```
*STEP 0 ;
/*1. Program Name:Vivek235_HW13_Program.sas.
                                                                    */
/* Program Location: C:\Users\vigupta\OneDrive\Learning\DataScience\Statistics Texas A&M
University\657\Homework\Assignment12\Vivek235_HW13_Program.sas          */
/* Date Created: 4/13/17
                                                                    */
/* Author: Vivek Kumar Gupta
                                                                */
/* Purpose:This assignment focuses on the techniques taught in the last few lectures of the semester. You
should have all of the information you need by the end of Lecture 22. You will be using the
ncaam06 and ncaam04 datasets that were used in previous assignments.;               */
/*************************************************************************************************
*****************/

/*Set up appropriate libnames need for the program*/
libname srcdata 'C:\Users\vigupta\OneDrive\Learning\DataScience\Statistics Texas A&M
University\657\Homework\Assignment04\SourceData' access=readonly;
filename pdfdev 'C:\Users\vigupta\OneDrive\Learning\DataScience\Statistics Texas A&M
University\657\Homework\Assignment13\Vivek235_HW13_Output.pdf';

/*Setp up system level options*/
option msglevel = i nodate nonumber;

/*Setp up the ODS for the output*/
ods pdf file=pdfdev  bookmarkgen=no ;

/*STEP 1. Use PROC SQL to create a table with columns seed, school, region, player, ppg, and rpg from
ncaam06 with
only schools that have 5 or more players listed in the dataset.*/
proc sql ;
/*STEP 1a. SQL commands to create the dataset*/
/*STEP 1b. The Start value of your format will come from the Team column.*/
/*STEP 1c. Use seed_ as the label in your format.*/
/*STEP 1d. Give the new format a name of your choosing*/
create table TeamSeed as
select distinct team as Start , put(seed_, 2.) as Label, "$TeamSeed" as FmtName from srcdata.ncaam04
order by label;

/*STEP 1e. After you have created the new dataset, use SQL to insert a row at the end. This row will
have values of start='other' and label='NA' along with the name of your format*/
```

```
insert into TeamSeed(Start, Label,FmtName)
values("other","NA","$TeamSeed");
quit;

/*STEP 1f. Use the format procedure to create from this dataset a user defined format in the work
library.*/
proc format library=work
       cntlin=TeamSeed ;
       select FmtName;
run;

/*STEP 1g. Use the format procedure to write the contents of the new format out to your output
document.*/
proc format library=work fmtlib;
      select $TeamSeed;
run;

/*STEP 2. Create a picture format that can be applied to PPG column*/
proc format library=work  ;
picture ppg
15-high = '(09.9)' (prefix='High(')
7.7 - < 15 = '(9.9)' (prefix='Medium(')
low - <7.7 ='(9.9)' (prefix='Low(');

run;

/*STEP 2. Use a SAS procedure to place a copy of the ncaam06 dataset in the WORK library. All future
references to ncaam06 will be to the WORK copy.*/

title "Descriptor Portion of ncaam06";

proc datasets library=srcdata nolist mtype=data;
copy out = work noclone;
select ncaam06;
run;


proc sql;

/*STEP 4a Create a character column labeled PPG Rating. This column needs to be wide enough to
```

```
accept all values of the picture format label created above.*/

/*STEP 4b Create a character column with a length of 2 labeled 2004 Seeding.*/
/*STEP 4c. Change the length and format of the School column so that it is wide enough to fully
display Wisconsin Milwaukee.*/

alter table work.ncaam06
add ppgchar char(20) label "PPG Rating"
add seed_04 char(2) label "2004 Seeding"
modify School char(40);

/*STEP 4d Populate the PPG Rating column using the put function to create values by applying the
picture format to the PPG column.*/
/*STE 4e. Correct the following school names*/
/*STEP 4f. Populate the 2004 Seeding column using the put function to create values by applying
the format from step 1 to the school column.*/

update work.ncaam06
set ppgchar=put(ppg, ppg.),
School =
      case(School)
      when 'Indania' then 'Indiana'
      when 'Boston Coll' then 'Boston College'
      when 'George mason' then 'George Mason'
      when 'Oral Robt-16' then 'Oral Roberts'
      when 'Wisc. Milwaukee' then 'Wisconsin Milwaukee'
      else School
end;

update work.ncaam06
set seed_04 = put(school,$TeamSeed.);

/*STEP 5. Create a composite index for ncaam06 using the variables player, school, and region in that
order. */
create index comp_idx on work.ncaam06(player, school, region);

quit;

/*STEP 5 Print the descriptor portion of the ncaam06 data set after the index has been created.*/
proc datasets library=work nolist ;
contents data= ncaam06;
```

```
quit;

/*STEP6.Question and answers */
title '6a. IDXWHERE on Player';

/*Yes the index will be used here because the composite index comp_idx
since we are forcing SAS to use the index regardless of whether sequential scan of the table will be
optimal or not
with the option idxwhere=yes.

The index will be used since comp_idx as player as the primary key*/

proc print data = work.ncaam06 (idxwhere=yes) label;
var Player School Region Seed ppgchar seed_04;
where player in ('Steve Burtt', 'Jared Dudley', 'Stanley Burrell');
run;


title '6b. IDXWHERE on School';

/*This step produces an  error since we are forcing SAS to use an index but the only index set on the table
does not have school as
the primary key variable. By the rules of index usage, this is not allowed.
*/

proc print data = work.ncaam06 (idxwhere=yes) label;
var Player School Region Seed ppgchar seed_04;
where school='Texas';
run;


title '6c. IDXWHERE on School';

/* This step produces an error because we are fociing SAS to use an index that is not applicable to the
query.
*/
proc print data = work.ncaam06 (idxname=comp_idx) label;
var Player School Region Seed ppgchar seed_04;
where school='Texas';
run;
```

```
title '6d. IDXWHERE on Player or School';

/* This is a case of compound optimization that does not occur
since both the where clause conditions are not connected by an AND operator . All other requirements of
compoud
optimization are satisifed however.
*/

proc print data = work.ncaam06 (idxwhere=yes) label;
var Player School Region Seed ppgchar seed_04;
where player in ('Steve Burtt', 'Jared Dudley', 'Stanley Burrell') or school='Indiana';
run;


title '6e. IDXWHERE on Player and School';

/* This is a case of compound optimization of the index.
Both the where clause conditions are connected by an AND operator . Keys used are the first 2
keys of the composite index and substr function begins at the beinning
*/

proc print data = work.ncaam06 (idxwhere=yes) label;
var Player School Region Seed ppgchar seed_04;
where substr(player,1,1)='S' and
school in ('Duke', 'Oral Roberts', 'Iona', 'Boston College',
'Gonzaga');
run;


/*Close the device*/
ods pdf close;

/*House cleaning*/
title;
footnote;
option date number msglevel=n;
```