```
964  ********************************************************************************;
965  *************** 80-character banner for column width reference ***************;
966  * (set window width to banner width to calibrate line length to 80 characters *;
967  ********************************************************************************;
968
969  *
970  This file prepares the dataset described below for analysis.
971  [Dataset Name] Dropouts by Race & Gender
972  [Experimental Units] California public K-12 schools
973  [Number of Observations] 58,876
974  [Number of Features] 20
975  [Data Source] The file http://dq.cde.ca.gov/dataquest/dlfile/dlfile.aspx?cLevel=
976  School&cYear=2014-15&cCat=Dropouts&cPage=filesdropouts was
977  downloaded as a text file and imported to Excel as tab-delimited text, then edited
978  to produce file project1_datasetv4.csv by setting all numeric values to number
979  types and the column labels and gender codes to text.
980  [Data Dictionary] http://www.cde.ca.gov/ds/sd/sd/fsdropouts.asp
981  [Unique ID] The columns  CDS_CODE, Ethnic(ity), and Gender form a
982  composite key
983  ;
984
985  * setup environmental parameters;
986  %let inputDatasetURL =
987  https://github.com/stat6250/team-5_project1/blob/master/project1_datasetv4.csv?raw=true
988  ;
989
990
991  * load raw dropout dataset over the wire;
992  filename tempfile TEMP;
993  proc http
994      method="get"
995      url="&inputDatasetURL."
996      out=tempfile
997      ;
998  run;

NOTE: PROCEDURE HTTP used (Total process time):
      real time           0.92 seconds
      cpu time            0.09 seconds


999  proc import
1000     file=tempfile
1001     out=project1_raw
1002     dbms=csv;
1003 run;

1004   /*************************************************************
1005   *   PRODUCT:   SAS
1006   *   VERSION:   9.4
1007   *   CREATOR:   External File Interface
1008   *   DATE:      15APR17
1009   *   DESC:      Generated SAS Datastep Code
1010   *   TEMPLATE SOURCE:  (None Specified.)
1011   *************************************************************/
1012     data WORK.PROJECT1_RAW    ;
```

```
1013      %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
1014      infile TEMPFILE delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=2 ;
1015         informat CDS_CODE best32. ;
1016         informat ETHNIC best32. ;
1017         informat GENDER $1. ;
1018         informat E7 best32. ;
1019         informat E8 best32. ;
1020         informat E9 best32. ;
1021         informat E10 best32. ;
1022         informat E11 best32. ;
1023         informat E12 best32. ;
1024         informat EUS best32. ;
1025         informat ETOT best32. ;
1026         informat D7 best32. ;
1027         informat D8 best32. ;
1028         informat D9 best32. ;
1029         informat D10 best32. ;
1030         informat D11 best32. ;
1031         informat D12 best32. ;
1032         informat DUS best32. ;
1033         informat DTOT best32. ;
1034         informat YEAR best32. ;
1035         format CDS_CODE best12. ;
1036         format ETHNIC best12. ;
1037         format GENDER $1. ;
1038         format E7 best12. ;
1039         format E8 best12. ;
1040         format E9 best12. ;
1041         format E10 best12. ;
1042         format E11 best12. ;
1043         format E12 best12. ;
1044         format EUS best12. ;
1045         format ETOT best12. ;
1046         format D7 best12. ;
1047         format D8 best12. ;
1048         format D9 best12. ;
1049         format D10 best12. ;
1050         format D11 best12. ;
1051         format D12 best12. ;
1052         format DUS best12. ;
1053         format DTOT best12. ;
1054         format YEAR best12. ;
1055      input
1056                  CDS_CODE
1057                  ETHNIC
1058                  GENDER $
1059                  E7
1060                  E8
1061                  E9
1062                  E10
1063                  E11
1064                  E12
1065                  EUS
1066                  ETOT
1067                  D7
1068                  D8
```

```
1069                    D9
1070                    D10
1071                    D11
1072                    D12
1073                    DUS
1074                    DTOT
1075                    YEAR
1076        ;
1077        if _ERROR_ then call symputx('_EFIERR_',1);  /* set ERROR detection macro variable */
1078        run;

NOTE: The infile TEMPFILE is:
      Filename=C:\Users\kd6274\AppData\Local\Temp\SAS Temporary Files\_TD4044_SCIENCE-30_\#LN00095,
      RECFM=V,LRECL=32767,File Size (bytes)=342709,
      Last Modified=15Apr2017:21:32:05,
      Create Time=15Apr2017:21:32:04

NOTE: 58875 records were read from the infile TEMPFILE.
      The minimum record length was 54.
      The maximum record length was 75.
NOTE: The data set WORK.PROJECT1_RAW has 58875 observations and 20 variables.
NOTE: DATA statement used (Total process time):
      real time           0.09 seconds
      cpu time            0.07 seconds


58875 rows created in WORK.PROJECT1_RAW from TEMPFILE.



NOTE: WORK.PROJECT1_RAW data set was successfully created.
NOTE: The data set WORK.PROJECT1_RAW has 58875 observations and 20 variables.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           0.17 seconds
      cpu time            0.14 seconds


1079  filename tempfile clear;
NOTE: Fileref TEMPFILE has been deassigned.
1080
1081
1082  * check raw dropout dataset for duplicates with respect to its composite key;
1083  proc sort nodupkey data=project1_raw dupout=project1_raw_dups out=_null_;
1084      by CDS_CODE Ethnic Gender;
1085  run;

NOTE: There were 58875 observations read from the data set WORK.PROJECT1_RAW.
NOTE: 0 observations with duplicate key values were deleted.
NOTE: The data set WORK.PROJECT1_RAW_DUPS has 0 observations and 20 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.05 seconds
      cpu time            0.04 seconds


1086
1087
```

```
1088  * build analytic dataset from dropout dataset with the least number of columns and
1089  minimal cleaning/transformation needed to address research questions in
1090  corresponding data-analysis files;
1091  data project1_analytic_file;
1092      drop
1093          year
1094      ;
1095
1096      set project1_raw;
1097  run;

NOTE: There were 58875 observations read from the data set WORK.PROJECT1_RAW.
NOTE: The data set WORK.PROJECT1_ANALYTIC_FILE has 58875 observations and 19 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.01 seconds
```