**STA563/HW2 NAME: LINA LEE**

**/\*header: Exercise Number 1 Read files using Import procedure and Data Step.**

**Author: Lina Lee**

**Purpose: read file Pizza.csv, print the last 4 observations from this dataset. And provide description for the contents of the data. And check the case that “Import procedure” cannot resolve.**

**input Pizza.csv**

**input variable:**

**SurveyNum : the respondents ID**

**Arugula PineNut Squash Shrimp Eggplant: Topping names\*/**

**/\*create a library\*/**

libname hw2 "C:\Users\linal\Desktop\2018\STA502\HW2";

**/\*EX1-(a) Import a data into SAS\*/**

proc import datafile="C:\Users\linal\Desktop\2018\STA502\HW2\Pizza.csv"

out=Pizimport\_excel

dbms=csv replace;

getnames=yes;

run;

**/\* EX1-(a) save a dataset in permernant library\*/**

data hw2.topping;

set work.Pizimport\_excel;

run;

**/\*EX1-(b) describes the contents of the data set\*/**

ods rtf file="C:\Users\linal\Desktop\2018\STA502\HW1\problem1" bodytitle style=journal;

title "EX1 Survey about Customer's Preferance for New Toppings";

proc contents data=hw2.topping;

run;

**/\*EX1-(a) select only last 4 observations\*/**

data topping2;

set hw2.topping nobs=\_\_nobs;

if \_n\_ gt \_\_nobs-4;

run;

**/\*EX1-(a) print last 4 observations\*/**

proc print data=topping2;

run;

ods rtf close;

ods rtf file="C:\Users\linal\Desktop\2018\STA502\HW1\problem1\_" bodytitle style=journal;

title "EX1 Survey about Customer's Preferance for New Toppings";

**/\*EX-(d) read the data using a DATA Step\*/**

data Pizza;

infile "C:\Users\linal\Desktop\2018\STA502\HW2\Pizza.csv" dsd missover firstobs=2 ;

input SurveyNum:$4. Arugula PineNut Squash Shrimp Eggplant;

run;

**/\*EX1-(e) describes the contents of the data set which is created by data steps\*/**

/\*ods rtf file="C:\Users\linal\Desktop\2018\STA502\HW1\problem1d" bodytitle style=journal;

title "EX1-(d) Pizza topping";\*/

proc contents data=Pizza;

run;

**/\*EX1-(e) select only last 4 observations\*/**

data topping3;

set Pizza nobs=\_\_nobs;

if \_n\_ gt \_\_nobs-4;

run;

**/\*EX1-(e) print last 4 observations\*/**

proc print data=topping3;

run;

ods rtf close;

**/\*header: Exercise Number 2,bar charts, a Box Plot, A graph.**

**Author: Lina Lee**

**Purpose: To create a bar chart of the responses to the first survey question from visit two, a box plot of BMI against the responses to the last survey question from the last visit. To make a graph of the regression line on top of the scatter plot using weight as the predictor and bmi as the response variable. To create another bar chart, each bar represents the average BMI for all people with responses to the first survey question.**

**input variable:**

**StudentID Height Weight BodyMassIndex**

**Q1~Q30: responses to six survey questions regarding weight loss were obtained at each of five visits for a total of 30 survey questions*\*/***

ods rtf file="C:\Users\linal\Desktop\2018\STA502\HW2\problem2" bodytitle style=journal;

title "EX2 weight loss in males";

**/\*2-(a) make a bar chart of the responses to the first survey question from visit two\*/**

proc sgplot data=hw2.Wls;

title "a bar chart of responses to the first survey question from visit2";

vbar Q7 /missing;

xaxis label="responses to the first survey question from visit2";

yaxis label="Number of people who respond to the first survey question from visit2";

run;

**/\*2-(b) make a box plot of bmi against the responses to the last survey\*/**

proc sgplot data=hw2.Wls;

title "a box plot of bmi against the responses to the last survey";

vbox bmi /category=Q30 missing;

xaxis label="responses to the last survey";

run;

**/\*2-(c) make a graph of the regression line on top of the scatter plot (predictor:weight, response: bmi)\*/**

proc sgplot data=hw2.Wls;

title " a graph of the regression line on top of the scatter plot (predictor:weight, response: bmi)";

reg x=weight y=bmi/nomarkers;

scatter x=weight y=bmi;

run;

**/\*2-(d) create another bar chart of the responses to the first survey question from visit two, each bar represents the average BMI for all people with the response\*/**

proc sgplot data=hw2.Wls;

title "Avrage BMI by response to question 1 in visit2";

vbar Q7 /response=bmi stat=mean missing;

xaxis label=" responses to the first survey question from visit2";

yaxis label="Avrage Body Max Index";

run;

ods rtf close;

***EX1-(a) Survey about Customer's Preferance for New Toppings, Four observations***

| *Obs* | *SurveyNum* | *Arugula* | *PineNut* | *Squash* | *Shrimp* | *Eggplant* |
| --- | --- | --- | --- | --- | --- | --- |
| *1* | 1207 | . | 1 | 1 | 5 |  |
| *2* | 1208 | . | 3 | 5 | 1 |  |
| *3* | 1209 | . | 4 | 5 | 3 |  |
| *4* | 1210 | . | 5 | 5 | 1 |  |

Answer for EX1-(a) : raw data file include 120 observations, 6 variables. Last 4 observation include Survey Number 1207~1210,

Each of Which give rate 1,3,4,5 for PineNut, 1,5,5,5 for Squash, 5,1,3,1 for Shrimp.

***EX1-(b) Survey about Customer's Preferance for New Toppings***

|  |
| --- |
| ***The CONTENTS Procedure*** |

|  |  |  |  |
| --- | --- | --- | --- |
| *Data Set Name* | HW2.TOPPING | *Observations* | 120 |
| *Member Type* | DATA | *Variables* | 6 |

| *Alphabetic List of Variables and Attributes* | | | | | |
| --- | --- | --- | --- | --- | --- |
| *#* | *Variable* | *Type* | *Len* | *Format* | *Informat* |
| *2* | Arugula | Num | 8 | BEST12. | BEST32. |
| *6* | Eggplant | Char | 1 | $1. | $1. |
| *3* | PineNut | Num | 8 | BEST12. | BEST32. |
| *5* | Shrimp | Char | 1 | $1. | $1. |
| *4* | Squash | Num | 8 | BEST12. | BEST32. |
| *1* | SurveyNum | Num | 8 | BEST12. | BEST32. |

Answer for EX1-(b) : Eggplant, Shrimp should be numeric, but here, appear as charaters.

Answer for EX1-(c) :. In the raw data, Missing values in Eggplant appear as “ ”(empty) eventhough missing values of other variables appear as “.”

Answer for EX1-(d) : all missing values appeared as “.” Not “ ”. Type of variables “Eggplant”, “Shrimp” is numeric now.

***EX1(e) Survey about Customer's Preferance for New Toppings***

|  |
| --- |
| ***The CONTENTS Procedure*** |

|  |  |  |  |
| --- | --- | --- | --- |
| *Data Set Name* | WORK.PIZZA | *Observations* | 120 |
| *Member Type* | DATA | *Variables* | 6 |

| *Alphabetic List of Variables and Attributes* | | | |
| --- | --- | --- | --- |
| *#* | *Variable* | *Type* | *Len* |
| *2* | Arugula | Num | 8 |
| *6* | Eggplant | Num | 8 |
| *3* | PineNut | Num | 8 |
| *5* | Shrimp | Num | 8 |
| *4* | Squash | Num | 8 |
| *1* | SurveyNum | Char | 4 |

***EX1(e) Survey about Customer's Preferance for New Toppings***

| *Obs* | *SurveyNum* | *Arugula* | *PineNut* | *Squash* | *Shrimp* | *Eggplant* |
| --- | --- | --- | --- | --- | --- | --- |
| *1* | 1207 | . | 1 | 1 | 5 | . |
| *2* | 1208 | . | 3 | 5 | 1 | . |
| *3* | 1209 | . | 4 | 5 | 3 | . |
| *4* | 1210 | . | 5 | 5 | 1 | . |

Answer for EX1-(e): still 120 observations and 6 variables. It contain one character variable “SurveyNum”, and 5 Numeric variables.

In the print of the last 4 observations in the new dataset, I can see that all missing values of Eggplant now appear as “.”

Exercise2-(a): A bar chart of responses to the first survey question from visit2



Answer for EX2-(a) : x-axis is responses to the first survey question from visit2. Some responded from 0 to 3 and the other didn’t respond. Number of people who responded “0” to the first survey question from visit2 is more than 80, some who said “1” is around 70, some who say “2” said around 80, more than 70 people said “3”.

Exercise2-(b) a box plot of BMI against the responses to the last survey



Answer for EX2-(a): X axis is responses to the last survey, Y axis is Body Mass Index. BMI Range for response “ ” is from 20 to around 28. Median of it is 23.5. BMI range for response “0” is from less than 20 to around 30. Median of it is around 24. BMI range for response “1” is from far less than 20 to around30. Median of it is around 23.5. BMI range for response “2” is from far less than 20 to around 29. BMI range for response “3” is from far less than 20 to around 32. Median of it is 24.

Response “3” has the biggest range of BMI. No response has the smallest range of BMI.

Exercise2-(c) a graph of the regression line on top of the scatter plot



Answer for EX2-(c): Data are scattered as two big parts. The intercept of the regression line is around 21. The slope is around 0.075.

Exercise2-(d) Average BMI by response to the question 1 in visit 2



Answer for EX2-(c): X axis is responses to the first survey question from visit2. Y axis is Number of people who respond to the first survey question from visit2. Average Body Max Index of people who respond to the first survey question from visit2 is around 24~25 across the responses.