Multivariate Analysis of Heart Risk Factors

Bill Qualls

Executive Summary

The purpose of the present study is to determine if trivial demographic / lifestyle data such as age, weight, gender, exercise, education, and use of tobacco can be used to assess an individual's heart risk. Data for 100 individuals was used in this study. Three different online heart risk calculators were used to derive three different heart risk scores for each individual. Several multivariate analysis techniques were used throughout the studying including Canonical Correlation Analysis, K-Means Clustering, Principal Component Analysis, and Linear Discriminant Analysis. The findings were as follows:

- The good news is that trivial demographic / lifestyle data can be used to assess an individual's heart risk.
- The bad news is that the two most significant contributors to heart risk are two things you cannot do anything about: age and gender.
- The study was inconclusive about the impact of the use of tobacco on heart risk.

Introduction

The purpose of this study is to determine if trivial demographic / lifestyle data can be used to assess an individual's heart risk. The data used for the analysis consists of demographic / lifestyle data (independent variables) and heart risk scores (dependent variables). The former came from Allan G. Bluman, Elementary Statistics - A Step by Step Approach, Wm. C. Brown Publishers (1992), and is described in detail in Appendix A. The latter was derived by entering selected data into each of three separate online heart risk calculators. Links to these calculator, and screenshots, can be found in Appendix B. The complete data set can be found in Appendix C. Here are the first five observations:

```
01 27 2 1 1 120 193 126 118 136 F M 27 -9  0.10
02 18 1 0 1 145 210 120 105 137 M S 50 -2  0.20
03 32 2 0 0 118 196 128 115 135 F M 39 -7  1.10
04 24 2 0 1 162 208 129 108 142 M M 50  2  0.30
05 19 1 2 0 106 188 119 106 133 F S 27 -9  0.10
```

The space-delimited data was imported into R Studio (see Appendix D). All analysis was done using R Studio.

Exploratory Data Analysis

Exploratory data analysis in the form of histograms and bar charts was performed on the data (see Appendix E). No anomalies were found, but based on this analysis several dummy variables were created:

- Based on the analysis of the education level variable (EDUC), a dummy variable COLLEGE was created where a value of 1 indicates College Degree or Grad Degree.

- Based on the analysis of the level of exercise variable (EXER), a dummy variable REALEXER was created where a value of 1 indicates moderate or heavy exercise.
- The SEX variable was coded as "M" or "F": a dummy variable MALE was created where a value of 1 indicates male.
- The MARITAL variable was coded as "M", "S", "W", or "D": a dummy variable MARRIER was created where a value of 1 indicates married.
- Based on the analysis of the level of smoking variable (SMOKE), a dummy variable ANYSMOKE was created where a value of 1 indicates any smoking, regardless of amount.

Models

As already stated, the purpose of this study is to determine if trivial demographic / lifestyle data can be used to assess an individual's heart risk. The models used herein will consist of several multivariate analysis techniques: Canonical Correlation Analysis, K-Means Clustering, Principal Component Analysis, and Linear Discriminant Analysis.

Canonical Correlation Analysis (CCA) can be used to measure the strength of the relationship between two sets of variables, presumably independent (x) variables and dependent (y) variables. This might sound like multiple linear regression: it is. The difference is multiple linear regression will have many x (predictor) variables and a single y (predicted) variable, whereas PCA will have many x variables and many y variables. (Good, concise information on CCA can be found at http://www.ats.ucla.edu/stat/r/dae/canonical.htm.)

➔ The present study will use CCA to measure the strength of the relationship between demographic / lifestyle variables (x's) and the outputs from three different heart risk calculators (y's). CCA will not be used for predictive purposes, only to measure the strength of the relationship.

K-Means Clustering is a technique which groups observations together. The analyst specifies the number of group's desired (k), and an iterative method is used to assign and reassign observations to groups based on their distance from (usually) the center of each group. The process continues until there is no change in group assignment.

➔ The present study will k-means clustering to assign each observation to one of three heart risk groups: high, medium, or low. There need be no mystery as to why three groups were chosen: high, medium, or low just make sense in the current context.

Principal Component Analysis (PCA) is one of many dimensionality reduction techniques. It is used to group variables into variates; that is, one or more variables can be combined into a single variate such that a simpler model is derived. Imagine two dimensions – blond hair and blue eyes – being reduced to a single dimension of blond hair with blue eyes. As a result of this dimensionality reduction, it is possible to produce a two-dimensional representation of more than two dimensions.

➔ The present study will use PCA to measure identify two variates (principal components) which capture the maximum variability in the original data. These two variates will be plotted on an XY axis, with each data point represented by the heart risk cluster that data

point was assigned to. This should provide visual confirmation of the models to date; that is, that trivial demographic / lifestyle data can indeed be used to assess an individual's heart risk.

Linear Discriminant Analysis (LDA) is a classification technique. As such it is also known as a supervised learning technique. The plan is to split the data into two data sets: one for training the model, and the other for testing the model. The training data is used to build the model, and the testing data is used to validate the model. Each row of the training data already has an assigned risk group (actual), but the model will determine a new (predicted) group for that record. A confusion matrix is also created, which summarizes the differences between the actual group and predicted group for each observation.

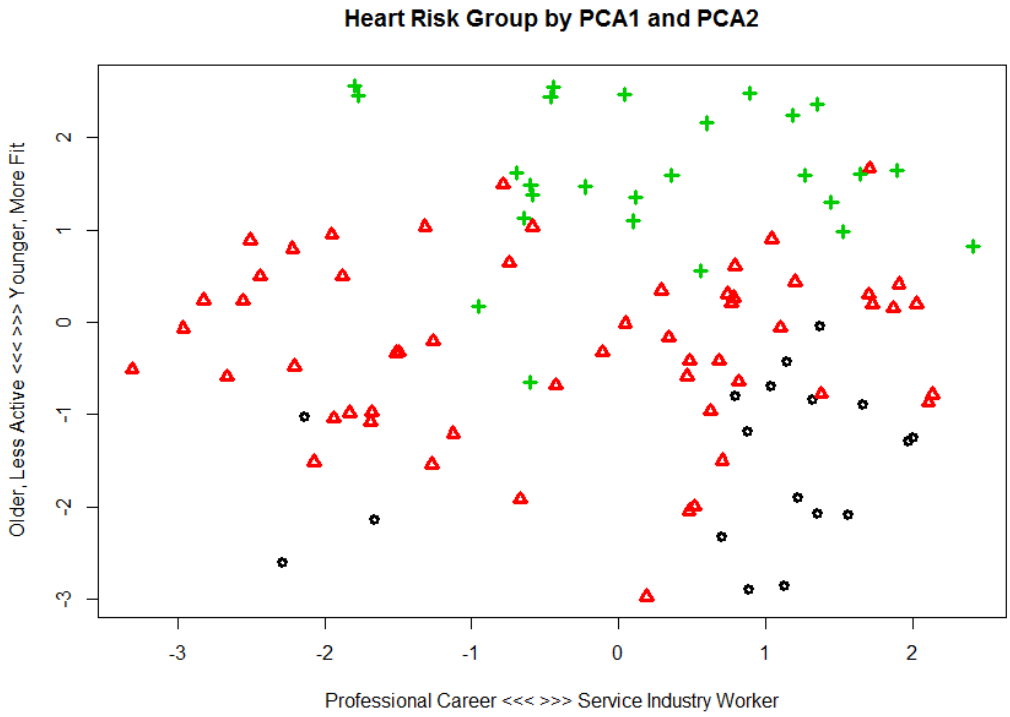
➔ The present study will use LDA to create a classification model which will assign each observation to a heart risk cluster (high / medium / low) based on non-medical demographic and lifestyle data. Whereas the models already mentioned will suggest that such a classification is doable, this one will actually do it!

Findings

On the one hand, the results of the Canonical Correlation Analysis (see Appendix F) were encouraging. The canonical correlation coefficient, R_c^2, was very high for the first two variates. The results suggest that there is, indeed a strong relationship between demographic / lifestyle data and heart risk. On the other hand, the strongest variables were age and gender, so the best thing you can do to lower your heart risk is stay young and stay female!

K-Means Clustering (see Appendix G) resulted in 18 observations being rated as high risk, 56 observations being rated as medium risk, and 26 observations being rated as low risk. (Note if this analysis were run again, these number might change slightly due to the use of pseudorandom numbers in the k-means algorithm: the algorithm is not deterministic.) Those group sizes seem reasonable, with high risk and low risk each occupying approximately 25% of the observations and the medium risk occupying approximately the middle half (χ^2=2.720, p=0.437).

Principal Component Analysis (see Appendix H) resulted in two variates which explained fifty percent of the variation in the data. The trick with PCA is to come up with "names" for the variates, which are a combination (sometimes nonsensical) of several variables. Given that the variables assigned to the first PCA component were highly negative for IQ and COLLEGE, and (slightly) positive for AGE and ANYSMOKE, this component was assigned the label "Service Industry Worker" at one extreme and "Professional / Career" at the other extreme. Given that the variables assigned to the second PCA component were highly negative in AGE and WEIGHT, and positive in REALEXER, this component was assigned the label "Yong and Fit" and one extreme and "Older / Less Active" at the other extreme. With the first component on the x-axis and the second component on the y-axis, each data point was plotted using a symbol indicating its heart risk group. The graphic shows some interesting clustering which is consistent with our earlier findings. Young people have lower risk. As people age, their risk increases.



Linear Discriminant Analysis (see Appendix I) resulted in a model which assigned observations to a heart risk group based on simple demographic / lifestyle data. Whereas (some of) the heart risk calculators used systolic blood pressure and total cholesterol to arrive at their scores, the LDA model did not. The LDA model did include weight. Only one of the twenty eight observations in the testing data was misclassified: it was classified as medium risk and should have been low risk.

```
> confusionMatrix
      pred
actual  1  2  3
     1  8  0  0
     2  0 13  0
     3  0  1  6
```

Further Study

There are opportunities for further study. These findings were inconclusive about impact of the use of tobacco on heart risk scores. Specifically, the results of the CCA and PCA were somewhat conflicting in this study. This could, however, be a result of misinterpreting the CCA and PCA results, which is so easy to do with any of the dimensionality reduction techniques.

While weight was included in the data, height was not, and one must assume that there is an interaction effect with weight and height or weight and gender (that is, 180 pounds might be a good weight if you are a six foot tall male, but an unsafe weight if you are a five foot tall female.)

Ethnicity and family heart health history are commonly recognized heart risk factors, but those data were unavailable.

Appendix A – Description of the data – Part 1

About the original data…

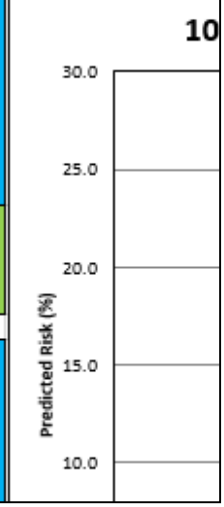
```
This is BLUMAN.TXT, Description of BLUMAN.DAT -
Source:  Allan G. Bluman, Elementary Statistics - A Step
    by Step Approach, Wm. C. Brown Publishers, 1992

Variables:                              Columns     Name

    1.  ID Number                         1-2       ID
    2.  Age (years)                       4-5       AGE
    3.  Educational level                 7-7       EDUC
        0 = no high school degree
        1 = high school graduate
        2 = college graduate
        3 = graduate degree
    4.  Smoking status                    9-9       SMOKE
        0 = does not smoke
        1 = less than one pack per day
        2 = one or more packs per day
    5.  Exercise                         11-11      EXER
        0 = none
        1 = light
        2 = moderate
        3 = heavy
    6.  Weight (pounds)                  13-15      WEIGHT
    7.  Serum Cholestrol in              17-19      CHOLEST
        milligram percent (mg%)
    8.  Systolic Pressure in             21-23      SYSTOL
        millimeters of mercury (mmHg)
    9.  IQ test score                    25-27      IQ
    10. Sodium in milliequivalents       29-31      SODIUM
        per liter (mEq/l)
    11. Sex (M/F)                        33-33      SEX
    12. Marital Status                   35-35      MARITAL
        M = Married
        S = Single
        W = Widowed
        D = Divorced

There are one hundred records.
```

Appendix B – Description of the data – Part 2

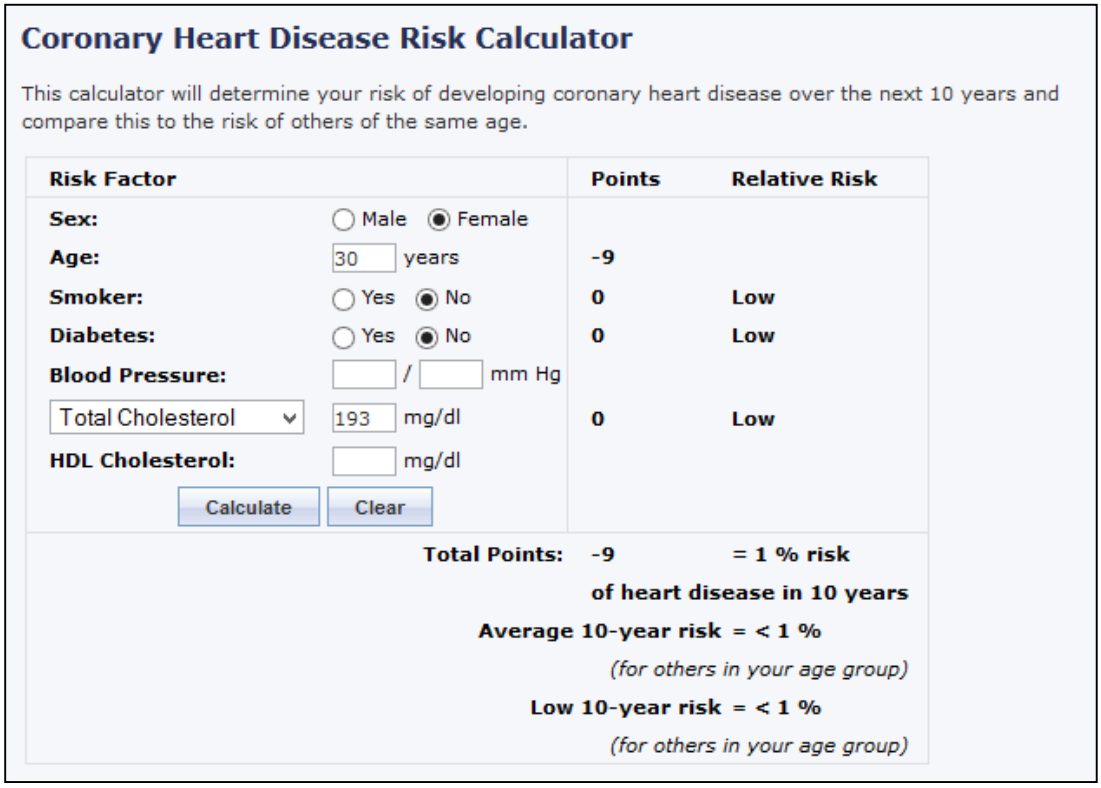
The CALC1, CALC2, and CALC3 columns are the results of entering selected fields into three different online heart risk calculators. Where necessary, ages were raised to the minimum required by the calculator. Where smoking was asked, the "moderate" level was selected. Where ethnicity was asked, "white" was selected. Where diabetes was asked, "no" was selected.

Source of CALC1:
http://my.americanheart.org/professional/StatementsGuidelines/PreventionGuidelines/Prevention-Guidelines_UCM_457698_SubHomePage.jsp

	A	B	C
1			Enter patient values in this
2	Risk Factor	Units	Value
3	Sex	M (for males) or F (for females)	F
4	Age	years	27
5	Race	AA (for African Americans) or WH (for whites or others)	WH
6	Total Cholesterol	mg/dL	193
7	HDL-Cholesterol	mg/dL	
8	Systolic Blood Pressure	mm Hg	126
9	Treatment for High Blood Pressure	Y (for yes) or N (for no)	N
10	Diabetes	Y (for yes) or N (for no)	N
11	Smoker	Y (for yes) or N (for no)	N
12			
13	Your 10-Year ASCVD Risk (%)	This calculator only provides 10-year risk estimates for individuals 40 to 79 years of age Enter 20-100 for HDL value	
14	10-Year ASCVD Risk (%) for Someone Your Age with Optimal Risk Factor Levels (shown above in column E)	This calculator only provides 10-year risk estimates for individuals 40 to 79 years of age	
15			
	Your Lifetime ASCVD Risk* (%)	27.0	



Source of CALC2:
http://www.mcw.edu/calculators/Coronary-Heart-Disease-Risk.htm



Coronary Heart Disease Risk Calculator

This calculator will determine your risk of developing coronary heart disease over the next 10 years and compare this to the risk of others of the same age.

Risk Factor		Points	Relative Risk
Sex:	○ Male ◉ Female		
Age:	30 years	-9	
Smoker:	○ Yes ◉ No	0	Low
Diabetes:	○ Yes ◉ No	0	Low
Blood Pressure:	/ mm Hg		
Total Cholesterol	193 mg/dl	0	Low
HDL Cholesterol:	mg/dl		

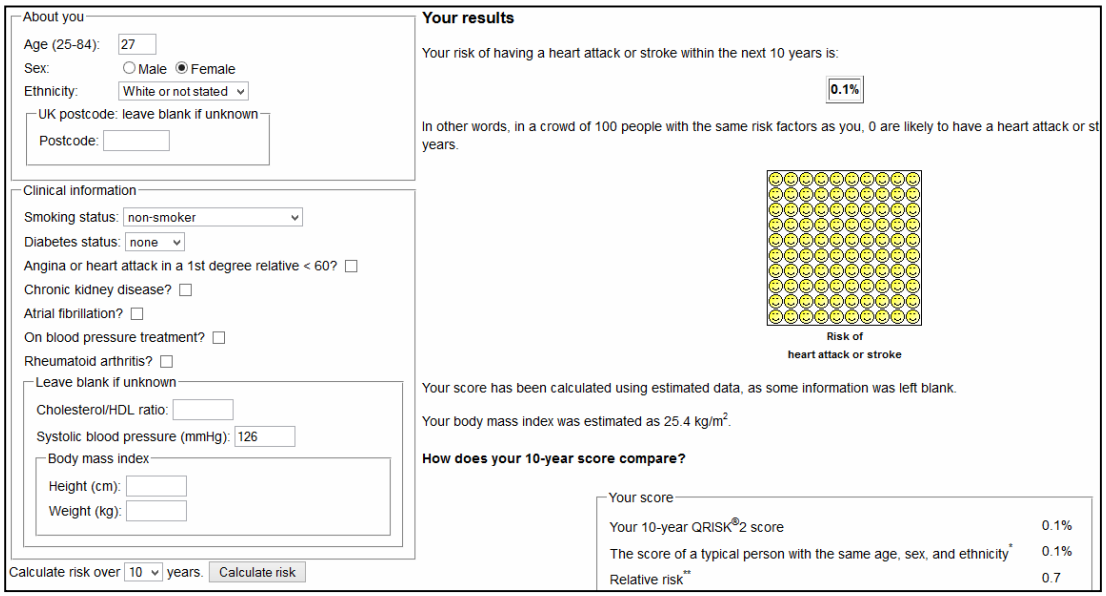
Calculate | Clear

Total Points: -9 **= 1 % risk of heart disease in 10 years**

Average 10-year risk = < 1 % *(for others in your age group)*

Low 10-year risk = < 1 % *(for others in your age group)*

Source of CALC3:
http://www.qrisk.org/index.php



About you
Age (25-84): 27
Sex: ○ Male ◉ Female
Ethnicity: White or not stated
UK postcode: leave blank if unknown
Postcode:

Clinical information
Smoking status: non-smoker
Diabetes status: none
Angina or heart attack in a 1st degree relative < 60? ☐
Chronic kidney disease? ☐
Atrial fibrillation? ☐
On blood pressure treatment? ☐
Rheumatoid arthritis? ☐
Leave blank if unknown
Cholesterol/HDL ratio:
Systolic blood pressure (mmHg): 126
Body mass index
Height (cm):
Weight (kg):

Calculate risk over 10 years. Calculate risk

Your results

Your risk of having a heart attack or stroke within the next 10 years is:

0.1%

In other words, in a crowd of 100 people with the same risk factors as you, 0 are likely to have a heart attack or st years.

Risk of heart attack or stroke

Your score has been calculated using estimated data, as some information was left blank.

Your body mass index was estimated as 25.4 kg/m^2.

How does your 10-year score compare?

Your score

Your 10-year QRISK®2 score	0.1%
The score of a typical person with the same age, sex, and ethnicity*	0.1%
Relative risk**	0.7

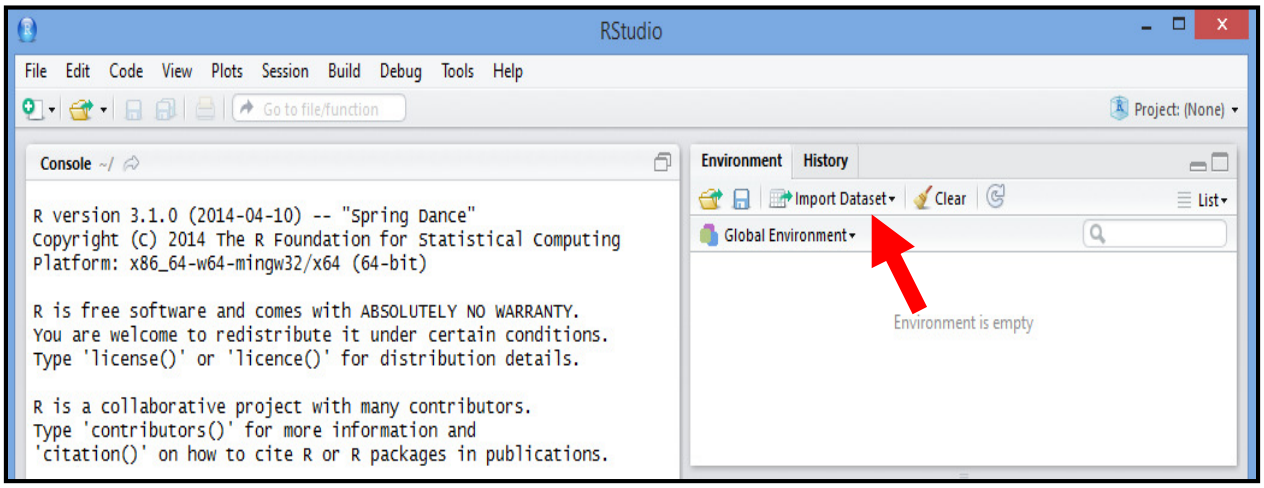
Appendix C – The Data

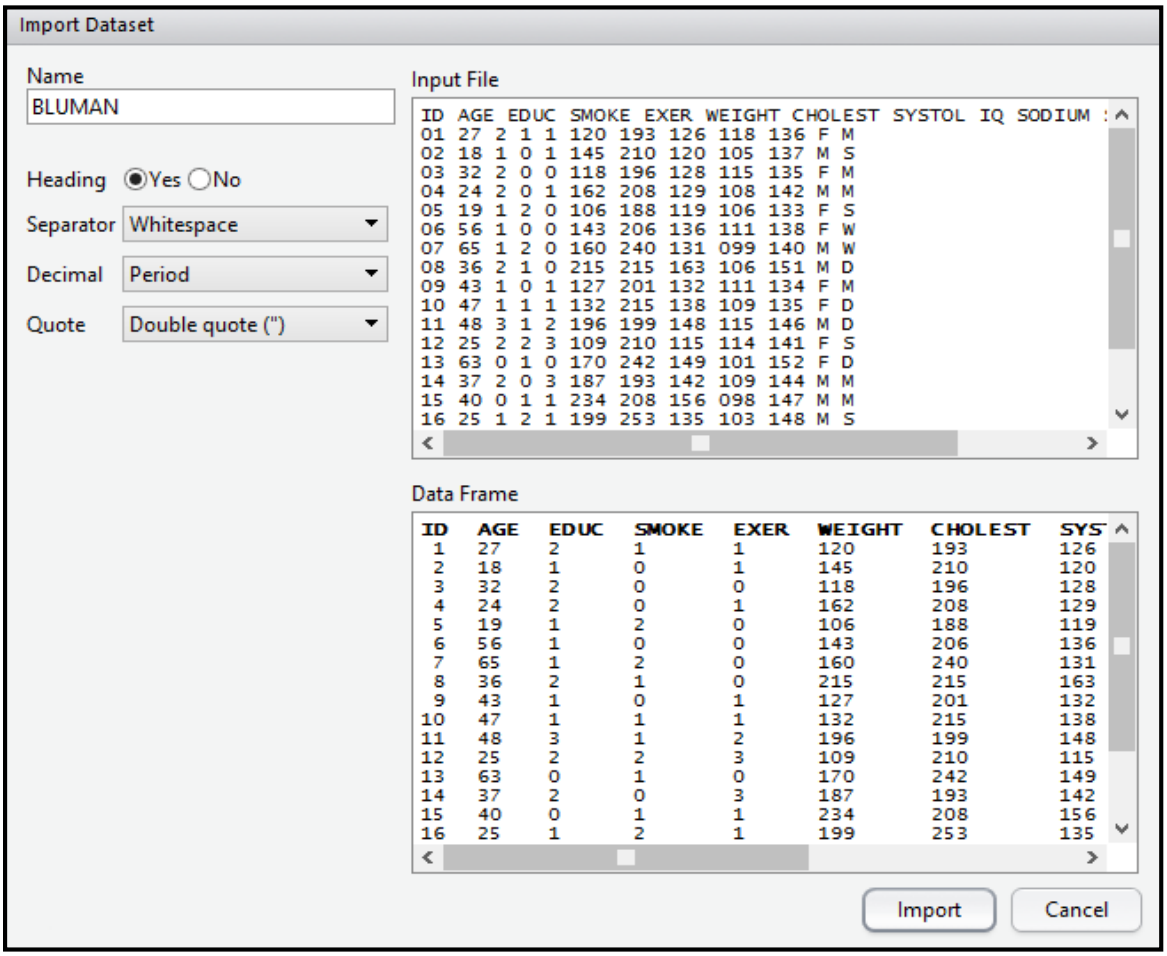
ID	AGE	EDUC	SMOKE	EXER	WEIGHT	CHOLEST	SYSTOL	IQ	SODIUM	SEX	MARITAL	CALC1	CALC2	CALC3
01	27	2	1	1	120	193	126	118	136	F	M	27	-9	0.10
02	18	1	0	1	145	210	120	105	137	M	S	50	-2	0.20
03	32	2	0	0	118	196	128	115	135	F	M	39	-7	1.10
04	24	2	0	1	162	208	129	108	142	M	M	50	2	0.30
05	19	1	2	0	106	188	119	106	133	F	S	27	-9	0.10
06	56	1	0	0	143	206	136	111	138	F	W	39	10	7.50
07	65	1	2	0	160	240	131	99	140	M	W	50	8	13.80
08	36	2	1	0	215	215	163	106	151	M	D	50	1	1.40
09	43	1	0	1	127	201	132	111	134	F	M	39	3	2.40
10	47	1	1	1	132	215	138	109	135	F	D	39	6	1.70
11	48	3	1	2	196	199	148	115	146	M	D	46	2	4.40
12	25	2	2	3	109	210	115	114	141	F	S	39	-8	0.10
13	63	0	1	0	170	242	149	101	152	F	D	39	9	7.80
14	37	2	0	3	187	193	142	109	144	M	M	50	2	2.70
15	40	0	1	1	234	208	156	98	147	M	M	46	2	2.10
16	25	1	2	1	199	253	135	103	148	M	S	50	1	0.10
17	72	0	0	0	143	288	156	103	145	F	M	50	13	26.30
18	56	1	1	0	156	164	153	99	144	F	D	39	7	4.50
19	37	2	0	2	142	214	122	110	135	M	M	50	3	2.20
20	41	1	1	1	123	220	142	108	134	F	M	39	1	1.00
21	33	2	1	1	165	194	122	112	137	M	S	36	-9	0.50
22	52	1	0	1	157	205	119	106	134	M	D	50	6	8.80
23	44	2	0	1	121	223	135	116	133	F	M	39	3	2.80
24	53	1	0	0	131	199	133	121	136	F	M	39	8	5.80
25	19	1	0	3	128	206	118	122	132	M	S	50	2	0.20
26	25	1	0	0	143	200	118	103	135	M	M	50	2	0.20
27	31	2	1	1	152	204	120	119	136	M	M	46	0	0.30
28	28	2	0	0	119	203	118	116	138	F	M	39	-6	0.40
29	23	1	0	0	111	240	120	105	135	F	S	50	-6	0.30
30	47	2	1	0	149	199	132	123	136	F	M	27	3	1.60
31	47	2	1	0	179	235	131	113	139	M	M	46	3	3.50
32	59	1	2	0	206	260	151	99	143	M	W	50	6	10.50
33	36	2	1	0	191	201	148	118	145	M	D	46	1	1.10
34	59	0	1	1	156	235	142	100	132	F	W	39	8	5.30
35	35	1	0	0	122	232	131	106	135	F	M	39	-1	1.10
36	29	2	0	2	175	195	129	121	148	M	M	50	1	0.70
37	43	3	0	3	194	211	138	129	146	M	M	50	4	4.90
38	44	1	2	0	132	240	130	109	132	F	S	39	1	1.20
39	63	2	2	1	188	255	156	121	145	M	M	50	7	8.10
40	36	2	1	1	125	220	126	117	140	F	S	39	-3	0.90
41	21	1	0	1	109	206	114	102	136	F	M	39	-6	0.30
42	31	2	0	2	112	201	116	123	133	F	M	39	-6	0.60
43	57	1	1	1	167	213	141	103	143	M	W	46	5	8.50
44	20	1	2	3	101	194	110	111	125	F	S	27	-9	0.10
45	24	2	1	3	106	188	113	114	127	F	D	27	-9	0.10
46	42	1	0	1	148	206	136	107	140	M	S	50	4	4.40
47	55	1	0	0	170	257	152	106	130	F	M	50	10	8.00
48	23	0	0	1	152	204	116	95	142	M	M	50	2	0.20
49	32	2	0	0	191	210	132	115	147	M	M	50	2	1.20
50	28	1	0	1	148	222	135	100	135	M	M	50	2	0.60

```
51 67 0 0 0 160 250 141 116 146 F W 50 11 17.60
52 22 1 1 1 109 220 121 103 144 F M 39 -8  0.10
53 19 1 1 1 131 231 117 112 133 M S 46  0  0.10
54 25 2 0 2 153 212 121 119 149 M D 50  2  0.20
55 41 3 2 2 165 236 130 131 152 M M 46  2  1.80
56 24 2 0 3 112 205 118 100 132 F S 39 -6  0.30
57 32 2 0 1 115 187 115 109 136 F S 39 -7  0.60
58 50 3 0 1 173 203 136 126 146 M M 50  6  8.60
59 32 2 1 0 186 248 119 122 149 M M 50  1  0.40
60 26 2 0 1 181 207 123 121 142 M S 50  2  0.30
61 36 1 1 0 112 188 117  98 135 F D 27 -4  0.40
62 40 1 1 0 130 201 121 105 136 F D 39  1  0.70
63 19 1 1 1 132 237 115 111 137 M S 46  0  0.10
64 37 2 0 2 179 228 141 127 141 F M 39 -1  1.60
65 65 3 2 1 212 220 158 129 148 M M 46  9 15.60
66 21 1 2 2  99 191 117 103 131 F S 27 -9  0.10
67 25 2 2 1 128 195 120 121 131 F S 27 -9  0.10
68 68 0 0 0 167 210 142  98 140 M W 50  9 26.80
69 18 1 1 2 121 198 123 113 136 F S 27 -9  0.10
70 26 0 1 1 163 235 128  99 140 M M 46  0  0.10
71 45 1 1 1 185 229 125 101 143 M M 46  3  2.70
72 44 3 0 0 130 215 128 128 137 F M 39  3  2.50
73 50 1 0 0 142 232 135 104 138 F M 39  9  4.60
74 63 0 0 0 166 271 143 103 147 F W 50 11 13.50
75 48 1 0 3 163 203 131 103 144 M M 50  5  7.10
76 27 2 0 3 147 186 118 114 134 M M 50  1  0.40
77 31 1 1 1 152 228 116 126 138 M D 46  0  0.30
78 28 2 0 2 112 197 120 123 133 F M 39 -7  0.40
79 36 2 1 2 190 226 123 121 147 M M 46  1  0.80
80 43 3 2 0 179 252 127 131 145 M D 50  3  2.20
81 21 1 0 1 117 185 116 105 137 F S 39 -7  0.30
82 32 2 1 0 125 193 123 119 135 F M 27 -9  0.30
83 29 2 1 0 123 192 131 116 131 F D 27 -9  0.20
84 49 2 2 1 185 190 129 127 144 M M 36  2  4.10
85 24 1 1 1 133 237 121 114 129 M M 46  0  0.10
86 36 2 0 2 163 195 115 119 139 M M 50  2  1.70
87 34 1 2 0 135 199 133 117 135 F M 27 -9  0.40
88 36 0 0 1 142 216 138  88 137 F M 39 -1  1.40
89 29 1 1 1 155 214 120  98 135 M S 46  0  0.20
90 42 0 0 2 169 201 123  96 137 M D 50  4  3.90
91 41 1 1 1 136 214 133 102 141 F D 39  1  0.90
92 29 1 1 0 112 205 120 102 130 F M 39 -8  0.20
93 43 1 1 0 185 208 127 100 143 M M 50  2  2.20
94 61 1 2 0 173 248 142 101 141 M M 50  7 11.30
95 21 1 1 3 106 210 111 105 131 F S 39 -8  0.10
96 56 0 0 0 149 232 142 103 141 F M 39 10  7.90
97 63 0 1 0 192 193 163  95 147 M M 50  5 14.30
98 74 1 0 0 162 247 151  99 151 F W 50 11 29.30
99 35 2 0 1 151 251 147 113 145 F M 50 -1  1.50
00 28 2 0 3 161 199 129 116 138 M M 50  1  0.60
```

Appendix D – Importing into R Studio

The following screen shots show how the data was imported into R Studio:





Use of the GUI to import the data as shown above results in the following generated code:

```
> BLUMAN <- read.table("C:/Users/Bill/Desktop/BQ/School/DePaul/CSC424 Advanced Data
Analysis/project/rawdata/BLUMAN.DAT", header=T, quote="\"")
>   View(BLUMAN)
```

I could also import with the following instructions. setwd = set working directory. Note use of forward slash in path, even for Windows systems.

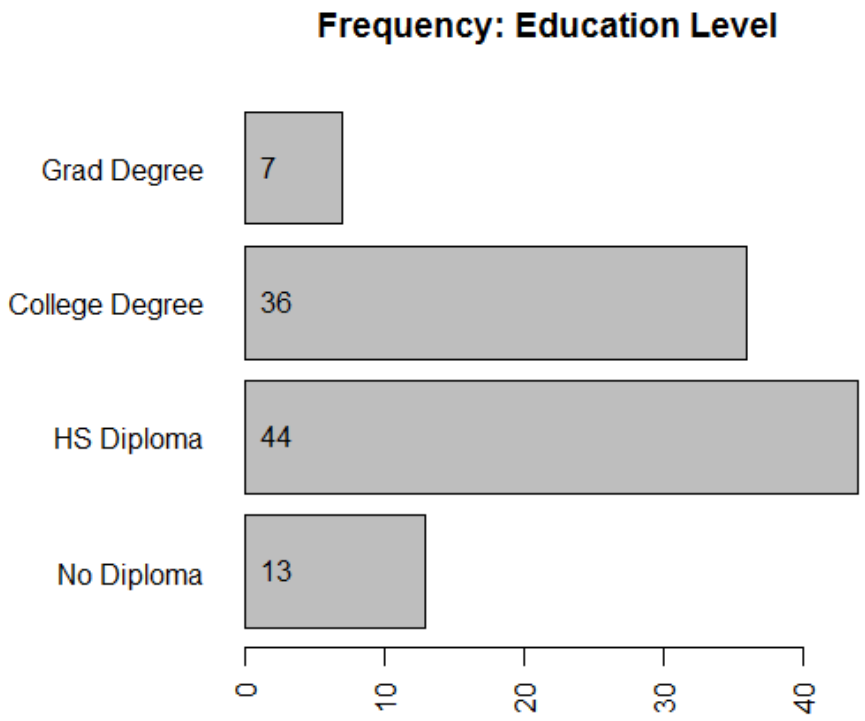
```
> setwd("C:/Users/Bill/Desktop/BQ/School/DePaul/CSC424 Advanced Data
Analysis/project/rawdata/")
> BLUMAN <- read.table("BLUMAN.DAT", header=T, quote="\"")
> View(BLUMAN)
```

The getwd() function will show the working directory.

```
> getwd()
[1] "C:/Users/Bill/Desktop/BQ/School/DePaul/CSC424 Advanced Data Analysis/project/rawdata"
```

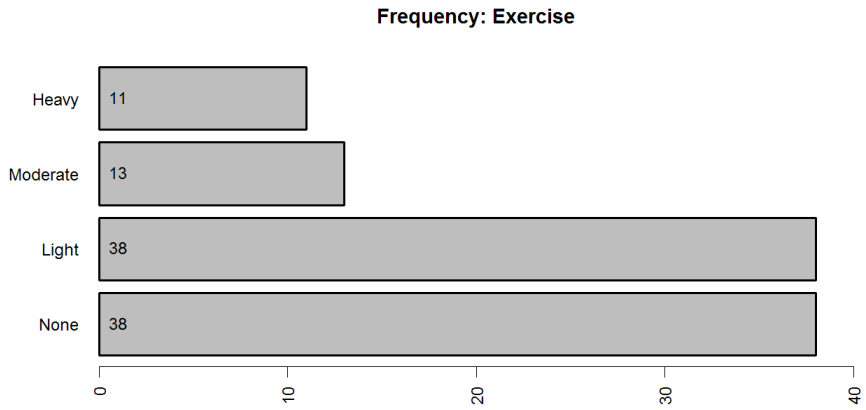
Reminder: R is case sensitive, so BLUMAN ≠ bluman.

Appendix E – Exploratory Data Analysis



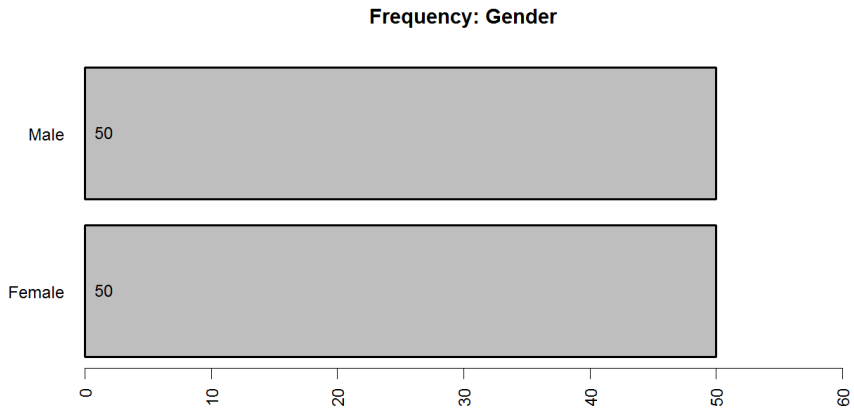

```
> educ.counts = table(BLUMAN$EDUC)
> educ.counts

 0  1  2  3
13 44 36  7

> names(educ.counts)=c("No Diploma", "HS Diploma", "College Degree", "Grad Degree")
> educ.counts

   No Diploma     HS Diploma College Degree    Grad Degree
           13             44             36              7

> par(las=2) # make label text perpendicular to axis
> par(mar=c(5,8,4,2)) # increase y-axis margin
> bp<-barplot(educ.counts, main="Frequency: Education Level", horiz=TRUE)  # plot and save plot object
> text(0, bp, educ.counts, pos=4)  # show counts on the bars
```

Appendix E – Exploratory Data Analysis (continued)



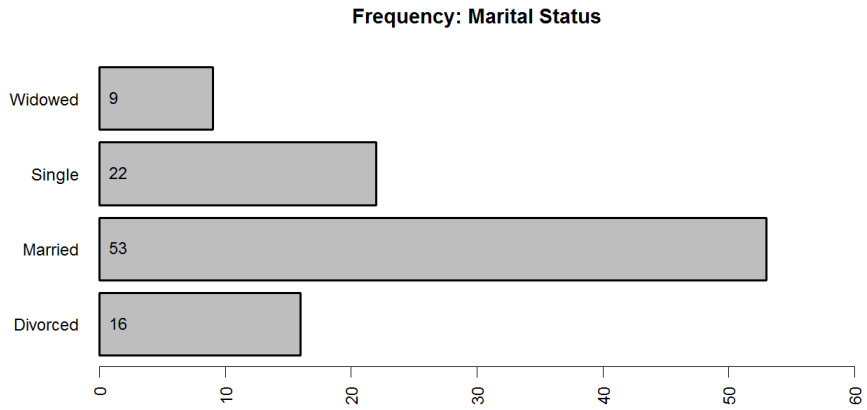

```
> exer.counts = table(BLUMAN$EXER)
> exer.counts

 0  1  2  3
38 38 13 11

> names(exer.counts)=c("None", "Light", "Moderate", "Heavy")
> exer.counts

   None    Light Moderate    Heavy
     38       38       13       11

> par(las=2) # make label text perpendicular to axis
> par(mar=c(5,8,4,2)) # increase y-axis margin
> bp<-barplot(exer.counts, main="Frequency: Exercise", horiz=TRUE, xlim=c(0,40))  # plot and save plot
object
> text(0, bp, exer.counts, pos=4)  # show counts on the bars
```

Appendix E – Exploratory Data Analysis (continued)




```
> sex.counts = table(BLUMAN$SEX)
> sex.counts

 F  M
50 50

> names(sex.counts)=c("Female", "Male")
> sex.counts

Female   Male
   50     50

> par(las=2) # make label text perpendicular to axis
> par(mar=c(5,8,4,2)) # increase y-axis margin
> bp<-barplot(sex.counts, main="Frequency: Gender", horiz=TRUE, xlim=c(0,60))  # plot and save plot
object
> text(0, bp, sex.counts, pos=4)  # show counts on the bars
```

Appendix E – Exploratory Data Analysis (continued)



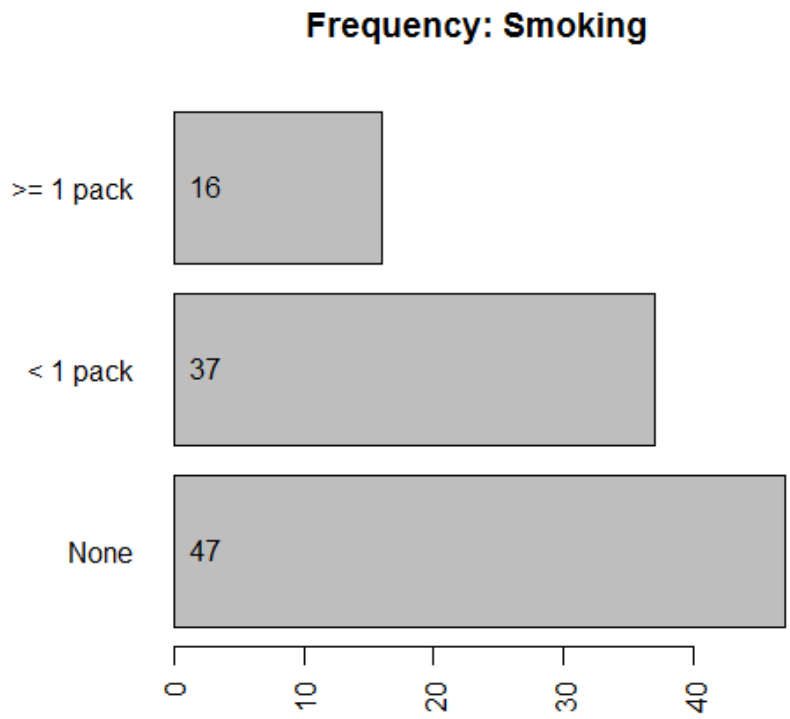

```
> marital.counts = table(BLUMAN$MARITAL)
> marital.counts

 D  M  S  W
16 53 22  9

> names(marital.counts)=c("Divorced", "Married", "Single", "Widowed")
> marital.counts

Divorced  Married   Single  Widowed
    16       53       22        9

> par(las=2) # make label text perpendicular to axis
> par(mar=c(5,8,4,2)) # increase y-axis margin
> bp<-barplot(marital.counts, main="Frequency: Marital Status", horiz=TRUE, xlim=c(0,60))  # plot and
save plot object
> text(0, bp, marital.counts, pos=4)  # show counts on the bars
```

Appendix E – Exploratory Data Analysis (continued)

Creating Dummy Variables

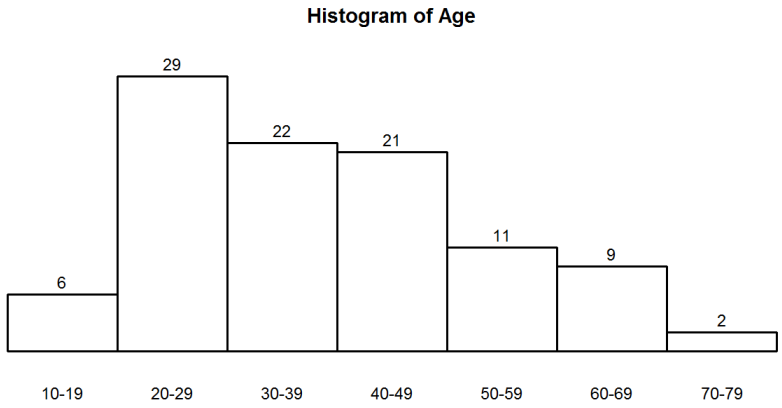
```
> BLUMAN$COLLEGE <- (BLUMAN$EDUC >= "2") * 1
> BLUMAN$ANYSMOKE <- (BLUMAN$SMOKE > "0") * 1
> BLUMAN$REALEXER <- (BLUMAN$EXER >= "2") * 1
> BLUMAN$MALE <- (BLUMAN$SEX == "M") * 1
> BLUMAN$MARRIED <- (BLUMAN$MARITAL == "M") * 1
> BLUMAN$HIGHBP <- (BLUMAN$SYSTOL >= 140) * 1
> names(BLUMAN) # show clumns
 [1] "ID"       "AGE"      "EDUC"     "SMOKE"    "EXER"
 [6] "WEIGHT"   "CHOLEST"  "SYSTOL"   "IQ"       "SODIUM"
[11] "SEX"      "MARITAL"  "COLLEGE"  "ANYSMOKE" "REALEXER"
[16] "MALE"     "MARRIED“  “HIGHBP"
```

Appendix E – Exploratory Data Analysis (continued)



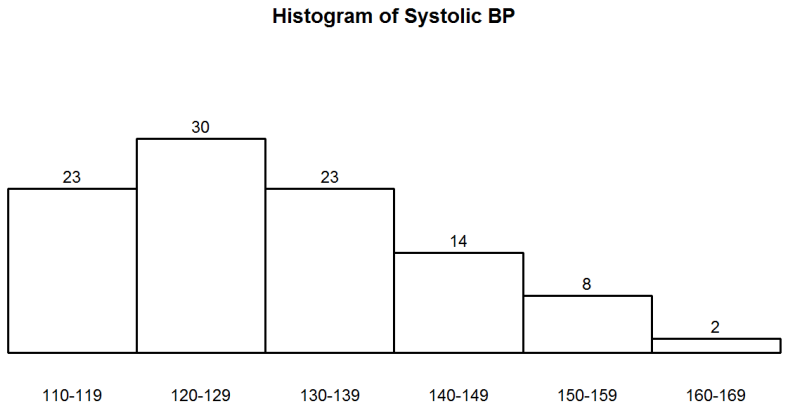

```
> smoke.counts = table(BLUMAN$SMOKE)
> smoke.counts

 0  1  2
47 37 16

> names(smoke.counts)=c("None", "< 1 pack", ">= 1 pack")
> smoke.counts

    None  < 1 pack >= 1 pack
      47       37       16

> par(las=2) # make label text perpendicular to axis
> par(mar=c(5,8,4,2)) # increase y-axis margin
> bp<-barplot(smoke.counts, main="Frequency: Smoking", horiz=TRUE)  # plot and save plot object
> text(0, bp, smoking.counts, pos=4)  # show counts on the bars
```

Appendix E – Exploratory Data Analysis (continued)




```
> summary(BLUMAN$AGE)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   26.75   36.00   38.41   47.25   74.00

> limits<-c("10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79")
> boundaries<-seq(9.5, 79.5, by=10)
> midpoints<-seq(14.5, 74.5, by=10)
> par(lwd=3, cex=1.5) # bar line width, bar label size
> hist(BLUMAN$AGE, breaks=boundaries, xaxt="n", yaxt="n", main="Histogram of Age", xlab='', ylab='',
labels=TRUE, ylim=c(0,30))
> axis(1, at=midpoints, labels=limits, tick=FALSE)
```

Appendix E – Exploratory Data Analysis (continued)




```
> summary(BLUMAN$SYSTOL)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  110.0   120.0   129.0   130.2   138.0   163.0

> limits<-c("110-119", "120-129", "130-139", "140-149", "150-159", "160-169")
> boundaries<-seq(109.5, 169.5, by=10)
> midpoints<-seq(114.5, 164.5, by=10)
> par(lwd=3, cex=1.5) # bar line width, bar label size
> hist(BLUMAN$SYSTOL, breaks=boundaries, xaxt="n", yaxt="n", main="Histogram of Systolic BP",
xlab='', ylab='', labels=TRUE, ylim=c(0,40))
> axis(1, at=midpoints, labels=limits, tick=FALSE)
```

Appendix F – Canonical Correlation Analysis

```
# required for CCA
install.packages("CCA")
require(CCA)

# perform CCA
cca<-cc(demog,calcs)

# create datasets for CCA
demog<-BLUMAN[, c("AGE", "WEIGHT", "COLLEGE", "ANYSMOKE", "REALEXER", "MALE", "MARRIED")]
calcs<-BLUMAN[, c("CALC1", "CALC2", "CALC3")]

# show canonical correlations
cca$cor

# show canonical coefficients
cca[3:4]

# loadings
loadings<-compute(demog,calcs,cca)
loadings[3:6]

# show canonical correlations
cca$cor

[1] 0.9538258 0.8223389 0.2556553

# standardized demog canonical coefficients
sdemog <- diag(sqrt(diag(cov(demog))))
sdemog %*% cca$xcoef

             [,1]         [,2]       [,3]
[1,] -0.891966653 -0.479187425 -0.6546250 ← AGE
[2,] -0.023130853  0.315718494  1.2769223 ← WEIGHT
[3,]  0.164180204 -0.084177876 -0.3991666 ← COLLEGE
[4,]  0.258673333 -0.433641137  0.2717584 ← ANYSMOKE
[5,]  0.004912276 -0.052179349 -0.1992531 ← REALEXER
[6,] -0.232307316  0.678926282 -0.8862088 ← MALE
[7,]  0.057299294 -0.004920603  0.5733980 ← MARRIED
```

Appendix F – Canonical Correlation Analysis (continued)

```
# standardized calcs canonical coefficients
scalcs <- diag(sqrt(diag(cov(calcs))))
scalcs %*% cca$ycoef
```

```
             [,1]        [,2]        [,3]
[1,]   0.01106031   1.1294676  -0.5922419  ← CALC1
[2,]  -0.63089763  -0.2340365   1.4967114  ← CALC2
[3,]  -0.46362770  -0.4268165  -1.2276703  ← CALC3
```

Appendix G – K-Means Clustering

```
# use heart risk values
calcs<-BLUMAN[, c("CALC1", "CALC2", "CALC3")]

# normalize
zcalcs<-scale(calcs,center=TRUE,scale=TRUE)

# cluster into three risk groups: high/medium/low?
clusters <- kmeans(zcalcs , 3)

# append assigned cluster
calcs$risk<-clusters$cluster

# how many in each risk cluster?
table(calcs$risk)
```

```
 1  2  3
18 56 26
```

```
# plot risk group against age and weight
plot(BLUMAN$AGE, BLUMAN$WEIGHT, pch=calcs$risk, col=calcs$risk, xlab="Age", ylab="Weight",
lwd=3, main="Heart Risk Group by Age and Weight")
```

Appendix H – Principal Component Analysis

```
analysis<-BLUMAN[, c("AGE", "WEIGHT", "IQ", "COLLEGE", "ANYSMOKE", "REALEXER", "MALE",
"MARRIED")]
pca<-prcomp(analysis, scale=TRUE)
pca

Standard deviations:
[1] 1.4634234 1.3688310 1.0960089 0.9994283 0.8803663 0.7426032
[7] 0.5526856 0.3906859

Rotation:
                PC1           PC2           PC3          PC4
            (Service)   (Young/Fit)  (Soccer Mom)
AGE         0.1109884  -0.48543951   0.31666694 -0.30900675
WEIGHT     -0.2512357  -0.63818495  -0.05920150  0.09208438
IQ         -0.5278131   0.15624816  -0.09820656 -0.42001097
COLLEGE    -0.5777760   0.12395660  -0.05572119 -0.27375837
ANYSMOKE    0.1145960  -0.13901867  -0.69667721 -0.40959588
REALEXER   -0.3337602   0.29312031  -0.11904455  0.50988843
MALE       -0.3082425  -0.45524835  -0.25344585  0.45211358
MARRIED    -0.3043171  -0.06923114   0.56542224 -0.11189617

                PC5           PC6           PC7          PC8
AGE         0.56929511 -0.266805384 -0.08964251 -0.39108709
WEIGHT      0.08993492 -0.001334428  0.18054422  0.69060114
IQ          0.14583284  0.173431572 -0.66023652  0.15390944
COLLEGE     0.11109862  0.168759230  0.68273213 -0.25654524
ANYSMOKE   -0.25013019 -0.495464139  0.03792305 -0.07018301
REALEXER    0.38439692 -0.615008924 -0.04001499  0.03047374
MALE       -0.23670494  0.215205648 -0.22907028 -0.52371567
MARRIED    -0.60650585 -0.447460342 -0.04266184 -0.02202220

summary(pca)

Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.4634 1.3688 1.0960 0.9994 0.88037
Proportion of Variance 0.2677 0.2342 0.1502 0.1249 0.09688
Cumulative Proportion  0.2677 0.5019 0.6521 0.7769 0.87381
                           PC6     PC7     PC8
Standard deviation     0.74260 0.55269 0.39069
Proportion of Variance 0.06893 0.03818 0.01908
Cumulative Proportion  0.94274 0.98092 1.00000
```

Half of the variability in the data can be accounted for by the first two variates.

Appendix H – Principal Component Analysis (continued)

```
# plot heart risk group by first two principal components
plot(pred[,1], pred[,2], pch=calcs$risk, col=calcs$risk, xlab="Professional Career <<< >>> Service Industry Worker", ylab="Older, Less Active <<< >>> Younger, More Fit",  lwd=3, main="Heart Risk Group by PCA1 and PCA2")
```

Appendix I – Linear Discriminant Analysis

```
# required package MASS
install.packages("MASS") # install required package
require(MASS)

# initialize random number seed
set.seed(4567)

# create analysis dataset with normalized risk score from clustering
analysis<-BLUMAN
analysis$risk<-calcs$risk

# create training and test datasets
ind<-sample(2, nrow(analysis), replace=TRUE, prob=c(0.7, 0.3))
trainData<-analysis[ind==1,]
testData<-analysis[ind==2,]
table(trainData$risk)
table(testData$risk)

# linear discriminant analysis
model<-lda(risk ~ AGE + WEIGHT + COLLEGE + ANYSMOKE + REALEXER + MALE + MARRIED,
data=analysis)
model

# predictions
predictions<-predict(model, testData)
class<-predictions$class

# create and display the confusion matrix
confusionMatrix<-table(testData$risk, class, dnn=c("actual", "pred"))
confusionMatrix
```

```
      pred
actual  1  2  3
     1  8  0  0
     2  0 13  0
     3  0  1  6
```

→ *Prediction resulted in only 1 misclassification!*