```
/************************** SAS Assingment-3 ************************************/
/*Name:- Alok Kumar Singh                                                       */
/*ID:-   19250990                                                              */
/*******************************************************************************/

/*Download the climate.csv dataset and read it into SAS. We will concentrate on the
variable air mean.
1. Explore air mean to identify missing values. Create a new dataset to summarise the total number
of observations and the number missing for each site and year.*/

proc import out= ST662Lib.climate
datafile= "/courses/d77u30vavpRs0h7u2Ms92/ST662_data/climate.csv"
dbms=csv replace;
getnames= yes;
run;

data st662lib.climate1;
    set st662lib.climate;
    count = 1;
run;

proc means data = st662lib.climate1;
    by site year;
    var count;
    output out = st662lib.climate2 sum(count) = N;
run;

proc sort data = st662lib.climate1;
    by site year;
run;

data st662lib.climate3;
    set st662lib.climate1;
    by site year;
    if air_mean = . then output;
    drop day month date precip air_min air_mean air_max var10 var11 var12;
run;
proc sort data = st662lib.climate3;
    by site year;
run;

proc means data = st662lib.climate3;
    by site year;
    var count;
    output out = st662lib.climate3 sum(count) = missing;
run;

data st662lib.climatefinal;
    merge  st662lib.climate2 st662lib.climate3;
    by site year;
    if missing <> . then  pmissing = (missing/N) * 100;
    drop _freq_ _type_;
run;

data st662lib.climatefinal;
    set st662lib.climatefinal;
    if missing <> . then  output;
run;

title "Question 1 data set in the format provided in the assingment";
proc print data = st662lib.climatefinal;
```

```
run;
/*2. Impute missing air mean values using the following guidelines*/

/*If a lot more than 5% of observations for a site in a year are missing, use the average of
air min and air max.*/

data st662lib.climatefinal1;
    merge  st662lib.climate1 st662lib.climatefinal;
    by site year;
    if pmissing > 5 then air_mean = (air_min + air_max)/2;
    if pmissing >5 and air_min = . then air_mean = air_max/2;
    if pmissing >5 and air_max = . then air_mean = air_min/2;
run;

/*Otherwise, use the average of all other air mean values for that site and year.*/


proc sort data= st662lib.climatefinal1;
    by site year;
run;
proc stdize data= st662lib.climatefinal1 out=st662lib.climatefinal1
            method= mean missing = mean reponly;
            by site year;
    var air_mean;
run;

/*3. For those sites and years that had some missing values, create a new dataset with
the average of air mean pre- and post- imputation. Provide a printout of the dataset.
Comment on what you have found.*/

data st662lib.climatefinal2;
    merge  st662lib.climate1 st662lib.climatefinal;
    by site year;
run;

proc means data = st662lib.climatefinal2 mean;
    by site year;
    var air_mean;
    where pmissing >0;
    output out = st662lib.climateoldmean mean(air_mean) = old_mean;
run;

proc means data = st662lib.climatefinal1 mean;
    by site year;
    var air_mean;
    where pmissing >0;
    output out = st662lib.climatenewmean mean(air_mean) = new_mean;
run;

data st662lib.climatemeanfinal;
    merge st662lib.climateoldmean st662lib.climatenewmean;
    drop _type_ _freq_;
run;

title "Question 3 data set";
proc print data = st662lib.climatemeanfinal;
run;

/*
In this dataset old mean and new mean are same for all the sites and year except for the sites
having 100% missing as there are no old mean for those cases.
*/
```