Homework 2

Group 3

Initial import and merging

```
PROC IMPORT OUT= WORK.prod
        DATAFILE= "H:\prod_tooth.xls"
        DBMS=EXCEL REPLACE;
    RANGE="Sheet1$";
    GETNAMES=YES;
    MIXED=NO;
    SCANTEXT=YES;
    USEDATE=YES;
    SCANTIME=YES;
RUN;

data grocery;
infile "H:\toothbr_groc" firstobs = 2;
input IRI_KEY WEEK SY GE VEND ITEM UNITS DOLLARS F$ D PR;
length UPC $ 17;
length SY2 $ 2;
length GE2 $ 2;
length VEND2 $ 5;
length ITEM2 $ 5;
SY2 = put(SY, z2.);
GE2 = put(GE,z2.);
VEND2 = put(VEND, z5.);
ITEM2 = put(ITEM, z5.);
UPC = catx('-', SY2, GE2, VEND2, ITEM2);
drop SY GE VEND ITEM;
run;

PROC PRINT DATA = grocery(obs=15);RUN;

data groc(rename=(SY2=SY GE2=GE VEND2=VEND ITEM2=ITEM));
set grocery;
RUN;

proc surveyselect data = groc
out = groc_out method = srs
sampsize = 100000;
run;

proc print data = groc_out (obs=5);
```

```
run;

proc print data = prod (obs=10);
run;

proc sort data = groc_out ; by UPC;
RUN;
proc sort data = prod ; by UPC;
RUN;


data groc_prod_details ;
merge groc_out (in = a)
prod (in=b);
by UPC;
if a and b;
run;

proc print data = groc_prod_details (obs=25) ;
run;
```

Questions

1. **What are the top 6 brands in the category in terms of dollar sales? What are the market shares of the 6 brands (assuming there are only 6 brands in the market)**

```
proc tabulate data = groc_prod_details out=top_six_brands_by_dollars  ;
class L5;
var dollars;
tables L5, dollars*(mean sum);
run;

proc sort data=top_six_brands_by_dollars;
by descending DOLLARS_Sum;
run;

data T6brnds;
set top_six_brands_by_dollars(obs=6);
run;

proc print data=T6brnds;
var L5 DOLLARS_Mean DOLLARS_Sum ;
run;
```

```
proc sql;
create table T6brndsTotal as
select L5,DOLLARS_Mean,DOLLARS_Sum, sum(DOLLARS_Sum) as TOTAL '' from
T6brnds;
quit;

proc print data = T6brndsTotal;
run;

data mktval;
set top_six_brands_by_dollars ;
MarketVal = DOLLARS_Sum*100/TOTAL;
run;

proc print data=mktval ;
run;
```

The top 6 brands are:

Obs	L5	DOLLARS_Mean	DOLLARS_Sum
1	PRIVATE LABEL	6.44654	62744.13
2	ORAL B ADVANTAGE	8.27610	61499.67
3	ORAL B CROSSACTION	9.77609	51207.18
4	ORAL B INDICATOR	6.97210	47082.60
5	COLGATE WAVE	7.97262	38627.35
6	COLGATE NAVIGATOR	8.96264	38566.24

Assuming there are only 6 brands in the market, market share of the 6 brands are:

Obs	L5	DOLLARS_Mean	DOLLARS_Sum	MarketVal
1	PRIVATE LABEL	6.44654	62744.13	20.9337
2	ORAL B ADVANTAGE	8.27610	61499.67	20.5186
3	ORAL B CROSSACTION	9.77609	51207.18	17.0846
4	ORAL B INDICATOR	6.97210	47082.60	15.7085
5	COLGATE WAVE	7.97262	38627.35	12.8875
6	COLGATE NAVIGATOR	8.96264	38566.24	12.8671

2. **Which companies are the major players in the category? Which company owns which brands?**

```
proc tabulate data = groc_prod_details out=major_companies  ;
class L4;
var dollars;
tables L4, dollars*(mean sum);
run;

proc sort data=major_companies;
by descending DOLLARS_Sum;
run;

proc print data=major_companies (obs=6) ;
run;


data majorCompaniesBrands;
set prod ;
if L4 = "PROCTER & GAMBLE" or L4 ="COLGATE ORAL PHARMACEUTICALS" or L4 = "JOHNSON & JOHNSON"  or  L4  =  "PRIVATE  LABEL"  or  L4="CHURCH  &  DWIGHT  CO  INC"  or  L4 ="GLAXOSMITHKLINE" ;
run;

 proc sql ;
 create table Company_Brands as
 select distinct L4,L5 from majorCompaniesBrands;
 quit;

 proc print data = Company_Brands;
 run;
```

Major players (Companies) in the category are as shown below.

Obs	L4	DOLLARS_Mean	DOLLARS_Sum
1	PROCTER & GAMBLE	7.61494	269591.76
2	COLGATE ORAL PHARMACEUTICALS	7.42638	161100.37
3	JOHNSON & JOHNSON	6.95940	93082.03
4	PRIVATE LABEL	6.44654	62744.13
5	CHURCH & DWIGHT CO INC	6.46638	41624.12
6	GLAXOSMITHKLINE	6.56979	37119.29

Below is the list that contain the brands and companies which own those brands

bs	L4	L5
1	CHURCH & DWIGHT CO INC	AIM
2	CHURCH & DWIGHT CO INC	CLOSE UP
3	CHURCH & DWIGHT CO INC	MENTADENT
4	CHURCH & DWIGHT CO INC	MENTADENT INTER SWEEP
5	CHURCH & DWIGHT CO INC	MENTADENT ORAL CARE
6	CHURCH & DWIGHT CO INC	MENTADENT PROCARE
7	CHURCH & DWIGHT CO INC	MENTADENT SMILE REPLENISHING
8	CHURCH & DWIGHT CO INC	MENTADENT SURROUND
9	CHURCH & DWIGHT CO INC	MENTADENT WHITE AND CLEAN
10	CHURCH & DWIGHT CO INC	PEPSODENT
11	CHURCH & DWIGHT CO INC	PEPSODENT PROFESSIONAL
12	COLGATE ORAL PHARMACEUTICALS	COLGATE
13	COLGATE ORAL PHARMACEUTICALS	COLGATE 360
14	COLGATE ORAL PHARMACEUTICALS	COLGATE ACTIVE ANGLE
15	COLGATE ORAL PHARMACEUTICALS	COLGATE BARBIE
16	COLGATE ORAL PHARMACEUTICALS	COLGATE BLUES CLUES
17	COLGATE ORAL PHARMACEUTICALS	COLGATE CLASSIC
18	COLGATE ORAL PHARMACEUTICALS	COLGATE COLOR CHANGE
19	COLGATE ORAL PHARMACEUTICALS	COLGATE DISNEY ATLANTIS
20	COLGATE ORAL PHARMACEUTICALS	COLGATE EXTRA CLEAN
21	COLGATE ORAL PHARMACEUTICALS	COLGATE GRIP EMS
22	COLGATE ORAL PHARMACEUTICALS	COLGATE HE MAN
23	COLGATE ORAL PHARMACEUTICALS	COLGATE LEGO
24	COLGATE ORAL PHARMACEUTICALS	COLGATE MASSAGER
25	COLGATE ORAL PHARMACEUTICALS	COLGATE NAVIGATOR
26	COLGATE ORAL PHARMACEUTICALS	COLGATE NICK JR DORA THE EXPL
27	COLGATE ORAL PHARMACEUTICALS	COLGATE NICK THE FRLY ODD PAR
28	COLGATE ORAL PHARMACEUTICALS	COLGATE PLUS
29	COLGATE ORAL PHARMACEUTICALS	COLGATE PLUS KOOL LOOKS
30	COLGATE ORAL PHARMACEUTICALS	COLGATE PLUS RIPPLED
31	COLGATE ORAL PHARMACEUTICALS	COLGATE PLUS ULTRA FIT

bs	L4	L5
32	COLGATE ORAL PHARMACEUTICALS	COLGATE POWER PUFF
33	COLGATE ORAL PHARMACEUTICALS	COLGATE SENSITIVE
34	COLGATE ORAL PHARMACEUTICALS	COLGATE SHREK
35	COLGATE ORAL PHARMACEUTICALS	COLGATE SPONGEBOB SQUAREPANTS
36	COLGATE ORAL PHARMACEUTICALS	COLGATE SUPER
37	COLGATE ORAL PHARMACEUTICALS	COLGATE SUPERMAN
38	COLGATE ORAL PHARMACEUTICALS	COLGATE TOTAL
39	COLGATE ORAL PHARMACEUTICALS	COLGATE TOTAL DESIGNS
40	COLGATE ORAL PHARMACEUTICALS	COLGATE TOTAL PROFESSIONAL
41	COLGATE ORAL PHARMACEUTICALS	COLGATE WAVE
42	COLGATE ORAL PHARMACEUTICALS	COLGATE WHITENING
43	COLGATE ORAL PHARMACEUTICALS	COLGATE ZIGZAG
44	COLGATE ORAL PHARMACEUTICALS	MY FIRST COLGATE
45	GLAXOSMITHKLINE	AQUAFRESH
46	GLAXOSMITHKLINE	AQUAFRESH DIRECT
47	GLAXOSMITHKLINE	AQUAFRESH DR SEUSS
48	GLAXOSMITHKLINE	AQUAFRESH EXTREME CLEAN
49	GLAXOSMITHKLINE	AQUAFRESH FLEX
50	GLAXOSMITHKLINE	AQUAFRESH FLEX DIRECT
51	GLAXOSMITHKLINE	AQUAFRESH FLEX FLEXOSAURUS
52	GLAXOSMITHKLINE	AQUAFRESH FLEX O FRIENDS FLEX
53	GLAXOSMITHKLINE	AQUAFRESH FLEX TIP
54	GLAXOSMITHKLINE	AQUAFRESH FLEX TIP MAX ACTIVE
55	GLAXOSMITHKLINE	AQUAFRESH FLEX TRIPLE ACTION
56	GLAXOSMITHKLINE	AQUAFRESH FLX O FRNDS FLEXOSA
57	GLAXOSMITHKLINE	AQUAFRESH MAX ACTIVE
58	GLAXOSMITHKLINE	AQUAFRESH XTENSIVE
59	GLAXOSMITHKLINE	SENSODYNE
60	GLAXOSMITHKLINE	SENSODYNE SEARCH
61	GLAXOSMITHKLINE	SOFTEX
62	JOHNSON & JOHNSON	JOHNSON & JOHNSON REACH
63	JOHNSON & JOHNSON	PLAX
64	JOHNSON & JOHNSON	PREVENT BY REACH
65	JOHNSON & JOHNSON	REACH
66	JOHNSON & JOHNSON	REACH ADVANCED

bs	L4	L5
67	JOHNSON & JOHNSON	REACH ADVANCED CLEAN ANGLE
68	JOHNSON & JOHNSON	REACH ADVANCED DESIGN
69	JOHNSON & JOHNSON	REACH ADVANCED GUM ACTION
70	JOHNSON & JOHNSON	REACH ANTIBACTERIAL
71	JOHNSON & JOHNSON	REACH BETWEEN
72	JOHNSON & JOHNSON	REACH CLEAN AND WHITEN
73	JOHNSON & JOHNSON	REACH CLEAN SWEEP
74	JOHNSON & JOHNSON	REACH CURIOUS GEORGE
75	JOHNSON & JOHNSON	REACH HARRY POTTER
76	JOHNSON & JOHNSON	REACH IN BETWEEN
77	JOHNSON & JOHNSON	REACH INTERDENTAL
78	JOHNSON & JOHNSON	REACH JIMMY NEUTRON
79	JOHNSON & JOHNSON	REACH JUSTICE LEAGUE
80	JOHNSON & JOHNSON	REACH MAX
81	JOHNSON & JOHNSON	REACH MAX BRIGHTENER
82	JOHNSON & JOHNSON	REACH MAX FRESH & CLEAN
83	JOHNSON & JOHNSON	REACH MAX TOOTH & GUM
84	JOHNSON & JOHNSON	REACH MAX TOOTH & GUM WONDERT
85	JOHNSON & JOHNSON	REACH PERFORMANCE
86	JOHNSON & JOHNSON	REACH PLAQUE BLASTER
87	JOHNSON & JOHNSON	REACH PLAQUE SWEEPER
88	JOHNSON & JOHNSON	REACH PLAQUE SWEEPER BETWEEN
89	JOHNSON & JOHNSON	REACH SPONGEBOB SQUAREPANTS
90	JOHNSON & JOHNSON	REACH SQUEEZE
91	JOHNSON & JOHNSON	REACH STRAWBERRY SHORTCAKE
92	JOHNSON & JOHNSON	REACH TOOTH & GUM CARE
93	JOHNSON & JOHNSON	REACH ULTRA CLEAN
94	JOHNSON & JOHNSON	REACH WONDER GRIP
95	PRIVATE LABEL	PRIVATE LABEL
96	PROCTER & GAMBLE	CREST
97	PROCTER & GAMBLE	CREST ADVANCED CARE
98	PROCTER & GAMBLE	CREST CLASSIC CLEAN
99	PROCTER & GAMBLE	CREST CLEAN EXPRESSIONS
100	PROCTER & GAMBLE	CREST COMPLETE
101	PROCTER & GAMBLE	CREST DEEP CLEAN ACTIVE CLEAN

bs	L4	L5
102	PROCTER & GAMBLE	CREST DEEP CLEAN GUM CARE
103	PROCTER & GAMBLE	CREST DEEP CLEAN SENSITIVITY
104	PROCTER & GAMBLE	CREST DEEP SWEEP
105	PROCTER & GAMBLE	CREST DUAL ACTION
106	PROCTER & GAMBLE	CREST EXTENDER
107	PROCTER & GAMBLE	CREST GUM CARE
108	PROCTER & GAMBLE	CREST MASSAGE PLUS
109	PROCTER & GAMBLE	CREST MULTICARE FLEX
110	PROCTER & GAMBLE	CREST MULTICLEAN
111	PROCTER & GAMBLE	CREST PREMIUM
112	PROCTER & GAMBLE	CREST SESAME STREET
113	PROCTER & GAMBLE	CREST TRIPLE EFFECT
114	PROCTER & GAMBLE	CREST WHITENING
115	PROCTER & GAMBLE	ORAL B
116	PROCTER & GAMBLE	ORAL B ADVANTAGE
117	PROCTER & GAMBLE	ORAL B ADVANTAGE ARTICA
118	PROCTER & GAMBLE	ORAL B ADVANTAGE PLUS
119	PROCTER & GAMBLE	ORAL B BLUES CLUES
120	PROCTER & GAMBLE	ORAL B CLASSIC
121	PROCTER & GAMBLE	ORAL B CROSSACTION
122	PROCTER & GAMBLE	ORAL B CROSSACTION VITALIZER
123	PROCTER & GAMBLE	ORAL B GRIPPER
124	PROCTER & GAMBLE	ORAL B INDICATOR
125	PROCTER & GAMBLE	ORAL B NICKELODEON
126	PROCTER & GAMBLE	ORAL B PULSAR
127	PROCTER & GAMBLE	ORAL B RADICAL CONTROL
128	PROCTER & GAMBLE	ORAL B RUGRATS
129	PROCTER & GAMBLE	ORAL B SENSITIVE ADVANTAGE
130	PROCTER & GAMBLE	ORAL B SESAME STREET
131	PROCTER & GAMBLE	ORAL B STAGES

3. **Create a 7th brand called “Other” that has all other brands that are not in the top 6.**

```
data Top6Brands;
set groc_prod_details;
if L5="PRIVATE LABEL" then Brand = "PRIVATE LABEL" ;
```

```
length Brand $35;
else if L5="ORAL B ADVANTAGE"  then Brand = "ORAL B ADVANTAGE" ;
else if L5="ORAL B CROSSACTION" then Brand = "ORAL B CROSSACTION";
else if L5="ORAL B INDICATOR" then Brand = "ORAL B INDICATOR";
else if L5="COLGATE NAVIGATOR" then Brand = "COLGATE NAVIGATOR";
else if L5="COLGATE WAVE" then Brand = "COLGATE WAVE";
else Brand = "OTHER";
run;

proc print data = Top6Brands(obs=10);
run;

proc sql ;
create table brnds as
select distinct Brand from Top6Brands ;
quit;

proc print data = brnds;
run;
```

Created a brand called 'Others' which that contain all other brands that are not in top 6.

Obs	Brand
1	COLGATE NAVIGATOR
2	COLGATE WAVE
3	ORAL B ADVANTAGE
4	ORAL B CROSSACTION
5	ORAL B INDICATOR
6	OTHER
7	PRIVATE LABEL

4. **Find average prices, display, features of each of the 7 brands.**

```
data Top6Brands_1;
set Top6Brands;
if F = "NONE" then FEAT = 0 ;
else FEAT = 1;
run;

proc print data = Top6Brands_1 (obs=10);
run;
```

```
proc means data = Top6Brands_1;
class Brand ;
var DOLLARS  FEAT  D ;
run;

proc tabulate data=Top6Brands_1;
class Brand;
var DOLLARS FEAT D;
table Brand, (DOLLARS FEAT D )* MEAN ;
run;
```

Below is the list that contains average prices (dollar/units/oz), display (1/0), features(1/0) of each of the 7 brands

	DOLLARS	FEAT	D
	Mean	Mean	Mean
Brand			
COLGATE NAVIGATOR	8.96	0.15	0.06
COLGATE WAVE	7.97	0.12	0.05
ORAL B ADVANTAGE	8.28	0.11	0.06
ORAL B CROSSACTION	9.78	0.09	0.10
ORAL B INDICATOR	6.97	0.07	0.04
OTHER	6.72	0.05	0.04
PRIVATE LABEL	6.45	0.05	0.02

5. **What are the top 5 regions in terms of dollar sales?**

Merging Delivery Stores data with the previous data.

```
/*START */

data dStores;
infile "H:/Delivery_Stores" firstobs= 2 missover ;
input @1 IRI_KEY best7. @9 OU $2. @12 EST_ACV best8. @21 Market_Name $20. @46 Open
best4. @51 Clsd best4. @56 MskdName $8. ;
run;

proc print data = dStores (obs=5);
run;

proc sql ;
create table GrocProdMkt as
select * from dStores , Top6Brands_1 where
```

```
dStores.IRI_KEY = Top6Brands_1.IRI_KEY;
quit;

proc print data = GrocProdMkt (obs=10);
run;
        /*end */


proc tabulate data=GrocProdMkt out =  Top5Regions;
class Market_Name;
var DOLLARS ;
table Market_Name , DOLLARS * (SUM MEAN);
run;

/*START OUR BRAND*/
DATA ourbrand_by_region;set GrocProdMkt; if L5 = "ORAL B CROSSACTION";
RUN;

proc tabulate data=ourbrand_by_region out =  Top5RegionsBYBrand;
class Market_Name;
var DOLLARS ;
table Market_Name, DOLLARS * (SUM MEAN);
run;

proc sort data = Top5RegionsBYBrand;
by descending DOLLARS_Sum;
run;


PROC PRINT DATA = Top5RegionsBYBrand(obs =5);
var Market_Name DOLLARS_Sum DOLLARS_Mean ;
run;
```

Top 5 regions in terms of dollar sales is as shown below

Obs	Market_Name	DOLLARS_Sum	DOLLARS_Mean
1	NEW YORK	71842.94	9.24620
2	LOS ANGELES	70791.61	8.29427
3	SAN FRANCISCO	31442.03	8.70006
4	WASHINGTON, DC	30287.39	8.82757
5	PORTLAND,OR	26549.00	8.56972

6. **What are the top 10 store chains that sell a lot of your category in terms of dollar sales?**

```
DATA Our_brand;
set GrocProdMkt;
if L5 = "ORAL B CROSSACTION";
RUN;

proc tabulate data=Our_brand out = Top10StoreChains;
class MskdName;
var DOLLARS ;
table MskdName , DOLLARS * (SUM MEAN);
run;


proc sort data = Top10StoreChains;
by descending DOLLARS_Sum;
run;

proc print data = Top10StoreChains(obs=10);
run;
```

Obs	MskdName	_TYPE_	_PAGE_	_TABLE_	DOLLARS_Sum	DOLLARS_Mean
1	Chain94	1	1	1	3157.83	12.8367
2	Chain124	1	1	1	2894.62	12.1623
3	Chain75	1	1	1	2757.49	20.5783
4	Chain114	1	1	1	2258.09	11.1787
5	Chain79	1	1	1	1868.56	12.7984
6	Chain134	1	1	1	1863.67	11.3638
7	Chain89	1	1	1	1768.67	9.5604
8	Chain117	1	1	1	1702.99	12.2517
9	Chain10	1	1	1	1561.15	8.5777
10	Chain98	1	1	1	1450.71	12.3992

The above output shows the top 10 store chains in terms of Dollar Sales.

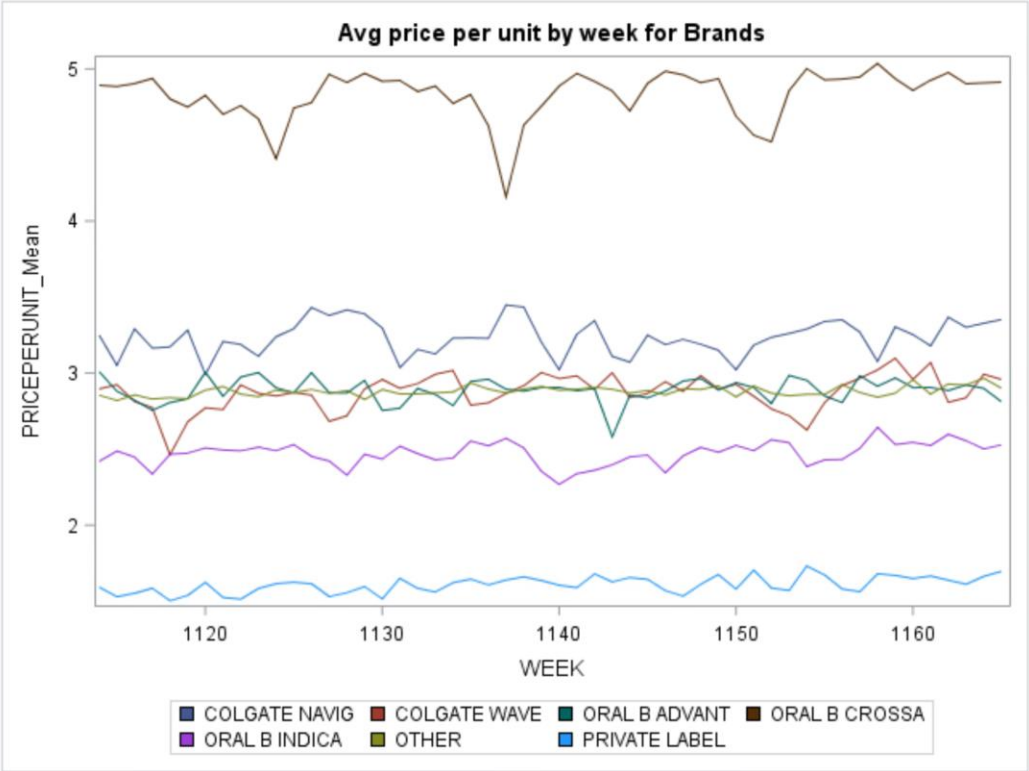
7. **What is the average price per unit of 7 brands by week? Plot the average price by week (I wish to see a line plot of price by week). Comment on your findings.**

```
data PricePerUnit ;
set  GrocProdMkt;
PRICEPERUNIT = DOLLARS/UNITS;
run;

proc print data=PricePerUnit (obs = 10) ;
run;

proc tabulate data = PricePerUnit out = AvgPriceByWeek ;
class Brand week;
var priceperunit;
table WEEK , priceperunit*MEAN*BRAND ;
run;


proc print data=AvgPriceByWeek (obs=10);
run;

/* Set the graphics environment */
goptions reset=all border cback=white
       htitle=12pt htext=10pt;

proc sgplot data=AvgPriceByWeek;
  title 'Avg price per unit by week for Brands';
  series x=week y=PRICEPERUNIT_Mean / group=Brand name='grouping';
  keylegend 'grouping' / type=linecolor;
run;
```

average price per unit of 7 brands by week

Obs	Brand	WEEK	_TYPE_	_PAGE_	_TABLE_	PRICEPERUNIT_Mean
1	COLGATE NAVIG	1114	11	1	1	3.24933
2	COLGATE WAVE	1114	11	1	1	2.89591
3	ORAL B ADVANT	1114	11	1	1	3.00920
4	ORAL B CROSSA	1114	11	1	1	4.89279
5	ORAL B INDICA	1114	11	1	1	2.41994
6	OTHER	1114	11	1	1	2.85450
7	PRIVATE LABEL	1114	11	1	1	1.59264
8	COLGATE NAVIG	1115	11	1	1	3.05124
9	COLGATE WAVE	1115	11	1	1	2.92379
10	ORAL B ADVANT	1115	11	1	1	2.88127

Plot of average price by week



We observe that the price per unit is highest for ORAL B CROSSACTION and least for PRIVATE LABEL. COLGATE WAVE and ORAL B ADVANTAGE have their price per units close to each other.

Among all the brands, ORAL B CROSSACTION has the most variation in price per unit.

a. Majority of tooth brush brands (i.e. except Oral B Crossaction and Private label) have average price range greater than $2 and less than $4 per unit. Also, among them three brands have very competitive average price per unit (near to $3)
b. Tooth brush brands under private label have the least average price per unit. While Oral B Crossaction has highest average price per unit.
c. From the line graph we can notice that no brand is stable in the market. All brands have a high and low every week and Oral B Crossaction has sharp low points every week.

8. **Assume you are manager of a brand (out of the top 6). Write a short paragraph stating what you learned from this descriptive analysis (steps 1-7).**

    a. Top 6 brands capture 42.32% of the total market
    b. The tooth brush brand we selected is Oral B Crossaction with market valuation over 7%
    c. For oral b advantage which has the highest market value had features on toothbrush for over 10% of the time and displayed over 5% of the time.
        i. Brand with highest feature % is Colgate navigator
        ii. brand with highest display % is Oral b crossaction
    d. We have a list of average prices in (dollar/unit), display, features of each of the 7 brands. We can use it to compare the prices of Oral B Crossaction competitors in the market
    e. Major market value is divided among far east and far west regions of the United States e.g. LA, SF, Portland is west coast and New York and Washington in the east coast. The distribution of highest dollar sales for our brand i.e. ORAL B CROSSACTION is same

9. **Do large stores (top 3 stores) have higher average price per unit than small stores (stores ranked 8-10) for brand 1 (the top brand in Q1). Test and report your results and comments.**

```
data LS;
set PricePerUnit;
if BRAND EQ "PRIVATE LABEL"  and (MskdName = "Chain94" or MskdName = "Chain114" or MskdName = "Chain124"  or MskdName = "Chain10" or MskdName = "Chain79" or MskdName = "Chain107")  ;
run;

data Large_Small_Store;
set LS;
if( MskdName = "Chain94" or MskdName = "Chain114" or MskdName = "Chain124" ) then STORE_SIZE = "Large" ;
if (MskdName = "Chain10" or MskdName = "Chain79" or MskdName = "Chain107") then STORE_SIZE = "Small" ;
RUN;
```

```
PROC PRINT DATA = Large_Small_Store(OBS=10);
RUN;

PROC TTEST data = Large_Small_Store SIDES = U TEST = DIFF;
VAR PRICEPERUNIT ;
CLASS STORE_SIZE;
RUN;
```

The TTEST Procedure

Variable: PRICEPERUNIT

STORE_SIZE	N	Mean	Std Dev	Std Err	Minimum	Maximum
Large	1041	1.8311	0.5740	0.0178	0.6300	3.7500
Small	939	1.7370	0.4576	0.0149	0.7900	2.9900
Diff (1-2)		0.0941	0.5221	0.0235		

STORE_SIZE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Large		1.8311	1.7962	1.8660	0.5740	0.5504	0.5998
Small		1.7370	1.7077	1.7663	0.4576	0.4378	0.4793
Diff (1-2)	Pooled	0.0941	0.0480	0.1402	0.5221	0.5063	0.5389
Diff (1-2)	Satterthwaite	0.0941	0.0486	0.1397			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1978	4.01	<.0001
Satterthwaite	Unequal	1949	4.05	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1040	938	1.57	<.0001

Null Hypothesis: average price per unit is for large stores <= small stores.

Alternate Hypothesis: large stores (top 3 stores) have higher average price per unit than small stores.

On performing the ttest procedure we observe that the Pr > F value in the equality of variance table it is <.0001 Since, 0.0001<0.05 hence we reject the null that variance is equal.

Now we check the value of Pr>|t| and it is <0.0001 for unequal variances. Since, 0.0001 < 0.05 we reject the null in favor of alternative.

Hence, the average price per unit for large stores is greater than small stores

10. **Develop three additional hypotheses linking useful variables to dollar sales, test them and report your findings.**

Hypothesis 1: Feature and Dollar sales.

```
proc means data = PricePerUnit;
class FEAT;
var DOLLARS;
run;

PROC TTEST DATA = PricePerUnit SIDES = U TEST = DIFF;
VAR DOLLARS ;
CLASS FEAT;
RUN;
```

Avg price per unit by week for Brands

The TTEST Procedure

Variable: DOLLARS

FEAT	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	93438	6.8107	7.3497	0.0240	0.3200	344.3
1	6706	11.5814	15.7005	0.1917	0.5500	494.3
Diff (1-2)		-4.7707	8.1796	0.1034		

FEAT	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		6.8107	6.7636	6.8578	7.3497	7.3165	7.3832
1		11.5814	11.2056	11.9572	15.7005	15.4392	15.9708
Diff (1-2)	Pooled	-4.7707	-4.9408	Infty	8.1796	8.1439	8.2156
Diff (1-2)	Satterthwaite	-4.7707	-5.0886	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	100142	-46.13	1.0000
Satterthwaite	Unequal	6917.4	-24.69	1.0000

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	6705	93437	4.56	<.0001

Null Hypothesis: Average Dollar Sales in absence of Feature <= Average Dollar Sales in case of Feature

Alternative Hypothesis: Average Dollar Sales in absence of Feature > Average Dollar Sales in case of Feature

On performing the ttest procedure we observe that the Pr > F value in the equality of variance table it is 0.0001. Since, 0.0001 < 0.05 hence reject the null in favor of alternative.

Now we check the value of Pr>|t| and it is 1.000 for equal variances. Since, 1.000 > 0.05 we fail to reject the null. There is no evidence to support the alternative.

Hence, we conclude based on the hypothesis that, Average Dollar Sales in presence of Feature is greater than average Dollar Sales in absence of Feature

Hypothesis 2: Display and Dollar sales.

```
proc sql ;
update PricePerUnit
set D = 1 where D = 2 ;
quit;

proc means data = PricePerUnit;
class D;
var DOLLARS;
run;

proc ttest data = PricePerUnit test =  diff side = u;
class D;
var DOLLARS;
run;
```

Null Hypothesis: Average Dollar Sales in presence of Display >= Average Dollar Sales in absence of Display

Alternative Hypothesis: Average Dollar Sales in presence of Display < Average Dollar Sales in absence of Display

On performing the ttest procedure we observe that the Pr > F value in the equality of variance table it is 0.0001. Since, 0.0001 < 0.05 hence reject the null in favor of alternative.

Now we check the value of Pr>|t| and it is 1.000 for unequal variances. Since, 1.000 > 0.05 we fail to reject the null. There is no evidence to support the alternative.

Hence, we conclude based on the hypothesis that, Average Dollar Sales in presence of Display is greater than average Dollar Sales in absence of Display

Avg price per unit by week for Brands

The TTEST Procedure

Variable: DOLLARS

D	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	96845	6.9263	7.5630	0.0243	0.3200	320.6
1	3299	13.1164	18.9227	0.3295	0.6600	494.3
Diff (1-2)		-6.1901	8.1919	0.1450		

D	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		6.9263	6.8786	6.9739	7.5630	7.5294	7.5968
1		13.1164	12.4704	13.7623	18.9227	18.4769	19.3907
Diff (1-2)	Pooled	-6.1901	-6.4287	Infty	8.1919	8.1562	8.2279
Diff (1-2)	Satterthwaite	-6.1901	-6.7336	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	100142	-42.68	1.0000
Satterthwaite	Unequal	3334	-18.74	1.0000

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3298	96844	6.26	<.0001

Hypothesis 3: TYPE_OF_BRUSH and Dollar sales.

```
data LSS ;
set PricePerUnit;
if L4 EQ "COLGATE ORAL PHARMACEUTICALS" ;
RUN;

PROC MEANS  DATA = LSS;
CLASS TYPE_OF_BRUSH;
VAR DOLLARS;
RUN;

DATA LSS2 ; SET LSS ;
IF TYPE_OF_BRUSH = "REGULAR" OR TYPE_OF_BRUSH = "SENSITIVE" ;


PROC MEANS DATA= LSS2;
CLASS TYPE_OF_BRUSH;
```

```
VAR DOLLARS;
RUN;

PROC TTEST DATA = LSS2 TEST =  DIFF SIDE = U;
CLASS TYPE_OF_BRUSH;
VAR DOLLARS;
RUN;
```

Null Hypothesis:
For "COLGATE ORAL PHARMACEUTICALS" Average Dollar Sales of a regular type of brush <= average sales of sensitive type Brush.

Alternative Hypothesis:
For "COLGATE ORAL PHARMACEUTICALS" Average Dollar Sales of a regular type of brush > average sales of sensitive type Brush.

On performing the ttest procedure we observe that the Pr > F value in the equality of variance table it is 0.0001. Since, 0.0001 < 0.05 hence reject the null in favor of alternative.

Now we check the value of Pr>|t| and it is 1.000 for unequal variances. Since, .901 > 0.05 we fail to reject the null. There is no evidence to support the alternative.

Hence, we conclude based on the hypothesis that, Average Dollar Sales for a regular type of brush is greater than the average sales of a sensitive type of Brush. Hence, It is beneficial for Colgate Oral Pharmaceuticals focus on increasing the sales of sensitive brushes.

The MEANS Procedure

Analysis Variable : DOLLARS						
TYPE OF BRUSH	N Obs	N	Mean	Std Dev	Minimum	Maximum
REGULAR	399	399	6.6615789	8.4391411	0.7700000	101.2700000
SENSITIVE	132	132	7.5184091	5.8900718	1.9900000	43.7800000

Variable: DOLLARS

TYPE_OF_BRUSH	N	Mean	Std Dev	Std Err	Minimum	Maximum
REGULAR	399	6.6616	8.4391	0.4225	0.7700	101.3
SENSITIVE	132	7.5184	5.8901	0.5127	1.9900	43.7800
Diff (1-2)		-0.8568	7.8850	0.7917		

TYPE_OF_BRUSH	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
REGULAR		6.6616	5.8310	7.4922	8.4391	7.8914	9.0692
SENSITIVE		7.5184	6.5042	8.5326	5.8901	5.2550	6.7011
Diff (1-2)	Pooled	-0.8568	-2.1614	Infty	7.8850	7.4372	8.3907
Diff (1-2)	Satterthwaite	-0.8568	-1.9527	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	529	-1.08	0.8602
Satterthwaite	Unequal	320.67	-1.29	0.9010

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	398	131	2.05	<.0001

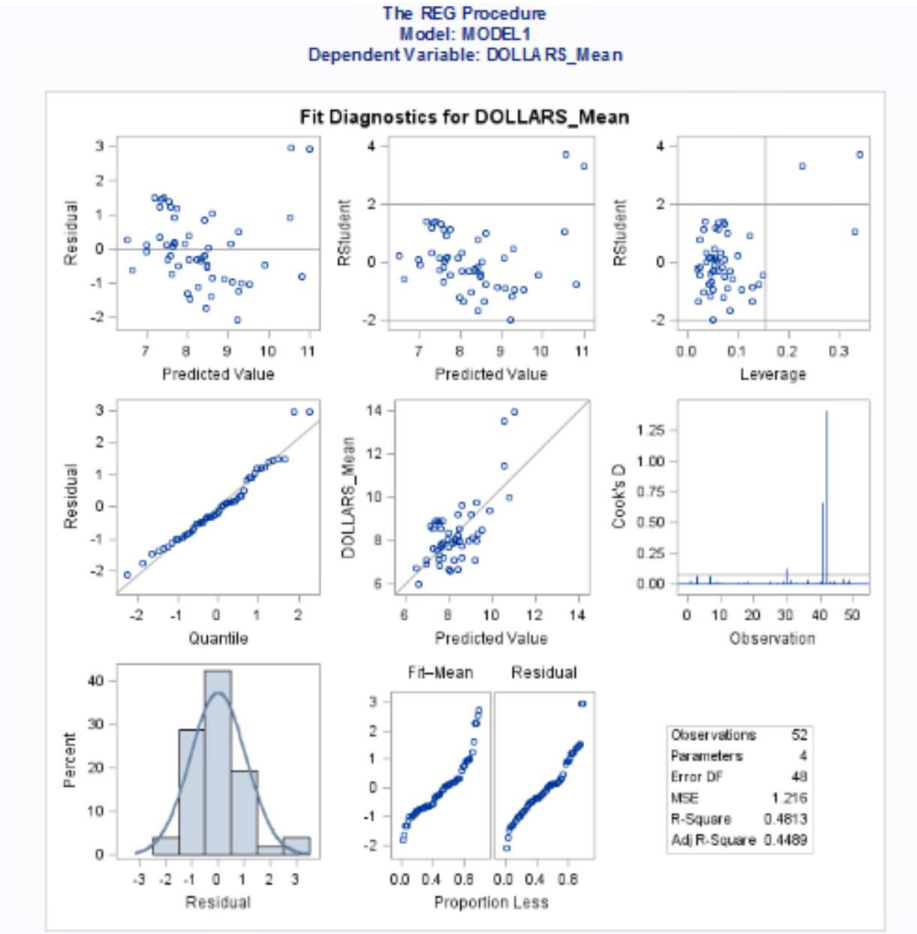
11. For the top brand: run a regression model with weekly dollar sales as dependent variable. Use average weekly price per unit, average display, average feature, and other useful variables in your regression and answer the following questions:

The top brand we considered is ORAL B ADVANTAGE.

```
DATA TopBrand;SET GrocProdMkt;
IF L5 ="ORAL B ADVANTAGE";
PRICEPERUNIT = DOLLARS/UNITS;
RUN;

PROC PRINT DATA= TopBrand(obs =10);RUN;


proc tabulate data = TopBrand out=MeanTopBrand;
class WEEK ;
var DOLLARS  FEAT  D  PRICEPERUNIT PR;
table WEEK, (DOLLARS FEAT D PRICEPERUNIT )* MEAN ;
run;

proc reg DATA = MeanTopBrand;
var DOLLARS_Mean  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean;
```

```
model DOLLARS_Mean =  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean;
RUN;

DATA test; SET MeanTopBrand;
RUN;

PROC PRINT DATA =test;
RUN;
```

A. What is the R-sq and adj R-sq of the model?



R-Square is 0.4813
Adj R-Square is 0.4489

a. Which coefficients are significant?

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS_Mean

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	54.15694	18.05231	14.85	<.0001
Error	48	58.36683	1.21598		
Corrected Total	51	112.52377			

Root MSE	1.10271	R-Square	0.4813
Dependent Mean	8.29407	Adj R-Sq	0.4489
Coeff Var	13.29520		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.88597	6.87424	0.42	0.6765
FEAT_Mean	1	9.86898	2.64254	3.73	0.0005
D_Mean	1	22.87124	5.93832	3.85	0.0003
PRICEPERUNIT_Mean	1	1.00521	2.32912	0.43	0.6680

Feature Mean coefficient 9.87 and Display Mean coefficient 22.87 are the significant coefficients as their p value is <0.05

b. Which variables are most important in explaining sales?
Feature and Display are most important in explaining sales as their coefficients are significant.

c. Interpret the meaning of the price coefficient? What is the price elasticity?
From the regression table we observe that price coefficient is insignificant (0.668 >0.005). Thus, price per unit doesn't much affect the dollar sale of the toothbrush ORAL B ADVANTAGE.
Price elasticity is (price coefficient * average price)/ average sales
= (1.005 * 3.01)/8.33457 = 0.362

d. Interpret the meaning of the display coefficient?
The display coefficient is significant.
We have assumed that display is a variable that takes the value 1 or 0. 1 represents both major and minor display whereas 0 represents no display.

Interpretation of the coefficient of display is that when there is a display, sales would increase by 22.87 dollars when compared to there is no display.

f. Test whether there is an interaction between display, feature and price. Comment on your findings.

```
DATA test; SET MeanTopBrand;
fd = D_Mean*FEAT_Mean;
pd = PRICEPERUNIT_Mean*D_Mean;
pf = PRICEPERUNIT_Mean*FEAT_Mean;
pdf = D_Mean*FEAT_Mean*PRICEPERUNIT_Mean;
RUN;

PROC PRINT DATA =test;
RUN;

proc reg DATA = test;
var DOLLARS_Mean  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean  fd;
model DOLLARS_Mean =  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean fd;
RUN;
proc reg DATA = test;
var DOLLARS_Mean  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean  pd;
model DOLLARS_Mean =  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean pd;
RUN;
proc reg DATA = test;
var DOLLARS_Mean  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean  pf;
model DOLLARS_Mean =  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean pf;
RUN;
proc reg DATA = test;
var DOLLARS_Mean  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean  pdf;
model DOLLARS_Mean =  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean pdf;
RUN;
```

Avg price per unit by week for Brands

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS_Mean

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	97.32367	24.33092	36.77	<.0001
Error	47	31.09689	0.66164		
Corrected Total	51	128.42056			

Root MSE	0.81341	R-Square	0.7579
Dependent Mean	8.32582	Adj R-Sq	0.7372
Coeff Var	9.76972		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.81021	5.59143	-0.14	0.8854
FEAT_Mean	1	-18.82848	5.41895	-3.47	0.0011
D_Mean	1	7.59749	3.89770	1.95	0.0573
PRICEPERUNIT_Mean	1	2.76373	1.88663	1.46	0.1496
fd	1	2.80104	0.41011	6.83	<.0001

As the p value is less than 0.05 which is significant for feature and display then there is an interaction between those two variables.

Avg price per unit by week for Brands

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS_Mean

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	72.34850	18.08713	15.16	<.0001
Error	47	56.07206	1.19302		
Corrected Total	51	128.42056			

Root MSE	1.09226	R-Square	0.5634
Dependent Mean	8.32582	Adj R-Sq	0.5262
Coeff Var	13.11889		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-8.80587	8.95901	-0.98	0.3307
FEAT_Mean	1	78.44060	28.96730	2.71	0.0094
D_Mean	1	8.45506	5.24277	1.61	0.1135
PRICEPERUNIT_Mean	1	5.25644	3.04973	1.72	0.0914
pf	1	-22.90450	10.30895	-2.22	0.0311

As the p value of price per unit and feature is less than 0.05(which is 0.0311) which is significant then there is an interaction between those two variables

Avg price per unit by week for Brands

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS_Mean

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	68.35769	17.08942	13.37	<.0001
Error	47	60.06287	1.27793		
Corrected Total	51	128.42056			

Root MSE	1.13046	R-Square	0.5323
Dependent Mean	8.32582	Adj R-Sq	0.4925
Coeff Var	13.57772		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-13.97871	15.31963	-0.91	0.3662
FEAT_Mean	1	14.03760	3.30707	4.24	0.0001
D_Mean	1	229.56732	182.05281	1.26	0.2135
PRICEPERUNIT_Mean	1	6.96155	5.22731	1.33	0.1894
pd	1	-75.72992	62.13281	-1.22	0.2290

As the P value (0.2290) is greater than 0.05 which is non significant then there is no interaction between those two variables

Avg price per unit by week for Brands

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS_Mean

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	71.91715	17.97929	14.96	<.0001
Error	47	56.50340	1.20220		
Corrected Total	51	128.42056			

Root MSE	1.09645	R-Square	0.5600
Dependent Mean	8.32582	Adj R-Sq	0.5226
Coeff Var	13.16925		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.60029	7.54633	0.48	0.6355
FEAT_Mean	1	1.03260	7.06663	0.15	0.8844
D_Mean	1	-8.25204	9.17408	-0.90	0.3730
PRICEPERUNIT_Mean	1	1.26575	2.53098	0.50	0.6193
pdf	1	70.66033	33.16272	2.13	0.0384

As the p value (0.0384) for display, feature and price is less than 0.05 whose values are significant which indicates that there is an interaction between display, feature and price

g. Test whether the effect of price is non-linear. Comment on your findings.

```
proc reg DATA = test;
var DOLLARS_Mean FEAT_Mean D_Mean PRICEPERUNIT_Mean price_square;
model DOLLARS_Mean = FEAT_Mean D_Mean PRICEPERUNIT_Mean price_square;
RUN;
```

Avg price per unit by week for Brands

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS_Mean

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	70.98896	17.74724	14.52	<.0001
Error	47	57.43159	1.22195		
Corrected Total	51	128.42056			

Root MSE	1.10542	R-Square	0.5528
Dependent Mean	8.32582	Adj R-Sq	0.5147
Coeff Var	13.27698		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	123.76888	63.63041	1.95	0.0578
FEAT_Mean	1	13.26289	3.27571	4.05	0.0002
D_Mean	1	7.41193	5.30014	1.40	0.1685
PRICEPERUNIT_Mean	1	-84.25716	44.58289	-1.89	0.0649
price_square	1	15.09364	7.83942	1.93	0.0602

As we know that x-square value is significant then the relation with the dependent variable is non-linear.
As the p value of price square(0.0602) is non-significant then the price doesn't have non linear relationship when we are considering 95% confidence level.
As there is no rule to strictly consider 95% confidence level if we are taking 90% confidence level then price is having non-linear relationship.
Actually price is showing non-linear relationship from 94% confidence level.

h. Test using VIF and COLLIN whether there is multicollinearity in the model? Comment on your findings.

```
proc reg DATA = test;
var DOLLARS_Mean  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean ;
model DOLLARS_Mean =  FEAT_Mean  D_Mean  PRICEPERUNIT_Mean / VIF tol
collin;
RUN;
quit;
```

Avg price per unit by week for Brands

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS_Mean

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66.45923	22.15308	17.16	<.0001
Error	48	61.96133	1.29086		
Corrected Total	51	128.42056			

Root MSE	1.13616	R-Square	0.5175
Dependent Mean	8.32582	Adj R-Sq	0.4874
Coeff Var	13.64622		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	2.12949	7.78686	0.27	0.7857	.	0
FEAT_Mean	1	14.46906	3.30466	4.38	<.0001	0.38031	2.62944
D_Mean	1	7.77241	5.44414	1.43	0.1599	0.86226	1.15975
PRICEPERUNIT_Mean	1	1.43999	2.62128	0.55	0.5853	0.41890	2.38719

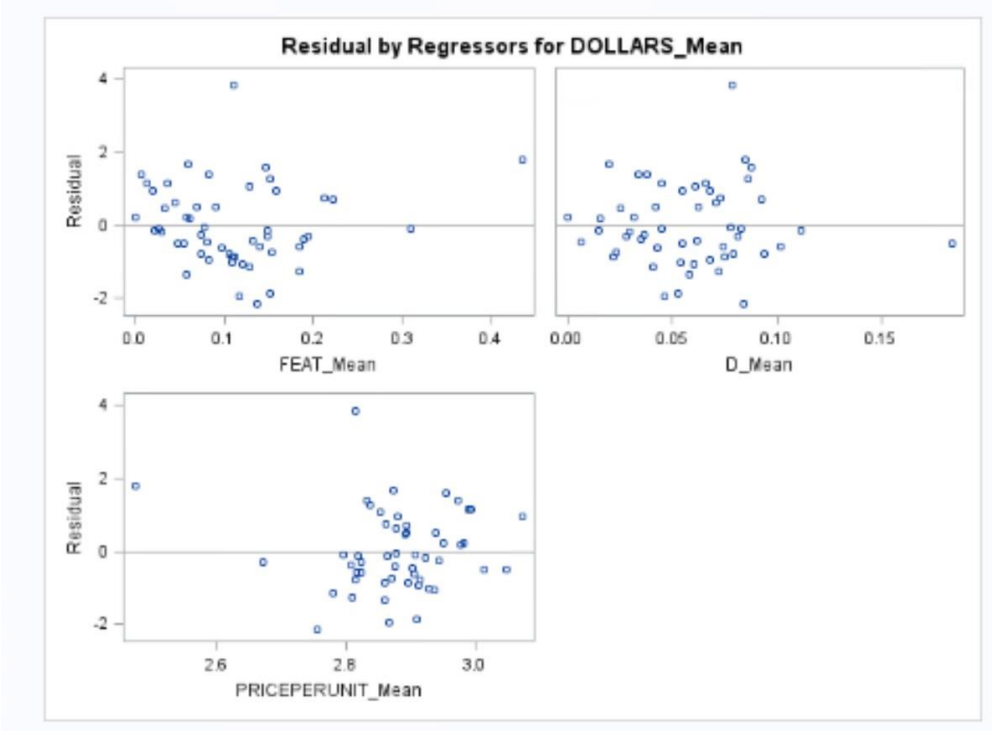
Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	FEAT_Mean	D_Mean	PRICEPERUNIT_Mean
1	3.60314	1.00000	0.00003008	0.00787	0.01310	0.00003169
2	0.24804	3.81139	0.00021791	0.35259	5.965584E-7	0.00029790
3	0.14862	4.92389	0.00024359	0.06273	0.97106	0.00024488
4	0.00021086	130.72060	0.99951	0.57681	0.01584	0.99943

As the VIF values (2.629, 1.15975, 2.38719) which are so less than 10 which is consider as benchmark for checking multicollinearity(normally greater than 10 considered as multicollinearity and between 3-8 considered to have some multicollinearity) there is no multicollinearity in the model(As the values are so less)

Similarly, for COLLIN test, the Condition Index isn't over 100 except in 1 case, where the eigen values is 0.00021086. In this case, there is a strong correlation between Price and the Feature mean values and so in this particular case of eigen value, to avoid multicollinearity, these can be combined instead of taking all of them in the model.

i. Test for presence of heteroscedasticity using White test. Do A WLS if needed. Comment on your findings.



The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS_Mean

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
9	7.36	0.5992

When the p value of chisq is significant then there is heteroscedasticity as the p value of chisq (0.5992) is not significant(greater than 0.05) then there is no presence of heteroscedasticity.

From the residual by regressors plot, we observe that the residuals of Feature mean, Display mean and price per unit are spread evenly. So there is no heteroscedasticity.