```
/*
 * Program: ST662 - Assignment 1.sas
 * Author : Sean O'Riogain (18145426)
 * Date   : 13th February 2019
 */

/* Import the CSV file (/home/seanoriogain200/ST662/Toenail.csv) into a Permanent Library (ST662) */
proc import out=tn1
    datafile="/home/seanoriogain200/ST662/Toenail.xlsx"
    dbms=xlsx replace;
    getnames=yes;
run;

/* Append and populate a row number variable (OBS) in dataset TN1 */
data tn1;
    set tn1;
    Obs = _n_;
run;

/* Analyse all non-numeric variables (TREAT, GENDER) */
proc freq;
    tables treat gender / nocum nopercent;
run;

/* Display the observations that have invalid values (including missing ones) for the non-numeric variables */

/* TREAT: Analyse its values */

proc print data=tn1;
    where prxmatch('/^(0|1).*$/', treat) = 0;
run;

/* Display the anomalies identified above in context of some of the surrounding data */
proc print data=tn1;
    where obs >= 1040 and obs <= 1060;
run;

/* Run a SQL query to determine if the Treat value is the same for all records of a particular patient */
proc sql;
    select id, treat, count(*) as count
    from tn1
    group by id, treat;
quit;

/* GENDER: Analyse its values */

proc print data=tn1;
    where prxmatch('/^(Male|Female).*$/', gender) = 0;
run;

/* Display the anomalies identified above in context of some of the surrounding data */
proc print data=tn1;
    where (obs >= 1110 and obs <= 1130) or (obs >= 1610 and obs <= 1640);
run;

/* Analyse all numeric variables (ID, TIME, Y) */
proc univariate data=tn1 plots;
    var ID TIME Y;
run;

proc freq data=tn1;
    tables ID TIME Y / nocum nopercent;
run;

/* Display the observations that have invalid values (including missing ones) for the numeric variables */

/* ID: Analyse its values */

/* Run a SQL query to identify any gaps in id values */
proc sql;
    select distinct a.obs, a.id, b.id as lag_value
    from tn1 a left join tn1 b
    on a.obs = b.obs + 1
    where a.id <> lag_value and a.id <> lag_value + 1 and lag_value <> .;
quit;
```

```
/* Display the anomalies identified above in context of some of the surrounding data */
proc print data=tn1;
    where obs >= 1750 and obs <= 1770;
run;

/* TIME: Analyse its values */

proc print data=tn1;
    where prxmatch('/^( 0| 1| 2| 3| 6| 9|12).*$/', put(time, 2. -R)) = 0;
run;

/* Display the anomalies identified above in context of some of the surrounding data */
proc print data=tn1;
    where obs >= 320 and obs <= 340;
run;

/* Y: Analyse its values */

proc print data=tn1;
    where prxmatch('/^(0|1).*$/', put(y, 1. -L)) = 0;
run;

/* Display the anomalies identified above in context of some of the surrounding data */
proc print data=tn1;
    where (obs >= 400 and obs <= 430) or (obs >= 780 and obs <= 790);
run;

/* Let's deal with the anomalies identified previously appropriately in a copy of dataset TN1 (TN2) */

data tn2;
    set tn1;

    if obs = 1052 and treat = 'A' then treat = '0';

    if obs = 1121 and gender = 'A' then gender = 'Female';

    if id = 252 and gender = 'A' then gender = '';

    if obs = 1758 and id = 722 then id = 272;

    if obs = 329 and time = 13 then time = 12;

    if prxmatch('/^(0|1).*$/', put(y, 1. -L)) = 0 then y = .;
run;

/* Display the observations that are still outside the valid range for all variables */
proc sql;
    select distinct a.obs, a.id, b.id as lag_value
    from tn2 a left join tn2 b
    on a.obs = b.obs + 1
    where a.id <> lag_value and a.id <> lag_value + 1 and lag_value <> .;
quit;

proc print data=tn2;
    where prxmatch('/^(0|1).*$/', treat) = 0
       or prxmatch('/^(Male|Female).*$/', gender) = 0
       or prxmatch('/^( 0| 1| 2| 3| 6| 9|12).*$/', put(time, 2. -R)) = 0
       or prxmatch('/^(0|1).*$/', put(y, 1. -L)) = 0;
run;
```