Loading Data into R

Dr. Sarah Hunter

3/15/2020

Loading Data into R

R was designed to analyze and manipulate data. Because of that, R can use many different types of data. The most R-friendly of data is a ".csv" file. This is a "comma separated variables" file This refers to the way in which the data are structured. While you can open a `.csv` file in Microsoft Excel, it is usually better just to save the file and an load it into R in order to use it. In fact, attempting to first open a file in Excel or Numbers can actually damage the file, leading to many errors in R. The best way to load a data file into R is to save it and leave it alone. To load a `.csv` file, use the following code, remembering to set your `working directory` to the file folder where your data are saved. You will need to do this every time you open R. Notice also, from our last R workshop, that we have to name the data set something. In this case, I named it `mydata`. However, it could have just as easily been `fred` or `bob`. Generally, I opt for more descriptive names. For example, this data set, which is data from the OECD on aid projects for 2012, could have been `oecd2012` or `aid2012`. Using descriptive names can help you in the future when you have multiple data sets loaded into R at the same time.

```
setwd("/Users/sarahhunter/Desktop/Data")

mydata<-read.csv("oda_crs_oecd2012.csv")
```

Summarizing Data

While you have your data loaded, you can first get the names of all the variables in the data set using the `names` command. The following code shows how it is done:

```
names(mydata)
```

```
##  [1] "X.1"                      "X"
##  [3] "Year"                     "DonorCode"
##  [5] "DonorName"                "AgencyCode"
##  [7] "AgencyName"               "CrsID"
##  [9] "ProjectNumber"            "InitialReport"
## [11] "RecipientCode"            "RecipientName"
## [13] "RegionCode"               "RegionName"
## [15] "IncomegroupCode"          "IncomegroupName"
## [17] "FlowCode"                 "FlowName"
## [19] "Bi_Multi"                 "Category"
## [21] "Finance_t"                "Aid_t"
## [23] "USD_Commitment"           "USD_Disbursement"
## [25] "USD_Received"             "USD_Commitment_Defl"
## [27] "USD_Disbursement_Defl"    "USD_Received_Defl"
## [29] "USD_Adjustment"           "USD_Adjustment_Defl"
## [31] "USD_AmountUntied"         "USD_AmountPartialTied"
```

```
## [33] "USD_AmountTied"             "USD_AmountUntied_Defl"
## [35] "USD_AmountPartialTied_Defl" "USD_Amounttied_Defl"
## [37] "USD_IRTC"                   "USD_Expert_Commitment"
## [39] "USD_Expert_Extended"        "USD_Export_Credit"
## [41] "CurrencyCode"               "Commitment_National"
## [43] "Disbursement_National"      "GrantEquiv"
## [45] "USD_GrantEquiv"             "ShortDescription"
## [47] "ProjectTitle"               "PurposeCode"
## [49] "PurposeName"                "SectorCode"
## [51] "SectorName"                 "ChannelCode"
## [53] "ChannelName"                "ChannelReportedName"
## [55] "ParentChannelCode"          "Geography"
## [57] "ExpectedStartDate"          "CompletionDate"
## [59] "LongDescription"            "SDGfocus"
## [61] "Gender"                     "Environment"
## [63] "PDGG"                       "Trade"
## [65] "RMNCH"                      "DRR"
## [67] "Nutrition"                  "Disability"
## [69] "FTC"                        "PBA"
## [71] "InvestmentProject"          "AssocFinance"
## [73] "Biodiversity"               "ClimateMitigation"
## [75] "ClimateAdaptation"          "Desertification"
## [77] "CommitmentDate"             "TypeRepayment"
## [79] "NumberRepayment"            "Interest1"
## [81] "Interest2"                  "Repaydate1"
## [83] "Repaydate2"                 "USD_Interest"
## [85] "USD_Outstanding"            "USD_Arrears_Principal"
## [87] "USD_Arrears_Interest"       "BudgetIdent"
## [89] "CapitalExpend"              "PSIflag"
## [91] "PSIAddType"                 "PSIAddAssess"
## [93] "PSIAddDevObj"
```

From there, you can get summary statistics from the entire data set using the `summary` command.

```
summary(mydata)
```

```
##       X.1                X                Year        DonorCode
##  Min.   :     1   Min.   :1413887   Min.   :2012   Min.   :   1.0
##  1st Qu.: 52163   1st Qu.:1466049   1st Qu.:2012   1st Qu.:  12.0
##  Median :106564   Median :1520450   Median :2012   Median : 302.0
##  Mean   :107665   Mean   :1521551   Mean   :2012   Mean   : 460.7
##  3rd Qu.:162960   3rd Qu.:1576846   3rd Qu.:2012   3rd Qu.: 905.0
##  Max.   :221053   Max.   :1634939   Max.   :2012   Max.   :1312.0
##
##   DonorName           AgencyCode      AgencyName           CrsID
##  Length:183482      Min.   : 1.000   Length:183482      Length:183482
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 2.000   Mode  :character   Mode  :character
##                     Mean   : 7.258
##                     3rd Qu.: 8.000
##                     Max.   :99.000
##
##  ProjectNumber      InitialReport  RecipientCode  RecipientName
##  Length:183482      Min.   :1.00   Min.   :  55   Length:183482
##  Class :character   1st Qu.:1.00   1st Qu.: 253   Class :character
```

```
##  Mode  :character   Median :3.00   Median : 389   Mode  :character
##                     Mean   :3.36   Mean   :1229
##                     3rd Qu.:3.00   3rd Qu.: 666
##                     Max.   :8.00   Max.   :9998
##                     NA's   :455
##    RegionCode     RegionName        IncomegroupCode IncomegroupName
##  Min.   :  298   Length:183482      Min.   :10016   Length:183482
##  1st Qu.:10003   Class :character   1st Qu.:10016   Class :character
##  Median :10006   Mode  :character   Median :10018   Mode  :character
##  Mean   :10428                      Mean   :10019
##  3rd Qu.:10009                      3rd Qu.:10019
##  Max.   :15006                      Max.   :10025
##
##     FlowCode       FlowName           Bi_Multi      Category
##  Min.   :11.00   Length:183482      Min.   :1.0   Min.   :10.00
##  1st Qu.:11.00   Class :character   1st Qu.:1.0   1st Qu.:10.00
##  Median :11.00   Mode  :character   Median :1.0   Median :10.00
##  Mean   :11.29                      Mean   :1.7   Mean   :10.43
##  3rd Qu.:11.00                      3rd Qu.:1.0   3rd Qu.:10.00
##  Max.   :14.00                      Max.   :7.0   Max.   :21.00
##
##    Finance_t        Aid_t           USD_Commitment      USD_Disbursement
##  Min.   :110.0   Length:183482      Min.   :-702.8310   Min.   : -59.094
##  1st Qu.:110.0   Class :character   1st Qu.:   0.0000   1st Qu.:   0.006
##  Median :110.0   Mode  :character   Median :   0.0135   Median :   0.042
##  Mean   :149.4                      Mean   :   1.3109   Mean   :   0.857
##  3rd Qu.:110.0                      3rd Qu.:   0.1283   3rd Qu.:   0.214
##  Max.   :623.0                      Max.   :1667.1500   Max.   :1733.270
##                                                         NA's   :7566
##   USD_Received   USD_Commitment_Defl USD_Disbursement_Defl USD_Received_Defl
##  Min.   :  0.0   Min.   :-758.6930   Min.   : -51.076      Min.   :   0.00
##  1st Qu.:  0.0   1st Qu.:   0.0000   1st Qu.:   0.005      1st Qu.:   0.00
##  Median :  0.0   Median :   0.0120   Median :   0.038      Median :   0.00
##  Mean   :  0.2   Mean   :   1.2113   Mean   :   0.797      Mean   :   0.19
##  3rd Qu.:  0.0   3rd Qu.:   0.1155   3rd Qu.:   0.195      3rd Qu.:   0.00
##  Max.   :960.1   Max.   :1799.6600   Max.   :1871.030      Max.   :1035.87
##  NA's   :74497                       NA's   :7566          NA's   :74497
##  USD_Adjustment   USD_Adjustment_Defl USD_AmountUntied  USD_AmountPartialTied
##  Min.   :-71.46   Min.   :-65.53      Min.   :-585.58   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:  0.00      1st Qu.:   0.00   1st Qu.:  0.00
##  Median :  0.00   Median :  0.00      Median :   0.00   Median :  0.00
##  Mean   : -2.31   Mean   : -2.11      Mean   :   0.78   Mean   :  0.07
##  3rd Qu.:  0.00   3rd Qu.:  0.00      3rd Qu.:   0.10   3rd Qu.:  0.00
##  Max.   :  0.00   Max.   :  0.00      Max.   :1581.36   Max.   :212.18
##  NA's   :183451   NA's   :183451      NA's   :64964     NA's   :113507
##  USD_AmountTied    USD_AmountUntied_Defl USD_AmountPartialTied_Defl
##  Min.   :-117.22   Min.   :-632.13       Min.   :  0.00
##  1st Qu.:   0.00   1st Qu.:   0.00       1st Qu.:  0.00
##  Median :   0.00   Median :   0.00       Median :  0.00
##  Mean   :   0.21   Mean   :   0.71       Mean   :  0.06
##  3rd Qu.:   0.00   3rd Qu.:   0.09       3rd Qu.:  0.00
##  Max.   : 521.39   Max.   :1165.76       Max.   :196.40
##  NA's   :83292     NA's   :64964         NA's   :113507
##  USD_Amounttied_Defl    USD_IRTC      USD_Expert_Commitment USD_Expert_Extended
```

```
##  Min.   :-126.53     Min.   : 0.00    Min.   :0.00          Min.   :0.00
##  1st Qu.:   0.00     1st Qu.: 0.00    1st Qu.:0.01          1st Qu.:0.01
##  Median :   0.00     Median : 0.00    Median :0.02          Median :0.02
##  Mean   :   0.21     Mean   : 0.09    Mean   :0.02          Mean   :0.04
##  3rd Qu.:   0.00     3rd Qu.: 0.03    3rd Qu.:0.03          3rd Qu.:0.03
##  Max.   : 431.76     Max.   :26.36    Max.   :0.06          Max.   :0.22
##  NA's   :83292       NA's   :178425   NA's   :183463        NA's   :183461
##  USD_Export_Credit  CurrencyCode   Commitment_National Disbursement_National
##  Min.   :-0.52     Min.   :  3.0   Min.   :-702.831    Min.   : -59.094
##  1st Qu.: 2.68     1st Qu.:302.0   1st Qu.:   0.000    1st Qu.:   0.003
##  Median : 5.73     Median :302.0   Median :   0.008    Median :   0.033
##  Mean   : 5.60     Mean   :499.4   Mean   :   1.394    Mean   :   0.957
##  3rd Qu.: 7.07     3rd Qu.:918.0   3rd Qu.:   0.118    3rd Qu.:   0.205
##  Max.   :16.84     Max.   :918.0   Max.   :3863.110    Max.   :3863.110
##  NA's   :183420                                        NA's   :7566
##  GrantEquiv     USD_GrantEquiv ShortDescription   ProjectTitle
##  Mode:logical   Mode:logical   Length:183482      Length:183482
##  NA's:183482    NA's:183482    Class :character   Class :character
##                                Mode  :character   Mode  :character
##
##
##
##
##   PurposeCode    PurposeName          SectorCode     SectorName
##  Min.   :  100   Length:183482      Min.   :100.0   Length:183482
##  1st Qu.:13040   Class :character   1st Qu.:130.0   Class :character
##  Median :16010   Mode  :character   Median :160.0   Mode  :character
##  Mean   :28267                      Mean   :283.8
##  3rd Qu.:33120                      3rd Qu.:331.0
##  Max.   :99820                      Max.   :998.0
##
##   ChannelCode    ChannelName        ChannelReportedName ParentChannelCode
##  Min.   :10000   Length:183482      Length:183482       Min.   :10000
##  1st Qu.:11000   Class :character   Class :character    1st Qu.:11000
##  Median :21000   Mode  :character   Mode  :character    Median :21000
##  Mean   :27767                                          Mean   :27764
##  3rd Qu.:40000                                          3rd Qu.:40000
##  Max.   :90000                                          Max.   :90000
##  NA's   :12203                                          NA's   :12203
##   Geography         ExpectedStartDate  CompletionDate     LongDescription
##  Length:183482      Length:183482      Length:183482      Length:183482
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  SDGfocus           Gender       Environment         PDGG           Trade
##  Mode:logical   Min.   :0.00    Min.   :0.00    Min.   :0.00    Min.   :0.00
##  NA's:183482    1st Qu.:0.00    1st Qu.:0.00    1st Qu.:0.00    1st Qu.:0.00
##                 Median :0.00    Median :0.00    Median :0.00    Median :0.00
##                 Mean   :0.39    Mean   :0.42    Mean   :0.64    Mean   :0.11
##                 3rd Qu.:1.00    3rd Qu.:1.00    3rd Qu.:1.00    3rd Qu.:0.00
##                 Max.   :2.00    Max.   :2.00    Max.   :2.00    Max.   :2.00
```

```
##                NA's   :64896   NA's   :61126   NA's   :61748   NA's   :96015
##      RMNCH          DRR           Nutrition      Disability          FTC
##  Min.   :0.00     Mode:logical   Mode:logical   Mode:logical   Min.   :1
##  1st Qu.:0.00     NA's:183482    NA's:183482    NA's:183482    1st Qu.:1
##  Median :0.00                                                  Median :1
##  Mean   :0.13                                                  Mean   :1
##  3rd Qu.:0.00                                                  3rd Qu.:1
##  Max.   :4.00                                                  Max.   :1
##  NA's   :165424                                                NA's   :112366
##       PBA          InvestmentProject  AssocFinance     Biodiversity
##  Min.   :1        Min.   :1         Min.   :1        Min.   :0.00
##  1st Qu.:1        1st Qu.:1         1st Qu.:1        1st Qu.:0.00
##  Median :1        Median :1         Median :1        Median :0.00
##  Mean   :1        Mean   :1         Mean   :1        Mean   :0.11
##  3rd Qu.:1        3rd Qu.:1         3rd Qu.:1        3rd Qu.:0.00
##  Max.   :1        Max.   :1         Max.   :1        Max.   :2.00
##  NA's   :171807   NA's   :158446    NA's   :182112   NA's   :77438
##  ClimateMitigation ClimateAdaptation Desertification CommitmentDate
##  Min.   :0.00      Min.   :0.0       Min.   :0.00    Length:183482
##  1st Qu.:0.00      1st Qu.:0.0       1st Qu.:0.00    Class :character
##  Median :0.00      Median :0.0       Median :0.00    Mode  :character
##  Mean   :0.09      Mean   :0.1       Mean   :0.05
##  3rd Qu.:0.00      3rd Qu.:0.0       3rd Qu.:0.00
##  Max.   :2.00      Max.   :2.0       Max.   :3.00
##  NA's   :75012     NA's   :75887     NA's   :87818
##  TypeRepayment    NumberRepayment   Interest1           Interest2
##  Min.   :1.00     Min.   : 1       Length:183482      Min.   :   0.00
##  1st Qu.:1.00     1st Qu.: 2       Class :character   1st Qu.:   0.00
##  Median :1.00     Median : 2       Mode  :character   Median :   0.00
##  Mean   :1.81     Mean   : 2                          Mean   :  89.27
##  3rd Qu.:2.00     3rd Qu.: 2                          3rd Qu.:   0.00
##  Max.   :5.00     Max.   :12                          Max.   :9860.00
##  NA's   :173169   NA's   :173726                      NA's   :175592
##   Repaydate1         Repaydate2         USD_Interest    USD_Outstanding
##  Length:183482      Length:183482      Min.   : -0.01   Min.   :   0.00
##  Class :character   Class :character   1st Qu.:  0.00   1st Qu.:   0.00
##  Mode  :character   Mode  :character   Median :  0.00   Median :   0.00
##                                        Mean   :  0.13   Mean   :   8.88
##                                        3rd Qu.:  0.03   3rd Qu.:   1.50
##                                        Max.   :121.12   Max.   :2856.66
##                                        NA's   :136238   NA's   :170569
##  USD_Arrears_Principal USD_Arrears_Interest  BudgetIdent    CapitalExpend
##  Min.   : 0.00         Min.   : 0.00        Min.   :11110   Mode:logical
##  1st Qu.: 0.00         1st Qu.: 0.00        1st Qu.:13040   NA's:183482
##  Median : 0.00         Median : 0.00        Median :16010
##  Mean   : 0.20         Mean   : 0.06        Mean   :28336
##  3rd Qu.: 0.00         3rd Qu.: 0.00        3rd Qu.:33120
##  Max.   :73.93         Max.   :40.90        Max.   :99820
##  NA's   :179654        NA's   :179648       NA's   :455
##  PSIflag        PSIAddType     PSIAddAssess   PSIAddDevObj
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA's:183482    NA's:183482    NA's:183482    NA's:183482
##
##
```

```
##
##
##
```

As seen above, the `summary` command can sometimes give too many lines of code. There are two commands you can use to make sure that your data set has loaded correctly: `dim` and `head`. The `dim` command gives you the numbers of rows and columns in the data set. In other words, this command gives you the number of observations and variables (including the identifying variables such as country and year). It looks like:

```
dim(mydata)
```

```
## [1] 183482     93
```

If you want to only see the first five rows of the data, to give you a better idea of the structure of the data, you can use the `head` command. It looks like:

```
head(mydata)
```

```
##   X.1       X Year DonorCode     DonorName AgencyCode
## 1   1 1413887 2012       302 United States          1
## 2   3 1413889 2012       301        Canada          1
## 3   4 1413890 2012       301        Canada          1
## 4   5 1413891 2012       302 United States          1
## 5   6 1413892 2012       974         UNFPA          1
## 6   7 1413893 2012         2       Belgium         10
##                                          AgencyName          CrsID
## 1                Agency for International Development     2012014531
## 2            Canadian International Development Agency 20030004330001
## 3            Canadian International Development Agency 20102005360001
## 4                Agency for International Development     2012008421
## 5                                              UNFPA     2012002520
## 6 Directorate General for Co-operation and Development
##   ProjectNumber InitialReport RecipientCode          RecipientName RegionCode
## 1      76_41179             1           251                Liberia      10003
## 2    A032123001             3           269                Senegal      10003
## 3    S064874001             3           269                Senegal      10003
## 4      76_35007             1           251                Liberia      10003
## 5                           1           251                Liberia      10003
## 6                          NA            64 Bosnia and Herzegovina      10010
##        RegionName IncomegroupCode IncomegroupName FlowCode
## 1 South of Sahara           10016            LDCs       11
## 2 South of Sahara           10016            LDCs       11
## 3 South of Sahara           10016            LDCs       11
## 4 South of Sahara           10016            LDCs       11
## 5 South of Sahara           10016            LDCs       11
## 6          Europe           10019           UMICs       14
##                                    FlowName Bi_Multi Category Finance_t Aid_t
## 1                                ODA Grants        1       10       110   C01
## 2                                ODA Grants        1       10       110   C01
## 3                                ODA Grants        1       10       110   D01
## 4                                ODA Grants        1       10       110   C01
## 5                                ODA Grants        4       10       110   C01
## 6 Other Official Flows (non Export Credit)        1       21       421
##   USD_Commitment USD_Disbursement USD_Received USD_Commitment_Defl
## 1       0.114020       0.03887000           NA          0.12308300
## 2       0.000000       0.00342274     0.000000          0.00000000
```

```
## 3       0.000000       0.02840270     0.000000          0.00000000
## 4       0.005840               NA           NA          0.00630417
## 5       0.121693       0.12169300     0.000000          0.11238300
## 6       0.000000       0.00000000     0.370032          0.00000000
##   USD_Disbursement_Defl USD_Received_Defl USD_Adjustment USD_Adjustment_Defl
## 1            0.04195950               NA             NA                   NA
## 2            0.00278702         0.000000             NA                   NA
## 3            0.02312740         0.000000             NA                   NA
## 4                    NA               NA             NA                   NA
## 5            0.11238300         0.000000             NA                   NA
## 6            0.00000000         0.345225             NA                   NA
##   USD_AmountUntied USD_AmountPartialTied USD_AmountTied USD_AmountUntied_Defl
## 1               NA                    NA        0.11402                   NA
## 2                0                     0        0.00000                    0
## 3                0                     0        0.00000                    0
## 4               NA                    NA        0.00584                   NA
## 5               NA                    NA             NA                   NA
## 6               NA                    NA             NA                   NA
##   USD_AmountPartialTied_Defl USD_Amounttied_Defl USD_IRTC USD_Expert_Commitment
## 1                         NA          0.12308300       NA                    NA
## 2                          0          0.00000000       NA                    NA
## 3                          0          0.00000000       NA                    NA
## 4                         NA          0.00630417       NA                    NA
## 5                         NA                  NA       NA                    NA
## 6                         NA                  NA       NA                    NA
##   USD_Expert_Extended USD_Export_Credit CurrencyCode Commitment_National
## 1                  NA                NA          302            0.114020
## 2                  NA                NA          301            0.000000
## 3                  NA                NA          301            0.000000
## 4                  NA                NA          302            0.005840
## 5                  NA                NA          302            0.121693
## 6                  NA                NA          302            0.000000
##   Disbursement_National GrantEquiv USD_GrantEquiv
## 1              0.038870         NA             NA
## 2              0.003420         NA             NA
## 3              0.028380         NA             NA
## 4                    NA         NA             NA
## 5              0.121693         NA             NA
## 6              0.000000         NA             NA
##                                                                                ShortDescription
## 1            LIBERIA MONITORING AND EVALUATION PROGRAM (L-MEP) - PROGRAM DESIGN AND LEARNING
## 2                                                          EDUCATION POLICY & ADMIN. MANAGEMENT
## 3 COLL\xc8GE MONTMORENCY - INTERNSHIPS 2010-2013 / COLL\xc8GE MONTMORENCY - STAGES 2010-2013
## 4                                                                ADMINISTRATION AND OVERSIGHT
## 5                                                 POP SVCS INT'L LIBERIA - YOUNG PEOPLE'S SRH
## 6                                                                              Semi-aggregates
##                                                                                   ProjectTitle
## 1            Liberia Monitoring and Evaluation Program (L-MEP) - Program Design and Learning
## 2
## 3 Coll\xe8ge Montmorency - Internships 2010-2013 / Coll\xe8ge Montmorency - Stages 2010-2013
## 4                                                                Administration and Oversight
## 5                                                 Pop Svcs Int'l Liberia - Young People's SRH
## 6
##   PurposeCode                                    PurposeName SectorCode
```

```
## 1       13010 Population policy and administrative management        130
## 2       11110  Education policy and administrative management        111
## 3       11110  Education policy and administrative management        111
## 4       13010 Population policy and administrative management        130
## 5       13010 Population policy and administrative management        130
## 6         240              II.4. Banking & Financial Services        240
##                                                 SectorName ChannelCode
## 1 I.3. Population Policies/Programmes & Reproductive Health       90000
## 2                      I.1.a. Education, Level Unspecified       90000
## 3                      I.1.a. Education, Level Unspecified       51000
## 4 I.3. Population Policies/Programmes & Reproductive Health       90000
## 5 I.3. Population Policies/Programmes & Reproductive Health       20000
## 6                       II.4. Banking & Financial Services          NA
##                                                                     ChannelName
## 1                                                                         Other
## 2                                                                         Other
## 3 University, college or other teaching institution, research institute or think-tank
## 4                                                                         Other
## 5                                Non-Governmental Organisation (NGO) and Civil Society
## 6
##      ChannelReportedName ParentChannelCode          Geography ExpectedStartDate
## 1                  Other             90000                         2010-08-27
## 2                  OTHER             90000                         2004-01-02
## 3 Coll\xe8ge Montmorency             51000                         2010-03-15
## 4                  Other             90000                         2012-01-01
## 5                     PN             20000 Liberia - Monrovia             <NA>
## 6                                       NA                             <NA>
##   CompletionDate
## 1     2010-08-27
## 2     2008-03-31
## 3     2013-02-28
## 4     2012-12-31
## 5           <NA>
## 6           <NA>
##
## 1
## 2
## 3 The International Youth Internship Program (IYIP) is an employment program for young Canadian profe
## 4
## 5
## 6
##   SDGfocus Gender Environment PDGG Trade RMNCH DRR Nutrition Disability FTC PBA
## 1       NA      1           2    0     0     2  NA        NA         NA  NA  NA
## 2       NA      1           0    0     0    NA  NA        NA         NA   1  NA
## 3       NA      1           1    1     0    NA  NA        NA         NA   1  NA
## 4       NA      0           0    0     0     2  NA        NA         NA  NA  NA
## 5       NA      0          NA   NA    NA    NA  NA        NA         NA  NA  NA
## 6       NA     NA          NA   NA    NA    NA  NA        NA         NA  NA  NA
##   InvestmentProject AssocFinance Biodiversity ClimateMitigation
## 1                NA           NA            2                 1
## 2                NA           NA            0                 0
## 3                NA           NA            0                 0
## 4                NA           NA            0                 0
## 5                NA           NA           NA                NA
```

```
## 6                NA           NA           NA                NA
##   ClimateAdaptation Desertification CommitmentDate TypeRepayment
## 1                 1              NA     2012-12-31            NA
## 2                 0               0           <NA>            NA
## 3                 0               0           <NA>            NA
## 4                 0              NA     2012-12-31            NA
## 5                NA              NA     2012-12-31            NA
## 6                NA              NA           <NA>            NA
##   NumberRepayment Interest1 Interest2 Repaydate1 Repaydate2 USD_Interest
## 1              NA        NA       <NA>       <NA>           NA
## 2              NA        NA       <NA>       <NA>           NA
## 3              NA        NA       <NA>       <NA>           NA
## 4              NA        NA       <NA>       <NA>           NA
## 5              NA        NA       <NA>       <NA>           NA
## 6              NA        NA       <NA>       <NA>           NA
##   USD_Outstanding USD_Arrears_Principal USD_Arrears_Interest BudgetIdent
## 1              NA                    NA                   NA       13010
## 2              NA                    NA                   NA       11110
## 3              NA                    NA                   NA       11110
## 4              NA                    NA                   NA       13010
## 5              NA                    NA                   NA       13010
## 6              NA                    NA                   NA          NA
##   CapitalExpend PSIflag PSIAddType PSIAddAssess PSIAddDevObj
## 1            NA      NA         NA           NA           NA
## 2            NA      NA         NA           NA           NA
## 3            NA      NA         NA           NA           NA
## 4            NA      NA         NA           NA           NA
## 5            NA      NA         NA           NA           NA
## 6            NA      NA         NA           NA           NA
```

The `summary` command returns descriptive statistics for all variables in the entire data set. Sometimes, this is too much at once and you only want summary statistics for a single variable. You can do this by specifying the variable you wish to summarize. R use the `$` symbol to find objects within a larger set of objects. For example, `mydata$variable1` tells R to find `variable1` in `mydata`. To summarize a single variable, use the following code:

```
summary(mydata$USD_Commitment)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -702.8310    0.0000    0.0135    1.3109    0.1283 1667.1500
```

The `summary` command gives us the range of the variable (the minimum `Min.` and the maximum `Max.`), the Median, the Mean, the first quartile (`1st Qu.`), the third quartile (`3rd Qu.`), and the number of missing values (`NA's`). The first and third quartile refer to statistics that you get from arranging the data from lowest value to highest, like when calculating the median, but instead of finding the halfway point of the data, you find the first 25% and the 75% respectively.

Notice that the `summary` command gives you a lot of information that you did not necessarily need. It also does not include the standard deviation, which is an important component of summarizing a variable. Instead of using the `summary` command, we can use more specific commands. The following section of code shows how to get the mean, median, and standard deviation of one variable. Notice how each section is preceded by a line with a `#` symbol and a desciption of the code below. This is symbol for a comment. A comment is just an annotation on your code to remind yourself or someone else what you are doing. R knows that nothing in a line with `#` should be run as code. Comments are very useful in R code because they are equivalent to notes. They can tell you which bit of code runs better or what each part of your script file should do. A good script file should have annotations.

```
#To get the Mean:
mean(mydata$USD_Commitment)
```

```
## [1] 1.310906
```

```
#To get the Median:
median(mydata$USD_Commitment)
```

```
## [1] 0.0135216
```

```
#To get the standard Deviation
sd(mydata$USD_Commitment)
```

```
## [1] 15.13912
```

However, the summary statistics above are only useful for continuous variables. Sometimes categorical variables are given numeric labels and R does not know that these are not continuous variables. This is because of the way R reads values. Variables are saved as "classes" or types: numeric, integer, and factor. Continuous variables will usually be saved as numeric. Categorical variables that use numbers as labels are usually saved as integers. Categorical variables with alphabetic names/labels are saved as factors. To find out how R saved a variable, you can use the `class` command:

```
class(mydata$USD_Commitment)
```

```
## [1] "numeric"
```

```
class(mydata$FlowCode)
```

```
## [1] "integer"
```

```
class(mydata$FlowName)
```

```
## [1] "character"
```

Notice that both `FlowName` and `FlowCode` contain the same information: `FlowCode` is a number label for `FlowName`. However, if you try to summarize `FlowCode`, it assumes this is a continuous variable and gives you the mean, median, range, and quartiles as usual:

```
summary(mydata$FlowCode)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   11.00   11.00   11.29   11.00   14.00
```

However, we know that this is not correct. A mean for a categorical variable does not tell us anything. The best way to summarize a categorical variable is to use a frequency table. To get R to do this for you, use the `table` command:

```
table(mydata$FlowCode)
```

```
##
##     11     13     14
## 160718  15511   7253
```

When R knows that it is dealing with a factor variable, the `summary` command gives you the same information as `table`:

```
summary(mydata$FlowName)
```

```
##    Length     Class      Mode
##    183482 character character
```

Loading Non- CSV Data

Data sets comes in many differ file types. One of the most comment file types, other than `.csv`, is a `.dta` file. This is a file from another statistical program called Stata. R itself cannot read or use this file. However, because R is an open source software, it works very well with user created packages. These packages allow you to use more functions than come standard with R. You can install a package using the `install.packages()` command. Once you have a package installed, you will only need to run the `install.packages()` command when you either update R or need to update the package (new versions are released periodically). The package needed to load a Stata data file is called the "foreign" package. You will need to install the foreign library. Then you will need to tell R that you want it to have access to a package by using the `library` command. You will need to use the `library` command each new session of R where you want to use that package. To install the "foreign" library, use the following code. I have included the `#` before the code to install the "foreign" library because I already have it on my computer. Be sure to take the `#` off before running this line of code yourself unless you already have the foreign library installed.

To tell R that you want to use the foreign package, use the following code:

```
library(foreign)
```

To load a Stata data file into R, you will use the foreign library and one of the commands in it: `read.dta`. This command follows many of the same rules as the `read.csv` command. The following code is used to load `.dta` files into R:

```
setwd("/Users/sarahhunter/Desktop/Data")

stata.data<-read.dta("nes2004subset2.dta")
```

From there, you can manipulate the Stata data using the same methods discussed above for `.csv` files.

```
names(stata.data)
```

```
##  [1] "religion"   "bush"       "female"     "unionhouse" "partyid"
##  [6] "eval_WoT"   "eval_HoE"   "ideology"   "bush_therm" "education"
## [11] "income"
```

```
summary(stata.data)
```

```
##     religion         bush           female        unionhouse
##  Min.   :1.00   Min.   :0.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:1.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :1.00   Median :1.000   Median :1.000   Median :0.0000
##  Mean   :2.31   Mean   :0.508   Mean   :0.533   Mean   :0.1716
##  3rd Qu.:2.00   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :7.00   Max.   :1.000   Max.   :1.000   Max.   :1.0000
##  NA's   :15     NA's   :401                     NA's   :6
##     partyid         eval_WoT          eval_HoE          ideology
##  Min.   :0.000   Min.   :-2.0000   Min.   :-2.0000   Min.   :1.00
##  1st Qu.:1.000   1st Qu.:-2.0000   1st Qu.:-2.0000   1st Qu.:3.00
##  Median :3.000   Median : 1.0000   Median :-1.0000   Median :4.00
##  Mean   :2.873   Mean   : 0.1533   Mean   :-0.3441   Mean   :4.27
##  3rd Qu.:5.000   3rd Qu.: 2.0000   3rd Qu.: 2.0000   3rd Qu.:6.00
##  Max.   :6.000   Max.   : 2.0000   Max.   : 2.0000   Max.   :7.00
##  NA's   :17      NA's   :31        NA's   :38        NA's   :292
##    bush_therm       education         income
##  Min.   :  0.00   Min.   :0.000   Min.   : 1.00
##  1st Qu.: 30.00   1st Qu.:3.000   1st Qu.:11.00
##  Median : 60.00   Median :4.000   Median :16.00
##  Mean   : 54.94   Mean   :4.303   Mean   :14.94
```

```
##  3rd Qu.: 85.00   3rd Qu.:6.000   3rd Qu.:20.00
##  Max.   :100.00   Max.   :7.000   Max.   :23.00
##  NA's   :5                        NA's   :142
```