Statistical Analysis of University of Illinois football team’s record
Name: Zixin Ouyang

The CONTENTS Procedure

Data Set Name	WORK.ILLINIFB16	**Observations**	125
Member Type	DATA	**Variables**	17
Engine	V9	**Indexes**	0
Created	03/10/2017 21:53:49	**Observation Length**	120
Last Modified	03/10/2017 21:53:49	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1090
Obs in First Data Page	125
Number of Data Set Repairs	0
Filename	/saswork/SAS_workE3950001C4E5_odaws04-prod-us/SAS_work66670001C4E5_odaws04-prod-us/illinifb16.sas7bdat
Release Created	9.0401M3
Host Created	Linux
Inode Number	13107226
Access Permission	rw-r--r--
Owner Name	zouyang70
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
11	AP_high	Num	3		Highest Rank
12	AP_post	Num	3		Final Rank
10	AP_pre	Num	3		Pre-season Rank
16	Bowl	Char	20		
17	BowlResult	Char	1		Bowl Result
14	Coach	Char	40		
3	Conf	Char	7		Conference
13	ConfTitle	Char	1		Conference Title
5	L	Num	3		Losses
1	Obs	Num	3		Observation
7	Pct	Num	5	5.3	Win Percentage
15	Record	Char	8		
9	SOS	Num	6	6.2	Schedule Strength
8	SRS	Num	6	6.2	Simple Rating
2	Season	Num	4		
6	T	Num	3		
4	W	Num	3		Wins

Description of the original data file:
I use the formatted input to read the raw data file into SAS. There are 125 observations and 17 variables in the data set. By looking at the raw data, I find that many values are missing and some values are invalid. When loading the data into SAS, I have created labels for some variables whose names are too simple and hard to understand.

The FREQ Procedure

Number of Variable Levels				
Variable	**Label**	**Levels**	**Missing Levels**	**Nonmissing Levels**
Obs	Observation	125	0	125
Season		125	0	125
Conf	Conference	3	0	3
W	Wins	11	0	11
L	Losses	13	1	12
T		4	0	4
Pct	Win Percentage	61	1	60
SRS	Simple Rating	123	0	123
SOS	Schedule Strength	119	0	119
AP_pre	Pre-season Rank	9	1	8
AP_high	Highest Rank	20	1	19
AP_post	Final Rank	12	1	11
ConfTitle	Conference Title	3	0	3
Coach		26	1	25
Record		76	1	75
Bowl		14	1	13
BowlResult	Bowl Result	3	1	2

According to the table above, we know that variable L, Pct, AP_pre, AP_high, AP_post, Coach, Record, Bowl and BowlResult have missing values. However, missing valued are allowed in variable AP_pre, AP_high, AP_post, Bowl and BowlResult. So we only need to clean the missing value in variable L, Pct, Coach, Record. Since Season has 125 levels, so each value of Season must be unique. Variable conference title has three nonmissing values, but it should only have two. So we need to check its values. The data set before cleaning is called illinifb16.

Coach	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Arthur Hall	6	4.84	6	4.84
Bill Cubit	1	0.81	7	5.65
Bob Blackman	6	4.84	13	10.48
E.K. Hall	2	1.61	15	12.10
Edgar Holt	2	1.61	17	13.71
Fred Lowenthal	1	0.81	18	14.52
Fred Smith	1	0.81	19	15.32
Gary Moeller	3	2.42	22	17.74
George Huff	5	4.03	27	21.77
George Woodruff	1	0.81	28	22.58
James Valek	4	3.23	32	25.81
John Mackovic	3	2.42	35	28.23
John Mackovic (6-5) Lou Tepper (0-1)	1	0.81	36	29.03
Justa Lindgren	1	0.81	37	29.84
Lou Tepper	5	4.03	42	33.87
Louis Vail	1	0.81	43	34.68
Lovie Smith	1	0.81	44	35.48
Mike White	8	6.45	52	41.94
Pete Elliott	7	5.65	59	47.58
Ray Eliot	18	14.52	77	62.10
Robert Zuppke	29	23.39	106	85.48
Ron Turner	8	6.45	114	91.94
Ron Zook	6	4.84	120	96.77
Ron Zook (6-6) Vic Koenning (1-0)	1	0.81	121	97.58
Tim Beckman	3	2.42	124	100.00
Frequency Missing = 1				

From the frequency report of Coach, we can know that there are not any typos in a coach's name. However, two records have two coaches in them.

Conference Title				
ConfTitle	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
N	111	88.80	111	88.80
Y	13	10.40	124	99.20
y	1	0.80	125	100.00

From the frequency report of Conference Title, we can find that there is a lowercase y in the variable. We should change it into uppercase.

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
SRS	Simple Rating	125	6.2811200	8.0841279	-12.9700000	24.0800000
SOS	Schedule Strength	125	5.3085600	4.7092842	-6.6000000	17.5500000
AP_pre	Pre-season Rank	13	12.4615385	7.2871506	3.0000000	22.0000000
AP_high	Highest Rank	32	10.4687500	6.8012540	2.0000000	25.0000000
AP_post	Final Rank	13	12.3846154	7.1477090	3.0000000	25.0000000

I use the means procedure to check whether there are extreme values in the data set. The values of SRS, SOS are in the normal range. And values of AP_pre, AP_high, and Ap_post are between 1 and 25, which are valid.

```
NOTE: Invalid data for L in line 90 19-19.
 RULE:     ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+-
---8----+----9----+----0
 90        90,1927,Western,7,O,1,0.938,15.77,3.27,,,,Y,Robert Zuppke,(7-0-1) 65
 Obs=90 Season=1927 Conf=Western W=7 L=. T=1 Pct=0.938 SRS=15.77 SOS=3.27 AP_pre=.
AP_high=. AP_post=. ConfTitle=Y
 Coach=Robert Zuppke Record=(7-0-1) Bowl=  BowlResult=  _ERROR_=1 _N_=90
 NOTE: 125 records were read from the infile
'~/my_courses/dunger_sas/midterm/illinifb16.dat'.
       The minimum record length was 46.
       The maximum record length was 98.
 NOTE: The data set WORK.ILLINIFB16 has 125 observations and 17 variables.
```

By reading the log of loading raw data file, we find that there is invalid data L in line 90 and this is because letter O is put in the data set instead of numeric value 0.

Also, I use where clause to find missing values which need cleaning.

Obs	Observation	Season	Wins	Losses	T	Win Percentage	Coach	Record
6	6	2011	7	6	0	0.538	Ron Zook (6-6) Vic Koenning (1-0)	
21	21	1996	2	9	0	.	Lou Tepper	(2-9)
26	26	1991	6	6	0	0.500	John Mackovic (6-5) Lou Tepper (0-1)	
113	113	1904	9	2	1	0.792		

According to the table above, We know that the missing values in observation 6 and 26 are due to two coach in one record. The missing value of percentage can be calculated from values in W, L and T. Regarding to the missing value in observation 113, I find the coach name and record online.

Obs	Wins	Losses	T	Win Percentage
2	5	7	0	0.417
3	6	7	0	0.462
4	4	8	0	0.333
5	2	10	0	0.167
6	7	6	0	0.538
7	7	6	0	0.538
9	5	7	0	0.417
10	9	4	0	0.692
11	2	10	0	0.167
12	2	9	0	0.182
13	3	8	0	0.273
14	1	11	0	0.083
15	5	7	0	0.417
16	10	2	0	0.833
17	5	6	0	0.455
18	8	4	0	0.667
19	3	8	0	0.273
21	2	9	0	.
22	5	5	1	0.500
23	7	5	0	0.583
24	5	6	0	0.455
25	6	5	1	0.542
27	8	4	0	0.667
28	10	2	0	0.833
29	6	5	1	0.542
30	3	7	1	0.318
31	4	7	0	0.364
32	6	5	1	0.542
33	7	4	0	0.636
34	10	2	0	0.833
35	7	5	0	0.583
36	7	4	0	0.636
37	3	7	1	0.418

Obs	Wins	Losses	T	Win Percentage
38	2	8	1	0.227
39	1	8	2	0.182
40	3	8	0	0.273
41	5	6	0	0.455
42	5	6	0	0.455
43	6	4	1	0.591
44	5	6	0	0.455
45	3	8	0	0.273
46	5	6	0	0.455
47	3	7	0	0.300
49	1	9	0	0.100
50	4	6	0	0.400
51	4	6	0	0.400
52	6	4	0	0.600
53	6	3	0	0.667
54	8	1	1	0.850
55	2	7	0	0.222
57	5	4	0	0.566
58	5	3	1	0.611
59	4	5	0	0.444
60	4	5	0	0.444
61	2	5	2	0.333
62	5	3	1	0.611
63	1	8	0	0.111
64	7	1	1	0.833
65	4	5	0	0.444
66	9	0	1	0.950
67	7	2	0	0.778
68	3	4	2	0.444
69	3	6	0	0.333
70	5	3	1	0.611
71	8	2	0	0.800
72	2	6	1	0.278
73	5	4	1	0.550
74	3	7	0	0.300
75	6	4	0	0.600
78	3	4	1	0.438
80	3	3	2	0.500

Obs	Wins	Losses	T	Win Percentage
81	4	3	1	0.563
85	5	4	0	0.556
88	6	1	1	0.813
90	7	.	1	0.938
93	6	1	1	0.813
95	2	5	0	0.286
96	3	4	0	0.429
97	5	2	0	0.714
98	6	1	0	0.857
99	5	2	0	0.714
100	5	2	1	0.688
101	3	3	1	0.500
102	5	0	2	0.857
104	4	2	1	0.643
105	3	3	1	0.500
106	6	2	1	0.643
108	5	2	0	0.714
109	5	1	1	0.786
110	3	2	0	0.600
111	1	3	1	0.300
112	5	4	0	0.556
113	9	2	1	0.792
114	8	6	0	0.571
115	10	2	1	0.808
116	8	2	0	0.800
117	7	3	2	0.667
118	3	5	1	0.389
119	4	5	0	0.444
121	4	2	1	0.643
122	4	2	1	0.643
123	4	3	0	0.571
124	3	2	3	0.563
125	7	4	1	0.625

I find that many values of W, L, T, and Pct do not coincide correctly , and I cleaned them later.

The data set after cleaning is called illinifb16_zouyang7. I have cleaned the missing values and invalid values in variable L, Pct, Coach, Record and ConfTitle. I reset the record value by concatenating W, L, T to match these entries directly. And I use function to set the value of winning percentage. In addition, I have created formats for rank, conference title, bowl, and bowl result to better understand values in these variables.

Number of Variable Levels

Variable	Label	Levels	Missing Levels	Nonmissing Levels
Obs	Observation	125	0	125
Season		125	0	125
Conf	Conference	3	0	3
W	Wins	11	0	11
L	Losses	12	0	12
T		4	0	4
Pct	Win Percentage	44	0	44
SRS	Simple Rating	123	0	123
SOS	Schedule Strength	119	0	119
AP_pre	Pre-season Rank	9	1	8
AP_high	Highest Rank	20	1	19
AP_post	Final Rank	12	1	11
ConfTitle	Conference Title	2	0	2
Coach		23	0	23
Record		79	0	79
Bowl		14	1	13
BowlResult	Bowl Result	3	1	2

The data set after cleaning is called illinifb16_zouyang7. I use the same freq procedure to verify that the data set is cleaned. There are no more missing values in variables L, Pct, Coach and Record. Also, the values of W, L, T, and Pct coincide correctly now.

	Wins
	Sum
Coach	
Arthur Hall	38.00
Bill Cubit	5.00
Bob Blackman	29.00
E.K. Hall	10.00
Edgar Holt	18.00
Fred Lowenthal	5.00
Fred Smith	7.00
Gary Moeller	6.00
George Huff	21.00
George Woodruff	8.00
James Valek	8.00
John Mackovic	30.00
Justa Lindgren	1.00
Lou Tepper	25.00
Louis Vail	4.00
Lovie Smith	3.00
Mike White	47.00
Pete Elliott	31.00
Ray Eliot	83.00
Robert Zuppke	131.00
Ron Turner	35.00
Ron Zook	35.00
Tim Beckman	12.00

According to the table above, we know that Robert Zuppke had the most wins in his career with the University of Illinois football team, and his wins are 131.

Obs	Season	Highest Rank
1	1964	2
2	1963	2
3	1952	2
4	1951	2
5	1953	3
6	1983	4
7	1960	4
8	1990	5
9	1954	5
10	1947	5
11	1946	5
12	1942	5
13	1950	6
14	2001	7
15	1989	8
16	1944	9
17	1985	11
18	1959	12
19	2007	13
20	1991	13
21	1956	13
22	1976	14
23	1974	14
24	1982	15
25	1957	15
26	2011	16
27	1955	16
28	2000	19
29	2008	20
30	1994	21
31	1999	24
32	1995	25

I have sorted the data by AP_high, and put the result in the data set called highestrank. According to the table above, we can find that Seasons 1964, 1963, 1952, and 1951 saw the football team with their highest ranking for the university across all seasons, and the highest ranking is 2.

The FREQ Procedure

Conference Title				
ConfTitle	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Lose	111	88.80	111	88.80
Win	14	11.20	125	100.00

By generating a frequency report of conference title, we can see that the number of times that Illinois won its conference title is 14.

	Wins
	Sum
Decade	
1890s	35.00
1900s	61.00
1910s	51.00
1920s	55.00
1930s	38.00
1940s	38.00
1950s	48.00
1960s	36.00
1970s	38.00
1980s	63.00
1990s	50.00
2000s	45.00
2010s	34.00

In order to identify which decade had the most wins in a decade, I create a data set called decadewins that contains the variable Decade and Wins. According to the result of tabulate procedure, 1980s had the most wins, which was 63.