STAT 506 Final Project

Zeru Zhou

Data Description:

This is a dataset that includes all tennis champion information in the history of tennis. All players in the final round of every match starting from 1968 to 2019 are included in this csv sheet. Information includes their height, weight, match score, hand preference, seed, citizenship, match information, etc. I obtained this dataset from Datahub.

Topics that I covered in this project:

1*, 3, 5*, 6, 7*, 8, 12*, 15*, 18, 22

Topic 1*

Code:

```
proc import datafile = "E:\SAS\PG2\data/ATP.csv" dbms = csv out = ATP; /* 1a */
run;

proc contents data = ATP varnum;    /* 1b */
run;
```

Output from Proc Contents:

The SAS System

The CONTENTS Procedure

Data Set Name	WORK.ATP	Observations	169690
Member Type	DATA	Variables	49
Engine	V9	Indexes	0
Created	04/28/2022 14:40:06	Observation Length	224
Last Modified	04/28/2022 14:40:06	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	582
First Data Page	1
Max Obs per Page	292
Obs in First Data Page	264
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\17654\AppData\Local\Temp\SAS Temporary Files_TD35484_DESKTOP-OHLK8RL_\atp.sas7bdat
Release Created	9.0401M7
Host Created	X64_10PRO
Owner Name	DESKTOP-OHLK8RL\17654
File Size	36MB
File Size (bytes)	38207488

Variables in Creation Order					
#	Variable	Type	Len	Format	Informat
1	best_of	Num	8	BEST12.	BEST32.
2	draw_size	Num	8	BEST12.	BEST32.
3	l_1stIn	Char	1	$1.	$1.
4	l_1stWon	Char	1	$1.	$1.
5	l_2ndWon	Char	1	$1.	$1.
6	l_SvGms	Char	1	$1.	$1.
7	l_ace	Char	1	$1.	$1.
8	l_bpFaced	Char	1	$1.	$1.
9	l_bpSaved	Char	1	$1.	$1.
10	l_df	Char	1	$1.	$1.
11	l_svpt	Char	1	$1.	$1.
12	loser_age	Num	8	BEST12.	BEST32.
13	loser_entry	Char	2	$2.	$2.
14	loser_hand	Char	1	$1.	$1.
15	loser_ht	Num	8	BEST12.	BEST32.
16	loser_id	Num	8	BEST12.	BEST32.
17	loser_ioc	Char	3	$3.	$3.
18	loser_name	Char	18	$18.	$18.
19	loser_rank	Char	1	$1.	$1.
20	loser_rank_points	Char	1	$1.	$1.
21	loser_seed	Num	8	BEST12.	BEST32.
22	match_num	Num	8	BEST12.	BEST32.
23	minutes	Char	1	$1.	$1.
24	round	Char	3	$3.	$3.
25	score	Char	19	$19.	$19.
26	surface	Char	5	$5.	$5.
27	tourney_date	Num	8	BEST12.	BEST32.
28	tourney_id	Char	8	$8.	$8.
29	tourney_level	Char	1	$1.	$1.
30	tourney_name	Char	16	$16.	$16.
31	w_1stIn	Char	1	$1.	$1.
32	w_1stWon	Char	1	$1.	$1.
33	w_2ndWon	Char	1	$1.	$1.
34	w_SvGms	Char	1	$1.	$1.
35	w_ace	Char	1	$1.	$1.
36	w_bpFaced	Char	1	$1.	$1.
37	w_bpSaved	Char	1	$1.	$1.
38	w_df	Char	1	$1.	$1.
39	w_svpt	Char	1	$1.	$1.
40	winner_age	Num	8	BEST12.	BEST32.
41	winner_entry	Char	1	$1.	$1.
42	winner_hand	Char	1	$1.	$1.
43	winner_ht	Num	8	BEST12.	BEST32.
44	winner_id	Num	8	BEST12.	BEST32.

45	winner_ioc	Char	3	$3.	$3.
46	winner_name	Char	20	$20.	$20.
47	winner_rank	Char	1	$1.	$1.
48	winner_rank_points	Char	1	$1.	$1.
49	winner_seed	Num	8	BEST12.	BEST32.

Topic 3:

Code:

```
proc print data = ATP (obs = 10); /* 3a 3c */
var round score surface match_num winner_name; /* 3b */
where winner_name like "%Federer"; /* 3d */
format round $4. match_num 3.; /* 3e */
run;
```

Output:

The SAS System

Obs	round	score	surface	match_num	winner_name
106429	R32	6-2 6-2	Hard	14	Roger Federer
106438	R16	6-1 7-6(5)	Hard	23	Roger Federer
107183	R32	7-6(1) 3-6 6-3	Hard	1	Roger Federer
107199	R16	6-7(6) 7-6(5) 7-6(5	Hard	17	Roger Federer
107411	R32	6-7(4) 7-5 7-6(3)	Carpe	14	Roger Federer
107420	R16	6-4 7-5	Carpe	23	Roger Federer
107796	RR	6-4 6-7(3) 6-3 6-4	Carpe	2	Roger Federer
109720	R32	6-4 6-3	Hard	9	Roger Federer
109855	R32	7-6(4) 6-1	Hard	3	Roger Federer
109893	R32	6-2 3-6 6-4	Carpe	10	Roger Federer

Topic 5*:

Code:

```
data ATP_new;
set ATP; /* 5a */
length older_better $ 3; /* 5e */
keep loser_name loser_age match_num winner_age winner_name winner_ht
loser_ht score set_num older_better; /* 5c */
where best_of = 5 & surface = "Grass" & winner_age is not null
& loser_age is not null & winner_ht is not null & loser_ht is not null; /* 5b */
format winner_age loser_age 2.; /* 5d */
if countw(score," ") >= 3 then set_num = countw(score," "); /* 5f */
if winner_age > loser_age then older_better = "Yes";
else older_better = "No";
if set_num in (3,4,5);
run;

proc print data = ATP_new (obs = 5); run;
```

Log information:

```
NOTE: There were 6576 observations read from the data set WORK.ATP.
      WHERE (best_of=5) and (surface='Grass') and (winner_age is not null) and (loser_age is not
      null) and (winner_ht is not null) and (loser_ht is not null);
NOTE: The data set WORK.ATP_NEW has 6576 observations and 10 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.00 seconds
```

Proc print output (test only):

The SAS System

Obs	loser_age	loser_ht	loser_name	match_num	score	winner_age	winner_ht	winner_name	older_better	set_num
1	19	178	Manuel Orantes	56	6-1 4-6 6-3 4-6 6-1	31	173	Barry Phillips Moore	Yes	5
2	22	180	Tom Gorman	3	8-6 6-4 6-3	45	185	E Victor Seixas	Yes	3
3	22	188	Sherwood Stewart	10	6-3 6-4 8-6	28	175	Jim Mcmanus	Yes	3
4	20	175	Tom Leonard	15	6-4 6-4 12-10	24	180	Patricio Cornejo	Yes	3
5	33	190	Barry Mackay	25	8-6 4-6 7-5 6-3	24	185	Jim Osborne	No	4

Topic 6:

Code:

```
data ATP_1;
set ATP;
length local_winner $ 5;
length Right_hand $ 5;
if winner_hand = 'R' then Right_hand = 'True'; /* 6a */
else Right_hand = 'False';
if tourney_name = 'Australian Chps.' & winner_ioc = 'AUS' then local_winner = 'True';/* 6b */
else if tourney_name = 'Australian Chps.'& winner_ioc ne 'AUS' then local_winner = 'False';
if winner_age < 25 then do; /* 6c */
winner_group = 'young age';
winner_future = 'bright';
end;
```

```
proc print data = ATP_1 (obs = 10);
var winner_group winner_future local_winner Right_hand;
where tourney_name = 'Australian Chps.' & winner_age < 25;
run;
```

Log output:

```
NOTE: There were 169690 observations read from the data set WORK.ATP.
NOTE: The data set WORK.ATP_1 has 169690 observations and 54 variables.
NOTE: DATA statement used (Total process time):
      real time           0.04 seconds
      cpu time            0.04 seconds
```

Proc print output:

The SAS System

Obs	winner_group	winner_future	local_winner	Right_hand
1	young age	bright	True	True
3	young age	bright	True	True
4	young age	bright	True	True
6	young age	bright	True	True
7	young age	bright	False	True
9	young age	bright	True	True
11	young age	bright	False	True
12	young age	bright	True	True
14	young age	bright	True	True
15	young age	bright	False	False

Topic 7*:

Code:

```
%let A = ATP information; /* 7d */
%let B = player and set information; /* 7d */
title "&A";/* 7a */
title2 "&B";/* 7b */
footnote "ATP players with age and set number";/* 7c */
proc print data = ATP_new (obs =10) label; /* 7e */
var set_num older_better;
label set_num = "Number of sets player win the game" older_better = "Older player win the game";
run;
title; /* 7f */
footnote; /* 7f */
```

Output:

ATP information
player and set information

Obs	Number of sets player win the game	Older player win the game
1	5	Yes
2	3	Yes
3	3	Yes
4	3	Yes
5	4	No
6	3	No
7	4	No
8	4	No
9	3	No
10	3	No

ATP players with age and set number

Topic 8:

Code:

```
proc freq data = ATP_new order = freq; /* 8a */
tables winner_age;
run;
proc freq data = ATP_new order = freq; /* 8b */
tables winner_age*set_num / nocol norow nopercent;
format winner_age 2.;
run;
```

Output for table 1:

The FREQ Procedure

winner_age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
26	651	9.97	651	9.97
24	645	9.88	1296	19.85
25	627	9.60	1923	29.46
23	620	9.50	2543	38.96
27	551	8.44	3094	47.40
28	531	8.13	3625	55.53
22	516	7.90	4141	63.43
29	432	6.62	4573	70.05
21	373	5.71	4946	75.77
30	352	5.39	5298	81.16
31	258	3.95	5556	85.11
20	227	3.48	5783	88.59
32	194	2.97	5977	91.56
19	126	1.93	6103	93.49
33	117	1.79	6220	95.28
34	85	1.30	6305	96.58
35	57	0.87	6362	97.46
18	50	0.77	6412	98.22
36	41	0.63	6453	98.85
17	22	0.34	6475	99.19
37	14	0.21	6489	99.40
40	14	0.21	6503	99.62
39	6	0.09	6509	99.71
41	6	0.09	6515	99.80
38	4	0.06	6519	99.86
42	3	0.05	6522	99.91
43	2	0.03	6524	99.94
44	2	0.03	6526	99.97
16	1	0.02	6527	99.98
45	1	0.02	6528	100.00

Output for table2:

The FREQ Procedure

Frequency	Table of winner_age by set_num			
	set_num			
winner_age	3	4	5	Total
26	330	236	85	651
24	301	255	89	645
25	308	242	77	627
23	309	241	70	620
27	257	228	66	551
28	248	214	69	531
22	223	208	85	516
29	191	178	63	432
21	186	141	46	373
30	175	138	39	352
31	125	101	32	258
20	105	87	35	227
32	93	80	21	194
19	54	49	23	126
33	60	44	13	117
34	42	32	11	85
35	31	20	6	57
18	22	23	5	50
36	27	11	3	41
17	7	12	3	22
37	7	4	3	14
40	7	6	1	14
39	5	1	0	6
41	5	1	0	6
38	2	2	0	4
42	1	1	1	3
43	1	0	1	2
44	1	1	0	2
16	0	1	0	1
45	1	0	0	1
Total	3124	2557	847	6528

Topic 12*:

Code:

```
proc sql;
  create table ATP_sql as /* 12a */
  select loser_name, winner_name, score, older_better, set_num
  from ATP_new
  where older_better = "Yes" /* 12b */
  order by set_num, winner_name desc; /* 12c */
  quit;

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

Log Notes:

```
NOTE: Table WORK.ATP_SQL created, with 3308 rows and 5 columns.

341!                                                /* 12c */
342  quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time               0.01 seconds
      cpu time                0.01 seconds
```

Output:

The SAS System

Obs	loser_name	winner_name	score	older_better	set_num
1	Brydan Klein	Yuichi Sugita	7-6(5) 6-3 6-0	Yes	3
2	Felix Mantilla	Younes El Aynaoui	7-6(5) 6-3 6-4	Yes	3
3	Andreas Vinciguerr	Younes El Aynaoui	6-1 6-4 7-6(4)	Yes	3
4	Mark Hilton	Younes El Aynaoui	6-3 6-2 6-4	Yes	3
5	Dominik Hrbaty	Yevgeny Kafelnikov	7-6 6-2(4) 6-2(7)	Yes	3
6	Roger Federer	Yevgeny Kafelnikov	7-5 7-5 7-6(6)	Yes	3
7	Juan Antonio Marin	Yevgeny Kafelnikov	6-4 6-2 6-0	Yes	3
8	Michal Przysiezny	Yen Hsun Lu	6-4 7-6(7) 6-3	Yes	3
9	Viktor Troicki	Yen Hsun Lu	7-6(5) 6-4 6-4	Yes	3
10	James Ward	Yen Hsun Lu	6-7(4) 6-4 7-6(11)	Yes	3

Topic 15*:

Code:

```
proc sort data = ATP out = ATP1; /* 15d */
by tourney_id;
run;

data ATP_sum;
set ATP1;
by tourney_id;
keep matches accumulated_matches ;
retain matches; /* 15a */
if first.tourney_id = 1 then do; /* 15c */
matches = 0;
accumulated_matches = 0;
end;
matches = matches+1;
accumulated_matches+matches; /* 15b */
run;
proc print data = ATP_sum (obs = 10); run;
```

Log note:

```
NOTE: There were 169690 observations read from the data set WORK.ATP.
NOTE: The data set WORK.ATP1 has 169690 observations and 49 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.06 seconds
      cpu time            0.06 seconds


NOTE: There were 169690 observations read from the data set WORK.ATP1.
NOTE: The data set WORK.ATP_SUM has 169690 observations and 2 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.03 seconds
```

Output:

The SAS System

Obs	matches	accumulated_matches
1	1	1
2	2	3
3	3	6
4	4	10
5	5	15
6	6	21
7	7	28
8	1	1
9	2	3
10	3	6

Topic 18:

Code:

```
data ATP_mod;
set ATP_new;
Lastname = scan(winner_name,-1); /* 18c */
First_set = substr(score,1,3); /* 18a */
winner_name = propcase(winner_name); /* 18d */
loser_name = lowcase(loser_name); /* 18e */
Namelength = length(winner_name); /* 18b */
if older_better = 'Yes' then Age_name = cats(winner_name, ',',
round(winner_age,1),',', 'older wins'); /* 18f */
else Age_name = cats(winner_name, ',', round(winner_age,1),
',', 'younger wins');
run;

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

Log note:

```
NOTE: There were 6528 observations read from the data set WORK.ATP_NEW.
NOTE: The data set WORK.ATP_MOD has 6528 observations and 14 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

Output:

The SAS System

Obs	loser_age	loser_ht	loser_name	match_num	score	winner_age	winner_ht	winner_name	older_better	set_num	Lastname	First_set	Namelength	Age_name
1	19	178	manuel orantes	56	6-1 4-6 6-3 4-6 6-1	31	173	Barry Phillips Moore	Yes	5	Moore	6-1	20	Barry Phillips Moore,31,older wins
2	22	180	tom gorman	3	8-6 6-4 6-3	45	185	E Victor Seixas	Yes	3	Seixas	8-6	15	E Victor Seixas,45,older wins
3	22	188	sherwood stewart	10	6-3 6-4 8-6	28	175	Jim Mcmanus	Yes	3	Mcmanus	6-3	11	Jim Mcmanus,28,older wins
4	20	175	tom leonard	15	6-4 6-4 12-10	24	180	Patricio Cornejo	Yes	3	Cornejo	6-4	16	Patricio Cornejo,24,older wins
5	33	190	barry mackay	25	8-6 4-6 7-5 6-3	24	185	Jim Osborne	No	4	Osborne	8-6	11	Jim Osborne,24,younger wins
6	24	175	peter burwash	27	6-1 6-1 6-3	21	180	Robert Lutz	No	3	Lutz	6-1	11	Robert Lutz,21,younger wins
7	31	185	andres gimeno	28	4-6 6-1 6-2 6-1	22	183	Raymond Moore	No	4	Moore	4-6	13	Raymond Moore,22,younger wins
8	35	188	robert perry	30	4-6 6-0 7-5 7-5	18	188	Dick Stockton	No	4	Stockton	4-6	13	Dick Stockton,18,younger wins
9	34	180	sammy giammalva	39	10-8 6-0 6-3	32	183	Roy Emerson	No	3	Emerson	10-	11	Roy Emerson,32,younger wins
10	28	175	jim mcmanus	45	6-1 6-4 14-12	25	188	Clark Graebner	No	3	Graebner	6-1	14	Clark Graebner,25,younger wins

Topic 22:

Code:

```
data ATP;
set ATP;
if countw(score," ") >= 3 then set_num = countw(score," ");
run;

libname pg2 base 'E:\SAS\PG2\data';
proc format library = pg2.hwformats; /* 22c */
value game 3 = 'Quick Game'  /* 22b */
           4 = 'Median Game'
           5 = 'Long Game';
value $field "Grass", "Hard" = "Fast"  /* 22a */
             "Carpet", "Clay" = "Slow";
run;

options fmtsearch = (pg2.hwformats);
proc freq data = ATP;  /* 22d */
tables surface set_num;
format surface $field. set_num game.;
run;
```

Proc freq output:

The SAS System

The FREQ Procedure

surface	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Carp	19861	11.88	19861	11.88
Slow	61368	36.71	81229	48.60
Fast	85842	51.36	167071	99.95
None	80	0.05	167151	100.00
Frequency Missing = 2539				

set_num	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Quick Game	64014	77.62	64014	77.62
Median Game	13707	16.62	77721	94.24
Long Game	4748	5.76	82469	100.00
Frequency Missing = 87221				