W203 Lab 1: Candidate Debt EDA

Yulia and Mitch

January 23, 2018

```r
# correcting variable names
var_names <- colnames(read.csv("CandidateDebt.csv", nrows = 1))

var_names_corrected <- c(var_names[1:grep("position", var_names)],
                         "position2",
                         var_names[(grep("position", var_names) + 1):(length(var_names) - 1)])
```

```r
# reading the data with correct headers
CandidateDebt <- read.csv("CandidateDebt.csv",
                          stringsAsFactors = FALSE,
                          col.names = var_names_corrected)
```

```r
str(CandidateDebt)
```

```
## 'data.frame':    1043 obs. of  28 variables:
##  $ reportnumber      : int  100495995 100496548 100498383 100495987 100496259 100496199 100496375 10
##  $ origin            : chr  "B.3" "B.3" "B.3" "B.3" ...
##  $ filerid           : chr  "RYU C  133" "THOMT  368" "FEY J  422" "STRAS  111" ...
##  $ filertype         : chr  "Candidate" "Candidate" "Candidate" "Candidate" ...
##  $ filername         : chr  "RYU CINDY S" "THOMAS TIMOTHY N JR" "FEY JACOB C" "STRACHAN STEVEN D" .
##  $ firstname         : chr  "CINDY" "TIMOTHY" "JACOB" "STEVEN" ...
##  $ middleinitial     : chr  "S" "N" "C" "D" ...
##  $ lastname          : chr  "RYU" "THOMAS" "FEY" "STRACHAN" ...
##  $ office            : chr  "STATE REPRESENTATIVE" "COUNTY COMMISSIONER" "STATE REPRESENTATIVE" "COU
##  $ legislativedistrict: chr  "STATE SENATOR" "STATE SENATOR" "STATE SENATOR" "STATE SENATOR" ...
##  $ position          : chr  "1" "1" "1" "1" ...
##  $ position2         : chr  "" "" "" "" ...
##  $ party             : chr  "REPUBLICAN" "REPUBLICAN" "REPUBLICAN" "REPUBLICAN" ...
##  $ jurisdiction      : chr  "LEG DISTRICT 01 - SENATE" "LEG DISTRICT 01 - SENATE" "LEG DISTRICT 01 -
##  $ jurisdictioncounty : chr  "KING" "KING" "KING" "KING" ...
##  $ jurisdictiontype  : chr  "Legislative" "Legislative" "Legislative" "Legislative" ...
##  $ electionyear      : chr  "2012" "2012" "2012" "2012" ...
##  $ amount            : chr  "283.25" "283.25" "283.25" "283.25" ...
##  $ recordtype        : chr  "DEBT" "DEBT" "DEBT" "DEBT" ...
##  $ fromdate          : chr  "6/1/12" "6/1/12" "6/1/12" "6/1/12" ...
##  $ thrudate          : chr  "7/16/12" "7/16/12" "7/16/12" "7/16/12" ...
##  $ debtdate          : chr  "7/3/12" "7/3/12" "7/3/12" "7/3/12" ...
##  $ code              : chr  "" "" "" "" ...
##  $ description       : chr  "RE-ORDER TEE SHIRTS" "RE-ORDER TEE SHIRTS" "RE-ORDER TEE SHIRTS" "RE-OR
##  $ vendorname        : chr  "HICKEY GAYLE" "HICKEY GAYLE" "HICKEY GAYLE" "HICKEY GAYLE" ...
##  $ vendoraddress     : chr  "PO BOX 2749" "PO BOX 2749" "PO BOX 2749" "PO BOX 2749" ...
##  $ vendorcity        : chr  "WOODINVILLE " "WOODINVILLE " "WOODINVILLE " "WOODINVILLE " ...
##  $ vendorstate       : chr  "WA" "WA" "WA" "WA" ...
```

```r
# remove records with missing record type (missing values are miscoded as #NA)
CandidateDebt <- CandidateDebt[!grepl('#N/A', CandidateDebt$recordtype),]
```

```r
# converting amount from character to numeric
CandidateDebt$amount <- as.numeric(CandidateDebt$amount)
```

```r
summary(CandidateDebt$amount)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.      Max.
##     3.24   283.25   300.00  1347.42  1210.50  19000.00
```

```r
# converting dates from character to dates
CandidateDebt$fromdate <- as.Date(CandidateDebt$fromdate, format = "%m/%d/%Y")
CandidateDebt$thrudate <- as.Date(CandidateDebt$thrudate, format = "%m/%d/%Y")
CandidateDebt$debtdate <- as.Date(CandidateDebt$debtdate, format = "%m/%d/%Y")

# creating a vector of variables to keep for analysis
keep_vars <- c("reportnumber", "filerid", "filername", "office", "legislativedistrict",
    "party", "jurisdiction", "jurisdictioncounty", "jurisdictiontype",
    "amount", "fromdate", "thrudate", "debtdate", "code", "description",
    "vendorname", "vendorcity", "vendorstate")

CandidateDebt <- CandidateDebt[keep_vars]
```

```r
str(CandidateDebt)
```

```
## 'data.frame':    987 obs. of  18 variables:
##  $ reportnumber       : int  100495995 100496548 100498383 100495987 100496259 100496199 100496375 10
##  $ filerid            : chr  "RYU C  133" "THOMT  368" "FEY J  422" "STRAS  111" ...
##  $ filername          : chr  "RYU CINDY S" "THOMAS TIMOTHY N JR" "FEY JACOB C" "STRACHAN STEVEN D" .
##  $ office             : chr  "STATE REPRESENTATIVE" "COUNTY COMMISSIONER" "STATE REPRESENTATIVE" "COU
##  $ legislativedistrict: chr  "STATE SENATOR" "STATE SENATOR" "STATE SENATOR" "STATE SENATOR" ...
##  $ party              : chr  "REPUBLICAN" "REPUBLICAN" "REPUBLICAN" "REPUBLICAN" ...
##  $ jurisdiction       : chr  "LEG DISTRICT 01 - SENATE" "LEG DISTRICT 01 - SENATE" "LEG DISTRICT 01 -
##  $ jurisdictioncounty : chr  "KING" "KING" "KING" "KING" ...
##  $ jurisdictiontype   : chr  "Legislative" "Legislative" "Legislative" "Legislative" ...
##  $ amount             : num  283 283 283 283 283 ...
##  $ fromdate           : Date, format: "12-06-01" "12-06-01" ...
##  $ thrudate           : Date, format: "12-07-16" "12-07-16" ...
##  $ debtdate           : Date, format: "12-07-03" "12-07-03" ...
##  $ code               : chr  "" "" "" "" ...
##  $ description        : chr  "RE-ORDER TEE SHIRTS" "RE-ORDER TEE SHIRTS" "RE-ORDER TEE SHIRTS" "RE-OR
##  $ vendorname         : chr  "HICKEY GAYLE" "HICKEY GAYLE" "HICKEY GAYLE" "HICKEY GAYLE" ...
##  $ vendorcity         : chr  "WOODINVILLE " "WOODINVILLE " "WOODINVILLE " "WOODINVILLE " ...
##  $ vendorstate        : chr  "WA" "WA" "WA" "WA" ...
```

```r
summary(CandidateDebt)
```

```
##   reportnumber        filerid            filername
##  Min.   :100346104   Length:987         Length:987
##  1st Qu.:100446276   Class :character   Class :character
##  Median :100471547   Mode  :character   Mode  :character
##  Mean   :100466089
##  3rd Qu.:100494036
##  Max.   :100599472
##     office           legislativedistrict    party
##  Length:987          Length:987           Length:987
##  Class :character    Class :character     Class :character
##  Mode  :character    Mode  :character     Mode  :character
##
##
##
```

```
##   jurisdiction       jurisdictioncounty jurisdictiontype
##   Length:987         Length:987         Length:987
##   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      amount            fromdate            thrudate
##   Min.   :    3.24   Min.   :9-10-01    Min.   :9-10-31
##   1st Qu.:  283.25   1st Qu.:11-10-01   1st Qu.:11-10-31
##   Median :  300.00   Median :12-02-01   Median :12-02-29
##   Mean   : 1347.42   Mean   :11-12-19   Mean   :12-01-20
##   3rd Qu.: 1210.50   3rd Qu.:12-06-01   3rd Qu.:12-07-16
##   Max.   :19000.00   Max.   :12-08-01   Max.   :12-08-31
##      debtdate            code            description
##   Min.   :8-10-29    Length:987         Length:987
##   1st Qu.:11-07-03   Class :character   Class :character
##   Median :12-02-29   Mode  :character   Mode  :character
##   Mean   :11-12-13
##   3rd Qu.:12-07-03
##   Max.   :12-08-31
##    vendorname          vendorcity          vendorstate
##   Length:987         Length:987         Length:987
##   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
# checking if there are any "#NA" values left
sapply(CandidateDebt, function(x) sum(grepl('#NA', x)))
```

```
##        reportnumber             filerid            filername
##                   0                   0                    0
##              office legislativedistrict                party
##                   0                   0                    0
##        jurisdiction  jurisdictioncounty     jurisdictiontype
##                   0                   0                    0
##              amount            fromdate             thrudate
##                   0                   0                    0
##            debtdate                code          description
##                   0                   0                    0
##          vendorname          vendorcity          vendorstate
##                   0                   0                    0
```

```r
# checking for presense of missing values
sum(is.na(CandidateDebt))
```

```
## [1] 0
```

```r
# checking some variables with smaller number of unique values
cat_vars <- c("office", "legislativedistrict", "party", "jurisdiction", "jurisdictioncounty",
              "jurisdictiontype", "fromdate", "thrudate", "debtdate", "code",
              "vendorcity", "vendorstate")
sapply(CandidateDebt[, cat_vars], function(x) table(x))
```

```
## $office
## x
##           APPEALS COURT JUDGE              ATTORNEY GENERAL
##                          4                            34
##               COUNTY ASSESSOR          COUNTY COMMISSIONER
##                         19                            72
##          COUNTY COUNCIL MEMBER               COUNTY SHERIFF
##                         15                            17
##                     GOVERNOR   PUBLIC LANDS COMMISSIONER
##                         42                            27
## PUBLIC UTILITY COMMISSIONER          SECRETARY OF STATE
##                          8                            11
##                STATE AUDITOR        STATE REPRESENTATIVE
##                          7                           528
##                STATE SENATOR STATE SUPREME COURT JUSTICE
##                        118                            28
##              STATE TREASURER          SUPERIOR COURT JUDGE
##                         24                            33
##
## $legislativedistrict
## x
##            ATTORNEY GENERAL              COUNTY ASSESSOR
##                         49                             2
##         COUNTY COMMISSIONER       COUNTY COUNCIL MEMBER
##                         44                             3
##               COUNTY SHERIFF                    GOVERNOR
##                          8                           101
##    PUBLIC LANDS COMMISSIONER PUBLIC UTILITY COMMISSIONER
##                         21                             2
##                STATE AUDITOR        STATE REPRESENTATIVE
##                          4                           384
##                STATE SENATOR STATE SUPREME COURT JUSTICE
##                        305                            12
##              STATE TREASURER          SUPERIOR COURT JUDGE
##                         28                            24
##
## $party
## x
##     DEMOCRAT  INDEPENDENT NON PARTISAN   REPUBLICAN
##          638            2           48          299
##
## $jurisdiction
## x
## ATTORNEY GENERAL, OFFICE OF   AUDITOR, OFFICE OF STATE
##                         49                             4
##                  CLALLAM CO          GOVERNOR, OFFICE OF
##                         10                           101
##                  ISLAND CO                 JEFFERSON CO
##                          6                             1
##                    KING CO    KING CO SUPERIOR COURT
##                          8                            21
##     LEG DISTRICT 01 - HOUSE    LEG DISTRICT 01 - SENATE
##                          3                           243
##     LEG DISTRICT 03 - HOUSE    LEG DISTRICT 03 - SENATE
```

```
##                          1                          7
##     LEG DISTRICT 05 - SENATE    LEG DISTRICT 08 - HOUSE
##                          3                          8
##     LEG DISTRICT 11 - HOUSE    LEG DISTRICT 11 - SENATE
##                         13                         30
##     LEG DISTRICT 14 - HOUSE    LEG DISTRICT 21 - HOUSE
##                          2                         55
##     LEG DISTRICT 22 - HOUSE    LEG DISTRICT 23 - HOUSE
##                          1                          3
##     LEG DISTRICT 24 - HOUSE    LEG DISTRICT 25 - HOUSE
##                          5                          4
##     LEG DISTRICT 26 - HOUSE    LEG DISTRICT 27 - HOUSE
##                         37                          6
##     LEG DISTRICT 28 - HOUSE    LEG DISTRICT 29 - HOUSE
##                         21                          2
##     LEG DISTRICT 32 - HOUSE    LEG DISTRICT 34 - HOUSE
##                          6                          7
##     LEG DISTRICT 36 - HOUSE    LEG DISTRICT 37 - HOUSE
##                          9                          1
##     LEG DISTRICT 40 - HOUSE    LEG DISTRICT 40 - SENATE
##                          6                         16
##     LEG DISTRICT 41 - HOUSE    LEG DISTRICT 41 - SENATE
##                          5                          5
##     LEG DISTRICT 43 - HOUSE    LEG DISTRICT 44 - HOUSE
##                         72                          1
##     LEG DISTRICT 45 - HOUSE    LEG DISTRICT 46 - HOUSE
##                         68                         17
##     LEG DISTRICT 46 - SENATE    LEG DISTRICT 47 - HOUSE
##                          1                         13
##     LEG DISTRICT 48 - HOUSE                    MASON CO
##                         18                         15
##  NATURAL RESOURCES, DEPT OF  OKANOGAN CO SUPERIOR COURT
##                         21                          1
##                   PIERCE CO    PIERCE CO SUPERIOR COURT
##                          5                          1
##                   SKAGIT CO               SUPREME COURT
##                         12                         12
##  THURSTON CO SUPERIOR COURT    THURSTON PUD 01, 02, 03
##                          1                          2
##  TREASURER, OFFICE OF STATE
##                         28
## 
## $jurisdictioncounty
## x
##            BENTON    CLALLAM     ISLAND  JEFFERSON       KING     KITSAP
##       215         8         15          6          1        544          3
##     MASON   OKANOGAN     PIERCE     SKAGIT  SNOHOMISH    SPOKANE   THURSTON
##        15          1         76         34         55          8          4
##    YAKIMA
##         2
## 
## $jurisdictiontype
## x
##    Judicial Legislative      Local    Statewide
```

5

```
##           36           689           59          203
##
## $fromdate
## x
##  9-10-01  9-11-01  9-12-01 10-01-01 10-02-01 10-09-01 10-10-01 10-11-01
##       16        5        1        1        6        2        6       16
## 10-12-01 11-01-01 11-02-01 11-03-01 11-04-01 11-05-01 11-06-01 11-07-01
##        6       10       13       28       20        2       58        5
## 11-08-01 11-09-01 11-10-01 11-11-01 11-12-01 12-01-01 12-01-06 12-01-17
##        7       41       44       76       10       24        4        4
## 12-02-01 12-02-06 12-03-01 12-04-01 12-04-13 12-04-26 12-05-01 12-05-19
##      102        1        9       40        3        5       56       10
## 12-05-25 12-06-01 12-07-17 12-07-31 12-08-01
##       11      283       31       30        1
##
## $thrudate
## x
##  9-10-31  9-11-30  9-12-31 10-01-31 10-02-28 10-09-30 10-10-31 10-11-30
##       16        5        1        1        6        2        6       16
## 10-12-31 11-01-31 11-02-28 11-03-31 11-04-30 11-05-31 11-06-30 11-07-31
##        6        8       15       28       18        4       58        5
## 11-08-31 11-09-30 11-10-31 11-11-30 11-12-31 12-01-31 12-02-29 12-03-31
##        7       41       44       76        9       32      102       10
## 12-04-30 12-05-31 12-06-30 12-07-16 12-07-30 12-08-31
##       49       77       14      269       31       31
##
## $debtdate
## x
##  8-10-29  9-02-01  9-02-28  9-04-30  9-11-30  9-12-31 10-01-22 10-09-30
##        4        2       18        4        1        1        1        2
## 10-10-31 10-11-05 10-11-30 10-12-31 11-01-21 11-01-26 11-01-28 11-01-31
##        6       21       18       10        4        1        4        7
## 11-02-17 11-02-22 11-02-28 11-03-31 11-04-26 11-04-30 11-05-02 11-05-06
##        4       14        6       10       12        4       31        1
## 11-05-24 11-05-31 11-06-27 11-06-30 11-07-01 11-07-06 11-07-27 11-07-29
##        1        1        1       57        1        9        1        1
## 11-07-31 11-08-25 11-08-31 11-09-30 11-10-16 11-10-18 11-10-31 11-11-02
##        1        1        6       19        8        1       25        2
## 11-11-11 11-11-30 11-12-23 11-12-31 12-01-25 12-01-31 12-02-01 12-02-24
##        5       69        1        1        1       29       19        1
## 12-02-28 12-02-29 12-03-30 12-03-31 12-04-02 12-04-11 12-04-18 12-04-27
##        1       88        5       14        2        1        2        5
## 12-04-30 12-05-01 12-05-15 12-05-31 12-06-24 12-06-27 12-06-30 12-07-03
##       41        1        4       47        3        4       13      241
## 12-07-04 12-07-13 12-07-16 12-07-19 12-07-30 12-08-28 12-08-29 12-08-31
##        1        2       19        2       18        1        1       24
##
## $code
## x
##                                  Fundraising    Management Services
##                    610                    5                     10
## Operation and Overhead
##                    362
##
```

```
## $vendorcity
## x
##                 BAINBRIDGE ISLAND                BELFAIR  CITY OF INDUSTRY
##                               24                       3                13                        1
##               DALLAS           FRIDAY HARBOR           GIG HARBOR                 ISSAQUAH
##                                5                       5                 5                        1
##            KENNEWICK                KIRKLAND       MERCER ISLAND           MOUNTAIN VIEW
##                                2                      38                 1                        3
##           OAK HARBOR                 OLYMPIA        PORT ORCHARD           PORT TOWNSEND
##                                6                      20                10                        1
##             PUYALLUP              SACRAMENTO            SAN JOSE             SANTA MONICA
##                                5                       3                 1                        3
##              SEATTLE                  SEQUIM             SHELTON                  SPOKANE
##                              452                       3                 2                        2
##       SPOKANE VALLEY                  TACOMA            TUMWATER         UNIVERSITY PLACE
##                                1                      24                 1                       11
##           WASHINGTON             WOODINVILLE
##                              100                     241
##
## $vendorstate
## x
##      CA   DC   TX   WA
##  25   10  100    5  847
```

```r
save(CandidateDebt, file = "CandidateDebt.RData")
```