Exercise 1

1. Find the pdf image of the NCS survey booklet and go to page 134 (that is the pdf page – the actual survey page in the print version is page 130). You will see that this page has two sections – Technology and Lifestyles and the Internet. Print this page out and put it in your Appendix.

Technology and Lifestyles and the Internet page in **Appendix**

1. Locate the question in pdf image the Technology section labeled “I’m always the first among my friends to have the latest in electronic equipment”. Then find the associated question in the excel data dictionary. Use this information to read in the five variables that correspond to the five responses for this question. Do a frequency for each of the five variables. If there are missing values, you should turn those values into something numeric and useful (e.g. zeroes).

I ran frequency procedure, for each of the five variables that correspond to the five responses for this question, to check if they are any missing values for each of the responses

| **first\_agree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 34 | 100.00 | 34 | 100.00 |
| **Frequency Missing = 966** | | | | |

| **first\_agree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 71 | 100.00 | 71 | 100.00 |
| **Frequency Missing = 929** | | | | |

| **first\_neither** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 216 | 100.00 | 216 | 100.00 |
| **Frequency Missing = 784** | | | | |

| **first\_disagree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 186 | 100.00 | 186 | 100.00 |
| **Frequency Missing = 814** | | | | |

| **first\_disagree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 446 | 100.00 | 446 | 100.00 |
| **Frequency Missing = 554** | | | | |

There are missing values for every response, meaning that when at least one of the responses has been selected, the others were left empty as missing values

I have made all the missing values to zeroes so now each response has only two types of values in it, 0 or 1; all missing values turned to zeroes are highlighted in yellow

| **first\_agree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 966 | 96.60 | 966 | 96.60 |
| **1** | 34 | 3.40 | 1000 | 100.00 |

| **first\_agree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 929 | 92.90 | 929 | 92.90 |
| **1** | 71 | 7.10 | 1000 | 100.00 |

| **first\_neither** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 784 | 78.40 | 784 | 78.40 |
| **1** | 216 | 21.60 | 1000 | 100.00 |

| **first\_disagree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 814 | 81.40 | 814 | 81.40 |
| **1** | 186 | 18.60 | 1000 | 100.00 |

| **first\_disagree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 554 | 55.40 | 554 | 55.40 |
| **1** | 446 | 44.60 | 1000 | 100.00 |

1. Check to see if any respondent checked more than one check box to this question. Show me the frequencies that back up your answer to this question. If they did check more than one box, then randomly select one of the answers and modify the data to reflect that. What is your conclusion about this?

| **mysum** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 47 | 4.70 | 47 | 4.70 |
| **1** | 953 | 95.30 | 1000 | 100.00 |

mysum is the sum of the responses for this particular question, so if mysum is greater than 1 that means the respondent checked more than 1 box for that question. Since the above table has only frequencies shown for 0 and 1, we can conclude that the respondents have selected either one box or none at all but not more than one box.

1. Combine the five variables into one variable so that it is coded 5=agree a lot, 4= agree a little, 3 = neither agree nor disagree, 2 = disagree a little 1 = disagree a lot. Do a frequency of that variable. Then find a way to validate that your combining of the five variables went ok and you didn’t mess it up somewhere.

The five variables are combined into one single variable, named “first”, of 5 categories. When I ran a frequency procedure, the frequencies of each of these 5 categories, shown in the below table, match with the frequencies of the 5 individual variables, shown in the tables following it. This concludes that everything was captured correctly while combining.

| **first** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 446 | 46.80 | 446 | 46.80 |
| **2** | 186 | 19.52 | 632 | 66.32 |
| **3** | 216 | 22.67 | 848 | 88.98 |
| **4** | 71 | 7.45 | 919 | 96.43 |
| **5** | 34 | 3.57 | 953 | 100.00 |
| **Frequency Missing = 47** | | | | |

| **first\_disagree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 554 | 55.40 | 554 | 55.40 |
| **1** | 446 | 44.60 | 1000 | 100.00 |

| **first\_disagree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 814 | 81.40 | 814 | 81.40 |
| **1** | 186 | 18.60 | 1000 | 100.00 |

| **first\_neither** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 784 | 78.40 | 784 | 78.40 |
| **1** | 216 | 21.60 | 1000 | 100.00 |

| **first\_agree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 929 | 92.90 | 929 | 92.90 |
| **1** | 71 | 7.10 | 1000 | 100.00 |

| **first\_agree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 966 | 96.60 | 966 | 96.60 |
| **1** | 34 | 3.40 | 1000 | 100.00 |

1. Find another question in the Lifestyles and the Internet section that suits your fancy. Wash, rinse and repeat steps 1 through 4 above for this variable.

In the Lifestyles and the Internet section, I choose the question, “**When I need information the first place I look is the internet**” highlighted in the image in **Appendix**

1. Find the pdf image of the NCS survey booklet and go to page 134 (that is the pdf page – the actual survey page in the print version is page 130). You will see that this page has two sections – Technology and Lifestyles and the Internet. Print this page out and put it in your Appendix.

Lifestyles and the Internet page in Appendix

1. Locate the question in pdf image the Technology section labeled “I’m always the first among my friends to have the latest in electronic equipment”. Then find the associated question in the excel data dictionary. Use this information to read in the five variables that correspond to the five responses for this question. Do a frequency for each of the five variables. If there are missing values, you should turn those values into something numeric and useful (e.g. zeroes).

I ran frequency procedure, for each of the five variables that correspond to the five responses for this question, to check if they are any missing values for each of the responses

| **inlife\_agree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 270 | 100.00 | 270 | 100.00 |
| **Frequency Missing = 730** | | | | |

| **inlife\_agree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 249 | 100.00 | 249 | 100.00 |
| **Frequency Missing = 751** | | | | |

| **inlife\_neither** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 153 | 100.00 | 153 | 100.00 |
| **Frequency Missing = 847** | | | | |

| **inlife\_disagree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 78 | 100.00 | 78 | 100.00 |
| **Frequency Missing = 922** | | | | |

| **inlife\_disagree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 185 | 100.00 | 185 | 100.00 |
| **Frequency Missing = 815** | | | | |

There are missing values for every response, meaning that when at least one of the responses has been selected, the others were left empty as missing values.

I have made all the missing values to zeroes so now each response has only two types of values in it; 0 or 1, all missing values turned to zeroes are highlighted in yellow

| **inlife\_agree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 730 | 73.00 | 730 | 73.00 |
| **1** | 270 | 27.00 | 1000 | 100.00 |

| **inlife\_agree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 751 | 75.10 | 751 | 75.10 |
| **1** | 249 | 24.90 | 1000 | 100.00 |

| **inlife\_neither** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 847 | 84.70 | 847 | 84.70 |
| **1** | 153 | 15.30 | 1000 | 100.00 |

| **inlife\_disagree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 922 | 92.20 | 922 | 92.20 |
| **1** | 78 | 7.80 | 1000 | 100.00 |

| **inlife\_disagree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 815 | 81.50 | 815 | 81.50 |
| **1** | 185 | 18.50 | 1000 | 100.00 |

1. Check to see if any respondent checked more than one check box to this question. Show me the frequencies that back up your answer to this question. If they did check more than one box, then randomly select one of the answers and modify the data to reflect that. What is your conclusion about this?

| **mysum** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 65 | 6.50 | 65 | 6.50 |
| **1** | 935 | 93.50 | 1000 | 100.00 |

mysum is the sum of the responses for this particular question, so if mysum is greater than 1 that means the respondent checked more than 1 box for that question. Since the above table has only frequencies shown for 0 and 1, we can conclude that the respondents have selected either one box or none at all but not more than one box.

1. Combine the five variables into one variable so that it is coded 5=agree a lot, 4= agree a little, 3 = neither agree nor disagree, 2 = disagree a little 1 = disagree a lot. Do a frequency of that variable. Then find a way to validate that your combining of the five variables went ok and you didn’t mess it up somewhere.

The five variables are combined into one single variable, named “inlife”, of 5 categories. When I ran a frequency procedure, the frequencies of each of these 5 categories, shown in the below table, match with the frequencies of the 5 individual variables, shown in the tables following it. This concludes that everything was captured correctly while combining.

| **inlife** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **1** | 185 | 19.79 | 185 | 19.79 |
| **2** | 78 | 8.34 | 263 | 28.13 |
| **3** | 153 | 16.36 | 416 | 44.49 |
| **4** | 249 | 26.63 | 665 | 71.12 |
| **5** | 270 | 28.88 | 935 | 100.00 |
| **Frequency Missing = 65** | | | | |

| **inlife\_disagree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 815 | 81.50 | 815 | 81.50 |
| **1** | 185 | 18.50 | 1000 | 100.00 |

| **inlife\_disagree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 922 | 92.20 | 922 | 92.20 |
| **1** | 78 | 7.80 | 1000 | 100.00 |

| **inlife\_neither** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 847 | 84.70 | 847 | 84.70 |
| **1** | 153 | 15.30 | 1000 | 100.00 |

| **inlife\_agree\_little** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 751 | 75.10 | 751 | 75.10 |
| **1** | 249 | 24.90 | 1000 | 100.00 |

| **inlife\_agree\_alot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 730 | 73.00 | 730 | 73.00 |
| **1** | 270 | 27.00 | 1000 | 100.00 |

**APPENDIX**

Code:

/\* create a library "mylib" \*/

libname mylib '\\client\C$\Users\prady\_000\Documents\Vishwa\MSDA\practicum1\sas data files';

/\* Import "first1k" dataset containing 1000 records into a new dataset "smallrec" \*/

filename smallrec '\\client\C$\Users\prady\_000\Documents\Vishwa\MSDA\practicum1\first1k.txt' lrecl =**65576**;

/\* create "mytemp", that copies data from smallrec but for only 6 variables, selectively, by specifying the positions from where

the data should be extracted\*/

**data** mytemp;

infile smallrec;

input my\_id **1**-**7** /\* ID field to uniquely identify each of the (1000) records \*/

first\_agree\_alot **6825**

first\_agree\_little **6842**

first\_neither **6876**

first\_disagree\_little **6893**

first\_disagree\_alot **6910**;

**run**;

/\* Run frequency tables for each of the five variables by using their information from mytemp file in temporary work library \*/

**proc** **freq** data = work.mytemp;

tables

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

**run**;

/\* create another dataset "mycalcs" that copies mytemp \*/

**data** mycalcs;

set mytemp;

/\* create an 2-D array of size 1X5 for one question and its 5 available options \*/

array missy(**1**,**5**)

first\_agree\_alot

first\_agree\_little

first\_neither

first\_disagree\_little

first\_disagree\_alot;

/\* loop through the array to check for missing values in any of the 5 cells (for each of 1000 records implicitly) and make it zero when found, else, do nothing\*/

do j = **1** to **5**;

if missy(**1**,j) = **.** then missy(**1**,j) = **0**;

end;

/\* compute the sum of each record by adding all its columns and drop the looping variable \*/

mysum = missy(**1**,**1**)+ missy(**1**,**2**)+missy(**1**,**3**) + missy(**1**,**4**) + missy(**1**,**5**);

drop j;

/\* define a variable to capture the option selected by the person when mysum is 1 else mark it as missing \*/

if mysum = **1** then

myvar = (missy(**1**,**1**)\***5**) + (missy(**1**,**2**)\***4**)+ (missy(**1**,**3**)\***3**) + (missy(**1**,**4**))\***2** + (missy(**1**,**5**)\***1**);

else

myvar = **.**;

/\* myvar has the either the option selected for the question or missing field indicating nothing is selected \*/

/\* naming it relevant to the question of interest \*/

first = myvar;

**run**;

**proc** **freq** data=mycalcs;

tables mysum; /\* the frequency table indicates they were only 1's and 0's hence no multiple options selected for same question\*/

tables first; /\* the frequency for the 5 categories of new variable match with the frequencies of the 5 variables indicating everything is captured into new variable\*/

tables first\_agree\_alot;

tables first\_agree\_little;

tables first\_neither;

tables first\_disagree\_little ;

tables first\_disagree\_alot;

**run**;

**QUESTION-5**

/\*\*\*\*\*\*\*\*\*\*\*\* LIFESTYLE/INTERNET QUESTION \*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* create "mytemp1", that copies data from smallrec but for only 6 variables, selectively, by specifying the positions from where

the data should be extracted\*/

**data** mytemp1;

infile smallrec;

input my\_id **1**-**7** /\* ID field to uniquely identify each of the 1000 records \*/

inlife\_agree\_alot **5439**

inlife\_agree\_little **5471**

inlife\_neither **5535**

inlife\_disagree\_little **5567**

inlife\_disagree\_alot **5599**;

**run**;

/\* Run frequency tables for each of the five variables by using their information from mytemp1 file in temporary work library \*/

**proc** **freq** data = work.mytemp1;

tables

inlife\_agree\_alot

inlife\_agree\_little

inlife\_neither

inlife\_disagree\_little

inlife\_disagree\_alot;

**run**;

/\* create another dataset "mycalcs1" that copies mytemp1 \*/

**data** mycalcs1;

set mytemp1;

/\* create a 2-D array of size 1X5 for one question and its 5 available options \*/

array missy(**1**,**5**)

inlife\_agree\_alot

inlife\_agree\_little

inlife\_neither

inlife\_disagree\_little

inlife\_disagree\_alot;

/\* loop through the array to check for missing values in any of the 5 cells (for each of 1000 records implicitly) and make it zero when found, else, do nothing\*/

do j = **1** to **5**;

if missy(**1**,j) = **.** then missy(**1**,j) = **0**;

end;

/\* compute the sum of each record by adding all its columns, drop the looping variables \*/

mysum = missy(**1**,**1**)+ missy(**1**,**2**)+missy(**1**,**3**) + missy(**1**,**4**) + missy(**1**,**5**);

drop j;

/\* define a variable to capture the option selected by the person when mysum is 1 else mark it as missing \*/

if mysum = **1** then

myvar = (missy(**1**,**1**)\***5**) + (missy(**1**,**2**)\***4**)+ (missy(**1**,**3**)\***3**) + (missy(**1**,**4**))\***2** + (missy(**1**,**5**)\***1**);

else

myvar = **.**;

/\* myvar has the either the option selected for the question or missing field indicating nothing is selected for it\*/

/\* naming it relevant to the question of interest \*/

inlife = myvar;

**run**;

**proc** **freq** data=mycalcs1;

tables mysum; /\* the frequency table indicates they were only 1's and 0's counts hence no multiple options selected for same question\*/

tables inlife; /\* the frequency for the 5 categories of new variable match with the frequencies of the 5 variables indicating that everything is captured correctly \*/

tables inlife\_agree\_alot;

tables inlife\_agree\_little;

tables inlife\_neither;

tables inlife\_disagree\_little ;

tables inlife\_disagree\_alot;

**run**;

