Fall W271 Lab 2

John Kenney

October 9, 2017

House Keeping

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

#Libraries required
library(car)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:car':
##
##     recode
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)#Used by author for 3D plotting
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
library(ggplot2)

library(effsize) #Used to calculate Cohen's D for T-Test
library(aod)    #Used for effect size of the logit model
```

```
##
## Attaching package: 'aod'
```

```
## The following object is masked from 'package:survival':
##
##     rats
```

```
library(mcprofile) #Used for confidence intervals
```

```
## Warning: package 'mcprofile' was built under R version 3.4.2
```

Exploratory Data Analysis

We begin with an exploration of the data. Let's first look at our data at a high level.

```
my.data = read.csv("C:\\Users\\jkenney\\Dropbox\\UCB\\Fall 2017\\W271\\Labs\\Lab 2\\lab2data.csv",
    header = TRUE)
glimpse(my.data)
```

```
## Observations: 1,000
## Variables: 12
## $ X               <int> 761, 620, 214, 373, 748, 1080, 1155, 1069, 116...
## $ Gender          <fctr> F, M, F, F, M, F, F, F, F, F, F, F, F, M, ...
## $ Class.Year      <int> 2002, 2002, 1982, 1992, 2002, 2012, 2012, 2012...
## $ Marital.Status  <fctr> M, S, M, M, S, S, S, S, S, M, S, S, S, D, M, ...
## $ Major           <fctr> Sociology, History, History, Anthropology, Ph...
## $ Next.Degree     <fctr> MSW, NONE, NONE, MS, NONE, JD, NONE, MS, NONE...
## $ AttendenceEvent <int> 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0...
## $ FY12Giving      <dbl> 50, 0, 100, 0, 0, 0, 0, 5, 0, 0, 0, 0, 10, 0, ...
## $ FY13Giving      <dbl> 51, 0, 0, 0, 0, 0, 0, 10, 0, 75, 0, 0, 0, 0, 0...
## $ FY14Giving      <dbl> 51, 0, 100, 0, 0, 0, 0, 25, 0, 0, 0, 0, 0, 0, ...
## $ FY15Giving      <dbl> 0, 0, 100, 0, 0, 0, 0, 25, 0, 0, 0, 0, 10, 0, ...
## $ FY16Giving      <dbl> 0, 0, 100, 0, 0, 0, 0, 50, 0, 60, 0, 0, 10, 15...
```

We have 1,000 observations, each with 12 associated variables.

```
# X Variable
describe(my.data$X)
```

```
## my.data$X
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0     1000        1    615.4    410.6    62.95   122.90
##      .25      .50      .75      .90      .95
##   308.75   613.00   917.25  1110.30  1174.05
##
## lowest :    1    2    3    4    5, highest: 1225 1226 1228 1229 1230
```

```
min(my.data$X)
```

```
## [1] 1
```

```
max(my.data$X)
```

```
## [1] 1230
```

It looks like this is the ID of the respondent, since there are 1000 unique numbers.

```
# Gender Variable
describe(my.data$Gender)
```

```
## my.data$Gender
##        n  missing distinct
```

```
##     1000        0        2
##
## Value          F     M
## Frequency    505   495
## Proportion 0.505 0.495
```

There is a reasonably equal proportion of men and women in the sample.

```
# Class Year Variable
describe(my.data$Class.Year)
```

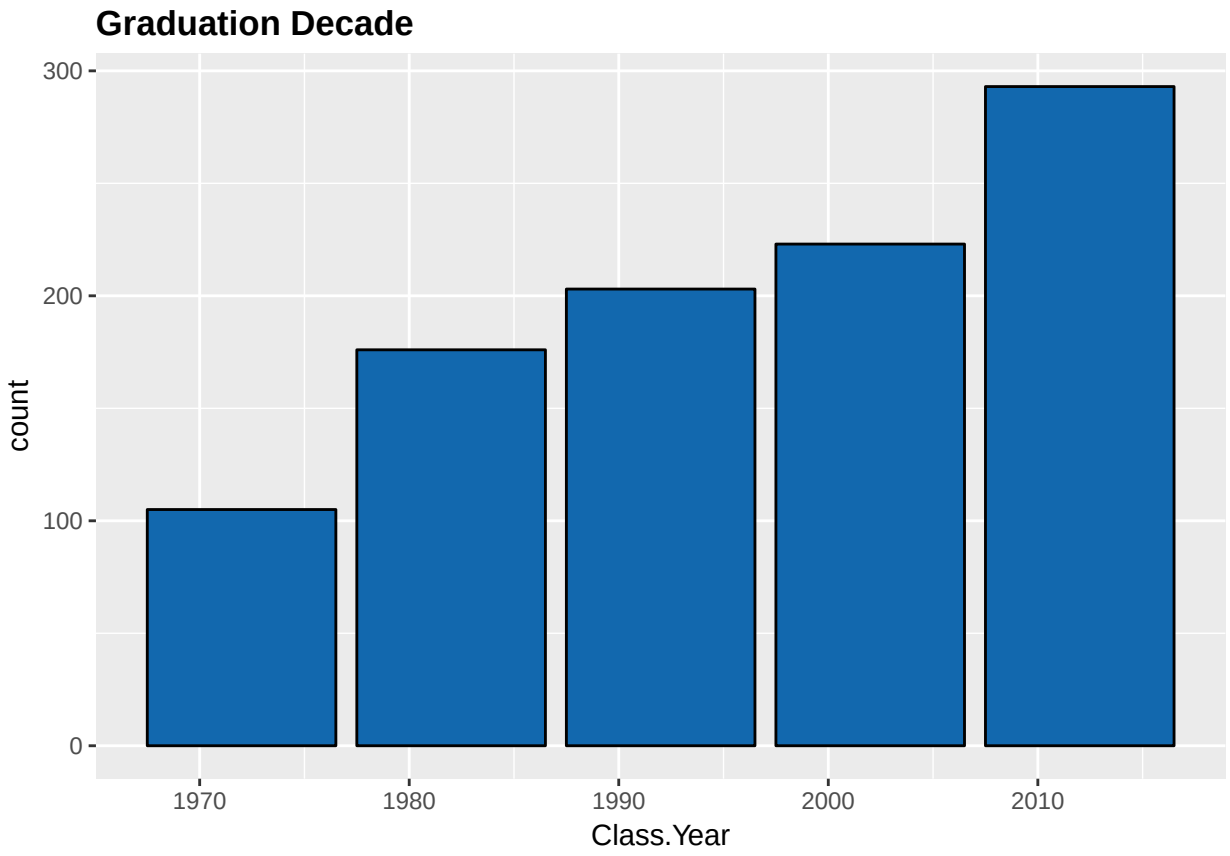
```
## my.data$Class.Year
##        n  missing distinct     Info     Mean      Gmd
##     1000        0        5    0.949     1996    15.07
##
## Value       1972  1982  1992  2002  2012
## Frequency    105   176   203   223   293
## Proportion 0.105 0.176 0.203 0.223 0.293
```

Here we have five classes for graduation year that look to be coded in ten year intervals:

1. 1972-1981
2. 1982-1991
3. 1992-2001
4. 2002-2011
5. 2012-Present

It seems the proportion of respondents is not independent of the graduation class. The more recently graduated students make up a greater proportion of the sample than older graduates. This distribution looks to be ordinal.

```
ggplot(my.data, aes(x = Class.Year)) + geom_bar(aes(y = ..count..),
    fill = "#1268AE", colour = "black") + ggtitle("Graduation Decade") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



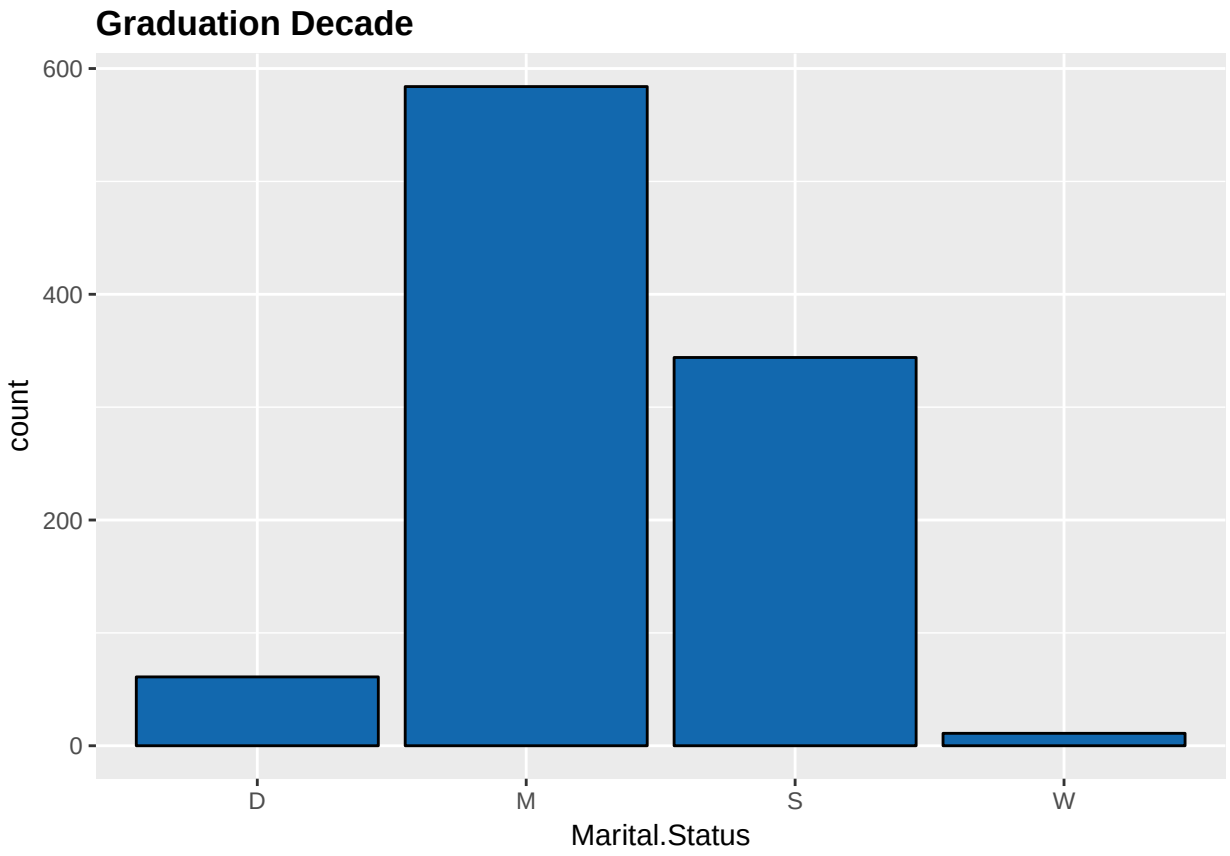

```
# Marital Status Variable
describe(my.data$Marital.Status)
```

```
## my.data$Marital.Status
##        n  missing distinct
##     1000        0        4
##
## Value          D     M     S     W
## Frequency     61   584   344    11
## Proportion 0.061 0.584 0.344 0.011
```

We have four classes for the Marital Status varible:

1. Divorced
2. Married
3. Single
4. Widowed

```
ggplot(my.data, aes(x = Marital.Status)) + geom_bar(aes(y = ..count..),
    fill = "#1268AE", colour = "black") + ggtitle("Graduation Decade") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```

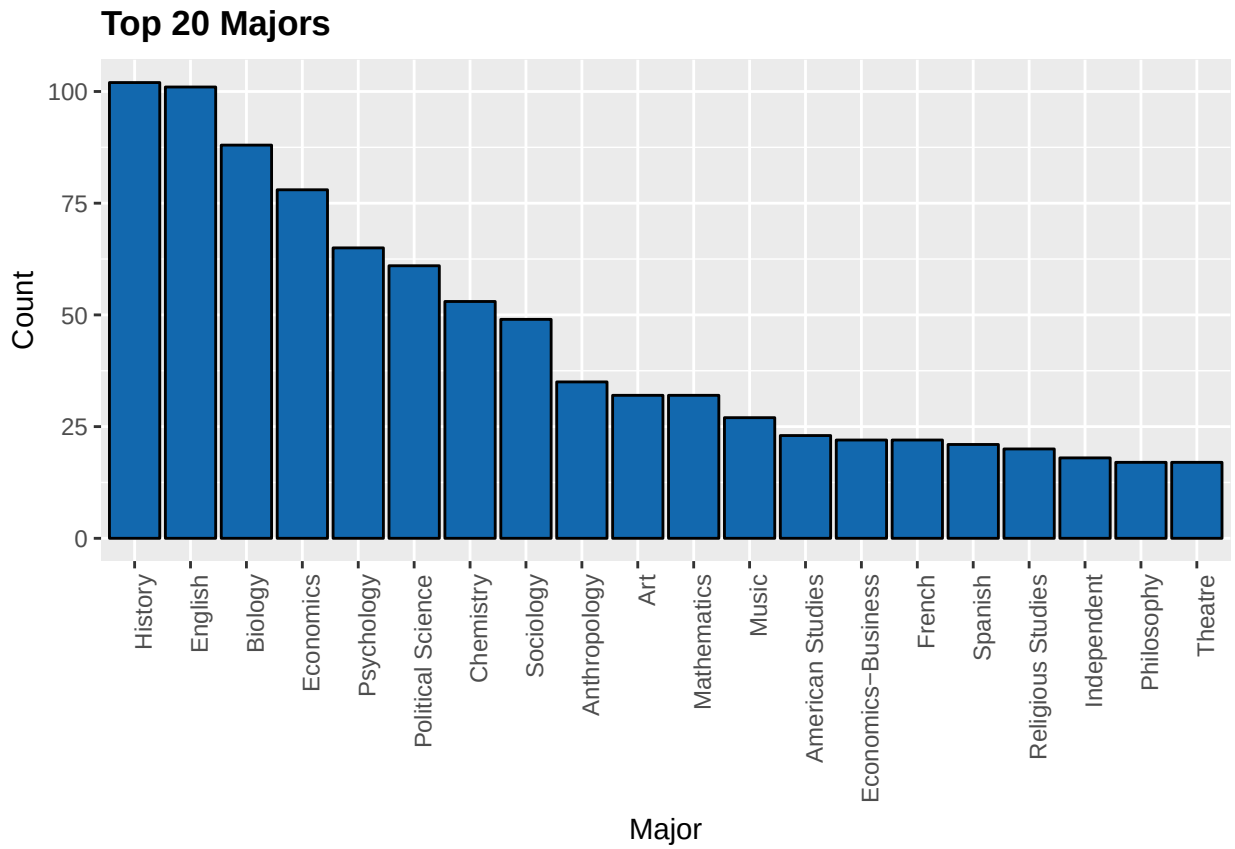



```
# Marital Status Variable
describe(my.data$Major)
```

```
## my.data$Major
##        n  missing distinct
##     1000        0       45
##
## lowest : American Studies     Anthropology         Art                  Biology              Chemistr
## highest: Spanish              Speech (Drama, etc.) Speech Correction    Theatre              Zoology
```

We have forty-five distinct classes for the Major! Wow. Let's take a look at a bar chart

```
my.data.major.count = as.data.frame(sort(table(my.data$Major),
    decreasing = TRUE)[1:20])
ggplot(my.data.major.count, aes(x = Var1, y = Freq)) + geom_bar(stat = "identity",
    fill = "#1268AE", colour = "black") + ggtitle("Top 20 Majors") +
    theme(plot.title = element_text(lineheight = 1, face = "bold")) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    xlab("Major") + ylab("Count")
```




```
# Next Degree Variable
summary(my.data$Next.Degree)
```

```
##    AA    BA   BAE    BD   BFA    BN    BS   BSN    DC   DDS   DMD    DO  DO2    DP    JD
##     1     4     1     1     1     2     2     3     1     1     1     2     1     1    90
##   LLB   LLD    MA   MA2   MAE  MALS   MAT   MBA   MCP    MD   MD2    ME   MFA   MHA    ML
##     1     1   108     1     1     1    10    34     1    42     9    17    14     1     1
##   MLS    MM   MPA   MPH    MS   MSM   MSW   NDA  NONE   PHD   STM    TC  UBDS  UDDS   UMD
##     9     1     6     4    53     1    11    58   378    78     1    22     6     4     6
##  UMDS  UNKD
##     2     6
```

```
my.data.next.degree = as.data.frame(sort(table(my.data$Next.Degree),
    decreasing = TRUE)[1:20])
my.data.next.degree
```

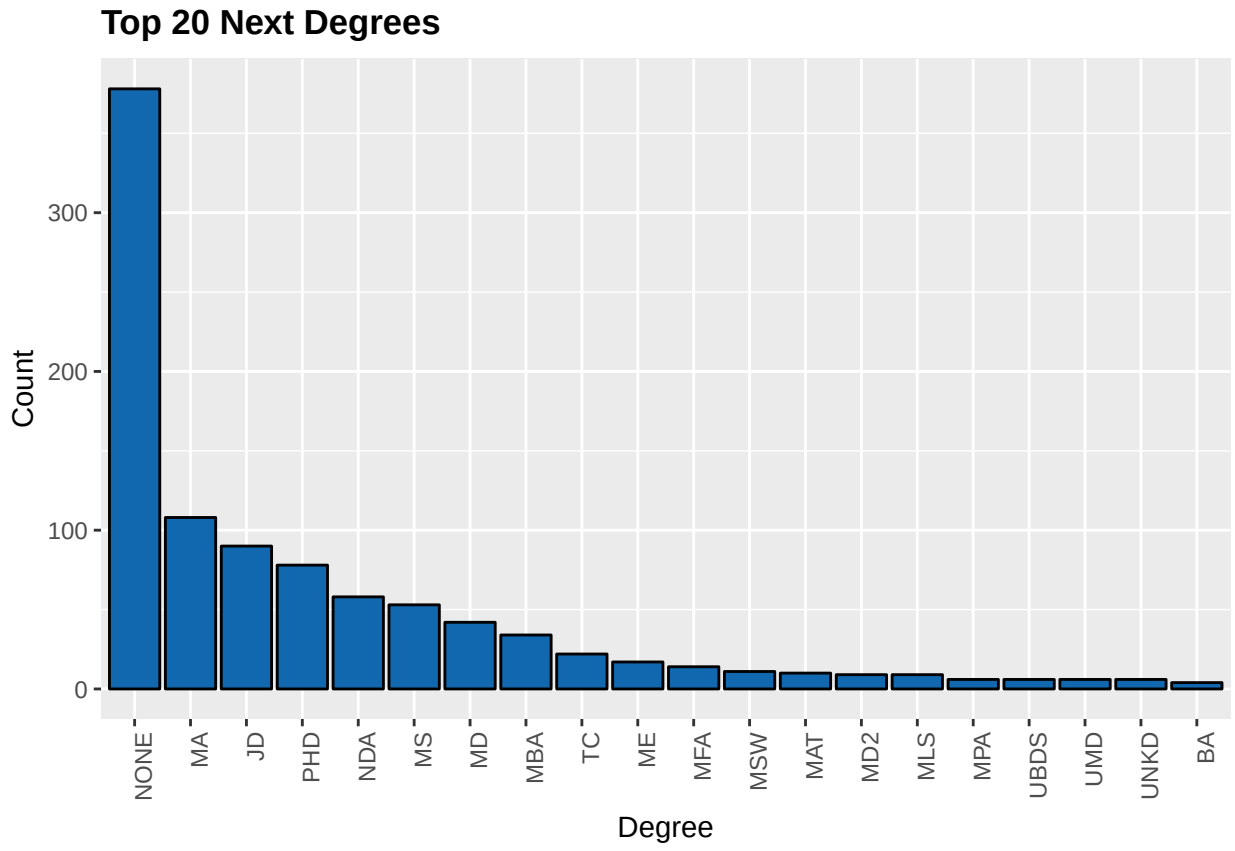
```
##     Var1 Freq
## 1   NONE  378
## 2     MA  108
## 3     JD   90
## 4    PHD   78
## 5    NDA   58
## 6     MS   53
## 7     MD   42
## 8    MBA   34
## 9     TC   22
## 10    ME   17
```

```
## 11  MFA   14
## 12  MSW   11
## 13  MAT   10
## 14  MD2    9
## 15  MLS    9
## 16  MPA    6
## 17 UBDS    6
## 18  UMD    6
## 19 UNKD    6
## 20   BA    4
```

```
my.data.next.degree$Freq
```

```
##  [1] 378 108  90  78  58  53  42  34  22  17  14  11  10   9   9   6   6
## [18]   6   6   4
```

```
ggplot(my.data.next.degree, aes(x = Var1, y = Freq)) + geom_bar(stat = "identity",
    fill = "#1268AE", colour = "black") + ggtitle("Top 20 Next Degrees") +
    theme(plot.title = element_text(lineheight = 1, face = "bold")) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    xlab("Degree") + ylab("Count")
```



Now on to Event Attendance...

```
describe(my.data$AttendenceEvent)
```

```
## my.data$AttendenceEvent
##        n  missing distinct     Info      Sum     Mean      Gmd
```

```
##     1000        0        2    0.717      605    0.605   0.4784
```

More in attendance than not with 60% attending an event.

Finally, on to giving, our dependent variables.

$ FY12Giving 50, 0, 100, 0, 0, 0, 0, 5, 0, 0, 0, 0, 10, 0, 0, 0, 0, 20, 100, 30, 100, 0, 0, 50, 50, 0, 0, 0, 0, 50, 500... $ FY13Giving 51, 0, 0, 0, 0, 0, 0, 10, 0, 75, 0, 0, 0, 0, 0, 5, 0, 50, 160, 0, 75, 0, 0, 100, 75, 0, 0, 0, 0, 50, 500, ... $ FY14Giving 51, 0, 100, 0, 0, 0, 0, 25, 0, 0, 0, 0, 0, 0, 0, 0, 0, 50, 200, 0, 0, 0, 0, 156, 50, 0, 0, 0, 0, 0, 50, 1000,... $ FY15Giving 0, 0, 100, 0, 0, 0, 0, 25, 0, 0, 0, 0, 10, 0, 0, 0, 0, 50, 150, 50, 0, 0, 0, 157, 80, 0, 0, 0, 500, 50, 15... $ FY16Giving

```
describe(my.data$FY12Giving)
```

```
## my.data$FY12Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       66    0.826    186.9    345.5        0        0
##      .25      .50      .75      .90      .95
##        0        0       60      200      350
##
## lowest :     0.00     5.00     6.50     7.00     8.00
## highest: 10000.00 12000.00 16959.99 20000.00 21000.00
```

```
describe(my.data$FY13Giving)
```

```
## my.data$FY13Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       78    0.864    311.5    590.4      0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0      0.0     75.0    210.5    400.0
##
## Value           0    500   1000   1500   2000   2500   3000   5000   5500
## Frequency     920     48     13      4      2      3      2      2      1
## Proportion  0.920  0.048  0.013  0.004  0.002  0.003  0.002  0.002  0.001
##
## Value        8000  12000  13000  14500 161500
## Frequency       1      1      1      1      1
## Proportion  0.001  0.001  0.001  0.001  0.001
```

```
describe(my.data$FY14Giving)
```

```
## my.data$FY14Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       80     0.83    142.6    255.5        0        0
##      .25      .50      .75      .90      .95
##        0        0       50      200      450
##
## lowest :     0.00     1.00     5.00     8.00    10.00
## highest:  5000.00  6000.00  8031.00 10000.00 11187.26
```

```
describe(my.data$FY15Giving)
```

```
## my.data$FY15Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       62    0.817    252.2    470.7      0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0      0.0     75.0    200.0    538.3
##
```

```
## lowest :     0.0     5.0    10.0    13.0    15.0
## highest: 10000.0 14776.0 15634.5 26500.0 58785.5
```

```
describe(my.data$FY16Giving)
```

```
## my.data$FY16Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       71    0.798      170    308.2        0        0
##      .25      .50      .75      .90      .95
##        0        0       75      216      500
##
## lowest :     0.00     5.00    10.00    15.00    18.00
## highest:  5000.00  6500.00 11500.00 11505.84 14655.25
```