Open*Mart Transaction History Analysis Using Only R

Nicholas Knauer
The Pennsylvania State University
May 16th, 2014

Table of Contents

Page #

1. Introduction……...4

2. Methods…….....4
   2.1 Hierarchal Clustering Method……...4
   Table 2.1(a): Total Units Bought of Item 5 Per Customer ID……...4
   Table 2.1(b): Cluster # Classification for Each Customer…….....5
   2.2 Demographic Breakdown……..6
   2.3 Exponential Smoothing……6
   Table 2.3(a): Holt-Winters Prediction - 1st 3 Months After 2-Year Span.7
   Table 2.3(b): Holt-Winter Prediction Error.....7

3. Results…….....8
   3.1 Hierarchal Cluster Dendogram at Height of 140.....8
   3.2 Plot of Amount of Customers in Each Cluster……...9
   3.3 Number of Units Bought By Cluster Number……..10
   3.4 Demographic Classifications…….....11
   3.5 Percentage of Each Demographic Category…..……...13
   3.6 Line Graph of Daily Amount of Units Bought Over 2 Years……...15
   3.7 Holt-Winters Filtering Graph w/ Prediction Lines……....16
   3.8 Difference of Logarithms Graph……...17
   3.9 Linear Filtering Graph w/ Weekly, Monthly, Quarterly Filtered Lines....18

4. Recommendations…….19

5. Conclusion…….21

6. Appendix……...23
   6.1 R Code for Entire Analysis……...24
   6.2 Data Dictionary for Retail Store……...28

7. Bibliography……32

1. Introduction

With the need to properly organize Open*Mart and understanding future demands, the board is hoping to obtain some information regarding the forecasting of one of their items and the buying patterns of its customers. In this case, the item being analyzed is ***cereal*** (Item 5). The ultimate goal is to determine whom to advertise cereal to as well as to discovering trends in sales to determine which is the best season to stock cereal at Open*Mart. The data received from Open*Mart contains the transaction and demographic details regarding each purchase during a two-year time span of all their customers.

2. Methods

In order to accomplish a full analysis of Item 5 purchases and get results that will maximize Open*Mart's revenue, three methods were used: hierarchal clustering method, demographic breakdown, and exponential smoothing methods. Using all these methods as a whole helped forecast sales of Item 5 and gives Open*Mart an idea on how to market Item 5 as well.

2.1 *Hierarchal Clustering Method*

Using just the transaction data in R, three aspects of the entire dataset were analyzed to make a hierarchal clustering dendogram: the Customer ID, the Item Type, and the number of Units Bought. After creating a table of these 3 aspects in R, another table was created based off of this new table. This new table contains a list of unique Customer ID's and the number of Item 5 Units Bought during the two-year time span for each customer. The table is below:

Total Units Bought of Item 5 per Customer ID

	row.names	5
1	15100503	2
2	15100677	12
3	15100768	2
4	15101246	3
5	15101519	61
6	15102004	24
7	15102202	0
8	15103002	0
9	15103887	5
10	15104307	11

(10 of 349 Customers)

TABLE 2.1(a)

From this point, hierarchal clustering was applied to the number of units bought. The linkage criteria used was "Ward's method". This method entails that it takes the two most closely related attributes, links them, and from there it finds the next closest related attribute from this linkage. So for example, Customer ID 15100503 and 15100768 both bought only one unit of Item 5 throughout the two-year span, therefore they would be grouped together. After making all these links, in the end there will only be one link to all the items almost like a pyramid structure. Hence, this is why it is called hierarchal clustering. On the hierarchal cluster dendogram, a line cutting the dendogram at a height of 140 was placed. This height was chosen because it seemed that if the dendogram was cut any higher, the amount of clusters would be too small and the data results would be very vague. If the dendogram was cut at a smaller height, there would have been too many clusters and the group that would be analyzed would be relatively small and the analysis results wouldn't have that much meaning for Open*Mart. This hierarchal cluster dendogram can be found in *Section 3.1*. Eight clusters were created from the cutoff height of 140 and below is a table representing the Customers in each cluster:

Cluster # Classification for Each Customer

	row.names	x
1	15100503	1
2	15100677	2
3	15100768	1
4	15101246	3
5	15101519	4
6	15102004	5
7	15102202	1
8	15103002	1
9	15103887	3
10	15104307	2

(10 of 349 Customers)

TABLE 2.1(b)

After creating the hierarchal cluster dendogram, the next step was to determine which was the best cluster to analyze. In order to do this, each group of customers were separated by cluster and analyzed separately, thus eight separate groups because there were eight clusters. Within each cluster, the total units bought of Item 5 were to be found within the two-year time span. The cluster with the most units bought was considered the best cluster to analyze because that cluster contains the customers whom are helping Open*Mart create the most revenue for Item 5. At this point, all of the clusters are separated by number of units bought for Item 5, and the analysis of the cluster with the most units bought can begin. A barplot of all the clusters with the best cluster highlighted in green is in *Section 3.2*.

2.2 *Demographic Breakdown*

Now that all of the clusters are separated by number of units bought, the cluster with the most units bought of Item 5 was analyzed. Only the demographics data was taken into account for this analysis. To match the demographic subsets with the customers in the cluster with the most units bought, a function called 'aggregate' was used in R. This tool was very useful because its output was a list of each demographic attribute classifications (ethnicity, family size, income, # of children, etc.) with a unique customer id. This was great because now one can look at each demographic group and determine what percentage bought Item 5. So for example, if you look at the 'family size' group, you can determine what percentages were a family size of 1, 2, 3, 4, 5, or 6 that bought Item 5. This can be useful in determining where to advertise Item 5. Therefore, if a certain family size bought the majority of Item 5, you can focus on that group in advertising. In *Section 3.3*, the demographic classification for each customer is shown and in *Section 3.4*, the percentages are shown for each demographic classification per customer who bought Item 5.

2.3 *Exponential Smoothing*

After looking at specific groups to determine advertising targets, the next step is to see how Item 5 sells for the two-year span for all customers. With this information, one would be able to see seasonal trends throughout the time span, difference in logarithms to see some irregularities in sales, linear filtering of purchases, and prediction of future sales.

The baseline of this analysis was to first create a line graph of all the Item 5 Units purchased on a daily scale. Since there were 104 weeks worth of data, a more accurate approach would be to analyze the sales by day, which ended up being 728 days. This line graph can be found in *Section 3.5*. Once this was accomplished a more in-depth approach was implemented to get more accurate results for a variety of statistical means. A difference of logarithms function was applied in R in order to see the variation of sales. The numbers that were farther away from zero meant that they were considered outliers and should be disregarded if you want a more accurate calculation whether it be average number of sales, seasonality results, etc. The difference in logarithm results are located in *Section 3.6.*

A predicting statistical method could be applied as well to the original data set of number of units bought daily. The approach used was the Holt-Winters method. The Holt-Winters method takes a weighted average of the previous estimates of the component's value and the value suggested by the new sales figure. The weights used are called the smoothing constants. The method will adapt more quickly to genuine changes in the sales pattern, but the only downside to this method is that it might also overreact to freak sales figures. In all, Holt-Winters was a good method to use prediction for this set of data. Below is a prediction for the next 92 days (3 months) and this can be seen visually in *Section 3.7*:

TABLE 2.3(a)

Holt-Winters Prediction for 1st 3 Months After 2-Year Span

	fit						
1	116.39603	23	83.13849	47	42.95672	71	130.51253
2	97.11729	24	104.46702	48	89.37963	72	121.16892
3	98.44614	25	123.48062	49	71.41864	73	82.69867
4	86.93486	26	133.32164	50	87.09854	74	96.61572
5	87.19467	27	114.07658	51	88.86756	75	88.22349
6	94.71321	28	110.81364	52	129.48352	76	88.36254
7	114.50994	29	88.17035	53	109.02443	77	102.44343
8	103.84314	30	93.55174	54	111.28120	78	91.16720
9	102.23035	31	102.91274	55	105.84489	79	102.82622
10	97.90906	32	133.20485	56	62.13182	80	128.67519
11	120.13487	33	112.33988	57	86.58730	81	122.34287
12	101.66769	34	118.41971	58	93.19177	82	106.68296
13	121.45884	35	88.55581	59	126.32095	83	104.74259
14	97.99126	36	115.16275	60	86.74211	84	108.03679
15	114.03282	37	131.88325	61	98.15780	85	92.61580
16	87.33254	38	159.76151	62	89.26916	86	108.95839
17	90.22430	39	157.41114	63	93.08955	87	100.74468
18	106.54198	40	97.74616	64	96.79877	88	109.59410
19	110.51934	41	110.03262	65	108.81947	89	96.70005
20	84.76361	42	89.83296	66	128.43836	90	113.57148
21	109.41129	43	95.76521	67	112.64154	91	125.24890
22	84.55603	44	105.82560	68	119.27242	92	87.77892
		45	106.92779	69	97.62822		
		46	137.73207	70	101.29614		

Along with these predictions, the accuracy of these predictions is something of interest. Below is a table of the smoothing parameters and the error of Table 3's data compared with the average amount of items sold per day (103.77):

TABLE 2.3(b)

Holt-Winter Prediction Error

Smoothing parameters: alpha: 0.03051736; beta: 0; gamma: 0.369717

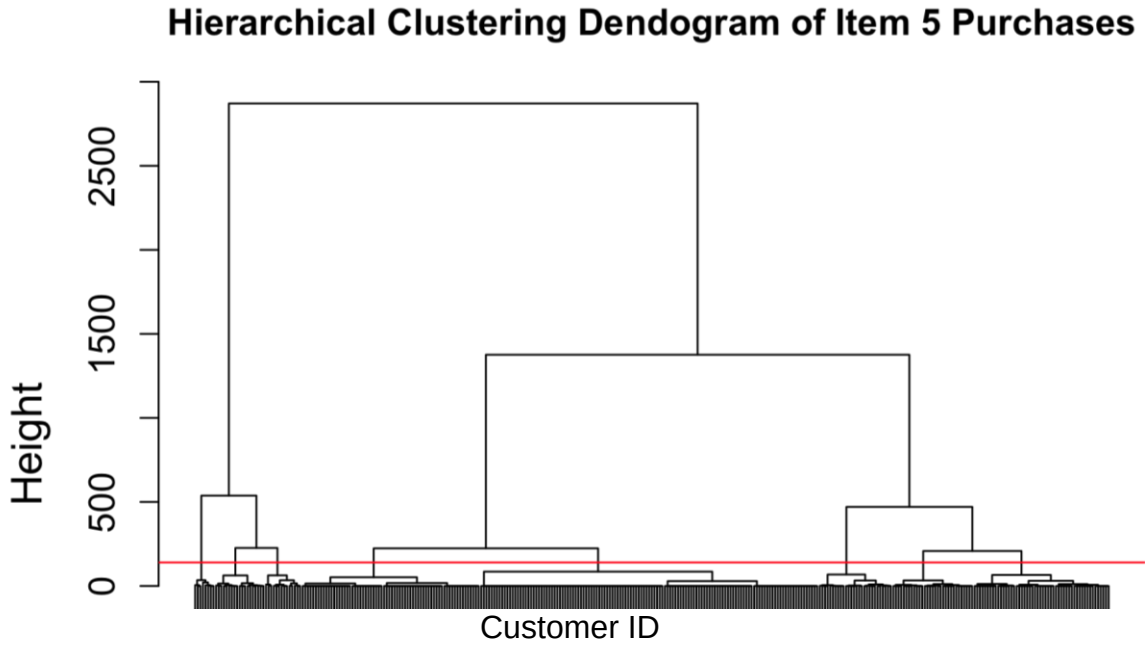
[,1]	s15 10.141136465	s31 -1.100482172	s47 -61.178037851
a 103.777738688	s16 -16.566736911	s32 29.184036033	s48 -14.771730849
s1 12.610696507	s17 -13.682574148	s33 8.311460620	s49 -32.739310792
s2 -6.675642087	s18 2.627508214	s34 14.383698022	s50 -17.067005947
s3 -5.354384120	s19 6.597277232	s35 -15.487799248	s51 -16.097585830
s4 -16.873265707	s20 -19.166057502	s36 11.111547441	s52 25.230781320
s5 -16.621047987	s21 5.474028421	s37 27.824446750	s53 4.844090222
s6 -9.110110868	s22 -19.388830628	s38 55.695114331	s54 7.093265999
s7 10.679027155	s23 -20.813963106	s39 53.337146363	s55 1.649354368
s8 0.004635449	s24 0.506968524	s40 -6.335431852	s56 -42.071312250
s9 -1.615758941	s25 19.512973642	s41 5.943438716	s57 -17.623426823
s10 -5.944637371	s26 29.346394439	s42 -14.263823839	s58 -11.026555341
s11 16.273571630	s27 10.093746169	s43 -8.339166403	s59 22.095032994
s12 -2.201200822	s28 6.823203777	s44 1.713629985	s60 -17.491407381
s13 17.582347010	s29 -15.827677573	s45 2.808222784	s61 -6.083307861
s14 -5.892827949	s30 -10.453886212	s46 33.604902328	s62 -23.979541793

s63	-11.166750876	s78	-13.203041111
s64	-7.563132448	s79	-1.551621218
s65	4.547971571	s80	24.289749670
s66	24.159265112	s81	17.949837977
s67	8.354850686	s82	2.282329020
s68	14.978134411	s83	0.334362125
s69	-6.681656378	s84	3.620965459
s70	-3.103340315	s85	-11.807622011
s71	26.195455580	s86	4.527375928
s72	16.844255704	s87	-3.693929036
s73	-21.633595513	s88	5.147896163
s74	-7.724138018	s89	-7.753757688
s75	-16.123962603	s90	9.110082937
s76	-15.992514963	s91	20.779901360
s77	-1.919218777	s92	-16.697674855

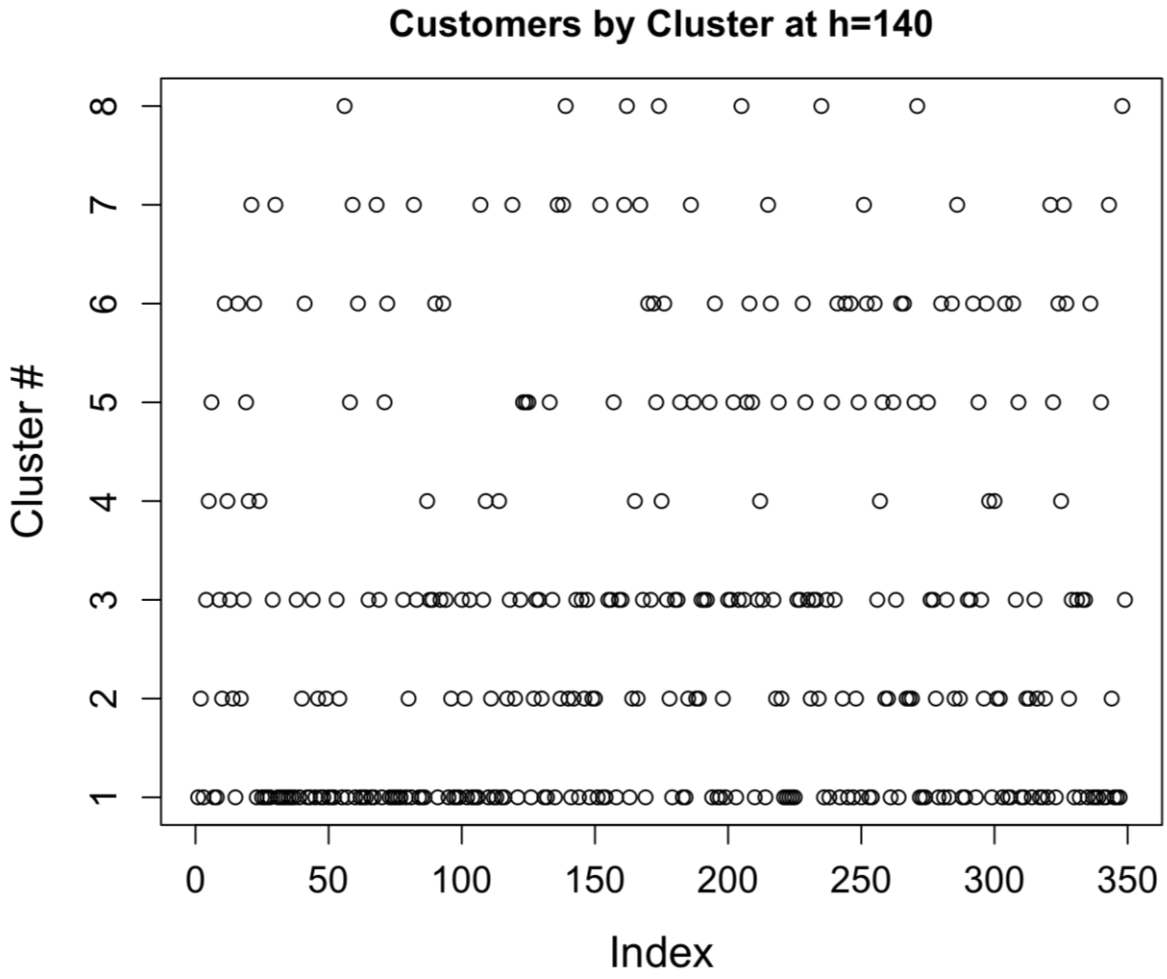
Finally linear filtering was used as well to help with the visual aspect of seeing the trends in daily sales. In order to do this, I used a function called 'filter' in R, which applies linear filtering to a univariate time series or to each series separately of a multivariate time series. It simply applies moving averages with equal weights across the graph. Within this function, I also created 3 separate filters: weekly, monthly, and quarterly. Each filter gives a different interpretation to the graph and in this case, quarterly is what was examined because it gave a clear trend on the graph. The graph of these 3 filters can be found in *Section 3.8*.

3. Results

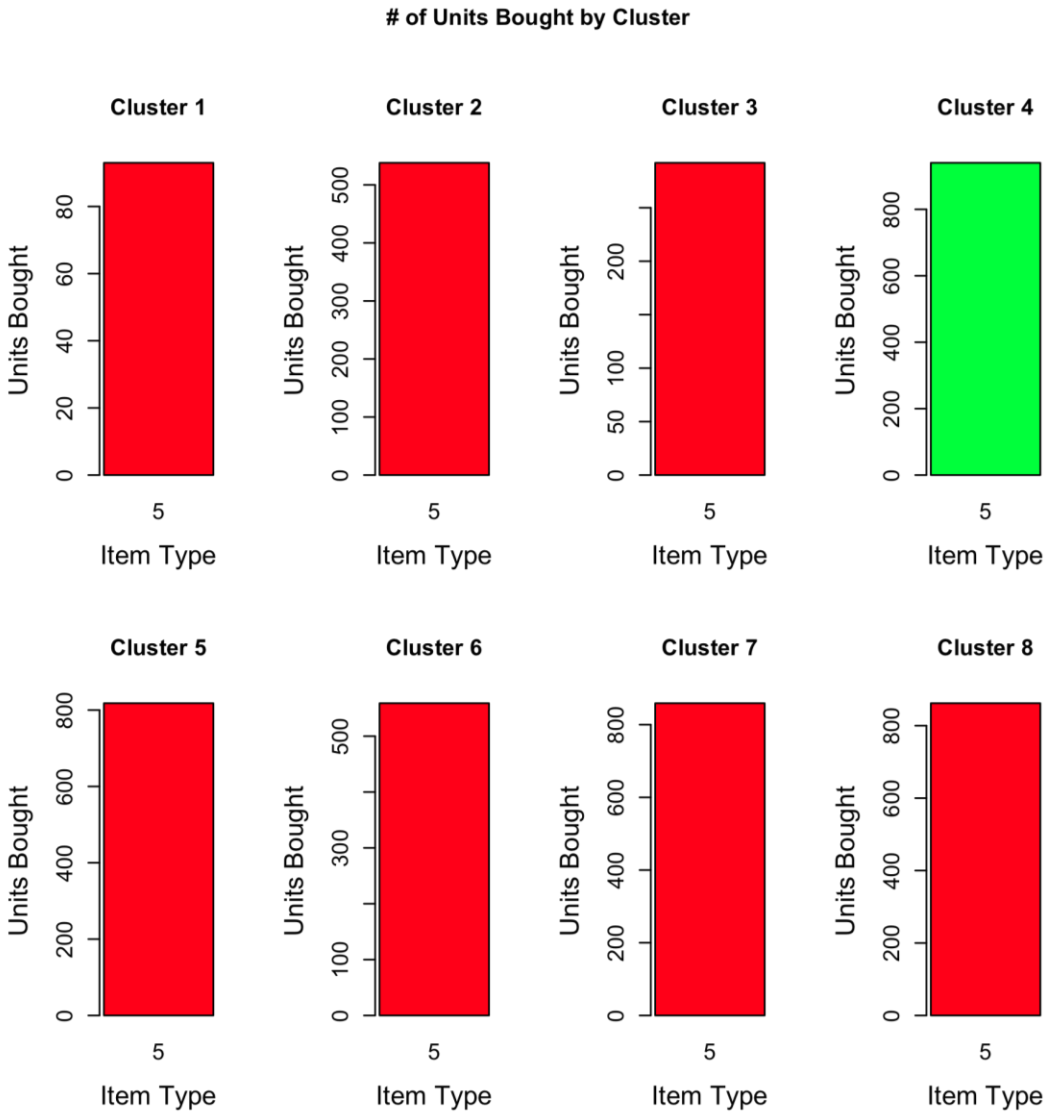
3.1 *Hierarchal Cluster Dendogram at Height of 140.*



3.2 *Plot of Amount of Customers in Each Cluster*



3.3 *Number of Units Bought By Cluster Number*



Cluster 1= 93 Units	Cluster 2=538 Units	Cluster 3=292 Units	**Cluster 4=940 Units**
Cluster 5 = 818 Units	Cluster 6=559 Units	Cluster 7=859 Units	Cluster 8=862 Units

3.4 *Demographic Classifications*

Female Occupation

	Customer.ID	Female.Occupation
1	15101519	10
2	15104398	4
3	15105965	13
4	15107169	10
5	15500074	4
6	15503847	4
7	15504167	4
8	15515742	4
9	15517862	4
10	15524504	10
11	15536094	10
12	15560615	2
13	15561118	13
14	15805564	4

Female Education

	Customer.ID	Female.Education
1	15101519	4
2	15104398	4
3	15105965	6
4	15107169	2
5	15500074	4
6	15503847	5
7	15504167	6
8	15515742	5
9	15517862	5
10	15524504	4
11	15536094	7
12	15560615	8
13	15561118	3
14	15805564	5

Female Work Hours

	Customer.ID	Female.Work.Hours
1	15101519	4
2	15104398	3
3	15105965	1
4	15107169	4
5	15500074	2
6	15503847	3
7	15504167	3
8	15515742	2
9	15517862	2
10	15524504	4
11	15536094	4
12	15560615	3
13	15561118	1
14	15805564	3

Female Age

	Customer.ID	Female.Age.
1	15101519	6
2	15104398	3
3	15105965	2
4	15107169	6
5	15500074	4
6	15503847	4
7	15504167	4
8	15515742	5
9	15517862	4
10	15524504	6
11	15536094	6
12	15560615	5
13	15561118	6
14	15805564	3

Male Occupation

	Customer.ID	Male.Occupation
1	15101519	10
2	15104398	5
3	15105965	11
4	15107169	10
5	15500074	1
6	15503847	7
7	15504167	7
8	15515742	1
9	15517862	2
10	15524504	10
11	15536094	11
12	15560615	6
13	15561118	7
14	15805564	2

Male Education

	Customer.ID	Male.Education
1	15101519	5
2	15104398	8
3	15105965	9
4	15107169	2
5	15500074	4
6	15503847	3
7	15504167	4
8	15515742	6
9	15517862	8
10	15524504	2
11	15536094	9
12	15560615	4
13	15561118	2
14	15805564	5

Male Work Hours

	Customer.ID	Male.Work.Hours
1	15101519	4
2	15104398	3
3	15105965	5
4	15107169	4
5	15500074	3
6	15503847	3
7	15504167	3
8	15515742	3
9	15517862	3
10	15524504	4
11	15536094	5
12	15560615	3
13	15561118	3
14	15805564	3

Male Age

	Customer.ID	Male.Age.
1	15101519	6
2	15104398	3
3	15105965	7
4	15107169	6
5	15500074	4
6	15503847	4
7	15504167	4
8	15515742	6
9	15517862	4
10	15524504	6
11	15536094	7
12	15560615	5
13	15561118	6
14	15805564	3

Dogs

	Customer.ID	Dogs
1	15101519	0
2	15104398	2
3	15105965	0
4	15107169	0
5	15500074	1
6	15503847	0
7	15504167	0
8	15515742	0
9	15517862	1
10	15524504	1
11	15536094	0
12	15560615	1
13	15561118	0
14	15805564	0

Cats

	Customer.ID	Cats
1	15101519	0
2	15104398	0
3	15105965	2
4	15107169	0
5	15500074	0
6	15503847	0
7	15504167	0
8	15515742	0
9	15517862	0
10	15524504	0
11	15536094	0
12	15560615	1
13	15561118	0
14	15805564	0

TVs

	Customer.ID	TVs
1	15101519	9
2	15104398	3
3	15105965	9
4	15107169	2
5	15500074	2
6	15503847	3
7	15504167	3
8	15515742	9
9	15517862	3
10	15524504	2
11	15536094	9
12	15560615	9
13	15561118	3
14	15805564	3

Income

	Customer.ID	Income
1	15101519	7
2	15104398	9
3	15105965	1
4	15107169	1
5	15500074	6
6	15503847	8
7	15504167	8
8	15515742	11
9	15517862	11
10	15524504	5
11	15536094	8
12	15560615	7
13	15561118	5
14	15805564	8

Family Size

	Customer.ID	Family.Size
1	15101519	2
2	15104398	6
3	15105965	5
4	15107169	2
5	15500074	4
6	15503847	4
7	15504167	2
8	15515742	3
9	15517862	4
10	15524504	2
11	15536094	2
12	15560615	2
13	15561118	2
14	15805564	4

Ethnicity

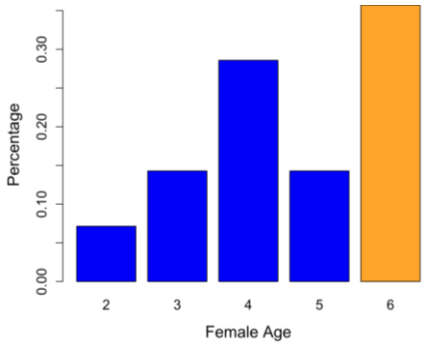
	Customer.ID	Ethnicity
1	15101519	1
2	15104398	1
3	15105965	1
4	15107169	1
5	15500074	1
6	15503847	1
7	15504167	1
8	15515742	1
9	15517862	1
10	15524504	1
11	15536094	1
12	15560615	1
13	15561118	1
14	15805564	1

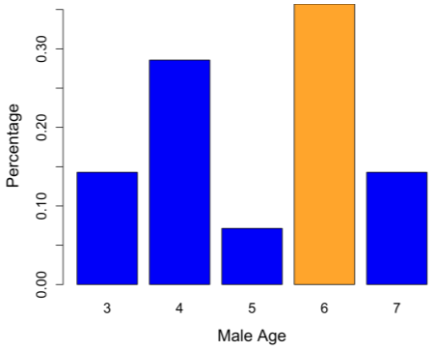
Children

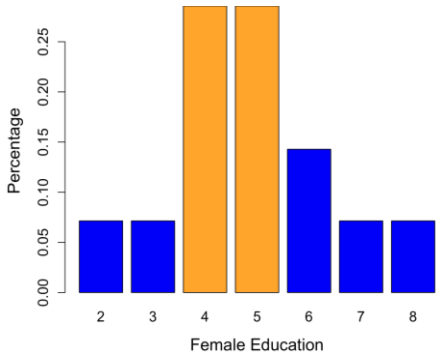
	Customer.ID	Children
1	15101519	8
2	15104398	3
3	15105965	6
4	15107169	8
5	15500074	8
6	15503847	8
7	15504167	8
8	15515742	8
9	15517862	3
10	15524504	8
11	15536094	8
12	15560615	8
13	15561118	8
14	15805564	3

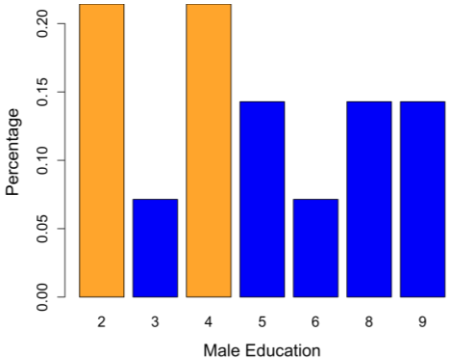
3.5 *Percentage of Each Demographic Category*

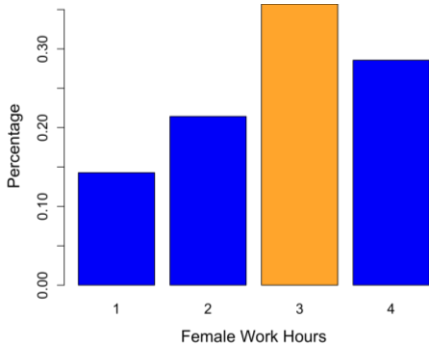
(Refer to Data Dictionary for Number Meanings on X-Axis's)

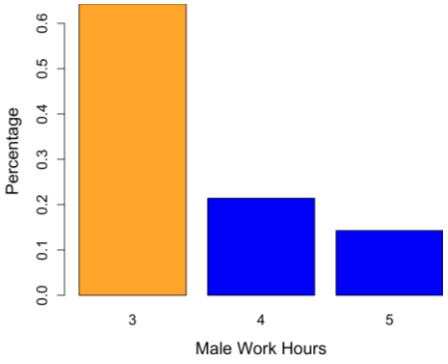


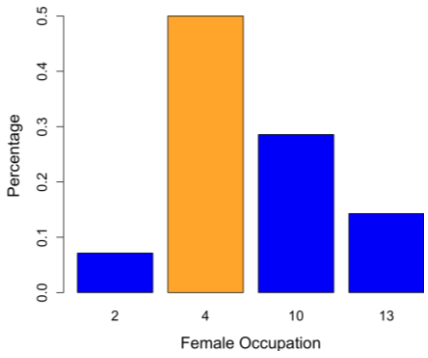


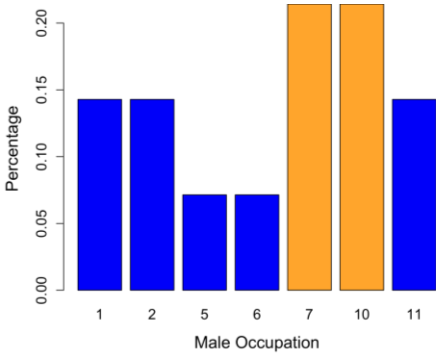


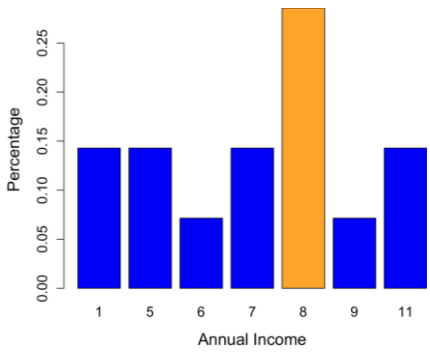


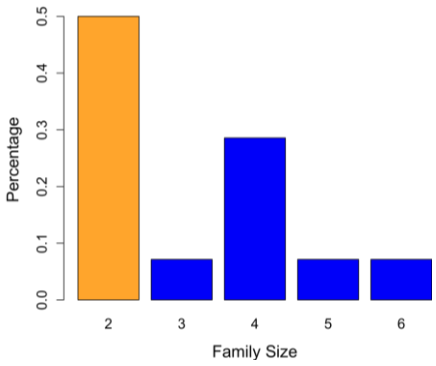


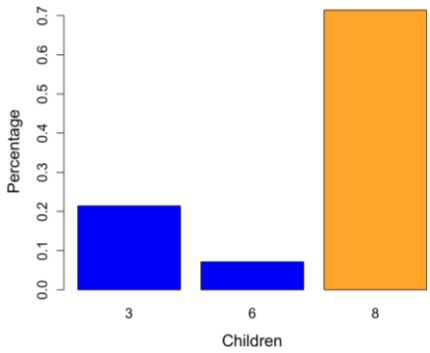


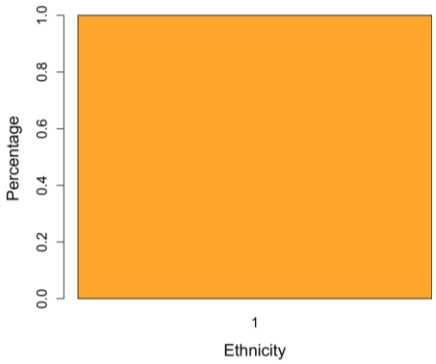


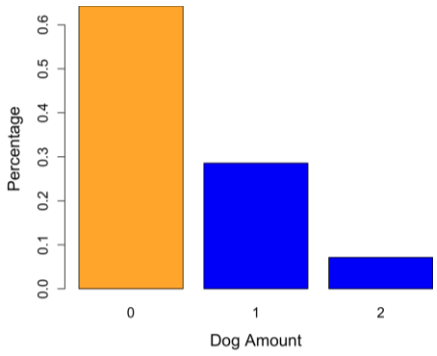


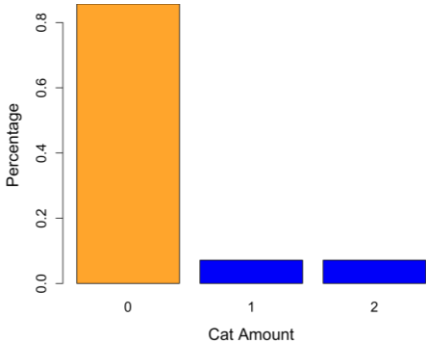

Percentage
0.00
0.05
0.10
0.15
0.20
0.25
1
5
6
7
8
9
11
Annual Income

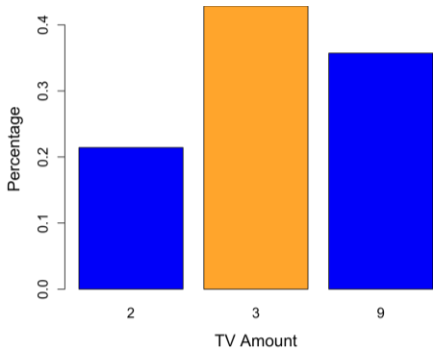

Percentage
0.0
0.1
0.2
0.3
0.4
0.5
2
3
4
5
6
Family Size


Percentage
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
3
6
8
Children

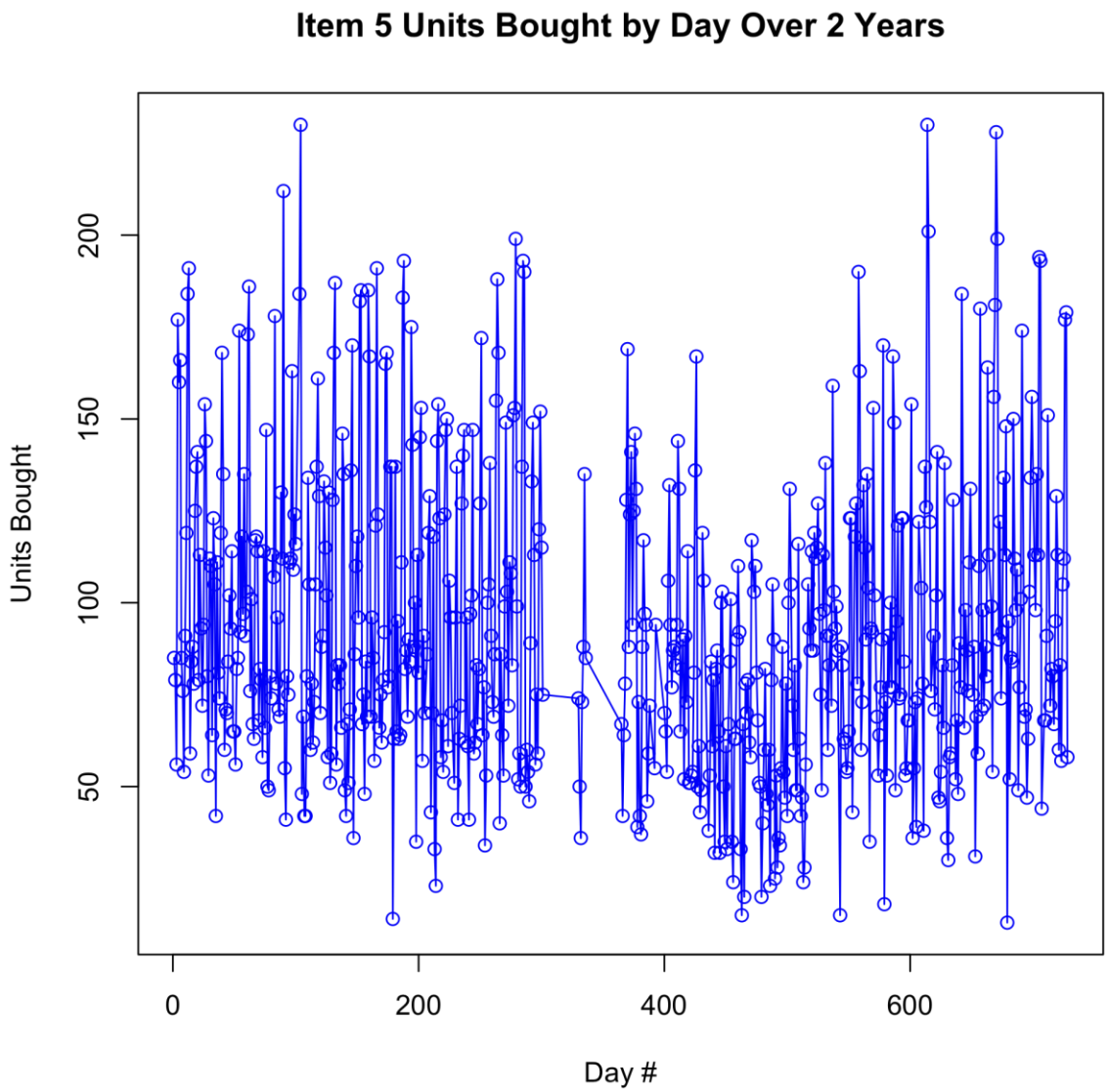

Percentage
0.0
0.2
0.4
0.6
0.8
1.0
1
Ethnicity


Percentage
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
Dog Amount

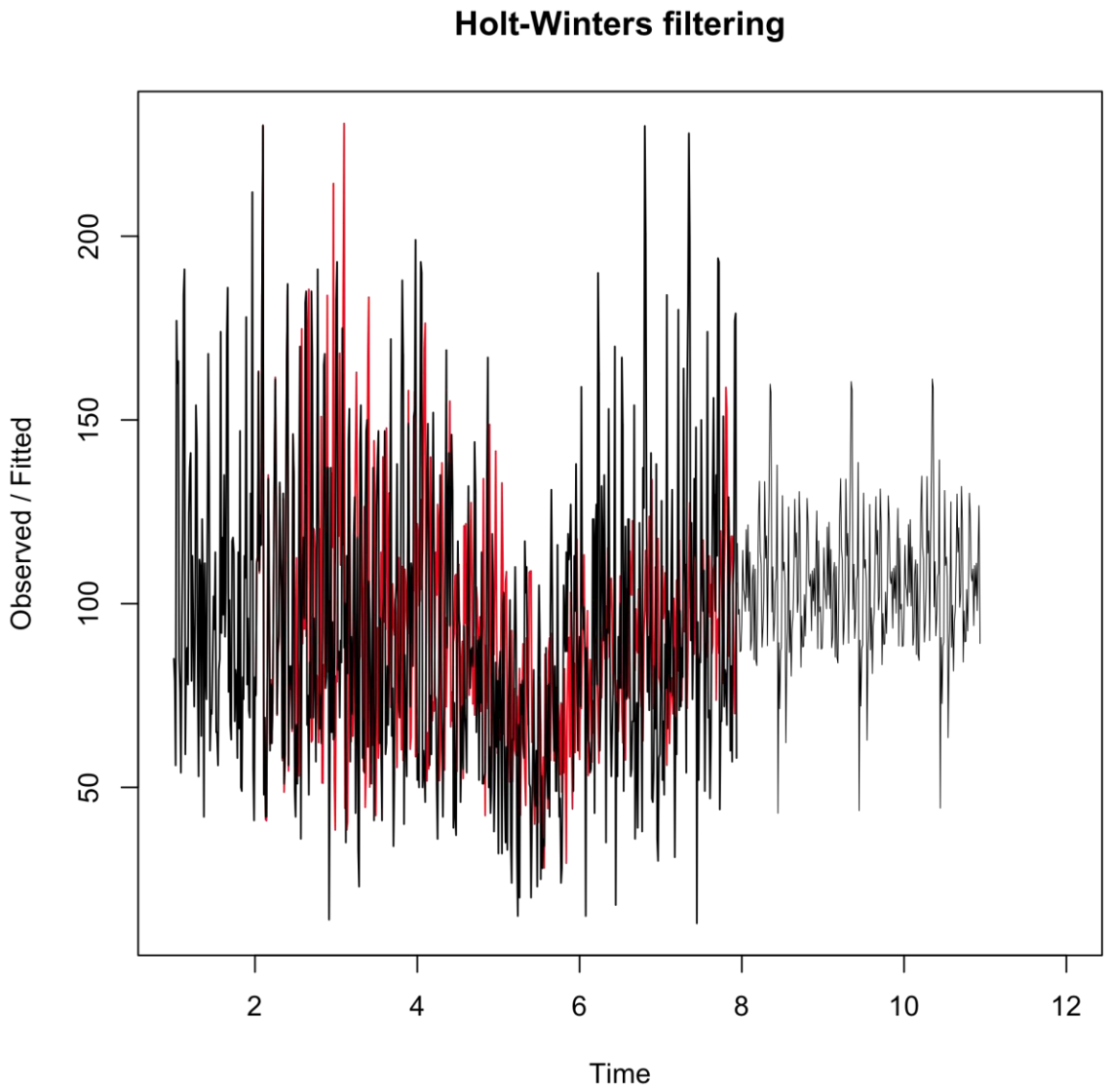

Percentage
0.0
0.2
0.4
0.6
0.8
0
1
2
Cat Amount


Percentage
0.0
0.1
0.2
0.3
0.4
2
3
9
TV Amount

3.6 ***Line Graph of Daily Amount of Units Bought Over 2 Years***

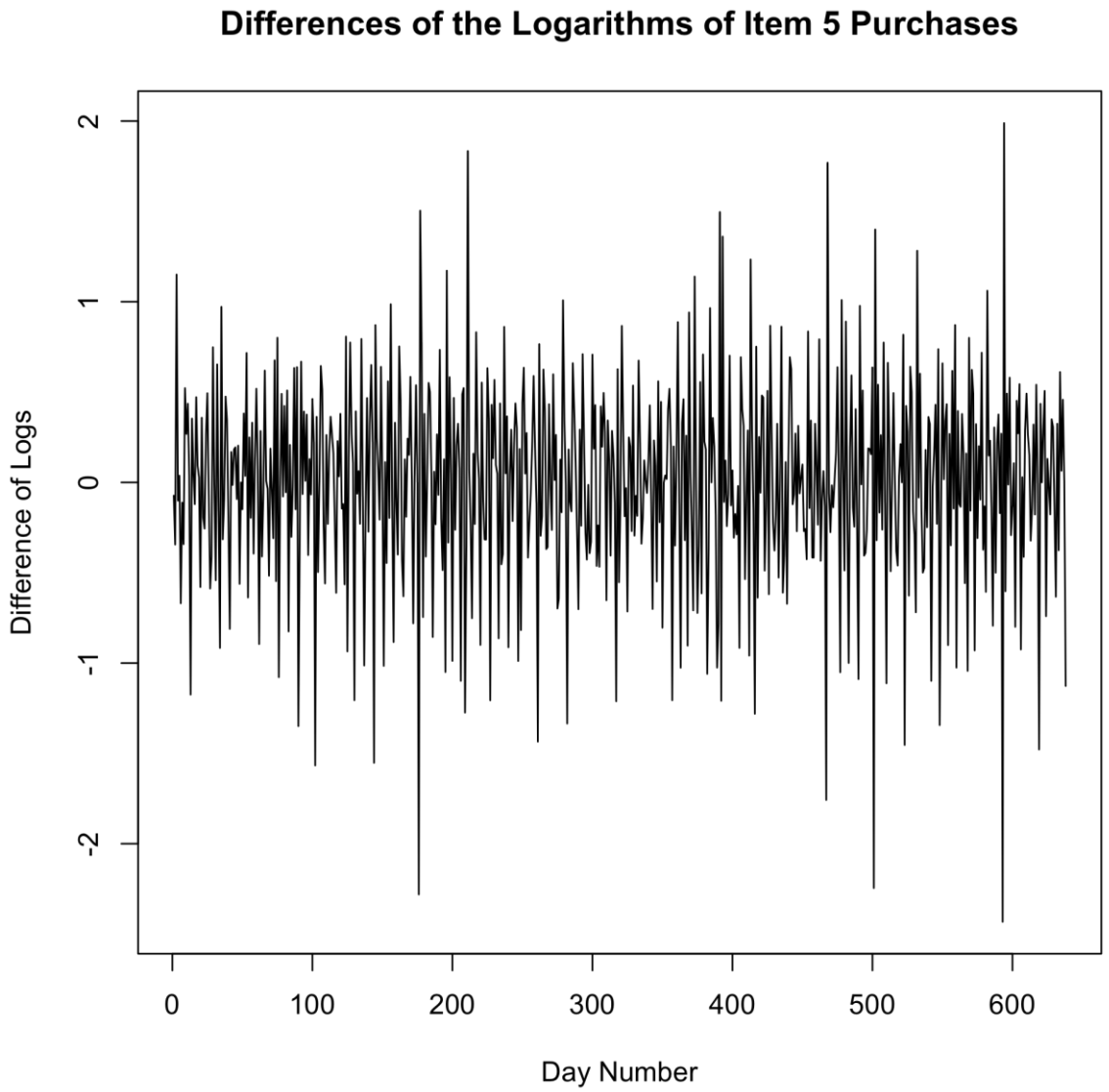


3.7 *Holt-Winters Filtering Graph w/ Prediction Lines*

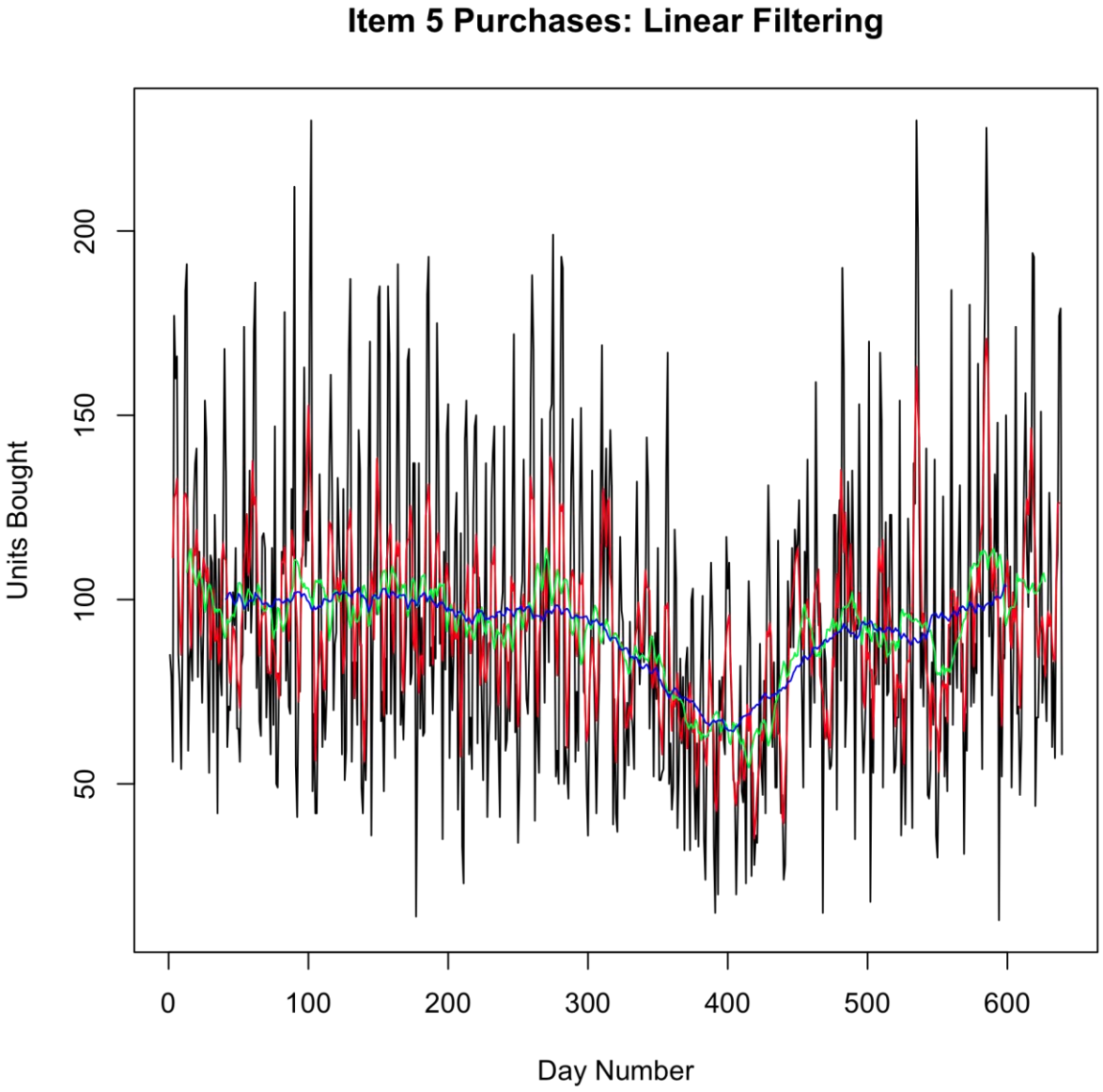


Prediction Time period is from Time 8-11 (Next 276 Days / 9 Months)

3.8 *Difference of Logarithms Graph*



3.9 *Linear Filtering Graph w/ Weekly, Monthly, Quarterly Filtered Lines*



Closing values and averages for Weekly (red), Monthly (green) and Quarterly (blue)
FIGURE 8

4. Recommendations

After creating a hierarchal cluster dendogram, determining the best cluster to analyze from this dendogram based on number of units bought, and finally evaluating the demographic percentages within this best cluster, a target customer base can be made to market Item 5 more. Based on the combination of these analytic methods, the following conclusions can be made about the customer base that bought Item 5 the most:

35.7% of females and males were between the ages of 65 to 99
28% were females who graduated with a high school degree or a college degree
21.4% of males graduated from either grade school or high school
35% of females work 35 hours or more per week
64.2% of males work 35 hours or more per week
50% of females have an occupation of clerical
21.4% of males have an occupation of either a laborer or retired
28.57% of a household makes an annual income of $45,000-$54,999
50% of households have a family size of 2 people
71.4% of households have children over 18 years old
100% of households are white
64.28% of households do not have dogs
85.7% of households do not have cats
42.8% have 3 or more TV's

Based off of these results, it may be advantageous for Open*Mart to make two separate targets for marketing Item 5: females and males. For females, since such a high percentage have a clerical occupation, then most likely they will be sitting behind a computer for most of the day. If Open*Mart were to extend their advertising of Item 5 to the internet, then that could really help their profits because that is the most likely place females will see advertisements for all companies during their 35 hr or more per week job. As for males, it was mostly evenly distributed but the customers who had occupations of a laborer or being retired bought the most. One suggestion would be to determine when a retired or laborer bought the most of Item 5 during the year. For someone that is retired, the most popular week to buy Item 5 was on week 84. Along with this week, other popular weeks to buy Item 5 for retirees were weeks 1, 4, 14, 41, and 85. A discount for Item 5 could be offered during these weeks and Open*Mart can have an even bigger customer base of retirees to increase Item 5 sales.

If Open*Mart is a chain store, they can use the demographic results found to target certain areas in the country to increase advertisement of Item 5. Based on the 2012 US Census data, states that have an average annual income between 45,000-54,999 are *Arizona, Florida, Georgia, Idaho, Indiana, Iowa, Kansas, Maine, Michigan, Missouri, Montana, Nebraska, Nevada, North Carolina, North Dakota, Ohio, Oregon, Pennsylvania, Rhode Island, South Dakota, Texas, Vermont, Wisconsin, and Wyoming*. Based on the 2010 US Census data, states that have the highest percentage of males and females with an average age of 65 and older in a household are *Florida, Iowa, Maine, North Dakota, Pennsylvania, Rhode Island, and West Virginia*. According to the 2010 US

Census, the states with the highest number of whites are *New York, Florida, California, Texas, Pennsylvania*. As a result from this set of census data, it seems that Pennsylvania and Florida would be the best areas to advertise Item 5 based on age, ethnicity, and average annual income.

If one were to determine the best time of year to market Item 5, the day with the most units sold is of interest. During the two-year time span, the top five days with the most Item 5 units sold were Day 90 (212 Units), Day 104 (230 Units), Day 614 (230 Units), Day 615 (201 Units), and Day 670 (228 Units). If one wanted to look at the seasonality sales, Open*Mart could order more items during a specific season that has many Item 5 units sold. In order to create the seasons, I divided up the days into every 92 days (3 months) to make up 7 seasons over the two-year time span. During the two year progress a negative trend can be seen for average number of units sold per day: Season 1 (100.6 Units), Season 2 (100.7 Units), Season 3 (96.6 Units), Season 4 (95.0 Units), Season 5 (81.3 Units), Season 6 (73.8 Units), and Season 7 (91.1 Units). If they used a specific advertising methodology during Season 1 or 2 that was different than the previous months, Open*Mart should revert back to advertising template. The most units sold was during the first two seasons.

From here, prediction of future sales using Holt Winters Method was used to determine when to stock Item 5 the most after this two year period. This is more accurate for sure than just seeing the averages during time two-year time span and just replicating it. From the results of the Holt Winters Method, it was proven that during the first 92 days after the two-year period, Day 38 will have the most units sold which is during Week 8. The least successful day for Item 5 Unit sales will be Day 47, so Open*Mart could really focus on that day to try and increase sales.

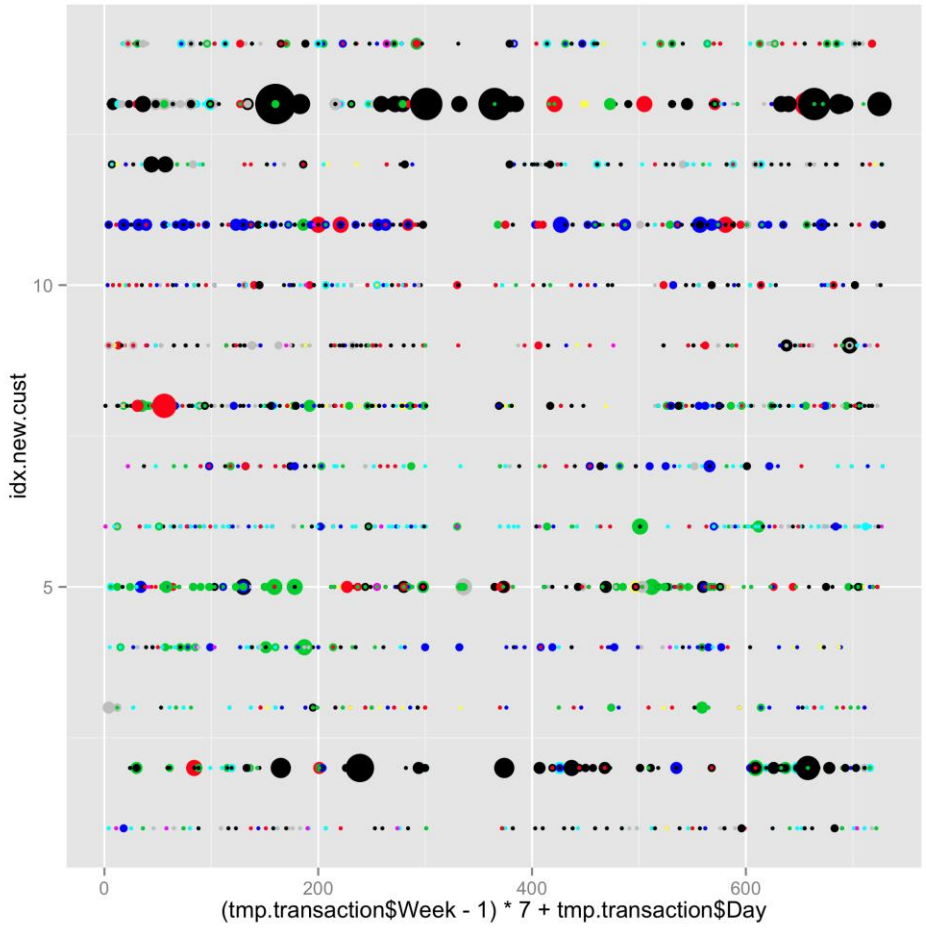
Examining the difference in logarithms graph, the variability can be determined between the number of units sold daily. The closer the results are to a zero value, the less variability occurs. So it would be advantageous to know which days are skewed beyond the average. Based on the results, Day 176, Day 501, Day 593 had the lowest number of sales and could be considered an outlier when analyzing the data. On the other hand, Day 211, Day 468, and Day 594 have the most sales. So when you are trying to figure out the average number of sales during a certain period of time, Open*Mart should void these days because it could skew the data.

Another concept that was looked at is the linear filtering of the number of units bought during the two-year time span. A graph was created in Figure 8 and three different linear filters were created based on moving averages. The quarterly moving average gave the best general trend of the Item 5 units bought. From Day 300 to Day 400, there was a negative trend. Those days can be analyzed in more depth by Open*Mart to determine how to increase sales during that time frame.

The potential for benefiting Open*Mart using all these analysis tools is huge. If Open*Mart really wanted to know in more detail all of their item sales, that can be done

as well. On the next page, an example of a *Beautiful Graph* is created with all of the items being looked at:

4.1 *Beautiful Graph of all Items Sold During Two Years*



This graph represents all of the transactions completed during 2 years. The x-axis is the time period and the y-axis is the customer id. Each circle represents every transaction on a daily scale and the color represents the item type sold. Also the size of the circle indicates how many units of that specified item was sold in that transaction.

5. Conclusion

Using all of these statistical results could definitely help Open*Mart with their customer base increase and help the rise in profits for Item 5/cereal. To review, the following statistical methods were used in this overall analysis on Item 5:

- Hierarchal Clustering
- Demographic Breakdown

- Time Series Visualization
- Holt-Winters Prediction Method
- Difference of Logarithms
- Linear Filtering

Also being aware of all the trends that occur with Item 5 sales can help keep Open*Mart on track and detect where loss in profits are coming from. Whether that be from a decrease in sales from a certain customer demographic to something beyond that, it is very beneficial to have these results on hand.

Open*Mart can ultimately save a great deal of money in the long run with these statistical tools, marketing success could increase drastically, and their competition could be in for a big surprise when these results are implemented.

6. Appendix

6.1 *R Code for Entire Analysis*

```
## Hierarchal Clustering Code
transaction<- read.csv("Transactiondata (2).csv")
subset<-subset(transaction, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(subset)
transaction2<-subset[good,]
CU.list<-sort(unique(subset$Customer.ID))
m<-matrix(0,nrow=length(CU.list), ncol=19)
for(i in 1:length(CU.list)){
for(j in 1:19)
m[i,j]<-sum(subset$Units.Bought[subset$Customer.ID==CU.list[i] & subset$Item.Type==j])
}
m1<-as.data.frame(m)
names(m1)<-1:19
row.names(m1)<-CU.list
Items5<-subset(m1, select = c(5))
tmp<-Items5
tmp<- Items5[,-1]
tmp[is.na(tmp)]<-0
fithclust<-hclust(dist(tmp), method="ward")
plot(fithclust)

## Plot of Customers By Cluster
new2<-cutree(fithclust, h=140)
plot(new2)

## Determining # of Units Bought By Cluster
m.8<-m1[new2==8,]
trans.8<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.8)),]
demographics<- read.csv("DemographicDatas (2).csv")
m.8demog<- merge(trans.8, demographics)
m.8subset<-subset(m.8demog, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(m.8subset)
m.8subset2<-m.8subset[good,]
CU.list2<-sort(unique(m.8subset2$Customer.ID))
m2<-matrix(0,nrow=length(CU.list2), ncol=19)
for(i in 1:length(CU.list2)){
for(j in 1:19)
m2[i,j]<-sum(m.8subset2$Units.Bought[m.8subset2$Customer.ID==CU.list2[i] & m.8subset2$Item.Type==j])
}
m3<-as.data.frame(m2)
names(m3)<-1:19
row.names(m3)<-CU.list2
m.8Item<- subset(m3, select = c(5))
cluster8totals<-colSums(m.8Item)
m.1<-m1[new2==1,]
trans.1<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.1)),]
m.1demog<- merge(trans.1, demographics)
m.8subset<-subset(m.8demog, select = c(Customer.ID, Item.Type, Units.Bought))
m.1subset<-subset(m.1demog, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(m.1subset)
m.1subset2<-m.1subset[good,]
CU.list3<-sort(unique(m.1subset2$Customer.ID))
m4<-matrix(0,nrow=length(CU.list3), ncol=19)
for(i in 1:length(CU.list3)){
for(j in 1:19)
m4[i,j]<-sum(m.1subset2$Units.Bought[m.1subset2$Customer.ID==CU.list3[i] & m.1subset2$Item.Type==j])
}
m5<-as.data.frame(m4)
names(m5)<-1:19
row.names(m5)<-CU.list3
m.1Item<- subset(m5, select = c(5))
cluster1totals<-colSums(m.1Item)
m.2<-m1[new2==2,]
```

```
trans.2<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.2)),]
m.2demog<- merge(trans.2, demographics)
m.2subset<-subset(m.2demog, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(m.2subset)
m.2subset2<-m.2subset[good,]
CU.list4<-sort(unique(m.2subset2$Customer.ID))
m6<-matrix(0,nrow=length(CU.list4), ncol=19)
for(i in 1:length(CU.list4)){
for(j in 1:19)
m6[i,j]<-sum(m.2subset2$Units.Bought[m.2subset2$Customer.ID==CU.list4[i] & m.2subset2$Item.Type==j])
}
m7<-as.data.frame(m6)
names(m5)<-1:19
names(m7)<-1:19
row.names(m7)<-CU.list4
m.2Item<- subset(m7, select = c(5))
cluster2totals<-colSums(m.2Item)
m.3<-m1[new2==3,]
trans.3<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.3)),]
m.3demog<- merge(trans.3, demographics)
m.3subset<-subset(m.3demog, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(m.3subset)
m.3subset2<-m.3subset[good,]
CU.list5<-sort(unique(m.3subset2$Customer.ID))
m8<-matrix(0,nrow=length(CU.list5), ncol=19)
for(i in 1:length(CU.list5)){
for(j in 1:19)
m8[i,j]<-sum(m.8subset2$Units.Bought[m.8subset2$Customer.ID==CU.list5[i] & m.2subset2$Item.Type==j])
}
for(i in 1:length(CU.list5)){
for(j in 1:19)
m8[i,j]<-sum(m.3subset2$Units.Bought[m.3subset2$Customer.ID==CU.list5[i] & m.3subset2$Item.Type==j])
}
m9<-as.data.frame(m8)
names(m9)<-1:19
row.names(m9)<-CU.list5
m.3Item<- subset(m9, select = c(5))
cluster3totals<-colSums(m.3Item)
m.4<-m1[new2==4,]
trans.4<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.4)),]
m.4demog<- merge(trans.4, demographics)
m.4subset<-subset(m.4demog, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(m.4subset)
m.4subset2<-m.4subset[good,]
CU.list6<-sort(unique(m.4subset2$Customer.ID))
m10<-matrix(0,nrow=length(CU.list6), ncol=19)
for(i in 1:length(CU.list6)){
for(j in 1:19)
m10[i,j]<-sum(m.4subset2$Units.Bought[m.4subset2$Customer.ID==CU.list6[i] & m.4subset2$Item.Type==j])
}
m11<-as.data.frame(m10)
names(m11)<-1:19
row.names(m11)<-CU.list6
m.4Item<- subset(m11, select = c(5))
cluster4totals<- colSums(m.4Item) ##Total Amount of Units Bought for Items 3 and 5 for Cluster 4
for(i in 1:length(CU.list6)){
for(j in 1:19)
m10[i,j]<-sum(m.4subset2$Units.Bought[m.4subset2$Customer.ID==CU.list6[i] & m.4subset2$Item.Type==j])
}
m11<-as.data.frame(m10)
names(m11)<-1:19
row.names(m11)<-CU.list6
m.4Item<- subset(m11, select = c(5))
cluster4totals<- colSums(m.4Item) ##Total Amount of Units Bought for Items 3 and 5 for Cluster 4
m.5<-m1[new2==5,]
trans.5<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.5)),]
m.5demog<- merge(trans.5, demographics)
m.5subset<-subset(m.5demog, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(m.5subset)
m.5subset2<-m.5subset[good,]
```

```
CU.list7<-sort(unique(m.5subset2$Customer.ID))
m12<-matrix(0,nrow=length(CU.list7), ncol=19)
for(i in 1:length(CU.list7)){
for(j in 1:19)
m12[i,j]<-sum(m.5subset2$Units.Bought[m.5subset2$Customer.ID==CU.list7[i] & m.5subset2$Item.Type==j])
}
m13<-as.data.frame(m12)
names(m13)<-1:19
row.names(m13)<-CU.list7
m.5Item<- subset(m13, select = c(5))
cluster5totals<- colSums(m.5Item) ##Total Amount of Units Bought for Items 3 and 5 for Cluster 5
m.6<-m1[new2==6,]
trans.6<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.6)),]
m.6demog<- merge(trans.6, demographics)
m.6subset<-subset(m.6demog, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(m.6subset)
m.6subset2<-m.6subset[good,]
CU.list8<-sort(unique(m.6subset2$Customer.ID))
m14<-matrix(0,nrow=length(CU.list8), ncol=19)
for(i in 1:length(CU.list8)){
for(j in 1:19)
m14[i,j]<-sum(m.6subset2$Units.Bought[m.6subset2$Customer.ID==CU.list8[i] & m.6subset2$Item.Type==j])
}
m15<-as.data.frame(m14)
names(m15)<-1:19
row.names(m15)<-CU.list8
m.6Item<- subset(m15, select = c(5))
cluster6totals<- colSums(m.6Item) ##Total Amount of Units Bought for Items 3 and 5 for Cluster 6
m.7<-m1[new2==7,]
trans.7<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.7)),]
m.7demog<- merge(trans.7, demographics)
m.7subset<-subset(m.7demog, select = c(Customer.ID, Item.Type, Units.Bought))
for(i in 1:length(CU.list8)){
for(j in 1:19)
m14[i,j]<-sum(m.6subset2$Units.Bought[m.6subset2$Customer.ID==CU.list8[i] & m.6subset2$Item.Type==j])
}
m15<-as.data.frame(m14)
names(m15)<-1:19
row.names(m15)<-CU.list8
m.6Item<- subset(m15, select = c(5))
cluster6totals<- colSums(m.6Item) ##Total Amount of Units Bought for Items 3 and 5 for Cluster 6
m.7<-m1[new2==7,]
trans.7<-transaction[transaction$Customer.ID %in% as.numeric(row.names(m.7)),]
m.7demog<- merge(trans.7, demographics)
m.7subset<-subset(m.7demog, select = c(Customer.ID, Item.Type, Units.Bought))
good<-complete.cases(m.7subset)
m.7subset2<-m.7subset[good,]
CU.list9<-sort(unique(m.7subset2$Customer.ID))
m16<-matrix(0,nrow=length(CU.list9), ncol=19)
for(i in 1:length(CU.list9)){
for(j in 1:19)
m16[i,j]<-sum(m.7subset2$Units.Bought[m.7subset2$Customer.ID==CU.list9[i] & m.7subset2$Item.Type==j])
}
m17<-as.data.frame(m16)
names(m17)<-1:19
row.names(m17)<-CU.list9
m.7Item<- subset(m17, select = c(5))
cluster7totals<- colSums(m.7Item) ##Total Amount of Units Bought for Items 3 and 5 for Cluster 7
par(mfrow=c(2,4))
barplot(cluster1totals, main="Cluster 1", xlab = "Item Type", ylab = "Units Bought", col='red')
barplot(cluster2totals, main="Cluster 2", xlab = "Item Type", ylab = "Units Bought", col='red')
barplot(cluster3totals, main="Cluster 3", xlab = "Item Type", ylab = "Units Bought", col='red')
barplot(cluster4totals, main="Cluster 4", xlab = "Item Type", ylab = "Units Bought", col='green')
barplot(cluster5totals, main="Cluster 5", xlab = "Item Type", ylab = "Units Bought", col='red')
barplot(cluster6totals, main="Cluster 6", xlab = "Item Type", ylab = "Units Bought", col='red')
barplot(cluster7totals, main="Cluster 7", xlab = "Item Type", ylab = "Units Bought", col='red')
barplot(cluster8totals, main="Cluster 8", xlab = "Item Type", ylab = "Units Bought", col='red')

## Demographic Classifications Code
m.4demog<- merge(trans.4, demographics)
```

```
m.4subset<-subset(m.4demog, select=c(Customer.ID, Item.Type, Family.Size, Income, Ethnicity, Male.Occupation,
  Female.Occupation))
install.packages('plyr')
library(plyr)
male.occupation4<-aggregate(Male.Occupation~Customer.ID, data=m.4subset, FUN=max)
female.occupation4<-aggregate(Female.Occupation~Customer.ID, data=m.4subset, FUN=max)
ethnicity4<-aggregate(Ethnicity~Customer.ID, data=m.4subset, FUN=max)
family.size4<-aggregate(Family.Size~Customer.ID, data=m.4subset, FUN=max)
income4<-aggregate(Income~Customer.ID, data=m.4subset, FUN=max)
dogs4<-aggregate(Dogs~Customer.ID, data=m.4subset, FUN=max)
cats4<-aggregate(Cats~Customer.ID, data=m.4subset, FUN=max)
tv4<-aggregate(TVs~Customer.ID, data=m.4subset, FUN=max)
male.age4<-aggregate(Male.Age.~Customer.ID, data=m.4subset, FUN=max)
female.age4<-aggregate(Female.Age.~Customer.ID, data=m.4subset, FUN=max)
children4<-aggregate(Children~Customer.ID, data=m.4subset, FUN=max)
male.work.hours4<-aggregate(Male.Work.Hours~Customer.ID, data=m.4subset, FUN=max)
female.work.hours4<-aggregate(Female.Work.Hours~Customer.ID, data=m.4subset, FUN=max)
male.education4<-aggregate(Male.Education~Customer.ID, data=m.4subset, FUN=max)
female.education4<-aggregate(Female.Education~Customer.ID, data=m.4subset, FUN=max)

##Percentage of Each Demographic Category Code
barplot(colSums(table(male.work.hours4))/14, col=c('orange','blue','blue'))
barplot(colSums(table(female.work.hours4))/14, col=c('blue', 'blue', 'orange', 'blue'))
barplot(colSums(table(female.education4))/14, col=c('blue', 'blue', 'orange', 'orange', 'blue', 'blue', 'blue'))
barplot(colSums(table(children4))/14, col=c('blue', 'blue', 'orange'))
barplot(colSums(table(female.age4))/14, col=c('blue', 'blue', 'blue', 'blue', 'orange'))
barplot(colSums(table(male.age4))/14, col=c('blue', 'blue', 'blue', 'orange', 'blue'))
barplot(colSums(table(tv4))/14, col=c('blue', 'orange', 'blue'))
barplot(colSums(table(cats4))/14, col=c('orange', 'blue', 'blue'))
barplot(colSums(table(dogs4))/14, col=c('orange', 'blue', 'blue'))
barplot(colSums(table(income4))/14, col=c('blue', 'blue', 'blue', 'blue', 'orange', 'blue', 'blue'))
barplot(colSums(table(family.size4))/14, col=c('orange', 'blue', 'blue', 'blue', 'blue'))
barplot(colSums(table(ethnicity4))/14, col=c('orange'))
barplot(colSums(table(male.occupation4))/14, col=c('blue', 'blue', 'blue', 'blue', 'orange', 'orange', 'blue'))

## Time Series Graph
daynum<-transaction
daynum1<-subset(daynum, select=c(Item.Type, Units.Bought, Week, Day))
daynum["Day Number"]<-NA
daynum1["Day Number"]<-NA
trial2<-(daynum1$Week-1)*7+daynum1$Day
daynum1["Day.Number"]<-NA
daynum1$Day.Number <- trial2
daynum1<-daynum1[,-5]
daynum<-daynum1[,-3]
daynum1<-daynum1[,-3]
daynum1<-daynum1[,-3]
daynum2<-daynum1[,-1]
daynum3<-aggregate(Units.Bought ~ Day.Number, data=daynum2, FUN=sum)
plot(daynum3, type="o", col="blue", main="Item 5 Units Bought by Day Over 2 Years", xlab="Day #", ylab="Units Bought")

## Difference in Logs Graph
plot(diff(log(daynum3$Units.Bought)),type="l")

## Linear Filtering
library(stats)
plot(daynum3$Units.Bought, type="l")
daynum.1<-filter(daynum3$Units.Bought,filter=rep(1/5,5))
daynum.2<- filter(daynum3$Units.Bought,filter=rep(1/25,25))
daynum.3<-filter(daynum3$Units.Bought, filter=rep(1/81,81))
lines(daynum.1,col="red")
lines(daynum.2,col="green")
lines(daynum.3,col="blue")

## Seasonal Trends
daynum3ts<-ts(daynum3$Units.Bought,start=0,freq=12)
plot(stl(log(daynum3ts),s.window="periodic"))

## Holt-Winters
daynum3ts<-ts(daynum3$Units.Bought,start=1,freq=92)
```

```
daynum3hw<-ts(daynum3$Units.Bought,start=1,freq=92)
HoltWinters(daynum3hw)
plot(daynum3hw)
daynum3hw<-HoltWinters(daynum3ts)
head(daynum3hw)
predict(daynum3hw,n.ahead=92)
plot(daynum3hw,xlim=c(1,12))
lines(predict(daynum3hw,n.ahead=276,col=4))
```

6.2 *Data Dictionary for Retail Store*

(1) Demographic data

Columns	Value	Description
Family Size	0	no response
	1	one person
	2	two people
	3	three people
	4	four people
	5	five people
	6	six or more people
Income	0	no response
	1	less than 10000
	2	10000 to 11999
	3	12000 to 14999
	4	15000 to 19999
	5	20000 to 24999
	6	25000 to 34999
	7	35000 to 44999
	8	45000 to 54999
	9	55000 to 64999
	10	65000 to 74999
	11	75000 and over
Race	0	no response
	1	white
	2	black
	3	hispanic
	4	oriental
	5	other
Dogs	0	none
	1	one
	2	two
	3	three
	4	four
	5	five or more
	9	no response
Cats	0	none
	1	one
	2	two
	3	three
	4	four
	5	five or more
	9	no response

TVs	0	none
	1	one cabled set
	2	two cabled sets
	3	three or more
	9	no response
Male age	0	no response
	1	18 to 29
	2	30 to 34
	3	35 to 44
	4	45 to 54
	5	55 to 64
	6	65 to 99
	7	no male
Female age	0	no response
	1	18 to 29
	2	30 to 34
	3	35 to 44
	4	45 to 54
	5	55 to 64
	6	65 to 99
	7	no female
Children	0	none
	1	children 0 to 5
	2	children 6 to 11
	3	children 12 to 18
	4	groups 1 and 2
	5	groups 1 and 3
	6	groups 2 and 3
	7	groups 1, 2 and 3
	8	children over 18
Male work hours	0	no response
	1	not employed
	2	less than 35 hours
	3	more than 35 hours
	4	retired
	5	no male
Male occupation	0	no response
	1	professional
	2	manager
	3	sales
	4	clerical
	5	craftsman
	6	operative
	7	laborer
	8	cleaning
	9	private household
	10	retired
	11	no male
	13	not employed
Male education	0	no response
	1	some grade school
	2	completed grade sch
	3	some high sch

	4	completed high sch
	5	some college
	6	completed college
	7	post graduate work
	8	technical school
	9	no male
Female work hours	0	no response
	1	not employed
	2	less than 35 hours
	3	more than 35 hours
	4	retired
	5	no female
Female occupation	0	no response
	1	professional
	2	manager
	3	sales
	4	clerical
	5	craftsman
	6	operative
	7	laborer
	8	cleaning
	9	private household
	10	retired
	11	no female
	13	not employed
Female education	0	no response
	1	some grade school
	2	completed grade sch
	3	some high sch
	4	completed high sch
	5	some college
	6	completed college
	7	post graduate work
	8	technical school
	9	no female

(2) Transaction data

Columns	Value	Description
Item Type	1	bacon
	2	bbq
	3	butter
	4	cat food
	5	cereal
	6	cleansers
	7	coffee
	8	cook
	9	crackers
	10	detergents
	11	dogs
	12	eggs

	13	ice cream
	14	nuts
	15	pill
	16	pizza
	17	snack
	18	soap
	19	soft
	20	softdrinks
	21	sugar
	22	tissue
	23	towel
	24	yogurt
UPC: System		UPC number system
Generation		Generation of reused UPC numbers
Vendor		Manufacturer identification
Item		UPC item identification
PANELIST		Customer ID
WEEK	1~104	a total of 104 weeks
DAY	1 ~ 7	day of week
STORE ID		Store number (expanded to 5 digits)
UNITS		Units purchased
COUPON ORIGIN	0	no coupon used
	1-18	NA
	19	ActNow
	20	Vendor (origin unknown)
	21	direct mail vendor
	22	magazine vendor
	23	sunday supplement vendor
	24	newspaper ad vendor
	25	in-pack (use next purchase)
	26	cross-ruff (in pack of other product)
	27	on-pack (use next purchase)
	28	free standing insert
	29	instant redeemable
	30	misredemption
	31-33	TV Guide test
	34-38	test
	39	vending machine and store coupon
	40	Vendor (origin unknown) and store cou
	41	direct mail vendor and store coupon

42	magazine vendor and store coupon
43	sunday supplement vendor and store co
44	newspaper ad vendor and store coupon
45	in-pack (use next purchase) and store
46	cross-ruff (in pack of other product)
47	on-pack (use next purchase) and store
48	free standing insert and store coupon
49	instant redeemable and store coupon
50	store (origin unknown)
51	newspaper ad store
52	store's flyer
53	free product (store coupon only)

COUPON VALUE Value of the deal in cents

7. Bibliography

Chen, Cheng-Bang. "R Lessons During Spring 2014 Semester." *Rm 233 Leonhard Building, The Pennsylvania State University* Spring 2014. Lecture.

Goodwin, Paul. "The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong." *Foresight* Fall 2010. PDF. <http://www.forecasters.org/pdfs/foresight/free/Issue19_goodwin.pdf>

Lindsay Hixson, Bradford B. Hepler, and Myoung Ouk Kim. "The White Population 2010." *2010 Census Briefs* September 2011. PDF <http://www.census.gov/prod/cen2010/briefs/c2010br-05.pdf>

Noss, Amanda. "Household Income 2012." *American Community Survey Briefs* September 2013. PDF <https://www.census.gov/prod/2013pubs/acsbr12-02.pdf>

Peng, Roger. "Computing For Data Analyis." *Johns Hopkins Course* 6 January 2014. Video Lectures.<https://www.coursera.org/course/compdata>

Walter Zucchini, Oleg Nenadi´c. "Time Series Analysis with R - Part I." University of Göttingen. PDF <http://www.statoek.wiso.uni-goettingen.de/veranstaltungen/zeitreihen/sommer03/ts_r_intro.pdf>