**ASSIGNMENT 1**

**QUESTION 1**

Open your SAS environment and copy the following program into a blank program editor and save the code as: grades.sas.

data GRADES; infile datalines;  
 input  
 student /\* student id \*/  
 gender $  
 exam1 /\* first exam mark \*/  
 exam2 /\* second exam mark \*/  
 coursework $ /\* coursework grade \*/  
 ;   
datalines;   
 1 M 70 82 A  
 2 M 62 54 B  
 3 F 72 73 A  
 4 M 60 65 B  
 5 F 52 50 B  
 6 F 81 78 B  
 7 F 40 49 C  
 8 M 59 60 C  
 9 M 64 66 C  
 10 F 64 63 B  
 11 M 69 70 A  
 12 M 75 72 A   
 13 M 38 50 B  
 14 F 63 69 B   
 15 M 61 45 C   
 ;  
run;  
  
proc print  
 data = GRADES;  
run;

1. Use the various sources of help to run the program and:
   1. Describe the content of the log.
2. Describe the output.
3. Find the WORK.GRADES data set, open it and describe its contents.

**ANSWER 1**

1. The log contains the inputted data in black and results in form of notes in blue.
   1. The notes states that there are 15 observations and 5 variables.
   2. It also shows the total time it took to process the data.
2. The output is a table visual of the inputted data titled The SAS System.
3. The WORK.GRADES data set can be found in the explorer>library>work tab. It contains a view table of the inputted datalines in rows and columns.

**QUESTION 2**

The data step names the data set (GRADES), uses *infile* to state where the raw data are stored, uses *input* to state which variables are to be read and *datalines* to list the data.

The log shows whether there was an error in the program. There should be no errors or warnings following execution of a successful program.

The output shows a table of the data. It should be identical to the *datalines*. The data set that SAS stores internally, should also be identical to the data.

**Self-assessment question**

1. Why might using a datalines statement be useful when developing SAS programs to analyse large data sets?
2. Why must the log contain no errors or warnings after successful execution of a program?
3. Why print the results of the data step, immediately afterward?

**Note: *Infile* references where the data is stored**

**ANSWER 2**

1. Datalines statement is used in adding the raw data set to be analysed.
2. A log must contain no errors or warning to ensure accuracy of the result in the output. An error or warning requires revision of the code written in the editor.
3. For best price, printing the results of the data step will help in detecting possible errors quickly.

**QUESTION 3**

Add the following code after the *proc print statement.*

proc sgplot data=GRADES;  
 scatter x=exam1 y=exam2 / group=coursework datalabel=student;  
 ellipse x=exam1 y=exam2;  
 reg x=exam1 y=exam2 / nomarkers;  
 xaxis min=0 max=100;  
 yaxis min=0 max=100;  
 run;

Run the code and then answer the following questions, when you see a graph:

1. Are there any students that have improved their performance from exam1 to exam2?
2. Are there any students with unexpected results for their coursework?
3. Can you state whether this is a statistically valid regression line? (Justify your answer.)
4. What further analyses might you do on each data item before accepting the results displayed in the graph?
5. What further analyses might you do following the regression analysis before accepting the results, displayed in the graph?

|  |
| --- |
| 1. For the toolbar, press ALT+F10 (PC) or ALT+FN+F10 (Mac). |

**ANSWER 3**

1) From the plot, students whose grades increased from a lower range in exam 1 to a higher range in exam2 improved in their performance. For instance,

Student 1 was in the 60 - 80 range in exam1 but improved to 80 - 100 range in exam2

Student 7 was in the 20 – 40 range in exam1 but improved to 40 – 60 range in exam2

Student 13 was in the 20 – 40 range in exam1 but improved to 40 – 60 range in exam2

2) Yes, there are students 1, 13 and 15 had significantly different grades in their coursework

3) Yes, it is a statistically valid regression line as most of the variables are closed except for 1 and 15 which are outliers. ﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿﻿

4) Using the line of regression, i will use the linear regression model to calculate the correlation of estimated values of the slope and intercepts.

Since the result of exam2 is depend on exam1 we can use the equation of regression line to calculate if our values are correct.

5) Hypothesis testing framework. if the probability of the results is true then we will say it’s a null hypothesis but if it is false then it’s an alternative hypothesis.

**QUESTION 4**

Open income.dat using Notepad++. From Notepad++'s menu, select: View|Show symbol|Show all characters. This should make all characters in the file visible.

The variables are stored in each row of the file are:

income1 /\* principal income \*/  
 income2 /\* secondary income \*/  
 size /\* family size \*/  
 ownrent /\* owns or rents house \*/  
 amount /\* rent/mortgage payment \*/  
 utility /\* average monthly utility payment \*/  
 location /\* city location \*/

##### Self assessment question

Describe the contents in the file. How are values separated? How are rows of data separated?

**ANSWER 4**

The contents of the file income.dat are variables stored in rows and columns with observations in numerical order.

The values are separated by red continuous dots between each value.

The row is separated with CRLF at the end of each line.

**QUESTION 5**

Copy the following program into an empty program editor and save it as: **income.sas**.

/\* pwd = Present Working Directory \*/

/\* The location for the data file \*/

filename pwd '/folders/myfolders/imat5168/2018/lesson01';

data HOUSE;

/\* ... using input from the income.dat file \*/

infile pwd(income.dat);

/\* ... creating the following variables \*/

input

income1 /\* numeric, principal income \*/

income2 /\* numeric, secondary income \*/

size /\* numeric, family size \*/

ownrent /\* numeric, owns or rents house \*/

amount /\* numeric, rent/mortgage payment \*/

utility /\* numeric, average monthly utility payment \*/

location /\* numeric, city location \*/

;

/\* ... which have the following display names \*/

label

income1 = 'Principal income'

income2 = 'Secondary income'

size = 'Size of family'

ownrent = 'Own or rent house'

amount = 'Amount of mortgage or rent payment'

utility = 'Average monthly utility payment'

location = 'Location in city'

;

run;

**Self assessment question**

Adjust the *Present Working Directory* (**pwd**) *filename* to use the location where you put **income.dat**. Run the program.

1. Did it run successfully? (Justify your answer)
2. What output did it produce?
3. Which data set did it produce?
4. Compare this data step with that in grades.sas; which do you prefer? (Justify your answer.)

**ANSWER 5**

1) No it did not. It returned an error stating that the file does not exist in my H: drive after changing it to my file location.

However, I solved the problem by copying my file path into the input function.

2) The output produces the inputted data steps in black and noted in blue showing the activity timestamps and number of observations as 50 and number of variables as 7.

3) It produces the data steps created in the program.

4) I prefer that of grades.sas, the data step is properly named and the input refers to the datalines which contains the raw data (income.dat) directly.

**Code for correct line**

/\* pwd = Present Working Directory \*/

/\* The location for the data file \*/

;

**data** HOUSE;

/\* ... using input from the income.dat file \*/

infile 'D:\Sas\income.dat';

/\* ... creating the following variables \*/

input

income1 /\* numeric, principal income \*/

income2 /\* numeric, secondary income \*/

size /\* numeric, family size \*/

ownrent /\* numeric, owns or rents house \*/

amount /\* numeric, rent/mortgage payment \*/

utility /\* numeric, average monthly utility payment \*/

location /\* numeric, city location \*/

;

/\* ... which have the following display names \*/

label

income1 = 'Principal income'

income2 = 'Secondary income'

size = 'Size of family'

ownrent = 'Own or rent house'

amount = 'Amount of mortgage or rent payment'

utility = 'Average monthly utility payment'

location = 'Location in city'

;

**run**;

**proc** **print** data =HOUSE;

**run**;

/\* tabulate a descriptive summary for principle income by location \*/

**proc** **tabulate** data=HOUSE;

/\* divide the incomes into groups by... \*/

class location;

/\* calculate descriptive statistics for... \*/

var income1;

/\* for each location show the summary \*/

table location all, income1 \* (n min p25 median p75 max mean std);

**run**;

**QUESTION 6**

Although it is not possible to check that each row in a data file is correct, there are useful checks that can be performed on numerical and categorical data. The code produces the following summary of numerical data:

* **n** - the number of data items (not part of the summary)
* **min** - the minimum value
* **p25** - the 25th percentile value
* **median** - the median
* **p75** - the 75th percentile value
* **max** - the maximum value
* **mean** - the mean
* **std** - the standard deviation
* for various locations in a city.

/\* tabulate a descriptive summary for principle income by location \*/

proc tabulate data=HOUSE;

/\* divide the incomes into groups by... \*/

class location;

/\* calculate descriptive statistics for... \*/

var income1;

/\* for each location show the summary \*/

table location all, income1 \* (n min p25 median p75 max mean std);

run;

**Self assessment question**

Run the program after adding this code after the data step.

1. How could these data be used in the validation of numerical data?
2. What do the data show?
3. Where would you expect people who are renting to live? (If it were reasonable to suppose that renting was associated with lower incomes.)

**ANSWER 6**

1) 1) By inputting and running the data in the SAS program.

2) The data shows the number of people living in each location. The minimum income, the percentile value, average income, maximum and standard deviation.

3) I would expect people who are renting to live in location 4 and 5 although there isn’t enough data from those locations to come to an exact conclusion the income form those locations are significantly lower than 1 and 2.

**QUESTION 7**

The following code explores the relationship between location and owning/renting a home.

/\* tabulate the frequency of owners/renters by location \*/  
 proc freq data = HOUSE;  
 tables location \* ownrent;  
 run;

**Self-assessment question**

Run the program.

1. Where does the highest proportion of renters live? (Justify your answer.)
2. What further information might you find useful?
3. Review both grades.sas and income.sas against the guidelines for good coding practise. What improvements would you make and why?

**ANSWER 7**

1) the highest proportion of renters live in location 4. Although the data is too little to make conclusion, location 4 has the highest renters because of their lower income.

2) Majority of the people in location 1 are owners

3) I would type my code in small for ease of reading

/\* would be used for multiline comments\*/

I will justify data and procedure statements after running them.