STATISTICAL PROGRAMMING FOR BUSINESS ANALYTICS

Assignment 3

Tanay Bhalerao

FEBRUARY 27, 2015

UNIVERSITY OF SOUTH FLORIDA
MANAGEMENT INFORMATION SYSTEMS

Homework 3

1. Refer to the USEDCARS data. Suppose that an econometrician wanted to use results applicable to the normal distribution to describe the prices of used cars. Used PROC UNIVARIATE to decide whether the prices or the logarithms of the prices more closely follow a normal distribution. Write down at least two findings from PROC UNIVARIATE to support your claim.

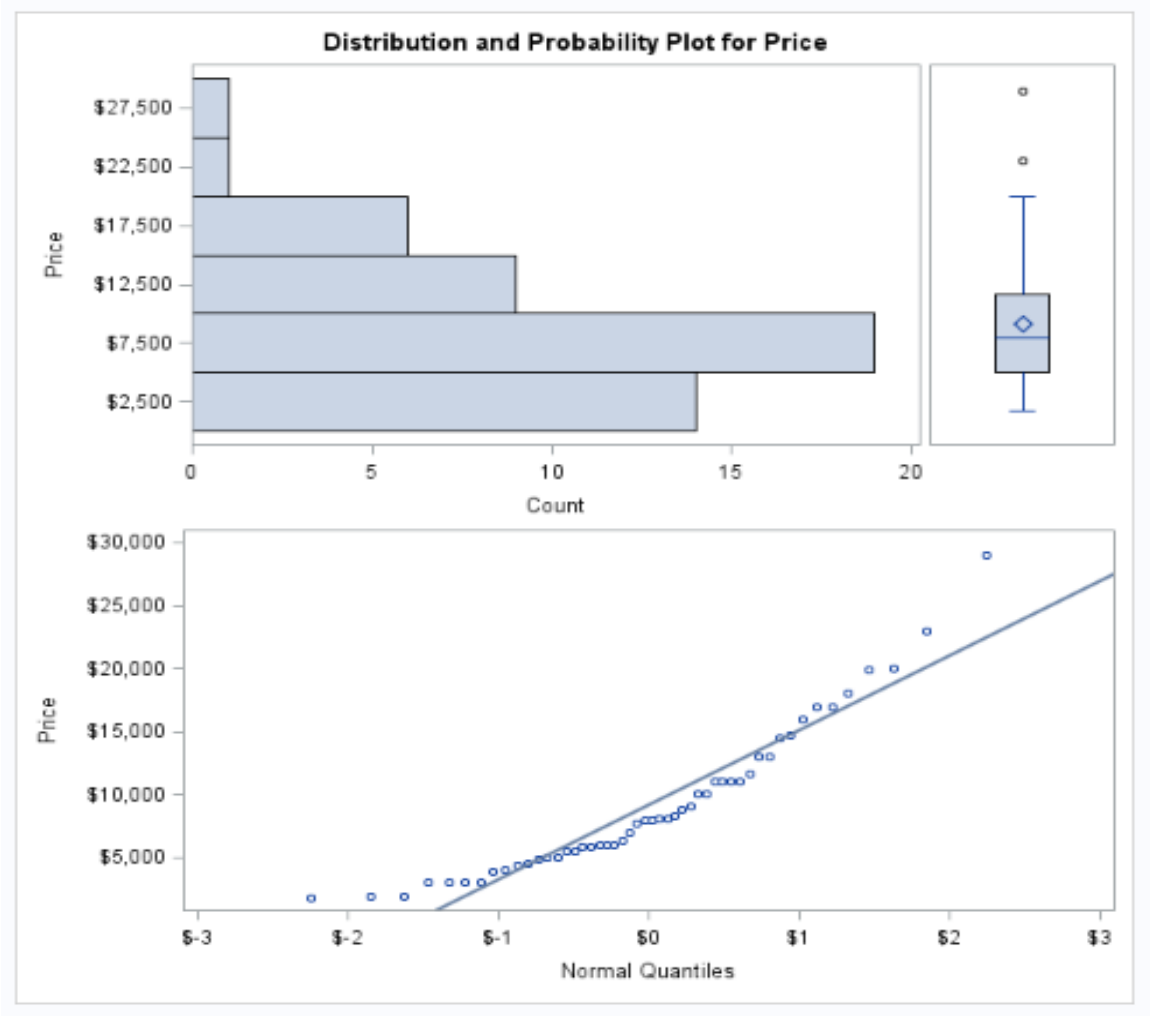
```
DATA used_cars;
      infile
"\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\usedcars.txt"
LRECL= 200 firstobs=2;
      INPUT @1 Year $2.
              @3 Manufacturer $18.
              @21 Model $15.
          @36 Miles Comma12.
          @48 Price Dollar10.
          @59 Dealer $32.;
      FORMAT Price Dollar10. Miles Comma9.;
    log_price=log(Price);
RUN;

PROC UNIVARIATE NORMAL PLOT DATA=used_cars;
var Price log_price;
TITLE "PROC UNIVARIATE-USED CARS";
histogram/ normal;
RUN;
```

PROC UNIVARIATE-USED CARS

The UNIVARIATE Procedure
Variable: Price

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.901447	Pr < W	0.0005
Kolmogorov-Smirnov	D	0.142256	Pr > D	0.0125
Cramer-von Mises	W-Sq	0.244169	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.437238	Pr > A-Sq	<0.0050

PROC UNIVARIATE Results for **tests for normality** on variable Price show us that the p-value for the different tests is a lot **less** than 0.05(arbitrary threshold). Hence we can conclude that the distribution is not normal.



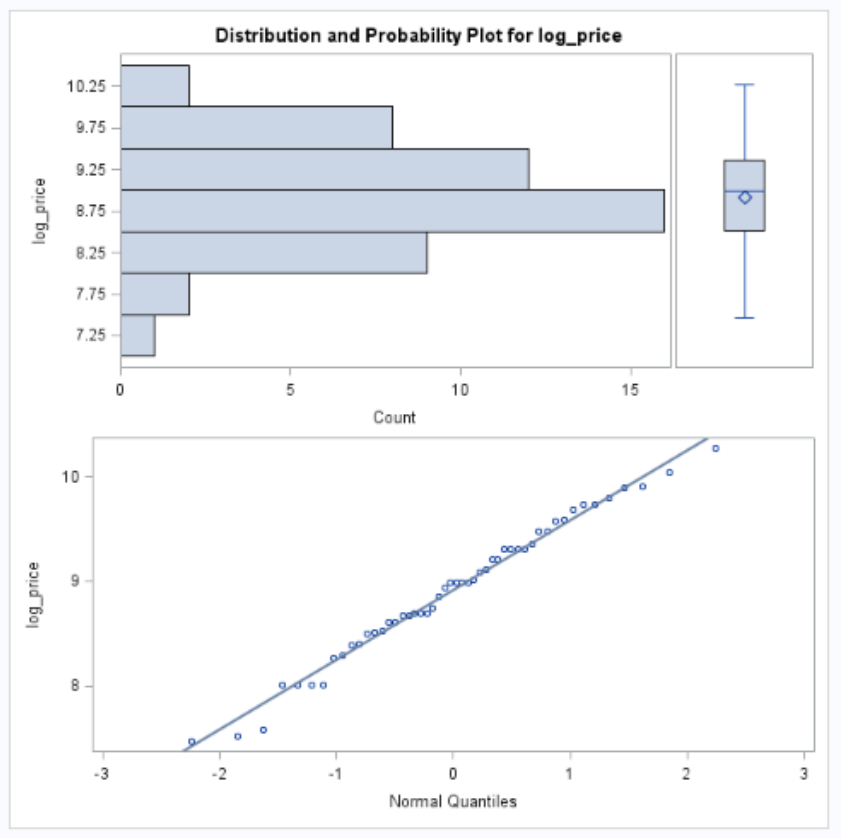
Looking at the **Distribution and probability plot** for Price, in the Q-Q plot, we see the data points are not close to the line mostly towards the edges. And hence it is not normal distribution.

PROC UNIVARIATE-USED CARS

The UNIVARIATE Procedure
Variable: log_price

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.983487	Pr < W	0.7051
Kolmogorov-Smirnov	D	0.059723	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.026728	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.210266	Pr > A-Sq	>0.2500

PROC UNIVARIATE Results for **tests for normality** on variable log_price show us that the p-value for the different tests is more than 0.05. Hence we can conclude that the distribution is **normal**.



Again when we observe the **Distribution and probability plot** for log_price, in the Q-Q plot, we see the data points are close to the line. And hence it can be considered as a **normal distribution**.

2. The midrange statistic is sometimes used to report a central value of a distribution. The midrange is defined as (minimum value + maximum value)/2. Refer to the LIMES data. Use PROC UNIVARIATE to calculate the midrange of the juice liquid volumes of the limes. For this problem, you must use SAS to perform all of the calculations; for example, you may not find the minimum and the maximum with SAS then calculate the midrange by hand. Use PROC IMPORT to import the data:

```
PROC IMPORT OUT= WORK.LIMES
            DATAFILE=
"\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\limes.txt"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

PROC UNIVARIATE DATA=WORK.LIMES noprint;
            var juice_vol;
                  output out= maxvol MAX=vol_max MIN=vol_min;
RUN;

DATA MidR;
Set maxvol;
Midrange=(vol_max+vol_min)/2;
RUN;

PROC PRINT DATA=MidR;
TITLE "CENTRAL VALUE OF A DISTRIBUTION"
      RUN;
```

CENTRAL VALUE OF A DISTRIBUTION

Obs	vol_max	vol_min	Midrange
1	67	3	35

3. Refer to CATS3 data. Suppose that the veterinarian wants to see if the treatment had altered kidney function within the first week after surgery. One way to do this is to perform a paired t-test. Calculate a new variable representing ((GFR of the untreated kidney in Week1) minus (GFR of the surgically-treated kidney in Week1)) for each of the 8 cats. Then, apply PROC UNIVARIATE to those differences. The p-value of the 2-sided t-test is the number marked Pr>|T|. Based on this value, would you decide that the surgery had an effect after one week?

```
DATA cats3;
      infile
"\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\cats3.txt" LRECL=
200 firstobs=2;
      Input Cat $ Side $ Week_0 Week_1 Week_2;
      GFR_val= (Week_0 - Week_1);
RUN;

PROC UNIVARIATE DATA=cats3;
      var GFR_val;
      TITLE "GFR VALUE T-Test";
      RUN;
```

GFR VALUE T-Test

The UNIVARIATE Procedure
Variable: GFR_val

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	4.483446	Pr > |t|	0.0029
Sign	M	4	Pr >= |M|	0.0078
Signed Rank	S	18	Pr >= |S|	0.0078

We observe the Pr >|t| value which is **0.0029** which is very less in comparison to the threshold of 5%. Hence the null hypothesis is rejected and we can state that surgery had an effect on the cats after one week.

4. Book page 157: Question 4.4, 4.6, 4.8, 4.10, 4.12

4.4

```
proc sort data = tanay.clinical;
      by PATIENT VISIT;
run;

data tanay.clinical_diff;
      set tanay.clinical;
      by PATIENT;
      DIFF_WEIGHT = WEIGHT - lag(WEIGHT);
      if not first.PATIENT then output;
run;
```

```
proc print data=tanay.clinical_diff;
run;
```

Obs	PATIENT	VISIT	DATE_VISIT	WEIGHT	GENDER	GROUP	DIFF_WEIGHT
1	1	2	11JUL2003	165	Male	C	8
2	1	3	26NOV2003	181	Male	C	16
3	3	2	12JUL2003	150	Male	B	-2
4	3	3	21DEC2003	139	Male	B	-11
5	3	4	23APR2004	146	Male	B	7
6	5	2	09AUG2003	126	Male	C	3
7	5	3	24JAN2004	123	Male	C	-3
8	7	2	04SEP2003	178	Male	A	13
9	7	3	25FEB2004	173	Male	A	-5
10	7	4	16SEP2004	171	Male	A	-2
11	9	2	26JUL2003	158	Female	B	5
12	10	2	10JUL2003	145	Female	A	4
13	12	2	02SEP2003	158	Female	A	3
14	13	2	19SEP2003	123	Male	C	-17
15	13	3	14JAN2004	136	Male	C	13
16	15	2	26JUN2003	161	Male	C	13
17	17	2	17OCT2003	149	Male	B	0
18	17	3	16NOV2003	160	Male	B	11
19	17	4	08APR2004	151	Male	B	-9
20	19	2	27JUL2003	155	Female	B	5
21	19	3	25OCT2003	155	Female	B	0
22	19	4	02FEB2004	132	Female	B	-23

23	20	2	05OCT2003	171	Female	C	15
24	21	2	29JUL2003	156	Female	A	11
25	23	2	10AUG2003	131	Male	A	4
26	23	3	11DEC2003	152	Male	A	21
27	23	4	21APR2004	157	Male	A	5

4.6

```
proc sort data = tanay.clinical;
      by PATIENT VISIT;
run;


data tanay.CHANGE;
      set tanay.clinical;
      by PATIENT;
      retain first_weight first_visit;
      if first.PATIENT and last.PATIENT then delete;
      if first.PATIENT then do;
            first_weight = WEIGHT;
            first_visit = VISIT;
      end;
      if last.PATIENT then do;
            WT_CHANGE = WEIGHT - first_weight;
            NO_OF_DAYS = VISIT - first_visit;
            output;
      end;
run;

proc print data=tanay.CHANGE;
      run;
```

Obs	PATIENT	VISIT	DATE_VISIT	WEIGHT	GENDER	GROUP	first_weight	first_visit	WT_CHANGE	NO_OF_DAYS
1	1	3	26NOV2003	181	Male	C	157	1	24	2
2	3	4	23APR2004	146	Male	B	152	1	-6	3
3	5	3	24JAN2004	123	Male	C	123	1	0	2
4	7	4	16SEP2004	171	Male	A	165	1	6	3

5	9	2	26JUL2003	158	Female	B	153	1	5	1
6	10	2	10JUL2003	145	Female	A	141	1	4	1
7	12	2	02SEP2003	158	Female	A	155	1	3	1
8	13	3	14JAN2004	136	Male	C	140	1	-4	2
9	15	2	26JUN2003	161	Male	C	148	1	13	1
10	17	4	08APR2004	151	Male	B	149	1	2	3
11	19	4	02FEB2004	132	Female	B	150	1	-18	3
12	20	2	05OCT2003	171	Female	C	156	1	15	1
13	21	2	29JUL2003	156	Female	A	145	1	11	1
14	23	4	21APR2004	157	Male	A	127	1	30	3

4.8

```
proc means data=tanay.clinical NOPRINT NWAY;
	class PATIENT;
	var WEIGHT;
	output out = tanay.weightStats
		MEAN = wMean
		MEDIAN = wMedian
		MIN = wMin
		MAX = wMax;
run;

proc print data=tanay.weightStats;
	run;
```

Obs	PATIENT	_TYPE_	_FREQ_	wtMean	wtMedian	wtMin	wtMax
1	1	1	3	167.667	165.0	157	181

2	3	1	4	146.750	148.0	139	152
3	4	1	1	144.000	144.0	144	144
4	5	1	3	124.000	123.0	123	126
5	6	1	1	133.000	133.0	133	133
6	7	1	4	171.750	172.0	165	178
7	8	1	1	148.000	148.0	148	148
8	9	1	2	155.500	155.5	153	158
9	10	1	2	143.000	143.0	141	145
10	12	1	2	156.500	156.5	155	158
11	13	1	3	133.000	136.0	123	140
12	15	1	2	154.500	154.5	148	161
13	16	1	1	148.000	148.0	148	148
14	17	1	4	152.250	150.0	149	160
15	18	1	1	133.000	133.0	133	133
16	19	1	4	148.000	152.5	132	155
17	20	1	2	163.500	163.5	156	171
18	21	1	2	150.500	150.5	145	156
19	22	1	1	158.000	158.0	158	158
20	23	1	4	141.750	141.5	127	157
21	24	1	1	172.000	172.0	172	172
22	25	1	1	175.000	175.0	175	175

4.10

```
proc means data=tanay.clinical noprint chartype;
      class GENDER GROUP;
      var WEIGHT;
      output out = tanay.summary
            MEAN = wtMean
            MEDIAN = wtMedian
            STD = wtSTD;
```

```
run;


data tanay.grand tanay.gender tanay.group tanay.gender_group;
      set tanay.summary;
      if _TYPE_ = '00' then output tanay.grand;
      else if _TYPE_ = '01' then output tanay.group;
      else if _TYPE_ = '10' then output tanay.gender;
      else if _TYPE_ = '11' then output tanay.gender_group;
run;

proc print data=tanay.grand;
      title "GRAND STATS";
run;



proc print data=tanay.group;
      title "GROUP STATS";
run;

proc print data=tanay.gender;
      title "GENDER STATS";
run;


proc print data=tanay.gender_group;
      title "GENDER AND GROUP STATISTICS";
run;
```

GRAND STATS

Obs	GENDER	GROUP	_TYPE_	_FREQ_	wtMean	wtMedian	wtSTD
1			00	49	150.490	151	14.8241

GROUP STATS

Obs	GENDER	GROUP	_TYPE_	_FREQ_	wtMean	wtMedian	wtSTD
1		A	01	15	153.467	155	14.6427
2		B	01	17	151.706	151	11.2903
3		C	01	17	146.647	148	17.8499

GENDER STATS

Obs	GENDER	GROUP	_TYPE_	_FREQ_	wtMean	wtMedian	wtSTD
1	Female		10	15	151.867	155.0	8.9910
2	Male		10	34	149.882	149.5	16.8555

GENDER AND GROUP STATISTICS

Obs	GENDER	GROUP	_TYPE_	_FREQ_	wtMean	wtMedian	wtSTD
1	Female	A	11	7	149.714	148.0	6.5756
2	Female	B	11	6	150.500	154.0	9.4393
3	Female	C	11	2	163.500	163.5	10.6066
4	Male	A	11	8	156.750	161.0	19.1143
5	Male	B	11	11	152.364	150.0	12.5720
6	Male	C	11	15	144.400	144.0	17.6101

4.12

```
data tanay.GRAND tanay.BY_GENDER tanay.BY_GROUP tanay.BY_GENDER_GROUP;
      set tanay.summaryData;
      if _TYPE_ = '00' then output tanay.GRAND;
      else if _TYPE_ = '01' then output tanay.BY_GROUP;
      else if _TYPE_ = '10' then output tanay.BY_GENDER;
      else if _TYPE_ = '11' then output tanay.BY_GENDER_GROUP;
run;


proc print data=tanay.GRAND;
      title "GRAND STATS";
run;


proc print data=tanay.BY_GROUP;
      title "GROUP STATS";
run;
```

```
proc print data=tanay.BY_GENDER;
      title "GENDER STATISTICS";
run;

proc print data=tanay.BY_GENDER_GROUP;
      title "GENDER AND GROUP STATISTICS";
run;
```

GRAND STATS

Obs	GENDER	GROUP	_TYPE_	_FREQ_	wtMean	wtMedian	wtSTD
1			00	49	150.490	151	14.8241

GROUP STATS

Obs	GENDER	GROUP	_TYPE_	_FREQ_	wtMean	wtMedian	wtSTD
1		A	01	15	153.467	155	14.6427
2		B	01	17	151.706	151	11.2903
3		C	01	17	146.647	148	17.8499

GENDER STAT

Obs	GENDER	GROUP	_TYPE_	_FREQ_	wtMean	wtMedian	wtSTD
1	Female		10	15	151.867	155.0	8.9910
2	Male		10	34	149.882	149.5	16.8555

GENDER AND GROUP STATISTICS

Obs	GENDER	GROUP	_TYPE_	_FREQ_	wtMean	wtMedian	wtSTD
1	Female	A	11	7	149.714	148.0	6.5756
2	Female	B	11	6	150.500	154.0	9.4393
3	Female	C	11	2	163.500	163.5	10.6066
4	Male	A	11	8	156.750	161.0	19.1143
5	Male	B	11	11	152.364	150.0	12.5720
6	Male	C	11	15	144.400	144.0	17.6101

5. Refer to USEDCARS dataset. Write a SAS program which reads the full dataset. Then create a dataset which contains only the least expensive car offered by each dealer. Print the new dataset, showing the year, manufacturer, model, price, and name of the dealer. This dataset should have 15 observations, and each dealer should appear exactly once.

```
DATA used_cars;
      infile
"\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\usedcars.txt"
LRECL= 200 firstobs=2;
      INPUT @1 Year $2.
              @3 Manufacturer $18.
              @21 Model $15.
          @36 Miles Comma12.
          @48 Price Dollar10.
          @59 Dealer $32.;
      FORMAT Price Dollar10. Miles Comma9.;
RUN;
PROC SORT DATA=used_cars out=used_car_sort;
      by Dealer Price;
RUN;

PROC PRINT DATA=used_car_sort;
RUN;

DATA TEMP;
      Set  used_car_sort;
      by Dealer Price;
      min_price=first.Dealer;
RUN;

PROC PRINT DATA=TEMP;
TITLE "ABC";
RUN;

PROC SORT Data= TEMP out= min_pri_dealer;
BY Dealer Price;
RUN;
```

```
Data low_price;
      set min_pri_dealer;
      by Dealer;
      if first.Dealer;
RUN;

PROC PRINT Data=low_price;
TITLE "Least Expensive Cars With Dealers";
RUN;
```

Least Expensive Cars With Dealers

Obs	Year	Manufacturer	Model	Miles	Price	Dealer	min_price
1	97	Geo	Metro	.	$6,988	Budget Car Sales	1
2	93	Dodge	Colt	.	$3,995	Bush Gator	1
3	88	Toyota	Tercel	60,000	$3,895	Gainesville Nissan	1
4	92	Geo	Prizm	.	$4,995	Gatorland Toyota	1
5	93	Mazda	Protege	.	$5,787	Hawes Chrysler Plymouth	1
6	88	Honda	Prelude	.	$4,900	Hometown Motors	1
7	94	Oldsmobile	Cutlass	.	$7,995	Kraft Motorcar	1
8	91	Geo	Storm	.	$2,995	Magic Imports	1
9	95	Ford	Aspire	52,000	$5,995	Santa Fe Ford	1
10	94	Toyota	Corolla	.	$7,665	Saturn of Gainesville	1
11	82	Volvo	240	.	$2,995	Taylor Volvo	1
12	96	Buick	LeSabre	25,000	$16,900	Tomlinson Motor Company	1
13	85	Pontiac	Grand Am	.	$1,750	University Auto	1
14	93	Pontiac	Grand Am	.	$5,485	Wade Raulerson	1
15	94	Ford	Ranger	.	$9,994	White Ford	1

For Dealer: Magic Import and Santa Fe Ford has different models of cars but their least expensive cars have duplicates. If I include all the values the number of observations go up to 18 records. Technically I algorithm generated a value for the temp column min_price as 1 for the first model ,hence I filtered it on the basis of the min_price column value generated.

24	91	Geo	Storm	.	$2,995	Magic Imports	1
25	93	Ford	Escort	.	$2,995	Magic Imports	0
26	92	Mitsubishi	Mirage	.	$2,995	Magic Imports	0
27	95	Ford	Aspire	52,000	$5,995	Santa Fe Ford	1
28	94	Ford	Tempo	39,000	$5,995	Santa Fe Ford	0
29	97	Ford	Aspire	26,000	$7,995	Santa Fe Ford	0
30	96	Ford	Explorer	32,000	$17,999	Santa Fe Ford	0

6. Refer to the BREAD dataset. Suppose that you need to create a reference list of bread recipes that do not use eggs (for dietary requirements or preferences, or perhaps you forgot to buy eggs). Create and print a permanent SAS dataset, using LIBNAME and associated commands, which contains only the recipes which use no eggs.

```
libname perm "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\";
DATA perm.bread;
infile "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\bread.txt"
LRECL= 200 firstobs=3 DLM=',';
Input dough_type $ water oil sugar salt dry_milk flour yeast wheat oregano
eggs;
RUN;
DATA perm.bread;
set perm.bread;
if eggs=0;
RUN;
PROC Print DATA=perm.bread;
TITLE "RECIPES WITHOUT EGGS";
RUN;
```

RECIPES WITHOUT EGGS

Obs	dough_type	water	oil	sugar	salt	dry_milk	flour	yeast	wheat	oregano	eggs
1	white	1.000	2.00	2.0	1.25	1.50	3.00	2.00	0.00	0	0
2	French	1.125	0.00	1.5	1.00	0.00	3.50	2.00	0.00	0	0
3	wheat	1.500	2.00	2.0	2.00	0.00	3.00	2.25	0.75	0	0
4	Italian	1.125	2.00	3.0	1.50	2.00	3.25	1.50	0.00	2	0
5	pizza	0.750	1.00	1.0	0.50	1.00	2.25	1.00	0.00	0	0
6	bagel	1.000	0.00	1.5	1.00	0.00	3.00	2.25	0.00	0	0
7	milk	1.000	2.00	0.5	1.50	5.33	3.00	1.75	0.00	0	0
8	foccacia	1.000	5.33	2.0	1.00	0.00	3.00	1.50	0.00	1	0

7. Refer to the CLINTON dataset. Gallup has conducted more polls to assess President Clinton's job approval rating since the CLINTON dataset was created. The data are shown below:

Date	Approve	Disapprove	No opinion
8-18-98	66	29	5
8-20-98	61	34	5
8-21-98	62	35	3
9-1-98	62	33	5
9-10-98	60	37	3
9-11-98	63	34	3

Create two datasets in SAS. One dataset should consist of the numbers in the file CLINTON.TXT and the second dataset should contain the numbers listed above. Combine the two datasets into a larger dataset with the appropriate commands, sort the observations in that dataset in

descending order by date (so that September 11, 1998 appears first), and print the larger dataset. Use an appropriate format to print the date variable.

```
DATA tanay.clinton;
      infile
"\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\clinton.txt"
firstobs=3;
      Input Day 7-8 Mo $ 10-12 Year 16-17 Approve 24-25 Disapprove 32-33
No_opinion 40-41;
      Month=int(month(input("01"||substr(Mo,1,3)||"2014",date9.)));
      drop Mo;
RUN;

DATA tanay.clinton2;
      infile
"\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\clinton2.txt"
firstobs=2 dlm='09'x;
      input Date $ Approve Disapprove No_opinion;
      Month = int((scan(Date,1,'-')));
      Day = int(scan(Date,2,'-'));
      Year = int(scan(Date,3,'-'));
      drop Date;
RUN;

proc sort data=tanay.clinton;
      by Year Month Day;
run;

proc sort data=tanay.clinton2;
      by Year Month Day;
run;

data tanay.merge_data;
      merge tanay.clinton tanay.clinton2;
      by Year Month Day;
run;
 proc sort data=tanay.merge_data;
      by descending Year descending Month descending Day;
run;

proc print data=tanay.merge_data;
Title "Clinton Data-Updated";
      run;
```

Clinton Data-Updated

Obs	Day	Year	Approve	Disapprove	No_opinion	Month
1	11	98	63	34	3	9
2	10	98	60	37	3	9
3	1	98	62	33	5	9
4	21	98	62	35	3	8
5	20	98	61	34	5	8
6	18	98	66	29	5	8
7	17	98	62	32	6	8
8	10	98	65	30	5	8
9	7	98	64	32	4	8
10	29	98	65	31	4	7
11	7	98	61	34	5	7
12	22	98	60	34	6	6
13	5	98	60	34	6	6
14	8	98	64	31	5	5
15	17	98	63	31	6	4
16	20	98	66	28	6	3
17	16	98	67	29	4	3
18	6	98	63	31	6	3
19	20	98	66	29	5	2
20	13	98	66	30	4	2
21	30	98	69	28	3	1
22	28	98	67	28	5	1
23	25	98	59	37	4	1
24	24	98	60	35	5	1
25	23	98	58	36	6	1
26	16	98	60	30	10	1

There were total 146 records after merge