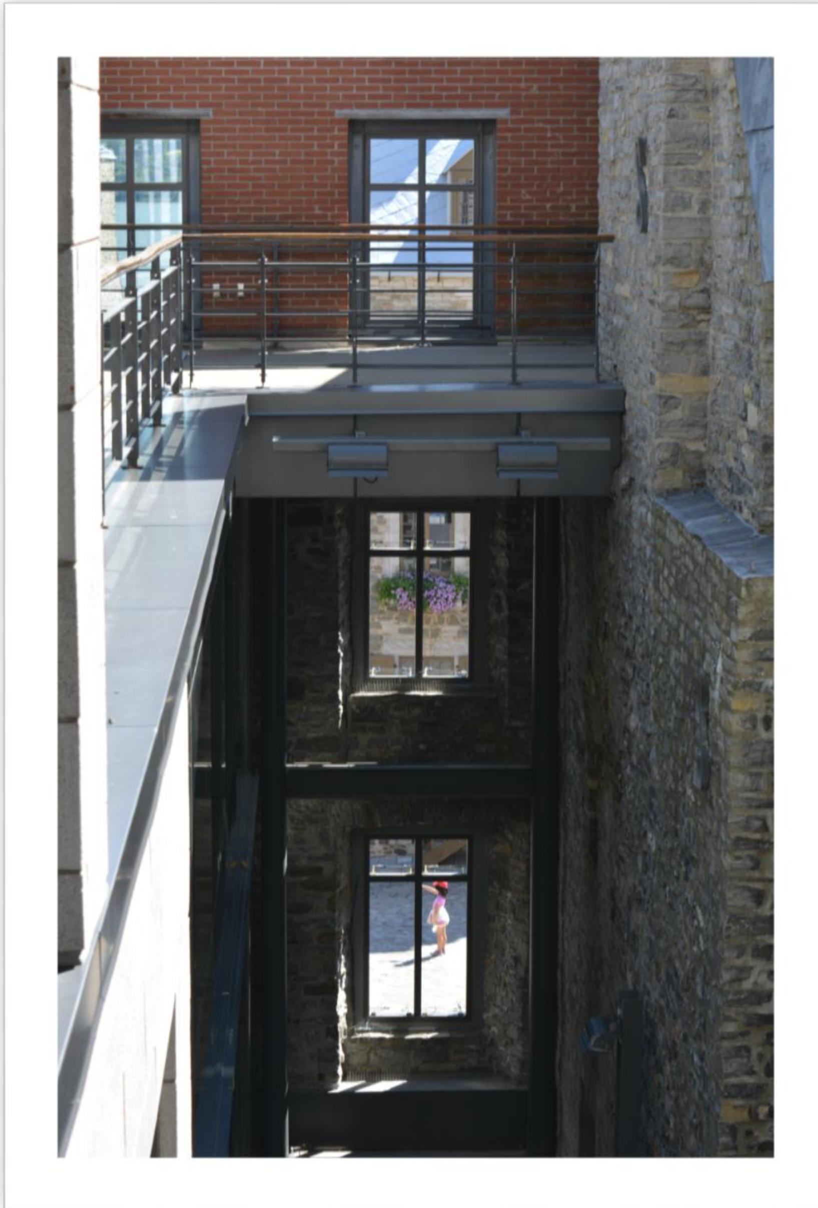


Reactive Predictive Ordering

RESPONSIVE TREND MONITORING

Robert Blacklidge | Springboard | August 1, 2016

Traditional Ordering

Most small businesses use instinct and past experiences in order to determine ordering needs. There are a multitude of factures that can influences consumer purchasing patterns. This variation in purchasing creates the opportunity for perpetually over or under ordering which leads to missed sales or waste. It is these issues I seek to resolve through identifying trends and key influencers.

I seek to discover relevant variables affecting Omaha metro consumer purchasing trends as it relates to sandwich shops. Customer demand is affected by a multitude of variables that are hard to determine. While I was operating a local sandwich shop it was nearly impossible for me to judge our demand for a particular day. This caused a multitude of issues for the company. As a food vendor we must order weekly due to the perishable nature of the product we produce. I hope to use key indicators to predict the demand on a particular day. Knowing the potential demand will allow for improved ordering which reduces waste and improves efficiencies.

I plan on using Stock Pricing, Fuel Prices, Sales Data, University of Nebraska – Omaha Crime Data, Temperatures, and Rain Fall. Stock Pricing, the closing stock price of multiple fortune 500 business in the Omaha metro as economic indicators. The variations in stock prices will help to identify fluctuations in the local market as the companies are closely tied to the local economy. Fuel Prices, this is a variable that can have a significant impact on the expendable budgets of consumers. Sales Data, I will be using historical sales data to develop models against.

This data will allow me to discover key influencers, if any, as compared to the Number of Customers. University of Nebraska – Omaha Crime Data, criminal data can have impacts that are not easily seen I will be utilizing this data as a way of looking for erroneous and unique factors. Temperatures, daily temperatures can impact travel and shopping habits; with the temperature we can see what key points have the largest impact on the consumer’s behavior. Rain Fall, daily rain can impact travel and shopping habits; with the rain data we can see what key points have the largest impact on the consumer’s behavior.

Datasets

Stocks contains fields containing date, close price, volume, open price, high, and low. The Stocks data sets are five historical stock information on Werner, Union Pacific, First Data, Walmart, and Conagra. Each data set contains the closing price for each of the companies, it is with this that I plan on using to model against. With increases and decreases in the prices compared to customer sales I plan on showing the impact each of these variables might have as influencers.

```
> summary(Conag)
        date           close           volume             open            high             low       
 1/11/2016:  1   Min.   :38.70   Min.   : 1234161   Min.   :38.47   Min.   :38.91   Min.   :37.97  
 1/12/2016:  1   1st Qu.:41.08   1st Qu.: 2092104   1st Qu.:41.15   1st Qu.:41.59   1st Qu.:40.66  
 1/13/2016:  1   Median :42.49   Median : 2726665   Median :42.48   Median :42.85   Median :42.10  
 1/14/2016:  1   Mean   :43.30   Mean   : 3053041   Mean   :43.28   Mean   :43.66   Mean   :42.90  
 1/15/2016:  1   3rd Qu.:45.75   3rd Qu.: 3571842   3rd Qu.:45.81   3rd Qu.:46.28   3rd Qu.:45.54  
 1/19/2016:  1   Max.   :48.39   Max.   :13021350   Max.   :48.24   Max.   :48.81   Max.   :48.02  
 (Other)   :248                                                                                    
> summary(First.Data)
        date           close           volume             open            high             low       
 1/11/2016:  1   Min.   : 8.67   Min.   :  284937   Min.   : 8.82   Min.   : 9.45   Min.   : 8.37  
 1/12/2016:  1   1st Qu.:12.12   1st Qu.: 2416484   1st Qu.:12.11   1st Qu.:12.38   1st Qu.:11.84  
 1/13/2016:  1   Median :12.96   Median : 3592138   Median :12.94   Median :13.19   Median :12.66  
 1/14/2016:  1   Mean   :13.41   Mean   : 4892647   Mean   :13.42   Mean   :13.71   Mean   :13.13  
 1/15/2016:  1   3rd Qu.:15.46   3rd Qu.: 5214864   3rd Qu.:15.48   3rd Qu.:15.88   3rd Qu.:15.21  
 1/19/2016:  1   Max.   :17.80   Max.   :64965690   Max.   :17.90   Max.   :17.99   Max.   :17.51  
 (Other)   :208                                                                                    
> summary(wern1)
        date           close           volume            open            high             low       
 1/11/2016:  1   Min.   :21.41   Min.   : 134798   Min.   :21.16   Min.   :21.84   Min.   :20.91  
 1/12/2016:  1   1st Qu.:24.13   1st Qu.: 546904   1st Qu.:24.09   1st Qu.:24.48   1st Qu.:23.66  
 1/13/2016:  1   Median :25.57   Median : 694635   Median :25.59   Median :26.03   Median :25.12  
 1/14/2016:  1   Mean   :25.40   Mean   : 817357   Mean   :25.38   Mean   :25.78   Mean   :24.97  
 1/15/2016:  1   3rd Qu.:26.83   3rd Qu.: 915358   3rd Qu.:26.82   3rd Qu.:27.25   3rd Qu.:26.32  
 1/19/2016:  1   Max.   :28.63   Max.   :6331999   Max.   :28.70   Max.   :28.95   Max.   :28.26  
 (Other)   :248                                                                                   
> summary(Union)
        date           close           volume             open            high             low       
 1/11/2016:  1   Min.   :68.79   Min.   : 2120102   Min.   :69.34   Min.   :69.77   Min.   :67.06  
 1/12/2016:  1   1st Qu.:79.41   1st Qu.: 3970607   1st Qu.:79.64   1st Qu.:80.50   1st Qu.:78.53  
 1/13/2016:  1   Median :84.79   Median : 4962014   Median :84.84   Median :85.39   Median :84.12  
 1/14/2016:  1   Mean   :84.44   Mean   : 5426220   Mean   :84.35   Mean   :85.31   Mean   :83.49  
 1/15/2016:  1   3rd Qu.:88.75   3rd Qu.: 6306068   3rd Qu.:88.55   3rd Qu.:89.51   3rd Qu.:87.97  
 1/19/2016:  1   Max.   :97.05   Max.   :19535860   Max.   :97.75   Max.   :98.28   Max.   :96.17  
 (Other)   :248                                                                                    
> summary(WMT)
        date           close           volume             open            high             low       
 1/11/2016:  1   Min.   :56.42   Min.   : 2483121   Min.   :56.39   Min.   :57.06   Min.   :56.30  
 1/12/2016:  1   1st Qu.:63.07   1st Qu.: 6843902   1st Qu.:62.83   1st Qu.:63.79   1st Qu.:62.04  
 1/13/2016:  1   Median :66.50   Median : 9222046   Median :66.61   Median :67.02   Median :65.89  
 1/14/2016:  1   Mean   :66.23   Mean   :10577852   Mean   :66.19   Mean   :66.75   Mean   :65.68  
 1/15/2016:  1   3rd Qu.:69.84   3rd Qu.:12439538   3rd Qu.:69.69   3rd Qu.:70.08   3rd Qu.:69.36  
 1/19/2016:  1   Max.   :74.30   Max.   :80751840   Max.   :74.94   Max.   :75.19   Max.   :73.87  
 (Other)   :248                                                                                    
```

Fuel Prices contains date and weekly U.S. all grades all formulations retail gasoline. Fuel prices have far reaching impacts on economics systems as a whole. With a submarine sandwich be considered a convenience item it is likely that a reduction in discretionary funds will impact customer sales. I seek to infer the significance that the fluctuations in the gas price may have on consumer behavior.

```
> summary(Fuel.Cost)
        Date    
 1/1/1996 :   1
 1/1/2001 :   1
 1/1/2007 :   1
 1/10/1994:   1
 1/10/2000:   1
 1/10/2005:   1
 (Other)  :1215
 Weekly.U.S..All.Grades.All.Formulations.Retail.Gasoline.Prices...Dollars.per.Gallon.
 Min.   :0.949
 1st Qu.:1.255
 Median :1.943
 Mean   :2.145
 3rd Qu.:2.911
 Max.   :4.165
```

Sales Data contains date, category, item, qty, modifiers applied, gross sales, tax, device name, and event type. I will be using historical sales data build the models as this will be the variable I am looking to influence. The number of transactions will be the important information; this can be found in the number of duplicate dates. Each unique transaction is represented by a new line. The total number of counts for a particular date represents the total consumer transactions for a particular day.

```
       Date           Time                              Time.Zone      Category
 1/4/2016 : 190   11:03:33:    6   Central Time (US & Canada):7298   12"   : 218
 2/22/2016: 176   11:16:06:    6                                     6"    :1547
 1/25/2016: 168   11:16:29:    6                                     None  :4377
 2/8/2016 : 167   11:33:26:    6                                     Salad:   68
 2/29/2016: 159   11:36:28:    6                                     Sides:1064
 1/11/2016: 154   11:43:20:    6                                     Soup :   24
 (Other)  :6284   (Other) :7262
           Item             Qty               Price.Point.Name    SKU
 Custom Amount:4377   Min.   :-1.000                 :4377     Mode:logical
 Chips        : 873   1st Qu.: 1.000   Regular       :  24     NA's:7298
 Blimpie Best : 502   Median : 1.000   Regular Price:2897
 Club         : 322   Mean   : 1.002
 Turkey       : 306   3rd Qu.: 1.000
 Tuna         : 109   Max.   : 3.000
 (Other)      : 809
    Modifiers.Applied  Gross.Sales      Discounts       Net.Sales         Tax
              :7066   $1.00  : 988   $0.00  :7296   $1.00  : 988   $0.00  :7282
 Bacon        :  60   $4.25  : 400   ($1.00):    2  $4.25  : 400   $0.07  :    4
 Extra Cheese :  51   $4.50  : 378                  $4.50  : 378   $0.45  :    3
 Pretzel Bread:  36   $8.63  : 290                  $8.63  : 289   $0.28  :    1
 Double Meat  :  35   $9.16  : 284                  $9.16  : 284   $0.42  :    1
 Guacamole    :  27   $6.48  : 259                  $6.48  : 259   $0.46  :    1
 (Other)      :  23   (Other):4699                  (Other):4700   (Other):    6
                  Transaction.ID                    Payment.ID              Device.Name
 HSk0tLFWpixWUV4OGPVGuDneV:    5   LmJAMhu9eoV2ggfWL5yfKQB :    5   Heather's iPad:    1
 1D1SJUdZBKOhskFG3L7la0zeV:    4   1TleRypt2LAHd1wCHNvXLQB :    4   iPad          :3154
 5JRFSdsk9yjMLxa0VjTshPpeV:    4   1VOe1J1G3cGwMbvOZfOvKQB :    4   iPhone        :4143
 74qE7oQGEIp0QAc3e2GItfjeV:    4   5hd5s10iDKssNIozIc9zJQB :    4
 b4OqbABn148n2iFk45iQcl5eV:    4   7gEHZP2iDugctcO0JAuKLQB :    4
 d9SNU0aiixlhEDDOEGnLMC1eV:    4   8lP5JAVoJpVc4bVWHBRvfyMF:    4
 (Other)                  :7273   (Other)                 :7273
                              Notes
                                 :6796
 Turkey and Ham                  : 255
 Ham, Salami, Capicola, Prosciuttini:  80
 Turkey Bacon                    :  80
 Turkey, Ham                     :  80
 Turkey, Bacon                   :   3
 (Other)                         :   4

                                                                                          D
etails
 https://squareup.com/dashboard/sales/transactions/HSk0tLFWpixWUV4OGPVGuDneV/by-unit/ANRYPESAPA
9KJ:    5
 https://squareup.com/dashboard/sales/transactions/1D1SJUdZBKOhskFG3L7la0zeV/by-unit/ANRYPESAPA
9KJ:    4
 https://squareup.com/dashboard/sales/transactions/5JRFSdsk9yjMLxa0VjTshPpeV/by-unit/ANRYPESAPA
9KJ:    4
 https://squareup.com/dashboard/sales/transactions/74qE7oQGEIp0QAc3e2GItfjeV/by-unit/ANRYPESAPA
9KJ:    4
 https://squareup.com/dashboard/sales/transactions/b4OqbABn148n2iFk45iQcl5eV/by-unit/ANRYPESAPA
9KJ:    4
 https://squareup.com/dashboard/sales/transactions/d9SNU0aiixlhEDDOEGnLMC1eV/by-unit/ANRYPESAPA
9KJ:    4
 (Other)
   :7273
   Event.Type     Location     Dining.Option   Customer.ID Customer.Name
 Payment:7294   Blondo:7298   Mode:logical       :7296         :7296
 Refund :    4                NA's:7298       , :    2      , :    2



 Customer.Reference.ID
   :7296
 , :    2
```

University of Nebraska – Omaha Crime data contains case #, incident code, reported, case status, start occurred, end occurred, building location, stolen

damaged, and description. I will be looking for unique outliers to find significant influencers that might impact the consumer sales count. I will be using this data to see if a crime in the local area on a particular day will impact the model.

```
     Case..                                      Incident.Code              Reported
 Min.   :20150256   MEDICAL EMERGENCY             : 29   2/1/2016 17:25 :  1
 1st Qu.:20160084   SUSPICIOUS PERSON             : 24   2/10/2016 13:30:  1
 Median :20160175   MISC - OTHER                  : 22   2/10/2016 19:34:  1
 Mean   :20160143   ACCIDENTS - P.D. H&R REPORTABLE: 20   2/10/2016 5:01 :  1
 3rd Qu.:20160264   LOST OR STOLEN ITEM           : 14   2/11/2016 10:45:  1
 Max.   :20160348   NARCOTICS - POSSESSION        : 13   2/11/2016 2:03 :  1
                    (Other)                       :195   (Other)        :311
                              Case.Status          Start.Occurred          End.Occurred
 Closed - Cleared by Arrest-Adult    :  9                  : 14                   : 49
 Closed - Cleared by Arrest-Juvenile:  1   2/10/2016 8:00:  2    4/5/2016 16:00 :  2
 Closed - Cleared by Exception      :  2   3/7/2016 9:00 :  2    6/9/2016 7:10  :  2
 Closed - Non-Criminal Case         :  4   4/6/2016 9:00 :  2    6/9/2016 7:30  :  2
 Closed - Unfounded                 :  3   6/9/2016 6:00 :  2    7/14/2016 5:15 :  2
 Open                               :298   7/14/2016 3:28:  2    1/26/2016 20:00:  1
                                           (Other)       :293    (Other)        :259
                             Building
 N/A                             : 33
 Arts & Science Hall             : 16
 HPER                            : 15
 Baxter Arena                    : 12
 Criss Library                   : 12
 Parking Structure 1 (EAST GARAGE): 12
 (Other)                         :217
                                                       Location        Stolen
 222 University Drive East (UNO ACADEMIC BUILDING)      : 12   $0.00     :278
 6323 Maverick Plaza (UNO ACADEMIC BUILDING)            : 10   $200.00   :  3
 310 University Drive East (GOV'T PARKING GARAGE (UNO)) :  8   $400.00   :  3
 6650 University Drive South (GOV'T PARKING GARAGE (UNO)):  8   $150.00   :  2
 6404 Shirley Street (UNO RESIDENCE HALL)               :  7   $600.00   :  2
 (GOV'T PARKING LOT (UNO))                              :  6   $1,200.00:  1
 (Other)                                                :266   (Other)   : 28
      Damaged
 $0.00     :312
 $1,380.00:  1
 $200.00   :  2
 $500.00   :  2



                          Description

                                : 11
 2-15-16 1210, residence at Scott Village building E room 204 reported being harassed.
                                :  1
 2/24/16 1010  - A student reported her wallet was stolen from ASH second floor women's restroo
m.                              :  1
 3-11-16 Health Services Staff reported his vehicle was hit and damaged while parked in lot G.
                                :  1
 3/16/16 0940 a student reported her ring was lost or stolen on campus.
                                :  1
 3/16/16 1235a while on patrol Scott Village a room was investigated for possible alcohol viola
tions. No citations were issued.:  1
 (Other)
                                :301
>
```

Temperature contains date, tempature High (˚F), tempature Low (˚F) precipitation MTD (Inch), percipitation YTD (Inch), snow MTD (Inch), snow YTD (Inch) and rain. The weather is an important variable as you know if it is raining, hot, or any number of weather related activities you may not want to venture out to a food vendor. I will be using this data to see if rain or temperature has an impact on the consumer sales.

Rain Fall is presented as a column in the temperature data set.

```
       Date      Tempature.High...F. Tempature.Low...F. Percipitation.MTD..Inch.
 1/1/2016 :  1   Min.   :33.00       Min.   :13.30      Min.   :0.00
 1/10/2016:  1   1st Qu.:37.90       1st Qu.:18.07      1st Qu.:0.36
 1/11/2016:  1   Median :58.00       Median :33.80      Median :0.79
 1/12/2016:  1   Mean   :57.24       Mean   :35.29      Mean   :1.29
 1/13/2016:  1   3rd Qu.:73.72       3rd Qu.:50.83      3rd Qu.:1.95
 1/14/2016:  1   Max.   :86.70       Max.   :65.00      Max.   :4.76
 (Other)  :176
 Percipitation.YTD..Inch. Snow.MTD..Inch. Snow.YTD..Inch.      Rain
 Min.   : 0.000           Min.   :0.000   Min.   : 0.00   Min.   :0.0000
 1st Qu.: 1.110           1st Qu.:0.120   1st Qu.: 9.37   1st Qu.:0.0000
 Median : 3.580           Median :0.790   Median :16.46   Median :0.0000
 Mean   : 5.226           Mean   :1.570   Mean   :13.14   Mean   :0.1374
 3rd Qu.: 8.850           3rd Qu.:2.882   3rd Qu.:17.52   3rd Qu.:0.0000
 Max.   :15.470           Max.   :6.100   Max.   :17.52   Max.   :1.0000
```

My data is extremely quantitative and far from absolute inclusion. I will not be able to account for any qualitative factors or external events.

Data Wrangling

I began prepared the RStudio environment by installing the necessary packages and loading the libraries.

```
install.packages(tidyr)          library(tidyr)

install.packages(dplyr)          library(dplyr)

install.packages(mice)           library(mice)

install.packages(ggplot2)
    library(ggplot2)
```

```
install.packages("gridExtra")
     library(gridExtra)

install.packages("reshape2")
     library(reshape2)
```

Next I loaded in the relevant data sets with the read.csv command. Once the data sets were loaded I used view() to view review each file and check for continuity. After reviewing the files I determined that I would need the following information

- From each of the stock datasets: Date, Close

- From the Temperatures dataset: Rain, Temperature high

-From the Daily Sales dataset: Date

-From the DailyCrimeLogSummary dataset: Start.Occurred

-From the Fuel Cost dataset: Weekly.U.S..All.Grades.All.Formulations.Retail.Gasoline.Prices...Dollars.per.Gallon.

“select” is used to remove the desired columns from the various datasets. For the stock datasets I used “names” renamed all of the “close” columns to represent respective company. Once I had cleaned the dataset and removed the wanted tiles I merged all of them usinf “left_joinn into one represented dataset “stocks”.

```
     date                Conagra          Walmart          Union            Werner
 Length:254         Min.   :38.70   Min.   :56.42   Min.   :68.79   Min.   :21.41
 Class :character   1st Qu.:41.08   1st Qu.:63.07   1st Qu.:79.41   1st Qu.:24.13
 Mode  :character   Median :42.49   Median :66.50   Median :84.79   Median :25.57
                    Mean   :43.30   Mean   :66.23   Mean   :84.44   Mean   :25.40
                    3rd Qu.:45.75   3rd Qu.:69.84   3rd Qu.:88.75   3rd Qu.:26.83
                    Max.   :48.39   Max.   :74.30   Max.   :97.05   Max.   :28.63

   First.Data
 Min.   : 8.67
 1st Qu.:12.12
 Median :12.96
 Mean   :13.41
 3rd Qu.:15.46
 Max.   :17.80
 NA's   :40
```

The sales data presented a separate set of difficulties as the number of customer sales per day was stored as multiple entries for individual dates. To overcome this I removed the relevant column "date" and applied "dplyr::count" thus reducing the count from 6284 rows to 153 rows.

```
> summary(`daily sales`)
        Date              Time
 1/4/2016 : 190   11:03:33:    6
 2/22/2016: 176   11:16:06:    6
 1/25/2016: 168   11:16:29:    6
 2/8/2016 : 167   11:33:26:    6
 2/29/2016: 159   11:36:28:    6
 1/11/2016: 154   11:43:20:    6
 (Other)  :6284   (Other) :7262
```

```
> summary(Sales)
        Date           n
 1/10/2016:  1   Min.   :  2.0
 1/11/2016:  1   1st Qu.: 21.0
 1/12/2016:  1   Median : 32.0
 1/13/2016:  1   Mean   : 45.9
 1/14/2016:  1   3rd Qu.: 49.0
 1/15/2016:  1   Max.   :190.0
 (Other)  :153
```

With the Temperatures dataset I used "select" to remove the "date" and "Rain" to create "Rain" dataset. To create the "Temp" dataset I used "select" to remove the "Date" and "Tempature.High...F"

```
> summary(Tempatures)
         Date      Tempature.High...F.  Tempature.Low...F.  Percipitation.MTD..Inch.
 1/1/2016 :  1   Min.   :33.00        Min.   :13.30       Min.   :0.00            
 1/10/2016:  1   1st Qu.:37.90        1st Qu.:18.07       1st Qu.:0.36            
 1/11/2016:  1   Median :58.00        Median :33.80       Median :0.79            
 1/12/2016:  1   Mean   :57.24        Mean   :35.29       Mean   :1.29            
 1/13/2016:  1   3rd Qu.:73.72        3rd Qu.:50.83       3rd Qu.:1.95            
 1/14/2016:  1   Max.   :86.70        Max.   :65.00       Max.   :4.76            
 (Other)  :176                                                                    
 Percipitation.YTD..Inch.  Snow.MTD..Inch.  Snow.YTD..Inch.       Rain       
 Min.   : 0.000           Min.   :0.000   Min.   : 0.00   Min.   :0.0000  
 1st Qu.: 1.110           1st Qu.:0.120   1st Qu.: 9.37   1st Qu.:0.0000  
 Median : 3.580           Median :0.790   Median :16.46   Median :0.0000  
 Mean   : 5.226           Mean   :1.570   Mean   :13.14   Mean   :0.1374  
 3rd Qu.: 8.850           3rd Qu.:2.882   3rd Qu.:17.52   3rd Qu.:0.0000  
 Max.   :15.470           Max.   :6.100   Max.   :17.52   Max.   :1.0000  
```

```
> summary(Temp)
         date      Tempature.High...F.
 1/1/2016 :  1   Min.   :33.00       
 1/10/2016:  1   1st Qu.:37.90       
 1/11/2016:  1   Median :58.00       
 1/12/2016:  1   Mean   :57.24       
 1/13/2016:  1   3rd Qu.:73.72       
 1/14/2016:  1   Max.   :86.70       
 (Other)  :176                       
```

```
> summary(Rain)
         date          Rain       
 1/1/2016 :  1   Min.   :0.0000  
 1/10/2016:  1   1st Qu.:0.0000  
 1/11/2016:  1   Median :0.0000  
 1/12/2016:  1   Mean   :0.1374  
 1/13/2016:  1   3rd Qu.:0.0000  
 1/14/2016:  1   Max.   :1.0000  
 (Other)  :176                   
```

With Crime dataset I used “tidyr::separate” to separate the “Start.Occurred” into two columns as it would difficult to count the number of individual days a crime was reported. Once this was accomplished “dplyr::count” was used to identify the number of crimes committed on a particular day.

```
> colnames(DailyCrimeLogSummary)
 [1] "Case.."          "Incident.Code"  "Reported"        "Case.Status"     "Start.Occurred"
 [6] "End.Occurred"    "Building"       "Location"        "Stolen"          "Damaged"       
[11] "Description"    
> colnames(Test)
 [1] "Case.."         "Incident.Code" "Reported"      "Case.Status"   "date"         
 [6] "time"           "End.Occurred"  "Building"      "Location"      "Stolen"       
[11] "Damaged"        "Description"  
```

```
> summary(Crimes)
     date           Number of crimes
 Length:161         Min.   : 1.000  
 Class :character   1st Qu.: 1.000  
 Mode  :character   Median : 1.000  
                    Mean   : 1.969  
                    3rd Qu.: 2.000  
                    Max.   :14.000  
```

I used “names” to ensure continuity in the name of the “date” column across all of the datasets. “names” was also used to change the names of the relevant columns to maintain continuity once the datasets are merged. At this point I feel the datasets are ready to be merged. I used “left_join” to merge the datasets into one.

```
> colnames(ds)
 [1] "date"                "Number of customers" "Conagra"             "Walmart"
 [5] "Union"               "Werner"              "First.Data"          "Rain"
 [9] "Tempature.High...F." "Number of crimes"    "Fuel Price"
```

Lastly I used “mice” to resolve the missing data points I used “complete(mice())” .

p

```
> summary(ds)
     date           Number of customers    Conagra         Walmart          Union      
 Length:159         Min.   :  2.0       Min.   :38.70   Min.   :60.84   Min.   :68.79  
 Class :character   1st Qu.: 21.0       1st Qu.:41.74   1st Qu.:65.88   1st Qu.:78.16  
 Mode  :character   Median : 32.0       Median :43.99   Median :67.41   Median :80.14  
                    Mean   : 45.9       Mean   :43.35   Mean   :67.09   Mean   :80.22  
                    3rd Qu.: 49.0       3rd Qu.:45.24   3rd Qu.:68.83   3rd Qu.:83.00  
                    Max.   :190.0       Max.   :47.15   Max.   :71.28   Max.   :89.63  
                                        NA's   :51      NA's   :51      NA's   :51     
     Werner        First.Data         Rain        Tempature.High...F. Number of crimes
 Min.   :21.41   Min.   : 8.67   Min.   :0.0000   Min.   :33.00       Min.   :1.000   
 1st Qu.:24.25   1st Qu.:11.91   1st Qu.:0.0000   1st Qu.:36.75       1st Qu.:1.000   
 Median :25.89   Median :12.65   Median :0.0000   Median :53.90       Median :2.000   
 Mean   :25.43   Mean   :12.54   Mean   :0.1447   Mean   :54.12       Mean   :2.019   
 3rd Qu.:26.77   3rd Qu.:13.22   3rd Qu.:0.0000   3rd Qu.:69.25       3rd Qu.:3.000   
 Max.   :28.48   Max.   :15.95   Max.   :1.0000   Max.   :82.40       Max.   :7.000   
 NA's   :51      NA's   :51                                           NA's   :55      
   Fuel Price   
 Min.   :1.834  
 1st Qu.:1.954  
 Median :2.135  
 Mean   :2.133  
 3rd Qu.:2.295  
 Max.   :2.482  
 NA's   :136    
> summary(ids1)
    Conagra         Walmart          Union           Werner        First.Data   
 Min.   :38.70   Min.   :60.84   Min.   :68.79   Min.   :21.41   Min.   : 8.67  
 1st Qu.:41.55   1st Qu.:64.80   1st Qu.:76.15   1st Qu.:23.66   1st Qu.:11.90  
 Median :43.77   Median :67.17   Median :80.01   Median :25.88   Median :12.69  
 Mean   :43.11   Mean   :66.75   Mean   :79.70   Mean   :25.30   Mean   :12.62  
 3rd Qu.:45.09   3rd Qu.:68.80   3rd Qu.:82.69   3rd Qu.:26.80   3rd Qu.:13.29  
 Max.   :47.15   Max.   :71.28   Max.   :89.63   Max.   :28.48   Max.   :15.95  
> summary(ids2)
 Number of crimes   Fuel Price   
 Min.   :1.000    Min.   :1.834  
 1st Qu.:1.000    1st Qu.:1.938  
 Median :2.000    Median :2.109  
 Mean   :2.126    Mean   :2.107  
 3rd Qu.:3.000    3rd Qu.:2.265  
 Max.   :7.000    Max.   :2.482  
```

Method / Proof

In an effort to ascertain the relevant data. I used linear regression to correlate the various values in the dataset. After multiple regression I was able to ascertain that the most relevant data set is that of “Rain”.

```
> summary(modelds)

Call:
lm(formula = ds1$`Number of customers` ~ ds1$Tempature.High...F. +
    ds1$Rain + ds1$Conagra + ds1$Walmart + ds1$Union + ds1$Werner +
    ds1$First.Data)

Residuals:
   Min     1Q Median     3Q    Max
-66.02 -23.25  -9.75  10.52 125.94

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              95.5281    93.8844   1.018  0.31054
ds1$Tempature.High...F.  -0.3729     0.2419  -1.541  0.12534
ds1$Rain                 27.1870     9.4039   2.891  0.00441 **
ds1$Conagra              -3.8276     2.7947  -1.370  0.17286
ds1$Walmart               0.1508     2.1585   0.070  0.94441
ds1$Union                 1.5030     0.9966   1.508  0.13359
ds1$Werner                0.2352     2.2567   0.104  0.91713
ds1$First.Data           -0.3302     2.5544  -0.129  0.89733
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.04 on 151 degrees of freedom
Multiple R-squared:  0.1336,	Adjusted R-squared:  0.09344
F-statistic: 3.326 on 7 and 151 DF,  p-value: 0.002532
```

To ensure that I had not missed any other possible were not overlooked I used “cor” to create a correlation matrix.

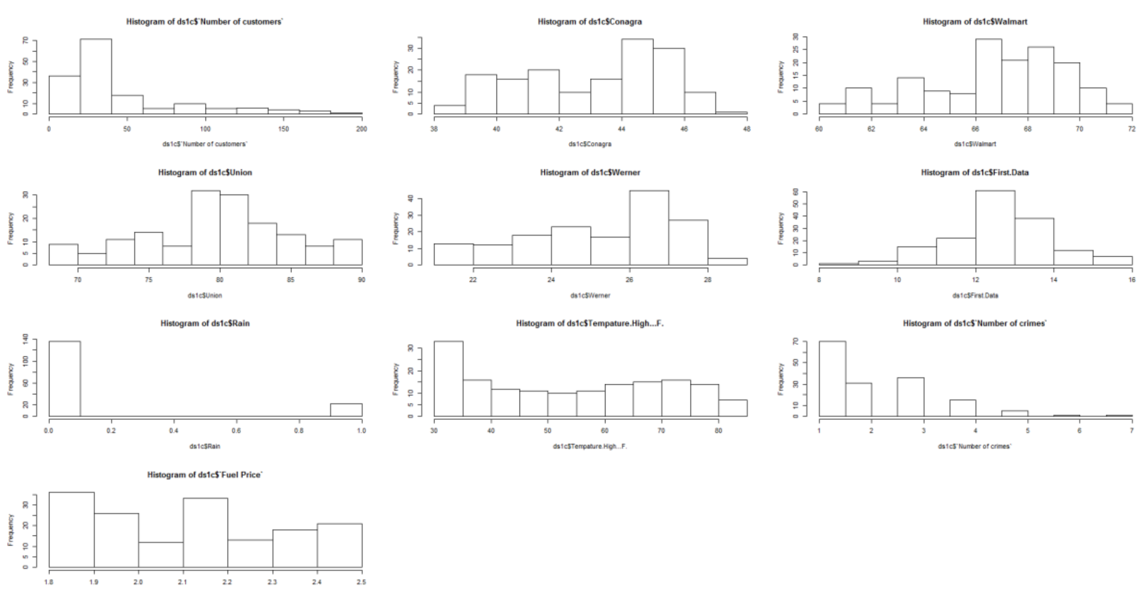
	Number of customers	Conagra	Walmart	Union	Werner
Number of customers	1.00000000	-0.169583110	-0.12044470	-0.07837886	-0.03178105
Conagra	-0.16958311	1.000000000	0.82445776	0.79339179	0.53317061
Walmart	-0.12044470	0.824457758	1.00000000	0.70026762	0.61083454
Union	-0.07837886	0.793391791	0.70026762	1.00000000	0.52222956
Werner	-0.03178105	0.533170606	0.61083454	0.52222956	1.00000000
First.Data	0.03229189	-0.339222296	-0.28519703	-0.34916387	-0.35410979
Rain	0.29465672	-0.157025367	-0.13489165	-0.18139707	-0.11855038
Tempature.High...F.	-0.26769756	0.491997112	0.36349527	0.42058513	0.02826956
Number of crimes	-0.16981699	0.061771538	0.03500631	0.07017178	0.10708547
Fuel Price	-0.04301254	0.006053378	-0.07666582	-0.01999653	-0.07639721

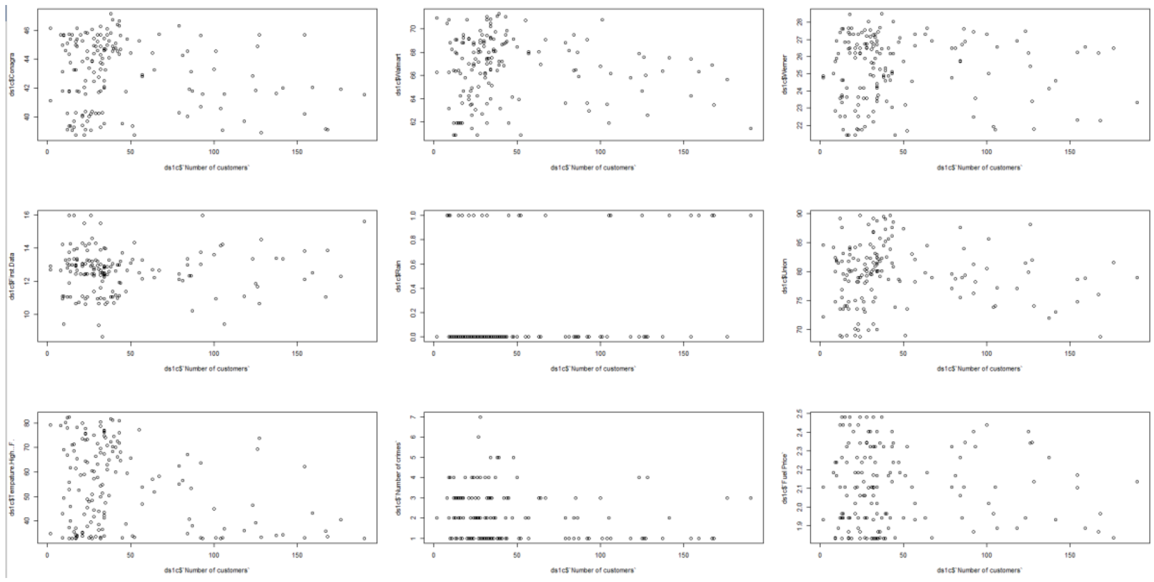
	First.Data	Rain	Tempature.High...F.	Number of crimes
Number of customers	0.03229189	0.29465672	-0.26769756	-0.16981699
Conagra	-0.33922230	-0.15702537	0.49199711	0.06177154
Walmart	-0.28519703	-0.13489165	0.36349527	0.03500631
Union	-0.34916387	-0.18139707	0.42058513	0.07017178
Werner	-0.35410979	-0.11855038	0.02826956	0.10708547
First.Data	1.00000000	0.04596759	-0.19053558	0.01060962
Rain	0.04596759	1.00000000	-0.37028999	0.01580007
Tempature.High...F.	-0.19053558	-0.37028999	1.00000000	-0.02533107
Number of crimes	0.01060962	0.01580007	-0.02533107	1.00000000
Fuel Price	0.14368027	-0.08561123	0.14764108	-0.29093477

	Fuel Price
Number of customers	-0.043012536
Conagra	0.006053378
Walmart	-0.076665824
Union	-0.019996531
Werner	-0.076397211
First.Data	0.143680271
Rain	-0.085611228
Tempature.High...F.	0.147641080
Number of crimes	-0.290934770
Fuel Price	1.000000000

Visuals

I used “hist” and “plot” in order to visualize the data. I used “par” in order to organize the histograms and plots.





Summary

I found a real interesting correlation or the lack there of. The only data that was correlational was rain. There is a clear correlation that in days without rain there is a significant increase in sales. The remaining data set had little to no impact on the customer sale this is actually a very exciting news. The information shows that the customer count clusters around 46 and typically remains steady. The lack of influence from the other data indicates that this product is not influenced by market economical fluctuations.

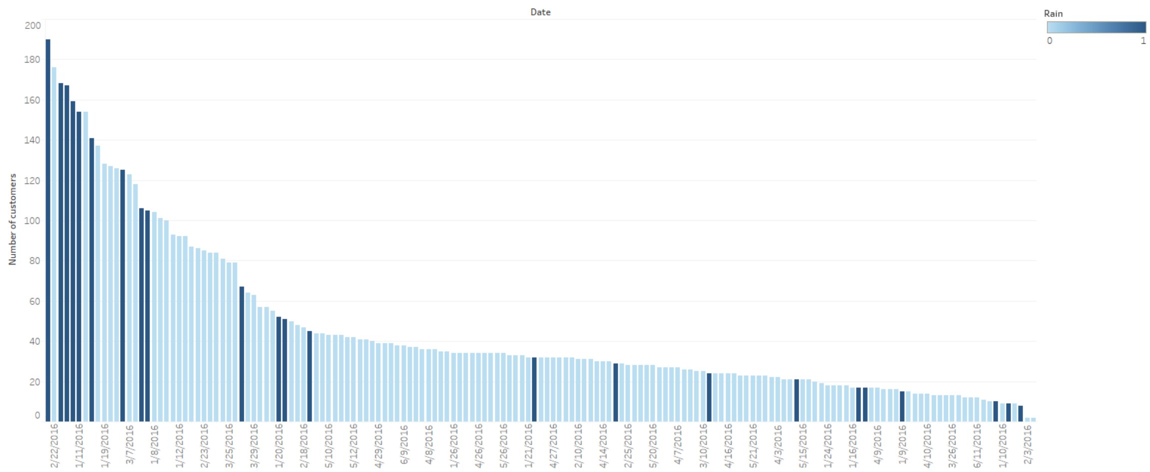
Recommendations

After this exhaustive review of relevant data I have multiple recommendations for the restraint owner.

It is my express recommendation that the owner plan to support 46 customers each day. This number will maintain, allowing for your ordering to be equally level.

It is my express recommendation that the owner observe weather predictions for rain.
On days without any rain there is a significant spike in customer count.

Number of customers vs Rain by date



Appendix A: RStudio Preprocessing Script

```
# this is the cap1 sript file
```

```
# install packages

install.packages(tidyr)
install.packages(dplyr)
install.packages(mice)
install.packages(Hmisc)
install.packages("gridExtra")
install.packages("reshape2")

#Load libraries

library(tidyr)
library(dplyr)
library(mice)
library(gridExtra)
library(ggplot2)
library(reshape2)


# Import data sets

 Conag <-
read.csv("D:/School/springboard/cap/Conag.csv")
```

```
First.Data <-
read.csv("D:/School/springboard/cap/First Data.csv")

wern1 <-
read.csv("D:/School/springboard/cap/wern1.csv")

Union <-
read.csv("D:/School/springboard/cap/Union.csv")

WMT <-
read.csv("D:/School/springboard/cap/WMT.csv")

`daily sales` <-
read.csv("D:/School/springboard/cap/items-2016-01-01-
2017-01-01.csv")

DailyCrimeLogSummary <-
read.csv("D:/School/springboard/cap/DailyCrimeLogSummary.
csv")

Fuel.Cost <-
read.csv("D:/School/springboard/cap/Fuel Cost.csv")

Tempatures <-
read.csv("D:/School/springboard/cap/Tempatures.csv")



# Wrangle Data


# View data to ensure continuity


View(Conag)

View(First.Data)

View(wern1)
```

```
View(Union)
View(WMT)
View('daily sales')
View(DailyCrimeLogSummary)
View(Fuel.Cost)
View(Tempatures)

summary(Conag)
summary(First.Data)
summary(wern1)
summary(Union)
summary(WMT)
summary('daily sales')
summary(DailyCrimeLogSummary)
summary(Fuel.Cost)
summary(Tempatures)


# Clean stock data sets to flter out unwanted
columns

Con <- select(Conag, date, close)
First <- select(First.Data, date, close)
Wer <- select(wern1, date, close)
```

```
UP <- select(Union, date, close)
Wal <- select(WMT, date, close)

#renamed close price columns in stock price data
sets to represent each individual company

names(Wal)[2] <- "Walmart"
names(UP)[2] <- "Union"
names(Wer)[2] <- "Werner"
names(First)[2] <- "First.Data"
names(Con)[2] <- "Conagra"

#Join all stock data sets into one data set

S1 <- left_join(Wal,UP, by = "date")
S2 <- left_join(Wer,First, by = "date")
S3 <- left_join(Con,S1, by = "date")
Stocks <- left_join(S3,S2, by = "date")

# clean sales data

'daily sales' <-
select(items.2016.01.01.2017.01.01, Date)
```

```
Sales<-dplyr::count(`daily sales`,Date)

names(Sales)[1] <- "date"

names(Sales)[2] <- "Number of customers"

# clean Tempatures

'Rain' <- select(Tempatures, Date, Rain)

names(Rain)[1] <- "date"

'Temp' <- select(Tempatures, Date,
Tempature.High...F.)

names(Temp)[1] <- "date"

# Clean Crime Dataset

Test <- tidyr::separate(DailyCrimeLogSummary,
Start.Occurred, c("date", "time" ),sep=" ")

't' <- select(Test, date)
```

```
Crimes<-dplyr::count(`t`,date)

names(Crimes)[2] <- "Number of crimes"

names(Crimes)[1] <- "date"


# Clean fuel data set

names(Fuel.Cost)[2] <- "Fuel Price"

names(Fuel.Cost)[1] <- "date"

# View wrangled datasets to validate and check for
continuity

View(Stocks)
View(`daily sales`)
view(Temp)
View(Rain)
View(Crimes)
View(Fuel.Cost)

summary(Stocks)
```

```
summary(Sales)

summary(Temp)

summary(Rain)

summary(Crimes)

summary(Fuel.Cost)



# Merging all of the datasets into one data set


m1 <- left_join(Sales,Stocks, by = "date")

m2 <- left_join(m1,Rain, by = "date")

m3 <- left_join(m2,Temp, by = "date")

m4 <- left_join(m3,Crimes, by = "date")

ds <- left_join(m4, Fuel.Cost, by = "date")


# Recitation


str(ds)


summary(ds)


# Multiple Imputation
```

```
sds1 <- ds[c("Conagra", "Walmart", "Union",
"Werner", "First.Data")]

sds2 <- ds[c("Number of crimes", "Fuel Price")]

set.seed(123)

ids1 <- complete(mice(sds1))

ids2 <- complete(mice(sds2))

summary(ids1)

summary(ids2)

# Create a new ds to keep seperation

ds1 = ds

# import imputed data back into new ds1 and ds2

ds1$Conagra = ids1$Conagra
```

```
ds1$Walmart = ids1$Walmart

ds1$Union = ids1$Union

ds1$Werner = ids1$Werner

ds1$First.Data = ids1$First.Data

ds1$`Number of crimes` = ids2$`Number of crimes`

ds1$`Fuel Price` = ids2$`Fuel Price`

summary(ds1)

str(ds1)

# Linear regression

model1 = lm(ds1$`Number of customers` ~
ds1$Tempature.High...F. )

summary(model1)

model1$residuals
```

```
sse1 = sum(model1$residuals^2)

sse1

model2 =lm(ds1$`Number of customers` ~ ds1$Tempature.High...F.+ ds1$Rain)

summary(model2) # rain is the only relevant factor

sse2 = sum(model2$residuals^2)

sse2

modelds = lm(ds1$`Number of customers` ~ ds1$Tempature.High...F.+ ds1$Rain + ds1$Conagra+ ds1$Walmart + ds1$Union + ds1$Werner + ds1$First.Data)

summary(modelds)

sseds = sum(modelds$residuals^2)

sseds

# Improving the model using coefficients
```

```
	modelds1 = lm(ds1$`Number of customers` ~
ds1$Tempature.High...F.+ ds1$Rain + ds1$Conagra +
ds1$Union + ds1$Werner + ds1$First.Data)

	summary(modelds1)

	modelds2 = lm(ds1$`Number of customers` ~
ds1$Tempature.High...F.+ ds1$Rain + ds1$Conagra +
ds1$Union + ds1$Werner)

	summary(modelds2)

	modelds3 = lm(ds1$`Number of customers` ~
ds1$Tempature.High...F.+ ds1$Rain + ds1$Conagra +
ds1$Union)

	summary(modelds3)

	# corralation matrix, set seed and remove date in
order to allow correlation

	ds1c <- select(ds1, -date)

	set.seed(123)

	cor(ds1c)

	 #EDA
```

```
#- Histograms

par(mfrow=c(4,3))
hist(ds1c$`Number of customers`)
hist(ds1c$Conagra)
hist(ds1c$Walmart)
hist(ds1c$Union)
hist(ds1c$Werner)
hist(ds1c$First.Data)
hist(ds1c$Rain)
hist(ds1c$Tempature.High...F.)
hist(ds1c$`Number of crimes`)
hist(ds1c$`Fuel Price`)

#scatter plots

par(mfrow=c(3,3))
plot(ds1c$`Number of customers`, ds1c$Conagra)
plot(ds1c$`Number of customers`, ds1c$Walmart)
plot(ds1c$`Number of customers`, ds1c$Union)
plot(ds1c$`Number of customers`, ds1c$Werner)
plot(ds1c$`Number of customers`, ds1c$First.Data)
plot(ds1c$`Number of customers`, ds1c$Rain)
```

```
plot(ds1c$`Number of customers`,
Ads1c$Tempature.High...F.)

plot(ds1c$`Number of customers`, ds1c$`Number of
crimes`)

plot(ds1c$`Number of customers`, ds1c$`Fuel
Price`)
```