SAS PROGRAMMING FOR BUSINESS ANALYTICS

Assignment 1

SUBMITTED BY

TANAY BHALERAO

U47707491

FEBRUARY 13, 2015

UNIVERSITY OF SOUTH FLORIDA

Management Information Systems

Homework 1

1. Refer to the CATS1 dataset. Write a SAS program to read the data file from your USB with anINFILE statement and print the dataset. There should be eight observations and two variables.There are 2 variables: the cat name and the direction. Make sure you look at the text file andsee where the columns are so that you specify them correctly. Copy your code and the resultingprint out in a word document and upload it to Canvas.

Solution:

```
DATA infile_cats1;
      infile "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\cats1.txt";
      LRECL= 200;
      input
      cats $ direction $
      ;
RUN;

PROC PRINT DATA=infile_cats1;
      TITLE "CATS using INFILE";
      VAR cats direction;
RUN;
```

CATS using INFILE

Obs	cats	direction
1	Garfield	Left
2	Felix	Right
3	Hobbes	Left
4	Catbert	Left
5	Bill	Right
6	Scratchy	Right
7	Stimpy	Right
8	Attilaz	Left

2. Refer to the DOGS1 dataset. Write a SAS program to read the data file from your USB with anINFILE statement and print the dataset. There should be 25 observations and six variables.

input dog $ 1-8 concent 16 sex $ 17 age 31-32 haircoat $ 33-37 weight45-48;

Solution:

```
DATA infile_dogs1;
      infile "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\dogs1.txt" LRECL=
200 firstobs=2;
      input
      dog $ 1-8 concent 16 sex $ 17 age 31-32 haircoat $ 33-37 weight 45-48;
RUN;

PROC PRINT DATA=infile_dogs1;
      TITLE "DOGS using INFILE";
      VAR dog concent sex age haircoat weight;
RUN;
```

DOGS using INFILE

Obs	dog	concent	sex	age	haircoat	weight
1	georgia	0	F	41	Med	11.10
2	max	0	M	22	Short	7.05
3	cai	0	M	12	Short	20.00
4	jessie	0	F	44	Med	17.00
5	pandora	0	F	19	Short	19.90
6	lucy	0	F	24	Short	6.30
7	simon	0	M	36	Short	5.75
8	baby	0	F	24	Short	16.40
9	cleo	0	F	10	Short	19.90
10	savannah	0	F	10	Short	8.05
11	cooper	1	M	10	Med	17.90
12	roxanne	1	F	39	Short	18.20
13	sheppy	1	M	90	Short	7.40
14	muttney	1	F	56	Short	11.80
15	bijou	1	F	14	Short	22.00
16	oreo	1	M	60	Med	21.80
17	tj	1	M	32	Med	18.90
18	lu	2	F	17	Med	16.00
19	rhea	2	F	54	Med	18.20
20	phoenix	2	F	67	Med	17.00
21	peewee	2	M	18	Short	8.45
22	penelope	2	F	9	Short	8.10
23	princess	2	F	36	Med	19.90
24	tanner	2	M	72	Short	10.00
25	elliott	2	M	54	Med	6.50

3. Refer to the DOGS2 dataset. Write a SAS program to read the data file from your USB with anINFILE statement, write a permanent SAS dataset onto your USB and print the dataset. Thereshould be 25 observations and four variables.

Solution:

```
Libname tanay "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\";
Data tanay.dogs2;

infile "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\dogs2.txt" DLM='09'x
firstobs=3;
input Dog_name $ Week0_wbc Week2_wbc Week4_wbc;

proc print DATA = tanay.dogs2;
title "DOGS using permanent SAS Dataset";
run;
```

DOGS using permanent SAS Dataset

Obs	Dog_name	Week0_wbc	Week2_wbc	Week4_wbc
1	baby	5600	5200	5900
2	bijou	11500	9500	8600
3	cai	12800	13000	11800
4	cleo	7300	7500	7600
5	cooper	7000	8000	7100
6	elliott	11400	9000	9800
7	georgia	5500	6900	6500
8	jessie	4400	6300	6100
9	lu	9400	17100	9000
10	lucy	9200	12700	9700
11	max	8700	7900	11100
12	muttney	19600	10700	12100
13	oreo	6400	9300	6800
14	pandora	9500	11500	18900
15	peewee	8300	7100	9900
16	penelope	8000	9900	8400
17	phoenix	7300	6200	4900
18	princess	8500	6900	8600
19	rhea	5100	6800	5600
20	roxanne	11200	12200	18600
21	savannah	9800	7000	7000
22	sheppy	5900	4200	5800
23	simon	7000	7600	8200
24	tanner	6000	4600	5600
25	tj	6800	4600	6700

4. Chapter 1 in the book: 1.2, 1.6, 1.8, 1.10

Question 1.2

Solution:

```
DATA Diet;
      Input SUBJ $ 1-3 HEIGHT 4-5 WT_INIT 6-8 WT_FINAL 9-11;
      BMI_INIT=((WT_INIT/2.2)/(HEIGHT*0.254))**2;
      BMI_FINAL=((WT_FINAL/2.2)/(HEIGHT*0.254))**2;
      BMI_DIFF=BMI_FINAL-BMI_INIT;
Datalines;
00768155150
00272250240
00563240200
00170345298
;
RUN;
```

```
PROC SORT DATA= Diet;
      BY SUBJ;
RUN;
PROC PRINT DATA=Diet;
      VAR SUBJ HEIGHT BMI_INIT BMI_FINAL BMI_DIFF;
RUN;
```

The SAS System

Obs	SUBJ	HEIGHT	BMI_INIT	BMI_FINAL	BMI_DIFF
1	001	70	77.7910	58.0395	-19.7515
2	002	72	38.6102	35.5832	-3.0270
3	005	63	46.4760	32.2750	-14.2010
4	007	68	16.6392	15.5830	-1.0562

Question 1.6

Solution:

```
DATA Survey;
      Input QUES1 $ 1 QUES2 $ 2 QUES3 $ 3 QUES4 $ 4 QUES5 $ 5;
Datalines;
ABCDE
AACCE
BBBBB
CABDA
DDAAC
CABBB
EEBBB
ACACA
;
RUN;
PROC FREQ DATA=Survey ORDER=FREQ;
      TABLES QUES1 QUES2 QUES3 QUES4 QUES5/ NOCUM;
RUN;
```

The SAS System

The FREQ Procedure

QUES1	Frequency	Percent
A	3	37.50
C	2	25.00
B	1	12.50
D	1	12.50
E	1	12.50

QUES2	Frequency	Percent
A	3	37.50
B	2	25.00
C	1	12.50
D	1	12.50
E	1	12.50

QUES3	Frequency	Percent
B	4	50.00
A	2	25.00
C	2	25.00

QUES4	Frequency	Percent
B	3	37.50
C	2	25.00
D	2	25.00
A	1	12.50

QUES5	Frequency	Percent
B	3	37.50
A	2	25.00
E	2	25.00
C	1	12.50

Question 1.8

Solution:

```
DATA Employee ;
      INPUT EMPID SALARY JCLASS;
      DLM='09'x;
      IF JCLASS EQ 1 THEN BONUS=0.1*SALARY;
      ELSE IF JCLASS EQ 2 THEN BONUS=0.15*SALARY;
      ELSE IF JCLASS EQ 3 THEN BONUS=0.2*SALARY;
      NEW_SALARY=SALARY + BONUS;
Datalines;
137   28000 1
214   98000 3
199   150000      3
355   57000 2
;
RUN;

PROC PRINT DATA=Employee;
      TITLE "NEW SALARY";
RUN;
```

NEW SALARY

Obs	EMPID	SALARY	JCLASS	DLM	BONUS	NEW_SALARY
1	137	28000	1		2800	30800
2	214	98000	3		19600	117600
3	199	150000	3		30000	180000
4	355	57000	2		8550	65550

Question 1.10

Solution:

```
DATA RAIN;
      INPUT CITY $ RAIN_JUNE RAIN_JULY RAIN_AUGUST;
      DLM='09'x;
      AVERAGE=(RAIN_JUNE+RAIN_JULY+RAIN_AUGUST)/3;
      PERCENT_JUNE=(RAIN_JUNE/AVERAGE)*100;
      PERCENT_JULY=(RAIN_JULY/AVERAGE)*100;
      PERCENT_AUGUST=(RAIN_AUGUST/AVERAGE)*100;
Datalines;
Trenton      23     25     30
Newark       18     27     22
Albany       22     21     27
;
RUN;
PROC SORT DATA=RAIN;
      BY CITY;
RUN;
PROC PRINT DATA=RAIN;
      ID CITY;
      VAR CITY RAIN_JUNE RAIN_JULY RAIN_AUGUST AVERAGE PERCENT_JUNE PERCENT_JULY
PERCENT_AUGUST;
RUN;
PROC MEANS DATA=RAIN MEAN STD ALPHA=0.05 MAXDEC=2;
      TITLE "STATISTICS";
RUN;
```

STATISTICS

CITY	CITY	RAIN_JUNE	RAIN_JULY	RAIN_AUGUST	AVERAGE	PERCENT_JUNE	PERCENT_JULY	PERCENT_AUGUST
Albany	Albany	22	21	27	23.3333	94.2857	90.000	115.714
Newark	Newark	18	27	22	22.3333	80.5970	120.896	98.507
Trenton	Trenton	23	25	30	26.0000	88.4615	96.154	115.385

STATISTICS

The MEANS Procedure

Variable	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
RAIN_JUNE	21.00	2.65	14.43	27.57
RAIN_JULY	24.33	3.06	16.74	31.92
RAIN_AUGUST	26.33	4.04	16.29	36.37
AVERAGE	23.89	1.90	19.18	28.60
PERCENT_JUNE	87.78	6.87	70.72	104.85
PERCENT_JULY	102.35	16.35	61.73	142.97
PERCENT_AUGUST	109.87	9.84	85.42	134.31

5. Chapter 2 in the book: 2.2, 2.4, 2.6, 2.8, 2.10

Question 2.2

Solution:

```
DATA CLINIC;
      INPUT ID $ 1-3 GENDER $ 4 RACE $ 5 HR 6-8 SBP 9-11 DBP 12-14 N_PROC 15-16;
      AVE_BP=DBP+((1/3)*(SBP-DBP));
Datalines;
001MW08013008010
002FW08811007205
003MB05018810002
004FB   10806801
005MW06812208204
006FB101   07404
007FW07810406603
008MW04811207006
009FB07719011009
010FB06616410610
;
RUN;
PROC MEANS DATA=CLINIC N MEAN STD CLM MEDIAN ALPHA=0.05;
VAR SBP DBP AVE_BP;
RUN;
```

The SAS System

The MEANS Procedure

Variable	N	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean	Median
SBP	9	136.4444444	34.8105986	109.6866496	163.2022392	122.0000000
DBP	10	82.8000000	16.4708497	71.0174639	94.5825361	77.0000000
AVE_BP	9	101.3333333	22.7986354	83.8087508	118.8579159	95.3333333

Question 2.4

Solution:

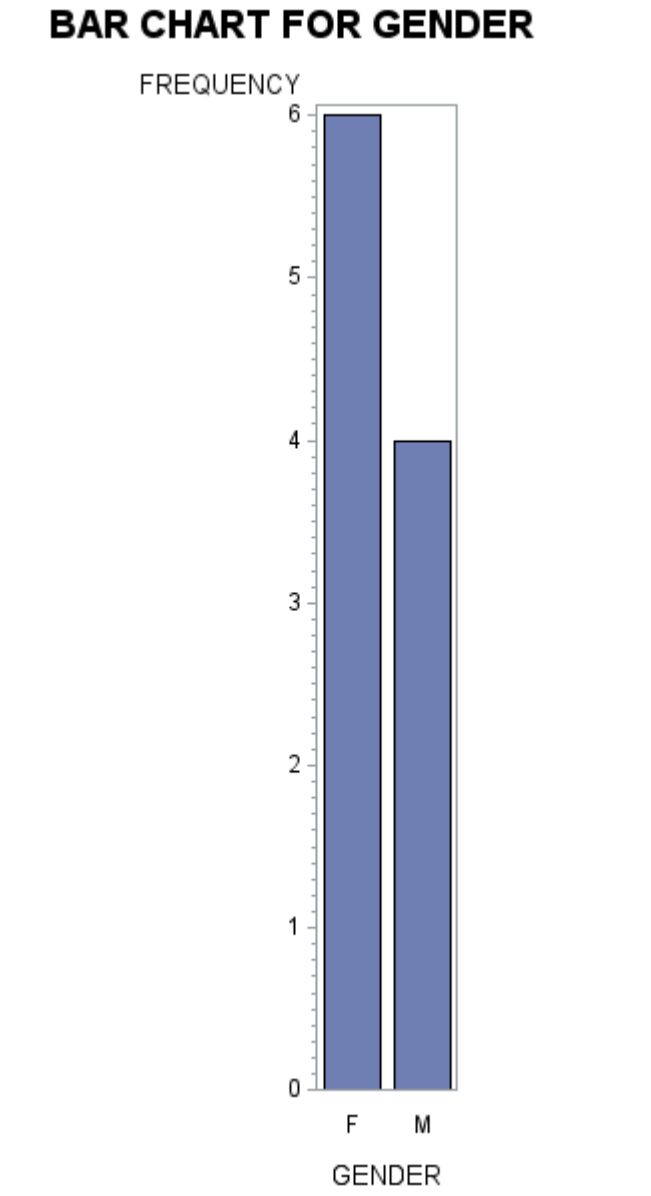
```
PROC FREQ DATA=CLINIC ORDER=FREQ;
      TABLES GENDER /NOCUM NOPERCENT;
RUN;
```

The SAS System

The FREQ Procedure

GENDER	Frequency
F	6
M	4

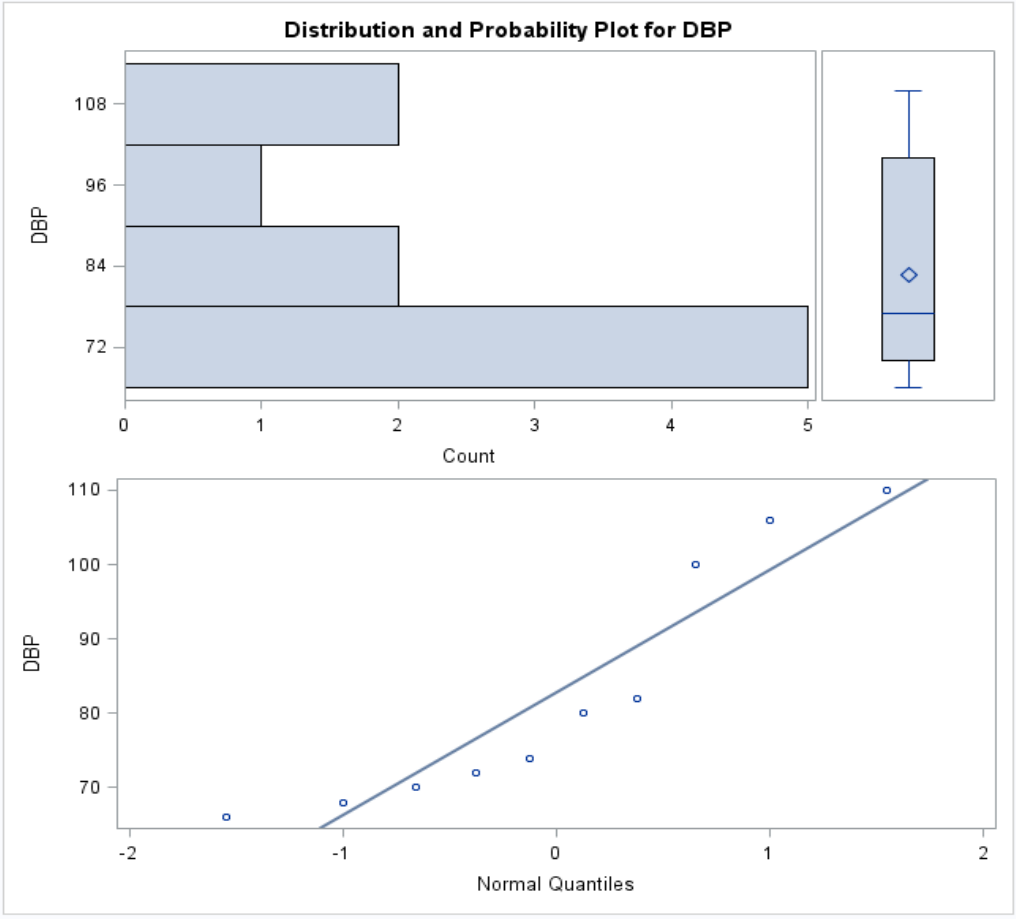
```
PROC GCHART DATA=CLINIC;
      TITLE "BAR CHART FOR GENDER";
      VBAR GENDER;
RUN;
```

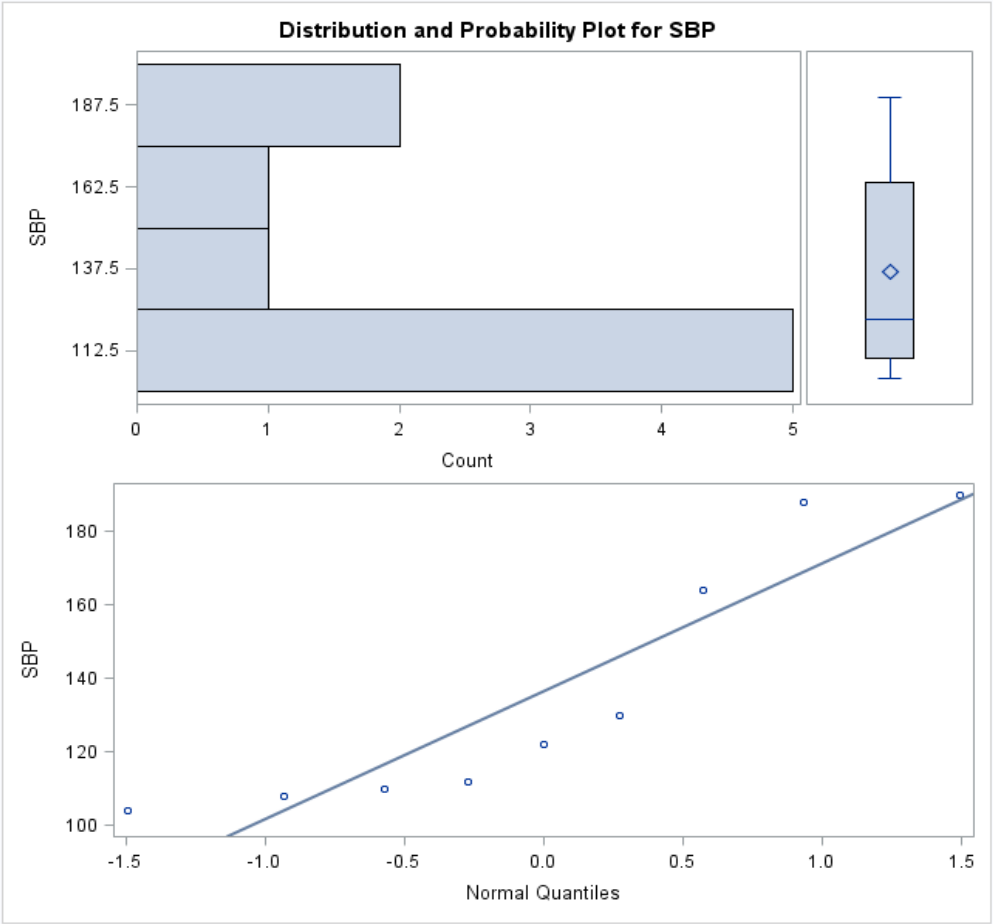


Question 2.6

Solution:

```
PROC UNIVARIATE DATA=CLINIC NORMAL PLOT;
      VAR SBP DBP;
      TITLE "STEM n LEAF AND BOX PLOT";
RUN;
```

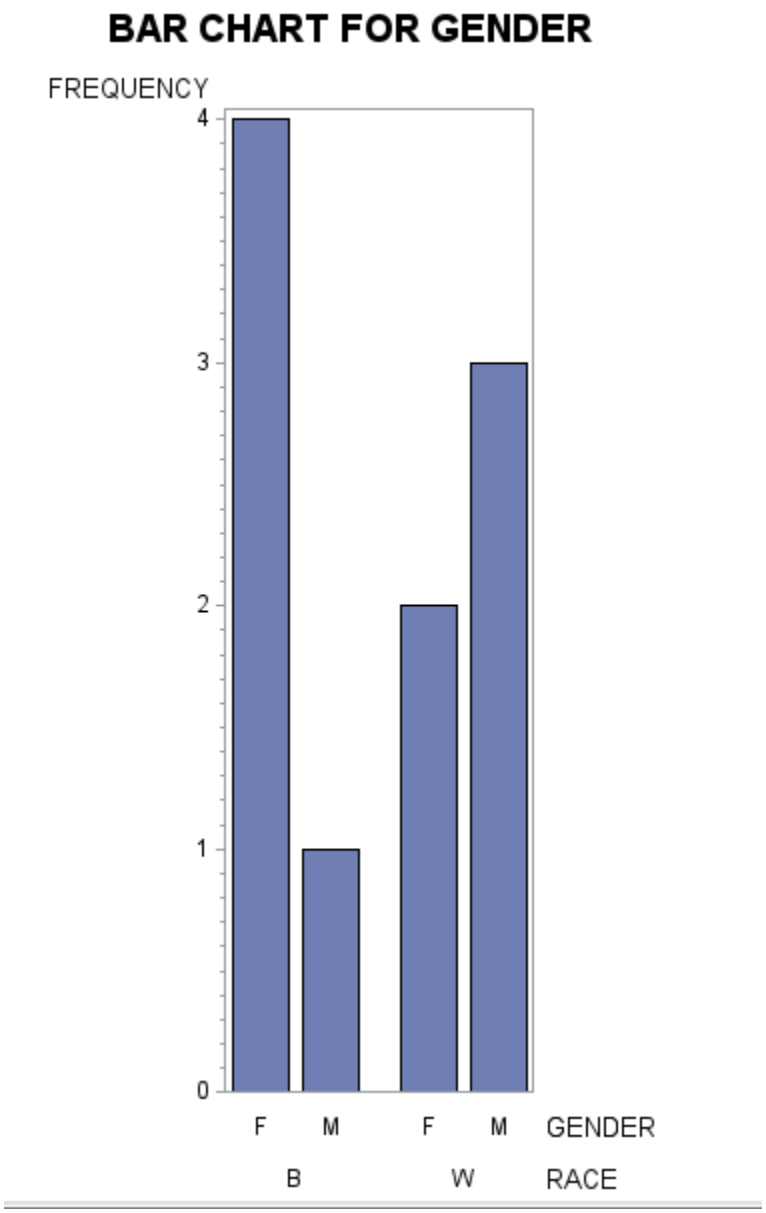

Distribution and Probability Plot for DBP
DBP
108
96
84
72
0
1
2
3
4
5
Count
DBP
110
100
90
80
70
-2
-1
0
1
2
Normal Quantiles


Distribution and Probability Plot for SBP
SBP
187.5
162.5
137.5
112.5
0
1
2
3
4
5
Count
SBP
180
160
140
120
100
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Normal Quantiles

Question 2.8

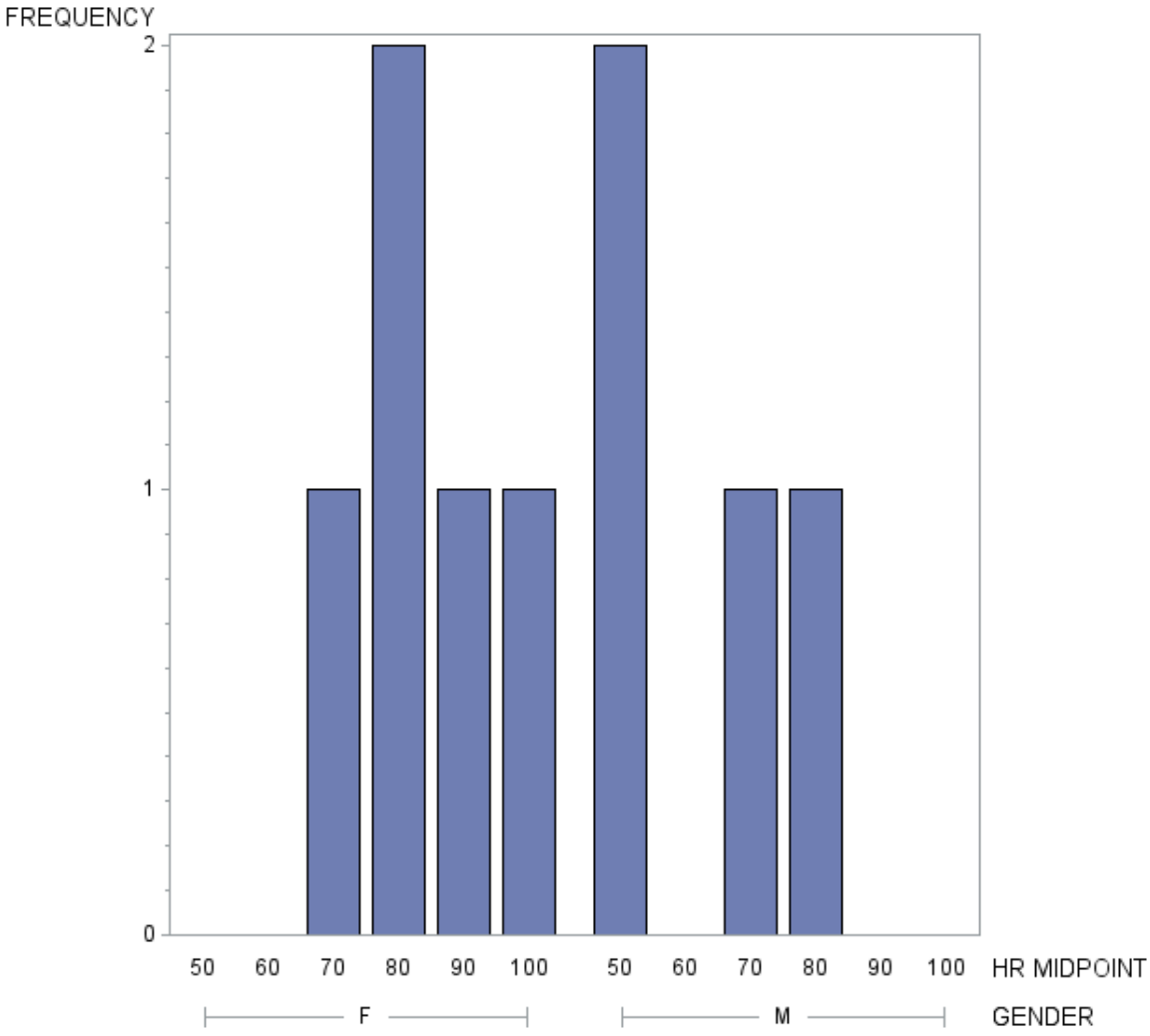
Solution:

```
PROC GCHART DATA=CLINIC;
      TITLE "BAR CHART FOR GENDER";
      VBAR GENDER/GROUP=RACE;
RUN;
```




```
PROC GCHART DATA=CLINIC;
      TITLE "BAR CHART FOR HEART RATE";
      VBAR HR/GROUP=GENDER MIDPOINTS=50 TO 100 BY 10;
RUN;
```

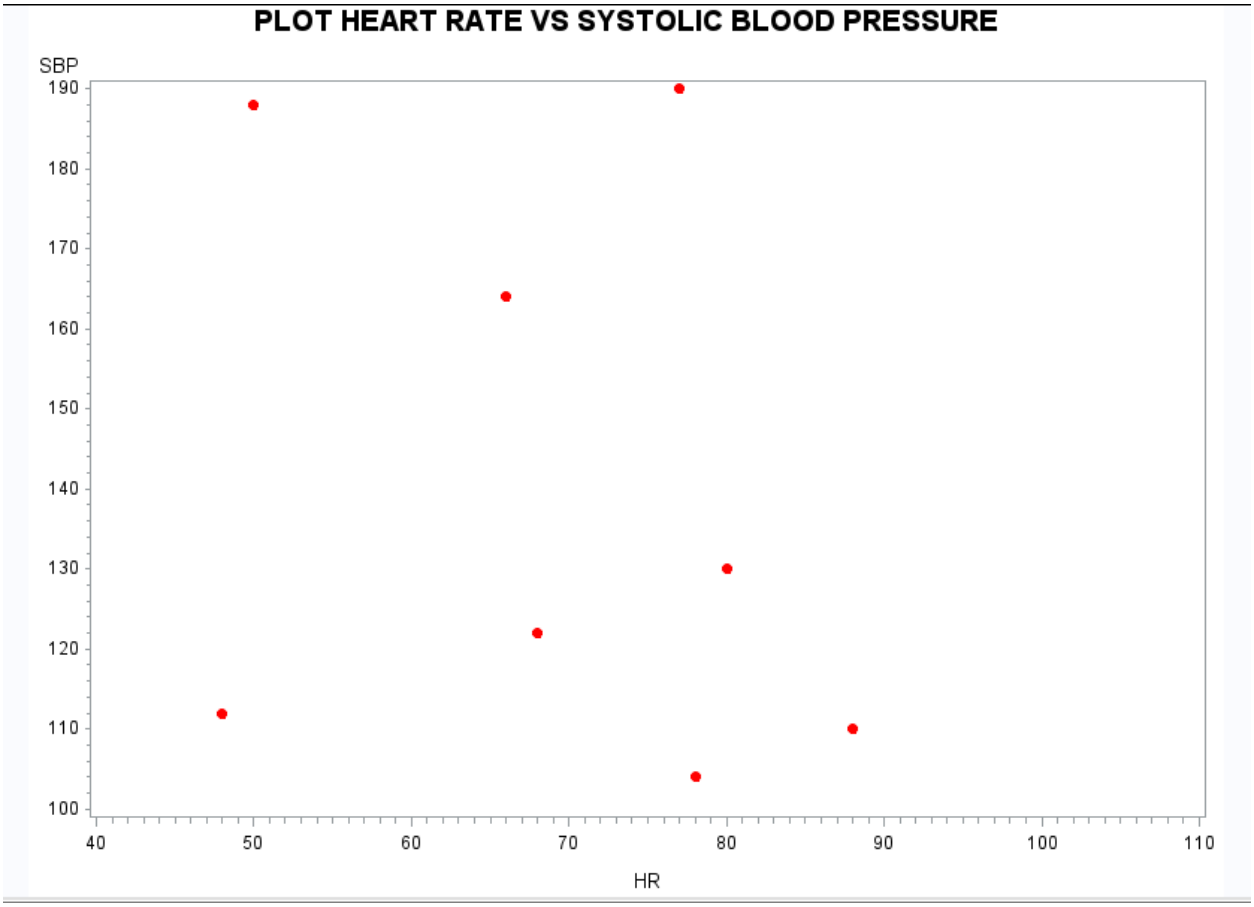
BAR CHART FOR HEART RATE



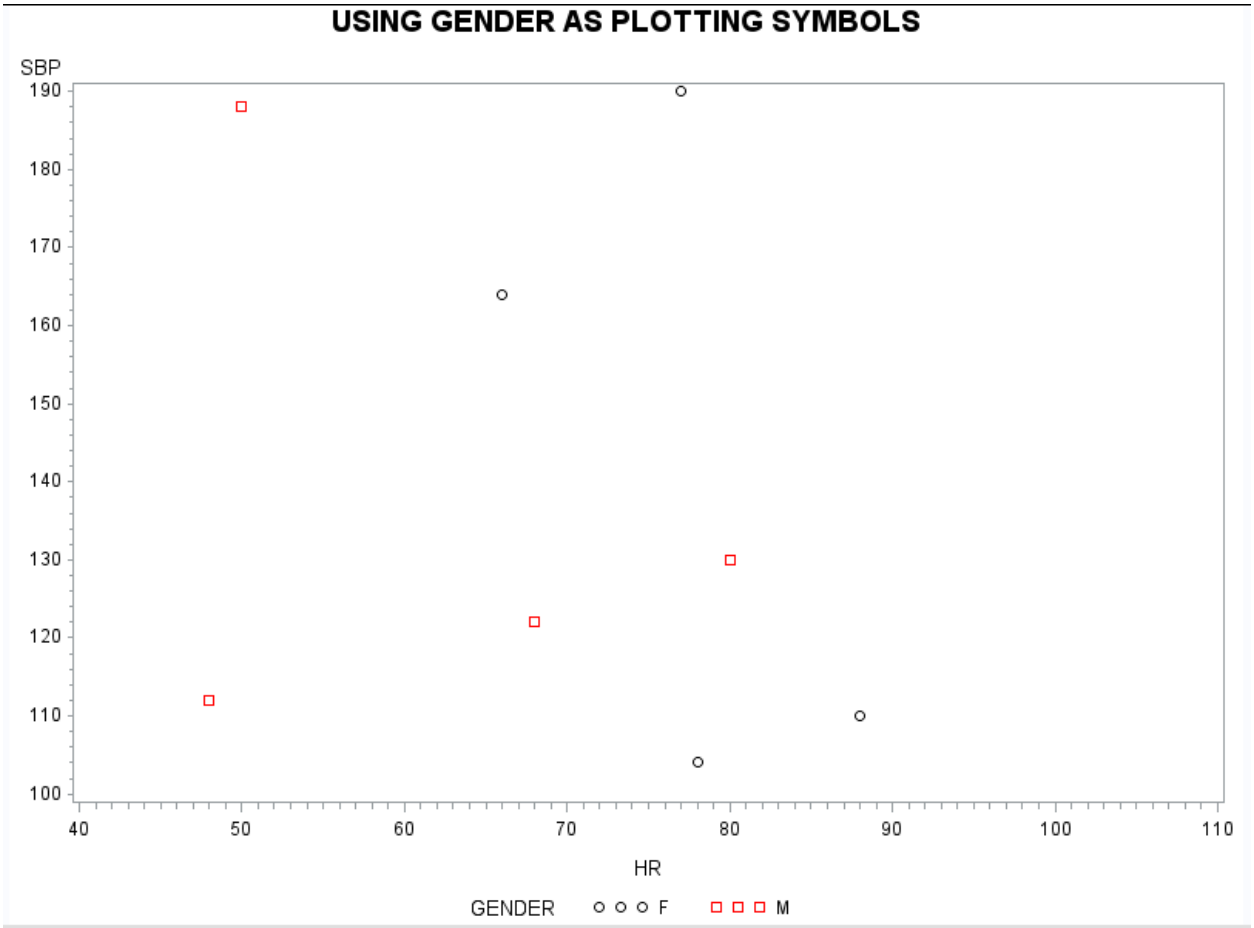
Question 2.10

Solution:

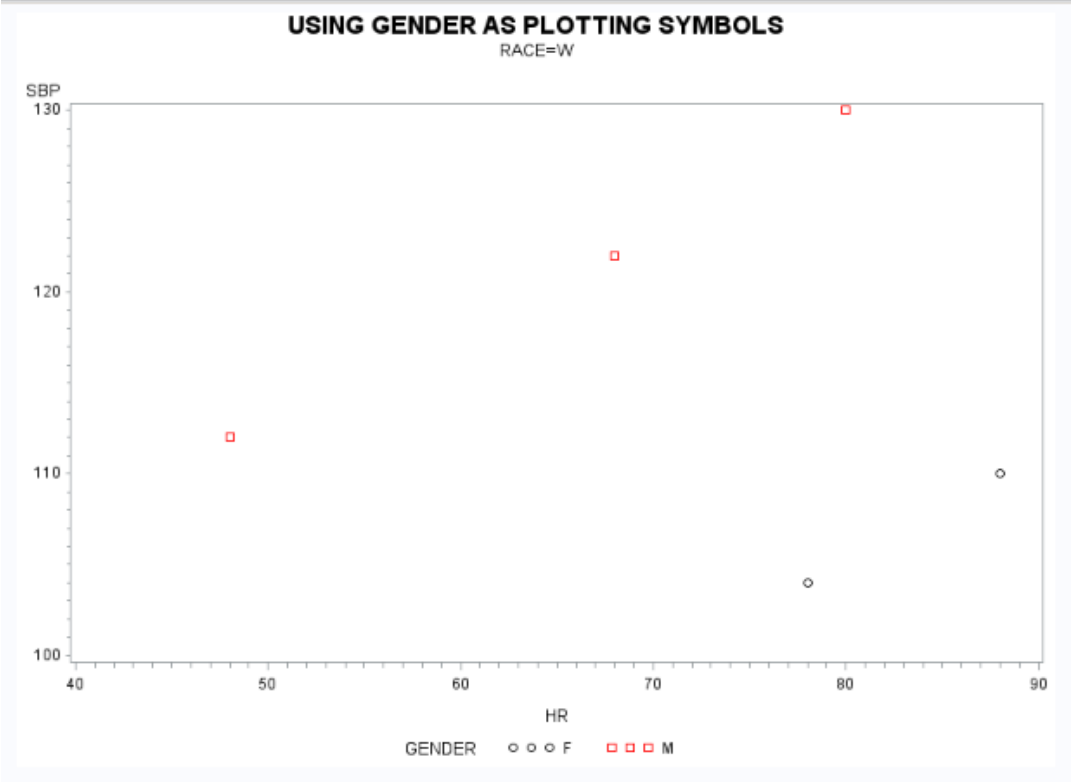
```
PROC GPLOT DATA=CLINIC;
      TITLE "PLOT HEART RATE VS SYSTOLIC BLOOD PRESSURE";
      SYMBOL VALUE= DOT COLOR =RED;
      PLOT SBP*HR;
RUN;
```

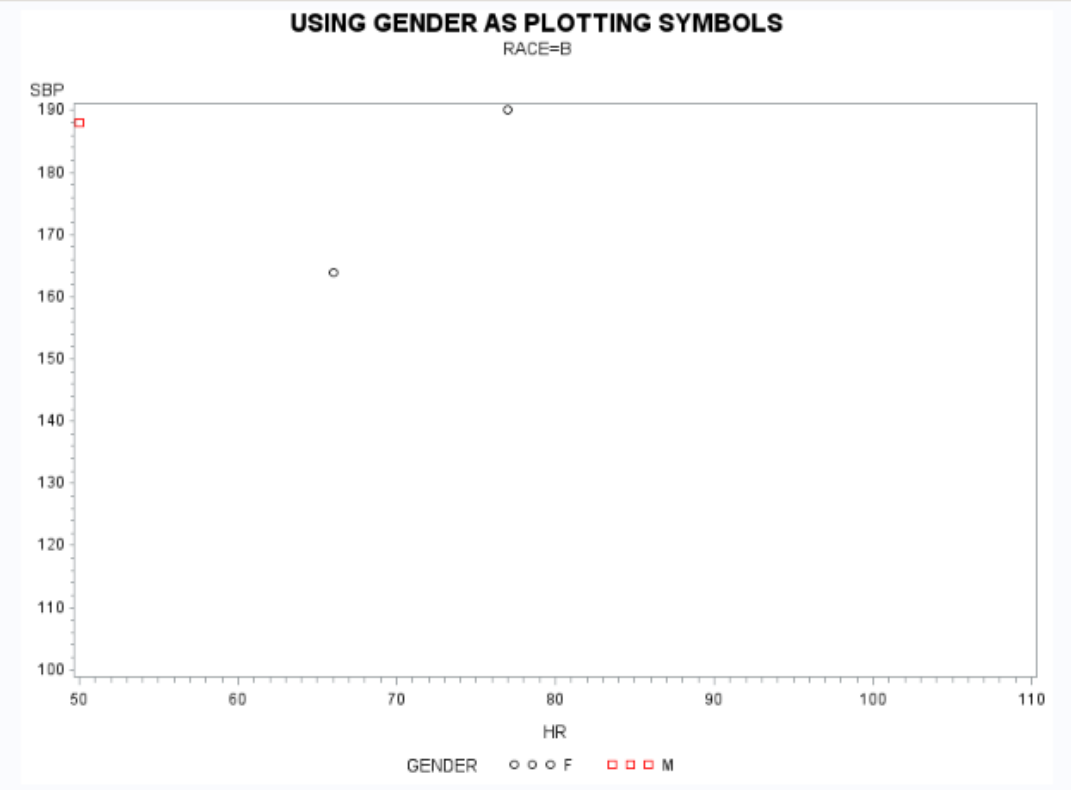


```
SYMBOL1 V= CIRCLE COLOR=BLACK;
SYMBOL2 V= SQUARE COLOR=RED;
PROC GPLOT DATA=CLINIC;
      TITLE "USING GENDER AS PLOTTING SYMBOLS";
      PLOT SBP*HR=GENDER;
RUN;
```



```
PROC SORT DATA=CLINIC;
      BY RACE;
RUN;
SYMBOL1 V= CIRCLE COLOR=BLACK;
SYMBOL2 V= SQUARE COLOR=RED;
PROC GPLOT DATA=CLINIC;
      BY RACE;
      TITLE "USING GENDER AS PLOTTING SYMBOLS";
      PLOT SBP*HR=GENDER;
RUN;
```



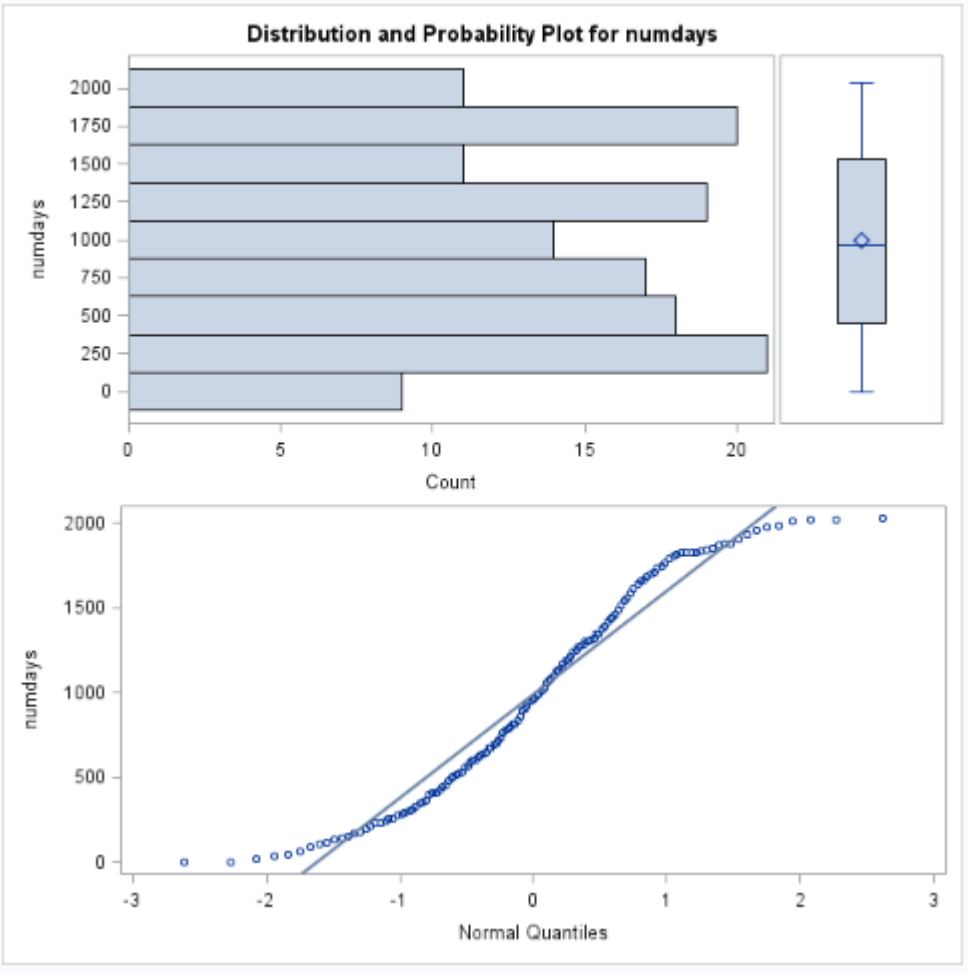


6. Refer to the CLINTON data. Write a SAS program which reads the data. Create a new variablewhich indicates the number of days elapsed between successive polls. Set the elapsed time forJanuary 24, 1993 (the day of the first poll) to be missing value. Then, use PROC UNIVARIATE toexamine the distribution of those elapsed times. Identify the 5 shortest and 5 longest elapsedtimes with dates, using an appropriate format for the date. What was happening to PresidentClinton when opinion polls were spaced very close together?

```
Libname tanay "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\";
DATA tanay.clinton;

infile "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\clinton.txt" DLM=" "
firstobs=3;
input Day $ Mo $ Year $ Approve Disapprove No_opinion;
mydate=CATT(Day,Mo,Year);
act_date=input(mydate,date9.);
format act_date mmddyy10.;
first_date=input('24Jan1993',date9.);
format first_date mmddyy10.;
numdays=intck('day',first_date,act_date);
RUN;

PROC PRINT DATA=tanay.clinton;
RUN;
PROC UNIVARIATE DATA=tanay.clinton NORMAL PLOT;
      TITLE "OPINION POLLS USING UNIVARIATE";
      VAR numdays;
RUN;
```



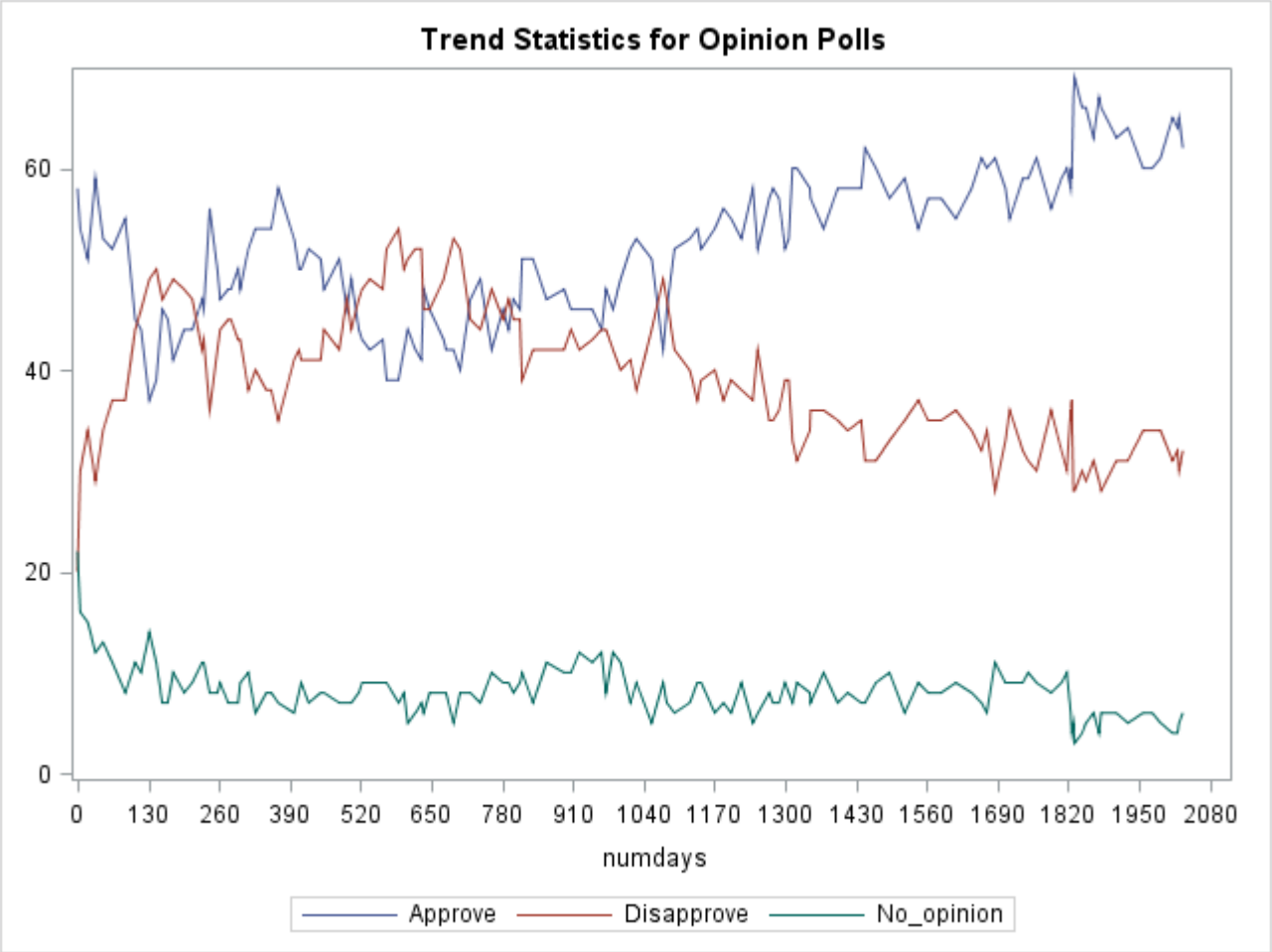
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	140	1990	5
5	139	2012	4
19	138	2021	3
33	137	2024	2
47	136	2031	1

What was happening to President Clinton:

TREND CHARTS to understand the opinion polls

```
TITLE "Trend Statistics for Opinion Polls";
   proc sgplot data=tanay.clinton;
      series x=numdays y=Approve  / lineattrs=(pattern=solid);
      series x=numdays y=Disapprove / lineattrs=(pattern=solid);
      series x=numdays y=No_opinion  / lineattrs=(pattern=solid);
      yaxis display=(nolabel);
        xaxis values=(0 to 2100 by 10);

   run;
```



When the opinions (Approve, Disapprove and Neutral) were plotted by the number of days and frequency it was observed that the Approval trend had a positive slope and it had an upward tendency. In contrast the Disapproval trend was falling having a negative slope.