Animesh Kansal

Jaahnavi Tiruthani

Medha Tiwary

Hanson Wu

Homework 1 Group 2

/\* Question 1a File examined by opening in notepad\*/

/\* Question 1b reading the file into SAS using DATA step\*/

**data** swFlu;

infile "E:\Users\axk169531\Downloads\HW 2 Data\SwineFlu2009.dat" truncover;

input I\_case\_Date\_ID **1**-**13** I\_case\_continent\_ID **14**-**27** Country $ **28**-**60** I\_Date\_reported yymmdd10. Apr\_reported\_cases **79**-**88** May\_reported\_cases **89**-**98** Jun\_reported\_cases **99**-**108** Jul\_reported\_cases **109**-**118** Aug\_reported\_cases **119**-**128** Tot\_Cum\_reported\_cases **129**-**144** I\_Death\_Date\_ID **145**-**154** I\_Death\_continent\_ID **155**-**168** I\_Date\_Death\_reported yymmdd10. May\_reported\_deaths **194**-**203** Jun\_reported\_deaths **204**-**213** Jul\_reported\_deaths **214**-**223** Aug\_reported\_deaths **224**-**233** Sep\_reported\_deaths **234**-**243** Oct\_reported\_deaths **244**-**253** Nov\_reported\_deaths **254**-**263** Dec\_reported\_deaths;

/\* Question 1c File located in work library and viewed the data\*/

/\* Question 1d Already stored date in the correct SAS date format (not as character). \*/

/\*Question 1e Create a permanent label for each variable based on the preceding descriptions\*/

**data** swFlu;

infile "E:\Users\axk169531\Downloads\HW 2 Data\SwineFlu2009.dat" truncover;

input I\_case\_Date\_ID **1**-**13** I\_case\_continent\_ID **14**-**27** Country $ **28**-**60** I\_Date\_reported yymmdd10. Apr\_reported\_cases **79**-**88** May\_reported\_cases **89**-**98** Jun\_reported\_cases **99**-**108** Jul\_reported\_cases **109**-**118** Aug\_reported\_cases **119**-**128** Tot\_Cum\_reported\_cases **129**-**144** I\_Death\_Date\_ID **145**-**154** I\_Death\_continent\_ID **155**-**168** I\_Date\_Death\_reported yymmdd10. May\_reported\_deaths **194**-**203** Jun\_reported\_deaths **204**-**213** Jul\_reported\_deaths **214**-**223** Aug\_reported\_deaths **224**-**233** Sep\_reported\_deaths **234**-**243** Oct\_reported\_deaths **244**-**253** Nov\_reported\_deaths **254**-**263** Dec\_reported\_deaths;

Label I\_case\_Date\_ID = “ By date: ID for sorting by first case date ”;

Label I\_case\_continent\_ID = “By continent: ID (X.YY) for sorting by first case date within a continent where X represents continent X, and YY represents the YYth country with the next first case”;

Label Country =”Country”;

Label I\_Date\_reported =”Date of first case reported”;

Label Apr\_reported\_cases = “ Number of cumulative cases reported on the first day of the month for April ”;

Label May\_reported\_cases =”Number of cumulative cases reported on the first day of the month for May”;

Label Jun\_reported\_cases = “Number of cumulative cases reported on the first day of the month for JUN ”;

Label Jul\_reported\_cases =”Number of cumulative cases reported on the first day of the month for July ”;

Label Aug\_reported\_cases =”Number of cumulative cases reported on the first day of the month for Aug ”;

Label Tot\_Cum\_reported\_cases = “Last reported cumulative number of cases reported to WHO as of August”;

Label I\_Death\_Date\_ID = “By date: ID for sorting by first death date”;

Label I\_Death\_continent\_ID = “By continent: ID (X.YY) for sorting by first death date within a continent where X represents continent X, and YY represents the YYth country with the next first death”;

Label I\_Date\_Death\_reported = “Date of first death”;

Label May\_reported\_deaths =”Number of cumulative deaths reported on the first day of the month for May”;

Label Jun\_reported\_deaths = “Number of cumulative deaths reported on the first day of the month for Jun”;

Label Jul\_reported\_deaths =”Number of cumulative deaths reported on the first day of the month for Jul”;

Label Aug\_reported\_deaths = “Number of cumulative deaths reported on the first day of the month for Aug”;

Label Sep\_reported\_deaths =”Number of cumulative deaths reported on the first day of the month for Sept”;

Label Oct\_reported\_deaths =”Number of cumulative deaths reported on the first day of the month for Oct”;

Label Nov\_reported\_deaths = “Number of cumulative deaths reported on the first day of the month for Nov”;

Label Dec\_reported\_deaths =”Number of cumulative deaths reported on the first day of the month for Dec”;

**proc** **contents** data=swFlu;

**run**;

/\* Question 1f Printing the contents of the data set including the labels created and other attributes of variables\*/

**proc** **contents** data=swFlu;

**run**;

/\* Question 2a loading the Pizza.csv file using the IMPORT Procedure\*/

**PROC** **IMPORT** OUT= WORK.PIZZA1

DATAFILE= "E:\Users\jxt163130\Downloads\HW 2 Data\Pizza.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=**2**;

**RUN**;

/\* Question 2b Printing the dataset\*/

**proc** **print** data =WORK.PIZZA1;

**RUN**;

/\*Describing the contents of data set \*/

**proc** **contents** data = WORK.PIZZA1;

**run**;

/\*Question 2c Opened the dataset in a WordPad and checked with the results on the SAS screen. The data has been read correctly using the IMPORT procedure. No problem were identified.\*/

/\*Question 2d\*/

**data** pizza;

infile 'E:\Users\jxt163130\Downloads\HW 2 Data\Pizza.csv' DLM =',' dsd firstobs=**2** truncover;

input SurveyNo Argula PineNut Squash Shrimp Eggplant;

**proc** **print** data = pizza;

**run**;

/\*Question 2e Create a new dataset with the average ratings for each topping\*/

**data** newpizza;

set Work.pizza;

;

**proc** **means** data = newpizza noprint;

var Argula PineNut Squash Shrimp Eggplant;

output out = summary(drop=\_TYPE\_ \_FREQ\_) mean = Avg\_Argula Avg\_PineNut Avg\_Squash Avg\_Shrimp Avg\_EggPlant;

**run**;

**proc** **print**;

**run**;

/\*Question 3a\*/

**data** datasas.hotel;

infile 'E:\Users\axk169531\Downloads\HW 2 Data\Hotel.dat' truncover;

input room\_num **1**-**5** num\_of\_guests **6**-**11** check\_in\_month **12**-**15** check\_in\_day **15**-**18** check\_in\_year **18**-**24** check\_out\_month **26**-**29** check\_out\_day **29**-**33** check\_out\_year **33**-**39** service\_used $ **40**-**47** number\_of\_days\_used **47**-**52** room\_type & $ **53**-**68** room\_rate **68**-**72**;

**proc** **print** data = datasas.hotel;

title " now this";

**run**;

/\*Question 3b\*/

**data** datasas.hotel\_2;

set datasas.hotel;

check\_in\_date=mdy(check\_in\_month,check\_in\_day,check\_in\_year);

check\_out\_date=mdy(check\_out\_month,check\_out\_day,check\_out\_year);

format check\_in\_date mmddyy10.;

format check\_out\_date mmddyy10.

run;

**proc** **print** data = datasas.hotel\_2;

**run**;

/\*Question 3C\*/

**data** datasas.hotel\_3;

set datasas.hotel\_2;

if service\_used = "yes" then subtotal = room\_rate\*(check\_out\_date - check\_in\_date) + **10**\*(num\_of\_guests-**1**) + **9.95** + **4.95**\*(number\_of\_days\_used);

else subtotal = room\_rate\*(check\_out\_day - check\_in\_day) + **10**\*(num\_of\_guests-**1**);

**proc** **print** data = datasas.hotel\_3;

**run**;

/\*Question 3d\*/

**data** datasas.hotel\_4;

set datasas.hotel\_3;

grandtotal = subtotal\***1.0775**;

**proc** **print** data = datasas.hotel\_4;

**run**;

/\*Question 3e\*/

data = datasas.hotel\_4;

**proc** **print** data = datasas.hotel\_4 ;

var grandtotal;

where room\_num = **211** ; ;

**run**;

/\*Q4. a. Uploading data using the tab delimited function.\*/

libname Homework 'C:\Users\mxt164730\Desktop\Hw2\_data\Homework';

**DATA** Homework.WLSurveys1;

INFILE 'C:\Users\mxt164730\Desktop\Hw2\_data\Homework\WLSurveys.dat' dsd dlm='09'x truncover;

INPUT SubjectID Height Weight1 Weight2 Weight3 Weight4 Weight5

S1Q1 S1Q2 S1Q3 S1Q4 S1Q5 S1Q6

S2Q1 S2Q2 S2Q3 S2Q4 S2Q5 S2Q6

S3Q1 S3Q2 S3Q3 S3Q4 S3Q5 S3Q6

S4Q1 S4Q2 S4Q3 S4Q4 S4Q5 S4Q6

S5Q1 S5Q2 S5Q3 S5Q4 S5Q5 S5Q6;

**PROC** **print** DATA =Homework.WLSurveys1;

**RUN**;

/\*Renaming the dataset.\*/

**DATA** WLSurveys;

set Homework.WLSurveys1;

**PROC** **print** DATA =WLSurveys;

**RUN**;

/\* Q4. b. Changing the wrongly entered survey value for questions 2, 3 and 5 from each survey. \*/

/\*Making a smaller dataset with the targeted variables to modify the values.\*/

**DATA** WL1;

set WLSurveys;

keep SubjectID S1Q2 S1Q3 S1Q5 S2Q2 S2Q3 S2Q5 S3Q2 S3Q3 S3Q5 S4Q2 S4Q3 S4Q5 S5Q2 S5Q3 S5Q5;

array change S1Q2--S5Q5;

do over change;

if change=**0** then change=**3**;

else if change=**1** then change=**2**;

else if change=**2** then change=**1**;

else if change=**3** then change=**0**;

end;

**run**;

/\*Creating another dataset without the above mentioned variables to make a consistent dataset\*/

**data** WL2;

set WLSurveys;

drop S1Q2 S1Q3 S1Q5 S2Q2 S2Q3 S2Q5 S3Q2 S3Q3 S3Q5 S4Q2 S4Q3 S4Q5 S5Q2 S5Q3 S5Q5;

**run**;

/\*Merging datasets WL1 and WL2 to create a new dataset WL3 with all the desired data values.\*/

**proc** **sort** data= WL1;by SubjectID;

**run**;

**proc** **sort** data= WL2;by SubjectID;

**run**;

**data** WLFinal;

retain SubjectID Height Weight1 Weight2 Weight3 Weight4 Weight5

S1Q1 S1Q2 S1Q3 S1Q4 S1Q5 S1Q6

S2Q1 S2Q2 S2Q3 S2Q4 S2Q5 S2Q6

S3Q1 S3Q2 S3Q3 S3Q4 S3Q5 S3Q6

S4Q1 S4Q2 S4Q3 S4Q4 S4Q5 S4Q6

S5Q1 S5Q2 S5Q3 S5Q4 S5Q5 S5Q6;

merge WL1 WL2;

by SubjectID;

**run**;

**proc** **print** data= WLFinal;

**run**;

/\*Q4. c. Using the ARRAY function to change the wrong datapoint -99 into missing values . for each of the

five surveys \*/

**DATA** WLsurveyMissing;

set WLFinal;

ARRAY WLFinal (**30**)

S1Q1 S1Q2 S1Q3 S1Q4 S1Q5 S1Q6

S2Q1 S2Q2 S2Q3 S2Q4 S2Q5 S2Q6

S3Q1 S3Q2 S3Q3 S3Q4 S3Q5 S3Q6

S4Q1 S4Q2 S4Q3 S4Q4 S4Q5 S4Q6

S5Q1 S5Q2 S5Q3 S5Q4 S5Q5 S5Q6;

DO i = **1** TO **30**;

IF WLFinal(i) = -**99** THEN WLFinal(i) = **.**;

END;

**proc** **print** data = WLsurveyMissing;

**run**;

/\* Q4. d. Calculating BMI of each visit.\*/

**DATA** BMI;

set WLSurveyMissing;

BMI1= (Weight1/ Height\*\***2**) \* **703**;

BMI2= (Weight2/ Height\*\***2**) \* **703**;

BMI3= (Weight3/ Height\*\***2**) \* **703**;

BMI4= (Weight4/ Height\*\***2**) \* **703**;

BMI5= (Weight5/ Height\*\***2**) \* **703**;

**RUN**;

**proc** **print** data= BMI;

**run**;

/\* Q4. e. Filtered data of patients with BMI 25 and above.\*/

**DATA** BMI25;

set BMI;

if BMI5 >= **25**;

**RUN**;

/\*Q4.f. View the resulting data.\*/

**proc** **print** data= BMI25;

**run**;

/\*Q4.f. Tenth observation ---> SubjectID = 13,

No. of question variables with missing values = 1,

BMI in the last visit= 25.9832 \*/

/\*Question 5a\*/

LIBNAME homework 'C:\Users\Hanson\Documents\Hanson\College\UTD Spring 2018\Predicive Analaysis using SAS\Datasets';

**DATA** AveCom;

SET homework.AVEPRICES;

**PROC** **MEANS** DATA = AveCom NOPRINT NWAY;

CLASS Year Commodity;

VAR Price;

OUTPUT OUT = Averages (drop=\_TYPE\_ \_FREQ\_) MEAN = AvgPrice;

**RUN**;

**PROC** **SORT** DATA=AveCom;

BY Year Commodity;

**PROC** **PRINT** DATA=Averages;

**RUN**;

/\*Question 5b\*/

**DATA** Month;

SET homework.AVEPRICES;

**PROC** **SORT** DATA=Month;

BY Year Commodity;

**RUN**;

**PROC** **TRANSPOSE** DATA=Month OUT=PricePerMonth (drop=\_NAME\_ \_LABEL\_);

BY Year Commodity;

VAR Price;

ID Month;

**PROC** **PRINT**;

**RUN**;

/\*Question 5c\*/

**DATA** Merged (rename=(\_1=Jan \_2=Feb \_3=Mar \_4=Apr \_5=May \_6=Jun \_7=Jul \_8=Aug \_9=Sep \_10=Oct \_11=Nov \_12=Dec));

MERGE Averages PricePerMonth;

BY Year Commodity;

**PROC** **PRINT** DATA=Merged;

**RUN**;

/\*Question 5d\*/

**Data** NewGas NewEgg NewMilk;

SET Merged;

SELECT (Commodity);

WHEN ('Gas') OUTPUT NewGas;

WHEN ('Egg') OUTPUT NewEgg;

WHEN ('Milk') OUTPUT NewMilk;

END;

**RUN**;

/\*Question 5e This Step was already accomplished in question 5c\*/

**PROC** **PRINT** DATA=NewGas;

**PROC** **PRINT** DATA=NewEgg;

**PROC** **PRINT** DATA=NewMilk;

**RUN**;

/\*Question 5f\*/

/\*Dataset NewGas contains 15 variables and 11 observations

Dataset NewEgg contains 15 variables and 11 observations

Dataset NewMilk contains 15 variables and 11 observations\*/

/\* Question 6a \*/

/\* reading the School Survey dataset\*/

libname Homework 'E:\Users\jxt163130\Documents\Homework';

**data** schooldata;

set Homework.schoolsurvey;

**run**;

/\*sorting the dataset by family id\*/

**proc** **sort** data = schooldata;

by Family\_id;

/\*printing the sorted dataset\*/

**proc** **print** data =schooldata;

title 'School Survey Data';

**run**;

/\* Question 6b \*/

/\* Create a data set that has one observation for each sixth grader \*/

**data** grade6data;

set schooldata;

if Grade ='6th';

**run**;

**proc** **print** data = grade6data;

title '6th graders data';

**run**;

/\* Question 6c \*/

/\* Combine the data set from part b) with the original data \*/

**data** grade6data1 (rename=(Student\_ID=stdnt\_id School=schname Grade=grad DOB=BirthDate));

set grade6data;

**run**;

**proc** **print** data=grade6data1;

**run**;

**data** agediff;

merge schooldata grade6data1;

by Family\_id;/\* create a variable that is the age difference in years between the sixth grader and the sibling.\*/

age\_difference = (Birthdate-DOB)/**365**;

**run**;

**proc** **print** data=agediff;

**run**;

/\* Question 6d \*/

/\* Count the number of older siblings and the number of younger siblings of the sixth grader\*/

**proc** **sql**;

create table eldersibling as

select Family\_id,count(age\_difference) as eldercount from agediff where age\_difference>**0**

group by Family\_id;

create table youngsibling as

select Family\_id,count(age\_difference) as youngcount from agediff where age\_difference<**0**

group by Family\_id;

**run**;

**data** noofsiblings;

merge grade6data eldersibling youngsibling;

by Family\_id;

**run**;

/\* printing the number of siblings of each 6th grader \*/

**proc** **print** data = noofsiblings;

title 'Siblings data';

**run**;

/\* Question 6e \*/

/\* compute the number of sixth graders, and the total number of younger and older siblings by school. \*/

**proc** **sql**;

create table countbyschool as

select School, count(Student\_ID) as grade6count, sum(eldercount) as count\_elder, sum(youngcount) as count\_young from noofsiblings;

group by School;

alter table countbyschool modify School char(**20**);

update countbyschool set School="Green Valley" where School="GV";

update countbyschool set School="Rachel Carson" where School="RC";

update countbyschool set School="Redwood Grove" where School="RG";

**proc** **print** data=countbyschool;

title 'Count of students by school';

**run**;

/\* Question 6f \*/

/\* Showing the age difference between the sixth grader and their youngest sibling, and the age difference between the sixth grader and their oldest sibling.\*/

**proc** **sql**;

create table min\_agediff as

select Family\_id, min(age\_difference) as diff\_youngest\_6th from agediff group by Family\_id;

create table max\_agediff as

select Family\_id, max(age\_difference) as diff\_eldest\_6th from agediff group by Family\_id;

**run**;

**data** grade6siblings;

merge noofsiblings min\_agediff max\_agediff;

by Family\_id;

**run**;

**proc** **print** data=grade6siblings;

title 'Grade 6 siblings count and age difference';

**run**;

/\* Question 6g \*/

/\* compute the average, minimum, and maximum age difference between the sixth graders and their youngest and oldest siblings. \*/

**proc** **sql**;

create table school\_stat\_summary as

select School, avg(diff\_youngest\_6th) as avgdiff\_6th\_young, min(diff\_youngest\_6th) as max\_agediff\_6th\_young,

max(diff\_youngest\_6th) as min\_agediff\_6th\_young, avg(diff\_eldest\_6th) as avgdiff\_6th\_elder,

min(diff\_eldest\_6th) as min\_agediff\_6th\_elder, max(diff\_eldest\_6th) as max\_agediff\_6th\_elder from grade6siblings

group by School;

alter table school\_stat\_summary modify School char(**20**);

update school\_stat\_summary set School="Green Valley" where School="GV";

update school\_stat\_summary set School="Rachel Carson" where School="RC";

update school\_stat\_summary set School="Redwood Grove" where School="RG";

**run**;

**proc** **print** data=school\_stat\_summary;

title ' School Statistics';

**run**;