W203 Lab 1: Candidate Debt EDA

Yulia and Mitch

January 29, 2018

1. Introduction

1.1 Introduction

1.2 Loading Data

```
CandidateDebt <- read.csv("CandidateDebt.csv", stringsAsFactors = FALSE)
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" ...
##  $ party              : chr  "" "" "" "" ...
##  $ jurisdiction       : chr  "REPUBLICAN" "REPUBLICAN" "REPUBLICAN" "REPUBLICAN" ...
##  $ jurisdictioncounty : chr  "LEG DISTRICT 01 - SENATE" "LEG DISTRICT 01 - SENATE" "LEG DISTRICT 01 -
##  $ jurisdictiontype   : chr  "KING" "KING" "KING" "KING" ...
##  $ electionyear       : chr  "Legislative" "Legislative" "Legislative" "Legislative" ...
##  $ amount             : chr  "2012" "2012" "2012" "2012" ...
##  $ recordtype         : chr  "283.25" "283.25" "283.25" "283.25" ...
##  $ fromdate           : chr  "DEBT" "DEBT" "DEBT" "DEBT" ...
##  $ thrudate           : chr  "6/1/12" "6/1/12" "6/1/12" "6/1/12" ...
##  $ debtdate           : chr  "7/16/12" "7/16/12" "7/16/12" "7/16/12" ...
##  $ code               : chr  "7/3/12" "7/3/12" "7/3/12" "7/3/12" ...
##  $ description        : chr  "" "" "" "" ...
##  $ vendorname         : chr  "RE-ORDER TEE SHIRTS" "RE-ORDER TEE SHIRTS" "RE-ORDER TEE SHIRTS" "RE-OI
##  $ vendoraddress      : chr  "HICKEY GAYLE" "HICKEY GAYLE" "HICKEY GAYLE" "HICKEY GAYLE" ...
##  $ vendorcity         : chr  "PO BOX 2749" "PO BOX 2749" "PO BOX 2749" "PO BOX 2749" ...
##  $ vendorstate        : chr  "WOODINVILLE " "WOODINVILLE " "WOODINVILLE " "WOODINVILLE " ...
##  $ vendorzip          : chr  "WA" "WA" "WA" "WA" ...
```

Problems with target variable *amount*:

```
table(CandidateDebt$amount)
```

```
##
## #N/A 2012
##   56  987
```

Resolution: shift column names:

```r
# get column names from row data
var_names <- colnames(read.csv("CandidateDebt.csv", nrows = 1))

# insert column after 'position' and remove last column
var_names_corrected <- c(var_names[1:grep("position", var_names)],
    "position2", var_names[(grep("position", var_names) + 1):(length(var_names) -
        1)])
```

Re-loading raw data:

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

Description of data set:

Blah Blah Blah

```r
dim(CandidateDebt)
```

```
## [1] 1043   28
```

```r
# Converting target variable to numeric
CandidateDebt$amount_num <- as.numeric(CandidateDebt$amount)
summary(CandidateDebt$amount_num)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##     3.24   283.25   300.00  1347.42  1210.50 19000.00 
##     NA's 
##       56
```

1.2 Exploring rows with missing debt data

```r
# creating flag for missing values (1 for missing)
CandidateDebt$missing_amount <- ifelse(is.na(CandidateDebt$amount_num),
    1, 0)
table(CandidateDebt$missing_amount)
```

```
## 
##   0   1 
## 987  56
```

While exploring 56 rows with missing data, we discovered that those rows are missing data in all columns except filer name and office they run for. Good news is we are losing only one candidate if we exclude those 56 rows from the analysis. No unique values of *office* variable are among 56 rows.

```r
# number of of unique filer ids (candidates in full dataset)
length(unique(CandidateDebt$filerid))
```

```
## [1] 141
```

```r
# number of unique filer ids (candidates) in data set without
# 56 rows with missing data:
length(unique(CandidateDebt[CandidateDebt$missing_amount == 0,
    ]$filerid))
```

```
## [1] 140
```

```r
# number of of unique values of office (candidates in full
# dataset)
length(unique(CandidateDebt$office))
```

```
## [1] 16
```

```r
# number of unique values of office in data set without 56
# rows with missing data:
length(unique(CandidateDebt[CandidateDebt$missing_amount == 0,
    ]$office))
```

```
## [1] 16
```

```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")
```

1.3 Creating analytic dataset

Exlcude variables:

- origin (one value = B.3)
- filertype (one value = Candidate)
- filername, firstname, middleinitial, lastname (will use filerid as a candidate identifier)
- position and position2 (values are not clear and were messed up in raw data)
- electionyear (one value = 2012)
- recordtype (one value = DEBT)

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

# removing 56 rows with missing data
CandidateDebtSub <- CandidateDebt[CandidateDebt$missing_amount ==
    0, ]
CandidateDebtSub <- CandidateDebtSub[keep_vars]
rm(keep_vars)
```

Looking at main analytic dataset:

```r
summary(CandidateDebtSub)
```

```
##   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_num          fromdate
##  Min.   :    3.24   Min.   :2009-10-01
##  1st Qu.:  283.25   1st Qu.:2011-10-01
##  Median :  300.00   Median :2012-02-01
##  Mean   : 1347.42   Mean   :2011-12-19
##  3rd Qu.: 1210.50   3rd Qu.:2012-06-01
##  Max.   :19000.00   Max.   :2012-08-01
##     thrudate             debtdate
##  Min.   :2009-10-31   Min.   :2008-10-29
##  1st Qu.:2011-10-31   1st Qu.:2011-07-03
##  Median :2012-02-29   Median :2012-02-29
##  Mean   :2012-01-20   Mean   :2011-12-13
##  3rd Qu.:2012-07-16   3rd Qu.:2012-07-03
##  Max.   :2012-08-31   Max.   :2012-08-31
##      code           description         vendorname
##  Length:987          Length:987         Length:987
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  vendoraddress       vendorcity        vendorstate
##  Length:987          Length:987         Length:987
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

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

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

1.4 Evaluating data quality

Calculating number of unique values per candidate for campaign related variable

```
aggr_office <- aggregate(amount_num ~ filerid + office, data = CandidateDebtSub,
    sum)
aggr_office <- aggregate(office ~ filerid, data = aggr_office,
```

```r
    length)

aggr_legdis <- aggregate(amount_num ~ filerid + legislativedistrict,
    data = CandidateDebtSub, sum)
aggr_legdis <- aggregate(legislativedistrict ~ filerid, data = aggr_legdis,
    length)

aggr_party <- aggregate(amount_num ~ filerid + party, data = CandidateDebtSub,
    sum)
aggr_party <- aggregate(party ~ filerid, data = aggr_party, length)

aggr_jur <- aggregate(amount_num ~ filerid + jurisdiction, data = CandidateDebtSub,
    sum)
aggr_jur <- aggregate(jurisdiction ~ filerid, data = aggr_jur,
    length)

aggr_jurc <- aggregate(amount_num ~ filerid + jurisdictioncounty,
    data = CandidateDebtSub, sum)
aggr_jurc <- aggregate(jurisdictioncounty ~ filerid, data = aggr_jurc,
    length)

aggr_jurt <- aggregate(amount_num ~ filerid + jurisdictiontype,
    data = CandidateDebtSub, sum)
aggr_jurt <- aggregate(jurisdictiontype ~ filerid, data = aggr_jurt,
    length)

aggr_comb <- cbind(aggr_office, aggr_legdis[, 2], aggr_party[,
    2], aggr_jur[, 2], aggr_jurc[, 2], aggr_jurt[, 2])

colnames(aggr_comb) <- c("filerid", "office", "legislativedistrict",
    "party", "jurisdiction", "jurisdictioncounty", "jurisdictiontype")
rm(list = c("aggr_office", "aggr_legdis", "aggr_party", "aggr_jur",
    "aggr_jurc", "aggr_jurt"))

# knitr::kable(summary(aggr_comb[, -1]), caption = 'Table
# with kable')
summary(aggr_comb[, -1])
```

```
##      office  legislativedistrict     party
##  Min.   :1   Min.   :1.000       Min.   :1.000
##  1st Qu.:1   1st Qu.:1.000       1st Qu.:1.000
##  Median :1   Median :3.000       Median :2.000
##  Mean   :1   Mean   :2.943       Mean   :1.836
##  3rd Qu.:1   3rd Qu.:4.000       3rd Qu.:2.000
##  Max.   :1   Max.   :8.000       Max.   :3.000
##   jurisdiction    jurisdictioncounty jurisdictiontype
##  Min.   : 1.000   Min.   :1.000      Min.   :1.000
##  1st Qu.: 2.000   1st Qu.:1.000      1st Qu.:1.000
##  Median : 3.000   Median :3.000      Median :2.000
##  Mean   : 4.457   Mean   :2.693      Mean   :2.057
##  3rd Qu.: 6.250   3rd Qu.:4.000      3rd Qu.:3.000
##  Max.   :14.000   Max.   :6.000      Max.   :4.000
```

The results of this preliminary analysis are not encouraging, and indicate that several fields that would otherwise be of interest to us are not completely accurate. More specifically, the variables legislative district, party, jurisdiction, jurisdictioncounty, and jurisditctiontype, each have instances in which the same candidate has more than one value in the dataset. Given that candidates can only have one value for each of these in a given election cycle, this suggests that some or all of the values contained in these columns is not reliable. To avoid making recommendations on inaccurate data, this analysis will exclude these variables, and provide guidance for how the [CLINT] can best improve data quality moving forward.

—Based on the above, we think all but *office* variables are unreliable

```
# creating flag variables for candidates with more than 1
# unique value
aggr_comb$legdist_mult <- ifelse(aggr_comb$legislativedistrict >
    1, 1, 0)
aggr_comb$party_mult <- ifelse(aggr_comb$party > 1, 1, 0)
aggr_comb$jur_mult <- ifelse(aggr_comb$jurisdiction > 1, 1, 0)
aggr_comb$jurc_mult <- ifelse(aggr_comb$jurisdictioncounty >
    1, 1, 0)
aggr_comb$jurt_mult <- ifelse(aggr_comb$jurisdictiontype > 1,
    1, 0)
aggr_comb$mult <- aggr_comb$legdist_mult + aggr_comb$party_mult +
    aggr_comb$jur_mult + aggr_comb$jurc_mult + aggr_comb$jurt_mult
table(aggr_comb$mult)
```

```
##
##  0  1  2  3  4  5
## 34  2  3  4 15 82
```

Only 34 candidates with “clean” data

```
# adding this flag variable to the main data set
CandidateDebtSub <- merge(CandidateDebtSub, aggr_comb[, c("filerid",
    "mult")], by = "filerid")
rm(aggr_comb)
```

```
# counting number of unique offices among those 34 candidates
length(unique(CandidateDebtSub$office[CandidateDebtSub$mult ==
    0]))
```

```
## [1] 9
```

```
# counting number of unique parties/offices among those 34
# candidates
aggr_party <- aggregate(amount_num ~ filerid + party + office,
    data = CandidateDebtSub[CandidateDebtSub$mult == 0, ], sum)
table(aggr_party$office, aggr_party$party)
```

```
##
##                                  DEMOCRAT NON PARTISAN
##   ATTORNEY GENERAL                      0            0
##   COUNTY COMMISSIONER                   3            2
##   GOVERNOR                              0            0
##   PUBLIC UTILITY COMMISSIONER           0            0
##   SECRETARY OF STATE                    0            0
##   STATE REPRESENTATIVE                  4            2
##   STATE SENATOR                         0            0
##   STATE SUPREME COURT JUSTICE           1            0
##   SUPERIOR COURT JUDGE                  3            1
##
```

```
##                                REPUBLICAN
##   ATTORNEY GENERAL                      1
##   COUNTY COMMISSIONER                   0
##   GOVERNOR                              1
##   PUBLIC UTILITY COMMISSIONER           2
##   SECRETARY OF STATE                    1
##   STATE REPRESENTATIVE                  9
##   STATE SENATOR                         2
##   STATE SUPREME COURT JUSTICE           0
##   SUPERIOR COURT JUDGE                  2
```

```
rm(aggr_party)
```

Based on the above, only “State Prepresentative” and “Superior Court Judge” had representatives of two major parties.This is suspect. Hence, we will eclude the following 5 variables from the analysis: *legislativedistrict*, *party*, *jurisdiction*, *jurisdictioncounty*, *jurisdictiontype*

1.5 Creating extra variables

Processing date variables

```
summary(CandidateDebtSub$debtdate)
```

```
##         Min.      1st Qu.       Median         Mean 
## "2008-10-29" "2011-07-03" "2012-02-29" "2011-12-13" 
##      3rd Qu.         Max. 
## "2012-07-03" "2012-08-31"
```

```
summary(CandidateDebtSub$fromdate)
```

```
##         Min.      1st Qu.       Median         Mean 
## "2009-10-01" "2011-10-01" "2012-02-01" "2011-12-19" 
##      3rd Qu.         Max. 
## "2012-06-01" "2012-08-01"
```

```
summary(CandidateDebtSub$thrudate)
```

```
##         Min.      1st Qu.       Median         Mean 
## "2009-10-31" "2011-10-31" "2012-02-29" "2012-01-20" 
##      3rd Qu.         Max. 
## "2012-07-16" "2012-08-31"
```

Based on the above we will assume that the election was in August 2012

```
# Number of months before election the debt occured
CandidateDebtSub$weeksindebt <- round(difftime(max(CandidateDebtSub$debtdate),
    CandidateDebtSub$debtdate, units = "weeks"))
CandidateDebtSub$monthsindebt <- round(CandidateDebtSub$weeksindebt/52 *
    12)
CandidateDebtSub$monthsindebt <- as.numeric(CandidateDebtSub$monthsindebt)
# capping months at 13 months (for exploratory reasons)
CandidateDebtSub$monthsindebt_cap <- ifelse(CandidateDebtSub$monthsindebt >
    12, 13, CandidateDebtSub$monthsindebt)
summary(CandidateDebtSub$monthsindebt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   2.000   6.000   8.583  14.000  46.000
```

```
summary(CandidateDebtSub$monthsindebt_cap)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    2.00    6.00    6.73   13.00   13.00
```

Recoding debt *description* variable to make it more digestable

```
creditcard <- c("AM EX", "AMERICAN EXPRESS", "AMERICAN EXPRESS LOWES",
    "AMEX", "CITI MASTERCARD", "MASTERCARD", "VISA", "CAPITOL ONE",
    "MASTER CARD")
consulting <- c("CONSULTING", "JANUARY SERVICES", "$750 PER MONTH THROUGH OCTOBER",
    "AUGUST CONSULTING", "CONSULTING ESTIMATE", "CONSULTING/PHOTOGRAPHY",
    "CONSULTING/TRAVEL", "MAY CONSULTING SERVICES", "MONTHLY CONSULTING FEE",
    "RETAINER", "APRIL RETAINER")
swag <- c("RE-ORDER TEE SHIRTS", "BUMPER STICKERS/FLYERS", "CONSULTING/YARD SIGNS",
    "YARD SIGNS", "OFFICE SUPPLIES/ WATER FOR KICKOFF")

CandidateDebtSub$description_aggr[grepl("TREASURY", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "TREASURY"
CandidateDebtSub$description_aggr[grepl("CAMPAIGN", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "CAMPAIGN MANAGEMENT"
CandidateDebtSub$description_aggr[grepl("FUND", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "FUNDRAISING"
CandidateDebtSub$description_aggr[grepl("CARRY FORWARD", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "CARRY FORWARD"
CandidateDebtSub$description_aggr[grepl("REIMB", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "REIMBURSEMENT"
CandidateDebtSub$description_aggr[grepl("ACCOUNTING", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "ACCOUNTING"
CandidateDebtSub$description_aggr[grepl("BONUS", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "BONUS"
CandidateDebtSub$description_aggr[grepl("DESIGN", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "DESIGN/PRINT"
CandidateDebtSub$description_aggr[grepl("PRINT", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "DESIGN/PRINT"
CandidateDebtSub$description_aggr[grepl("POLLING", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "POLLING"
CandidateDebtSub$description_aggr[grepl("CREDIT", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "CREDIT CARD"
CandidateDebtSub$description_aggr[CandidateDebtSub$vendorname %in%
    creditcard] <- "CREDIT CARD"
CandidateDebtSub$description_aggr[CandidateDebtSub$description %in%
    consulting] <- "CONSULTING"
CandidateDebtSub$description_aggr[CandidateDebtSub$description %in%
    swag] <- "SWAG"
CandidateDebtSub$description_aggr[grepl("MAIL", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "MAIL"
CandidateDebtSub$description_aggr[grepl("POSTAGE", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "MAIL"
CandidateDebtSub$description_aggr[grepl("STAMPS", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "MAIL"
CandidateDebtSub$description_aggr[grepl("DATA", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "DATA/TECH/AD"
CandidateDebtSub$description_aggr[grepl("DISPLAY", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "DATA/TECH/AD"
```

```
CandidateDebtSub$description_aggr[grepl("WEB", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "DATA/TECH/AD"
CandidateDebtSub$description_aggr[grepl("ADVERTISEMENT", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "DATA/TECH/AD"
CandidateDebtSub$description_aggr[grepl("COMPUTER", CandidateDebtSub$description,
    ignore.case = TRUE)] <- "DATA/TECH/AD"
CandidateDebtSub$description_aggr[is.na(CandidateDebtSub$description_aggr)] <- "OTHER"

rm(list = c("creditcard", "consulting", "swag"))
table(CandidateDebtSub$description_aggr)
```

```
##
##          ACCOUNTING                 BONUS CAMPAIGN MANAGEMENT
##                  79                    22                  10
##       CARRY FORWARD            CONSULTING         CREDIT CARD
##                  17                   130                  42
##        DATA/TECH/AD          DESIGN/PRINT         FUNDRAISING
##                  30                    36                  45
##                MAIL                 OTHER             POLLING
##                  14                    24                   5
##       REIMBURSEMENT                  SWAG            TREASURY
##                  54                   261                 218
```

```
# table(CandidateDebtSub$description[CandidateDebtSub$description_aggr
# == 'OTHER'])
```

Now we are ready to explore!

2. Univariate Analysis

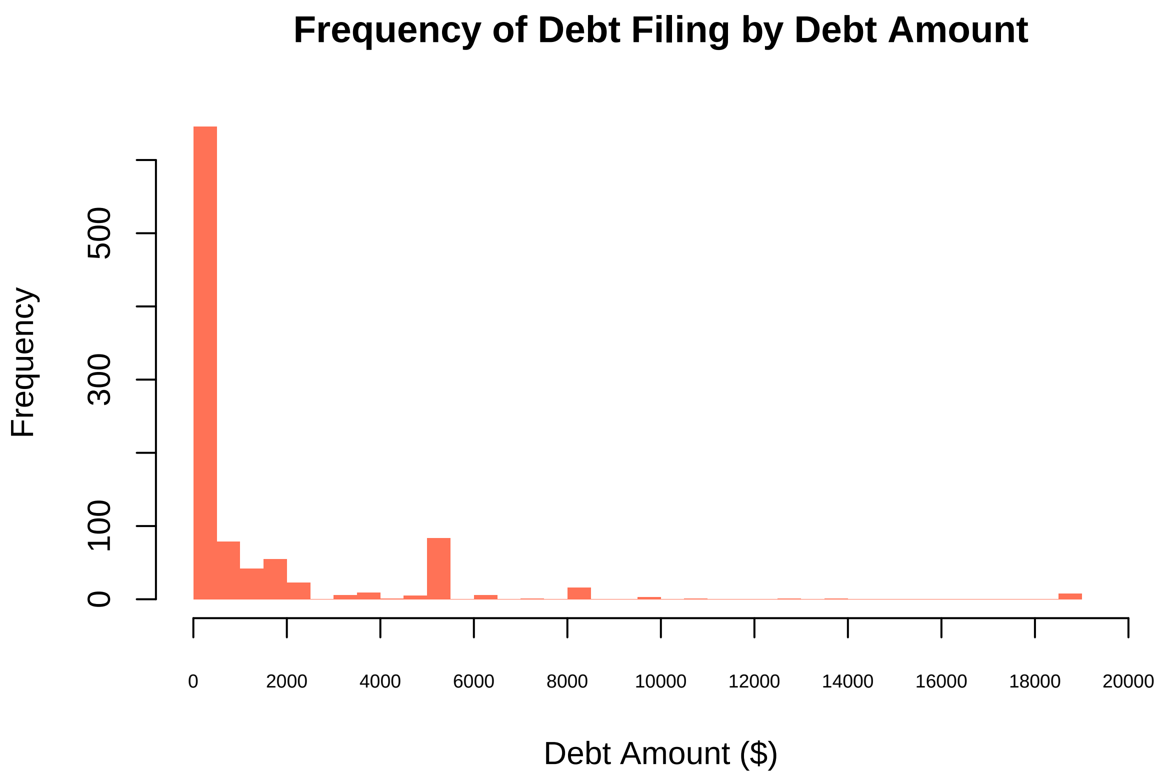
Univariate analysis was conducted on the variables that were not determined to have faulty data. The objective of this subset of analysis is to better understand the behavior of each variable and to identify specific variables that may be informative in a bivariate analysis. Specifically, the variables to be examined in this section are: - Amount: The amount of the debt incurred or order placed. - Office: The office sought by the candidate - WeeksinDebt - Code: The type of debt - description_aggr (Derived): A derived field categorizing the type of expense of the debt, based on the debt description field - weeksindebt (Derived): A derived field showing the length time the debt was held for, based on the debt date and

2.1 Univariate Analysis - Amount

```
summary(CandidateDebtSub$amount_num)
```

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

```
# Amount

hist(CandidateDebtSub$amount, breaks = 50, main = "Frequency of Debt Filing by Debt Amount",
    col = "coral1", border = NA, xlim = range(0:20000), xaxt = "n",
    xlab = "Debt Amount ($)")
axis(1, at = seq(0, 20000, 2000), labels = seq(0, 20000, 2000),
    cex.axis = 0.6)
```



The amounts associated with each filing are between $3.24 and $19,000, with the majority being less than $500.

```
CandidateDebtSub[CandidateDebtSub$amount_num >= 19000, c("office",
    "amount_num", "description")]
```

```
##                          office amount_num description
## 176            ATTORNEY GENERAL      19000  CONSULTING
## 360        STATE REPRESENTATIVE      19000  CONSULTING
## 366        STATE REPRESENTATIVE      19000  CONSULTING
## 378          COUNTY COMMISSIONER      19000  CONSULTING
## 462         SUPERIOR COURT JUDGE      19000  CONSULTING
## 492 STATE SUPREME COURT JUSTICE      19000  CONSULTING
## 549                STATE SENATOR      19000  CONSULTING
## 593        STATE REPRESENTATIVE      19000  CONSULTING
```

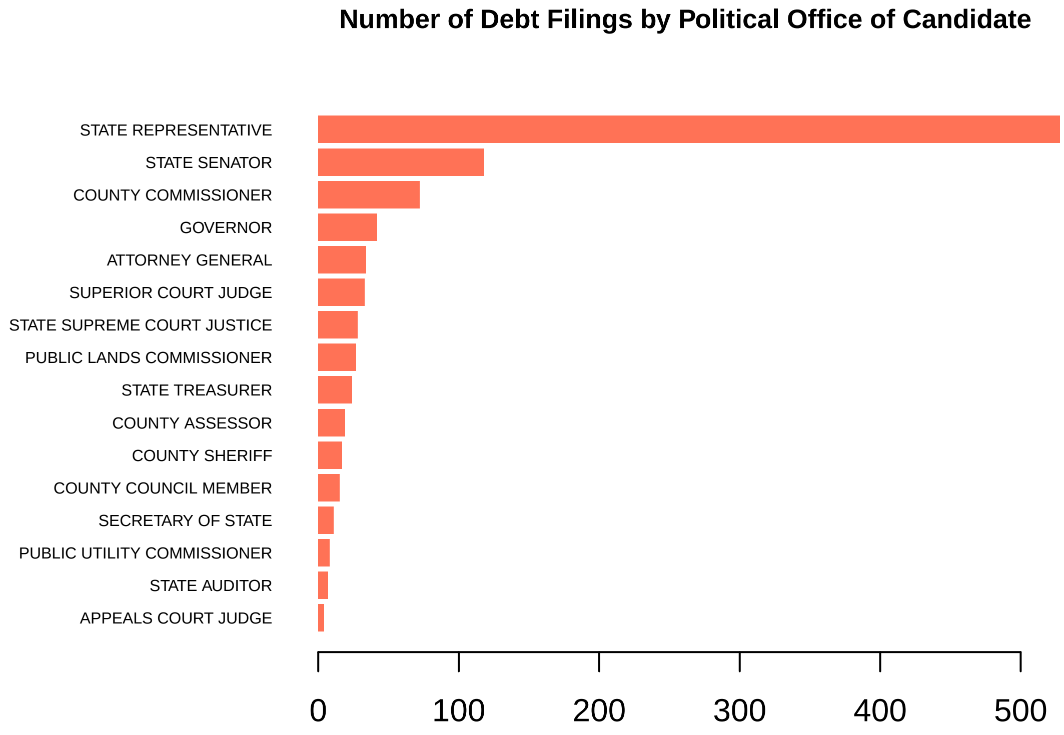
There are two notable observations when looking at a histogram of the variable: the outlier group of 8 filings of $19,000, and the large cluster of amounts just over $5,000.

2.2 Univariate Analysis - Office

YZ: this chart needs a legend

```
par(mar = c(4, 8, 4, 5))
barplot(sort(table(CandidateDebtSub$office)), horiz = TRUE, las = 1,
    cex.names = 0.5, col = "coral1", border = NA)
```

```
title("Number of Debt Filings by Political Office of Candidate",
    cex.main = 0.8)
```



The vast majority of filings in the existing dataset are from respondents running for or currently serving as State REpresentatives and State Senators. Given that there are many more seats for those positions, this finding is to be expected.

2.3 Univariate Analysis - Vendor State & Vendor City

```
table(CandidateDebtSub$vendorstate)
```

```
##
##       CA  DC  TX  WA
##   25  10 100   5 847
```

It appears there are 25 values for which there is no listed State for the deb holder. Looking at those values, they all appear to be associated with Credit Card debt.

```
DF <- as.data.frame(CandidateDebtSub)
table(DF$vendorname[DF$vendorstate == ""])
```

```
##
##        AMERICAN EXPRESS AMERICAN EXPRESS LOWES
##                      10                     11
##         CABELA'S (VISA)            CAPITOL ONE
##                       1                      1
##         CITI MASTERCARD
##                       2
```

```
rm(DF)
```

```
# table(CandidateDebtSub$vendorcity)
aggr_city <- aggregate(reportnumber ~ vendorcity, CandidateDebtSub,
    length)
colnames(aggr_city) <- c("Vendor City", "Number of Reports")
aggr_city[order(-aggr_city$`Number of Reports`), ]
```

```
##          Vendor City Number of Reports
## 21           SEATTLE               452
## 30      WOODINVILLE                241
## 29        WASHINGTON               100
## 10          KIRKLAND                38
## 1                                   24
## 26            TACOMA                24
## 14           OLYMPIA                20
## 3            BELFAIR                13
## 28  UNIVERSITY PLACE                11
## 15      PORT ORCHARD                10
## 13       OAK HARBOR                  6
## 5             DALLAS                 5
## 6      FRIDAY HARBOR                 5
## 7         GIG HARBOR                 5
## 17          PUYALLUP                 5
## 2  BAINBRIDGE ISLAND                 3
## 12     MOUNTAIN VIEW                 3
## 18        SACRAMENTO                 3
## 20      SANTA MONICA                 3
## 22            SEQUIM                 3
## 9          KENNEWICK                 2
## 23           SHELTON                 2
## 24           SPOKANE                 2
## 4   CITY OF INDUSTRY                 1
## 8           ISSAQUAH                 1
## 11     MERCER ISLAND                 1
## 16     PORT TOWNSEND                 1
## 19          SAN JOSE                 1
## 25    SPOKANE VALLEY                 1
## 27          TUMWATER                 1
```

```
rm(aggr_city)
```

2.4 Univariate Analysis - Code

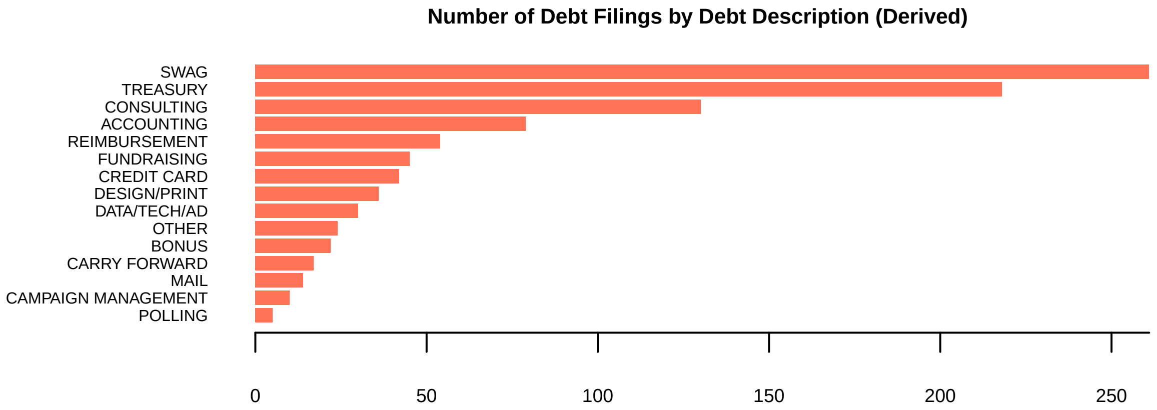
```
# Code
aggr_code <- aggregate(reportnumber ~ code, CandidateDebtSub,
    length)
colnames(aggr_code) <- c("Code", "Number of Reports")
aggr_code[order(-aggr_code$`Number of Reports`), ]
```

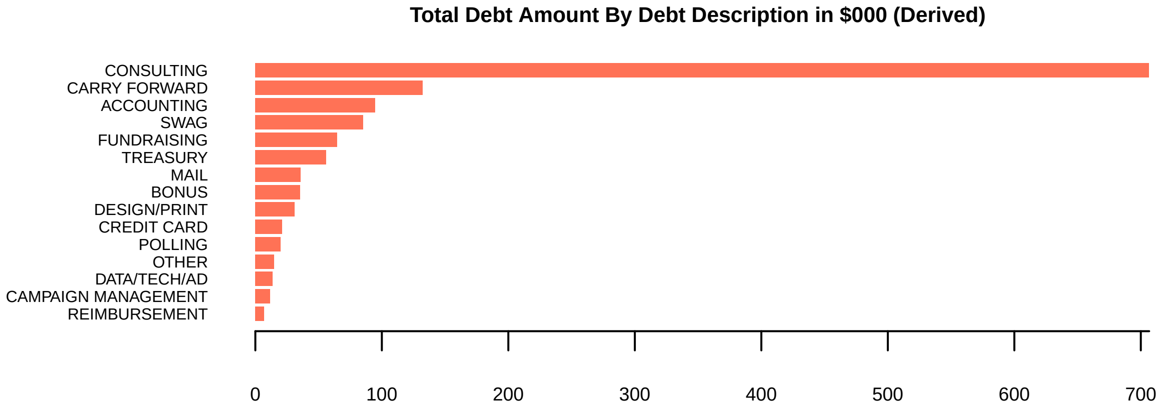
```
##                      Code Number of Reports
## 1                                       610
## 4 Operation and Overhead                362
## 3    Management Services                 10
## 2            Fundraising                  5
```

```
rm(aggr_code)
```

2.5 Univariate Analysis - description_aggr

```
# description_aggr

aggr_descr <- aggregate(amount_num ~ description_aggr, data = CandidateDebtSub,
    sum)
aggr_descr <- aggr_descr[order(aggr_descr$amount_num), ]

par(mar = c(2, 7, 2, 2), mfrow = c(2, 1))
barplot(sort(table(CandidateDebtSub$description_aggr)), horiz = TRUE,
    las = 2, cex.names = 0.5, col = "coral1", border = NA, axes = FALSE)
axis(1, at = seq(0, 300, 50), cex.axis = 0.6, las = 1)
title("Number of Debt Filings by Debt Description (Derived)",
    cex.main = 0.7)

# par(mar=c(4,12,4,4))
barplot(aggr_descr$amount_num, names.arg = aggr_descr$description_aggr,
    horiz = TRUE, las = 2, cex.names = 0.5, axes = FALSE, col = "coral1",
    border = NA)
axis(1, at = seq(0, 8e+05, 1e+05), cex.axis = 0.6, labels = seq(0,
    800, 100), las = 1)
title("Total Debt Amount By Debt Description in $000 (Derived)",
    cex.main = 0.7)
```

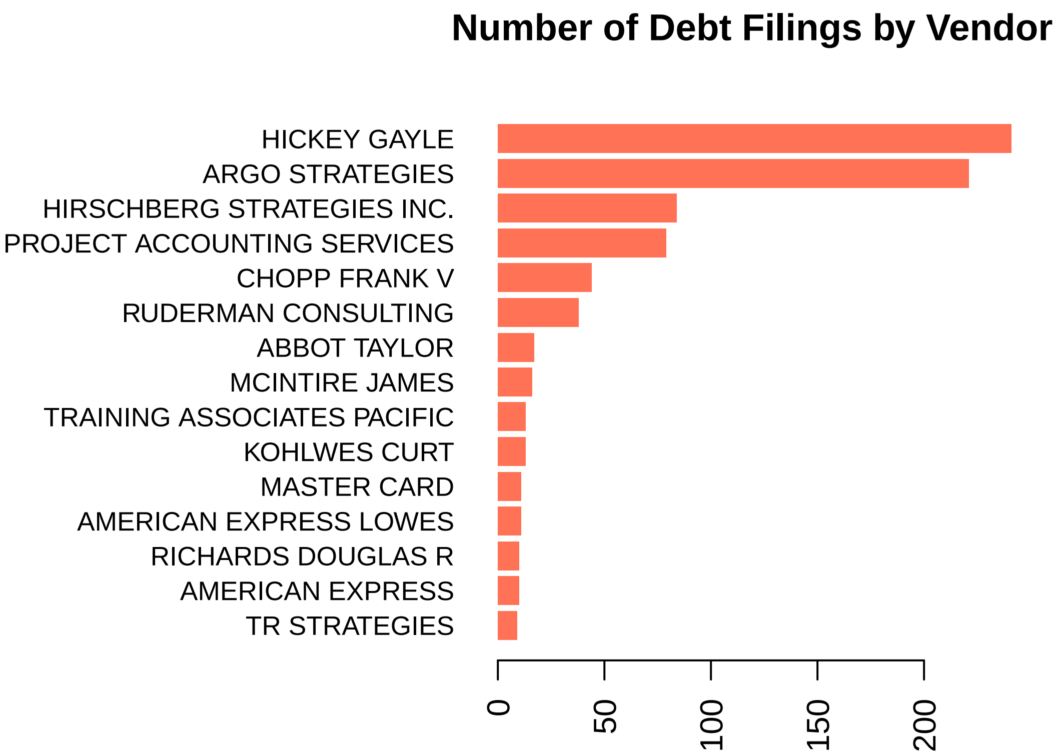




```
rm(aggr_descr)
```

2.6 Univariate Analysis - Vendor

```
par(mar = c(4, 15, 4, 4))
vendorTable <- sort(table(CandidateDebtSub$vendorname), decreasing = TRUE)
barplot(sort(vendorTable[1:15]), horiz = TRUE, las = 2, cex.names = 0.8,
    main = "Number of Debt Filings by Vendor", col = "coral1",
    border = NA)
```



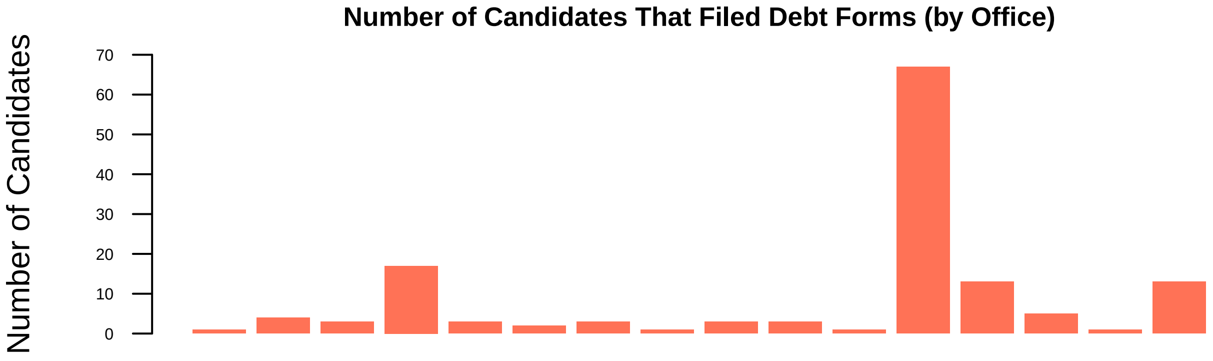
3. Analysis of Key Relationships

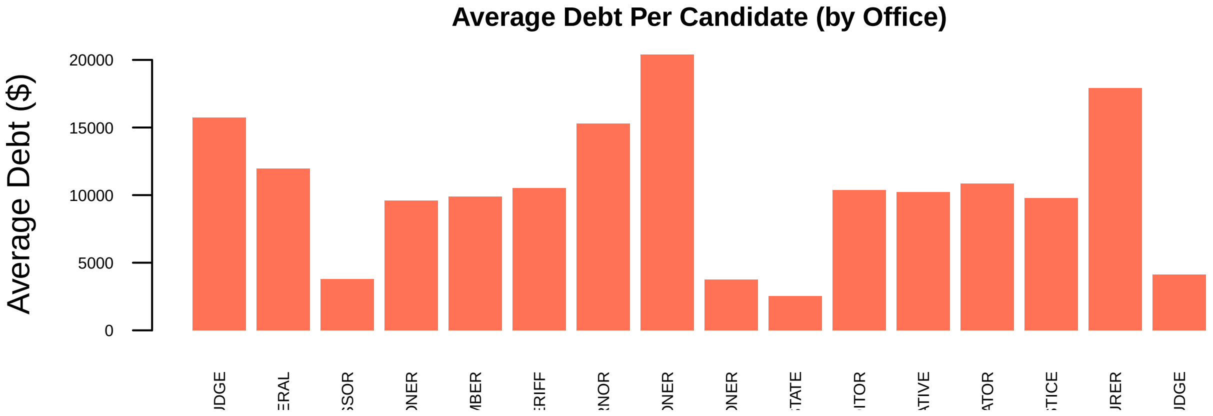
3.1 Average debt per candidate

Let's first look at how many candidates within each office filed debt reports.

```
aggr_office <- aggregate(amount_num ~ filerid + office, data = CandidateDebtSub, sum)
aggr_office <- aggregate(filerid ~ office, data = aggr_office, length)
aggr_office2 <- aggregate(amount_num ~ office, data = CandidateDebtSub, sum)
aggr_office <- cbind(aggr_office, aggr_office2[,2])
colnames(aggr_office)[3] <- c("amount_num")
aggr_office$amount_p_cand <- aggr_office$amount_num / aggr_office$filerid

par(mar = c(2,4,2,0),
    #oma = c(0,0,0,0),
    mfrow = c(2,1))
barplot(aggr_office$filerid,
        #names.arg = aggr_office$office,
        cex.names = 0.5,
        cex.axis = 0.5,
        border = NA,
        las = 2,
        ylim = range(0, 70),
        ylab = "Number of Candidates",
        col = "coral1")
```

```
title("Number of Candidates That Filed Debt Forms (by Office)",
      cex.main = 0.8)

barplot(aggr_office$amount_p_cand,
        names.arg = aggr_office$office,
        cex.names = 0.5,
        cex.axis = 0.5,
        border = NA,
        las = 2,
        ylab = "Average Debt ($)",
        col = "coral1")
title("Average Debt Per Candidate (by Office)",
      cex.main = 0.8)
```



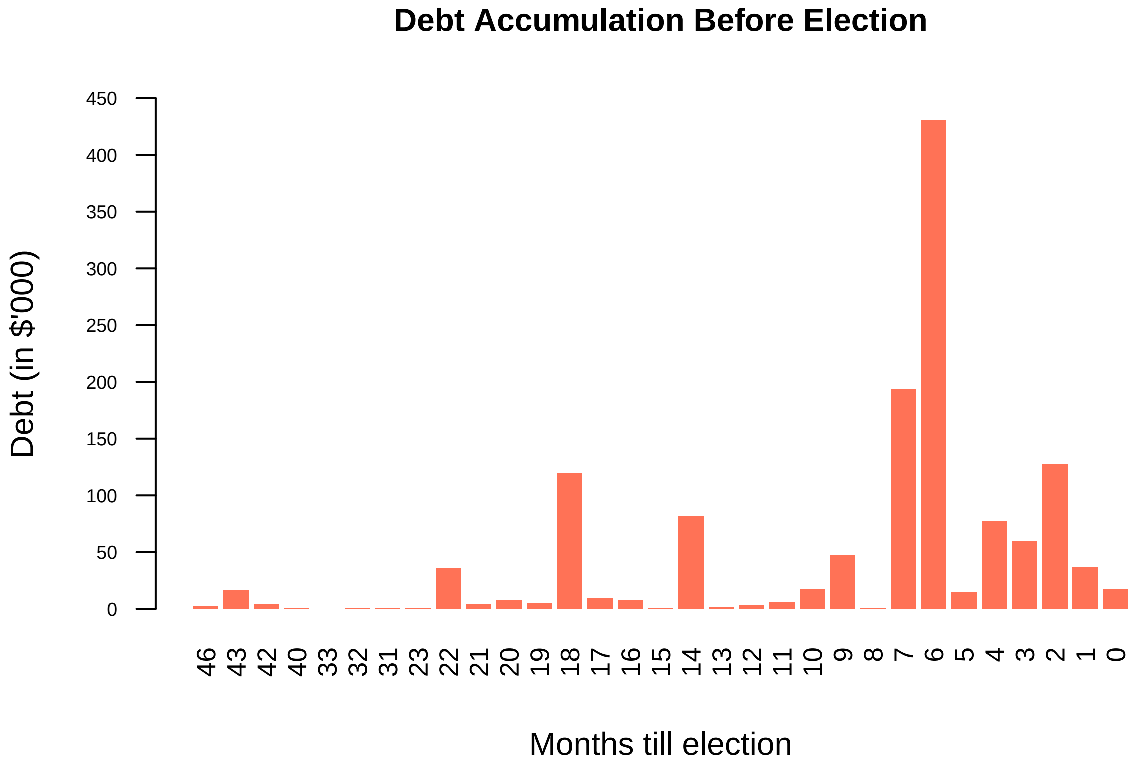


```
rm(list = c("aggr_office", "aggr_office2"))
```

Based on the above, “State Representative” had by far the largest number of candidates in this election. “County Commissioner”, “State Senator” and “Superior Court Judge” had similar number of candidates (between 15 and 20).

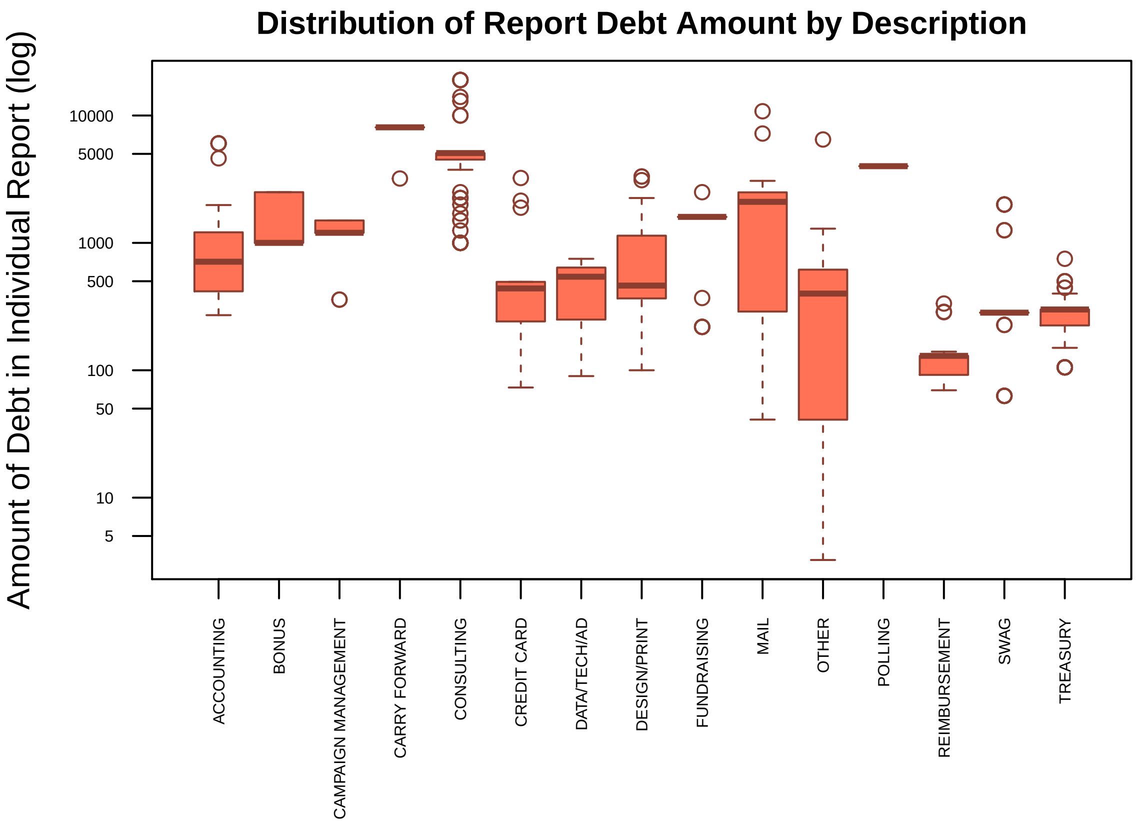
3.2 Timing Analysis of the Debt

```
aggr_months <- aggregate(amount_num ~ monthsindebt, data = CandidateDebtSub,
    sum)
aggr_months <- aggr_months[order(-aggr_months$monthsindebt),
    ]

barplot(aggr_months$amount_num/1000, names.arg = aggr_months$monthsindebt,
    cex.names = 0.8, cex.axis = 0.8, ylim = range(0, 450), las = 2,
    border = NA, axes = FALSE, xlab = "Months till election",
    ylab = "Debt (in $'000)", col = "coral1")
axis(2, at = seq(0, 450, 50), cex.axis = 0.6, las = 1)
title("Debt Accumulation Before Election", cex.main = 1)
```



```
rm(aggr_months)
```

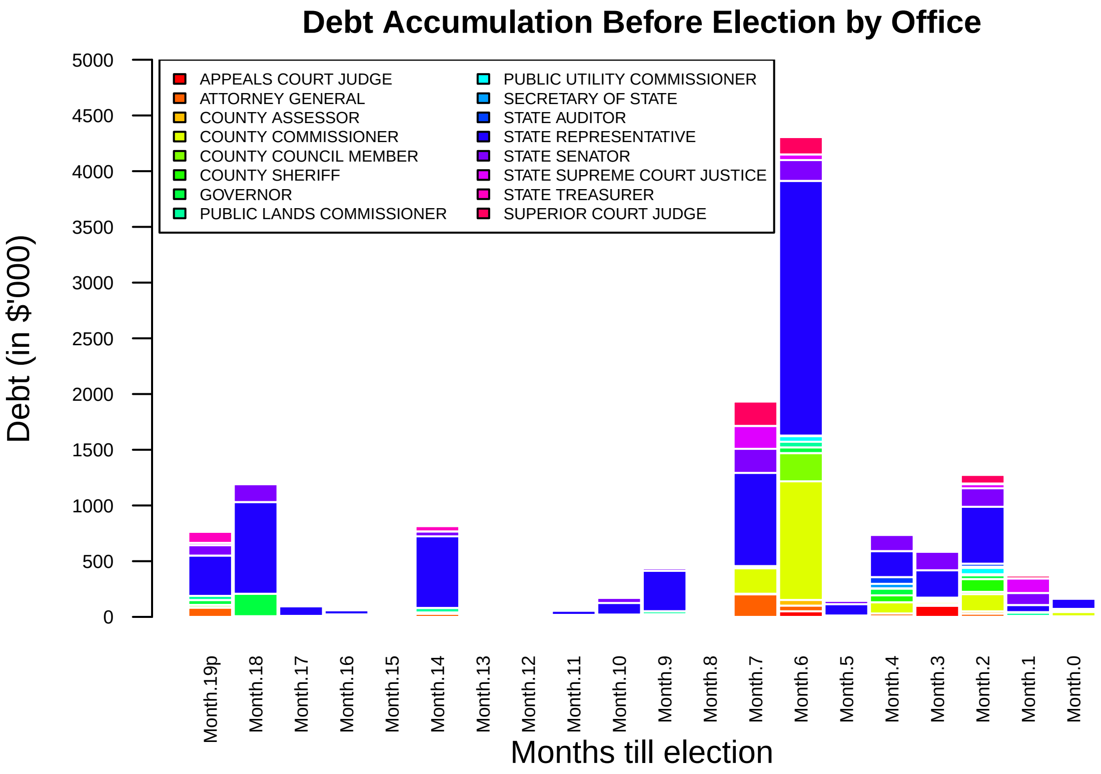
```
aggr_descr0 <- aggregate(amount_num ~ description_aggr, data = CandidateDebtSub,
    sum)
aggr_descr0 <- aggr_descr0[order(-aggr_descr0$amount_num), ]

par(mar = c(7, 4, 2, 3), oma = c(0, 0, 0, 0), xpd = TRUE)
boxplot(amount_num ~ description_aggr, data = CandidateDebtSub,
    log = "y", ylab = "Amount of Debt in Individual Report (log)",
    las = 2, cex.names = 0.5, cex.axis = 0.5, col = "coral1",
    border = "coral4")
title("Distribution of Report Debt Amount by Description", cex.main = 1)
```



4. Analysis of Secondary Effects

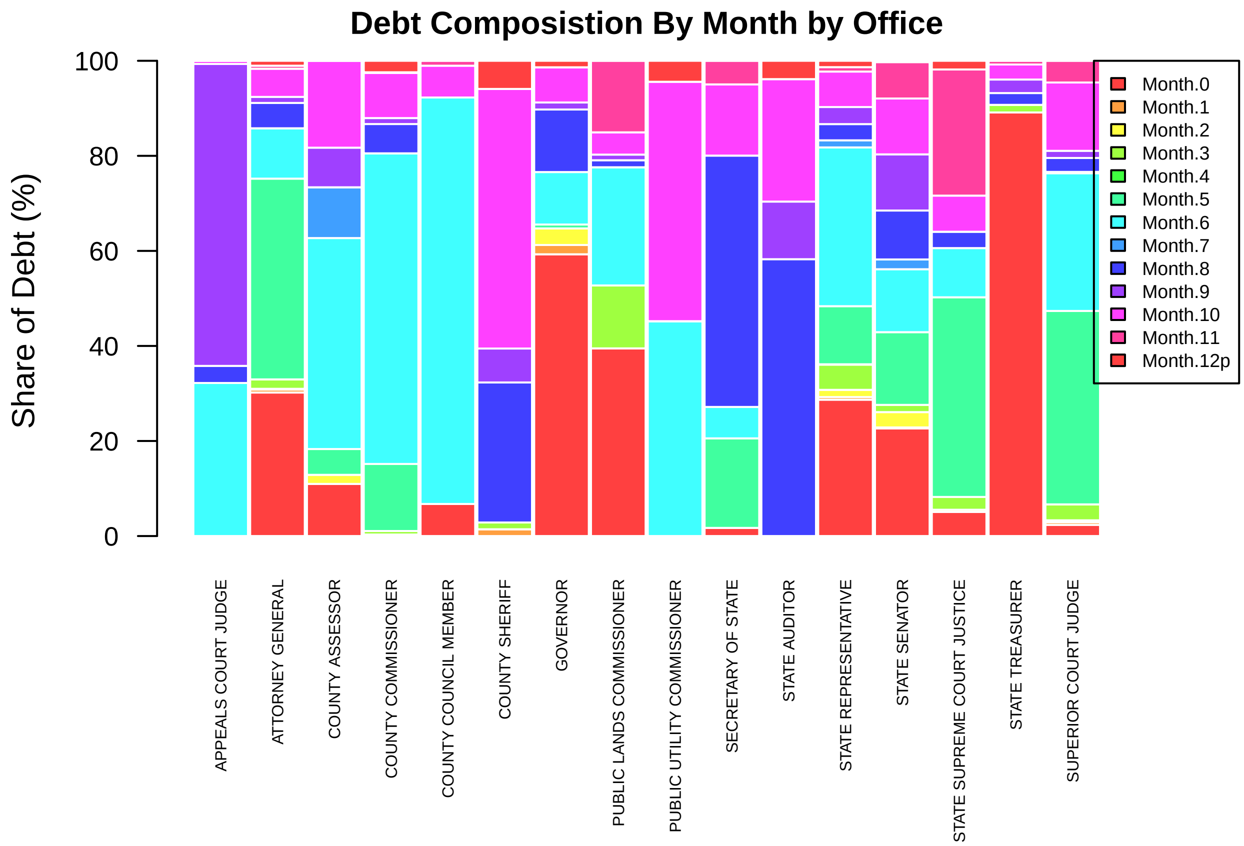
4.1 Timing Analysis of the Debt by Office

```
aggr_months_office <- aggregate(amount_num ~ monthsindebt + office,
    data = CandidateDebtSub, sum)
aggr_months_office <- reshape(aggr_months_office, v.names = "amount_num",
    idvar = "office", timevar = "monthsindebt", direction = "wide")

aggr_months_office[is.na(aggr_months_office)] <- 0

aggr_months_office$amount_num.19plus = aggr_months_office$amount_num.19 +
    aggr_months_office$amount_num.20 + aggr_months_office$amount_num.21 +
    aggr_months_office$amount_num.22 + aggr_months_office$amount_num.23 +
    aggr_months_office$amount_num.31 + aggr_months_office$amount_num.32 +
    aggr_months_office$amount_num.33 + aggr_months_office$amount_num.40 +
    aggr_months_office$amount_num.42 + aggr_months_office$amount_num.43 +
    aggr_months_office$amount_num.46

keep_vars <- c("office", "amount_num.19plus", "amount_num.18",
    "amount_num.17", "amount_num.16", "amount_num.15", "amount_num.14",
    "amount_num.13", "amount_num.12", "amount_num.11", "amount_num.10",
    "amount_num.9", "amount_num.8", "amount_num.7", "amount_num.6",
    "amount_num.5", "amount_num.4", "amount_num.3", "amount_num.2",
    "amount_num.1", "amount_num.0")
new_vars <- c("office", "Month.19p", "Month.18", "Month.17",
    "Month.16", "Month.15", "Month.14", "Month.13", "Month.12",
    "Month.11", "Month.10", "Month.9", "Month.8", "Month.7",
    "Month.6", "Month.5", "Month.4", "Month.3", "Month.2", "Month.1",
    "Month.0")

aggr_months_office <- aggr_months_office[, keep_vars]
colnames(aggr_months_office) <- new_vars
aggr_months_office2 <- aggr_months_office[, -1]
rownames(aggr_months_office2) <- aggr_months_office[, 1]

par(mar = c(4, 4, 2, 1), oma = c(1, 1, 1, 1))
barplot(as.matrix(aggr_months_office2), border = "white", space = 0.04,
    cex.names = 0.6, las = 2, cex.axis = 0.6, col = rainbow(16),
    axes = FALSE, ylim = range(0, 5e+05), xlab = "Months till election",
    ylab = "Debt (in $'000)")
axis(2, at = seq(0, 5e+05, 50000), cex.axis = 0.6, labels = seq(0,
    5000, 500), las = 1)
legend("topright", legend = rownames(aggr_months_office2), fill = rainbow(16),
    inset = c(0.365, 0), ncol = 2, cex = 0.5)
title("Debt Accumulation Before Election by Office", cex.main = 1)
```



```
rm(list = c("keep_vars", "new_vars", "aggr_months_office2"))
```

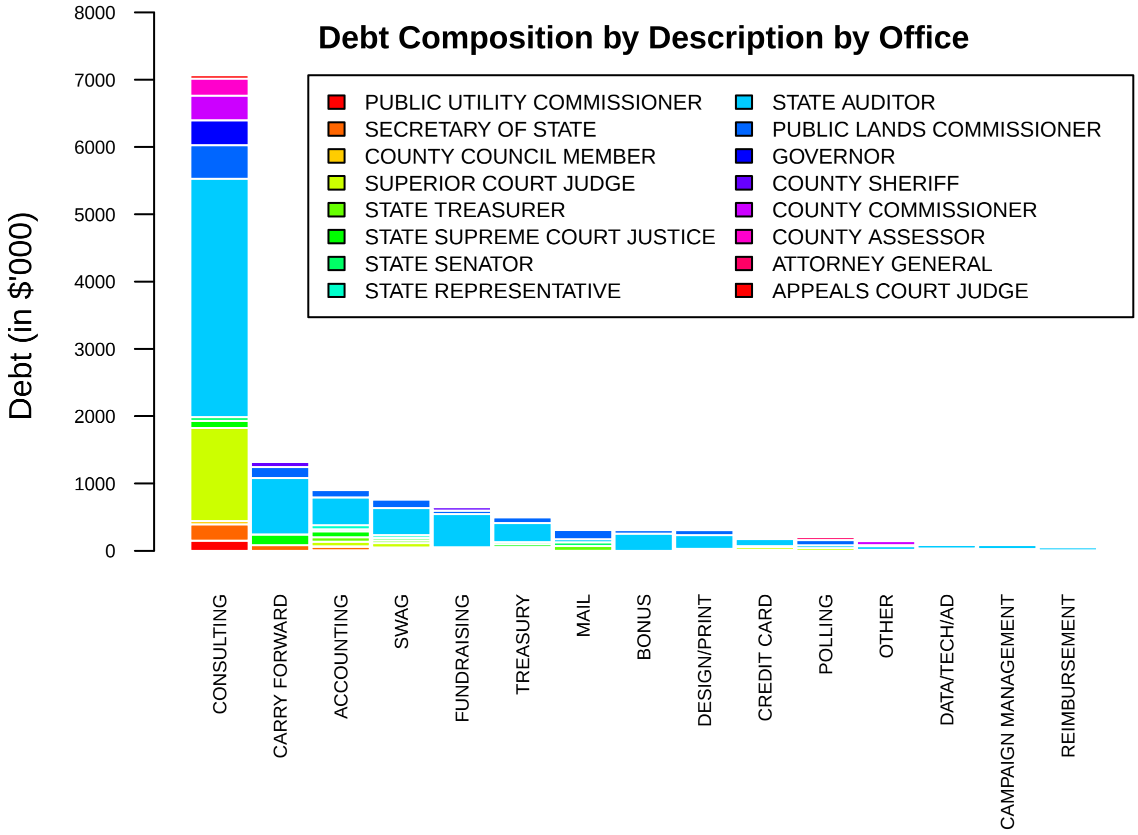
```r
amount_by_office <- aggregate(amount_num ~ office, data = CandidateDebtSub,
    sum)
aggr_months_office <- cbind(aggr_months_office, amount_by_office[,
    2])
colnames(aggr_months_office)[22] <- "amount"
aggr_months_office$Month.12p <- 100 * (aggr_months_office$Month.19p +
    aggr_months_office$Month.18 + aggr_months_office$Month.17 +
    aggr_months_office$Month.16 + aggr_months_office$Month.15 +
    aggr_months_office$Month.14 + aggr_months_office$Month.13 +
    aggr_months_office$Month.12)/aggr_months_office$amount
aggr_months_office$Month.11 <- 100 * aggr_months_office$Month.11/aggr_months_office$amount
aggr_months_office$Month.10 <- 100 * aggr_months_office$Month.10/aggr_months_office$amount
aggr_months_office$Month.9 <- 100 * aggr_months_office$Month.9/aggr_months_office$amount
aggr_months_office$Month.8 <- 100 * aggr_months_office$Month.8/aggr_months_office$amount
aggr_months_office$Month.7 <- 100 * aggr_months_office$Month.7/aggr_months_office$amount
aggr_months_office$Month.6 <- 100 * aggr_months_office$Month.6/aggr_months_office$amount
aggr_months_office$Month.5 <- 100 * aggr_months_office$Month.5/aggr_months_office$amount
aggr_months_office$Month.4 <- 100 * aggr_months_office$Month.4/aggr_months_office$amount
aggr_months_office$Month.3 <- 100 * aggr_months_office$Month.3/aggr_months_office$amount
aggr_months_office$Month.2 <- 100 * aggr_months_office$Month.2/aggr_months_office$amount
aggr_months_office$Month.1 <- 100 * aggr_months_office$Month.1/aggr_months_office$amount
aggr_months_office$Month.0 <- 100 * aggr_months_office$Month.0/aggr_months_office$amount

new_vars <- c("Month.12p", "Month.11", "Month.10", "Month.9",
    "Month.8", "Month.7", "Month.6", "Month.5", "Month.4", "Month.3",
    "Month.2", "Month.1", "Month.0")

aggr_months_office2 <- aggr_months_office[, new_vars]
aggr_months_office3 <- as.data.frame(t(aggr_months_office2))
colnames(aggr_months_office3) <- aggr_months_office[, 1]

par(mar = c(8, 4, 2, 3), oma = c(0, 0, 0, 0), xpd = TRUE)
barplot(as.matrix(aggr_months_office3), border = "white", space = 0.04,
    cex.names = 0.5, las = 2, cex.axis = 0.8, col = rainbow(12,
        s = 0.75), ylab = "Share of Debt (%)")
legend("topright", legend = rev(rownames(aggr_months_office3)),
    fill = rainbow(12, s = 0.75), inset = c(-0.105, 0), ncol = 1,
    cex = 0.6)
title("Debt Composistion By Month by Office", cex.main = 1)
```



```
rm(list = c("new_vars", "aggr_months_office", "aggr_months_office2",
    "aggr_months_office3", "amount_by_office"))
```

4.2 Composition of Debt by Office by Description

```
aggr_descr0 <- aggregate(amount_num ~ description_aggr, data = CandidateDebtSub,
    sum)
aggr_descr0 <- aggr_descr0[order(-aggr_descr0$amount_num), ]
aggr_descr <- aggregate(amount_num ~ office + description_aggr,
    data = CandidateDebtSub, sum)
aggr_descr_office <- reshape(aggr_descr, v.names = "amount_num",
    idvar = "office", timevar = "description_aggr", direction = "wide")
aggr_descr_office[is.na(aggr_descr_office)] <- 0

colnames(aggr_descr_office) <- sub("amount_num.", "", colnames(aggr_descr_office))
aggr_descr_office2 <- aggr_descr_office[, aggr_descr0[, 1]]
rownames(aggr_descr_office2) <- aggr_descr_office[, 1]

par(mar = c(8, 4, 2, 3), oma = c(0, 0, 0, 0), xpd = TRUE)
barplot(as.matrix(aggr_descr_office2), border = "white", space = 0.04,
    cex.names = 0.6, las = 2, cex.axis = 0.6, col = rainbow(15),
    axes = FALSE, ylab = "Debt (in $'000)")
axis(2, at = seq(0, 8e+05, 1e+05), labels = seq(0, 8000, 1000),
    cex.axis = 0.6, las = 1)
legend("topright", legend = rev(rownames(aggr_descr_office2)),
    fill = rainbow(15), ncol = 2, cex = 0.7)
title("Debt Composition by Description by Office", cex.main = 1)
```



```
rm(list = c("aggr_descr0", "aggr_descr", "aggr_descr_office",
    "aggr_descr_office2"))
```