**Homework 3**

**STAT 6494 Fall 2017**

**Haim Bar**

**Due Date: September 24, 2017**

**(by 23:59, via HuskyCT)**

1. **SAS**: You have a CSV file called employee.csv. This file contains the following information:

|  |  |  |
| --- | --- | --- |
| Variable | Description | Desired Informat |
| ID | Employee ID | $3. |
| Name | Employee Name | $20. |
| Depart | Department | $8. |
| DateHire | Hire date | MMDDY10. |
| Salary | Yearly salary | DOLLAR8. |

Use list input to read data from this file. You will need an informat to read most of these values correctly (i.e., DateHire needs a date informat). You need to do this in two ways. **First** is to include an INFORMAT statement to associate each variable with the appropriate informat. **The other** is to use the colon modifier and supply the informats directly in the INPUT statement. Create a temporary SAS data set(Employ) from this data file. Use PROC PRINT to list the observations in your data set and the appropriate procedure to compute frequencies for the variable Depart.

A list of the raw data file is:

File employee.csv

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 123 | Harold Wilson | Acct | 1/15/1989 | $78 | 123 |
| 128 | Julia Child | Food | 8/29/1988 | $89 | 123 |
| 7 | James Bond | Security | 2/1/2000 | $82 | 100 |
| 828 | Roger Doger | Acct | 8/15/1999 | $39 | 100 |
| 900 | Earl Davenport | Food | 9/9/1989 | $45 | 399 |
| 906 | James Swindler | Acct | 12/21/1978 | $78 | 200 |

filename emp 'P:\STAT-6494-Data Management in SAS and R\data\employee.csv';

\*Including INFORMAT statement way;

**data** Employ;

infile emp dsd;

informat ID $3.

Name $20.

Depart $8.

DateHire MMDDYY10.

Salary DOLLAR8.;

input ID Name Depart DateHire Salary;

format DateHire MMDDYY10. Salary DOLLAR8.;

**run**;

\*Using colon modifier and supplying the informats directly in the INPUT statement;

**data** Employ;

infile emp dsd;

input ID : $3.

Name : $20.

Depart : $8.

DateHire : MMDDYY10.

Salary : DOLLAR8.;

format DateHire MMDDYY10. Salary DOLLAR8.;

**run**;

**proc** **print** data=employ;

**run**;

**proc** **freq** data=employ;

table depart;

**run**;

1. **R**: create a data frame by reading the data from employee.csv. Choose your own variable names. Print out a sentence for each person like the following: (use the paste function)

*The occupation of Harold Wilson is Acct, he/she has been in our company since 1-15-1989. His/her salary is $78/hour.*

employ<-read.csv("P:/STAT-6494-Data Management in SAS and R/data/employee.csv",header=F,stringsAsFactors = F)

colnames(employ)<-c("ID","Name","Depart","DateHire","Salary")

employ$DateHire=as.Date(employ$DateHire, "%m/%d/%Y")

paste("The occupation of ", employ$Name, " is ", employ$Depart,", he/she has been in our company since ",

format(employ$DateHire,"%m-%d-%Y"), ". His/her salary is ", employ$Salary, "/hour.", sep="")

1. Create a permanent **SAS** data set called Voter, data is listed below:

23 D 1 1 2 2

45 R 5 5 4 1

67 D 2 4 3 3

39 R 4 4 4 4

19 D 2 1 2 1

75 D 3 3 2 3

57 R 4 3 4 4

Use column names col1 – col6.

libname hw3 'P:\STAT-6494-Data Management in SAS and R\data';

**data** hw3.voter;

input col1 col2 $ col3-col6;

datalines;

23 D 1 1 2 2

45 R 5 5 4 1

67 D 2 4 3 3

39 R 4 4 4 4

19 D 2 1 2 1

75 D 3 3 2 3

57 R 4 3 4 4

;

**run**;

1. Use LABEL to add labels to variables of this dataset. The label for first column is Age, second is Party, column 3 – 6 as follows:

Column3 The president is doing a good job

Column4 Congress is doing a good job

Column5 Taxes are too high

Column6 Government should cut spending

**data** hw3.voter;

set hw3.voter;

label col1='Age'

col2='Party'

col3='The president is doing a good job'

col4='Congress is doing a good job'

col5='Taxes are too high'

col6='Government should cut spending';

**run**;

1. Use PROC PRINT to list the observations in this data. Add formats for Age (0-30, 31-50, 51-70, 71+), Party (D = Democrat, R = Republican), and Col3 – Col6 (1 = Strongly Disagree, 2 = Disagree, 3 = No Opinion, 4 = Agree, 5 = Strongly Agree). Use Age as the first column in the output. And use labels defined in a. as column headings.

libname myfmt 'P:\STAT-6494-Data Management in SAS and R\data';

**proc** **format** library=myfmt;

value Age low-**30** = '0-30'

**31**-**50** = '31-50'

**51**-**70** = '51-70'

**71**-high = '71+';

value $Party 'D' ='Democrat'

'R' = 'Republican';

value opinion **1**='Strongly Disagree'

**2** = 'Disagree'

**3** = 'No Opinion'

**4** = 'Agree'

**5** = 'Strongly Agree';

**run**;

**proc** **print** data=hw3.voter label;

id col1;

format col1 Age. col2 $Party. col3-col6 opinion.;

**run**;

1. What if those ages are given as decimal points, how will you add the format?

**proc** **format** library=myfmt;

value AGE low<-**30**='(0, 30]'

**30**<-**50**='(30, 50]'

**50**<-**70**='(50, 70]'

**70**<-high='greater than 70';

**run**

1. You want to see frequencies for Questions 1 to 4 (column 3 – 6). However, you want only three categories: Generally Disagree (combine Strongly Disagree and Disagree), No Opinion, and Generally Agree (combine Agree and Strongly Agree). Accomplish this using a new format for column 3 - 6. Use the corresponding question as title.

**proc** **format** library=myfmt;

value three **1**,**2**='Generally Disagree'

**3** = 'No Opinion'

**4**,**5** = 'Generally Agree';

**run**;

**proc** **freq** data=hw3.voter;

title "Question Frequencies Using the three categories";

table col3-col6;

format col3-col6 three.;

**run**;

1. Suppose you received this dataset and associated formats file from someone else. Load this permanent SAS data set into a temporary SAS data set with formats, called Voter2.

libname hw3 'P:\STAT-6494-Data Management in SAS and R\data';

libname myfmt 'P:\STAT-6494-Data Management in SAS and R\data';

options fmtsearch=(myfmt work library);

**data** voter2;

set hw3.voter;

format col1 Age. col2 $Party. col3-col6 opinion.;

**run**;

1. Run the following program to create a **SAS** data set called Colors:

**data** colors;

input Color : $1. @@;

datalines;

R R B G Y Y . . B G R B G Y P O O V V B

;

Use a format to group the colors as follows:

R, B, G = Group 1

Y, O = Group 2

Missing = Not Given

All others = Group 3

Use PROC FREQ to list the frequencies of the color groups.

**proc** **format**;

value $group 'R','B','G' = 'Group 1'

'Y','O' = 'Group 2'

' ' = 'Not Given'

other = 'Group 3';

**run**;

**proc** **freq** data=colors;

format color $group.;

**run**;

1. **SAS**: Write the necessary statements to make permanent formats in a library of your choice. Use the FMTLIB option to list YESNO and $YESNO(using 2 different ways). The formats are defined as follows:

YESNO 1 = Yes, 0 = No

$YESNO Y = Yes, N = No

$Gender M = Male, F = Female

Age20yr low-20 = 1, 21-40 = 2, 41-60 = 3, 61-80 = 4, 81-high = 5

libname fmt 'P:\STAT-6494-Data Management in SAS and R\data';

**proc** **format** library=fmt;

value YESNO **1** ='Yes' **0** = 'No';

value $YESNO 'Y' ='Yes' 'N' = 'No';

value $Gender 'M' = 'Male' 'F' = 'Female';

value Age20yr low-**20**= '1' **21**-**40** = '2' **41**-**60** = '3' **61**-**80** = '4' **81**-high = '5';

**run**;

options fmtsearch=(myfmts);

**proc** **format** library=myfmts fmtlib;

select YESNO $YESNO;

**run**;

1. Create a data frame in **R**, using the input data in question 3. Choose your own variable names.

votertxt<-"

23 D 1 1 2 2

45 R 5 5 4 1

67 D 2 4 3 3

39 R 4 4 4 4

19 D 2 1 2 1

75 D 3 3 2 3

57 R 4 3 4 4

"

voter<-read.table(text=votertxt, stringsAsFactors=FALSE)

1. Change the name of your variables to “AGE”, “PARTY”, “Q1”, “Q2”, “Q3, “Q4”.

colnames(voter)<-c("AGE", "PARTY", "Q1", "Q2", "Q3", "Q4")

1. For PARTY, convert D to Democrat, R to Republican, as factor.

voter$PARTY[voter$PARTY=="D"]<-'Democrat'

voter$PARTY[voter$PARTY=="R"]<-'Republician'

voter$PARTY<-factor(voter$PARTY)

1. For Q1 – Q4, convert 1 to Strongly Disagree, 2 to Disagree, 3 to No Opinion, 4 to Agree, 5 to Strongly Agree, as ordinal.

for (i in 1:4){

voter[,2+i][voter[,2+i]==1]<-'Strongly Disagree'

voter[,2+i][voter[,2+i]==2]<-'Disagree'

voter[,2+i][voter[,2+i]==3]<-'No Opinion'

voter[,2+i][voter[,2+i]==4]<-'Agree'

voter[,2+i][voter[,2+i]==5]<-'Strongly Agree'

voter[,2+i]<-factor(voter[,2+i],ordered=T,

levels=c('Strongly Disagree','Disagree','No Opinion','Agree','Strongly Agree'))

}

1. For AGE, use two different ways to convert numerical values to categorical (factor): [0,55] (55,75] (75,Inf].

#method1

agecat<-NA

agecat[voter$AGE>75]<-"(75,Inf]"

agecat[voter$AGE>55 & voter$AGE<=75]<-"(55,75]"

agecat[voter$AGE<=55]<-"[0,55]"

voter$AGE <-factor(agecat)

#method2

brks<-c(0,55,75,Inf)

age.range<-cut(voter$AGE,breaks=brks,include.lowest = TRUE)

voter$AGE=factor(age.range,levels=c("[0,55]","(55,75]","(75,Inf]"))